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

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

- Maximum Likelihood Estimator (MLE):

$$\Theta_{MLE} = \arg\max_{\Theta} p(\mathbf{X}; \Theta) \iff \Theta_{MLE} = \arg\max_{\Theta} \log p(\mathbf{X}; \Theta)$$

Basic Inference Problem

Maximum Likelihood Estimator

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

- Maximum Likelihood Estimator (MLE):

$$\Theta_{MLE} = \arg\max_{\Theta} p(\mathbf{X}; \Theta) \iff \Theta_{MLE} = \arg\max_{\Theta} \log p(\mathbf{X}; \Theta)$$

Basic Inference Problem

Maximum Likelihood Estimator

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

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

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

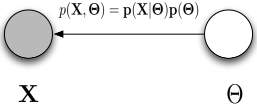
- Drawback:
  It does not take into account parameter and model uncertainty.

Bayesian Inference Problem

Graphical Model

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

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



Bayesian Inference Problem

Bayesian Estimator

Cost function:

$$\mathcal{C}(\Theta, \hat{\Theta})$$

Example: squared error:

$$\mathcal{C}(\Theta, \hat{\Theta}) = ||\Theta - \hat{\Theta}||^2$$

Bayesian Estimator:

$$\hat{\Theta} = \arg\min_{\hat{\Theta}} \mathbb{E}[\mathcal{C}(\Theta, \hat{\Theta}) | \mathbf{X}]$$

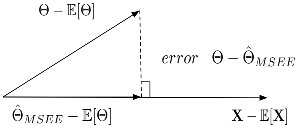
Bayesian Inference Problem

Minimum Mean Square Error Estimator

- Minimum Mean Square Error Estimator (MMSE):

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

proof see Appendix 1



- Problem: How to calculate the posterior $p(\Theta|\mathbf{X})$?

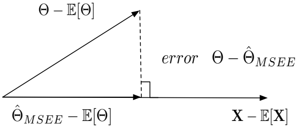
Bayesian Inference Problem

Minimum Mean Square Error Estimator

- Minimum Mean Square Error Estimator (MMSE):

$$\begin{aligned} \hat{\Theta}_{MMSE} &= \mathbb{E}[\Theta|\mathbf{X}] = \int \Theta p(\Theta|\mathbf{X}=\mathbf{x}) d\Theta \Big|_{\mathbf{x}=\mathbf{X}} \\ &= \arg\min_{\hat{\Theta}} \int ||\Theta - \hat{\Theta}||^2 p(\Theta|\mathbf{X}) d\Theta \end{aligned}$$

proof see Appendix 1



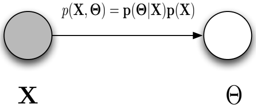
- Problem: How to calculate the posterior $p(\Theta|\mathbf{X})$?

Bayesian Inference Problem

Calculation of Posterior

- By Bayesian Theorem:

$$p(\Theta|\mathbf{X}) = \frac{p(\mathbf{X},\Theta)}{p(\mathbf{X})} = \frac{p(\mathbf{X}|\Theta)p(\Theta)}{\int p(\mathbf{X}|\Theta)p(\Theta)d\Theta}$$



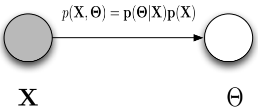
- Problem: Intractability
  The posterior is difficult to calculate. For example:
  $p(\mathbf{X}) = \int p(\mathbf{X}|\Theta)p(\Theta)d\Theta$ is very difficult to be marginalized.

Bayesian Inference Problem

Calculation of Posterior

- By Bayesian Theorem:

$$p(\Theta|\mathbf{X}) = \frac{p(\mathbf{X}, \Theta)}{p(\mathbf{X})} = \frac{p(\mathbf{X}|\Theta)p(\Theta)}{\int p(\mathbf{X}|\Theta)p(\Theta)d\Theta}$$



- Problem: Intractability
  The posterior is difficult to calculate. For example:
  $p(\mathbf{X}) = \int p(\mathbf{X}|\Theta)p(\Theta)d\Theta$ is very difficult to be marginalized.

Outline

Approximate Bayesian Inference

Possible Solutions

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

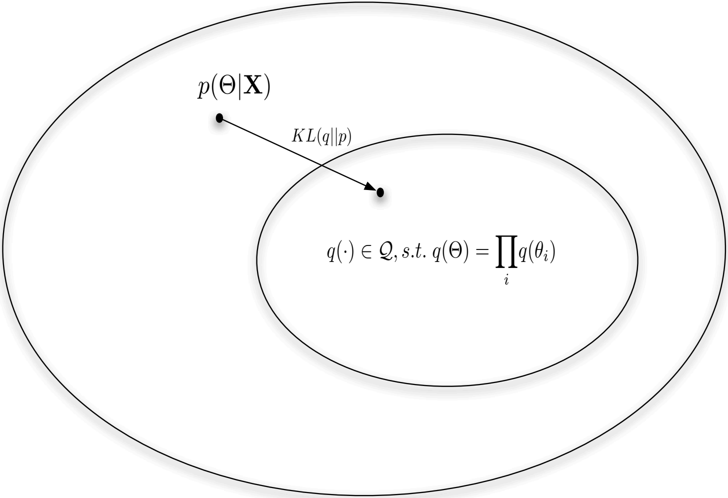
Variational Bayesian Inference

- Goal:
  Approximate $p(\Theta|\mathbf{X})$ by variational distribution $q(\Theta)$
- Variational Method:
  - Concept: functional derivative
  - It is to restrict the range of functions over the which the optimization is performed.
  - Confine the family of $q(\Theta)$, minimize the divergence between $q(\Theta)$ and $p(\Theta|\mathbf{X})$

Variational Bayesian Inference

- Goal:
  Approximate $p(\Theta|\mathbf{X})$ by variational distribution $q(\Theta)$
- Variational Method:
  - Concept: functional derivative
  - It is to restrict the range of functions over the which the optimization is performed.
  - Confine the family of $q(\Theta)$, minimize the divergence between $q(\Theta)$ and $p(\Theta|\mathbf{X})$

Variational Bayesian Inference



Variational Bayesian Inference

Object:

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

Subject to:

$$q(\Theta) \in \mathcal{Q}, \;\; s.t. \;\; q(\Theta) = \prod_j q(\theta_j)$$

- The constraint condition ensures tractability

Variational Bayesian Inference

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

$$\begin{aligned}
\ln p(\mathbf{X}) &= \ln\left(\frac{p(\mathbf{X},\Theta)}{p(\Theta|\mathbf{X})}\right) = \int q(\Theta)\left\{\ln\left(\frac{p(\mathbf{X},\Theta)}{q(\Theta)} \times \frac{q(\Theta)}{p(\Theta|\mathbf{X})}\right)\right\} d\Theta \\
&= \int q(\Theta)\ln\left(\frac{p(\mathbf{X},\Theta)}{q(\Theta)}\right) d\Theta + \left\{-\int q(\Theta)\ln\left(\frac{p(\Theta|\mathbf{X})}{q(\Theta)}\right) d\Theta\right\} \\
&= \mathcal{L}(q) + KL(q||p)
\end{aligned}$$

- 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

The solution to the problem in previous slides is:

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

,where the $q(\Theta \backslash j)$ is the variational distribution of all element in Θ except θ_j.
The proof is in the Appendix 2.

- The whole idea of VBI is to approximate the intractable $p(\Theta|\mathbf{X})$ 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:

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

- The whole idea of VBI is to approximate the intractable $p(\Theta|\mathbf{X})$ 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:

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

- The whole idea of VBI is to approximate the intractable $p(\Theta|\mathbf{X})$ 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:

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

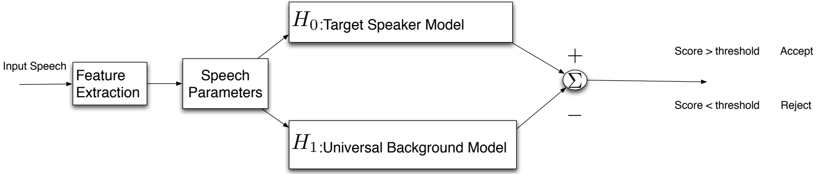
Outline

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

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
   - $Score = \log \frac{p(\mathbf{X}|TargetModel)}{p(\mathbf{X}|UBM)} \underset{H_1}{\overset{H_0}{\gtrless}} Threshold$
     - ⋆ Note: $Score = \log p(\mathbf{X}|TargetModel) - \log p(\mathbf{X}|UBM)$

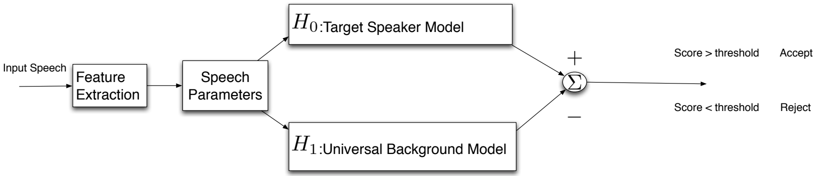
Speaker Verification Model

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
   - $Score = \log \frac{p(\mathbf{X}|TargetModel)}{p(\mathbf{X}|UBM)} \underset{H_1}{\overset{H_0}{\gtrless}} Threshold$
     - ★ Note: $Score = \log p(\mathbf{X}|TargetModel) - \log p(\mathbf{X}|UBM)$

Speaker Verification Model

Base line system

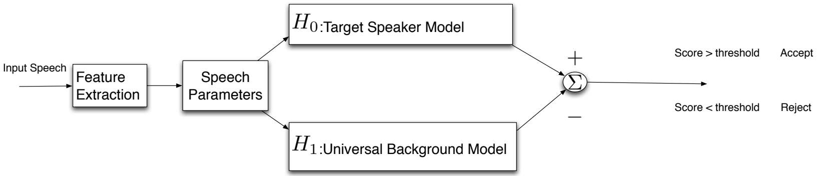


How to solve the following problem?

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

Speaker Verification Model

Base line system

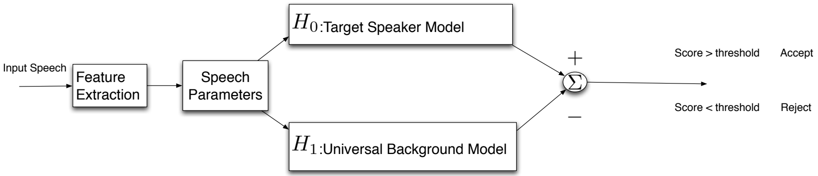


How to solve the following problem?

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

Speaker Verification Model

Base line system

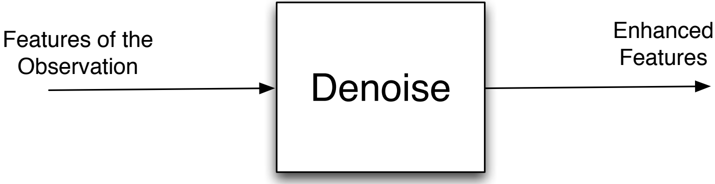


How to solve the following problem?

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

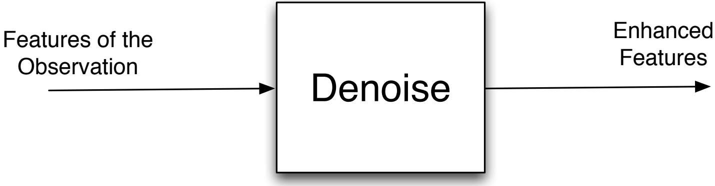
Outline

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 $\leftrightharpoons$ 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 $\leftrightharpoons$ 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 $\leftrightharpoons$ 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

Log Spectrum Feature Extraction

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

$$y[t] = h[t] * s[t] + n[t]$$

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

$$\mathbf{y} \approx \mathbf{s} + \mathbf{h} + \log(1 + \exp(\mathbf{n} - \mathbf{h} - \mathbf{s}))$$

- Approximately:

$$\boxed{\mathbf{y} \approx \mathbf{s} + \log(1 + \exp(\mathbf{n} - \mathbf{s}))}$$

by assumption that we can mitigate channel effects

Outline

Observation Likelihood

The speech feature model:

$$\mathbf{y} \approx \mathbf{s} + \log(1 + \exp(\mathbf{n} - \mathbf{s}))$$

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

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

Then, the likelihood of the observation $\mathbf{y}$ is:

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

Speaker Dependent Prior

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

$$p(\mathbf{s}|\ell) = \sum_{m=1}^{M_s} \pi^s_{\ell m} \mathcal{N}(\mathbf{s}; \mu^s_{\ell m}, \Sigma^s_{\ell m})$$

,where :

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

$$\sum_{m=1}^{M_s} \pi^s_{\ell m} = 1$$

Speaker Dependent Prior

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

$$
\begin{aligned}
p(\mathbf{s}) &= \sum_{\ell} p_\ell \times p(\mathbf{s}|\ell) \\
&= p \times \sum_{m=1}^{M_s} \pi^s_{m_{target}} \mathcal{N}(\mathbf{s}; \mu^s_{m_{target}}, \Sigma^s_{m_{target}}) \\
&+ (1-p) \times \sum_{m=1}^{M_s} \pi^s_{m_{UBM}} \mathcal{N}(\mathbf{s}; \mu^s_{m_{UBM}}, \Sigma^s_{m_{UBM}})
\end{aligned}
$$

- Note: $|\mathcal{L}| = 2$

Speaker Dependent Prior

GMM for clean speech

We obtain the Gaussian Mixture distribution for clean speech:

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

, where $\pi^s = \begin{pmatrix} \pi_1^s \\ \vdots \\ \pi_{M_s|\mathcal{L}|}^s \end{pmatrix} = \begin{pmatrix} p\pi_{Target} \\ (1-p)\pi_{UBM} \end{pmatrix}$

Speaker Dependent Prior

Indicator Variable

- 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

$$\mathbf{z}_s^T = (\underbrace{0, \cdots, 0, \underset{i^{th}}{1}, 0, \cdots, 0}_{TargetSpeakerModel}, \underbrace{0 \cdots 0}_{UBM})$$

- Relationship between indicator $\mathbf{z}_s$ and mixture coefficients π^s:
  - $p(z_{s,i} = 1) = \pi_i^s$
  - $p(\mathbf{z}_s) = \prod_{i=1}^{M_s|\mathcal{L}|} \pi_i^{z_{s,i}}$

Speaker Dependent Prior

Indicator Variable

- 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

$$\mathbf{z}_s^T = (\underbrace{0, \cdots, 0, \underset{i^{th}}{1}, 0, \cdots, 0,}_{TargetSpeakerModel} \underbrace{0 \cdots 0}_{UBM})$$

- Relationship between indicator $\mathbf{z}_s$ and mixture coefficients π^s:
  - $p(z_{s,i} = 1) = \pi_i^s$
  - $p(\mathbf{z}_s) = \prod_{i=1}^{M_s|\mathcal{L}|} \pi_i^{z_{s,i}}$

Speaker Dependent Prior

Indicator Variable

- 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

$$\mathbf{z}_s^T = (\underbrace{0, \cdots, 0, \underset{i^{th}}{1}, 0, \cdots, 0,}_{TargetSpeakerModel} \underbrace{0 \cdots 0}_{UBM})$$

- Relationship between indicator $\mathbf{z}_s$ and mixture coefficients π^s:
  - $p(z_{s,i} = 1) = \pi_i^s$
  - $p(\mathbf{z}_s) = \prod_{i=1}^{M_s|\mathcal{L}|} \pi_i^{z_{s,i}}$

Speaker Dependent Prior

Speaker Dependent Prior

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

Then we can obtain

$$p(\mathbf{s}|\mathbf{z}_s) = \prod_{i=1}^{M_s|\mathcal{L}|} \mathcal{N}(\mathbf{s}; \mu_i^s, \Sigma_i^s)^{z_{s,i}} \quad (1)$$

Speaker Dependent Prior

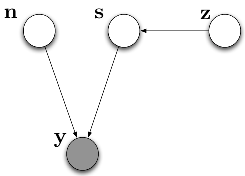
We can write: $p(\mathbf{s}|z_{s,i}=1)=\mathcal{N}(\mathbf{s};\mu_i^s,\Sigma_i^s)$

Then we can obtain

$$p(\mathbf{s}|\mathbf{z}_s)=\prod_{i=1}^{M_s|\mathcal{L}|}\mathcal{N}(\mathbf{s};\mu_i^s,\Sigma_i^s)^{z_{s,i}} \tag{1}$$

Probability Model

Markov Chain between variables



$$\boxed{p(\mathbf{y}, \mathbf{s}, \mathbf{z}_s, \mathbf{n}) = p(\mathbf{y}|\mathbf{s}, \mathbf{n}) \times p(\mathbf{s}|\mathbf{z}_s) \times p(\mathbf{z}_s) \times p(\mathbf{n})}$$

Probability Model

Markov Chain between variables

The joint distribution:

$$
\begin{aligned}
p(\mathbf{y}, \mathbf{s}, \mathbf{z}_s, \mathbf{n}) &= p(\mathbf{y}|\mathbf{s}, \mathbf{n}, \mathbf{z}_s) \times p(\mathbf{s}, \mathbf{n}|\mathbf{z}_s) \times p(\mathbf{z}_s) \\
&\overset{(a)}{=} p(\mathbf{y}|\mathbf{s}, \mathbf{n}, \mathbf{z}_s) \times p(\mathbf{s}|\mathbf{z}_s) \times p(\mathbf{n}|\mathbf{z}_s) \times p(\mathbf{z}_s) \\
&\overset{(b)}{=} p(\mathbf{y}|\mathbf{s}, \mathbf{n}) \times p(\mathbf{s}|\mathbf{z}_s) \times p(\mathbf{z}_s) \times p(\mathbf{n})
\end{aligned}
$$

(a) is because that given $\mathbf{z}_s$, $\mathbf{s}$ and $\mathbf{n}$ are conditionally independent
(b) is because of Markov property

Outline

Problem Reiterate

- Speech model in log spectrum features:

$$\mathbf{y} \approx \mathbf{s} + \log(1 + \exp(\mathbf{n} - \mathbf{s}))$$

- Probability Model:

$$p(\mathbf{y}, \mathbf{s}, \mathbf{z}_s, \mathbf{n}) = p(\mathbf{y}|\mathbf{s}, \mathbf{n}) \times p(\mathbf{s}|\mathbf{z}_s) \times p(\mathbf{z}_s) \times p(\mathbf{n})$$

  - $p(\mathbf{y}|\mathbf{s}, \mathbf{n}) = \mathcal{N}(\mathbf{y}|\mathbf{s} + \log(1 + \exp(\mathbf{n} - \mathbf{s})),\ \psi)$
  - $p(\mathbf{s}|\mathbf{z}_s) = \prod_{i=1}^{M_s|\mathcal{L}|} \mathcal{N}(\mathbf{s}; \mu_i^s, \Sigma_i^s)^{z_{s,i}}$
  - $p(\mathbf{z}_s) = \prod_{i=1}^{M_s|\mathcal{L}|} \pi_i^{z_{s,i}} = \prod_{i=1}^{M_s|\mathcal{L}|} \gamma_i^{z_{s,i}}$
  - $p(\mathbf{n}) = \mathcal{N}(\mathbf{n}; \mu_n, \Sigma_n)$ by assumption

Problem Reiterate

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

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

- We need to replace $p(\Theta|\mathbf{y})$ by $q(\Theta)$ as approximate posterior by VB method.
- Calculate $q^*(\mathbf{s})$, $q^*(\mathbf{z_s})$ and $q^*(\mathbf{n})$, then $\hat{\Theta}_{MMSE} = \{\mu_s^*, \Sigma_s^*, \mu_n^*, \Sigma_n^*, \gamma_i^*\}$ for $i \in \{1, \ldots, M_s|\mathcal{L}|\}$

Problem Reiterate

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

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

- We need to replace $p(\Theta|\mathbf{y})$ by $q(\Theta)$ as approximate posterior by VB method.
- Calculate $q^*(\mathbf{s})$, $q^*(\mathbf{z_s})$ and $q^*(\mathbf{n})$, then $\hat{\Theta}_{MMSE} = \{\mu_s^*, \Sigma_s^*, \mu_n^*, \Sigma_n^*, \gamma_i^*\}$ for $i \in \{1, \ldots, M_s|\mathcal{L}|\}$

Problem Reiterate

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

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

- We need to replace $p(\Theta|\mathbf{y})$ by $q(\Theta)$ as approximate posterior by VB method.
- Calculate $q^*(\mathbf{s})$, $q^*(\mathbf{z_s})$ and $q^*(\mathbf{n})$, then $\hat{\Theta}_{MMSE} = \{\mu_s^*, \Sigma_s^*, \mu_n^*, \Sigma_n^*, \gamma_i^*\}$ for $i \in \{1, \ldots, M_s|\mathcal{L}|\}$

Problem Reiterate

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

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

- We need to replace $p(\Theta|\mathbf{y})$ by $q(\Theta)$ as approximate posterior by VB method.
- Calculate $q^*(\mathbf{s})$, $q^*(\mathbf{z_s})$ and $q^*(\mathbf{n})$, then $\hat{\Theta}_{MMSE} = \{\mu_s^*, \Sigma_s^*, \mu_n^*, \Sigma_n^*, \gamma_i^*\}$ for $i \in \{1, \ldots, M_s|\mathcal{L}|\}$

Review General VB solution in previous slides:

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

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:
  let $\Theta = \{\mathbf{s}, \mathbf{z}_s, \mathbf{n}\}$ and $q(\Theta) = q(\mathbf{s})q(\mathbf{z}_s)q(\mathbf{n})$
  - $$\begin{aligned} q^*(\mathbf{n}) &= \mathbb{E}\{\log p(\mathbf{y},\ \mathbf{s}, \mathbf{z}_s, \mathbf{n})\}_{q(\mathbf{z}_s)q(\mathbf{s})} + C_1 \\ &= \mathbb{E}\{\log p(\mathbf{y}|\mathbf{s}, \mathbf{n})\}_{q(\mathbf{s})} + \\ &\quad \mathbb{E}\{\log p(\mathbf{s}|\mathbf{z}_s)\}_{q(\mathbf{z}_s)q(\mathbf{s})} + \mathbb{E}\{\log p(\mathbf{z}_s)\}_{q(\mathbf{z}_s)} + C_1 \end{aligned}$$
  - $$\begin{aligned} q^*(\mathbf{s}) &= \mathbb{E}\{\log p(\mathbf{y},\ \mathbf{s}, \mathbf{z}_s, \mathbf{n})\}_{q(\mathbf{z}_s)q(\mathbf{n})} + C_2 \\ &= \mathbb{E}\{\log p(\mathbf{y}|\mathbf{s}, \mathbf{n})\}_{q(\mathbf{n})} + \\ &\quad \mathbb{E}\{\log p(\mathbf{s}|\mathbf{z}_s)\}_{q(\mathbf{z}_s)} + \mathbb{E}\{\log p(\mathbf{z}_s)\}_{q(\mathbf{z}_s)} + \\ &\quad \mathbb{E}\{\log p(\mathbf{n})\}_{q(\mathbf{n})} + C_2 \end{aligned}$$
  - $$\begin{aligned} q^*(\mathbf{z}_s) &= \mathbb{E}\{\log p(\mathbf{y},\ \mathbf{s}, \mathbf{z}_s, \mathbf{n})\}_{q(\mathbf{s})q(\mathbf{n})} + C_2 \\ &= \mathbb{E}\{\log p(\mathbf{y}|\mathbf{s}, \mathbf{n})\}_{q(\mathbf{n})q(\mathbf{s})} + \mathbb{E}\{\log p(\mathbf{s}|\mathbf{z}_s)\}_{q(\mathbf{s})} \\ &\quad \mathbb{E}\{\log p(\mathbf{n})\}_{q(\mathbf{n})} + C_3 \end{aligned}$$

Likelihood Linearization

- Observation Likelihood:

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

  with non linear mean value $\mathbf{s} + \log(1 + \exp(\mathbf{n} - \mathbf{s}))$

- New Problem Arises: How to calculate $\mathbb{E}\{\log p(\mathbf{y}|\mathbf{s},\mathbf{n})\}_{q(\Theta\setminus j)}$?

Likelihood Linearization

- Observation Likelihood:

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

  with non linear mean value $\mathbf{s} + \log(1+\exp(\mathbf{n}-\mathbf{s}))$

- New Problem Arises: How to calculate $\mathbb{E}\{\log p(\mathbf{y}|\mathbf{s},\mathbf{n})\}_{q(\Theta\setminus j)}$?

Likelihood Linearization

- Linearized likelihood:
  $\hat{p}(\mathbf{y}|\mathbf{s},\mathbf{n}) = \mathcal{N}\left(\mathbf{y}\middle|(\mathbf{s} + g([\mathbf{s}_0,\mathbf{n}_0]) + \mathbf{G} \times ([\mathbf{s},\mathbf{n}] - [\mathbf{s}_0,\mathbf{n}_0])\,),\ \ \psi\right)$
- Linearization is by the first order Taylor series expansion around the point $[\mathbf{s}_0,\mathbf{n}_0]$

$$g([\mathbf{s},\mathbf{n}]) = \log(1+\exp(\mathbf{n}-\mathbf{s}) \overset{(c)}{\approx} g([\mathbf{s}_0,\mathbf{n}_0]) + \mathbf{G} \times ([\mathbf{s},\mathbf{n}] - [\mathbf{s}_0,\mathbf{n}_0])$$

$\mathbf{G} = [\mathbf{G}_s, \mathbf{G}_n] \overset{def}{=} \nabla g([\mathbf{s}_0,\mathbf{n}_0])$, and

$$\mathbf{G}_s = diag[\frac{-\exp(n_0^1 - s_0^1)}{1+\exp(n_0^1 - s_0^1)}, \cdots, \frac{-\exp(n_0^N - s_0^N)}{1+\exp(n_0^N - s_0^N)}]$$
$$\mathbf{G}_n = diag[\frac{\exp(n_0^1 - s_0^1)}{1+\exp(n_0^1 - s_0^1)}, \cdots, \frac{\exp(n_0^N - s_0^N)}{1+\exp(n_0^N - s_0^N)}]$$

,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 $[\mathbf{s}_0, \mathbf{n}_0] = [\mu_s^*, \mu_n^*]$;

E-STEP: Compute $\mathbf{G} = [\mathbf{G}_s, \mathbf{G}_n]$ and $g([\mathbf{s}_0, \mathbf{n}_0])$;

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

Update γ_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

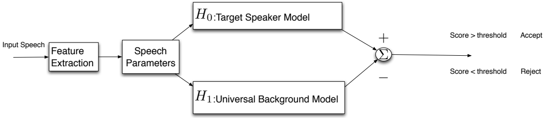
Enhancement Summary

$$\boxed{\hat{\mathbf{n}}_{MSEE} = \mathbb{E}[\mathbf{n}|\mathbf{y}]} \longleftarrow \boxed{p(\mathbf{n}|\mathbf{y})} \xleftarrow{\text{VB Algorithm}} \boxed{q^*(\mathbf{y})}$$

$$\boxed{\hat{\mathbf{s}}_{MSEE} = \mathbb{E}[\mathbf{s}|\mathbf{y}]} \longleftarrow \boxed{p(\mathbf{s}|\mathbf{y})} \xleftarrow{\text{VB Algorithm}} \boxed{q^*(\mathbf{s})}$$

$$\boxed{\hat{\mathbf{z}}_{\mathbf{s}MSEE} = \mathbb{E}[\mathbf{z}_{\mathbf{s}}|\mathbf{y}]} \longleftarrow \boxed{p(\mathbf{z}_{\mathbf{s}}|\mathbf{y})} \xleftarrow{\text{VB Algorithm}} \boxed{q^*(\mathbf{z}_{\mathbf{s}})}$$

Speaker Verification using Enhanced Features



$$Score = \log p(\mathbf{X}_{enhanced} | TargetModel) - \log p(\mathbf{X}_{enhanced} | UBM)$$

- Train the library model $\mathcal{L} = \{TargetSpeaker, UBM\}$:
  - 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 based on the intuition that clean speech improves the performance of speaker verification.
- It introduce speaker dependent priors for feature enhancement based on Algonquin Algorithm.
- It derive Variational Bayesian Algorithm to obtain the approximate posterior for clean speech.

Appendix 1

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

- Proof:

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

Take partial derivative to find minimum MSE:

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

Then we have:

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

Appendix 2

The optimal solution is:

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

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

Appendix 2

- Proof:

$$
\begin{aligned}
\mathcal{L}(q) &= \int q(\Theta) \ln \left(\frac{p(\mathbf{X}, \Theta)}{q(\Theta)} \right) d\Theta \\
&= \int \prod_i q_i \left\{ \ln p(\mathbf{X}, \Theta) - \sum_i \ln q_i \right\} d\Theta \\
&= \int q_j \left\{ \ln p(\mathbf{X}, \Theta) \prod_{i \neq j} q_i - (\sum_i \ln q_i) \prod_{i \neq j} q_i \right\} d\Theta \\
&= \int q_j \left\{ \ln p(\mathbf{X}, \Theta) \prod_{i \neq j} q_i d\Theta_i \right\} d\Theta_j - \int q_j \ln q_j d\Theta_j + Const \\
&= \int q_i \mathbb{E}[\ln p(\mathbf{X}, \Theta)]_{q(\Theta \setminus j)} d\Theta_j - \int q_j \ln q_j d\Theta_j + Const \\
&= -KL \left(q_i \middle\| \mathbb{E}[\ln p(\mathbf{X}, \Theta)]_{q(\Theta \setminus j)} \right) + Const
\end{aligned}
$$

Appendix 2

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

Appendix 3 I

$$\mathbf{y} \approx \mathbf{s} + \mathbf{h} + \log(1 + \exp(\mathbf{n} - \mathbf{h} - \mathbf{s}))$$

- Proof:
  Given $Y[k] = H[k]S[k] + N[k]$, we have:

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

$$= |H[k]|^2|S[k]|^2 + |N[k]|^2 + 2Re\{(H[k]S[k]) \times N[k]^*\} \tag{3}$$

$$\approx |H[k]|^2|S[k]|^2 + |N[k]|^2 \tag{4}$$

  Let $\mathbf{y} = \log|Y[:]|^2$, then $|Y[:]|^2 = \exp(\mathbf{y})$ and similarly for $\mathbf{s}$,$\mathbf{h}$ and $\mathbf{n}$.
  We can rewrite (3) as:

Appendix 3 II

$$\begin{aligned}\exp(\mathbf{y}) &= \exp(\mathbf{s}+\mathbf{h}) + \exp(\mathbf{n}) \\ &= \exp(\mathbf{s}+\mathbf{h}) \circ (1 + \exp(\mathbf{n}-\mathbf{s}-\mathbf{h}))\end{aligned}$$

Taking log for both sides, we have:

$$\mathbf{y} \approx \mathbf{s} + \mathbf{h} + \log(1 + \exp(\mathbf{n} - \mathbf{h} - \mathbf{s}))$$

Appendix 4: Compute G I

Given$\mathbf{G} = [\mathbf{G}_s, \mathbf{G}_n] \stackrel{def}{=} \nabla g([\mathbf{s}_0, \mathbf{n}_0])$, we have

$$\mathbf{G} = \nabla g([\mathbf{s}_0, \mathbf{n}_0]) = \nabla g([\mathbf{s}, \mathbf{n}])\Big|_{[\mathbf{s},\mathbf{n}]=[\mathbf{s}_0,\mathbf{n}_0]}$$
$$= \nabla(\log(\mathbf{1} + \exp(\mathbf{n} - \mathbf{s}))\Big|_{[\mathbf{s},\mathbf{n}]=[\mathbf{s}_0,\mathbf{n}_0]}$$

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

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

similarly,

Appendix 4: Compute G II

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

Therefore:

$$\mathbf{G}_s = diag[\frac{-\exp(n_0^1 - s_0^1)}{1 + \exp(n_0^1 - s_0^1)}, \cdots, \frac{-\exp(n_0^N - s_0^N)}{1 + \exp(n_0^N - s_0^N)}]$$

$$\mathbf{G}_n = diag[\frac{\exp(n_0^1 - s_0^1)}{1 + \exp(n_0^1 - s_0^1)}, \cdots, \frac{\exp(n_0^N - s_0^N)}{1 + \exp(n_0^N - s_0^N)}]$$

Appendix 5

$$q^*(\mathbf{s}) = \mathbb{E}\{\log p(\mathbf{y},\ \mathbf{s}, \mathbf{z}_s, \mathbf{n})\}_{q(\mathbf{z}_s)q(\mathbf{n})} + C_1 = \mathcal{N}(\mathbf{s}; \mu_s^*, \Sigma_s^*)$$

with

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

$$\mu_s^* = \Sigma_s^*[(\mathbf{I} + \mathbf{G}_s^T)\psi^{-1}(\mathbf{y} - g([\mathbf{s}_0, \mathbf{n}_0]) - \mathbf{G}_n\mu_n^* + \mathbf{G}_s\mathbf{s}_0 + \mathbf{G}_n\mathbf{n}_0) + \sum_{i=1}^{M_s|\mathcal{L}|}\gamma_i(\Sigma_i^s)^{-1}\mu_i^s]$$

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

$$\Sigma_n^* = [\mathbf{G}_n^T \psi^{-1} \mathbf{G}_n + \Sigma_n^{-1}]^{-1}$$

$$\mu_n^* = \Sigma_n^* [\mathbf{G}_n^T \psi^{-1} (\mathbf{y} - \mu_s^* - g([\mathbf{s}_0, \mathbf{n}_0]) - \mathbf{G}_s \mu_s^* + \mathbf{G}_s \mathbf{s}_0 + \mathbf{G}_n \mathbf{n}_0) + \Sigma_n^{-1} \mu_n]$$

Appendix 5

$$q^*(\mathbf{z}_s) = \mathbb{E}\{\log p(\mathbf{y},\ \mathbf{s}, \mathbf{z}_s, \mathbf{n})\}_{q(\mathbf{s})q(\mathbf{n})} + C_3 = \prod_{i=1}^{M_s|\mathcal{L}|} (\gamma_i)^{z_{s,i}}$$

with

$$\gamma_i = \frac{\rho_i}{\prod_{i=1}^{M_s|\mathcal{L}|} \rho_i}$$

$$\begin{aligned} \log \rho_i = & -\frac{1}{2}(\mu_s^* - \mu_i^s)^T (\Sigma_i^s)^{-1} (\mu_s^* - \mu_i^s) \\ & - \frac{1}{2} \log |\Sigma_i^s| - \frac{1}{2} Tr((\Sigma_i^s)^{-1} \Sigma_s^*) + \log \pi_i^s \end{aligned}$$

Ciira wa Maina, John MacLaren Walsh
Log Spectra Enhancement using Speaker Dependent Priors for Speaker Verification
IEEE Trans. Audio, Speech, Language Processing., submitted March 14, 2011. Revised July 15, 2011.

Ciira wa Maina
Approximate Bayesian Inference for Robust Speech Processing
Ph.D Thesis, June 2011

Christopher M. Bishop
Pattern Recognition and Machine Learning
8^{th} edition, 2009, Springer

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