Log Spectra Enhancement for Speaker Verification

Mengke HU

ECE Department
Drexel University

ASPITRG Group Meeting
Outline

1. Variational Bayesian Inference
   - Bayesian Inference
   - Variational Bayesian Inference

2. Speaker Verification
   - Base Line System
   - Robust Speech Processing

3. Log Spectra Enhancement for Speaker Verification
   - Feature Extraction and Speech Model
   - Probabilistic Model
   - VBI for feature enhancement
Outline

1. Variational Bayesian Inference
   - Bayesian Inference
   - Variational Bayesian Inference

2. Speaker Verification
   - Base Line System
   - Robust Speech Processing

3. Log Spectra Enhancement for Speaker Verification
   - Feature Extraction and Speech Model
   - Probabilistic Model
   - VBI for feature enhancement
Basic Inference Problem
Maximum Likelihood Estimator

Problem Description:
Given observation $X = \{x_1, x_2, \ldots, x_N\}$ and probabilistic model $p(X; \Theta)$, we want to estimate the unknown parameters $\Theta$.

Maximum Likelihood Estimator (MLE):

$$\Theta_{MLE} = \arg \max_{\Theta} p(X; \Theta) \iff \Theta_{MLE} = \arg \max_{\Theta} \log p(X; \Theta)$$
Basic Inference Problem
Maximum Likelihood Estimator

- Problem Description:
  Given observation \( X = \{x_1, x_2, \ldots, x_N\} \) and probabilistic model \( p(X; \Theta) \), we want to estimate the unknown parameters \( \Theta \).

- Maximum Likelihood Estimator (MLE):

\[
\Theta_{MLE} = \arg \max_{\Theta} p(X; \Theta) \iff \Theta_{MLE} = \arg \max_{\Theta} \log p(X; \Theta)
\]
Basic Inference Problem
Maximum Likelihood Estimator

Example:
Samples in \( X = \{x_1, x_2, \ldots, x_N\} \) are i.i.d., \( p(X_i; \theta) \sim \mathcal{N}(x_i; \mu, \sigma) \) and \( p(X; \theta) = \prod_{i=1}^{N} p(x_i; \Theta) \sim \prod_{i=1}^{N} \mathcal{N}(x_i; \mu, \sigma) \), then we have:

\[
\mu_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i
\]

\[
\sigma_{MLE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_{MLE})^2
\]

Drawback:
It does not take into account parameter and model uncertainty.
Bayesian Inference Problem

Graphical Model

- Treat parameters $\Theta$ as random variable with $\Theta \sim p(\Theta)$, then the model becomes:

$$p(X, \Theta) = p(X|\Theta)p(\Theta) = \text{likelihood} \times \text{prior}$$

![Graphical Model Diagram]
Bayesian Inference Problem
Bayesian Estimator

Cost function:

\[ C(\Theta, \hat{\Theta}) \]

Example: squared error:

\[ C(\Theta, \hat{\Theta}) = ||\Theta - \hat{\Theta}||^2 \]

Bayesian Estimator:

\[ \hat{\Theta} = \arg \min_{\hat{\Theta}} \mathbb{E}[C(\Theta, \hat{\Theta})|X] \]
Bayesian Inference Problem
Minimum Mean Square Error Estimator

- Minimum Mean Square Error Estimator (MMSE):

\[
\hat{\Theta}_{MMSE} = \mathbb{E}[\Theta|X] = \int \Theta p(\Theta|X = x) d\Theta \bigg|_{x=X} = \arg\min_{\hat{\Theta}} \int ||\Theta - \hat{\Theta}||^2 p(\Theta|X) d\Theta
\]

proof see Appendix 1

Problem: How to calculate the posterior \( p(\Theta|X) \)?
Bayesian Inference Problem
Minimum Mean Square Error Estimator

- Minimum Mean Square Error Estimator (MMSE):

$$\hat{\Theta}_{MMSE} = \mathbb{E}[\Theta|X] = \int \Theta p(\Theta|X = x) d\Theta \bigg|_{x=X}$$

$$= \arg\min_{\hat{\Theta}} \int ||\Theta - \hat{\Theta}||^2 p(\Theta|X) d\Theta$$

proof see Appendix 1

- Problem: How to calculate the posterior $p(\Theta|X)$?
Bayesian Inference Problem

**Calculation of Posterior**

- **By Bayesian Theorem:**

\[ p(\Theta|X) = \frac{p(X, \Theta)}{p(X)} = \frac{p(X|\Theta)p(\Theta)}{\int p(X|\Theta)p(\Theta)d\Theta} \]

- **Problem: Intractability**
  
  The posterior is difficult to calculate. For example:
  
  \[ p(X) = \int p(X|\Theta)p(\Theta)d\Theta \] is very difficult to be marginalized.
Bayesian Inference Problem
Calculation of Posterior

- By Bayesian Theorem:

\[
p(\Theta | X) = \frac{p(X, \Theta)}{p(X)} = \frac{p(X|\Theta)p(\Theta)}{\int p(X|\Theta)p(\Theta)d\Theta}
\]

- Problem: Intractability
  The posterior is difficult to calculate. For example:
  \[
p(X) = \int p(X|\Theta)p(\Theta)d\Theta
\]
  is very difficult to be marginalized.
Outline

1. Variational Bayesian Inference
   - Bayesian Inference
   - Variational Bayesian Inference

2. Speaker Verification
   - Base Line System
   - Robust Speech Processing

3. Log Spectra Enhancement for Speaker Verification
   - Feature Extraction and Speech Model
   - Probabilistic Model
   - VBI for feature enhancement
Approximate Bayesian Inference

Possible Solutions

- Solutions:
  - Using tractable approximation to replace the intractable $p(\Theta|X)$
    - Variational Bayesian Inference
    - Expectation Propagation (EP)
  - Using the samples of $p(\Theta|X)$
    - Markov Chain Monte Carlo Methods, ex. Gibbs Sampler
Goal:
Approximate \( p(\Theta|X) \) by variational distribution \( q(\Theta) \)

Variational Method:
- Concept: functional derivative
- It is to restrict the range of functions over which the optimization is performed.
- Confine the family of \( q(\Theta) \), minimize the divergence between \( q(\Theta) \) and \( p(\Theta|X) \)
Variational Bayesian Inference

- Goal:
  Approximate \( p(\Theta | X) \) by variational distribution \( q(\Theta) \)

- Variational Method:
  - Concept: functional derivative
  - It is to restrict the range of functions over which the optimization is performed.
  - Confine the family of \( q(\Theta) \), minimize the divergence between \( q(\Theta) \) and \( p(\Theta | X) \)
Variational Bayesian Inference

\[ q(\Theta) = \prod_i q(\theta_i) \]

\[ KL(q||p) \]
Variational Bayesian Inference

Object:

\[ q^* = \arg\min_q KL(q \| p) \]

Subject to:

\[ q(\Theta) \in Q, \text{ s.t. } q(\Theta) = \prod_j q(\theta_j) \]

- The constraint condition ensures tractability
Variational Bayesian Inference

Input $q = q(\Theta)$; Output $p = p(\Theta|X)$; $p(X)$ is fixed, then:

\[
\ln p(X) = \ln \left( \frac{p(X, \Theta)}{p(\Theta|X)} \right) = \int q(\Theta) \left\{ \ln \left( \frac{p(X, \Theta)}{q(\Theta)} \times \frac{q(\Theta)}{p(\Theta|X)} \right) \right\} d\Theta
\]

\[
= \int q(\Theta) \ln \left( \frac{p(X, \Theta)}{q(\Theta)} \right) d\Theta + \left\{ -\int q(\Theta) \ln \left( \frac{p(\Theta|X)}{q(\Theta)} \right) d\Theta \right\}
\]

\[
= \mathcal{L}(q) + KL(q||p)
\]

- Ideal case: $\min KL(q||p) = 0$, when $q = p$.
- $\max_q \mathcal{L}(q) \iff \min_q KL(q||p)$
- We can use $q$ that minimizes $KL$ divergence to approximate $p$
Variational Bayesian Inference

General Solution

The solution to the problem in previous slides is:

$$
\ln q^*(\theta_j) = \mathbb{E}_{q(\Theta \setminus j)}[\ln p(X, \Theta)] + \text{Const}.
$$

where the $q(\Theta \setminus j)$ is the variational distribution of all element in $\Theta$ except $\theta_j$.

The proof is in the Appendix 2.
The whole idea of VBI is to approximate the intractable $p(\Theta | X)$ by tractable distribution $q(\Theta)$.

Optimization problem: find $q(\Theta)$ to minimize the KL divergence.

Confine the family of $q(\Theta)$ s.t. $q(\Theta) = \prod_{j} q(\theta_j)$, we have the optimum solution:

$$\ln q^*(\theta_j) = \mathbb{E}_{q(\Theta \setminus j)}[\ln p(X, \Theta)] + \text{Const.}$$
The whole idea of VBI is to approximate the intractable $p(\Theta|X)$ by tractable distribution $q(\Theta)$.

Optimization problem: find $q(\Theta)$ to minimize the KL divergence.

Confine the family of $q(\Theta)$ s.t. $q(\Theta) = \prod_j q(\theta_j)$, we have the optimum solution:

$$\ln q^*(\theta_j) = \mathbb{E}_{q(\Theta \setminus j)}[\ln p(X, \Theta)] + \text{Const.}$$
The whole idea of VBI is to approximate the intractable \( p(\Theta | X) \) by tractable distribution \( q(\Theta) \).

Optimization problem: find \( q(\Theta) \) to minimize the KL divergence.

Confine the family of \( q(\Theta) \) s.t. \( q(\Theta) = \prod q(\theta_j) \), we have the optimum solution:

\[
\ln q^*(\theta_j) = \mathbb{E}_{q(\Theta \setminus j)}[\ln p(X, \Theta)] + \text{Const}.
\]
Outline

1 Variational Bayesian Inference
   - Bayesian Inference
   - Variational Bayesian Inference

2 Speaker Verification
   - Base Line System
   - Robust Speech Processing

3 Log Spectra Enhancement for Speaker Verification
   - Feature Extraction and Speech Model
   - Probabilistic Model
   - VBI for feature enhancement
Feature Extraction

- **Purpose:** In order to identify speakers, we need to extract information in speech signal.

- **FFT:**
  - Feature dimension is too high to extract speech information
  - It does not compress the relevant information in each speech frame

- **MFCC:**
  - It is not sensitive to noise.
  - It takes into account the non-linear processing of sound in the ear (characterize the timber).

- **Log Spectra:**
  - Separate clean speech from noise and channel (Production to Addition)
  - Compare to MFCC, it is easier to clean speech
Speaker Verification Model

Base line system
Speaker Verification Model

Base line system

Given a speech segment $X$, we test 2 hypotheses:

- $H_0$: $X$ is from claimed target speaker $S$ (GMM)
- $H_1$: $X$ is not from speaker $S$, it is from the background (UBM)

Decision Rule

- \[ \text{Score} = \log \frac{p(X|\text{TargetModel})}{p(X|\text{UBM})} \]

- $H_0 > \text{Threshold} \iff H_1$

Note: \[ \text{Score} = \log p(X|\text{TargetModel}) - \log p(X|\text{UBM}) \]
Speaker Verification Model

Base line system

1 Given a speech segment $X$, we test 2 hypotheses:
   - $H_0$: $X$ is from claimed target speaker $S$ (GMM)
   - $H_1$: $X$ is not from speaker $S$, it is from the background (UBM)

2 Decision Rule

   - $Score = \log \frac{p(X|\text{TargetModel})}{p(X|\text{UBM})}$
   - $H_0 \quad > \quad Threshold \quad < \quad H_1$

★ Note: $Score = \log p(X|\text{TargetModel}) - \log p(X|\text{UBM})$
How to solve the following problem?

1. Input speech has additive noise.
2. Mismatch between training and operation conditions.
Speaker Verification Model
Base line system

How to solve the following problem?

1. Input speech has additive noise.
2. Mismatch between training and operation conditions.
How to solve the following problem?

1. Input speech has additive noise.
2. Mismatch between training and operation conditions.
Outline

1. Variational Bayesian Inference
   • Bayesian Inference
   • Variational Bayesian Inference

2. Speaker Verification
   • Base Line System
   • Robust Speech Processing

3. Log Spectra Enhancement for Speaker Verification
   • Feature Extraction and Speech Model
   • Probabilistic Model
   • VBI for feature enhancement
Feature Domain Robust Speech Processing

- Algonquin algorithm
- NAP for feature compensation
Feature Domain Robust Speech Processing

- Algonquin algorithm
- NAP for feature compensation
Joint Speech Enhancement and Speaker Verification

- **Intuition:**
  Cleaner speech ⇔ Better speaker verification

- **General Idea**
  Jointly obtain the clean speech and speaker identity by using the prior distribution of the speech (i.e. speaker dependent)

- **Principle Model**
  Model this idea as variational Bayesian (VB) inference problem
Joint Speech Enhancement and Speaker Verification

- **Intuition:**
  Cleaner speech $\Rightarrow$ Better speaker verification

- **General Idea**
  Jointly obtain the clean speech and speaker identity by using the prior distribution of the speech (i.e. speaker dependent)

- **Principle Model**
  Model this idea as variational Bayesian (VB) inference problem
Joint Speech Enhancement and Speaker Verification

- **Intuition:**
  Cleaner speech ⇔ Better speaker verification

- **General Idea**
  Jointly obtain the clean speech and speaker identity by using the prior distribution of the speech (i.e. speaker dependent)

- **Principle Model**
  Model this idea as variational Bayesian (VB) inference problem
Outline

1 Variational Bayesian Inference
   - Bayesian Inference
   - Variational Bayesian Inference

2 Speaker Verification
   - Base Line System
   - Robust Speech Processing

3 Log Spectra Enhancement for Speaker Verification
   - Feature Extraction and Speech Model
   - Probabilistic Model
   - VBI for feature enhancement
Log Spectrum Feature Extraction

Assume clean speech $s[t]$ is corrupted by the channel $h[t]$ and additive noise $n[t]$:

$$y[t] = h[t] \ast s[t] + n[t]$$

Take DFT for both sides (frame by frame):

$$Y[k] = H[k]S[k] + N[k]$$

Note: frame size $\geq$ length of $h[t]$
Log Spectrum Feature Extraction

- Let log spectra features:
  \[ y = \log |Y[:]|^2, s = \log |S[:]|^2, h = \log |H[:]|^2 \text{ and } n = \log |N[:]|^2, \]
  we can show (proof in Appendix 3):

  \[ y \approx s + h + \log(1 + \exp(n - h - s)) \]

- Approximately:

  \[
  y \approx s + \log(1 + \exp(n - s))
  \]

  by assumption that we can mitigate channel effects
Outline

1. Variational Bayesian Inference
   - Bayesian Inference
   - Variational Bayesian Inference

2. Speaker Verification
   - Base Line System
   - Robust Speech Processing

3. Log Spectra Enhancement for Speaker Verification
   - Feature Extraction and Speech Model
   - Probabilistic Model
   - VBI for feature enhancement
Observation Likelihood

The speech feature model:

\[ y \approx s + \log(1 + \exp(n - s)) \]

Assuming the approximation errors in formula (2) are Gaussian, that is

\[ \mathcal{E} = y - (s + \log(1 + \exp(n - s))) \sim \mathcal{N}(0, \psi) \]

Then, the likelihood of the observation \( y \) is:

\[ p(y|s, n) = \mathcal{N}(y|s + \log(1 + \exp(n - s)), \psi) \]
Speaker Dependent Prior

Let the library $\mathcal{L} = \{\text{TargetSpeaker}, \text{UBM}\}$, then given $\ell \in \mathcal{L}$, the mixture Gaussian distribution for $s$ is:

$$p(s|\ell) = \sum_{m=1}^{M_s} \pi_{\ell m}^s \mathcal{N}(s; \mu_{\ell m}^s, \Sigma_{\ell m}^s)$$

where:

- $M_s$ is the number of Gaussian mixture coefficients
- $\pi_{\ell m}^s$ is $m^{th}$ mixture coefficient for speech $s$ in the library $\ell$

$$\sum_{m=1}^{M_s} \pi_{\ell m}^s = 1$$
Speaker Dependent Prior

Given $p_{\ell}(\text{Target}) = p$, by Total Probability Theory:

\[
p(s) = \sum_{\ell} p_{\ell} \times p(s|\ell)
\]

\[
= p \times \sum_{m=1}^{M_s} \pi^s_{m_{\text{target}}} \mathcal{N}(s; \mu^s_{m_{\text{target}}}, \Sigma^s_{m_{\text{target}}})
\]

\[
+ (1 - p) \times \sum_{m=1}^{M_s} \pi^s_{m_{\text{UBM}}} \mathcal{N}(s; \mu^s_{m_{\text{UBM}}}, \Sigma^s_{m_{\text{UBM}}})
\]

• Note: $|\mathcal{L}| = 2$
We obtain the Gaussian Mixture distribution for clean speech:

\[ p(s) = \sum_{i=1}^{M_s|\mathcal{L}|} \pi_i^s \mathcal{N}(s; \mu_i^s, \Sigma_i^s) \]

where \( \pi^s = \left( \begin{array}{c} \pi_1^s \\ \vdots \\ \pi_{M_s|\mathcal{L}|}^s \end{array} \right) = \left( \begin{array}{c} p\pi_{\text{Target}} \\ (1-p)\pi_{\text{UBM}} \end{array} \right) \)
Speaker Dependent Prior
Indicator Variable

- Let $z_s$ be an indicator of dimension $M_s|\mathcal{L}| \times 1$

- Example: If for target speech model, $i^{th}$ mixture coefficient is active, then

$$z_s^T = (0, \cdots, 0, 1, 0, \cdots, 0, 0 \cdots 0)$$

- Relationship between indicator $z_s$ and mixture coefficients $\pi^S$:
  - $p(z_{s,i} = 1) = \prod_{i=1}^{M_s|\mathcal{L}|} \pi_{i}^{z_{s,i}}$
  - $p(z_s) = \prod_{i=1}^{M_s|\mathcal{L}|} \pi_{i}^{z_{s,i}}$
Speaker Dependent Prior

Indicator Variable

- Let \( z_s \) be an indicator of dimension \( M_s|\mathcal{L}| \times 1 \)

- Example: If for target speech model, \( i^{th} \) mixture coefficient is active, then

\[
\begin{align*}
  z_s^T &= (0, \cdots, 0, 1, 0, \cdots, 0, 0 \cdots 0) \\
  &\quad \text{UBM} \\
  &\quad \text{TargetSpeakerModel}
\end{align*}
\]

- Relationship between indicator \( z_s \) and mixture coefficients \( \pi^s \):

  - \( p(z_{s,i} = 1) = \pi^s_i \)
  - \( p(z_s) = \prod_{i=1}^{M_s|\mathcal{L}|} \pi^s_{z_s,i} \)
Let $\mathbf{z}_s$ be an indicator of dimension $M_s|\mathcal{L}| \times 1$

Example: If for target speech model, $i^{th}$ mixture coefficient is active, then

$$\mathbf{z}_s^T = (0, \cdots, 0, 1, 0, \cdots, 0, 0 \cdots 0)$$

Relationship between indicator $\mathbf{z}_s$ and mixture coefficients $\pi^s$:

- $p(z_{s,i} = 1) = \pi^s_i$
- $p(\mathbf{z}_s) = \prod_{i=1}^{M_s|\mathcal{L}|} \pi^s_{i^{th}}$
We can write: $p(s|z_{s,i} = 1) = \mathcal{N}(s; \mu_i^s, \Sigma_i^s)$

Then we can obtain

$$p(s|z_s) = \prod_{i=1}^{M_s|\mathcal{L}|} \mathcal{N}(s; \mu_i^s, \Sigma_i^s)^{z_{s,i}} \quad (1)$$
We can write: $p(s|z_{s,i} = 1) = \mathcal{N}(s; \mu_i^s, \Sigma_i^s)$

Then we can obtain

$$p(s|z_s) = \prod_{i=1}^{M_s \cdot |\mathcal{L}|} \mathcal{N}(s; \mu_i^s, \Sigma_i^s)^{z_{s,i}}$$

(1)
Probability Model

Markov Chain between variables

\[ p(y, s, z, n) = p(y|s, n) \times p(s|z) \times p(z) \times p(n) \]
Probability Model
Markov Chain between variables

The joint distribution:

\[ p(y, s, z_s, n) = p(y|s, n, z_s) \times p(s, n|z_s) \times p(z_s) \]

\[ = p(y|s, n, z_s) \times p(s|z_s) \times p(n|z_s) \times p(z_s) \quad (a) \]

\[ = p(y|s, n) \times p(s|z_s) \times p(z_s) \times p(n) \quad (b) \]

(a) is because that given \( z_s \), \( s \) and \( n \) are conditionally independent
(b) is because of Markov property
Outline

1. Variational Bayesian Inference
   - Bayesian Inference
   - Variational Bayesian Inference

2. Speaker Verification
   - Base Line System
   - Robust Speech Processing

3. Log Spectra Enhancement for Speaker Verification
   - Feature Extraction and Speech Model
   - Probabilistic Model
   - VBI for feature enhancement
Problem Reiterate

- Speech model in log spectrum features:
  \[ y \approx s + \log(1 + \exp(n - s)) \]
- Probability Model:
  \[ p(y, s, z_s, n) = p(y|s, n) \times p(s|z_s) \times p(z_s) \times p(n) \]
  - \[ p(y|s, n) = \mathcal{N}(y|s + \log(1 + \exp(n - s)), \psi) \]
  - \[ p(s|z_s) = \prod_{i=1}^{\mathcal{M}_s|\mathcal{L}|} \mathcal{N}(s; \mu_i^s, \Sigma_i^s)^{z_s,i} \]
  - \[ p(z_s) = \prod_{i=1}^{\mathcal{M}_s|\mathcal{L}|} \pi_i^{z_s,i} = \prod_{i=1}^{\mathcal{M}_s|\mathcal{L}|} \gamma_i^{z_s,i} \]
  - \[ p(n) = \mathcal{N}(n; \mu_n, \Sigma_n) \text{ by assumption} \]
Purpose: We want to obtain enhanced features $\hat{s}$ for clean speech.

We need to estimate $\Theta = \{s, z_s, n\}$ by

$$\hat{\Theta}_{\text{MMSE}} = \mathbb{E}[\Theta|y] = \int \Theta p(\Theta|X = x) d\Theta \bigg|_{x=X}$$

We need to replace $p(\Theta|y)$ by $q(\Theta)$ as approximate posterior by VB method.

Calculate $q^*(s)$, $q^*(z_s)$ and $q^*(n)$, then

$$\hat{\Theta}_{\text{MMSE}} = \{\mu_s^*, \Sigma_s^*, \mu_n^*, \Sigma_n^*, \gamma_i^*\} \text{ for } i \in \{1, \ldots, M_s|L|\}$$
Problem Reiterate

- **Purpose:** We want to obtain enhanced features $\hat{s}$ for clean speech.
- We need to estimate $\Theta = \{s, z_s, n\}$ by

$$
\hat{\Theta}_{\text{MMSE}} = \mathbb{E}[\Theta | y] = \int \Theta p(\Theta | X = x) d\Theta \bigg|_{x=X}
$$

- We need to replace $p(\Theta | y)$ by $q(\Theta)$ as approximate posterior by VB method.
- Calculate $q^*(s)$, $q^*(z_s)$ and $q^*(n)$, then

$$
\hat{\Theta}_{\text{MMSE}} = \{\mu_s^*, \Sigma_s^*, \mu_n^*, \Sigma_n^*, \gamma_i^*\} \text{ for } i \in \{1, \ldots, M_s|\mathcal{L}|\}$$
Purpose: We want to obtain enhanced features \( \hat{s} \) for clean speech.

We need to estimate \( \Theta = \{s, z_s, n\} \) by

\[
\hat{\Theta}_{\text{MMSE}} = \mathbb{E}[\Theta|y] = \int \Theta p(\Theta|X = x) d\Theta \bigg|_{x=X}
\]

We need to replace \( p(\Theta|y) \) by \( q(\Theta) \) as approximate posterior by VB method.

Calculate \( q^*(s), q^*(z_s) \) and \( q^*(n) \), then

\[
\hat{\Theta}_{\text{MMSE}} = \{\mu_s^*, \Sigma_s^*, \mu_n^*, \Sigma_n^*, \gamma_i^*\} \text{ for } i \in \{1, \ldots, M_s|\mathcal{L}|\} \]
Purpose: We want to obtain enhanced features $\hat{s}$ for clean speech.

We need to estimate $\Theta = \{s, z_s, n\}$ by

$$\hat{\Theta}_{MMSE} = \mathbb{E}[\Theta | y] = \int \Theta p(\Theta | X = x) d\Theta \bigg|_{x=X}$$

We need to replace $p(\Theta | y)$ by $q(\Theta)$ as approximate posterior by VB method.

Calculate $q^*(s)$, $q^*(z_s)$ and $q^*(n)$, then

$$\hat{\Theta}_{MMSE} = \{\mu_s^*, \Sigma_s^*, \mu_n^*, \Sigma_n^*, \gamma_i^*\} \text{ for } i \in \{1, \ldots, M_s | \mathcal{L}|\}$$
Approximate Posterior
Review general VB solution

Review General VB solution in previous slides:

\[ \ln q^*(\theta_j) = \mathbb{E}_{q(\Theta \setminus j)}[\ln p(X, \Theta)] + \text{Const.} \]

with \( q(\Theta) \) is the element in the tractable family, s.t.

\[ q(\Theta) = \prod_j q(\theta_j) \]
Approximate Posterior

Apply to our problem:

- Apply to our problem:
  
  \[ \Theta = \{ s, z_s, n \} \text{ and } q(\Theta) = q(s)q(z_s)q(n) \]

  \[
  q^*(n) = \mathbb{E}\{ \log p(y, s, z_s, n) \}_{q(z_s)q(s)} + C_1
  = \mathbb{E}\{ \log p(y|s, n) \}_{q(s)} + \mathbb{E}\{ \log p(z_s) \}_{q(z_s)} + C_1
  
  q^*(s) = \mathbb{E}\{ \log p(y, s, z_s, n) \}_{q(z_s)q(n)} + C_2
  = \mathbb{E}\{ \log p(y|s, n) \}_{q(n)} + \mathbb{E}\{ \log p(z_s) \}_{q(z_s)} + \mathbb{E}\{ \log p(s|z_s) \}_{q(z_s)} + C_2
  
  q^*(z_s) = \mathbb{E}\{ \log p(y, s, z_s, n) \}_{q(s)q(n)} + C_2
  = \mathbb{E}\{ \log p(y|s, n) \}_{q(n)q(s)} + \mathbb{E}\{ \log p(s|z_s) \}_{q(s)} + \mathbb{E}\{ \log p(n) \}_{q(n)} + C_3
  
  \]
Likelihood Linearization

- Observation Likelihood:

\[ p(y|s, n) = \mathcal{N}(y|s + \log(1 + \exp(n - s)), \psi) \]

with non linear mean value \( s + \log(1 + \exp(n - s)) \)

- New Problem Arises: How to calculate \( \mathbb{E}\{\log p(y|s, n)\}_q(\Theta|\gamma) \)?
Likelihood Linearization

- Observation Likelihood:

\[ p(y|s, n) = \mathcal{N}(y|s + \log(1 + \exp(n - s)), \psi) \]

with non linear mean value \( s + \log(1 + \exp(n - s)) \)

- New Problem Arises: How to calculate \( \mathbb{E}\{\log p(y|s, n)\}_q(\Theta_j) \)?
Likelihood Linearization

- Linearized likelihood:
  \[
  \hat{p}(y|s, n) = \mathcal{N}\left(y \mid s + g([s_0, n_0]) + G \times ([s, n] - [s_0, n_0]) \right), \psi
  \]

- Linearization is by the first order Taylor series expansion around the point \([s_0, n_0]\)

\[
g([s, n]) = \log(1 + \exp(n - s)) \approx g([s_0, n_0]) + G \times ([s, n] - [s_0, n_0])
\]

\[
G = [G_s, G_n] \overset{def}{=} \nabla g([s_0, n_0]), \text{ and}
\]

\[
G_s = \text{diag}\left[\frac{-\exp(n_0^1 - s_0^1)}{1 + \exp(n_0^1 - s_0^1)}, \ldots, \frac{-\exp(n_0^N - s_0^N)}{1 + \exp(n_0^N - s_0^N)}\right]
\]

\[
G_n = \text{diag}\left[\frac{\exp(n_0^1 - s_0^1)}{1 + \exp(n_0^1 - s_0^1)}, \ldots, \frac{\exp(n_0^N - s_0^N)}{1 + \exp(n_0^N - s_0^N)}\right]
\]

, where \(N\) is the dimension of feature vector. (See Appendix 4)
Variational Bayesian Algorithm

for $k = 1, \cdots, K$ frame do

Initialize the posterior distribution parameters \( \{\mu_s^*, \Sigma_s^*, \mu_n^*, \Sigma_n^*, \gamma_i\} \)

for $n = 1$ to Number of Iterations do

Set \([s_0, n_0] = [\mu_s^*, \mu_n^*] \);

E-STEP: Compute \( G = [G_s, G_n] \) and \( g([s_0, n_0]) \);

M-STEP: Update \( \{\mu_s^*, \Sigma_s^*, \mu_n^*, \Sigma_n^*\} \);

Update \( \gamma_i \)

end

end

Return \( \{\mu_s^*, \Sigma_s^*, \mu_n^*, \Sigma_n^*, \gamma_i\} \) for enhanced features after last iteration

Expressions of \( \{\mu_s^*, \Sigma_s^*, \mu_n^*, \Sigma_n^*, \gamma_i\} \) are in Appendix 5
Enhancement Summary

\[
\hat{n}_{MSEE} = \mathbb{E}[n|y]
\]

\[
\hat{s}_{MSEE} = \mathbb{E}[s|y]
\]

\[
\hat{z}_{sMSEE} = \mathbb{E}[z_s|y]
\]

VB Algorithm

\[
p(n|y) \rightarrow \hat{n}_{MSEE}
\]

\[
p(s|y) \rightarrow \hat{s}_{MSEE}
\]

\[
p(z_s|y) \rightarrow \hat{z}_{sMSEE}
\]

VB Algorithm

\[
q^*(y)
\]

\[
q^*(s)
\]

\[
q^*(z_s)
\]
Speaker Verification using Enhanced Features

\[ \text{Score} = \log p(X_{\text{enhanced}} \mid \text{TargetModel}) - \log p(X_{\text{enhanced}} \mid \text{UBM}) \]

- Train the library model \( \mathcal{L} = \{ \text{TargetSpeaker}, \text{UBM} \} \):
  - Target speaker is known but all other speakers are unknown
  - \( \mathcal{L} \) varies depending on which Target Speaker is for each verification test.
  - Use adapted GMM and adapted UBM to train the library model due to inadequacy of the data.
Summary

- This work is based on the intuition that clean speech improves the performance of speaker verification.
- It introduces speaker-dependent priors for feature enhancement based on the Algonquin Algorithm.
- It derives Variational Bayesian Algorithm to obtain the approximate posterior for clean speech.
Appendix 1

\[ \hat{\Theta}_{MMSE} = \mathbb{E}[\Theta|X] \]

- Proof:

\[ MSE = \int (\Theta - \hat{\Theta}(X))^2 p(\Theta|X)d\Theta \]

Take partial derivative to find minimum MSE:

\[ \frac{\partial MSE}{\partial \Theta} = \int 2(\Theta - \hat{\Theta}(X))p(\Theta|X)d\Theta = 0 \]

Then we have:

\[ \hat{\Theta}_{MMSE} = \int \Theta p(\Theta|X) = \mathbb{E}[\Theta|X] \]
Appendix 2

The optimal solution is:

$$\ln q^*(\theta_j) = \mathbb{E}_{q(\Theta \setminus j)}[\ln p(X, \Theta)] + \text{Const.}$$

with $q(\Theta) = \prod_j q(\theta_j)$
Appendix 2

Proof:

\[ \mathcal{L}(q) = \int q(\Theta) \ln \left( \frac{p(X, \Theta)}{q(\Theta)} \right) d\Theta \]

\[ = \int \prod_i q_i \left\{ \ln p(X, \Theta) - \sum_i \ln q_i \right\} d\Theta \]

\[ = \int q_j \left\{ \ln p(X, \Theta) \prod_{i \neq j} q_i - \left( \sum_i \ln q_i \right) \prod_{i \neq j} q_i \right\} d\Theta \]

\[ = \int q_j \left\{ \ln p(X, \Theta) \prod_{i \neq j} q_id\Theta_i \right\} d\Theta_j - \int q_j \ln q_j d\Theta_j + \text{Const} \]

\[ = \int q_i \mathbb{E}[\ln p(X, \Theta)]_{q(\Theta \setminus j)} d\Theta_j - \int q_j \ln q_j d\Theta_j + \text{Const} \]

\[ = -KL \left( q_i \left| \| \mathbb{E}[\ln p(X, \Theta)]_{q(\Theta \setminus j)} \right) \right) + \text{Const} \]
Appendix 2

- go on proof:
  Therefore
  \[
  q_i^* = \mathbb{E}[\ln p(X, \Theta)]_{q(\Theta \setminus j)} + \text{Const}
  \]
  will minimize the KL divergence
Appendix 3 I

\[ y \approx s + h + \log(1 + \exp(n - h - s)) \]

Proof:

Given \( Y[k] = H[k]S[k] + N[k] \), we have:

\[ |Y[k]|^2 = Y[k] \times Y[k]^* = (H[k]S[k] + N[k]) \times (H[k]S[k] + N[k])^* \]

\[ = |H[k]|^2|S[k]|^2 + |N[k]|^2 + 2\text{Re}\{(H[k]S[k]) \times N[k]^*}\]  

\[ \approx |H[k]|^2|S[k]|^2 + |N[k]|^2 \]

Let \( y = \log |Y[:]|^2 \), then \( |Y[:]|^2 = \exp(y) \) and similarly for \( s, h \) and \( n \).

We can rewrite (3) as:
\[ \exp(y) = \exp(s + h) + \exp(n) \]
\[ = \exp(s + h) \circ (1 + \exp(n - s - h)) \]

Taking log for both sides, we have:

\[ y \approx s + h + \log(1 + \exp(n - h - s)) \]
Appendix 4: Compute $G_I$

Given $G = [G_s, G_n] \overset{def}{=} \nabla g([s_0, n_0])$, we have

$$G = \nabla g([s_0, n_0]) = \nabla g([s, n]) \bigg|_{[s, n]=[s_0, n_0]}$$

$$= \nabla (\log(1 + \exp(n - s))) \bigg|_{[s, n]=[s_0, n_0]}$$

For $i^{th}$ element $i \in \{1, \cdots, N\}$

$$G_s(i) = \frac{d}{ds^i} \log(1 + \exp(n^i - s^i)) \bigg|_{s^i=s_0^i; n^i=n_0^i}$$

$$= \frac{-\exp(n_0^i - s_0^i)}{1 + \exp(n_0^i - s_0^i)}$$

similarly,
Appendix 4: Compute $G_{II}$

$$G_n(i) = \frac{\exp(n_0^i - s_0^i)}{1 + \exp(n_0^i - s_0^i)}$$

Therefore:

$$G_s = diag\left[\frac{-\exp(n_0^1 - s_0^1)}{1 + \exp(n_0^1 - s_0^1)}, \ldots, \frac{-\exp(n_0^N - s_0^N)}{1 + \exp(n_0^N - s_0^N)}\right]$$

$$G_n = diag\left[\frac{\exp(n_0^1 - s_0^1)}{1 + \exp(n_0^1 - s_0^1)}, \ldots, \frac{\exp(n_0^N - s_0^N)}{1 + \exp(n_0^N - s_0^N)}\right]$$
$q^*(s) = \mathbb{E}\{\log p(y, s, z_s, n)\}^q(z_s)q(n) + C_1 = \mathcal{N}(s; \mu^*_s, \Sigma^*_s)$

with

$$\Sigma^*_s = [\psi^{-1} + G_s^T \psi^{-1} G_s + \psi^{-1} G_s + G_s^T \psi^{-1} + \sum_{i=1}^{M_s|\mathcal{L}|} \gamma_i (\Sigma_i^s)^{-1}]^{-1}$$

$$\mu^*_s = \Sigma^*[((I + G_s^T) \psi^{-1}(y - g([s_0, n_0]) - G_n \mu^*_n + G_s s_0 + G_n n_0)$$

$$+ \sum_{i=1}^{M_s|\mathcal{L}|} \gamma_i (\Sigma_i^s)^{-1} \mu_i^s]$$
Appendix 5

\[ q^*(\mathbf{n}) = \mathbb{E}\{\log p(\mathbf{y}, \mathbf{s}, \mathbf{z}_s, \mathbf{n})\} q(\mathbf{s}) q(\mathbf{z}_s) + C_2 = \mathcal{N}(\mathbf{n}; \mu^*_n, \Sigma^*_n) \]
with

\[ \Sigma^*_n = [G_n^T \psi^{-1} G_n + \Sigma_n^{-1}]^{-1} \]

\[ \mu^*_n = \Sigma^*_n [G_n^T \psi^{-1} (\mathbf{y} - \mu^*_s - g([\mathbf{s}_0, \mathbf{n}_0]) - G_s \mu^*_s + G_s \mathbf{s}_0 + G_n \mathbf{n}_0) + \Sigma_n^{-1} \mu_n] \]
Appendix 5

\[ q^*(z_s) = \mathbb{E}\{\log p(y, s, z_s, n)\} q(s)q(n) + C_3 = \sum_{i=1}^{M_s|L|} (\gamma_i)^{z_{s,i}} \]

with

\[ \gamma_i = \frac{\rho_i}{M_s|L|} \prod_{i=1}^{\rho_i} \prod_{i=1}^{\rho_i} \]

\[ \log \rho_i = -\frac{1}{2}(\mu_s^* - \mu_i^*)^T (\Sigma_i^s)^{-1}(\mu_s^* - \mu_i^*) \]

\[ -\frac{1}{2} \log |\Sigma_i^s| - \frac{1}{2} \text{Tr}((\Sigma_i^s)^{-1}\Sigma_s^*) + \log \pi_i^s \]
Ciira wa Maina, John MacLaren Walsh

*Log Spectra Enhancement using Speaker Dependent Priors for Speaker Verification*


Ciira wa Maina

*Approximate Bayesian Inference for Robust Speech Processing*


Christopher M. Bishop

*Pattern Recognition and Machine Learning*


B.J. Frey, T.T. Kristyansson, L. Deng, and A. Acero

*ALGONQUIN Learning dynamic noise models from noisy speech for robust speech recognition*

In Advances in Neural Information Processing Systems 14, pages 1165-1172, January 2002