Statistical Speech and Audio Processing

Makoto Yamada (Sugiyama Lab PD)
http://sugiyama-www.cs.titech.ac.jp/~yamada/
yamada@sg.cs.titech.ac.jp

2010/5/18

Contents

Contents

Statistical Speech & Audio Processing

Speech and audio processing applications

- Speech recognition
- Speech / singing voice synthesis
- Voice conversion
- Sound classification (Speaker identification, audio identification)
- Source separation
- etc.

Statistical Speech & Audio Processing (1)

Speech recognition

- Information retrieval based on Speech (Google)
- Automatic speech to text transformation in conference / meeting

Statistical Speech & Audio Processing (2)

Speech / Singing voice synthesis

- Text-to-speech sysnthesis
- DeskTop Music (DTM)

Statistical Speech & Audio Processing (3)

Voice conversion

- Intercommunication system (Answering with male voice)
- Singing male / female song

Statistical Speech & Audio Processing (4-1)

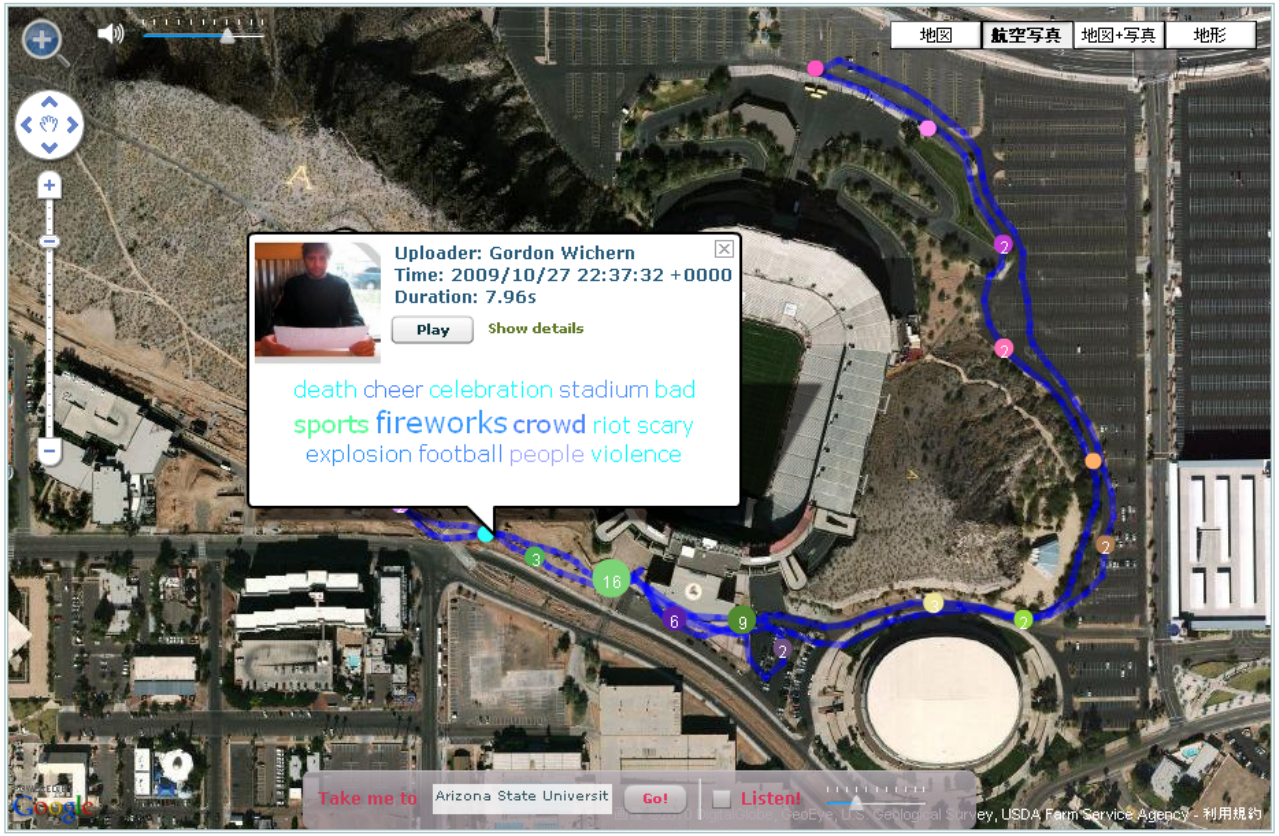
Speaker identification

- Security system
- Humanoid robots
- Diarization in meeting / conference (Recording who spoke when)

Statistical Speech & Audio Processing (4-2)

Audio identification

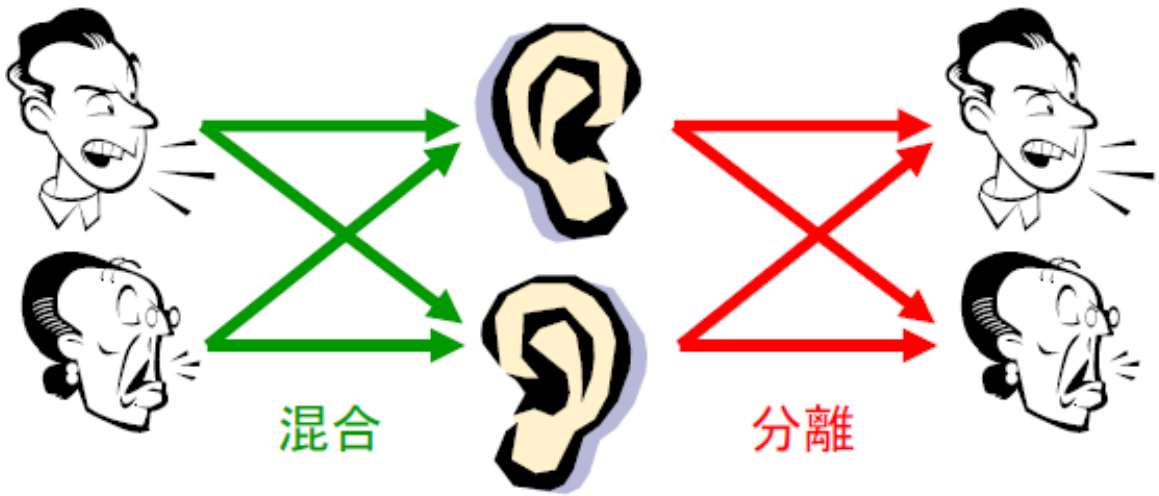
- Automatic internet audio archive generation



Statistical Speech & Audio Processing (5)

Source separation

- Pre-processing of speech recognition /speaker identification
- Sound recording
- Audio upmixing (2ch audio → 5ch audio)



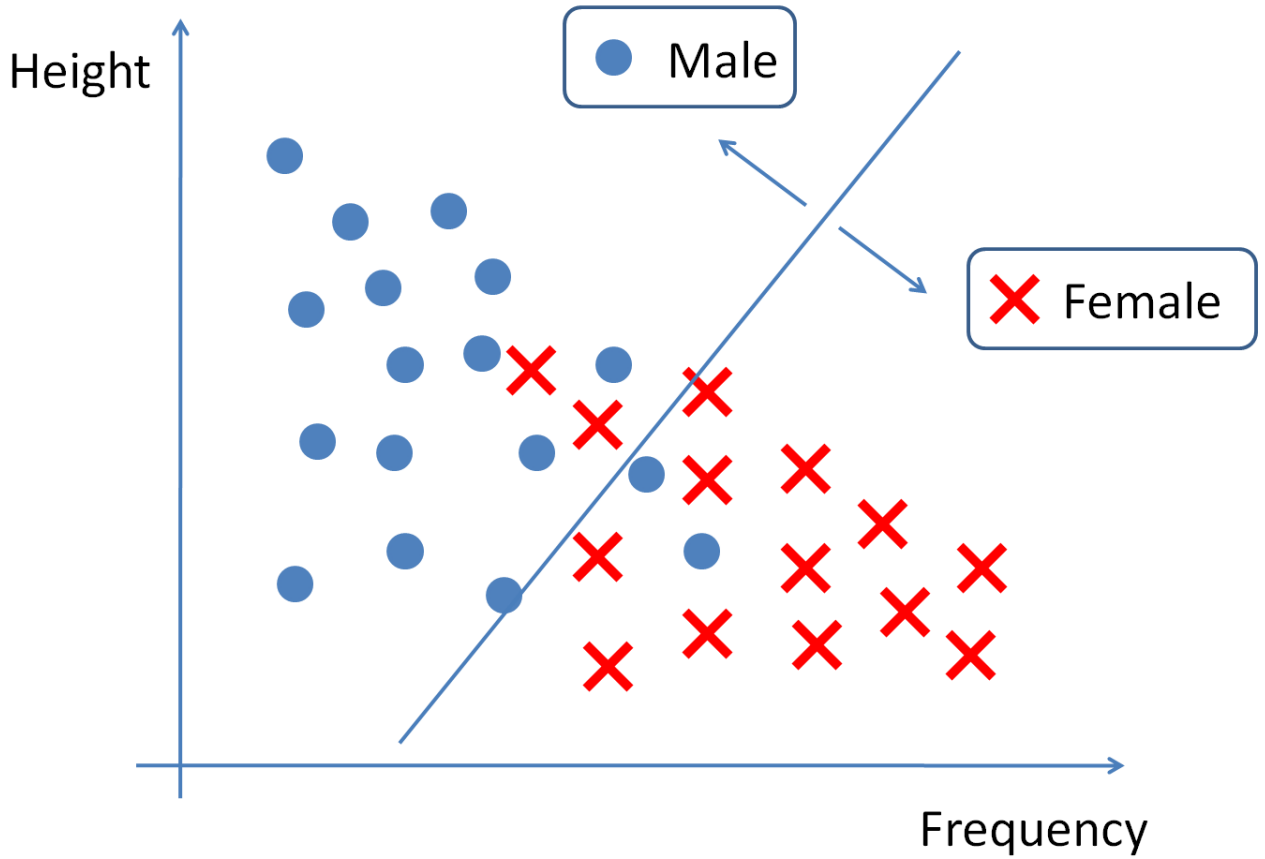
Contents

Definition of variables

Random variable: X, Y
Observed variable : $\boldsymbol{x} = [x_1, \ldots, x_d]^\top \in \mathbb{R}^d$, y
Matrix : $\mathrm{X} = [\boldsymbol{x}_1, \ldots \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$

Pattern recognition ?

Example：Male and female identification



Bayes Theorem (1)

$p(X)$: X (Frequency, Height)
$p(Y)$: Y (Male = 1, Female = 2)
$p(X, Y)$: Joint probability
$p(X|Y)$: ($p(X|Y = 1)$ Male's probability distribution)
$p(Y|X)$: Posterior probability

$$p(X, Y) = p(Y|X)p(X)$$

Bayes Theorem (2)

Marginalization (**周辺化**)

$$p(X) = \sum_{Y'} p(X, Y')$$

Bayes Theorem (**ベイズの定理**)

$$\begin{aligned} p(Y|X) &= \frac{p(X,Y)}{p(X)} \\ &= \frac{p(X|Y)p(Y)}{\sum_{Y'} p(X,Y')} \\ &= \frac{p(X|Y)p(Y)}{\sum_{Y'} p(X|Y')p(Y')} \\ &= \frac{p(X|Y)}{\sum_{Y'} p(X|Y')}, \ (p(Y) = p(Y') \ \forall Y') \end{aligned}$$

Maximum Likelihood Estimation (1)

Let us denote $\boldsymbol{z}_i = (x_i, y_i)$, then the joint probability of the feature $\{\boldsymbol{x}_i\}_{i=1}^N$ and class label $\{y_i\}_{i=1}^N$ can be written as

$$\begin{aligned} p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N) &= p(\boldsymbol{z}_1) \ldots p(\boldsymbol{z}_N) \\ &= \prod_{i=1}^{N} p(\boldsymbol{z}_i) \\ &= \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i) p(\boldsymbol{x}_i) \\ &\propto \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i) \end{aligned}$$

Maximum Likelihood Estimation (2)

Negative log likelihood function

$$\begin{aligned} \mathcal{P}^{\log}(\mathcal{Z}; \mathrm{V}) &= -\log \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i, \mathrm{V}) \\ &= -\sum_{i=1}^{N} \log p(y_i|\boldsymbol{x}_i, \mathrm{V}) \end{aligned}$$

Maximizing the joint distribution w.r.t V

$$\frac{\partial}{\partial \mathrm{V}} \mathcal{P}^{\log}(\mathcal{Z}; \mathrm{V}) = -\sum_{i=1}^{N} \frac{1}{p(y_i|\boldsymbol{x}_i)} \frac{\partial}{\partial \mathrm{V}} p(y_i|\boldsymbol{x}_i; \mathrm{V}) = 0$$

Fourier Transformation

Time → Frequency transformation

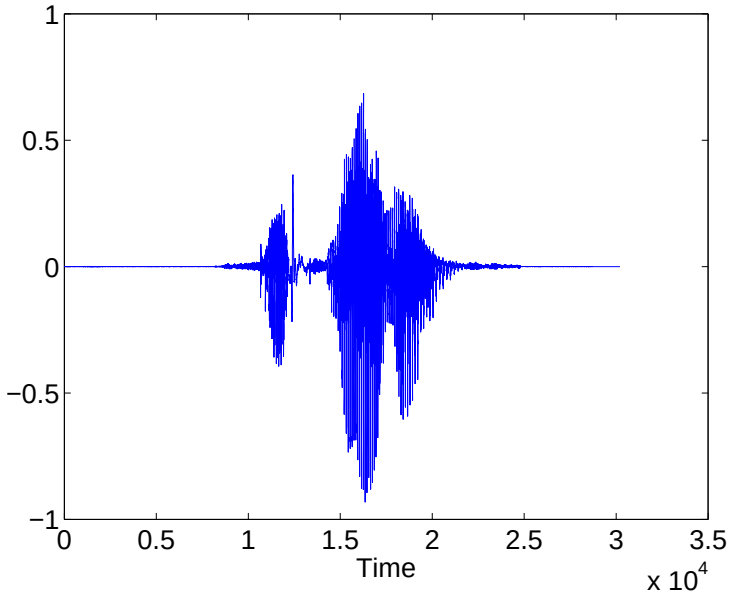


Figure: Speech signal (time)

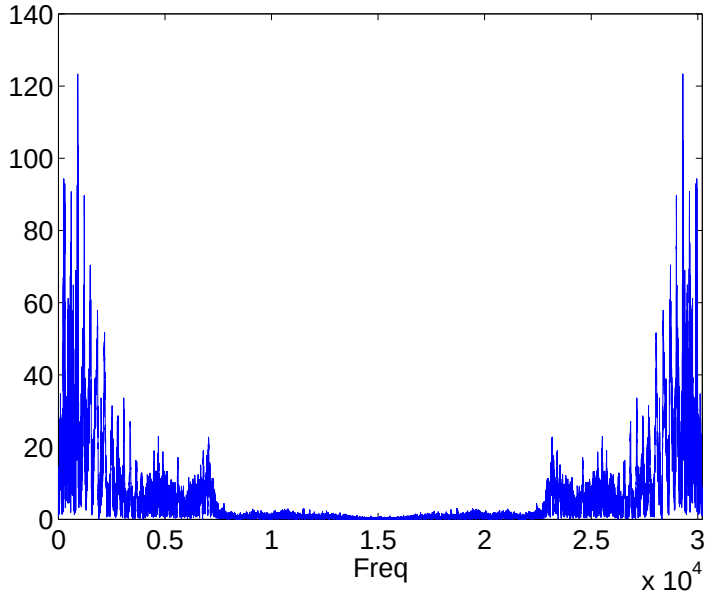
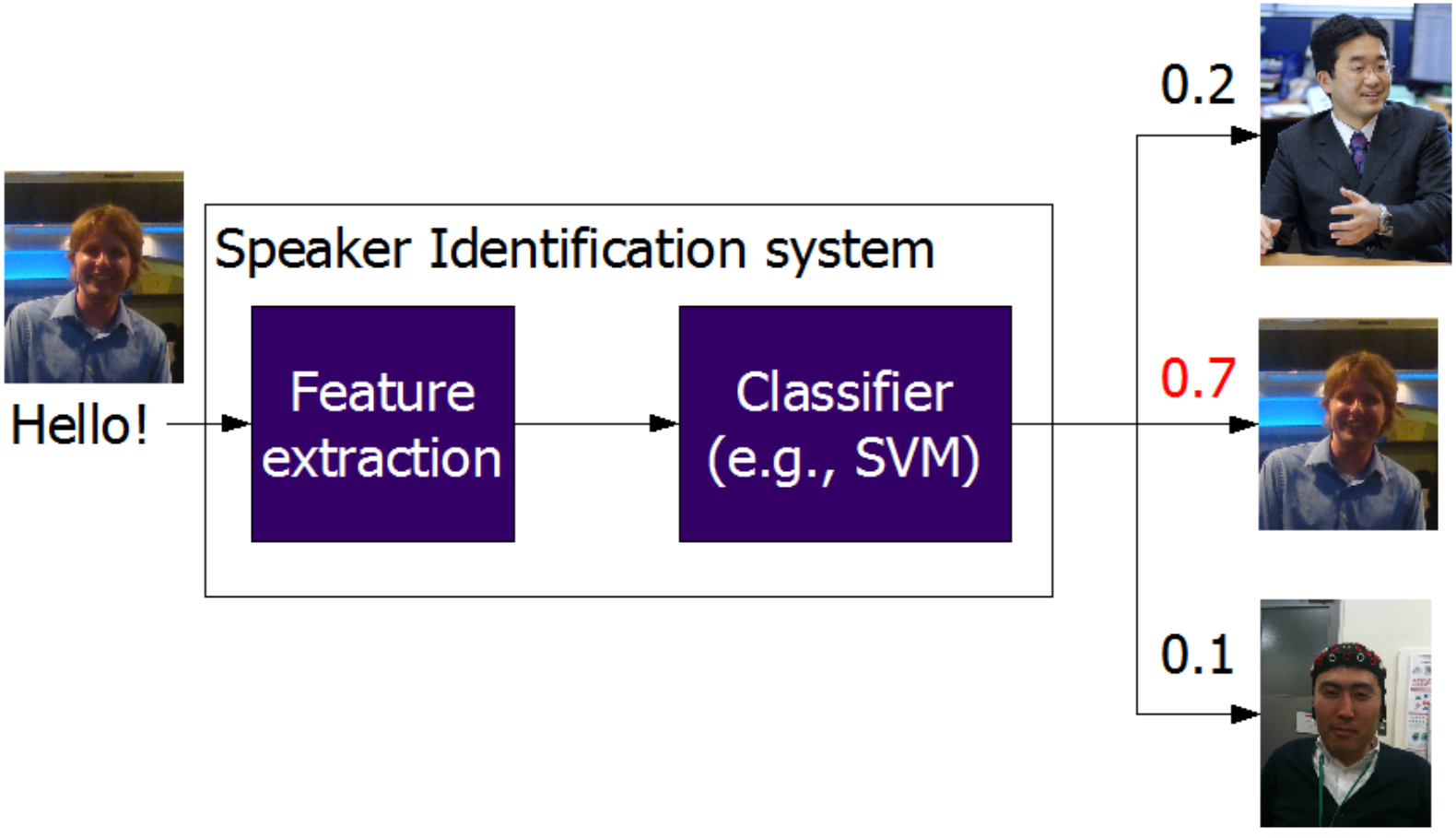


Figure: Speech signal (frequency)

Contents

What is Speaker Identification?

Speaker Identification: Identifying speakers from speech



Applications of Speaker Identification

- Security system (Home security, etc.)
- Speaker search from databases
  e.g., speaker diarization in conferences/meetings
- Singer search from databases
  e.g., searching the favorite singer
- Speaker Identification for robots

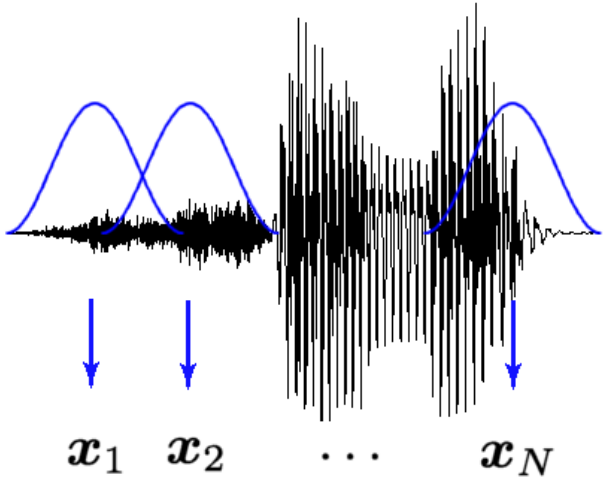
Feature Extraction: MFCC

Mel-Frequency Cepstrum Coefficients (MFCC):

- Framing, windowing, and Fast Fourier Transform (FFT)
- Mel filter bank
- Taking logarithm (Spectrum $\rightarrow$ cepstrum)
- Discrete Cosine Transform (DCT)

Δ MFCC

- The derivative of MFCC



Classifier Types for Speaker Identification

- Vector Quantization (since 1980)
- Gaussian Mixture Models (GMM) (since 1990)
- Kernel methods (since 2000)
  Support Vector Machine, Kernel Logistic Regression
  →Better identification performance than GMM
  →Proposed various kernels for speech processing

Kernel method

Kernel model:

$$\widehat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i)$$

$k(\boldsymbol{x}, \boldsymbol{x}')$: Kernel function
e.g., Gaussian kernel

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

Speaker Identification Problem Formulation

Speech feature (MFCC):

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

n_{tr} labeled training samples:

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

Speaker index:

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

Bayes Theorem

Bayes Theorem (**ベイズの定理**)

$$
\begin{aligned}
p(Y|X) &= \frac{p(X,Y)}{p(X)} \\
&= \frac{p(X|Y)p(Y)}{\sum_{Y'} p(X,Y')} \\
&= \frac{p(X|Y)p(Y)}{\sum_{Y'} p(X|Y')p(Y')} \\
&= \frac{p(X|Y)}{\sum_{Y'} p(X|Y')}, \ (p(Y) = p(Y') \ \forall Y')
\end{aligned}
$$

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 (Kernel model):

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

Kernel-based Speaker Identification

Mean Operator Sequence Kernel (MOSK)[1]:

$$\mathcal{K}(\mathrm{X}, \mathrm{X}') = \frac{1}{NN'} \sum_{p=1}^{N} \sum_{p'=1}^{N'} k(\boldsymbol{x}_p, \boldsymbol{x}'_{p'}),$$

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

Kernels can be designed w.r.t data. (speech, image, bioinfo, etc.)

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

Maximum Likelihood Estimation (1)

Let us denote $\boldsymbol{z}_i = (x_i, y_i)$, then the joint probability of the feature $\{\boldsymbol{x}_i\}_{i=1}^N$ and class label $\{y_i\}_{i=1}^N$ can be written as

$$\begin{aligned} p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N) &= p(\boldsymbol{z}_1) \ldots p(\boldsymbol{z}_N) \\ &= \prod_{i=1}^{N} p(\boldsymbol{z}_i) \\ &= \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i) p(\boldsymbol{x}_i) \\ &\propto \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i) \end{aligned}$$

Maximum Likelihood Estimation (2)

Negative log likelihood function

$$\begin{aligned} \mathcal{P}^{\log}(\mathcal{Z}; \mathrm{V}) &= -\log \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i, \mathrm{V}) \\ &= -\sum_{i=1}^{N} \log p(y_i|\boldsymbol{x}_i, \mathrm{V}) \end{aligned}$$

Maximizing the joint distribution w.r.t V

$$\frac{\partial}{\partial \mathrm{V}} \mathcal{P}^{\log}(\mathcal{Z}; \mathrm{V}) = -\sum_{i=1}^{N} \frac{1}{p(y_i|\boldsymbol{x}_i)} \frac{\partial}{\partial \mathrm{V}} p(y_i|\boldsymbol{x}_i; \mathrm{V}) = 0$$

Kernel Logistic Regression (KLR)

Negative log-likelihood function:

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

Gram matrix:

$$\mathrm{K} = [\mathcal{K}(\mathrm{X}_i, \mathrm{X}_j)]_{i,j=1}^{n_{tr}}$$

Regularizer:

$$\frac{\delta}{2}\mathrm{trace}(\mathrm{VKV}^{\top})$$

Negative log-likelihood function is convex $\rightarrow$ can be efficiently solved via (quasi-)Newton methods.

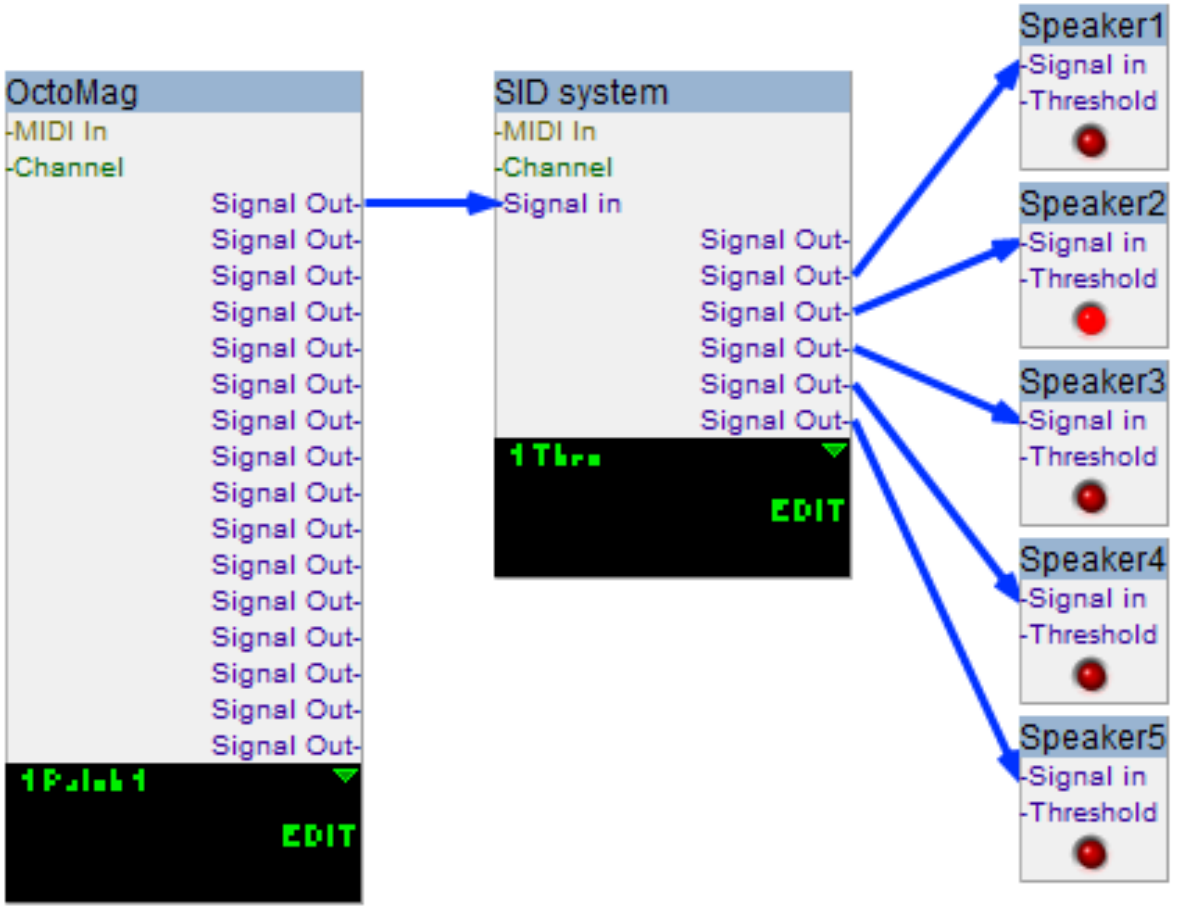
Realtime Speaker Identification

PC spec

- Core2 Qued 2.0GHz
- 2G byte memory
- 16kHz sampling
- Virtual Studio Technology (VST)
- C++

Realtime speaker identification

Demo



Contents

Problems in Speaker Identification

Variation of speech feature degrades the identification performance.

Feature variation types

- Sound recording environment change
- Physical condition/emotion
- Noise
- Session variation

Solutions

Recording several sessions of speech

- Labeling for new dataset is required
  → Very expensive!

Semi-supervised learning

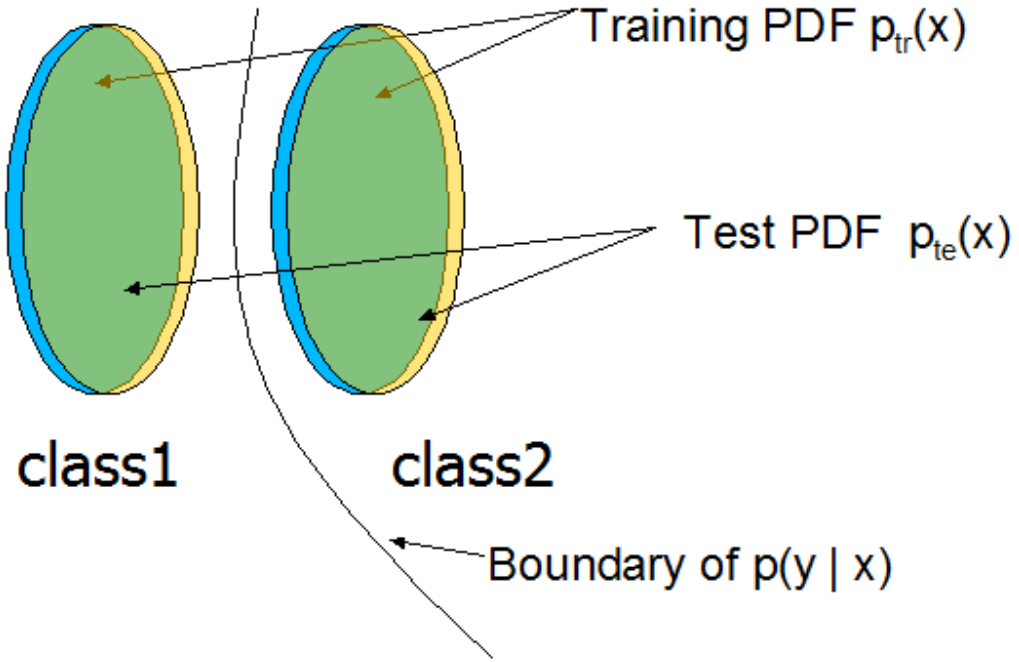
- Using labeled + unlabeled data for training
  e.g., labeled data: speech recorded in 2009/05
  unlabeled data: speech recorded in 2009/06
- Labeling for new dataset is NOT required.
  → Reasonable solution!

We assume the speech data follows covariate shift

Supervised Learning

Assumption in supervised learning

Training and test probability density functions are same.

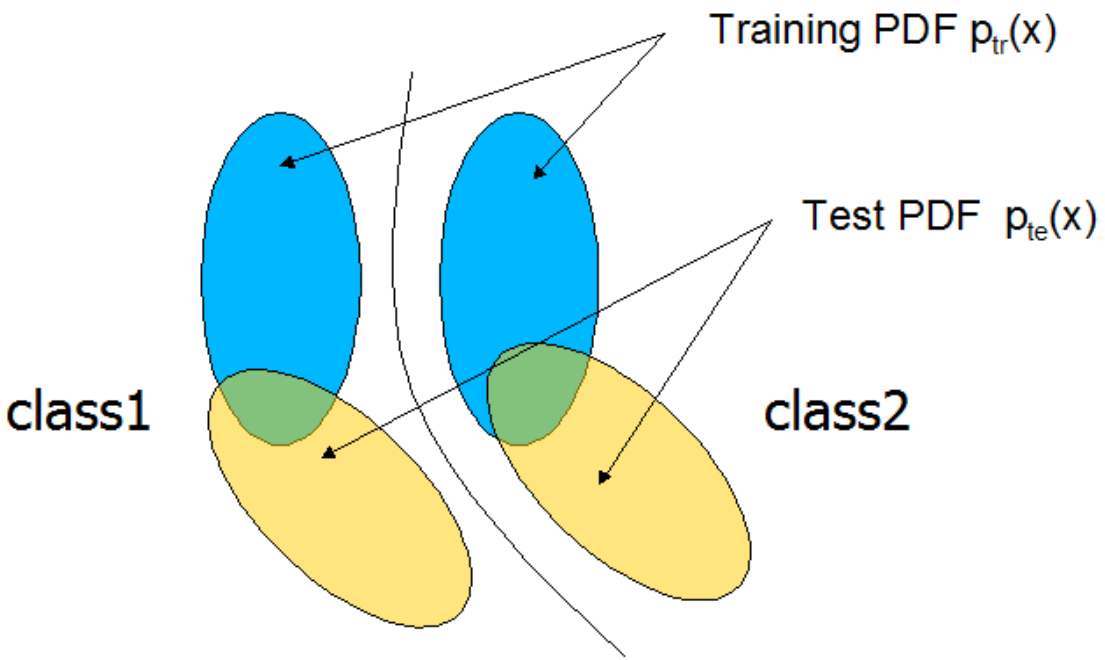


Is this assumption acceptable in practice? NO!

Covariate Shift

Assumption in covariate shift

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



e.g., Training data (clean signal), Test data (noisy signal)

Covariate Shift

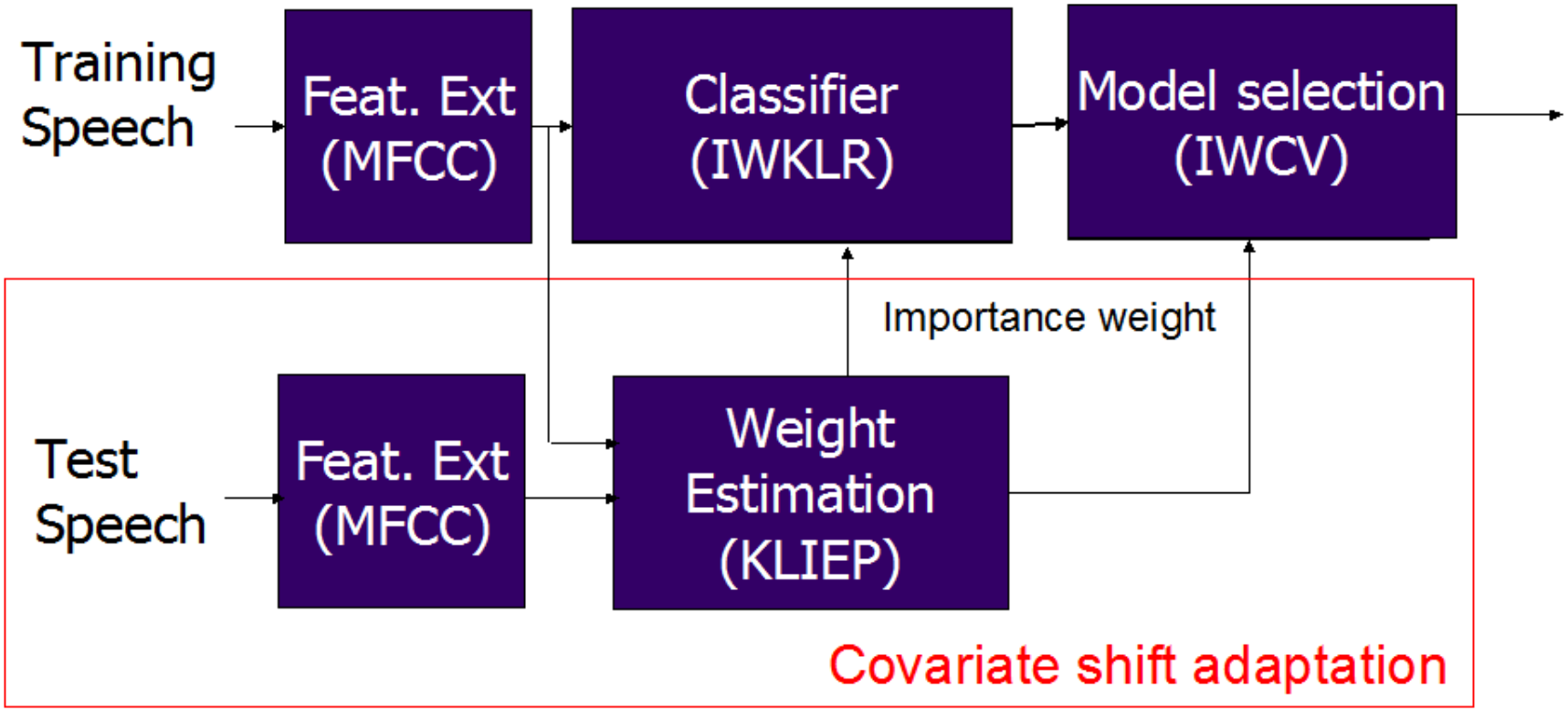
Define the cost function over testing data
However, difficult to compute the expectation over test data
$\rightarrow$ importance sampling!

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

Importance:

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

Proposed Framework



Proposed method is consistent under covariate shift!

Speaker Identification Problem Formulation under Covariate Shift

Speech feature (MFCC):

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

n_{tr} labeled training samples, n_{te} unlabeled test samples:

$$\begin{aligned} \mathcal{Z}_{tr} &= \{\mathrm{X}_i, y_i\}_{i=1}^{n_{tr}} \\ \mathcal{Z}_{te} &= \{\mathrm{X}_i\}_{i=1}^{n_{te}} \end{aligned}$$

Speaker index:

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

Importance Weighted Kernel Logistic Regression (IWKLR)

Negative regularized importance weighted log-likelihood:

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

Gram matrix:

$$\mathrm{K} = [\mathcal{K}(\mathrm{X}_i,\mathrm{X}_j)]_{i,j=1}^{n_{tr}}$$

Regularizer:

$$\frac{\delta}{2}\mathrm{trace}(\mathrm{VKV}^{\top})$$

Negative log-likelihood is convex
→ Easy to compute via Newton method.
http://sugiyama-www.cs.titech.ac.jp/˜yamada/iwklr.html

Simulation condition

- 10 speakers
- Training data (1990/12)
- Test data 1991/3, 1991/6, 1991/9
- 16kHz sampling
- Speech length 10sec $\times$ 10 speakers
- 12 MFCC + ΔMFCC + log power + Δlog power
- Cepstrum Mean Normalization (CMN)
- 5-fold CV (KLR, GMM)
- 5-fold IWCV (IWKLR)

Evaluation

Text-independent speaker identification:

	1991/3			1991/6			1991/9		
	IWKLR	KLR	GMM	IWKLR	KLR	GMM	IWKLR	KLR	GMM
Time	$(1.4, 10^{-2})$	$(1.0, 10^{-2})$	(16)	$(1.3, 10^{-4})$	$(1.0, 10^{-2})$	(16)	$(1.2, 10^{-4})$	$(1.0, 10^{-2})$	(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

Text-dependent speaker identification:

	1991/3			1991/6			1991/9		
	IWKLR	KLR	GMM	IWKLR	KLR	GMM	IWKLR	KLR	GMM
Time	$(1.2, 10^{-4})$	$(1.0, 10^{-2})$	(16)	$(1.2, 10^{-4})$	$(1.0, 10^{-2})$	(16)	$(1.2, 10^{-4})$	$(1.0, 10^{-2})$	(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

Conclusion and Future Works

Conclusion

- Proposed the semi-supervised speaker identification
- Session dependent variation was alleviated by using the covariate shift adaptation

Future works

- Modeling the physical condition/emotion using covariate shift.
- Speaker identification in conference system, robotics, etc.
- Real-time implementation of adaptation method.

Statistical Speech & Audio Processing

Speech and audio processing applications

- Speech recognition
- Speech / singing voice synthesis
- Voice conversion
- Sound classification (Speaker identification, audio identification)
- Source separation
- etc.

Contact

If you have any questions about my research or something, please feel free to come to E504 or sending an e-mail.

- http://sugiyama-www.cs.titech.ac.jp/˜yamada
- yamada@sg.cs.titech.ac.jp