Senone I-Vectors for Robust Speaker Verification

Zhili Tan[1], Yingke Zhu[2], Manwai Mak[1], Brian Mak[2]

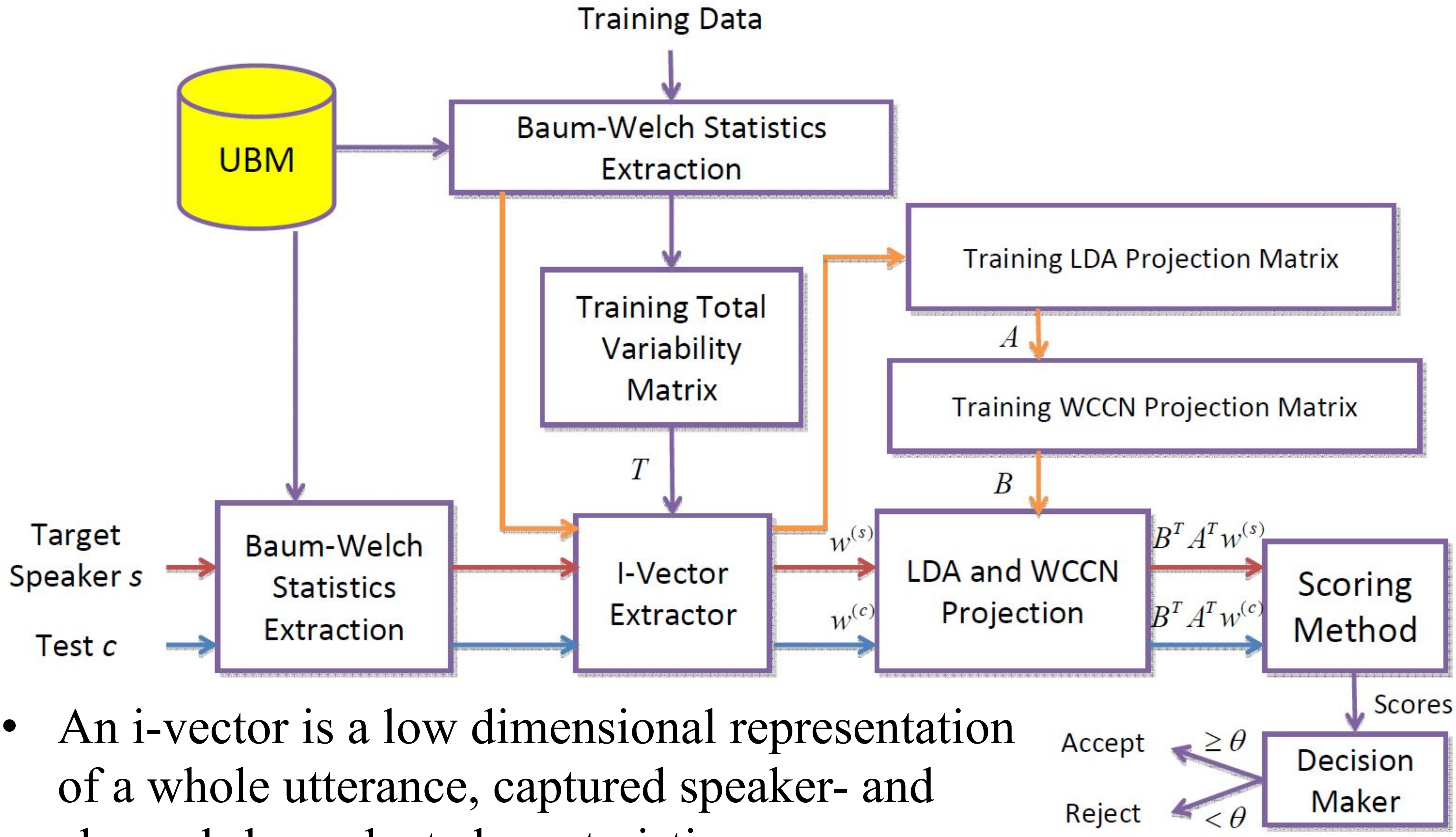
[1]The Hong Kong Polytechnic University

[2]The Hong Kong University of Science and Technology

Contents

1. I-Vector for Speaker Recognition
2. Motivation of Work
3. Conventional I-Vectors
4. DNN I-Vectors
5. Experiments on NIST 2012 Evaluation Set
6. Conclusions and Future Work

I-Vector Based Speaker Verification



- An i-vector is a low dimensional representation of a whole utterance, captured speaker- and channel-dependent characteristics.

Motivation

- Integrates phonetic information into i-vectors by DNN
  - Extracts bottleneck (BN) features
  - Estimates senone posteriors

- Denoises the MFCC vectors through the deoising autoencoder

- This architecture allows us to extract BN features and estimates senone posteriors given noisy MFCCs as input, resulting in robust senone i-vectors.

Sufficient Statistics

- Given the observed acoustic feature vectors of speaker *i*, $O_i = \{o_1, ..., o_{iT_i}\}$, we can calculate the sufficient statistics corresponding to the Gaussian mixture *c*:

Posterior: $$\gamma_c(\boldsymbol{o}_{it}) = \frac{\lambda_c^{(b)} \mathcal{N}(\boldsymbol{o}_{it} | \boldsymbol{\mu}_c^{(b)}, \boldsymbol{\Sigma}_c^{(b)})}{\sum_{j=1}^{C} \lambda_j^{(b)} \mathcal{N}(\boldsymbol{o}_{it} | \boldsymbol{\mu}_j^{(b)}, \boldsymbol{\Sigma}_j^{(b)})}$$

0th order statistic $$N_{ic} = \sum_t \gamma_c(\boldsymbol{o}_{it})$$

1st order statistic $$\tilde{\boldsymbol{f}}_{ic} = \sum_t \gamma_c(\boldsymbol{o}_{it})(\boldsymbol{o}_{it} - \boldsymbol{\mu}_c^{(b)})$$

2nd order statistic $$\boldsymbol{S}_{ic} = \sum_t \gamma_c(\boldsymbol{o}_{it})(\boldsymbol{o}_{it} - \boldsymbol{\mu}_c)(\boldsymbol{o}_{it} - \boldsymbol{\mu}_c)^\mathsf{T}$$

Total Variability Matrix Training

- I-vector model: $\boldsymbol{\mu}_i = \boldsymbol{\mu}^{(b)} + \boldsymbol{T}\boldsymbol{w}_i + \boldsymbol{\epsilon}_i$
- E-step:

$$\langle \boldsymbol{w}_i | \mathcal{O}_i \rangle = \boldsymbol{L}_i^{-1} \sum_c \boldsymbol{T}_c^{\intercal} (\boldsymbol{\Sigma}_c^{(b)})^{-1} \tilde{\boldsymbol{f}}_{ic}$$

$$\langle \boldsymbol{w}_i \boldsymbol{w}_i^{\intercal} | \mathcal{O}_i \rangle = \boldsymbol{L}_i^{-1} + \langle \boldsymbol{w}_i | \mathcal{O}_i \rangle \langle \boldsymbol{w}_i | \mathcal{O}_i \rangle^{\intercal}$$

$$\boldsymbol{L}_i = \boldsymbol{I} + \boldsymbol{T}^{\intercal} (\boldsymbol{\Sigma}^{(b)})^{-1} \boldsymbol{N}_i \boldsymbol{T}$$

All we need:

N_{ic}

$\tilde{\boldsymbol{f}}_{ic}$

$\boldsymbol{\Sigma}_c^{(b)}$

- M-step:

$$\boldsymbol{T}_c = \left[\sum_i \tilde{\boldsymbol{f}}_{ic} \langle \boldsymbol{w}_i | \mathcal{O}_i \rangle^{\intercal} \right] \left[\sum_i N_{ic} \langle \boldsymbol{w}_i \boldsymbol{w}_i^{\intercal} | \mathcal{O}_i \rangle \right]^{-1}$$

Mean Vector and Covariance Matrix

- In most systems, $\{\boldsymbol{\mu}_c\}$ and $\{\boldsymbol{\Sigma}_c\}$ are obtained from the UBM.
- However, they can also be obtained using the sufficient statistics:

$$\boldsymbol{\mu}_c = \frac{\sum_i \sum_t \gamma_c(\boldsymbol{o}_{it})\boldsymbol{o}_{it}}{\sum_i \sum_t \gamma_c(\boldsymbol{o}_{it})}$$

$$\boldsymbol{\Sigma}_c = \frac{\sum_i \sum_t \gamma_c(\boldsymbol{o}_{it})(\boldsymbol{o}_{it} - \boldsymbol{\mu}_c)(\boldsymbol{o}_{it} - \boldsymbol{\mu}_c)^{\mathsf{T}}}{\sum_i \sum_t \gamma_c(\boldsymbol{o}_{it})}$$

General Type of I-Vector

- Only the acoustic vectors $\boldsymbol{o}_{it}$ and the mixture posteriors $\gamma_c(\boldsymbol{o}_{it})$ are necessary for i-vector extraction.
- Given the speech signal of the t-th frame in the i-th utterance $\boldsymbol{s}_{it}$
  - The MFCC could be replaced by other types of acoustic features:

$$\boldsymbol{o}_{it} = f(\boldsymbol{s}_{it})$$

  - The mixture posteriors could be estimated from other model rather than GMM:

$$\gamma_c(\boldsymbol{s}_{it}) = P(c|\boldsymbol{s}_{it})$$

Senone I-Vector

- The general type of i-vector allows the integration of supervised signal, with the standard backends remaining unchanged.
  - The supervised information is brought by $f(\bullet)$ and c.
- For example, a deep neural network (DNN) trained for ASR can help to integrate the phonetic information into i-vectors.
  - The BN features replaces MFCCs as acoustic features;
  - The posteriors of senones replaces the GMM mixture posteriors.
- Furthermore, a denoising autoencoder integrated into i-vector extraction may help to improve the noise robustness.

DNN I-Vectors

- BN feature vectors: $\boldsymbol{o}_{it} = \mathrm{BN}(\boldsymbol{s}_{it})$
- Senone posteriors: $\gamma_c(\boldsymbol{s}_{it}) = P_{DNN}(c|\boldsymbol{s}_{it})$
  - The output of the c-th node in the softmax output layer.
- Baum-Welch statistics:

$$N_{ic} = \sum_t P_{DNN}(c|\boldsymbol{s}_{it})$$

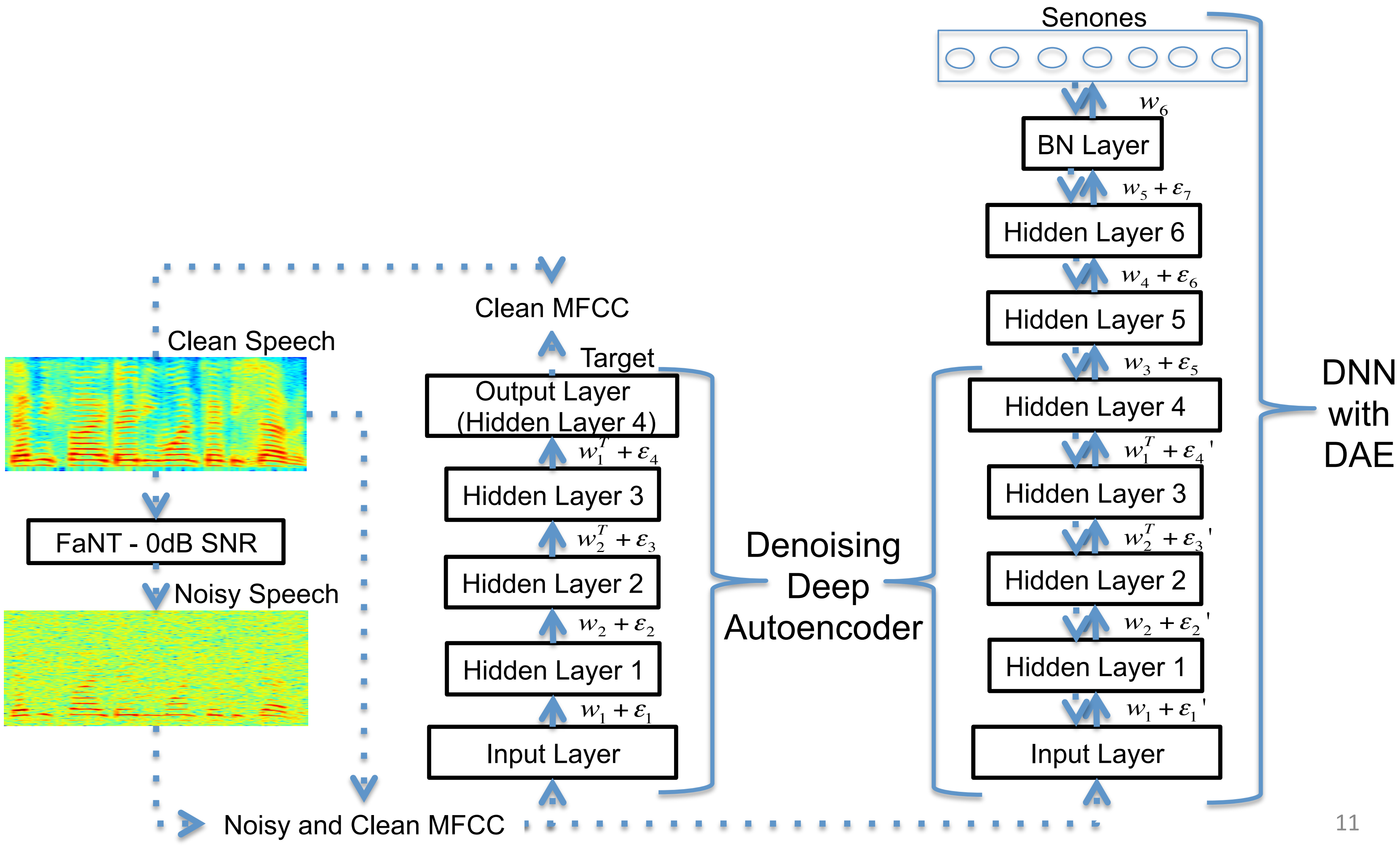
$$\tilde{\boldsymbol{f}}_{ic} = \sum_t P_{DNN}(c|\boldsymbol{s}_{it})(\mathrm{BN}(\boldsymbol{s}_{it}) - \boldsymbol{\mu}_c)$$

$$\boldsymbol{S}_{ic} = \sum_t P_{DNN}(c|\boldsymbol{s}_{it})(\mathrm{BN}(\boldsymbol{s}_{it}) - \boldsymbol{\mu}_c)(\mathrm{BN}(\boldsymbol{s}_{it}) - \boldsymbol{\mu}_c)^{\mathsf{T}}$$

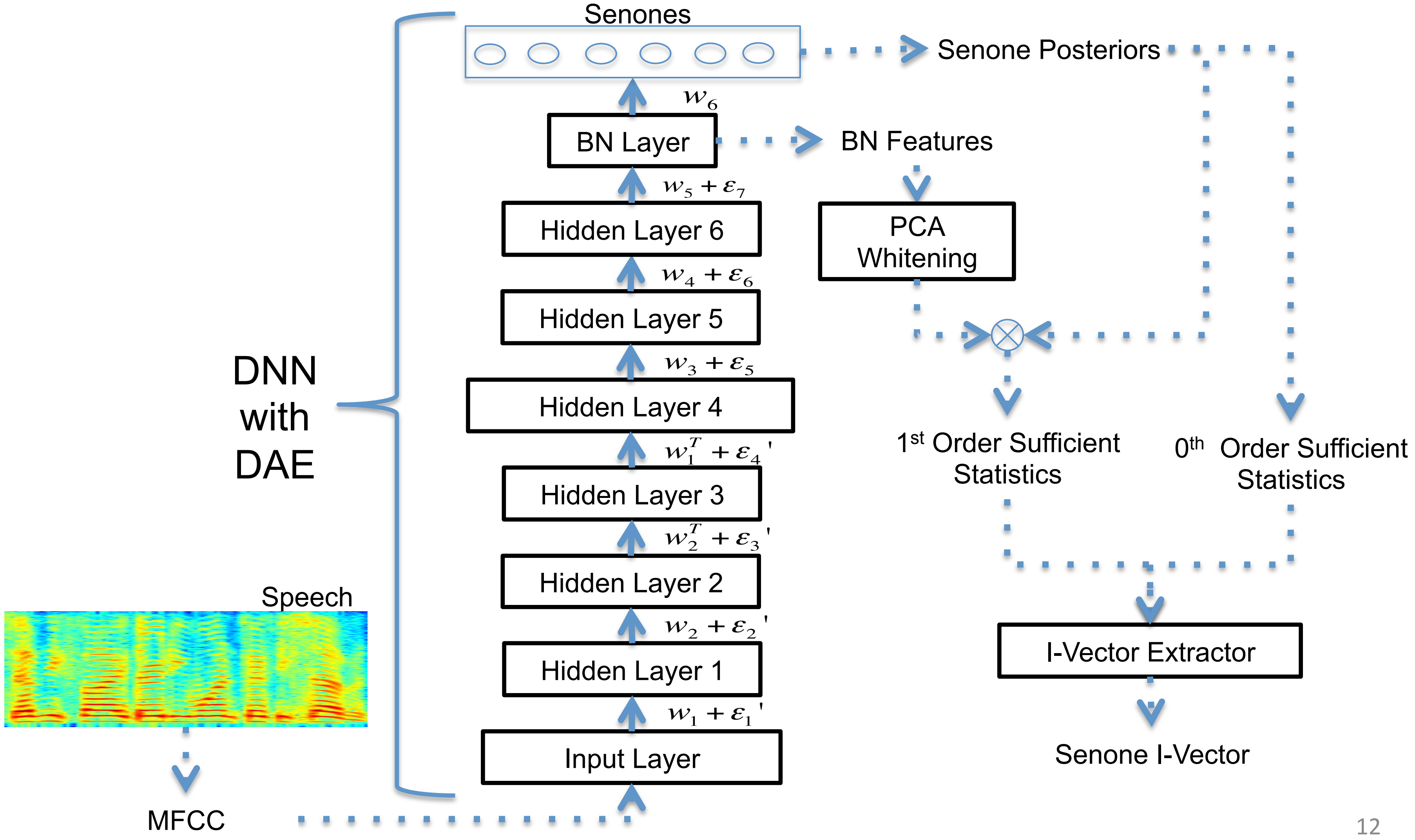
where:

$$\boldsymbol{\mu}_c = \frac{\sum_i \sum_t \gamma_c(\boldsymbol{s}_{it})\mathrm{BN}(\boldsymbol{s}_{it})}{\sum_i N_{ic}}$$

Denoising Classifier Training



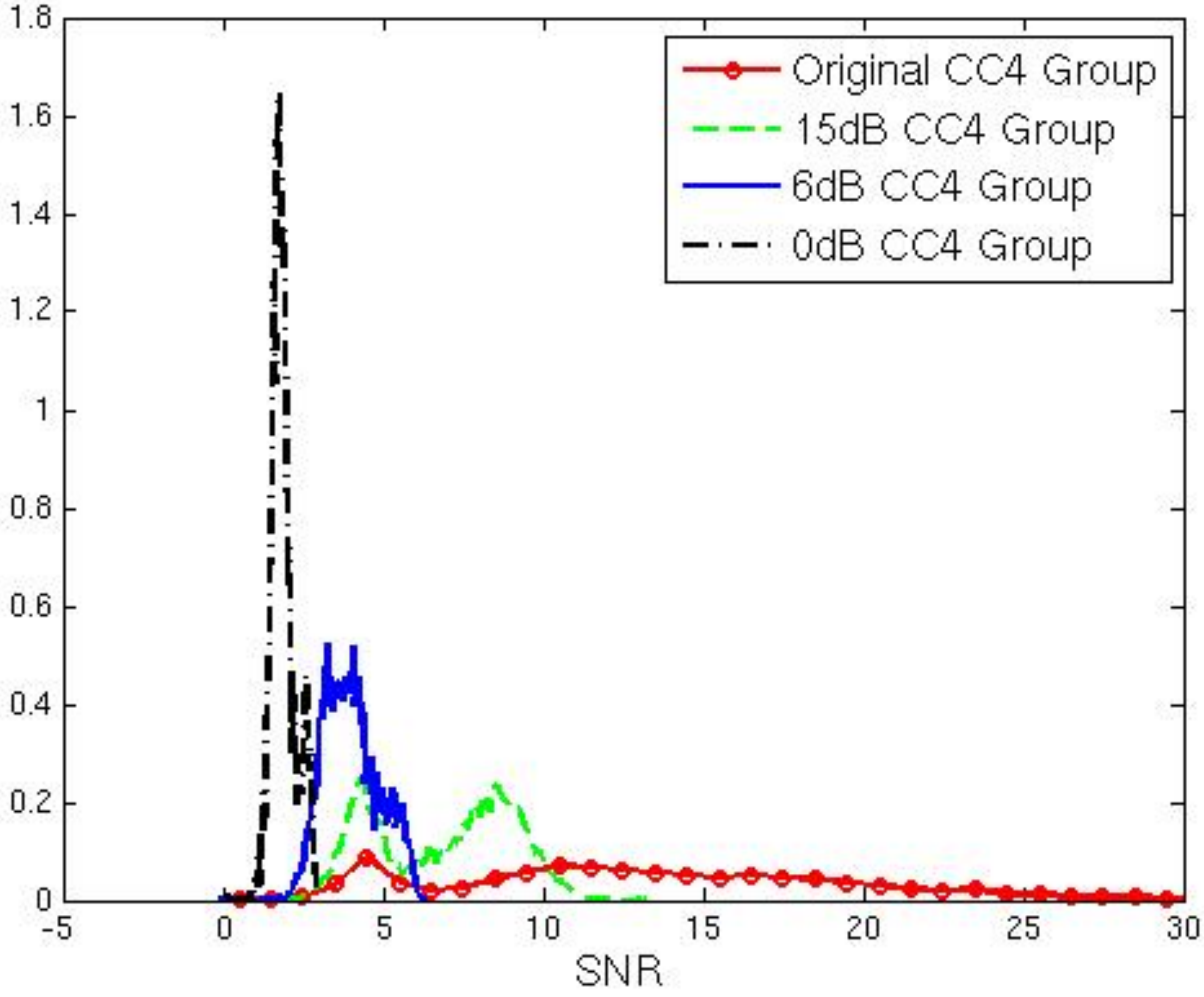
Denoising Senone I-vector Extraction



Experimental Setup

- Evaluation dataset: NIST 2012 SRE CC4
  - Add babble noise at SNR of 15dB, 6dB and 0dB
- Baseline:
  - Acoustic features: 19 MFCCs together with energy plus their 1st and 2nd derivatives→60-Dim with feature warping
  - Posteriors: from GMM with 1024 mixtures
- T-matrix
  - 500-dimensional subspace trained by clean data
- PLDA
  - 150 latent variables

SNR Distribution of Evaluation Set

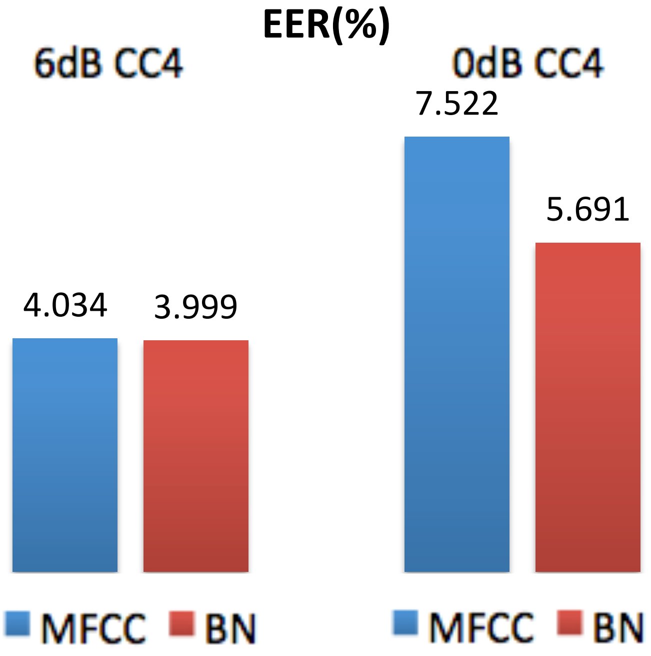


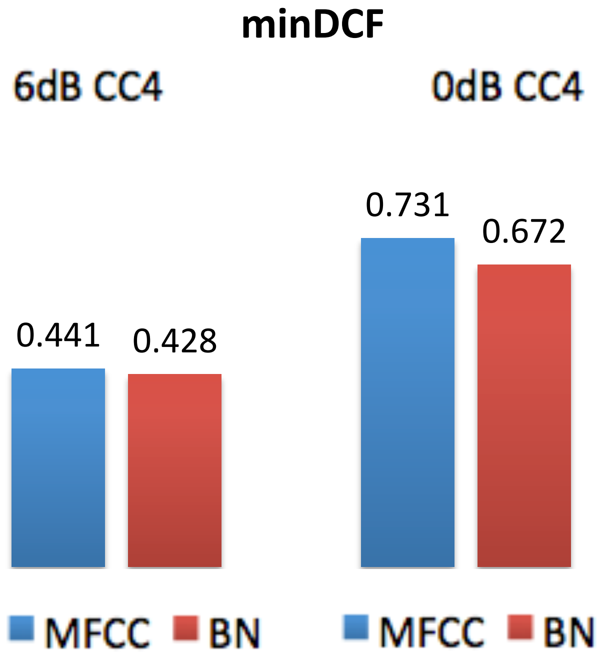
DBN Setup

- Structure: $\mathcal{D}$-256-256-256-$\mathcal{D}$-256-256-60(BN)-2000(Senone)
  - $\mathcal{D}$ represents the dimension of the input vectors
  - RBM pretraining: two Gaussian-Bernoulli RBMs and one Bernoulli-Gaussian RBM
  - BP fine-tuning: two linear activated layers
- Input of DNN:
  - 11 frames of 20-Dim MFCC without Feature Warping
  - Normalization by z-norm
- Decorrelation for BN features:
  - PCA whitening
  - GMM-UBM with diagonal covariance matrix

Result – Power of BN Features

- Posteriors from GMM-UBM are used
- Acoustic features are (1) MFCC, and (2) BN features

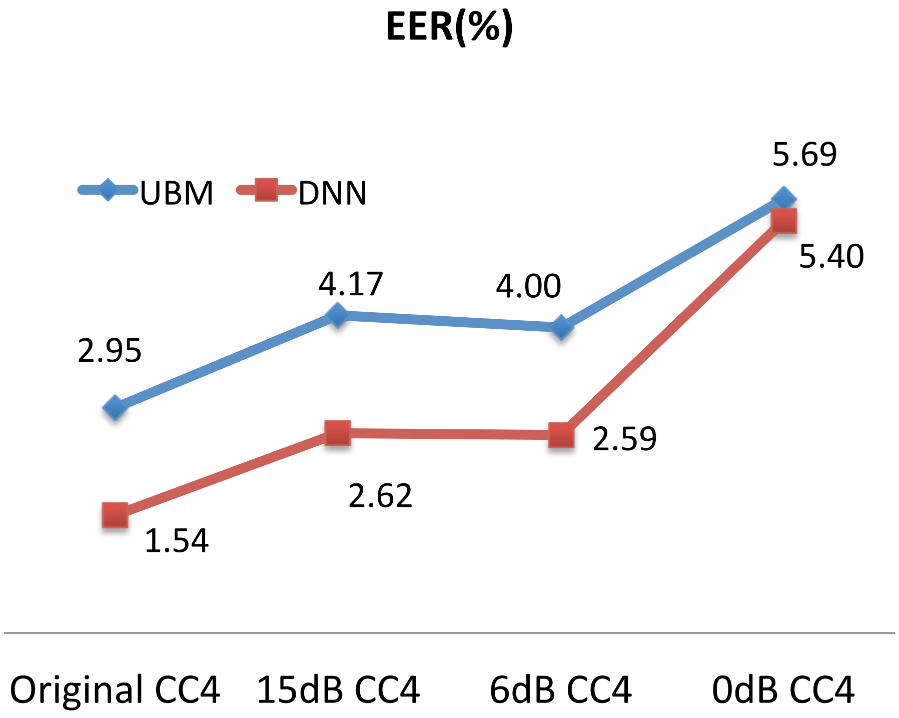


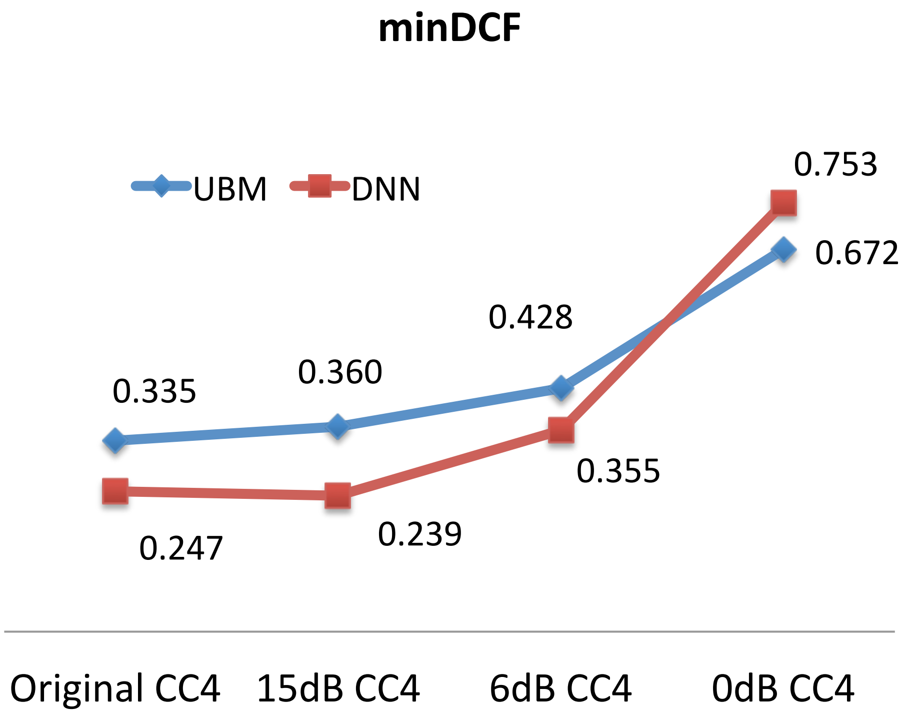


- With posteriors from GMM-UBM, BN features outperform MFCC under noisy conditions

Result – Power of Senone Posteriors

- BN features are used
- Posteriors are obtained from (1) UBM, and (2) DNN

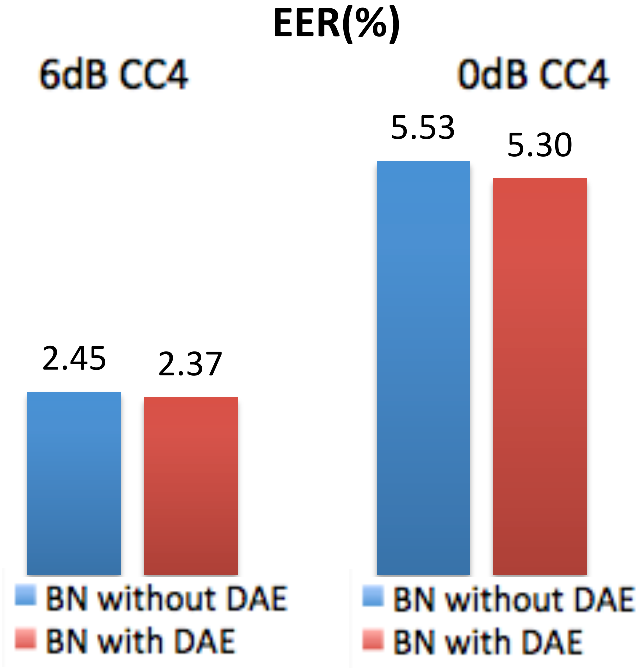


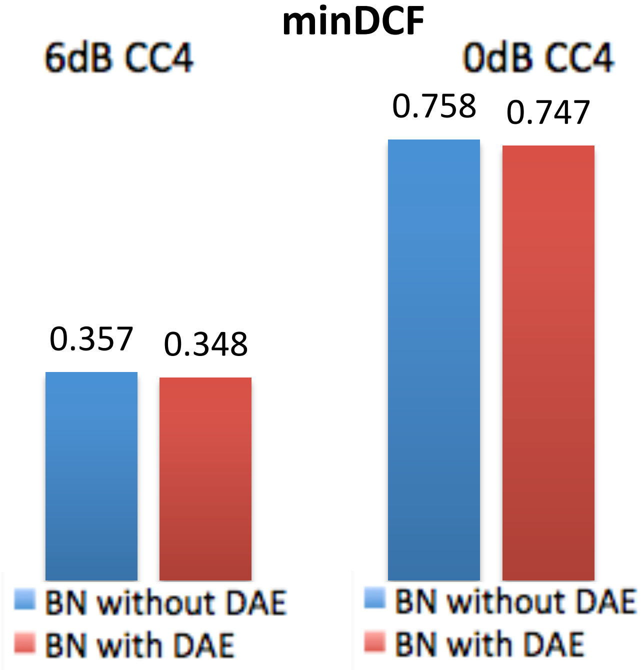


- With BN features, DNN posteriors outperform UBM posteriors

Result – Power of Denoising AE

- To verify the power of Denoising Autoencoder, another DNN without DAE training was built.
- Posteriors from DNN without DAE training
- BN features are (1) with DAE training, and (2) without DAE training





- The denoising autoencoder improves noise robustness

Conclusions and Future Work

- The senone i-vectors outperforms the conventional i-vectors under all of the SNR conditions.
- The senone information benefits i-vector extraction.
- The denoising autoencoder improves noise robustness.
- The original NIST 12 CC4 evaluation set already contains noisy speech. Experiments on only clean speech are necessary in the future.

THANKS !

Q & A

Full Result Table

		Original CC4		15dB CC4		6dB CC4		0dB CC4	
Acoustic Features	Posteriors from	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
MFCC	UBM	2.664	0.2830	3.600	0.3633	4.034	0.4412	7.522	0.7313
BN with DAE	UBM	2.945	0.3352	4.167	0.3595	3.999	0.4279	5.691	**0.6722**
BN with DAE	DNN with DAE	1.537	0.2468	2.616	0.2387	2.591	0.3545	5.404	0.7529
BN with DAE	DNN without DAE	1.476	0.2345	2.369	0.2289	**2.370**	0.3481	**5.297**	0.7465
BN without DAE	DNN with DAE	**1.330**	0.2319	**2.305**	**0.2171**	2.522	**0.3372**	5.423	0.7495
BN without DAE	DNN without DAE	1.506	**0.2219**	2.440	0.2264	2.446	0.3573	5.531	0.7575