Semi-supervised Speaker Identification under Covariate Shift

Makoto Yamada (myamada0321@gmail.com)
Department of Computer Science,
Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
and
Department of Statistical Science,
The Graduate University for Advanced Studies
4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan

Masashi Sugiyama (sugi@cs.titech.ac.jp)
http://sugiyama-www.cs.titech.ac.jp/∼sugi/
Department of Computer Science,
Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan

Tomoko Matsui (tmatsui@ism.ac.jp)
Department of Statistical Modeling,
The Institute of Statistical Mathematics
4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan

Abstract

In this paper, we propose a novel semi-supervised speaker identification method that can alleviate the influence of non-stationarity such as session dependent variation, the recording environment change, and physical conditions/emotions. We assume that the voice quality variants follow the *covariate shift* model, where only the voice feature distribution changes in the training and test phases. Our method consists of weighted versions of kernel logistic regression and cross validation and is theoretically shown to have the capability of alleviating the influence of covariate shift. We experimentally show through text-independent/dependent speaker identification simulations that the proposed method is promising in dealing with variations in voice quality.

Keywords

Speaker identification, covariate shift, semi-supervised learning, kernel logistic regression, importance estimation.

1 Introduction

Speaker identification methods are widely used in various real-world situations such as access control of information service systems and speaker detection in speech dialog and speaker indexing problems with large audio archives [1]. Recently, the speaker identification and indexing problems in meeting attracted a great deal of attention.

Popular methods of text-independent speaker identification are based on the Gaussian mixture model (GMM) [2] or kernel methods such as the support vector machine (SVM) [3, 4]. In these supervised learning methods, it is implicitly assumed that training and test data follow the same probability distribution. However, since the speech features vary over time due to session dependent variation, the recording environment change, and physical conditions/emotions, the training and test distributions are not necessarily the same in practice. In the paper [5], the influence of the session dependent variation of voice quality in speaker identification problems has been investigated and the identification performance was shown to decrease significantly over 3 months—the major cause for the performance degradation was the voice source characteristic variations.

To alleviate the influence of session dependent variation, it is popular to use several sessions of speaker utterance samples [6, 7] or to use *cepstral mean normalization* (CMN) [8]. However, gathering several sessions of speaker utterance data and assigning the speaker ID to the collected data are expensive both in time and cost and therefore not realistic in practice. Moreover, it is not possible to perfectly remove the session dependent variation by CMN alone.

A more practical/effective setup would be *semi-supervised learning*, where unlabeled samples are additionally given from the testing environment. In semi-supervised learning, it is required that the probability distributions of training and test are related to each other in some sense; otherwise we may not be able to learn anything about the test probability distribution from the training samples. A common modeling assumption is called *covariate shift*, where the input (feature) probability distributions are different in the training and test phases but the conditional probability distribution of labels remains unchanged. In many real-world applications such as robot control [9, 10, 11], bioinformatics [12, 13], spam filtering [14], natural language processing [15, 16], brain-computer interfacing [17, 18], and econometrics [19], the covariate shift model has been shown to be useful. Covariate shift is also naturally induced in selective sampling or active learning scenarios [20, 21, 22, 23, 24]. For this reason, learning under covariate shift is receiving a great deal of attention these days in the machine learning community [25].

In this paper, we formulate the semi-supervised speaker identification problem in the covariate shift framework and propose a method that can cope with voice quality variants. Under covariate shift, standard maximum likelihood estimation is no longer consistent. The influence of covariate shift can be asymptotically canceled by weighting the log-likelihood terms according to the *importance* [26]:

$$w(\mathrm{X}) = \frac{p_{te}(\mathrm{X})}{p_{tr}(\mathrm{X})},$$

where $p_{te}(\mathrm{X})$ and $p_{tr}(\mathrm{X})$ are test and training input densities. We apply this weighting idea in kernel logistic regression (KLR). The importance weight $w(\mathrm{X})$ is unknown in practice and needs to be estimated from data. For weight estimation, we utilize the Kullback-Leibler importance estimation procedure (KLIEP) [27, 28] since it is equipped with a built-in model selection procedure. The (regularized) kernel logistic regression model contain two tuning parameters: the kernel width and the regularization parameter. Usually those tuning parameters are optimized based on cross validation (CV). However, ordinary CV is no longer unbiased due to covariate shift and therefore is not reliable as a model selection method. To cope with this problem, we use importance weighted CV [18] for unbiased model selection. The validity of our approach is experimentally shown through text-independent/dependent speaker identification simulations.

The rest of this paper is structured as follows. Section 2 formulates the semi-supervised speaker identification problem and review existing methods such as KLR and CV. In Section 3, importance weighting techniques for covariate shift adaptation are introduced. Experimental results are reported in Section 4. Section 5 concludes with a summary of our contributions and possible future work.

2 Problem Formulation

In this section, we formulate the speaker identification problem from a machine learning point of view.

2.1 Kernel-based Speaker Identification

An utterance feature X pronounced by a speaker is expressed as a set of N mel-frequency cepstrum coefficient (MFCC) [29] vectors of d dimensions:

$$\mathrm{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{d \times N}. \tag{1}$$

For training, we are given n_{tr} labeled utterance samples

$$\mathcal{Z}^{tr} = \{\mathrm{X}_i, y_i\}_{i=1}^{n_{tr}}, \tag{2}$$

where $y_i \in \{1, \ldots, K\}$ denotes the index of the speaker who pronounced X_i. The goal of speaker identification is to predict the speaker index of a test utterance sample X based on the training samples. We predict the speaker index c of the test sample X following the Bayes decision rule:

$$P(y = c|\mathrm{X}) > P(y = i|\mathrm{X}) \quad \forall\, i \neq c. \tag{3}$$

For approximating the class-posterior probability, we use the following parametric model $p(y = c|\mathrm{X}, \mathrm{V})$:

$$p(y = c|\mathrm{X}, \mathrm{V}) = \frac{\exp f_{\boldsymbol{v}_c}(\mathrm{X})}{\sum_{l=1}^{K} \exp f_{\boldsymbol{v}_l}(\mathrm{X})}, \tag{4}$$

where $\mathrm{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K]^\top \in \mathbb{R}^{K \times n_{tr}}$ is the parameter, $^\top$ denotes the transpose, and $f_{\boldsymbol{v}_l}$ is a discriminant function corresponding to the speaker l. This model is known as the softmax function and widely used in multiclass logistic regression. We use the following kernel regression model as the discriminant function $f_{\boldsymbol{v}_l}$ [7]:

$$f_{\boldsymbol{v}_l}(\mathrm{X}) = \sum_{i=1}^{n_{tr}} v_{l,i} \mathcal{K}(\mathrm{X}, \mathrm{X}_i) \quad l = 1, \ldots, K, \tag{5}$$

where $\boldsymbol{v}_l = (v_{l,1}, \ldots, v_{l,n_{tr}})^\top \in \mathbb{R}^{n_{tr}}$ are parameters corresponding the speaker l and $\mathcal{K}(\mathrm{X}, \mathrm{X}')$ is a kernel function. In this paper, we use the *sequence kernel* [4] as the kernel function since it allows us to handle features with different size; for two utterance samples $\mathrm{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{d \times N}$ and $\mathrm{X}' = [\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{N'}] \in \mathbb{R}^{d \times N'}$ (generally $N \neq N'$), the sequence kernel is defined as

$$\mathcal{K}(\mathrm{X}, \mathrm{X}') = \frac{1}{NN'} \sum_{i=1}^{N} \sum_{i'=1}^{N'} k(\boldsymbol{x}_i, \boldsymbol{x}'_{i'}), \tag{6}$$

where $k(\boldsymbol{x}, \boldsymbol{x}')$ is a vectorial kernel; we use the Gaussian kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(\frac{-\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2} \right). \tag{7}$$

Note that kernel logistic regression is a modeling assumption, thus the true class-conditional probability may not be exactly realized by the kernel logistic regression model. This implies that there exists a model error, i.e., even when the parameter is chosen optimally, there remains an approximation error. This setup is not of course preferable, but more or less there exists a model error in practice since it is not generally possible to have an exact model in reality. Traditional machine learning theories often assume that the model at hand is correct (i.e., no model error exists). However, this is not realistic and not useful in practice, so in this paper we explicitly take into account *model misspecification.*

2.2 Kernel Logistic Regression

Kernel logistic regression (KLR) is a kernelized variant of *logistic regression.* In KLR, we map the input vector to a high-dimensional space (feature space) and solve the logistic regression problem in the feature space; the similarity in feature space can be implicitly computed via the *kernel trick.* The kernel trick allows one to non-linearize a linear algorithm without sacrificing computational simplicity of the linear algorithm. Below, we briefly review KLR following the paper [30].

We employ maximum likelihood estimation for learning the parameter V. The negative log-likelihood function $\mathcal{P}_\delta^{\log}(\mathrm{V}; \mathcal{Z}^{tr})$ for the kernel logistic regression model is given by

$$\mathcal{P}_\delta^{\log}(\mathrm{V}; \mathcal{Z}^{tr}) = -\sum_{i=1}^{n_{tr}} \log P(y_i | \mathrm{X}_i, \mathrm{V}) + \frac{\delta}{2} \mathrm{trace}(\mathrm{V}\mathrm{K}\mathrm{V}^\top), \tag{8}$$

where $\mathrm{trace}(\mathrm{V}\mathrm{K}\mathrm{V}^\top)$ is a regularizer to avoid overfitting, δ is the regularization parameter that controls strength of regularization, and $\mathrm{K} = [\mathcal{K}(\mathrm{X}_i, \mathrm{X}_j)]_{i,j=1}^{n_{tr}}$ is the kernel Gram matrix. The negative log-likelihood function is convex and the unique minimizer can be obtained by, e.g., the Newton method. In the Newton method, the parameter matrix V is updated iteratively as

$$\mathrm{V} \leftarrow \mathrm{V} - \epsilon \Delta \mathrm{V}, \tag{9}$$

where ϵ is the step size and $\Delta\mathrm{V}$ is defined as

$$\mathrm{vec}\Delta\mathrm{V} = [\nabla^2 \mathcal{P}_\delta^{\log}(\mathrm{V}; \mathcal{Z})]^{-1} \mathrm{vec} \nabla \mathcal{P}_\delta^{\log}(\mathrm{V}; \mathcal{Z}). \tag{10}$$

'vec' denotes the vectorization operator, $\nabla \mathcal{P}_\delta^{\log}(\mathrm{V}; \mathcal{Z})$ is the gradient of Eq.(8) with respect to V, and $\nabla^2 \mathcal{P}_\delta^{\log}(\mathrm{V}; \mathcal{Z})$ is the Hessian of Eq.(8) with respect to V. The gradient and Hessian are given as

$$\nabla \mathcal{P}_\delta^{\log}(\mathrm{V}; \mathcal{Z}) = (\mathrm{P}(\mathrm{V}) - \mathrm{Y} + \delta \mathrm{V})\mathrm{K}, \tag{11}$$

$$\nabla^2 \mathcal{P}_\delta^{\log}(\mathrm{V}; \mathcal{Z}) = \sum_{i=1}^{n_{tr}} (\mathrm{diag}(\boldsymbol{p}(\mathrm{X}_i)) - \boldsymbol{p}(\mathrm{X}_i)\boldsymbol{p}(\mathrm{X}_i)^\top) \otimes \boldsymbol{k}(\mathrm{X}_i)\boldsymbol{k}(\mathrm{X}_i)^\top + (\mathrm{K}^\top \otimes \mathrm{I}), \tag{12}$$

where

$$\mathrm{P}(\mathrm{V}) = [\boldsymbol{p}(\mathrm{X}_1), \ldots, \boldsymbol{p}(\mathrm{X}_{n_{tr}})] \in \mathbb{R}^{K \times n_{tr}} \tag{13}$$

is a matrix whose n-th column is a vector of the class-posterior probabilities $\boldsymbol{p}(\mathrm{X}_n)$,

$$\boldsymbol{p}(\mathrm{X}) = [p(y = 1|\mathrm{X}, \mathrm{V}), \ldots, p(y = K|\mathrm{X}, \mathrm{V})]^\top \in \mathbb{R}^K \tag{14}$$

denotes the class-posterior probabilities for all classes given X,

$$\mathrm{Y} = [\boldsymbol{e}_{y^1}, \ldots, \boldsymbol{e}_{y^N}] \in \mathbb{R}^{K \times n_{tr}}, \tag{15}$$

whose n-th column $\boldsymbol{e}_{y^n}$ is a unit vector with all zeros except for element y^n being 1, $\mathrm{diag}(a, \ldots, b)$ denotes the diagonal matrix with diagonal elements $a, \ldots, b$,

$$\boldsymbol{k}(\mathrm{X}) = [\mathcal{K}(\mathrm{X}, \mathrm{X}_1), \ldots, \mathcal{K}(\mathrm{X}, \mathrm{X}_{n_{tr}})]^\top \in \mathbb{R}^{n_{tr}} \tag{16}$$

is a vector whose elements are given by the sequence kernel, $\otimes$ denotes the Kronecker product, and I denotes the identity matrix.

In order to estimate the update matrix $\Delta\mathrm{V}$, the inverse of the Hessian needs to be computed at every iteration. This is computationally expensive so we approximate $\Delta\mathrm{V}$ by the conjugate gradient method; an approximation $\widehat{\Delta\mathrm{V}}$ can be estimated by solving the following linear equation [30]:

$$\nabla^2 \mathcal{P}_\delta^{\log}(\mathrm{V}; \mathcal{Z}) \mathrm{vec} \widehat{\Delta\mathrm{V}} = \mathrm{vec} \nabla \mathcal{P}_\delta^{\log}(\mathrm{V}; \mathcal{Z}). \tag{17}$$

Substituting Eqs.(11) and (12) into Eq.(17) and using the transformation

$$\mathrm{vec}(\mathrm{ABC}) = (\mathrm{C}^\top \otimes \mathrm{A})\mathrm{vec}(\mathrm{B}), \tag{18}$$

we have

$$\sum_{i=1}^{n_{tr}} (\mathrm{diag}(\boldsymbol{p}(\mathrm{X}_i)) - \boldsymbol{p}(\mathrm{X}_i)\boldsymbol{p}(\mathrm{X}_i)^\top)\widehat{\Delta\mathrm{V}}\boldsymbol{k}(\mathrm{X}_i)\boldsymbol{k}(\mathrm{X}_i)^\top = (\mathrm{P}(\mathrm{V}) - \mathrm{Y} + \delta\mathrm{V})\mathrm{K}. \tag{19}$$

2.3 Model Selection in KLR

The above KLR method includes two tuning parameters: the Gaussian width σ and the regularization parameter δ. One of the popular approaches to model selection is cross validation (CV).

Let us divide the training set $\mathcal{Z}^{tr} = \{(\mathrm{X}_i, y_i)\}_{i=1}^{n_{tr}}$ into k disjoint non-empty subsets $\{\mathcal{Z}_i^{tr}\}_{i=1}^{k}$. Let $\widehat{y}_{\mathcal{Z}_j^{tr}}(\mathrm{X})$ be an estimate of a speaker of a test utterance sample X obtained from $\{\mathcal{Z}_i^{tr}\}_{i\neq j}$ (i.e., without $\mathcal{Z}_j^{tr}$). Then the k-fold CV (kCV) score is given by

$$\widehat{R}_{kCV}^{\mathcal{Z}^{tr}} = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{|\mathcal{Z}_j^{tr}|}\sum_{(\mathrm{X},y)\in\mathcal{Z}_j^{tr}} I(y = \widehat{y}_{\mathcal{Z}_j^{tr}}(\mathrm{X})), \tag{20}$$

where $|\mathcal{Z}_j^{tr}|$ is the number of samples in the subset $\mathcal{Z}_j^{tr}$ and $I(\cdot)$ denotes the indicator function.

2.4 KLR, CV, and Covariate Shift

Here, we show potential limitations of KLR and CV in the light of model misspecification.

The use of KLR and CV could be theoretically justified when the training utterance features and the test utterance features independently follow the *same* probability distribution with density $p(\mathrm{X})$ and the class label y follows the *common* conditional probability distribution $p(y|\mathrm{X})$ in the training and test phases. Indeed, if the above conditions are met, KLR is shown to be *consistent*, i.e., the learned parameter converges to the optimal value:

$$\lim_{n_{tr}\to\infty} \widehat{\mathrm{V}} = \mathrm{V}^*, \tag{21}$$

where $\widehat{\mathrm{V}}$ is the parameter learned by KLR and V^* is the optimal parameter that minimizes the expected prediction error for test samples:

$$\mathrm{V}^* = \underset{\mathrm{V}}{\mathrm{argmin}} \iint I(y = \widehat{y}(\mathrm{X}|\mathrm{V}))p(y|\mathrm{X})p(\mathrm{X})dyd\mathrm{X}. \tag{22}$$

$\widehat{y}(\mathrm{X}|\mathrm{V})$ is an estimate of speaker of an utterance feature X for parameter V. Also, when $p(\mathrm{X})$ and $p(y|\mathrm{X})$ are common in the training and test phases, kCV is (almost) *unbiased* [31]:

$$\mathrm{E}_{\mathcal{Z}^{tr}}\left[\widehat{R}_{kCV}^{\mathcal{Z}^{tr}} - R^{\mathcal{Z}^{tr}}\right] \approx 0, \tag{23}$$

where $\mathrm{E}_{\mathcal{Z}^{tr}}$ is the expectation over the training set $\mathcal{Z}^{tr}$ and $R^{\mathcal{Z}^{tr}}$ is the expected prediction error defined by

$$R^{\mathcal{Z}^{tr}} = \iint I(y = \widehat{y}(\mathrm{X}; \mathcal{Z}^{tr}))p(y|\mathrm{X})p(\mathrm{X})dyd\mathrm{X}. \tag{24}$$

$\widehat{y}(\mathrm{X}; \mathcal{Z}^{tr})$ is a learned function from the training set $\mathcal{Z}^{tr}$.

However, in practical speaker identification, speech features are not stationary due to time-dependent voice variation, the recording environment change, and physical conditions/emotion. Thus, the training and test feature distributions are not the same. Then, the above good theoretical properties are no longer true[1].

In this paper, we explicitly deal with such changing environment via the *covariate shift* model [26]—the input distributions change between the training and test phases, $p_{tr}(\mathrm{X}) \neq p_{te}(\mathrm{X})$, but the conditional distribution $p(y|\mathrm{X})$ remains unchanged.

3 Importance Weighting Techniques for Covariate Shift Adaptation

In this section, we show how to cope with covariate shift.

3.1 Parameter Learning and Model Selection under Covariate Shift

Here we show how KLR and CV could be extended and justified even under covariate shift.

3.1.1 Importance Sampling

In the absence of covariate shift, the expectation over test samples can be consistently estimated by the expectation over training samples since they are drawn from the same distribution. However, under covariate shift, the difference of input distributions should be explicitly taken into account. A basic technique for compensating for the distribution

[1] If the KLR model is exactly correct, consistency of KLR and almost unbiasedness of CV still holds even when the feature distributions change between the training and test stages. However, the correct model assumption is not satisfied in reality.

change is *importance sampling* [32], i.e., the expectation over training samples is weighted according to their importance in the test distribution. Indeed, for the importance weight

$$w(\mathrm{X}) = \frac{p_{te}(\mathrm{X})}{p_{tr}(\mathrm{X})}, \tag{25}$$

the expectation of some function $F(\mathrm{X})$ over the probability density $p_{te}(\mathrm{X})$ can be computed by

$$\mathrm{E}_{p_{te}(\mathrm{X})}[F(\mathrm{X})] = \int F(\mathrm{X}) p_{te}(\mathrm{X}) d\mathrm{X} = \int F(\mathrm{X}) w(\mathrm{X}) p_{tr}(\mathrm{X}) d\mathrm{X} = \mathrm{E}_{p_{tr}(\mathrm{X})}[F(\mathrm{X}) w(\mathrm{X})]. \tag{26}$$

3.1.2 Importance Weighted Kernel Logistic Regression

If the importance sampling technique is applied to LKR, we have the following importance weighted KLR (IWKLR) [26]:

$$\widetilde{\mathcal{P}}_{\delta}^{\log}(\mathrm{V}; \mathcal{Z}^{tr}) = -\sum_{i=1}^{n_{tr}} w(\mathrm{X}_i) \log P(y_i | \mathrm{X}_i, \mathrm{V}). \tag{27}$$

IWKLR is consistent even under covariate shift:

$$\lim_{n_{tr} \to \infty} \widetilde{\mathrm{V}} = \mathrm{V}^*, \tag{28}$$

where $\widetilde{\mathrm{V}}$ is the parameter learned by IWKLR and V^* is the optimal parameter that minimizes the expected prediction error for test samples:

$$\mathrm{V}^* = \operatorname*{argmin}_{\mathrm{V}} \iint I(y = \widehat{y}(\mathrm{X}|\mathrm{V})) p(y|\mathrm{X}) p_{te}(\mathrm{X}) dy d\mathrm{X}. \tag{29}$$

In practice, we may include a regularizer:

$$\widetilde{\mathcal{P}}_{\delta}^{\log}(\mathrm{V}; \mathcal{Z}^{tr}) = -\sum_{i=1}^{n_{tr}} w(\mathrm{X}_i) \log P(y_i | \mathrm{X}_i, \mathrm{V}) + \frac{\delta}{2} \mathrm{trace}(\mathrm{V}\mathrm{K}\mathrm{V}^\top), \tag{30}$$

where δ is the *regularization parameter*.

The Newton update rule for IWKLR is given by the same form as Eq.(9); the gradient and Hessian of (30) are given by

$$\nabla \mathcal{P}_{\delta}^{\log}(\mathrm{V}; \mathcal{Z}) = \{(\mathrm{P}(\mathrm{V}) - \mathrm{Y})\mathrm{W} + \delta \mathrm{V}\}\mathrm{K}, \tag{31}$$

$$\nabla^2 \mathcal{P}_{\delta}^{\log}(\mathrm{V}; \mathcal{Z}) = \sum_{i=1}^{n_{tr}} w(\mathrm{X}_i)(\mathrm{diag}(\boldsymbol{p}(\mathrm{X}_i)) - \boldsymbol{p}(\mathrm{X}_i)\boldsymbol{p}(\mathrm{X}_i)^\top) \otimes \boldsymbol{k}(\mathrm{X}_i)\boldsymbol{k}(\mathrm{X}_i)^\top + (\mathrm{K}^\top \otimes \mathrm{I}), \tag{32}$$

where

$$\mathrm{W} = \mathrm{diag}(w(\mathrm{X}_1), \ldots, w(\mathrm{X}_{n_{tr}})) \in \mathbb{R}^{n_{tr} \times n_{tr}}. \tag{33}$$

An approximation $\widetilde{\Delta \mathrm{V}}$ of the update factor is given as the solution of the following linear equation:

$$\sum_{i=1}^{n_{tr}} w(\mathrm{X}_i)(\mathrm{diag}(\boldsymbol{p}(\mathrm{X}_i)) - \boldsymbol{p}(\mathrm{X}_i)\boldsymbol{p}(\mathrm{X}_i)^\top) \widetilde{\Delta \mathrm{V}} \boldsymbol{k}(\mathrm{X}_i)\boldsymbol{k}(\mathrm{X}_i)^\top = \{(\mathrm{P}(\mathrm{V}) - \mathrm{Y})\mathrm{W} + \delta \mathrm{V}\}\mathrm{K}. \tag{34}$$

3.1.3 Importance Weighted Cross Validation

In a similar way as IWKLR, CV could also be enhanced based on the importance weighting technique [18]:

$$\widetilde{R}^{\mathcal{Z}^{tr}}_{kIWCV} = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{|\mathcal{Z}^{tr}_j|}\sum_{(\mathrm{X},y)\in\mathcal{Z}^{tr}_j} w(\mathrm{X})I(y = \widetilde{y}_{\mathcal{Z}^{tr}_i}(\mathrm{X})). \tag{35}$$

We refer to this method as k-fold importance-weighted CV (kIWCV). Even under covariate shift, kIWCV is almost unbiased:

$$\mathrm{E}_{\mathcal{Z}^{tr}}\left[\widetilde{R}^{\mathcal{Z}^{tr}}_{kIWCV} - R^{\mathcal{Z}^{tr}}\right] \approx 0. \tag{36}$$

3.2 Importance Weight Estimation

As shown above, the importance weight $w(\mathrm{X})$ plays a central role in covariate shift adaptation. However, the importance weight is usually unknown, thus it needs to be estimated from samples. Here, we assume that in addition to the training input samples $\mathcal{X}^{tr} = \{\mathrm{X}_i\}_{i=1}^{n_{tr}}$, we are given (unlabeled) test samples $\mathcal{X}^{te} = \{\mathrm{X}_i\}_{i=1}^{n_{te}}$ drawn independently from $p_{te}(\mathrm{X})$ (i.e., the semi-supervised setup).

Under this setup, the importance weight may be simply approximated by estimating $p_{tr}(\mathrm{X})$ and $p_{te}(\mathrm{X})$ from training and test samples separately and then taking their ratio. However, density estimation is known to be a hard problem and taking the ratio of estimated quantities tends to magnify the estimation error. Thus this two-shot process is not reliable in practice. Below, we introduce a method that allows us to directly learn the importance weight function without going through density estimation. The method is called the *Kullback Leibler Importance Estimation Procedure (KLIEP)* [27, 28].

3.2.1 Direct Importance Weight Estimation

Let us model the importance function $w(\mathrm{X})$ by the following linear model:

$$\widehat{w}(\mathrm{X}) = \sum_{l=1}^{b}\alpha_l\varphi(\mathrm{X},\mathrm{C}_l), \tag{37}$$

where $\{\alpha_l\}_{l=1}^{b}$ are parameters to be learned from data samples, C_l is a template point randomly chosen from the test input set $\{\mathrm{X}_i\}_{i=1}^{n_{te}}$, and $\varphi(\mathrm{X},\mathrm{X}')$ is a basis function chosen as

$$\varphi(\mathrm{X},\mathrm{X}') = \frac{1}{NN'}\sum_{i=1}^{N}\sum_{i'=1}^{N'}k(\boldsymbol{x}_i,\boldsymbol{x}'_{i'}). \tag{38}$$

We use the Gaussian kernel for $k(\boldsymbol{x},\boldsymbol{x}')$:

$$k(\boldsymbol{x},\boldsymbol{x}') = \exp\left(\frac{-\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\tau^2}\right). \tag{39}$$

Using the model $\widehat{w}(\mathrm{X})$, we can estimate the test input density $p_{te}(\mathrm{X})$ by

$$\widehat{p}_{te}(\mathrm{X}) = \widehat{w}(\mathrm{X})p_{tr}(\mathrm{X}). \tag{40}$$

Based on this, we estimate $\{\alpha_l\}_{l=1}^{b}$ so that the Kullback-Leibler divergence from $p_{te}(\mathrm{X})$ to $\widehat{p}_{te}(\mathrm{X})$ minimized:

$$\begin{aligned} KL[p_{te}(\mathrm{X})\|\widehat{p}_{te}(\mathrm{X})] &= \int p_{te}(\mathrm{X}) \log \frac{p_{te}(\mathrm{X})}{p_{tr(\mathrm{X})}\widehat{w}(\mathrm{X})} d\mathrm{X} \\ &= \int p_{te}(\mathrm{X}) \log \frac{p_{te}(\mathrm{X})}{p_{tr}(\mathrm{X})} d\mathrm{X} - \int p_{te}(\mathrm{X}) \log \widehat{w}(\mathrm{X}) d\mathrm{X}. \end{aligned} \tag{41}$$

The first term in the above equation is independent of $\{\alpha_l\}_{l=1}^{b}$, thus it can be ignored and we concentrate on the second term; we define the second term as J_{KLIEP}:

$$J_{KLIEP} = \int p_{te}(\mathrm{X}) \log \widehat{w}(\mathrm{X}) d\mathrm{X} \approx \frac{1}{n_{te}} \sum_{\mathrm{X} \in \mathcal{X}^{te}} \log \widehat{w}(\mathrm{X}), \tag{42}$$

where the expectation over the test input distribution is approximated by the empirical average of test input samples. Since $\widehat{p}_{te}(\mathrm{X})$ is a probability density, the following equation should hold.

$$1 = \int \widehat{p}_{te}(\mathrm{X}) d\mathrm{X} = \int p_{tr}(\mathrm{X})\widehat{w}(\mathrm{X}) d\mathrm{X} \approx \frac{1}{n_{tr}} \sum_{\mathrm{X} \in \mathcal{X}^{tr}} \widehat{w}(\mathrm{X}), \tag{43}$$

where the expectation over the training input distribution is approximated by the empirical average of training input samples. We determine the coefficient $\{\alpha_l\}_{l=1}^{b}$ by solving the following optimization problem:

$$\begin{aligned} &\max_{\{\alpha_l\}_{l=1}^{b}} \quad \left[\sum_{\mathrm{X} \in \mathcal{X}^{te}} \log \left(\sum_{l=1}^{b} \alpha_l \varphi(\mathrm{X}, \mathrm{C}_l) \right) \right] \\ &\text{s.t} \quad \sum_{\mathrm{X} \in \mathcal{X}^{tr}} \sum_{l=1}^{b} \alpha_l \varphi(\mathrm{X}, \mathrm{C}_l) = n_{tr} \ \text{ and } \ \alpha_1, \ldots, \alpha_b \geq 0. \end{aligned} \tag{44}$$

This optimization problem is convex and thus the global solution can be obtained by simply performing gradient ascent and feasibility satisfaction iteratively. Note that the solution $\{\widehat{\alpha}_l\}_{l=1}^{b}$ tends to be sparse, which contributes to reducing the computational cost in the test phase.

3.2.2 Model Selection of KLIEP by Likelihood Cross Validation

The choice of the Gaussian width τ in KLIEP heavily affects the performance of importance weight estimation. Here, we explain a practical way to chose reasonable basis functions from data samples. Since KLIEP is based on the maximization of the score

J_{KLIEP}, it is natural to select the model such that J_{KLIEP} is maximized. The expectation over $p_{te}(\mathrm{X})$ involved in J_{KLIEP} can be numerically approximated by *likelihood cross validation* (LCV) as follows: First divide the test samples $\mathcal{X}^{te}$ into K disjoint subsets $\{\mathcal{X}_i^{te}\}_{i=1}^K$. Then obtain an importance estimate $\widehat{w}_k(\mathrm{X})$ from $\{\mathcal{X}_j^{te}\}_{j \neq k}$ (i.e., without $\mathcal{X}_k^{te}$) and approximate the score J_{KLIEP} using $\mathcal{X}_k^{te}$ as

$$\widehat{J}_k = \frac{1}{|\mathcal{X}_k^{te}|} \sum_{\mathrm{X} \in \mathcal{X}_k^{te}} \log \widehat{w}_k(\mathrm{X}). \tag{45}$$

This procedure is repeated for $k = 1, \ldots, K$ and the average of $\widehat{J}_k$ over all k is used as an estimate of J:

$$\widehat{J} = \frac{1}{K} \sum_{k=1}^{K} \widehat{J}_k. \tag{46}$$

For model selection, $\widehat{J}$ is computed for all model candidates (the Gaussian width τ in the current setting) and choose the one that maximizes $\widehat{J}$.

One of the potential limitations of CV in general is that it is not reliable in small sample cases since data splitting by CV further reduces the sample size. A key advantage of the LCV procedure described above is that, not the training samples, but the test input samples are cross-validated. This contributes greatly to improving model selection accuracy since the number of training samples is typically limited while a large number of test input samples are available.

3.3 Illustrative Examples

Here, we illustrate the behavior of IWKLR, IWCV, and KLIEP in covariate shift adaptation.

Figure 1 illustrates a two-dimensional binary classification problem under covariate shift. In this experiment, we define the optimal class posterior probability as follows:

$$p(y = +1|\boldsymbol{x}) = \frac{1 + \tanh(x^{(1)} - \min(0, x^{(2)}))}{2}, \tag{47}$$

$$p(y = -1|\boldsymbol{x}) = 1 - p(y = +1|\boldsymbol{x}), \tag{48}$$

where $\boldsymbol{x} = [x^{(1)}, x^{(2)}]^\top \in \mathbb{R}^2$ is the input vector. Data samples were generated from mixtures of Gaussian distributions as follows:

$$p_{tr}(\boldsymbol{x}) = \sum_{k=1}^{2} \pi_k^{tr} \mathcal{N}(\mathrm{X}|\boldsymbol{\mu}_k^{tr}, \Sigma_k^{tr}),$$

$$p_{te}(\boldsymbol{x}) = \sum_{k=1}^{2} \pi_k^{te} \mathcal{N}(\mathrm{X}|\boldsymbol{\mu}_k^{te}, \Sigma_k^{te}),$$

Table 1: Setup of illustrative examples.

	$p_{tr}(\boldsymbol{x})$		$p_{te}(\boldsymbol{x})$	
	Mixture 1	Mixture 2	Mixture 1	Mixture 2
π	0.5	0.5	0.5	0.5
$\boldsymbol{\mu}$	$(-2, 2.5)$	$(2, 2.5)$	$(-3.5, -0.5)$	$(0.5, -0.5)$
Σ	$\begin{pmatrix} 0.5 & 0 \\ 0 & 2.5 \end{pmatrix}$	$\begin{pmatrix} 0.5 & 0 \\ 0 & 2.5 \end{pmatrix}$	$\begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$	$\begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$

where π_k^{tr} and π_k^{te} are mixing coefficients of training and test distributions, and $\mathcal{N}(\mathrm{X}|\boldsymbol{\mu}, \Sigma)$ denotes the Gaussian density with mean $\boldsymbol{\mu} \in \mathbb{R}^2$ and covariance matrix $\Sigma \in \mathbb{R}^{2\times 2}$. In this experiment, we set the mixing coefficients, means, and covariances as described in Table 1.

Let the number of training and test samples be $n_{tr} = 1000$ and $n_{te} = 2000$. We use KLR/IWKLR with the linear kernel and employ CV/IWCV for tuning the regularization parameter δ. The value δ chosen by CV and IWCV for KLR and IWKLR were 10^{-6} and 1, respectively. The importance weights used in IWKLR and IWCV are learned by KLIEP and LCV is used for choosing the Gaussian width τ in KLIEP. Figure 1 shows the decision boundaries obtained by KLR+CV and IWKLR+IWCV. For references, we also showed 'OPT', which is the optimal decision boundary given by Eqs.(47) and (48). As the figure clearly shows, IWKLR+IWCV gives the decision boundary that is closer to OPT for the test samples than plain KLR+CV. The correct classification rate of KLR+CV is 93.6%, while that of IWKLR+IWCV is 96.1%. This illustrates that, under covariate shift, the prediction performance can be improved by employing the importance weighting techniques.

4 Experiments

In this section, we report the results of speaker identification in the light of covariate shift adaptation.

4.1 Data and System Description

Training and test samples were collected from 10 male speakers, and we have conducted two types of experiment—text-dependent and text-independent speaker identification. In text-dependent speaker identification, the training and test sentences are common to all speakers. On the other hand, in text-independent speaker identification, the training sentences are common to all speakers, but the test sentences are different from training sentences.

Each speaker uttered several Japanese sentences for text-dependent and text-independent speaker identification evaluation. The following three sentences are used as training and test samples in the text-dependent speaker identification experiments

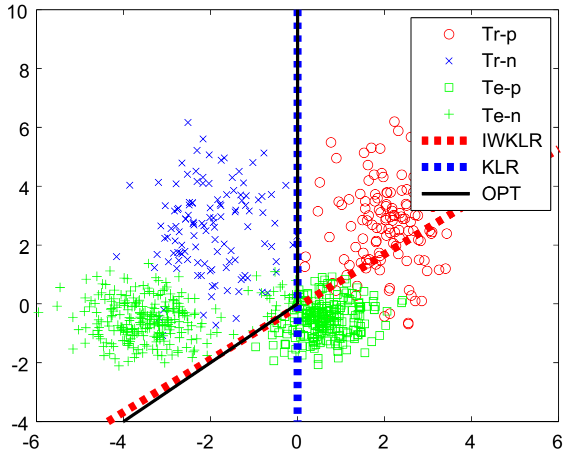

Figure 1: Decision boundaries obtained by IWKLR+IWCV and KLR+CV (red and blue dashed lines) and the optimal decision boundary (black solid line). '◦' and '×' are positive and negative training samples, while '□' and '+' are positive and negative test samples. Note that the input-output test samples are not used in the training of KLR and the output test samples are not used in the training of IWKLR—they are plotted in the figure for illustration purposes.

(Japanese sentences written using the Hepburn system of Romanization):

- seno takasawa hyakunanajusseNchi hodode mega ookiku yaya futotteiru,
- oogoeo dashisugite kasuregoeni natte shimau,
- tashizaN hikizaNwa dekinakutemo eha kakeru.

In the text-independent speaker identification experiments, the following three sentences are used as training samples:

- seno takasawa hyakunanajusseNchi hodode mega ookiku yaya futotteiru,
- oogoeo dashisugite kasuregoeni natte shimau,
- tashizaN hikizaNwa dekinakutemo eha kakeru,

and the following five sentences are used as test samples:

- tobujiyuuwo eru kotowa jiNruino yume datta,
- hajimete ruuburubijutsukaNe haittanowa juuyoneNmaeno kotoda,
- jibuNno jitsuryokuwa jibuNga ichibaN yoku shitteiru hazuda,
- koremade shouneNyakyuu mamasaN bareenado chiikisupootsuo sasae shimiNni micchakushite kitanowamusuuno boraNtiadatta,

- giNzakeno tamagoo yunyuushite fukasase kaichuude sodateru youshokumo hajimatteiru.

The utterance samples for training were recorded in 1990/12, while the utterance samples for testing were recorded in 1991/3, 1991/6, and 1991/9, respectively. Since the recording time is different between training and test utterance samples, the voice quality variation is expected to be included. Thus, the target speaker identification problem is a challenging task.

The total duration of the training sentences is about 9 sec. The durations of the test sentences for text-dependent and text-independent speaker identifications are 9 sec and 24 sec, respectively. There are approximately 10 vowels in the sentences for every 1.5 sec.

The input utterance is sampled at 16kHz. A feature vector consists of 26 components: 12 MFCCs, the normalized log energy, and their first derivatives. Feature vectors are derived at every 10 ms over the 25.6-ms Hamming-windowed speech segment, and the *cepstral mean normalization* (CMN) is applied over the features to remove channel effects. We divide each utterance into 300-ms disjoint segments, each of which corresponds to a set of features of size 26 × 30. Thus the training set is given as $\mathcal{X}^{tr} = \{\mathrm{X}_i\}_{i=1}^{411}$ for text-independent and text-dependent speaker identification evaluations. For text-independent speaker identification, the sets of test samples for 1991/3, 1991/6, and 1991/9 are given as $\mathcal{X}_1^{te1} = \{\mathrm{X}_i\}_{i=1}^{907}$, $\mathcal{X}_1^{te2} = \{\mathrm{X}_i\}_{i=1}^{919}$, and $\mathcal{X}_1^{te3} = \{\mathrm{X}_i\}_{i=1}^{906}$, respectively. For text-dependent speaker identification, the sets of test data are given as $\mathcal{X}_2^{te1} = \{\mathrm{X}_i\}_{i=1}^{407}$, $\mathcal{X}_2^{te2} = \{\mathrm{X}_i\}_{i=1}^{407}$, and $\mathcal{X}_2^{te3} = \{\mathrm{X}_i\}_{i=1}^{412}$, respectively.

We compute the speaker identification rate at every 1.5s, 3.0s, and 4.5s and identify the speaker from the average posterior probability

$$p(\mathrm{X}_t|\mathrm{V}) = \frac{1}{m}\sum_{i=1}^{m} p(\mathrm{X}_{t-i}|V), \tag{49}$$

where $m = 5, 10$, and 15, respectively.

4.2 The Results of Speaker Identification under Covariate Shift

We compared GMM, KLR, and IWKLR by computing the speaker identification rates on the 1991/3, 1991/6, and 1991/9 datasets (NTT dataset [33]), respectively. For GMM and KLR training, we only use the 1990/12 dataset (inputs $\mathcal{X}^{tr}$ and their labels).

For GMM training, the means, diagonal covariance matrices, and mixing coefficients are initialized by the results of k-means clustering on all training sentences for all speakers; then these parameters are estimated via the EM algorithm [34] for each speaker. The number of mixtures is determined by 5-fold CV. In the test phase of GMM, we compare the probability $p(\mathrm{X}_t|\boldsymbol{\mu}_k, \Sigma_k) = \prod_{j=1}^{p} p(\boldsymbol{x}_{t-j}|\boldsymbol{\mu}_k, \Sigma_k), k = 1, \ldots, 10$, where $\boldsymbol{\mu}_k$ and Σ_k are the means and covariance matrices for speaker k.

For IWKLR training, we use unlabeled samples $\mathcal{X}^{te1}$, $\mathcal{X}^{te2}$, $\mathcal{X}^{te3}$ in addition to the training inputs $\mathcal{X}^{tr}$ and their labels (i.e., semi-supervised). We first estimate the importance weight from the training and test dataset pairs ($\mathcal{X}^{tr}$, $\mathcal{X}^{te1}$), ($\mathcal{X}^{tr}$, $\mathcal{X}^{te2}$), or ($\mathcal{X}^{tr}$,

$\mathcal{X}^{te3}$) by KLIEP with 5-fold LCV, and we use 5-fold IWCV to decide the kernel band width σ and regularization parameter δ.

In our preliminary experiments, we observed that the kCV and kIWCV scores tend to be heavily affected by the way the data samples are split into k disjoint subsets (we used $k = 5$). We conjecture that this is due to non-i.i.d. nature of the MFCC features, which is different from the theory. To obtain reliable experimental results, we decided to repeat the CV procedure 50 times with different random data splits and use the highest score for model selection.

Table 2 shows the text-independent speaker identification rates in percent for 1991/3, 1991/6, and 1991/9. IWKLR refers to IWKLR with σ and δ chosen by 5-fold IWCV, KLR refers to KLR with σ and δ chosen by 5-fold CV, and GMM refers to GMM with the number of mixtures chosen by 5-fold CV. The chosen values of these hyper-parameters are described in the bracket. 'Std' in the bottom line refers to the standard deviation of the estimated importance weights $\{w(\mathrm{X}_i)\}_{i=1}^{n_{tr}}$; the smaller the standard deviation is, the 'flatter' the importance weights are. Flat importance weights imply that there is no significant distribution change between the training and test phases. Thus, the standard deviation of the estimated importance weights may be regarded as a rough indicator of the degree of distribution change.

As can be seen from the table, IWKLR+IWCV outperforms GMM+CV and KLR+CV for all sessions. This result implies that importance weighting is useful in coping with the influence of non-stationarity in practical speaker identification such as utterance variation, the recording environment change, and physical conditions/emotions.

Table 3 summarizes the text-dependent speaker identification rates in percent for 1991/3, 1991/6, and 1991/9, showing that IWKLR+IWCV and KLR+CV slightly outperform GMM and are highly comparable to each other. The result that IWKLR+IWCV and KLR+CV are comparable in this experiment would be a reasonable consequence since the standard deviation of the estimated importance weights is very small in all three cases—implying that there is no significant distribution change and therefore no adaptation is necessary. This result indicates that the proposed method does not degrade the performance when there is no significant distribution change.

Overall, the proposed method tends to improve the performance when there exists a significant distribution change and it tends to maintain the good performance of the baseline method when no distribution change exists. Based on these experimental results, we conclude that the proposed method is a promising approach to handling session dependent variation.

Table 2: Correct classification rates for text-independent speaker identification. All values are in percent. IWKLR refers to IWKLR with σ and δ chosen by 5-fold IWCV, KLR refers to KLR with σ and δ chosen by 5-fold CV, and GMM refers to GMM with the number of mixtures chosen by 5-fold CV. The chosen values of these hyper-parameters are described in the bracket. 'Std' refers to the standard deviation of estimated importance weights $\{w(\mathrm{X}_i)\}_{i=1}^{n_{tr}}$, roughly indicating the degree of distribution change.

	1991/3			1991/6			1991/9		
Time	IWKLR (1.4, 10^{-2})	KLR (1.0, 10^{-2})	GMM (16)	IWKLR (1.3, 10^{-4})	KLR (1.0, 10^{-2})	GMM (16)	IWKLR (1.2, 10^{-4})	KLR (1.0, 10^{-2})	GMM (16)
1.5s	**91.0**	88.2	89.7	**91.0**	87.7	90.2	**94.8**	91.7	92.1
3.0s	**95.0**	92.9	94.4	**95.3**	91.1	94.0	**97.9**	96.3	95.0
4.5s	**97.7**	96.1	94.6	**97.4**	93.4	96.1	**98.8**	98.3	95.8
Std	0.34	n/a	n/a	0.37	n/a	n/a	0.35	n/a	n/a

Table 3: Correct classification rates for text-dependent speaker identification. All values are in percent. IWKLR refers to IWKLR with σ and δ chosen by 5-fold IWCV, KLR refers to KLR with σ and δ chosen by 5-fold CV, and GMM refers to GMM with the number of mixtures chosen by 5-fold CV. The chosen values of these hyper-parameters are described in the bracket. 'Std' refers to the standard deviation of estimated importance weights $\{w(\mathrm{X}_i)\}_{i=1}^{n_{tr}}$, roughly indicating the degree of distribution change.

	1991/3			1991/6			1991/9		
Time	IWKLR (1.2, 10^{-4})	KLR (1.0, 10^{-2})	GMM (16)	IWKLR (1.2, 10^{-4})	KLR (1.0, 10^{-2})	GMM (16)	IWKLR (1.2, 10^{-4})	KLR (1.0, 10^{-2})	GMM (16)
1.5s	**100.0**	98.9	96.8	97.5	96.2	**97.8**	**100.0**	**100.0**	98.2
3.0s	**100.0**	**100.0**	97.7	97.5	97.2	**98.1**	**100.0**	**100.0**	98.4
4.5s	**100.0**	**100.0**	97.9	**98.9**	97.4	98.3	**100.0**	**100.0**	98.5
Std	0.05	n/a	n/a	0.05	n/a	n/a	0.05	n/a	n/a

5 Conclusions

In this paper, we proposed a novel semi-supervised speaker identification method that can alleviate the influence of non-stationarity such as session dependent variation, the recording environment change, and physical conditions/emotions. Under such non-stationary environment, standard machine learning techniques such as kernel logistic regression (KLR) and cross validation (CV) or Gaussian mixture models (GMM) and CV do not work properly due to changing environment.

Our assumption was that voice quality variants follow the *covariate shift* model—the voice feature distribution changes between the training and test phases, but the conditional distribution of the speaker index given voice features is unchanged. Under this covariate shift model, we employed the importance weighted KLR (IWKLR) method, where the importance weights are estimated by using the Kullback-Leibler importance estimation procedure (KLIEP) with likelihood CV (LCV). By combining IWKLR and KLIEP, classification accuracy under covariate shift is highly improved. Moreover, the kernel width and the regularization parameter of IWKLR are tuned based on importance weighted CV (IWCV), which is guaranteed to be almost unbiased even under covariate shift. To verify the validity of our approach, we conducted text-independent/dependent speaker identification simulations and experimentally found that the covariate shift formulation with IWKLR, IWCV, and KLIEP is a promising approach.

Following the current line of research, there are several remaining issues to be pursued for further improving the identification performance. For example, the IWCV method appeared to be rather unstable in experiments when the degree of distribution shift is very high. In such cases, further regularization of the IWCV method is expected to be useful, e.g., following the line of the paper [35]. Another challenging issue is to weaken the covariate shift assumption. The covariate shift model where only the input distribution changes could be rather restrictive in practice—the conditional distribution may also change in speaker identification tasks. In such cases, however, it is not possible to learn well in principle in the semi-supervised setup since there is no information on the test output distribution. To cope with this situation, we need to change the problem setup from semi-supervised learning to *transfer learning* [36], where a small number of test output samples are also available. We expect that a similar weighting approach is still useful even in the transfer learning scenarios.

The proposed approach, IWKLR+IWCV, is a multi-class classification method. Thus, in principle, it can be applied to identification of *any* number of speakers. In our experiments, the proposed method was shown to work well for distinguishing 10 speakers. Our future challenge is to investigate whether the same approach is still applicable to larger-scale identification problems.

Acknowledgments: MS thanks MEXT Grant-in-Aid for Young Scientists (A), 20680007, Support Center for Advanced Telecommunications Technology Research Foundation, and the Asian Office of Aerospace Research and Development for financial support.

References

[1] D. A. Reynolds and R. C. Rose, "Speaker verification using adapted Gaussian mixture models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[3] W. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Proceedings of the IEEE International Conference on Audio Speech and Signal Processing*, Orland, Florida, USA, 2002, pp. 161–164.

[4] J. Mariethoz and S. Bengio, "A kernel trick for sequences applied to text-independent speaker verification systems," *Pattern Recognition*, vol. 40, no. 8, pp. 2315–2324, 2007.

[5] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 29, no. 3, pp. 342–350, 1986.

[6] T. Matsui and K. Aikawa, "Robust model for speaker verification against session-dependent utterance variation," *IEICE Transactions on Information and Systems*, vol. E86-D, no. 4, pp. 712–718, 2003.

[7] T. Matsui and K. Tanabe, "Comparative study of speaker identification methods: dPLRM, SVM, and GMM," *IEICE Transactions on Information and Systems*, vol. E89-D, no. 3, pp. 1066–1073, 2006.

[8] S. Furui, "Cepstral analysis technique for automatic speaker verification," *Journal of Acoustical Society of America*, vol. 55, pp. 1204–1312, June, 1974.

[9] R. S. Sutton and G. A. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.

[10] H. Hachiya, T. Akiyama, M. Sugiyama, and J. Peters, "Adaptive importance sampling with automatic model selection in value function approximation," in *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI2008)*, Chicago, USA, 2008, pp. 1351–1356.

[11] H. Hachiya, T. Akiyama, M. Sugiyama, and J. Peters, "Adaptive importance sampling for value function approximation in off-policy reinforcement learning," *Neural Networks*, 2009, to appear.

[12] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach*, MIT Press, Cambridge, MA, 1998.

[13] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Interesting structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[14] S. Bickel and T. Scheffer, "Dirichlet-enhanced spam filtering based on biased samples," in *Advances in Neural Information Processing Systems*, Cambridge, MA, 2007, pp. 161–168, MIT Press.

[15] J. Jing and Z. ChengXiang, "Instance weighting for domain adaptation in NLP," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, 2007, pp. 264–271.

[16] Y. Tsuboi, H. Kashima, S. Hido, S. Bickel, and M. Sugiyama, "Direct density ratio estimation for large-scale covariate shift adaptation," *IPSJ Journal*, vol. 50, no. 4, pp. 1–19, 2009.

[17] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[18] M. Sugiyama, M. Krauledat, and K.-R. Müller, "Covariate shift adaptation by importance weighted cross validation," *Journal of Machine Learning Research*, vol. 8, pp. 985–1005, 2007.

[19] J. J. Heckman, "Sample selection bias as a specification error," *Econometrica*, vol. 47, no. 1, pp. 153–162, 1979.

[20] D. A. Chon, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.

[21] V. V. Fedorov, *Theory of Optimal Experiments*, Academic Press, New York, 1972.

[22] D. P. Wiens, "Robust weights and designs for biased regression models: Least squares and generalized M-estimation," *Journal of Statistical Planning and Inference*, vol. 83, no. 2, pp. 395–412, 2000.

[23] T. Kanamori and H. Shimodaira, "Active learning algorithm using the maximum weighted log-likelihood estimator," *Journal of Statistical Planning and Inference*, vol. 116, no. 1, pp. 149–162, 2003.

[24] M. Sugiyama, "Active learning in approximately linear regression based on conditional expectation of generalization error," *Journal of Machine Learning Research*, vol. 7, pp. 141–166, 2006.

[25] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*, MIT Press, Cambridge, MA, 2009.

[26] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of Statistical Planning and Inference*, vol. 90, no. 2, pp. 227–244, 2000.

[27] M. Sugiyama, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Advances in Neural Information Processing Systems*, Cambridge, MA, 2008, pp. 1433–1440, MIT Press.

[28] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, "Direct importance estimation for covariate shift adaptation," *Annals of the Institute of Statistical Mathematics*, vol. 60, no. 4, pp. 699–746, 2008.

[29] L. Rabiner and B-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[30] K. Tanabe, "Penalized logistic regression machines: New methods for statistical prediction 1," Tech. Rep. 143, Institute of Statistical Mathematics, 2001.

[31] B. Schölkopf and A. J. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.

[32] G. S. Fishman, *Monte Carlo: Concepts, Algorithms, and Applications*, Springer-Verlag, Berlin, 1996.

[33] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," in *Proceedings of the IEEE International Conference on Audio Speech and Signal Processing*, Minneapolis, USA, 1993, pp. 391–394.

[34] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[35] M. Sugiyama, M. Kawanabe, and K.-R. Müller, "Trading variance reduction with unbiasedness: The regularized subspace information criterion for robust model selection in kernel regression," *Neural Computation*, vol. 16, no. 5, pp. 1077–1104, 2004.

[36] S. Thrun and L. Pratt, Eds., *Learning to Learn*, Kluwer Academic Publishers, Norwell, MA, USA, 1998.