Covariate Shift Adaptation for Semi-supervised Speaker Identification

2009/4/22

M. Yamada, M. Sugiyama, and T. Matsui

Speaker Identification

Task: Identify speaker using voice.

Hello!

Speaker identification system

Feature extraction

Classifier (e.g., SVM)

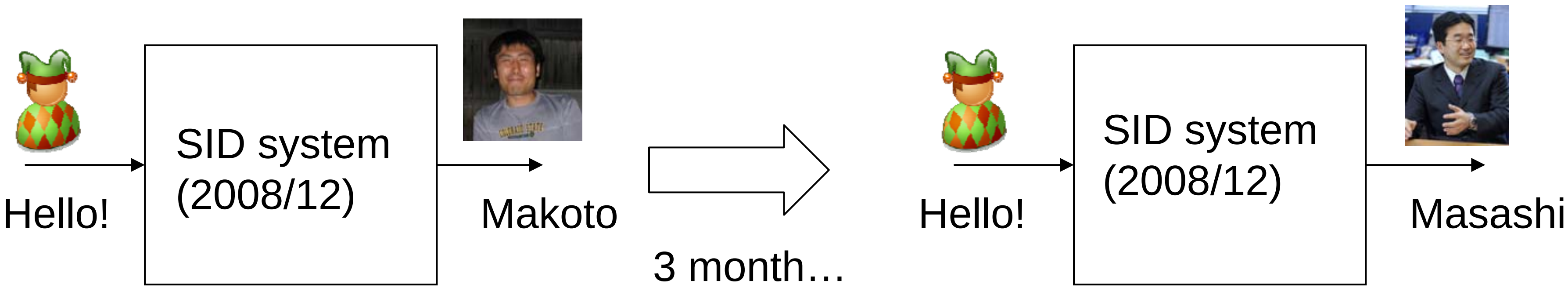
0.1

0.9

makoto

Problems in speaker identification

- Feature variation
  - Sound recording environment change
  - Physical condition/emotion
  - Noise
  - <u>Session dependent variation</u>



Can we use the same system trained in 2008/12 for future? NO!
⇒ Speech feature (e.g., MFCC) changes in 3 months.

Solutions

1. Recording several sessions of speeches
   - Labeling is required.

   ⇒ Very expensive!
2. Semi-supervised learning
   - We use unlabeled data for training.
   - No labeling process required.

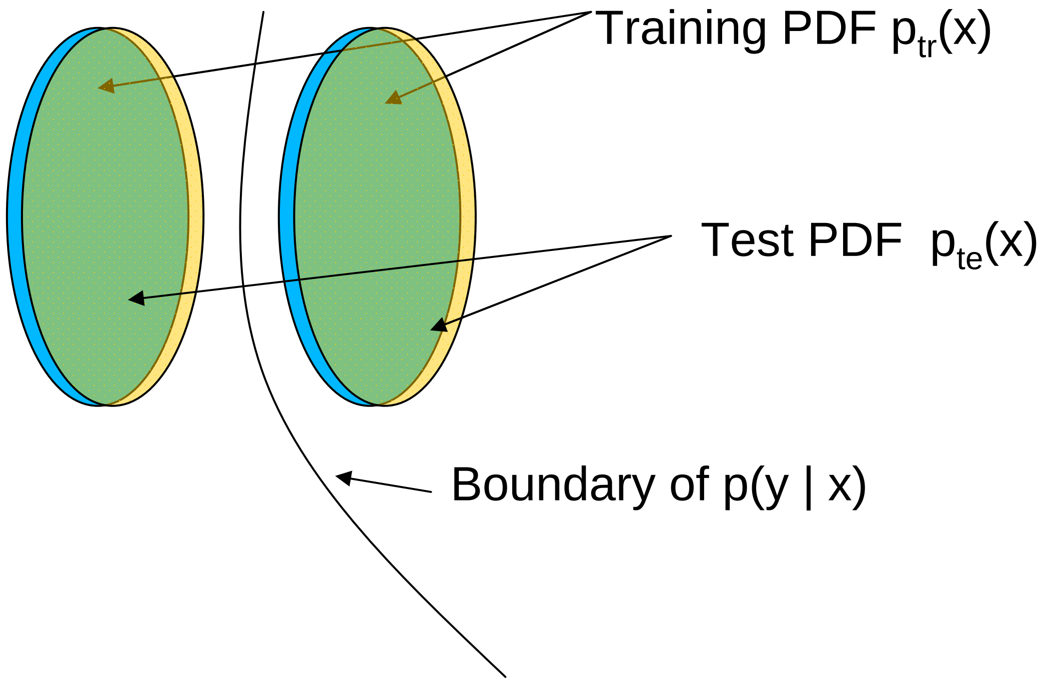
   ⇒ Reasonable.

We assume the speech data follows covariate shift
⇒We model the covariate shift by using
Importance Weighted Kernel Logistic Regression (IWKLR)

Supervised Learning

Assumption in supervised learning

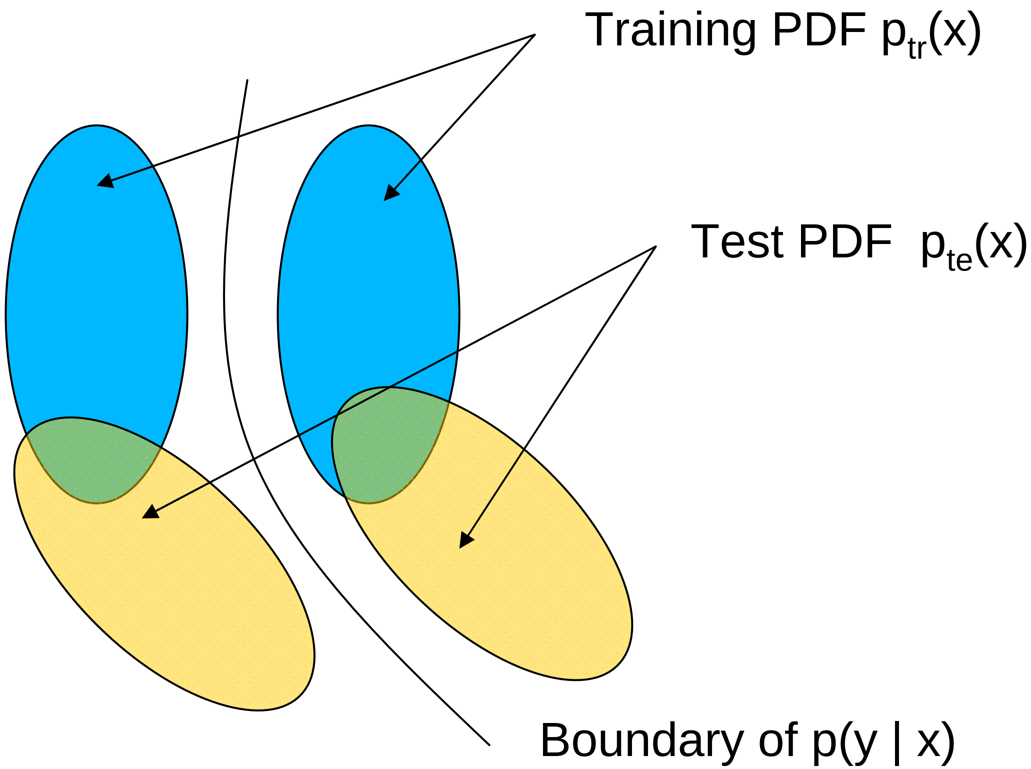
Training and test probability density functions are same.



Is this assumption acceptable in practice？ NO!

Covariate shift [1]

- Input probability density changes: $p_{tr}(x) \neq p_{te}(x)$
- Conditional probability density remains unchanged: $p(y \mid x)$



[1] H. Shimodaira, '' Improving predictive inference under covariate shift by weighting the log-likelihood function,'' JSPI, 90, 227-244, 2000

Covariate shift

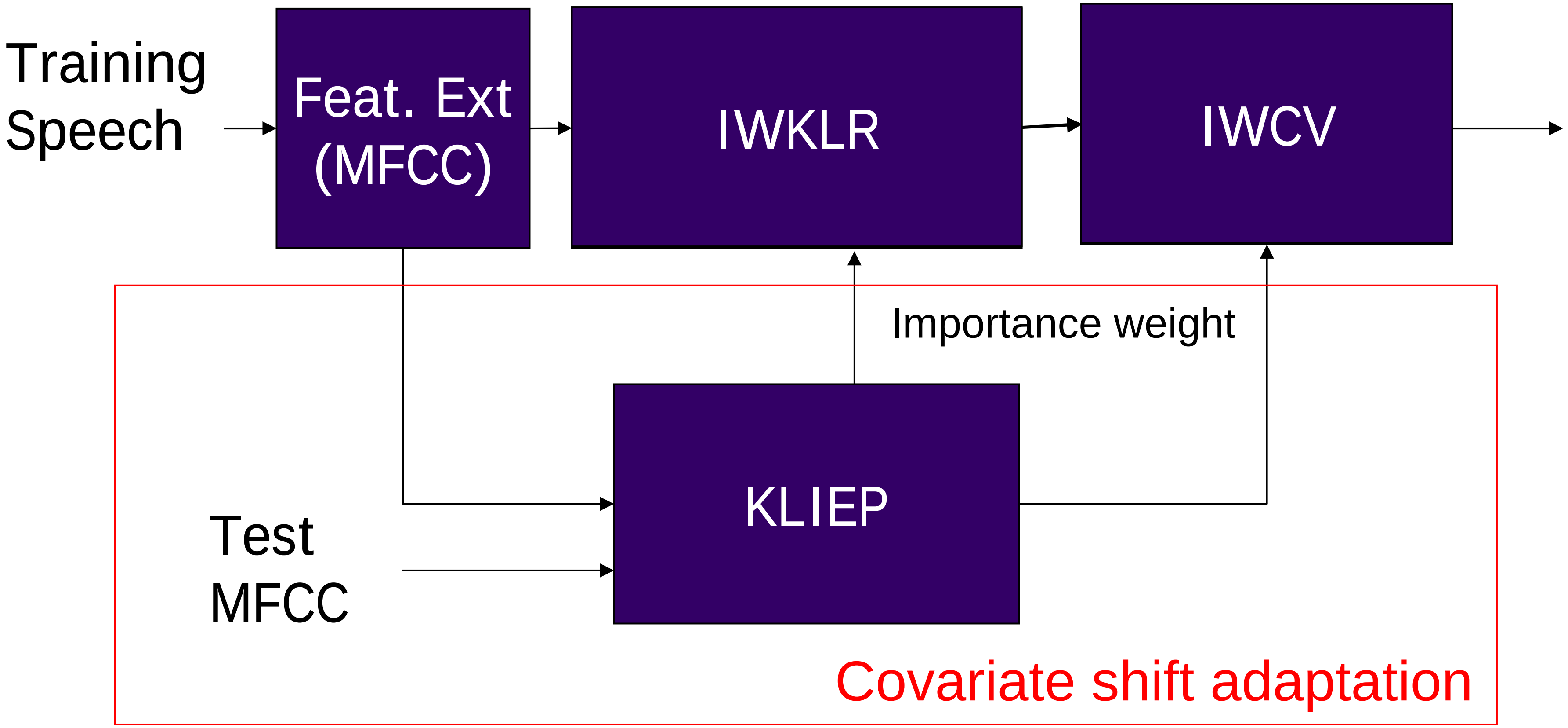
Cost function under covariate shift:

Taking the expectation over test probability density.

$$
\begin{aligned}
\mathrm{E}_{p_{te}(\mathrm{X})}[F(\mathrm{X})] &= \int F(\mathrm{X}) p_{te}(\mathrm{X}) d\mathrm{X} \\
&= \int F(\mathrm{X}) \boxed{w(\mathrm{X}) p_{tr}(\mathrm{X})} d\mathrm{X} \\
&= \mathrm{E}_{p_{tr}(\mathrm{X})}[F(\mathrm{X}) w(\mathrm{X})].
\end{aligned}
$$

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

Proposed framework



Proposed method is consistent under covariate shift!

[2] M. Sugiyama, et.al, ``Covariate shift adaptation by importance weighted cross validation.,'' JMLR, vol. 8 (May), pp.985-1005, 2007

[3] M. Sugiyama, et.al., `` Direct importance estimation for covariate shift adaptation,'' AISM, vol. 60, no.4, pp.699-746, 2008

Problem formulation

Mel-frequency cepstrum coefficient (MFCC):

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

m labeled samples and (*n*−*m*) unlabeled samples:

$$\mathcal{Z}_{tr} = \{\mathrm{X}_i, y_i\}_{i=1}^{m}$$
$$\mathcal{Z}_{te} = \{\mathrm{X}_i\}_{i=m+1}^{n}$$

Speaker index:

$$\boxed{y_i \in \{1, \ldots, K\}}$$

Kernel based Speaker Identification

Posterior probability:

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

Discriminative function:

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

Sequence kernel[4]:

$$\mathcal{K}(\mathrm{X},\mathrm{X}') = \frac{1}{NN'}\sum_{i=1}^{N}\sum_{i'=1}^{N'} \exp\left(\frac{-||\boldsymbol{x}_i - \boldsymbol{x}'_{i'}||^2}{2\sigma^2}\right).$$

[4] J. Mariethoz and S. Bengio, '' A kernel trick for sequences applied to text-independent speaker verification systems,'' Pattern Recognition, 40, 2315-2324, 2007

Importance weighted kernel logistic regression (IWKLR)

Negative regularized importance weighted log-likelihood:

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

Regularizer : $\frac{\delta}{2}\mathrm{tr}(\mathrm{V}\mathrm{K}\mathrm{V}^\top)$

Gram matrix: $\mathrm{K} = [\mathcal{K}(\mathrm{X}_i,\mathrm{X}_j)]_{i,j=1}^{n}$

Importance weight: $w(\mathrm{X})$

Negative log likelihood is convex

⇒ Easy to compute via Newton method.

Importance Weighted Cross Validation (IWCV)

Model parameters in IWKLR

Kernel width : σ
Regularization parameter : δ

k-fold importance weighted cross validation (IWCV):

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

$\{\mathcal{Z}_i\}_{i=1}^{k}$: Subset of $\mathcal{Z} = \{(\mathrm{X}_i, y_i)\}_{i=1}^{n}$

$|\mathcal{Z}_j|$: Number of samples in the subset

$I(\cdot)$: Indicator function

$w(\mathrm{X})$: Importance weight

Kullback–Leibler Importance Estimation Procedure (KLIEP)[3]

Model:

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

Basis:

$$\varphi(\mathrm{X}, \mathrm{X}') = \frac{1}{NN'} \sum_{i=1}^{N} \sum_{i'=1}^{N'} \exp\left(\frac{-||\boldsymbol{x}_i - \boldsymbol{x}_{i'}||^2}{2\tau^2} \right)$$

Cost function:

$$\max_{\{\alpha_l\}_{l=1}^{b}} \left[\sum_{i=1}^{n_{te}} \log \left(\sum_{l=1}^{b} \alpha_l \varphi(\mathrm{X}_i^{te}, \mathrm{C}_l) \right) \right]$$

$$\text{s.t.} \sum_{i=1}^{n_{tr}} \sum_{l=1}^{b} \alpha_l \varphi(\mathrm{X}_i^{tr}, \mathrm{C}_l) = n_{tr} \ \text{ and } \ \alpha_1, \ldots, \alpha_b \geq 0.$$

[3] M. Sugiyama, et.al., `` Direct importance estimation for covariate shift adaptation,'' AISM, vol. 60, no.4, pp.699-746, 2008

Evaluation

- Simulation condition
  - 10 speakers
  - Training data (1990/12)
  - Test data (1991/3, 1991/6, 1991/9)
  - 16kHz sampling
  - Speech length 12sec × 10 speakers
  - 12 MFCC + ΔMFCC + log power + Δlog power
  - Utterance data (300ms) $\mathrm{X}_i \in \mathbb{R}^{26\times 30}$
  - 5-fold CV (KLR)
  - 5-fold IWCV (IWKLR)

Evaluation result

Date	IWKLR + IWCV (1.5s)	KLR + CV (1.5s)	IWKLR + IWCV (4.5s)	KLR + CV (4.5s)
1991/3	**86.8** (1.2, 0.0001)	86.1 (1.2, 0.0001)	**92.6** (1.2,0.0001)	92.3 (1.2, 0.0001)
1991/6	**83.9** (1.3,0.0001)	82.0 (1.2, 0.0001)	**93.7** (1.3,0.0001)	92.7 (1.2, 0.0001)
1991/9	**92.0** (1.2, 0.0001)	91.7 (1.2, 0.0001)	**99.9** (1.2,0.0001)	99.7 (1.2, 0.0001)
Average	**87.6**	86.6	**95.4**	94.9

* σ and δ are in the bracket.
Model parameters are selected CV/IWCV.

Conclusion & Future works

- Conclusion
  - Propose the semi-supervised speaker identification
  - Session dependent variation was alleviated by using the covariate shift adaptation
- Future works
  - Detection of Covariate shift
  - Modeling the physical condition/emotion using Covariate shift.