# Two-stage Noise Aware Training Using Asymmetric Deep Denoising Autoencoder

**Kang Hyun Lee, Shin Jae Kang, Woo Hyun Kang and Nam Soo Kim**
School of Electrical and Computer Engineering and INMC, Seoul National University, Seoul, Korea
E-mail: {khlee, sjkang, whkang}@hi.snu.ac.kr, nkim@snu.ac.kr

## Introduction

- <u>Deep neural network (DNN) and progress in automatic speech recognition (ASR)</u>
  - In acoustic modeling, appearance of DNN-hidden Markov model (HMM) system is considered as a breakthrough.
  - Capability in automatically learning complicated non-linear mapping from the input to the target vectors.
  - Expanded to the robust speech recognition area.

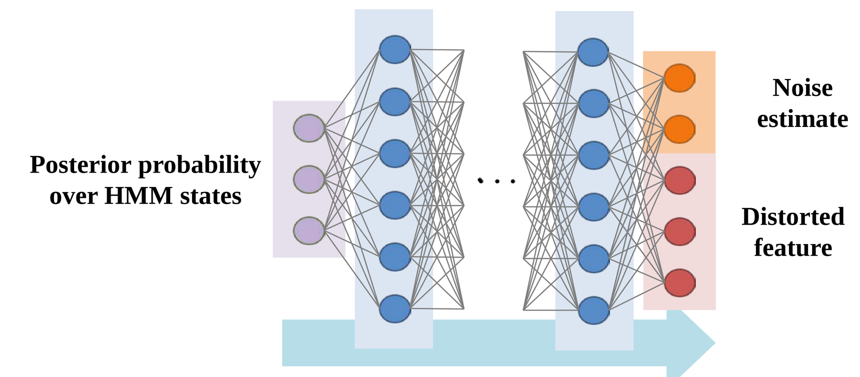
- <u>DNN-based robust speech recognition</u>
  - Feature-based approach
    - Directly trains an arbitrary unknown mapping from the noisy to the clean speech features
    - Deep denoising autoencoder (DDAE) has demonstrated its superiority in reconstructing the clean features from noisy features
  - Model-based approach
    - Let the DNN parameters find out the relationship between the observed speech and the phonetic targets
    - Noise-aware training (NAT) attained the state-of-the-art results on Aurora-4 task

- <u>Properties of noise aware training (NAT)</u>
  - Follows the general procedure of the multi-condition DNN-HMM, except for the input structure of network
  - Augments the input signal by concatenating the distorted feature and the noise estimate
    - Enables the DNN to learn the relationship among noisy input, noise features and target vectors corresponding to the phonetic identity

- <u>Remaining issues on NAT</u>
  - Is NAT an optimal method for sufficiently utilizing the inherent robustness of DNN?
    - Performance of NAT in adverse environment is still far from that in clean condition
    - A promising way to improve the NAT is to extract some hidden representation relevant to clean speech features and then to implement the mapping from this representation to the phonetic targets

- <u>What we propose?</u>
  - A novel approach to DNN training which can be a solution to the aforementioned issue of NAT
    - Let the DNN clarify the relationship among noisy features, noise estimates and phonetic targets only after reconstructing the clean features.

## Brief review on noise aware training




- We assume that there exists an unknown underlying function that approximates the posterior probabilities of the HMM states given as follows:

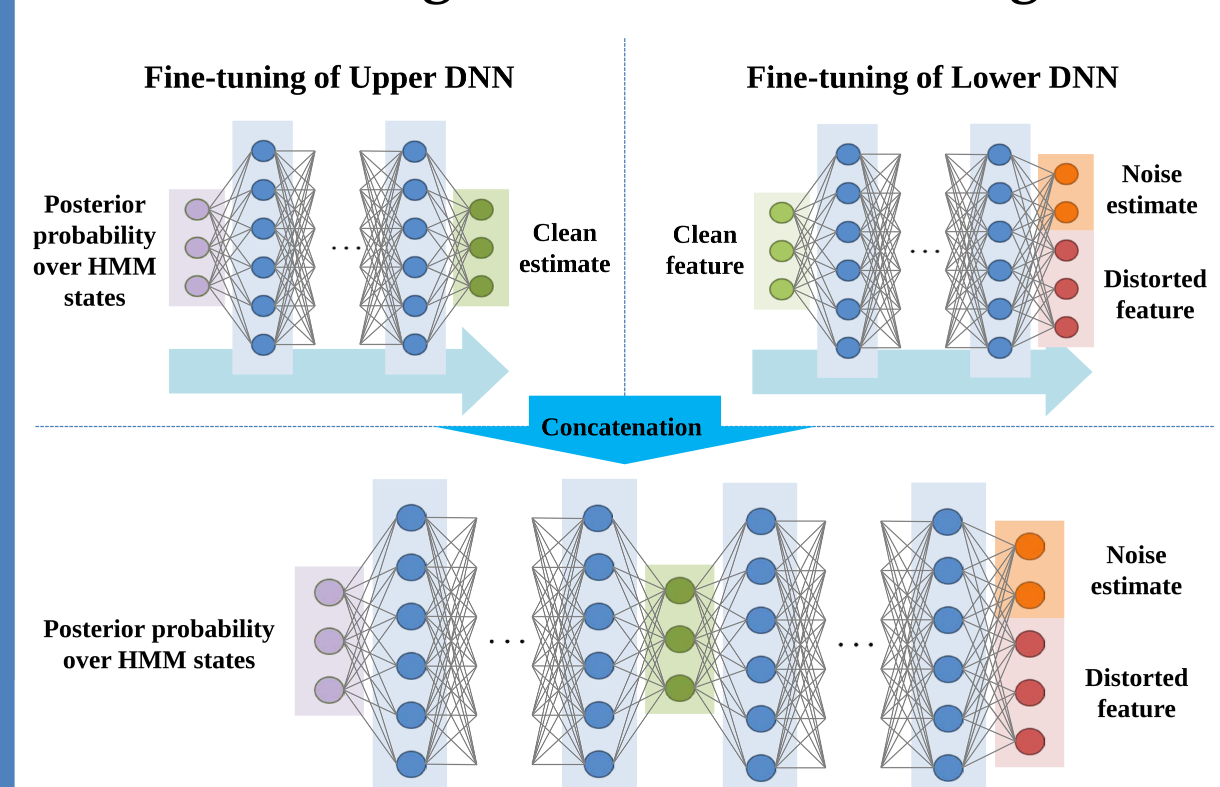
$$p(\boldsymbol{s}_t|\boldsymbol{y}_1^T) \cong f(\boldsymbol{y}_{t-\tau}^{t+\tau}, \boldsymbol{n}_{t-\tau}^{t+\tau})$$

$\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{n}_t, \boldsymbol{s}_t$ : Noisy feature, clean feature, noise feature, and HMM state identity extracted at the $t$–th frame
$\boldsymbol{y}_{m_1}^{m_2}$ : Subsequence of noisy feature vectors from frame index $m_1$ to $m_2$

- NAT replaces them with a single noise estimate
  - The input vector of NAT is formed by augmenting the noise estimate with a window of consecutive frames of noisy feature, i.e.,

$$\boldsymbol{v}_t = [\boldsymbol{y}_{t-\tau}^{t+\tau}, \hat{\boldsymbol{n}}_t]$$

## Two-stage noise aware training




Final structure

- The assumption that the function $f(\cdot)$ can be expressed as a composition of two separate functions as follows:

$$p(\boldsymbol{s}_t|\boldsymbol{y}_1^T) \cong f(\boldsymbol{y}_{t-\tau}^{t+\tau}, \boldsymbol{n}_{t-\tau}^{t+\tau}) \cong h \circ g(\boldsymbol{y}_{t-\tau}^{t+\tau}, \boldsymbol{n}_{t-\tau}^{t+\tau})$$

Where the output of $g(\cdot)$ is a clean feature vector stream,

$$\boldsymbol{x}_{t-\tau}^{t+\tau} \cong g(\boldsymbol{y}_{t-\tau}^{t+\tau}, \boldsymbol{n}_{t-\tau}^{t+\tau}),$$

and

$$p(\boldsymbol{s}_t|\boldsymbol{y}_1^T) \cong h(\boldsymbol{x}_{t-\tau}^{t+\tau})$$

$g(\cdot)$ : function which deals with the mapping from the noisy and noise features to the clean speech features
$h(\cdot)$ : function predicting the phonetic target based on the clean speech feature stream.

- Lower DNN
  - For initializing the lower DNN, DDAE is applied
  - Noise-related nodes are excluded in the output layer at the fine-tuning phase
    - The DDAE is designed to have an asymmetric structure where the dimensions of the input and output vector are different

$$\hat{\boldsymbol{v}}_t = [\hat{\boldsymbol{x}}_{t-\tau}^{t+\tau}]$$

  - A time-varying environmental estimation approach based on the interacting multiple model algorithm is utilized for noise estimation(Han, Kang and Kim, 2009)

- Upper DNN
  - The network learns the mapping between the output vector of the lower DNN $\hat{\boldsymbol{v}}_t$ and the corresponding one-hot encoding label which contains information of the HMM states.

## Experiment

- Aurora-5 task
  - Noise and reverberation on hands-free, speech digit
  - Training set : 8623 utterances (4 hours)
  - Evaluation set : 8700 utterances per each condition

| | Non-filtered | | | G. 712 filtered | | | |
|---|---|---|---|---|---|---|---|
| Noise | Interior noise | | | Car Noise | | | Street noise |
| SNR (dB) | Interior | Hands-free in office | Hands-free in living room | Car | Hands-free in car | Hands-free in car & GSM | GSM |
| Clean | Clean | Clean | Clean | Clean | Clean | Clean | Clean |
| 15 | 15 | 15 | 15 | 15 | 15 | 15 | 15 |
| 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

- Clean-condition GMM-HMM setting
  - Feature : 39 dim. MFCC feature + CMN
  - Language model : uniform unigram
  - Number of HMM states : 179-dim

- Tested DNN-based acoustic modeling methods
  - Multi-condition DNN-HMM (*Baseline*)
  - Noise aware training (*NAT*)
  - Two-stage Noise aware training (*TS-NAT*)

- Structure of DNNs
  - Lower DNN (*TS-NAT*)
    - Input vector: 69-dim. Log mel-filter bank (LMFB) feature, context window size 5, noise estimate (828 dim.)
    - 5 hidden layers with 2048 nodes , sigmoid activation
    - Target vector : 69-dim. clean LFMB feature, context window size 5 (759 dim.)
  - Upper DNN (*TS-NAT*)
    - Input vector: Reconstructed vector lower DNN (759 dim.)
    - 5 hidden layers with 2048 nodes , sigmoid activation
    - Target vector: 179 HMM state labels, softmax activation
  - *Baseline*
    - Input vector: 69-dim. LMFB feature, context window size 5 (759 dim.)
    - 11 hidden layers with 2048 nodes, sigmoid activation
    - Target vector : 179 HMM state labels, softmax activation
  - *NAT*
    - Input vector: 69-dim. LMFB feature, context window size 5, noise estimate (828 dim.)
    - Same with *Baseline* in hidden layer and target vector

- Performance evaluations
  - WER (%) on Aurora-5 task

| SNR (DB) | Non-filtered | | | G.712 filtered | | |
|---|---|---|---|---|---|---|
| Method | Baseline | NAT | TS-NAT | Baseline | NAT | TS-NAT |
| Clean | 1.32 | 1.25 | **0.89** | 0.90 | 0.87 | **0.71** |
| 15 | 1.88 | 1.95 | **1.51** | 1.28 | 1.21 | **0.94** |
| 10 | 3.33 | 3.42 | **2.88** | 2.09 | 1.94 | **1.60** |
| 5 | 7.83 | 8.09 | **7.14** | 4.71 | 4.36 | **4.06** |
| 0 | 20.85 | 20.67 | **19.64** | 13.13 | 11.94 | **11.92** |
| Avg. | 7.04 | 7.08 | **6.41** | 4.42 | 4.06 | **3.85** |

  - WER (%) on Aurora-5 task with dropout (20%)

| SNR (DB) | Non-filtered | | | G.712 filtered | | |
|---|---|---|---|---|---|---|
| Method | Baseline | NAT | **TS-NAT** | Baseline | NAT | **TS-NAT** |
| Clean | 1.32 | 1.05 | **0.91** | 0.84 | 0.78 | **0.85** |
| 15 | 1.87 | 1.78 | **1.52** | 0.90 | 1.15 | **0.92** |
| 10 | 3.29 | 3.18 | **2.59** | 1.89 | 1.88 | **1.31** |
| 5 | 7.77 | 7.62 | **6.63** | 4.33 | 3.97 | **3.68** |
| 0 | 20.60 | 19.92 | **19.30** | 11.92 | 11.57 | **11.36** |
| Avg. | 6.97 | 6.71 | **6.19** | 3.98 | 3.87 | **3.62** |

## Conclusions

- We have proposed a DNN-based acoustic model for effective usage of multi-condition data and its noise estimate
  - Addresses the mapping from noisy speech and noise estimates to phonetic targets effectively by concatenating two DNNs
    - Clean feature reconstruction
    - Prediction of posterior probability over HMM states
  - Proposed technique outperforms NAT in word accuracy on Aurora-5