

Factor Analysis for Speaker Recognition

Presented by:

Kong Aik LEE

Institute for Infocomm Research (I²R)
A*STAR, Singapore

Overview

- Introduction
- Factor analysis
 - General formulation
 - Marginal and conditional distribution
- Total variability model
 - Tying across frames (single Gaussian)
 - Tying across frames and mixtures (multi-Gaussian factor analyser)
- Interesting research topics
 - Using non-standard Gaussian prior for i-vector extraction
 - From total to local variability model

A brief Introduction to

SPEAKER RECOGNITION

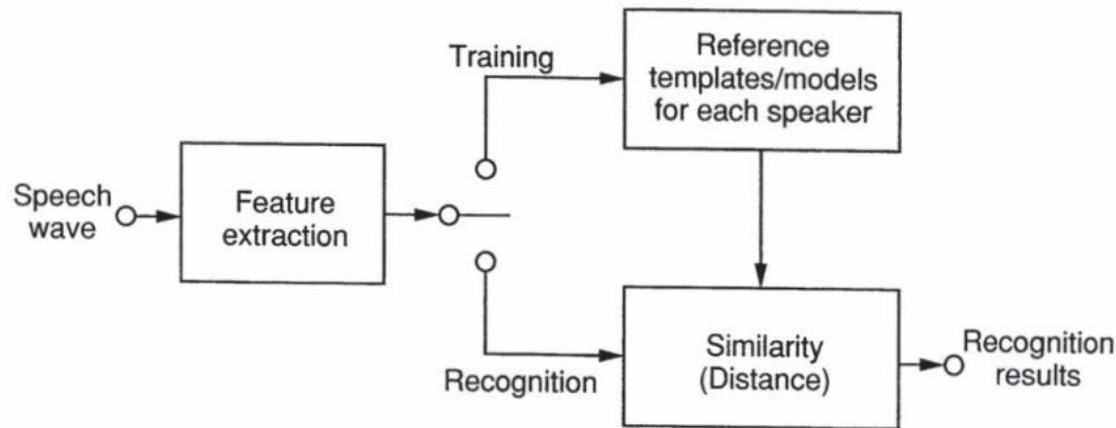
Individual characteristics

- **Speaker recognition** refers to the **automatic recognition** of a **speaker** (or talker) through measurements of characteristics arising in the speaker's **voice** signal.
- A spoken message conveys information about the speaker in addition to the meaning of the message.
- Individual speaker is characterized by a variety of voice attributes:
 - **High-level** attributes:
 - Prosody (i.e., pitch intonation)
 - Accents
 - Choice of words

-
- **Low-level** attributes:
 - Vocal tract spectrum
 - Pitch period
 - Formant trajectory
 - High-level attributes are related to the **behavioral** differences in the manner of speaking and are difficult to extract by machine for automatic speaker recognition (though fairly easy for human)
 - Low-level attributes are related to **physiological** aspect of vocal organ (mostly the vocal tract) and are more measurable given their acoustic nature.

Structure of speaker recognition system

- Feature parameters extracted from a speech wave are compared with stored **models** (or reference templates) for each registered speakers.
- The recognition decision is made according to the **distance** (or similarity) values. The distance measure is closely related to the type of model and algorithm used.



Input and decision modes

- **Input modes**

- **Text dependent:**

- The content of the speech is known (e.g., speaker prompted with text to speak)
 - More accurate

- **Text independent:**

- Unconstrained but less accurate
 - Can be applied to 'found' speech (no control over the content)

- **Decision modes**

- **Identification**

- Speech sample from an unknown speaker is compared with models of registered speakers. The unknown speaker is identified as the speaker whose model best matches the input speech sample.
 - One-to-many matching

– Verification

- An identity claim is made by an unknown speaker. Speech sample from the speaker is compared with the model of the claimed identity. The identity claim is accepted if the match is good enough (i.e., passes a given threshold).
- One-to-one matching
- For an **identification** task, speaker model with maximum score is selected.
- For a **verification** task, the decision is based on log-likelihood ratio of the following form

$$\Lambda(Y) = s(Y | \theta) - s(Y | \theta_{bg})$$

- θ is the speaker model while θ_{bg} represents a **background** model.
- The log-likelihood ratio $\Lambda(Y)$ is to be compared to a threshold α so as to **accept** (if $\Lambda(Y) \geq \alpha$) or **reject** (if $\Lambda(Y) < \alpha$) the identity claim.

Gaussian mixture model

- A speaker **model** (or reference template) is constructed using **enrollment** utterances from that speaker. Each enrollment utterance X is a **sequence of feature vectors** $\{\mathbf{x}_t\}_{t=1}^T$ generated by the feature extraction front-end.
- For the case of text-independent speaker recognition, where the system has no prior knowledge of the text of speaker's utterance, **Gaussian mixture models** (GMMs) have proven to be effective.

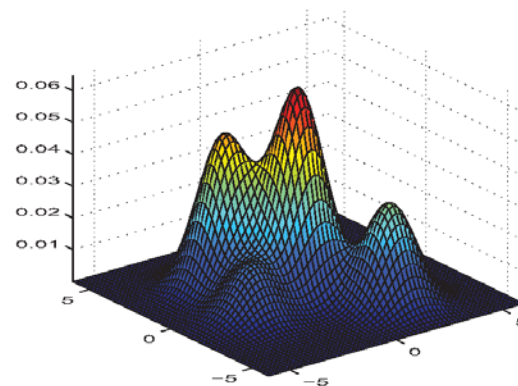
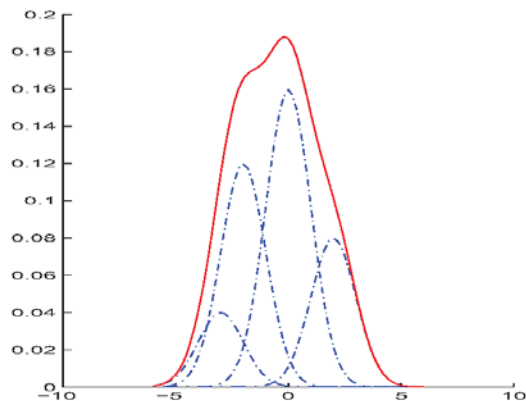
$$p(\mathbf{x} | \theta) = \sum_{k=1}^K w_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- The **weights** w_k , for $k = 1, 2, \dots, K$, always sum to 1 such that the resulting mixture $p(\mathbf{x} | \theta)$ is a legitimate probability distribution.
- The set $\theta = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, w_k\}_{k=1}^K$ represents the **parameters** of the distribution

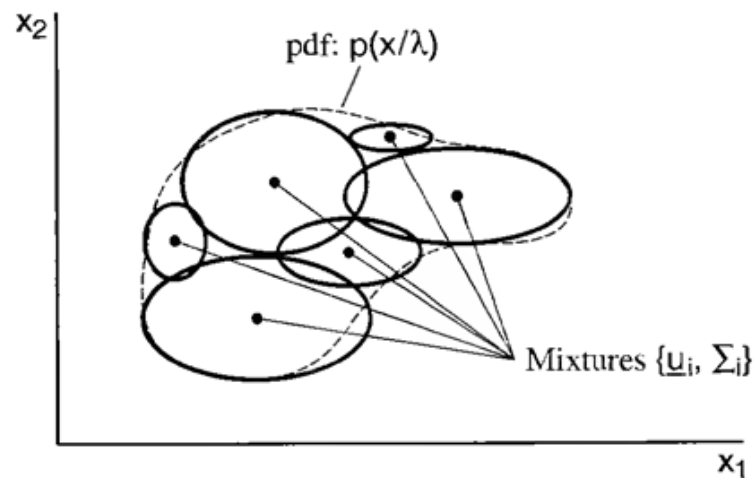
-
- The Gaussian or normal density is given by

$$\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}_k|}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right]$$

- Feature vectors of the enrollment utterance $\{\mathbf{x}_t\}_{t=1}^T$ are assumed to be drawn from a probability density function (pdf) that is a **mixture** of K Gaussians.



-
- A GMM can be interpreted as representation of various **acoustic classes** the make up the sounds of a speaker.
 - Each **component** density can be thought of as an **acoustic class**, each representing one speech sound (e.g., a particular phoneme or senone) or a set of speech sounds (voiced, unvoiced, fricative, diphthong etc.).



Expectation maximization (EM) for GMM

- Given the enrollment data $\{\mathbf{x}_t\}_{t=1}^T$, the maximum likelihood estimate of $\theta = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, w_k\}_{k=1}^K$ can be obtained using the EM algorithm.
- E-step:** compute the membership of each feature vector to the K Gaussians

$$\lambda_k(\mathbf{x}_t) = \frac{\mathcal{N}(\mathbf{x}_t | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \cdot w_k}{\sum_{k=1}^K \mathcal{N}(\mathbf{x}_t | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \cdot w_k}$$

- M-step:** Update the mean $\boldsymbol{\mu}_k$, covariance matrices $\boldsymbol{\Sigma}_k$, and weights w_k based on the membership information $\lambda_k(\mathbf{x}_t)$ of each frame

$$\boldsymbol{\mu}_k = \frac{1}{n_k} \times \sum_{t=1}^T \lambda_k(\mathbf{x}_t) \cdot \mathbf{x}_t$$

$$\boldsymbol{\Sigma}_k = \frac{1}{n_k} \times \sum_{t=1}^T \lambda_k(\mathbf{x}_t) \cdot (\mathbf{x}_t - \boldsymbol{\mu}_k)(\mathbf{x}_t - \boldsymbol{\mu}_k)^T$$

$$w_k = \frac{1}{T} \underbrace{\sum_{t=1}^T \lambda_k(\mathbf{x}_t)}_{n_k}$$

Speaker recognition

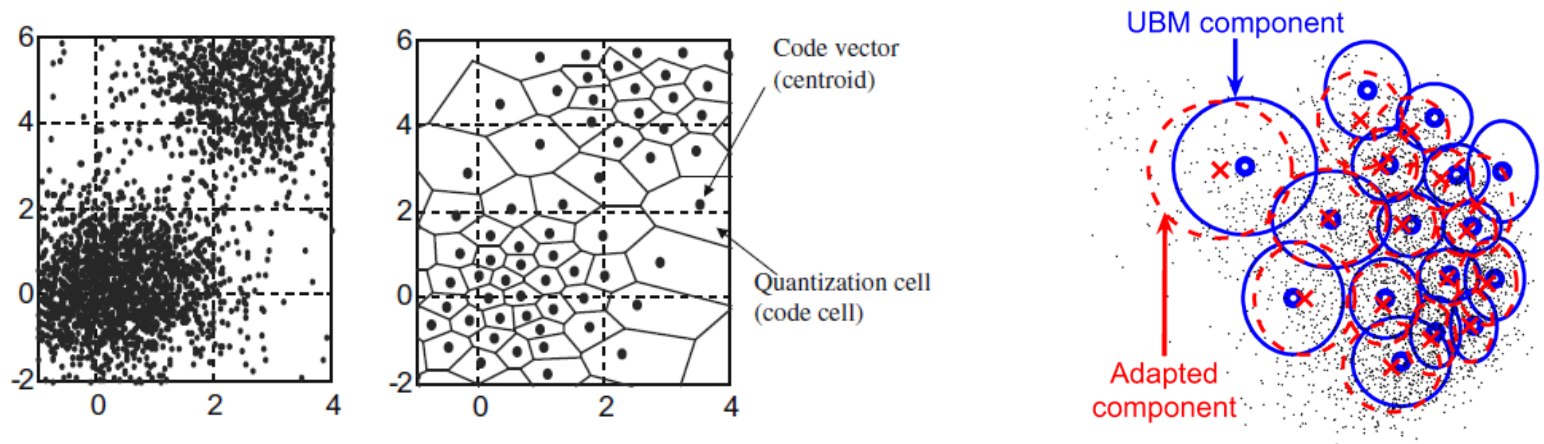
- Recognizing persons from their voices
- Decision modes
 - Identification (one-to-many matching)
 - Verification (one-to-one matching)
- Input modes
 - text-independent (e.g., NIST SREs)
 - text-dependent (fixed passphrase – common, unique, free-choice)
 - **RSR2015** [<https://www.etpl.sg/innovation-offerings/ready-to-sign-licenses/rsr2015-overview-n-specifications>]
 - ASVspoof2017
 - text-prompted (random)
 - **RedDots** project [<https://sites.google.com/site/thereddotsproject/>]

Factor Analysis for Speaker Recognition

I-VECTOR PLDA PIPELINE

Progresses in text-independent SR (1/3)

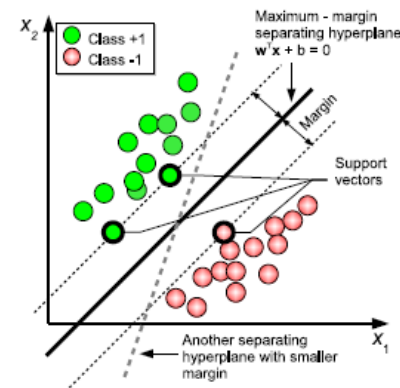
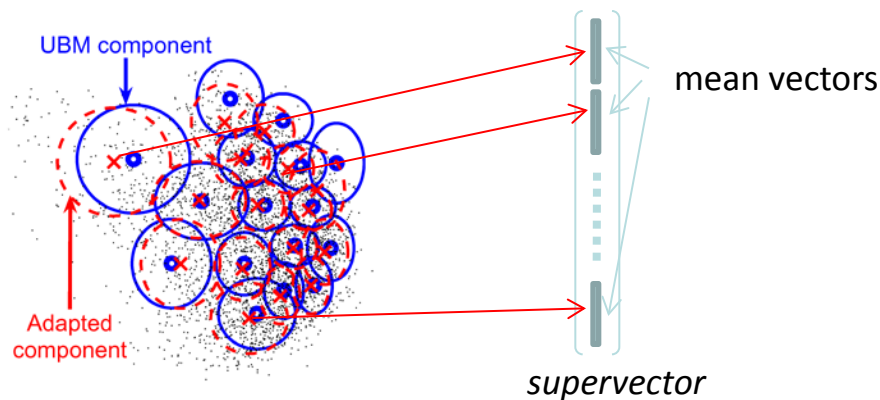
- Mostly driven by NIST Speaker Recognition Evaluations (SREs): 1996 – 2006, 2008, 2010, 2012, 2016.
- Vector Quantization (VQ)
- GMM-UBM using MAP adaptation [D. Reynolds, **2000**]



- <http://www.itl.nist.gov/iad/mig/tests/spk/>
- T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: from features to supervectors,” *Speech Communication*, vol. 52, no. 1, pp. 12-40, Jan. 2010.
- D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, 2000.

Progress in text-independent SR (2/3)

- Support vector machine (SVM)
 - GLDS (generalized linear discriminant analysis) [W. Campbell, **2002**]
 - **GMM KL kernel** [W. Campbell et al, **2006**]
 - GMM Bhattacharya kernel [C. You et al, **2009**]
 - Discrete kernel [Lee et al, **2011**]



- W. Campbell, “Generalized linear discriminant sequence kernel for speaker recognition,” in *Proc. ICASSP*, pp. 161-164, 2002.
- W. M. Campbell, D. E. Sturim, and D. A. Reynolds, “Support vector machines using GMM supervectors for speaker recognition,” *IEEE Signal Processing Lett.*, vol. 13, no. 5, pp. 308-311, May 2006.
- C. You, K. A. Lee, and H. Li, “An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition,” *IEEE Signal Processing Lett.*, vol. 16, no. 1, pp. 49-52, 2009.
- K. A. Lee, C. You, and H. Li, “Using discrete probabilities with Bhattacharyya measure for SVM-based speaker verification,” *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 4, pp. 861-870, May 2011.

Progress in text-independent SR (3/3)

- Subspace method
 - **Principle component analysis (PCA)**
 - Nuisance Attribute Projection (NAP) [W. Campbell et al, **2006**]
 - **Factor analysis (FA)**
 - Eigenchannel [Kenny et al, **Eurospeech 2003**]
 - Joint factor analysis (JFA) [Kenny et al, **TR 2006, TASL 2007, TASL 2008**]
 - Total variability model
 - I-vector cosine distance [Dehak et al, **IS 2009, TASL 2011**]
 - I-vector PLDA [Kenny et al, **Odyssey 2010**]
 - Whitening + Length normalization + PLDA [Garcia-Romero et al, **IS 2011**]
 - Local variability model [Chen et al, **Odyssey 2014**]
 - PLDA for multi-session [Lee et al, **IS 2013**] [Chen et al, **ICASSP 2014**]
 - Informative Prior [Sheptone et al, **ICASSP 2015**]
 - Rapid I-vector extraction [Xu et al, **Odyssey 2016**]
 - etc.

I-vector extraction

- **Compression** process – an i-vector is a fixed-length low-dimensional representation of a variable-length speech utterance [Dehak et al, 2011].
- **i-vector** = speaker + language + recording device + transmission channel + acoustic environment
- **MAP estimate** – posterior mean of the latent variable \mathbf{h} in a multi-Gaussian factor analysis model (i.e., total variability model)

$$\phi = \operatorname{argmax}_{\mathbf{h}} \left[\prod_{c=1}^C \prod_{t=1}^{N_c} \mathcal{N}(o_t | \boldsymbol{\mu}_c + \mathbf{W}_c \mathbf{h}, \boldsymbol{\Phi}_c) \right] p(\mathbf{h})$$

- The alignment of frames to Gaussian could be accomplished using GMM [Kenny et al, 2008] or senone posteriors [Lei *et al*, 2014].

The mechanic of i-vector extraction

- We need two sets of quantities in order to compute an i-vector:
 - Total variability matrix $\mathbf{T} = [\mathbf{W}_1; \mathbf{W}_2; \dots; \mathbf{W}_C]$, UBM's mean vectors $\boldsymbol{\mu}_c$ and covariance matrices $\boldsymbol{\Phi}_c$.
 - Zero and first order sufficient statistics computed from feature vector sequences
- Given an observed sequence of feature vectors $\{o_1, o_2, \dots, o_T\}$

First-order statistics (centred)

i-vector \rightarrow

$$\phi = \underbrace{\left(\sum_{c=1}^C N_c \mathbf{W}_c^T \boldsymbol{\Phi}_c^{-1} \mathbf{W}_c + I \right)}_{\mathbf{L}^{-1}} \sum_{c=1}^C \mathbf{W}_c^T \boldsymbol{\Phi}_c^{-1} \left(\sum_{t=1}^T \gamma_t(o_t - \boldsymbol{\mu}_c) \right)$$

Zero-order statistics

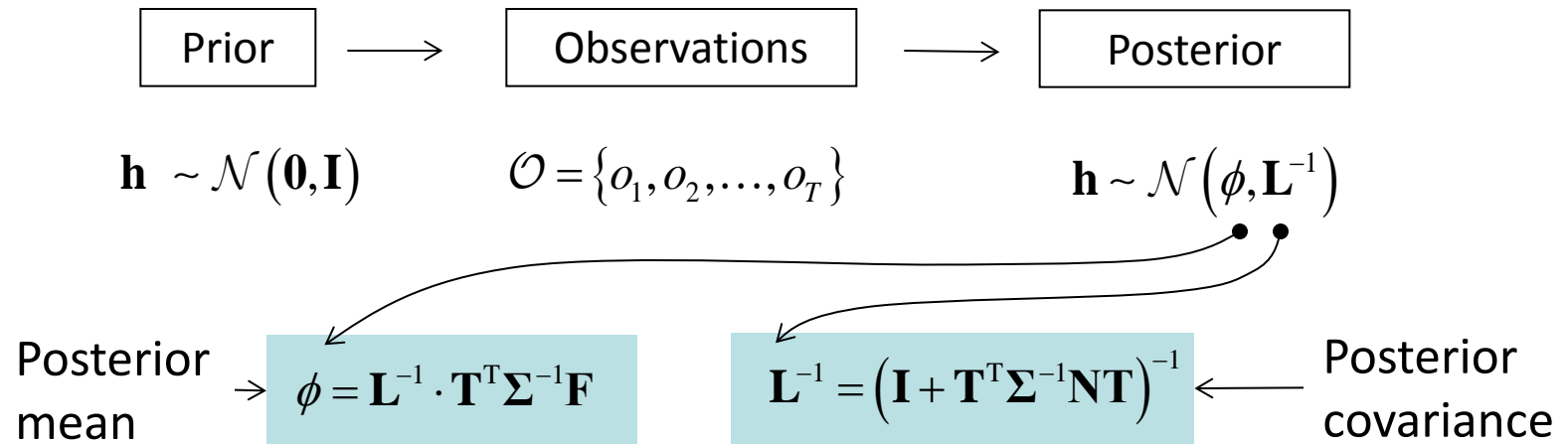
Alignment of frame o_t to Gaussian c :

$$N_{r,c} = \sum_{t=1}^T \underbrace{\gamma_t(c)}_{\rightarrow}$$

$$\gamma_t(c) = \frac{\omega_c \mathcal{N}(o_t | \boldsymbol{\mu}_c, \boldsymbol{\Phi}_c)}{\sum_{j=1}^C \omega_j \mathcal{N}(o_t | \boldsymbol{\mu}_j, \boldsymbol{\Phi}_j)}$$

Posterior mean estimation

- l-vector extraction is a **posterior inference** process

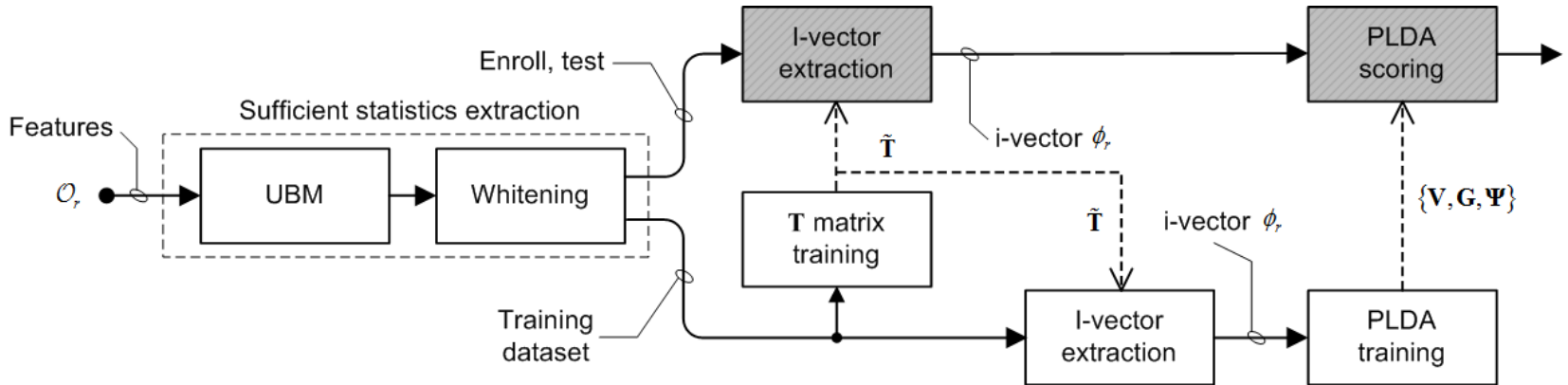


- Vector notation**

$$\begin{aligned} \boldsymbol{\phi} &= \left(\sum_{c=1}^C N_c \mathbf{W}_c^T \boldsymbol{\Phi}_c^{-1} \mathbf{W}_c + \mathbf{I} \right)^{-1} \sum_{c=1}^C \mathbf{W}_c^T \boldsymbol{\Phi}_c^{-1} \left(\sum_{t=1}^T \gamma_t (o_t - \boldsymbol{\mu}_c) \right) \\ &= \underbrace{(\mathbf{T}^T \boldsymbol{\Sigma}^{-1} \mathbf{N} \mathbf{T} + \mathbf{I})^{-1}}_{\mathbf{L}^{-1}} \cdot \mathbf{T}^T \boldsymbol{\Sigma}^{-1} (\mathbf{F} - \mathbf{N} \mathbf{m}_o) \end{aligned}$$

I-vector PLDA pipeline

- I-vector extraction followed by PLDA scoring



[K. A. Lee and H. Li, Interspeech 2017]

- Pre-whitened statistics [Matejka *et al*, 2011]

$$\begin{aligned}\phi_r &= (\mathbf{T}^T \Sigma^{-1} \mathbf{N}_r \mathbf{T} + \mathbf{I})^{-1} \cdot \mathbf{T}^T \Sigma^{-1} (\mathbf{F}_r - \mathbf{N}_r \mathbf{m}_o) \\ &= (\tilde{\mathbf{T}}^T \tilde{\mathbf{N}}_r \tilde{\mathbf{T}} + \mathbf{I})^{-1} \cdot \tilde{\mathbf{T}}^T \tilde{\mathbf{F}}_r\end{aligned}$$

Factor Analysis for Speaker Recognition

FACTOR ANALYSIS

Normal (Gaussian) distribution

- A Gaussian distribution is defined by its mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
- Consider a set of T observations of D dimensions, $O = \{o_1, o_2, \dots, o_T\}$ where $o_t \in \mathbb{R}^D$, we are interested in fitting a Gaussian density function to O .
- Formally, we intend to estimate the set of parameters $\{\boldsymbol{\mu}_o, \boldsymbol{\Sigma}\}$ that best describes the observations, as follow

$$p(o_t) = N(o_t | \boldsymbol{\mu}_o, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp \left[-\frac{1}{2} (o_t - \boldsymbol{\mu}_o)^T \boldsymbol{\Sigma}^{-1} (o_t - \boldsymbol{\mu}_o) \right]$$

- The maximum likelihood estimates of $\{\boldsymbol{\mu}_o, \boldsymbol{\Sigma}\}$ are given by the sample mean vector and covariance matrix computed from O .
- In our application, the observations consist of the feature vectors, e.g., the MFCC feature vector.

Full vs. diagonal covariance matrices

- In the case where the training data is **sparse**, the covariance matrix could be restricted to be of **diagonal** form.
- Naive Bayes assumption – the D dimensions of the observed vector o are **uncorrelated**.
- Setting all the diagonal elements to have equal value, $\Sigma = \alpha \mathbf{I}$, leads to an isotropic covariance matrix.
- Diagonal covariance matrix
 - Reduces the **number of parameters** to be estimated, stored, and manipulated
 - **Inversion** of matrix is much simpler than that of a full covariance matrix
 - Practically useful for the case when D is extremely large, though restrictive in its representational power

Structured covariance matrix

- The goal is to model the covariance with a complexity **controllable** between a **diagonal** and a **full** covariance form:

$$p(o) = \mathcal{N}(o | \boldsymbol{\mu}_o, \underbrace{\mathbf{W}\mathbf{W}^T}_{\boldsymbol{\Sigma}} + \boldsymbol{\Phi})$$


- \mathbf{W} is a $K \times D$ **rectangular** matrix. The columns of \mathbf{W} span the subspace in which the observed features correlate the most.
- $\boldsymbol{\Phi}$ is a **diagonal** matrix accounts for the remaining variations (residual covariance).
- The dimensionality K of the subspace is kept much lower than D of the observed feature vector.
- \mathbf{W} is a tall and thin rectangular matrix in portrait form that models the major part of the correlation.
- No closed-form solution is available to estimate $\{\mathbf{W}, \boldsymbol{\Phi}\}$

Factor analysis

- Structured covariance matrix is achieved with a factor analysis model.
- Introduce the latent variable \mathbf{h} through a conditional distribution:

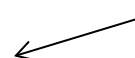
$$p(o) = \int p(o|\mathbf{h}) p(\mathbf{h}) d\mathbf{h}$$



$p(o|\mathbf{h}) = \mathcal{N}(o|\boldsymbol{\mu}_o + \mathbf{W}\mathbf{h}, \Phi)$



- Assuming a standard normal prior on the latent variable \mathbf{h} , we obtain a Gaussian distribution from marginalization:

$$p(o) = \int \mathcal{N}(o|\boldsymbol{\mu}_o + \mathbf{W}\mathbf{h}, \Phi) \mathcal{N}(\mathbf{h}|0, I) d\mathbf{h}$$

Prior distribution 

Conditional distribution  $= \mathcal{N}(o|\boldsymbol{\mu}_o, \mathbf{W}\mathbf{W}^T + \Phi)$ Marginal distribution 

- The latent variable \mathbf{h} has no significant role except in the E-step of the parameter estimation.

Prior and posterior

- Given an observation o , the posterior distribution is Gaussian with mean \mathbf{m} and covariance \mathbf{L}^{-1} (**Note**: one latent variable for each observed vector)

| Prior | Posterior |
|---|--|
| $p(\mathbf{h}) = \mathcal{N}(\mathbf{h} \mathbf{0}, \mathbf{I})$ | $p(\mathbf{h} o) = \mathcal{N}(\mathbf{h} \mathbf{m}, \mathbf{L}^{-1})$ $\mathbf{m} = \mathbf{L}^{-1} \mathbf{W}^T \mathbf{\Phi}^{-1} (o - \boldsymbol{\mu}_o)$ $\mathbf{L}^{-1} = (\mathbf{W}^T \mathbf{\Phi}^{-1} \mathbf{W} + \mathbf{I})^{-1}$ |
| $p(\mathbf{h}) = \mathcal{N}(\mathbf{h} \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$ | $p(\mathbf{h} o) = \mathcal{N}(\mathbf{h} \mathbf{m}, \mathbf{L}^{-1})$ $\mathbf{m} = \mathbf{L}^{-1} [\mathbf{W}^T \mathbf{\Phi}^{-1} (o - \boldsymbol{\mu}_o) + \boldsymbol{\Sigma}_h^{-1} \boldsymbol{\mu}_h]$ $\mathbf{L}^{-1} = (\mathbf{W}^T \mathbf{\Phi}^{-1} \mathbf{W} + \boldsymbol{\Sigma}_h^{-1})^{-1}$ |

Non-standard Gaussian prior

- The latent variable is assumed to follow a **standard Gaussian prior**.
- A **non-standard Gaussian prior** could always be absorbed by the mean and the subspace matrix

$$\begin{aligned} p(o) &= \int \mathcal{N}(o | \boldsymbol{\mu}_o + \mathbf{W}\mathbf{h}, \boldsymbol{\Phi}) \mathcal{N}(\mathbf{h} | \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h) d\mathbf{h} \\ &= \mathcal{N}(o | \boldsymbol{\mu}_o + \mathbf{W}\boldsymbol{\mu}_h, \mathbf{W}\boldsymbol{\Sigma}_h\mathbf{W}^T + \boldsymbol{\Phi}) \\ &= \mathcal{N}(o | \tilde{\boldsymbol{\mu}}_o, \tilde{\mathbf{W}}\tilde{\mathbf{W}}^T + \boldsymbol{\Phi}) \\ &= \int \mathcal{N}(o | \tilde{\boldsymbol{\mu}}_o + \tilde{\mathbf{W}}\mathbf{h}, \boldsymbol{\Phi}) \mathcal{N}(\mathbf{h} | 0, \mathbf{I}) d\mathbf{h} \end{aligned}$$

- **Multiple** non-standard Gaussian priors are useful, for instance, to model and compensate for the negative effects of heterogeneous dataset.

-
- S. E. Shepstone, K. A. Lee, H. Li, Z.-H. Tan, S. H. Jensen, “Source-specific informative prior for i-vector extraction,” in *Proc. ICASSP*, pp. 4185 – 4189, Apr. 2015.
 - S. E. Shepstone, K. A. Lee, H. Li, Z.-H. Tan, S. H. Jensen, “Total variability modeling using source-specific priors,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 3, pp. 504 - 517, Mar 2016.
-

Parameter estimation with EM

- Likelihood function

$$l(O) = \prod_{t=1}^T \int \mathcal{N}(o_t | \boldsymbol{\mu}_o + \mathbf{W}\mathbf{h}_t, \boldsymbol{\Phi}) \mathcal{N}(\mathbf{h}_t | 0, \mathbf{I}) d\mathbf{h}$$

- Maximum likelihood estimation

- **E-step** (Expectation, posterior inference)

$$\mathbf{E}[\mathbf{h}_t] = \mathbf{L}^{-1} \mathbf{W}^T \boldsymbol{\Phi}^{-1} (o_t - \boldsymbol{\mu}_o)$$

$$\mathbf{L}^{-1} = (\mathbf{W}^T \boldsymbol{\Phi}^{-1} \mathbf{W} + \mathbf{I})^{-1} \Rightarrow \mathbf{E}[\mathbf{h}_t \mathbf{h}_t^T] = \mathbf{L}^{-1} + \mathbf{E}[\mathbf{h}_t] \mathbf{E}[\mathbf{h}_t^T]$$

- **M-step** (Maximization, parameter update)

$$\mathbf{W} = \left(\sum_{t=1}^T (o_t - \boldsymbol{\mu}_o) \mathbf{E}[\mathbf{h}_t^T]^T \right) \left(\sum_{t=1}^T \mathbf{E}[\mathbf{h}_t \mathbf{h}_t^T] \right)^{-1}$$

$$\boldsymbol{\Phi} = \frac{1}{T} \text{diag} \left[\sum_{t=1}^T (o_t - \boldsymbol{\mu}_o) (o_t - \boldsymbol{\mu}_o)^T - \mathbf{W} \mathbf{E}[\mathbf{h}_t] (o_t - \boldsymbol{\mu}_o)^T \right]$$

Generative process

- A convenient way to look at a factor analysis model is by expressing the **observed variable** as

$$o_t = \boldsymbol{\mu}_o + \mathbf{W}\mathbf{h}_t + \boldsymbol{\varepsilon}_t \quad \text{for } t = 1, 2, \dots, T$$

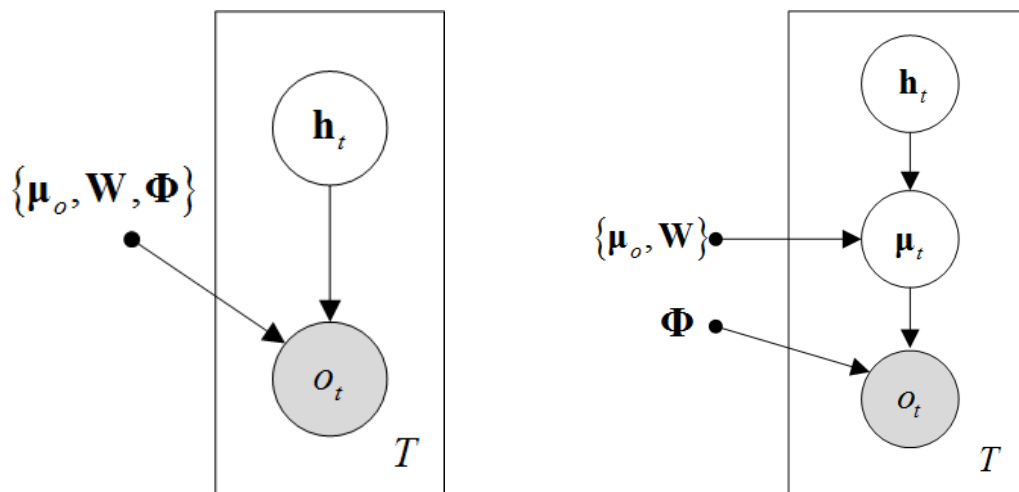
- Taking the expectation on both sides, we obtain the mean and covariance of the marginal density $p(o)$

$$\begin{aligned} E\{o_t\} &= E\{\boldsymbol{\mu}_o + \mathbf{W}\mathbf{h}_t + \boldsymbol{\varepsilon}_t\}, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \boldsymbol{\Phi}) \longleftarrow \text{Residual covariance} \\ &= E\{\boldsymbol{\mu}_o\} + \mathbf{W}E\{\mathbf{h}_t\} + E\{\boldsymbol{\varepsilon}_t\} = \boldsymbol{\mu}_o \end{aligned}$$

$$\begin{aligned} \boldsymbol{\Sigma} &= E\left\{ (o_t - \boldsymbol{\mu}_o)(o_t - \boldsymbol{\mu}_o)^T \right\} \\ &= E\left\{ (\mathbf{W}\mathbf{h}_t + \boldsymbol{\varepsilon}_t)(\mathbf{W}\mathbf{h}_t + \boldsymbol{\varepsilon}_t)^T \right\} \\ &= \underbrace{\mathbf{W} E\{\mathbf{h}_t \mathbf{h}_t^T\}}_{\mathbf{I}} \mathbf{W}^T + 2\mathbf{W} E\{\mathbf{h}_t \boldsymbol{\varepsilon}_t^T\} + \underbrace{E\{\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^T\}}_{\boldsymbol{\Phi}} = \mathbf{W}\mathbf{W}^T + \boldsymbol{\Phi} \end{aligned}$$

Graphical model representation

- Ancestral sampling process:
 - For each observation o_t , we first draw a sample \mathbf{h}_t from a standard normal prior $N(\mathbf{h}|0, I)$.
 - The observed vector o_t is then produced as a sample from $N(o|\boldsymbol{\mu}_o + \mathbf{W}\mathbf{h}_t, \boldsymbol{\Phi})$.



- Alternatively, we could also look at the intermediate variable $\boldsymbol{\mu}_t = \boldsymbol{\mu}_o + \mathbf{W}\mathbf{h}_t$ lies in the affine subspace span by \mathbf{W}

Factor Analysis for Speaker Recognition

TOTAL VARIABILITY MODEL

Latent variable tying across frames (1/2)

- A factor analysis model is defined per observation basis in the sense that one latent variable is drawn for each observed frame.

$$l(O) = \prod_{t=1}^T \int \mathcal{N}(o_t | \boldsymbol{\mu}_o + \mathbf{W}\mathbf{h}_t, \boldsymbol{\Phi}) \mathcal{N}(\mathbf{h}_t | 0, I) d\mathbf{h}$$

- Each observation o_t is associated with a latent variable \mathbf{h}_t . The feature vector sequence of an utterance will be described with a sequence of latent variable.
- **Latent variable tying** – associate one latent variable \mathbf{h} to all frames $O = \{o_1, o_2, \dots, o_T\}$ in an utterance, i.e., tying of the observed frames conditioned on the same latent variable)

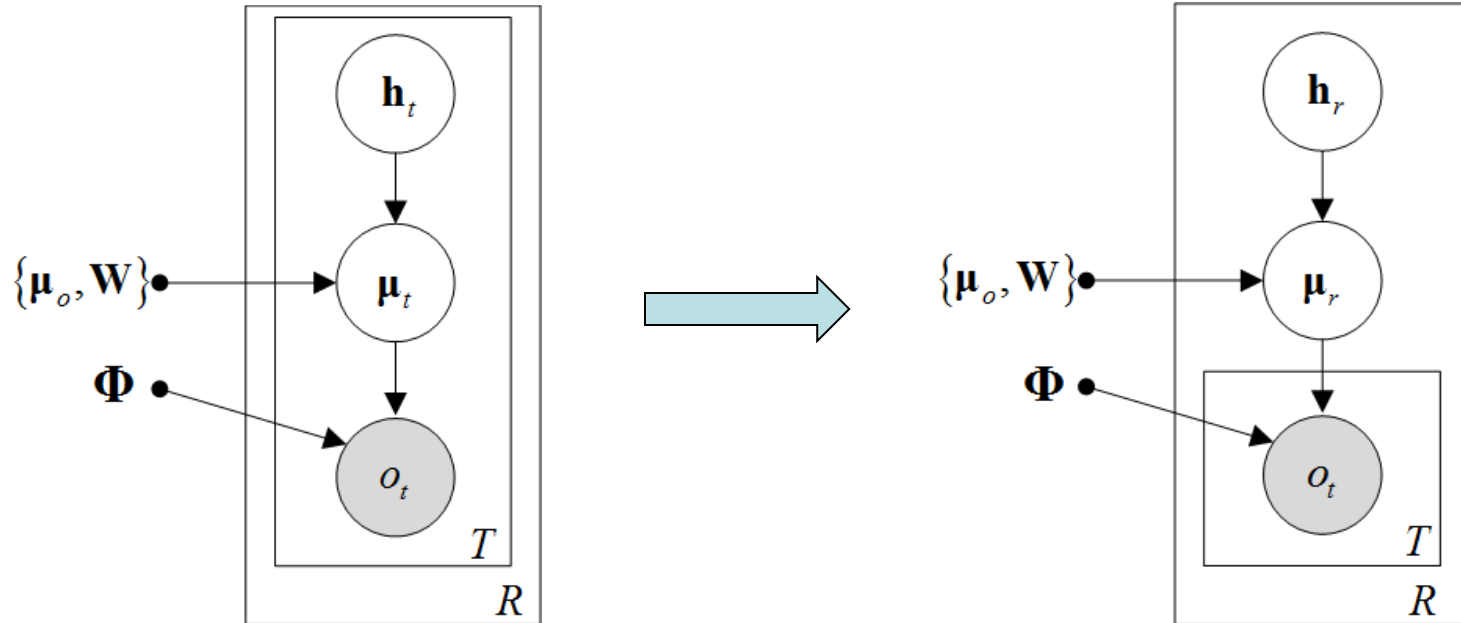
$$l(O) = \int \left(\prod_{t=1}^T \mathcal{N}(o_t | \boldsymbol{\mu}_o + \mathbf{W}\mathbf{h}, \boldsymbol{\Phi}) \right) \mathcal{N}(\mathbf{h} | 0, I) d\mathbf{h}$$

Latent variable tying across frames (2/2)

- Our intention of using **one** latent variable is to describe, for instance, speaker, language or channel effects, which are homogeneous throughout the entire utterance O .
- Parameter sharing (or tying) – sharing of parameters across different parts of a model – an useful idea found in machine learning in the 80s.
- For a given R number of utterances, we assign one latent variable \mathbf{h}_r to each utterance O_r

$$l(O_1, O_2, \dots, O_R) = \prod_{r=1}^R \int \left(\prod_{t=1}^T \mathcal{N}(o_t | \boldsymbol{\mu}_o + \mathbf{W}\mathbf{h}_r, \Phi) \right) \mathcal{N}(\mathbf{h}_r | 0, I) d\mathbf{h}$$

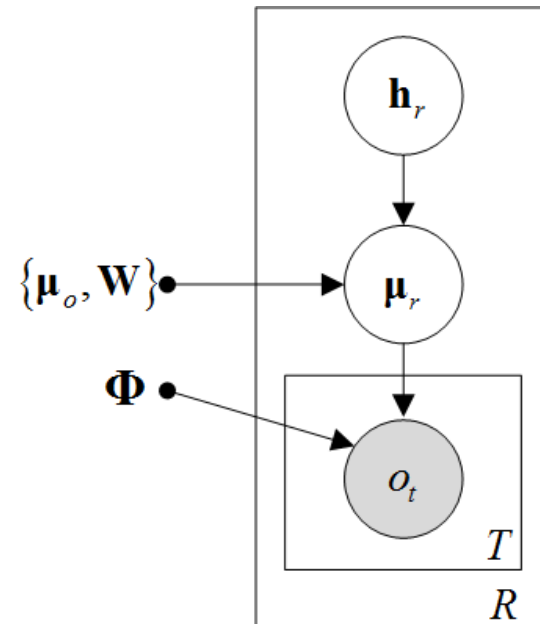
Graphical model representation (1/2)



| | |
|---------------------------|--|
| Classical factor analysis | $l(O_1, O_2, \dots, O_R) = \prod_{r=1}^R \prod_{t=1}^T \int \mathcal{N}(o_t \mu_o + \mathbf{W}\mathbf{h}_t, \Phi) \mathcal{N}(\mathbf{h}_t 0, I) d\mathbf{h}$ |
| TVM (single Gaussian) | $l(O_1, O_2, \dots, O_R) = \prod_{r=1}^R \int \left(\prod_{o_t \in O_r} \mathcal{N}(o_t \mu_o + \mathbf{W}\mathbf{h}_r, \Phi) \right) \mathcal{N}(\mathbf{h}_r 0, I) d\mathbf{h}$ |

Graphical model representation (2/2)

- Total variability model (single Gaussian)
 - The inner rectangular box with a value T indicates that there are T observed vectors per latent variable \mathbf{h}_r .
 - The outer box with a value R indicates the number of utterances.
 - A separate latent variable \mathbf{h}_r was drawn for each utterance r , where $r = 1, 2, \dots, R$.
 - For brevity of notations, we have assumed that each utterance consists of T frames.
 - The values of T depend on the duration of the utterances which are generally different in actual implementation.



TVM with single Gaussian (1/3)

- Factor analysis (**generative equation**):

$$o_t = \boldsymbol{\mu}_o + \mathbf{W}\mathbf{h}_t + \boldsymbol{\varepsilon}_t \text{ for } t=1,2,\dots,T$$

- Factor analysis with tying across frames (associate one latent variable to all frames in an utterance):

$$o_t = \boldsymbol{\mu}_o + \mathbf{W}\mathbf{h} + \boldsymbol{\varepsilon}_t \text{ for } t=1,2,\dots,T$$

- Stacking all the observed frames in an utterance, we obtain a **generative equation** that takes the form of classical factor analysis model:

$$\begin{bmatrix} o_1 \\ o_2 \\ \vdots \\ o_T \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_o \\ \boldsymbol{\mu}_o \\ \vdots \\ \boldsymbol{\mu}_o \end{bmatrix} + \begin{bmatrix} \mathbf{W} \\ \mathbf{W} \\ \vdots \\ \mathbf{W} \end{bmatrix} \mathbf{h}_r + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_T \end{bmatrix} \Rightarrow \tilde{o}_r = \tilde{\boldsymbol{\mu}}_o + \tilde{\mathbf{W}}\mathbf{h}_r + \tilde{\boldsymbol{\varepsilon}}_t \text{ for } r=1,2,\dots,R$$

TVM with single Gaussian (2/3)

- In probabilistic form, we obtain the following **conditional** and **prior** probabilities in a similar form as classical factor analysis

$$p(\tilde{o}_r | \mathbf{h}_r) = \mathcal{N}(\tilde{o} | \tilde{\boldsymbol{\mu}}_o + \tilde{\mathbf{W}}\mathbf{h}_r, \tilde{\boldsymbol{\Phi}})$$

$$p(\mathbf{h}_r) = \mathcal{N}(\mathbf{h}_r | 0, \mathbf{I})$$

- The observed variable \tilde{o}_r is an $DT \times 1$ vector containing all T frames.
- Using the result from classical FA, we have the posterior inference given by

$$\mathbb{E}[\mathbf{h}_r] = (\tilde{\mathbf{W}}^T \tilde{\boldsymbol{\Phi}}^{-1} \tilde{\mathbf{W}} + \mathbf{I})^{-1} \tilde{\mathbf{W}}^T \tilde{\boldsymbol{\Phi}}^{-1} (\tilde{o}_r - \tilde{\boldsymbol{\mu}}_o)$$

- Solve out the compound terms, the **posterior mean** and **covariance** can be obtained:

$$\mathbb{E}[\mathbf{h}_r] = \underbrace{\left(\sum_{t=1}^T \mathbf{W}^T \boldsymbol{\Phi}^{-1} \mathbf{W} + \mathbf{I} \right)^{-1}}_{\mathbf{L}_r^{-1}} \left(\sum_{t=1}^T \mathbf{W}^T \boldsymbol{\Phi}^{-1} (o_t - \boldsymbol{\mu}_o) \right)$$

TVM with single Gaussian (3/3)

- The **posterior covariance** L^{-1} become sharper for larger T :

$$\begin{aligned} E[\mathbf{h}_r] &= \left(\sum_{t=1}^T \mathbf{W}^T \mathbf{\Phi}^{-1} \mathbf{W} + \mathbf{I} \right)^{-1} \left(\sum_{t=1}^T \mathbf{W}^T \mathbf{\Phi}^{-1} (o_t - \boldsymbol{\mu}_o) \right) \\ &= \underbrace{\left(T \mathbf{W}^T \mathbf{\Phi}^{-1} \mathbf{W} + \mathbf{I} \right)^{-1}}_{\mathbf{L}_r^{-1}} \mathbf{W}^T \mathbf{\Phi}^{-1} \sum_{t=1}^T (o_t - \boldsymbol{\mu}_o) \end{aligned}$$

 **i-vector (single Gaussian)**

- Assumptions:
 - Conditional independent of observed frames (on the same latent variable)
 - The compound residual covariance matrix $\mathbf{\Phi}$ is at most block diagonal

Tying across frames and mixtures

- Consider R utterances and one Gaussian distribution, tying a latent variable \mathbf{h}_r across frames in an utterance, we obtain:

$$l(O_1, O_2, \dots, O_R) = \prod_{r=1}^R \int \left(\prod_{t=1}^T \mathcal{N}(o_t | \boldsymbol{\mu}_o + \mathbf{W}\mathbf{h}_r, \boldsymbol{\Phi}) \right) \mathcal{N}(\mathbf{h}_r | 0, I) d\mathbf{h}$$

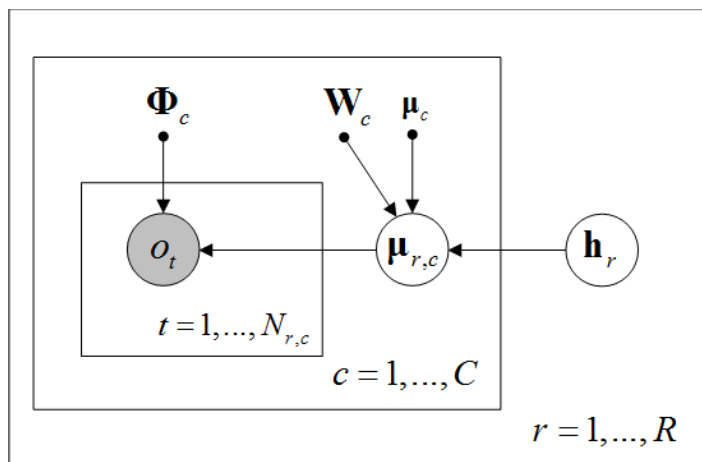
- Consider R utterances and C Gaussian components, tying \mathbf{h}_r across frames in an utterance and Gaussians, we arrive at the **total variability model** (TVM) with multiple Gaussian

$$l(O_1, O_2, \dots, O_R) = \prod_{r=1}^R \int \left(\prod_{c=1}^C \prod_{t=1}^{N_{r,c}} \mathcal{N}(o_t | \boldsymbol{\mu}_c + \mathbf{W}_c \mathbf{h}_r, \boldsymbol{\Phi}_c) \right) \mathcal{N}(\mathbf{h}_r | 0, I) d\mathbf{h}$$

- For each utterance, latent variable \mathbf{h}_r is used for all C mixtures and $N_{r,c}$ frames from each mixture, assuming we know the alignment of frames to mixtures.

Graphical model representation

- A **total variability model** (TVM) is best described as a **tied factor analysis model** (a bag of Gaussians) instead of a Gaussian mixture model (GMM). No GMM weights are involved, except when computing the sufficient statistics.
- The latent variable \mathbf{h}_r is tied/shared across the observed frames and the Gaussian components.



$$p(\mathbf{h}_r) = \mathcal{N}(\mathbf{h}_r | 0, I)$$
$$p(o_t | c) = \mathcal{N}(o_t | \underbrace{\mu_c + \mathbf{W}_c \mathbf{h}_r}_{\mu_{r,c}}, \Phi_c)$$

- L. Chen, K. A. Lee, B. Ma, W. Guo, H. Li, and L. -R. Dai, "Exploration of local variability in text-independent speaker verification," J. Sign. Process. Syst., vol. 82, no. 2, pp 217–228, Feb. 2016.

TVM with multiple Gaussians (1/3)

- Factor analysis is **performed on feature vector** o_t with additional tying across frames and Gaussian components sharing the same latent variable.

$$o_t = \boldsymbol{\mu}_c + \mathbf{W}_c \mathbf{h}_r + \boldsymbol{\varepsilon}_t \quad \text{for } t=1,2,\dots,T \text{ and } c \in \{1,2,\dots,C\}$$

- In probabilistic terms (assuming we know the alignment of frames to Gaussian):

$$p(o_t | c) = \mathcal{N}(o_t | \underbrace{\boldsymbol{\mu}_c + \mathbf{W}_c \mathbf{h}_r}_{\boldsymbol{\mu}_{r,c}}, \Phi_c)$$

$$\boldsymbol{\mu}_{r,c} = \boldsymbol{\mu}_c + \mathbf{W}_c \mathbf{h}_r \quad \text{for } c = 1, 2, \dots, C$$

$$\begin{bmatrix} \boldsymbol{\mu}_{r,1} \\ \vdots \\ \boldsymbol{\mu}_{r,C} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_C \end{bmatrix} + \begin{bmatrix} \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_C \end{bmatrix} \mathbf{h}_r \Rightarrow \boxed{\mathbf{m}_r = \mathbf{m}_o + \mathbf{T} \mathbf{h}_r}$$

mean supervector Total variability matrix
Latent variable

TVM with multiple Gaussians (2/3)

- Re-write the generative equation

$$o_t = \boldsymbol{\mu}_c + \mathbf{W}_c \mathbf{h}_r + \boldsymbol{\varepsilon}_t \text{ for } t=1,2,\dots,T \text{ and } c \in \{1,2,\dots,C\}$$

to resemble the classical factor analysis model by stacking up all the T observed vectors and assuming we know the alignment of frames $\{o_1, o_2, \dots, o_T\}$ to mixtures:

$$\begin{bmatrix} o_1 \\ \vdots \\ o_t \\ \vdots \\ o_T \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_c \\ \vdots \\ \boldsymbol{\mu}_C \end{bmatrix} + \begin{bmatrix} \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_c \\ \vdots \\ \mathbf{W}_C \end{bmatrix} \mathbf{h}_r + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_c \\ \vdots \\ \boldsymbol{\varepsilon}_C \end{bmatrix} \Rightarrow \tilde{o}_r = \tilde{\boldsymbol{\mu}}_o + \tilde{\mathbf{W}} \mathbf{h}_r + \tilde{\boldsymbol{\varepsilon}}_t \text{ for } r=1,2,\dots,R$$

- The same $\boldsymbol{\mu}_c$ or \mathbf{W}_c appears $N_{r,c}$ times depending on the alignment of frames to Gaussians.

TVM with multiple Gaussians (3/3)

- Using the result from classical FA, we have the posterior inference given by

$$\begin{aligned}
 E[\mathbf{h}_r] &= (\tilde{\mathbf{W}}^T \tilde{\mathbf{\Phi}}^{-1} \tilde{\mathbf{W}} + \mathbf{I})^{-1} \tilde{\mathbf{W}}^T \tilde{\mathbf{\Phi}}^{-1} (\tilde{o}_r - \tilde{\boldsymbol{\mu}}_o) \\
 &= \left(\sum_{c=1}^C N_{r,c} \mathbf{W}_c^T \mathbf{\Phi}_c^{-1} \mathbf{W}_c + \mathbf{I} \right)^{-1} \left(\sum_{c=1}^C \mathbf{W}_c^T \mathbf{\Phi}_c^{-1} \sum_{t=1}^{N_{r,c}} (o_t - \boldsymbol{\mu}_c) \right) \\
 \phi &= \underbrace{(\mathbf{T}^T \boldsymbol{\Sigma}^{-1} \mathbf{N}_r \mathbf{T} + \mathbf{I})^{-1}}_{\mathbf{L}_r^{-1}} \mathbf{T}^T \boldsymbol{\Sigma}^{-1} (\mathbf{F}_r - \mathbf{N}_r \mathbf{m}_o) \quad \leftarrow \text{i-vector = posterior mean of the latent variable}
 \end{aligned}$$

- The total variability matrix \mathbf{T} is obtained by stacking up the loading matrices from each Gaussian:

UBM mean vectors and covariance matrices

$$\mathbf{T} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \vdots \\ \mathbf{W}_C \end{bmatrix} \quad \mathbf{m}_o = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_C \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Phi}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{\Phi}_2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & \boldsymbol{\Phi}_C \end{bmatrix}$$

Factor analysis vs. total variability model

- Factor analysis

$$\mathbf{E}[\mathbf{h}_t] = \underbrace{(\mathbf{W}^T \mathbf{\Phi}^{-1} \mathbf{W} + \mathbf{I})^{-1}}_{\mathbf{L}^{-1}} \mathbf{W}^T \mathbf{\Phi}^{-1} (o_t - \boldsymbol{\mu}_o)$$

- Total variability model (single Gaussian)

$$\mathbf{E}[\mathbf{h}_r] = \underbrace{(T \mathbf{W}^T \mathbf{\Phi}^{-1} \mathbf{W} + \mathbf{I})^{-1}}_{\mathbf{L}_r^{-1}} \mathbf{W}^T \mathbf{\Phi}^{-1} \underbrace{\sum_{t=1}^T (o_t - \boldsymbol{\mu}_o)}_{\text{Zero-order statistics}}$$

Zero-order statistics

- Total variability model (multi Gaussian)

First-order statistics (centred)

$$\mathbf{E}[\mathbf{h}_r] = \underbrace{\left(\sum_{c=1}^C N_{r,c} \mathbf{W}_c^T \mathbf{\Phi}_c^{-1} \mathbf{W}_c + I \right)^{-1}}_{\mathbf{L}_r^{-1}} \sum_{c=1}^C \mathbf{W}_c^T \mathbf{\Phi}_c^{-1} \underbrace{\left(\sum_{t=1}^{N_c} (o_t - \boldsymbol{\mu}_c) \right)}_{\text{i-vector}}$$

Zero-order statistics

Parameter estimation with EM

- Factor analysis

$$\mathbf{W} = \left(\sum_{t=1}^T (o_t - \mu_o) E[\mathbf{h}_t^T] \right) \left(\sum_{t=1}^T E[\mathbf{h}_t \mathbf{h}_t^T] \right)^{-1}$$

One latent variable per frame

- Total variability model (single Gaussian)

$$\mathbf{W} = \left(\sum_{r=1}^R \underbrace{\left[\sum_{t=1}^T (o_{r,t} - \mu_o) \right]} E[\mathbf{h}_r^T] \right) \left(\sum_{r=1}^R \underbrace{T}_{\uparrow} E[\mathbf{h}_r \mathbf{h}_r^T] \right)^{-1}$$

First-order statistics (centred)

Zero-order statistics
(number of frames)

- Total variability model (multi Gaussian)

$$\mathbf{W}_c = \left(\sum_{r=1}^R \underbrace{\left[\sum_{t=1}^{N_{r,c}} (o_{r,t} - \mu_o) \right]} E[\mathbf{h}_r^T] \right) \left(\sum_{r=1}^R \underbrace{N_{r,c}}_{\nwarrow} E[\mathbf{h}_r \mathbf{h}_r^T] \right)^{-1} \quad (c = 1, 2, \dots, C)$$

First-order statistics (centred)

Zero-order statistics (number of frames per Gaussian)

Interesting Research Topics

USING NON-STANDARD GAUSSIAN PRIOR

I-vector extraction with informative prior

- Conventional i-vector extraction assumes a standard Gaussian prior on the latent variable \mathbf{x}

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

- Consider a more general case where the prior on \mathbf{x} has mean $\boldsymbol{\mu}_p$ and covariance $\boldsymbol{\Sigma}_p$

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$$

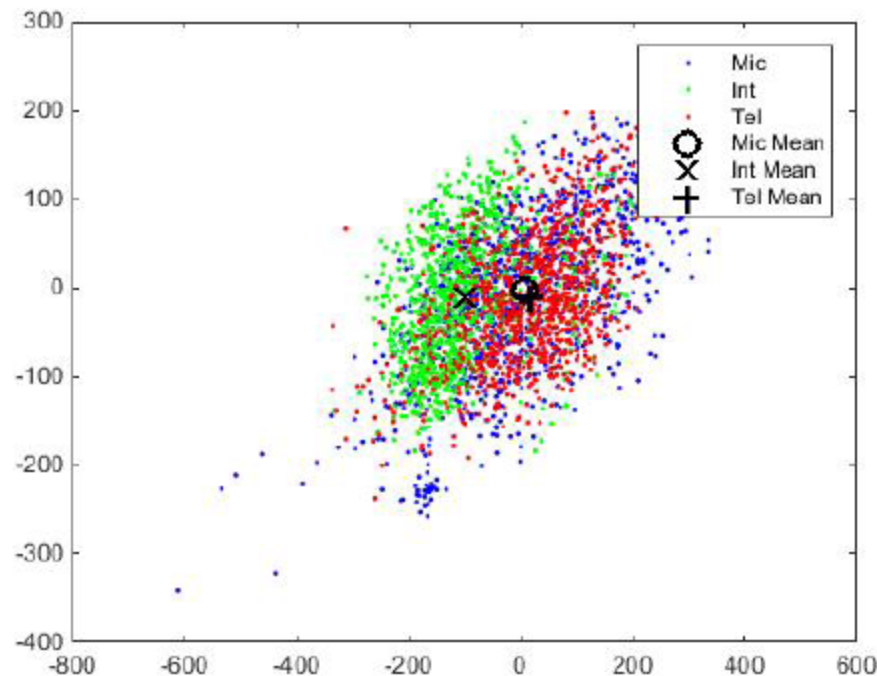
- I-vector extraction with non-standard Gaussian prior

$$\phi = \mathbf{L}^{-1} \cdot \left(\mathbf{T}^T \tilde{\mathbf{F}} + \boxed{\boldsymbol{\Sigma}_p^{-1} \mu_p} \right)$$

$$\mathbf{L}^{-1} = \left(\boxed{\boldsymbol{\Sigma}_p^{-1}} + \mathbf{T}^T \mathbf{N} \mathbf{T} \right)^{-1}$$

Heterogeneous dataset

- For heterogeneous dataset, we could consider using **multiple priors**, one for each homogeneous subset, e.g., NIST SRE'08 and SRE'10 consists of
 - Telephone
 - Microphone
 - Interview



Source compensation

- Direct implementation causes the differences due to the source being emphasized
- Solution: bring together the sources to the same origin
 - The following marginalization leads to the same result

$$\begin{aligned}\Pi(c) &= \int \mathcal{N}(O|\mathbf{m}_0(c) + \mathbf{T}_c \mathbf{w}, \Sigma_0) \mathcal{N}(\mathbf{w}|\mu_{\mathbf{p}}, \Sigma_{\mathbf{p}}) d\mathbf{w} \\ &= \int \mathcal{N}(O|\mathbf{m}_0(c) + \mathbf{T}_c \mu_{\mathbf{p}} + \mathbf{T}_c \mathbf{w}, \Sigma_0) \mathcal{N}(\mathbf{w}|0, \Sigma_{\mathbf{p}}) d\mathbf{w}\end{aligned}$$

- Lifting of prior mean to the acoustic space (high dimensional) by applying the transform for each Gaussian
- Priors now differ only with regard with one another covariance
- Sources have common mode at the origin

Estimation of informative prior

- Estimated from i-vectors extracted from initial \mathbf{T} matrix.
- For each source, seek the distribution to minimize the KL divergence from a given set of posterior distributions.
- Minimum divergence criterion:

$$D(\theta_{\text{MD}}) = \sum_{i=1}^I E \left\{ \log \frac{\mathcal{N}(\mathbf{w} | \phi_i, \mathbf{L}_i^{-1})}{\mathcal{N}(\mathbf{w} | \boldsymbol{\mu}_{\mathbf{P}}, \boldsymbol{\Sigma}_{\mathbf{P}})} \right\}$$

- Closed form solution:

$$\boldsymbol{\mu}_{\mathbf{P}} = \frac{1}{I} \sum_{i=1}^I \phi_i$$

$$\boldsymbol{\Sigma}_{\mathbf{P}} = \frac{1}{I} \sum_{i=1}^I (\phi_i - \boldsymbol{\mu}_{\mathbf{P}})(\phi_i - \boldsymbol{\mu}_{\mathbf{P}})^T + \frac{1}{I} \sum_{i=1}^I \mathbf{L}_i^{-1}$$

Re-centring of first-order statistics

- Algorithmically, the source compensation is applied by re-centring the first order statistics:

$$\begin{aligned}\tilde{\tilde{\mathbf{F}}}(c) &= \sum_t \gamma_t(c) (o_t - \mathbf{m}_o(c) - \mathbf{T}_c \mu_p) \\ &= \tilde{\mathbf{F}}(c) - N(c) \mathbf{T}_c \mu_p\end{aligned}$$

- Extract i-vector using zero-mean informative prior for each source

$$\begin{aligned}\phi &= \mathbf{L}^{-1} \left(\mathbf{T}^T \mathbf{\Sigma}^{-1} (\tilde{\mathbf{F}} - \mathbf{N} \mathbf{T}_c \mu_p) \right) \\ &= \mathbf{L}^{-1} \left(\mathbf{T}^T \mathbf{\Sigma}^{-1} \tilde{\tilde{\mathbf{F}}} \right) \\ \mathbf{L} &= \mathbf{T}^T \mathbf{N} \mathbf{\Sigma}^{-1} \mathbf{T} + \mathbf{\Sigma}_p^{-1}\end{aligned}$$

Experiment

| | CC1: int-int | | CC2: int-int | | CC3: int-int | | CC4: int-tel | | CC5: tel-mic | | CC6: tel-tel | | CC7: tel-tel | | CC8: tel-tel | |
|----------------|--------------|-------------|--------------|-------------|--------------|-------------|--------------|------|--------------|------|--------------|-------------|--------------|-------------|--------------|-------------|
| EER | F | M | F | M | F | M | F | M | F | M | F | M | F | M | F | M |
| Telephone only | 3.51 | 2.84 | 1.50 | 0.32 | 3.61 | 2.97 | 5.69 | 4.07 | 6.65 | 4.17 | 5.85 | 4.67 | 2.73 | 2.32 | 3.24 | 1.43 |
| Pooled | 3.22 | 2.54 | 1.28 | 0.33 | 3.29 | 2.64 | 4.65 | 3.89 | 5.62 | 3.05 | 5.86 | 4.15 | 2.84 | 1.60 | 3.32 | 1.04 |
| Cascade | 3.17 | 3.01 | 1.25 | 0.41 | 3.26 | 3.22 | 5.38 | 4.27 | 6.10 | 4.12 | 5.86 | 4.06 | 2.98 | 1.66 | 3.81 | 1.32 |
| 2-prior | 2.34 | 1.95 | 1.32 | 0.32 | 2.39 | 2.04 | 4.32 | 3.91 | 5.37 | 3.21 | 5.79 | 3.84 | 2.87 | 1.39 | 3.27 | 0.90 |

Table 1. SRE'08 Performance comparison for the sub-task short2-short3. Left: FEMALE Trials, Right: MALE Trials

| | CC1: int-int-same-mic | | CC2: int-int-diff-mic | | CC3: int-tel | | CC4: int-mic | | CC5: nve-nve-diff-tel | | CC6: nve-hve-diff-tel | | CC7: nve-hve-mic | | CC8: nve-lve-diff-tel | | CC9: nve-lve-mic | |
|----------------|-----------------------|-------------|-----------------------|-------------|--------------|------|--------------|-------------|-----------------------|------|-----------------------|------|------------------|-------------|-----------------------|------|------------------|-------------|
| EER | F | M | F | M | F | M | F | M | F | M | F | M | F | M | F | M | F | M |
| Telephone only | 3.06 | 2.02 | 5.65 | 3.45 | 4.21 | 3.55 | 3.96 | 2.72 | 3.59 | 3.47 | 8.09 | 4.64 | 8.49 | 4.91 | 2.01 | 1.14 | 2.46 | 1.54 |
| Pooled | 3.16 | 2.22 | 5.13 | 3.14 | 3.34 | 2.82 | 3.78 | 2.54 | 3.00 | 2.60 | 7.13 | 4.01 | 7.98 | 4.95 | 1.66 | 1.54 | 2.55 | 1.34 |
| Cascade | 3.12 | 2.29 | 5.60 | 3.29 | 4.01 | 2.62 | 4.04 | 2.87 | 3.41 | 3.13 | 7.10 | 4.33 | 8.19 | 5.25 | 1.83 | 1.68 | 3.08 | 1.61 |
| 2-prior | 2.43 | 1.67 | 4.44 | 2.25 | 3.87 | 3.19 | 3.33 | 2.22 | 3.00 | 2.89 | 7.11 | 4.13 | 7.49 | 4.16 | 1.59 | 1.56 | 2.48 | 1.15 |

Table 2. SRE'10 Performance comparison for the sub-task core-core. Left: FEMALE Trials, Right: MALE Trials

- S. E. Shepstone, K. A. Lee, H. Li, Z.-H. Tan, S. H. Jensen, “Source-specific informative prior for i-vector extraction,” in *Proc. ICASSP*, pp. 4185 – 4189, Apr. 2015.
- S. E. Shepstone, K. A. Lee, H. Li, Z.-H. Tan, S. H. Jensen, “Total variability modeling using source-specific priors,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 3, pp. 504 - 517, Mar 2016.

Interesting Research Topics

LOCAL VARIABILITY MODEL

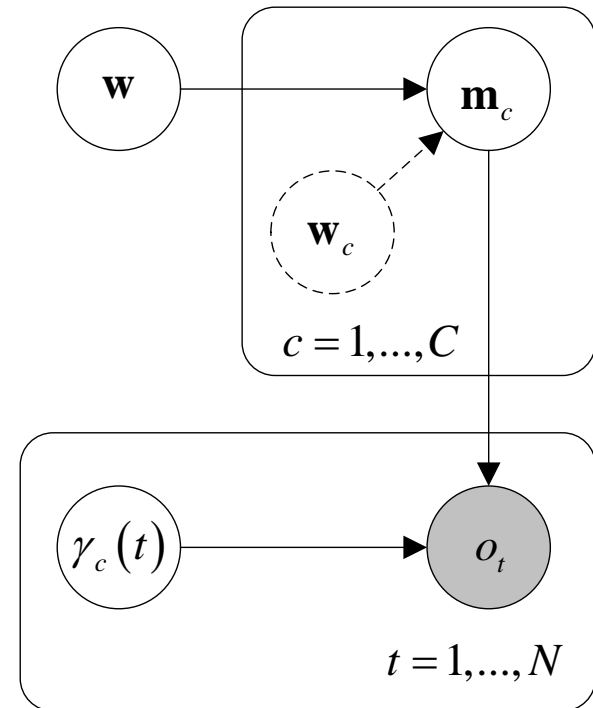
Total vs. local variability models

- **Total** variability model

- latent variable \mathbf{w} is tied/shared across
- the observed frames (speaker and recording channel remain the same throughout an utterance)
- the Gaussians

- **Local** variability model

- latent variable is moved inside the rectangular box
- relaxes the tying across Gaussians, where one latent variable \mathbf{w}_c is assigned to each Gaussian



Local variability vector

- Total variability model

$$l_{\text{TVM}}(\theta) = \prod_{r=1}^R \int \left(\prod_{c=1}^C \prod_{t=1}^{N_{r,c}} \mathcal{N}(o_t | \boldsymbol{\mu}_c + \mathbf{T}_c \mathbf{w}_r, \boldsymbol{\Sigma}_c) \right) \mathcal{N}(\mathbf{w}_r | \mathbf{0}, \mathbf{I}) d\mathbf{w}_r$$

- Local variability model

$$l_{\text{LVM}}(\theta) = \prod_{r=1}^R \prod_{c=1}^C \int \left(\prod_{t=1}^{N_{r,c}} \mathcal{N}(o_t | \boldsymbol{\mu}_c + \mathbf{T}_c \mathbf{w}_{r,c}, \boldsymbol{\Sigma}_c) \right) \mathcal{N}(\mathbf{w}_{r,c} | \mathbf{0}, \mathbf{I}) d\mathbf{w}_{r,c}$$

- The posterior means of the latent variables form the local variability vector.
- The local variability vector captures the information dedicated to individual Gaussian.

Results

Table I: Performance comparison of total variability model (TVM), local variability model (LVM) and score fusion on DET6 of *short2-short3* task in NIST SRE'08.

| Male | | | |
|--------|---------|----------|----------|
| | EER (%) | minDCF08 | minDCF10 |
| TVM | 3.6182 | 0.2130 | 0.6820 |
| LVM | 4.7559 | 0.2596 | 0.7895 |
| fusion | 3.3700 | 0.1943 | 0.6042 |
| Female | | | |
| | EER (%) | minDCF08 | minDCF10 |
| TVM | 5.3908 | 0.2767 | 0.9972 |
| LVM | 6.6144 | 0.3367 | 0.9950 |
| fusion | 5.4505 | 0.2707 | 0.9961 |

Table II: Performance comparison of total variability model (TVM), local variability model (LVM) and score fusion on CC5 of *core-core* tests in NIST SRE'10.

| Male | | | |
|--------|---------|----------|----------|
| | EER (%) | minDCF08 | minDCF10 |
| TVM | 3.0836 | 0.1253 | 0.3654 |
| LVM | 3.7590 | 0.1453 | 0.5439 |
| fusion | 2.5136 | 0.1212 | 0.3626 |
| Female | | | |
| | EER (%) | minDCF08 | minDCF10 |
| TVM | 2.6743 | 0.1458 | 0.3239 |
| LVM | 4.3068 | 0.2317 | 0.6119 |
| fusion | 2.5399 | 0.1488 | 0.3521 |

L. Chen, K. A. Lee, B. Ma, W. Guo, H. Li, and L. R. Dai, "Local variability modelling for text-independent speaker verification," in *Proc. Odyssey*, 2014.

L. Chen, K. A. Lee, B. Ma, W. Guo, H. Li, and L. R. Dai, "Local variability vector for text-independent speaker verification," in *Proc. ISCSLP*, 2014.

Kong Aik LEE

*Scientist, Institute for Infocomm Research, A*STAR, Singapore*

THANK YOU

Learning Diary

- A. An i-vector is a compressed representation of variable-duration utterances. It is widely described with the following expression

$$\mathbf{m}_r = \mathbf{m}_o + \mathbf{T}\mathbf{h}_r$$

Explain the role of variable \mathbf{h}_r in this model.

- B. The generative equation of a single Gaussian total variability model could be expressed as follows.

$$\begin{bmatrix} o_1 \\ o_2 \\ \vdots \\ o_T \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_o \\ \boldsymbol{\mu}_o \\ \vdots \\ \boldsymbol{\mu}_o \end{bmatrix} + \begin{bmatrix} \mathbf{W} \\ \mathbf{W} \\ \vdots \\ \mathbf{W} \end{bmatrix} \mathbf{h}_r + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_T \end{bmatrix}$$

Explain the rationale of latent variable tying across frames in the model.