DNN-Based Feature Enhancement Using DOA-Constrained ICA for Robust Speech Recognition

Ho-Yong Lee, Ji-Won Cho, Minook Kim, and Hyung-Min Park

Department of Electronic Engineering, Sogang University, Seoul, Republic of Korea

hpark@sogang.ac.kr

Abstract

Recently, deep-neural-network(DNN)-based speech feature enhancement (FE) approaches have attracted much attention owing to their powerful modeling capabilities. However, DNN-based approaches are unable to achieve remarkable performance improvements for speech with severe distortion in the test environments different from training environments. We propose a DNN-based FE method where the DNN inputs include pre-enhanced spectral features computed from multi-channel input signals to reconstruct noise-robust features. The pre-enhanced spectral features are obtained by direction-of-arrival(DOA)-constrained independent component analysis (DCICA) followed by Bayesian FE using a hidden-Markov-model(HMM) prior, to exploit the capabilities of efficient online target speech extraction and efficient FE with prior information for robust ASR. In addition, noise spectral features computed from DCICA are included for further improvement. Therefore, the DNN is trained to reconstruct a clean spectral feature vector, from a sequence of corrupted input feature vectors in addition to the corresponding pre-enhanced and noise feature vectors. Experimental results demonstrate that the proposed method significantly improves recognition performance, even in mismatched noise conditions.

Introduction

Robust automatic speech recognition (ASR)
- The performance of most ASR systems is seriously degraded owing to differences between training and testing environments.
- Although many algorithms have been proposed to compensate for the mismatch under specific conditions, most of them frequently fail to attain high-recognition performances in real-world environments with various noises.

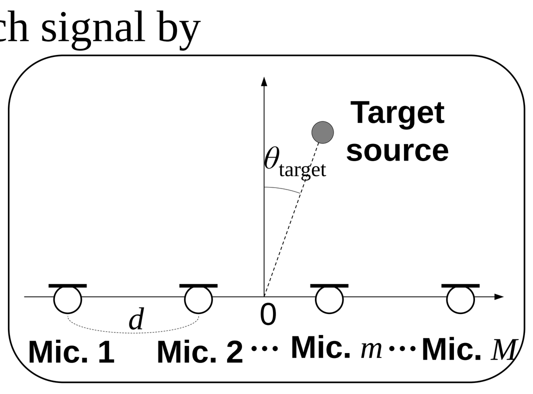
Deep learning
- Recently emerged as a breakthrough for acoustic modeling.
- Applied to speech enhancement or preprocessing for robust ASR.
  - Denoising autoencoder to reconstruct a clean speech signal from a noisy input.
- One common problem of DNN-based algorithms
  - Degraded in mismatched noise conditions.
  - Multicondition training including many different noise types in the training set.
  - Noise-aware training (NAT) including estimated noise information in DNN inputs.
  - DNN-based binary mask estimation in the time-frequency domain by training in a wide range of acoustic environments : extended to ratio mask estimation.
  - Various feature combinations based on mask estimation using multichannel inputs.

Proposed method
- DNN-based feature enhancement (FE) method using multichannel inputs for robust ASR.
- FE of logarithmic mel-frequency power spectral coefficients (LMPSCs) for efficiency.
- DNN is trained to reconstruct a clean-speech-feature vector, from a sequence of corrupted input feature vectors in addition to the corresponding preenhanced-speech- and estimated-noise-feature vectors.
  - Preenhanced spectral features by direction-of-arrival(DOA)-constrained independent component analysis (DCICA) followed by Bayesian FE based on a hidden-Markov-model(HMM) prior.
  - Noise spectral features computed from DCICA.

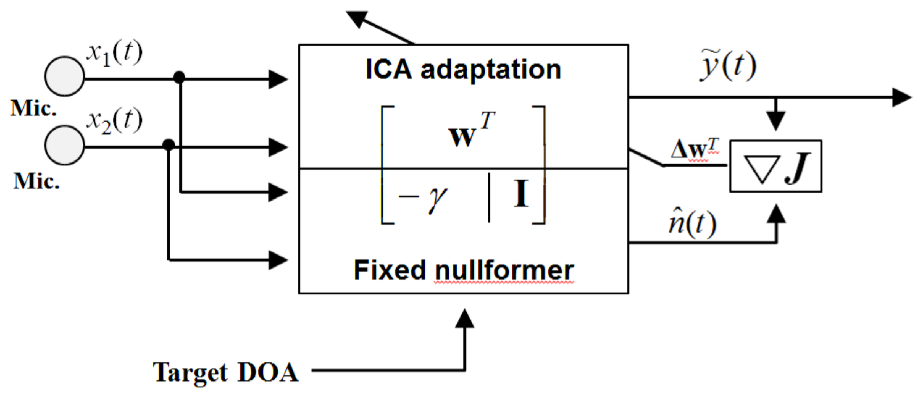
DCICA-FE of Corrupted Speech

DCICA
- Efficient online target speech extraction without any permutation problem.
- Dummy outputs : noise estimation by canceling a target speech signal by

$$U^m_{i,j} = X^m_{i,j} - \exp\left\{ j\omega_j \frac{d(m-1)\sin\theta_{\text{target}}}{c} \right\} X^1_{i,j}, \quad m = 2, \cdots, M.$$



- Target speech output estimated by minimizing the dependency between $Y_{i,j}$ and $U^m_{i,j}$.



$$\Delta \mathbf{w}_j \propto -\frac{\phi(Y_{i,j})}{\sqrt{\xi_{i,j}}} \left[\sum_{m=2}^{M} \exp\left\{ j\omega_j \frac{d(m-1)\sin\theta_{\text{target}}}{c} \right\} (U^m_{i,j})^*, \ -(U^2_{i,j})^*, \ \cdots, \ -(U^M_{i,j})^* \right].$$

Bayesian FE
- Bayesian inference to estimate clean features.
- Target speech output employed as noisy speech to be processed for further enhancement.
- Applying the kth band mel-scale filter on $|Y_{i,j}|^2$, the LMPSC

$$y_{i,k} = \log[\exp(s_{i,k}) + \exp(n_{i,k})].$$

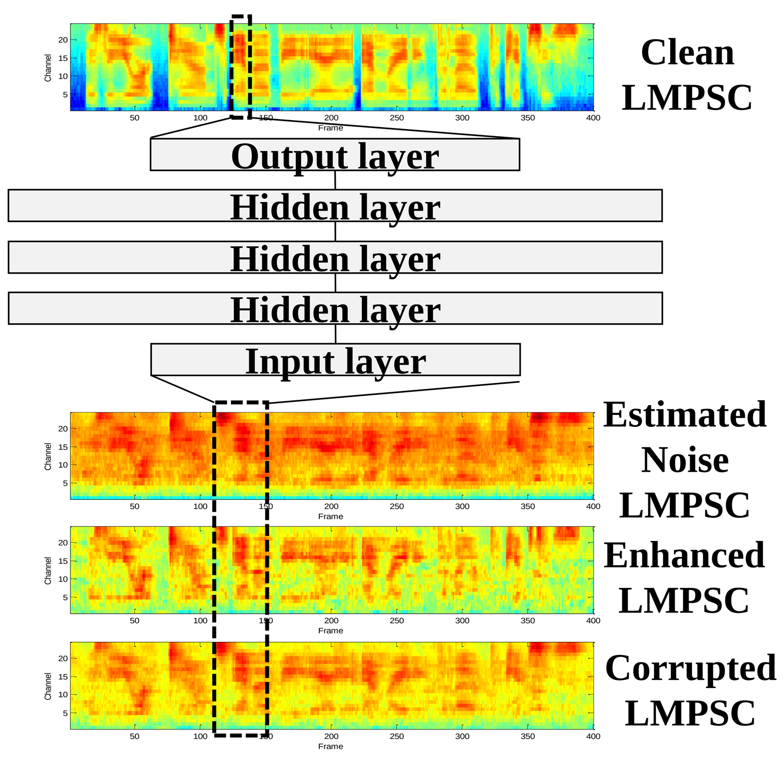
- Bayesian FE accomplished by the MMSE estimate

$$\hat{\mathbf{s}}_i = \arg\min_{\hat{\mathbf{s}}_i} E[(\mathbf{s}_i - \hat{\mathbf{s}}_i)^2 \mid \mathbf{y}_{1:i}] = E[\mathbf{s}_i \mid \mathbf{y}_{1:i}].$$

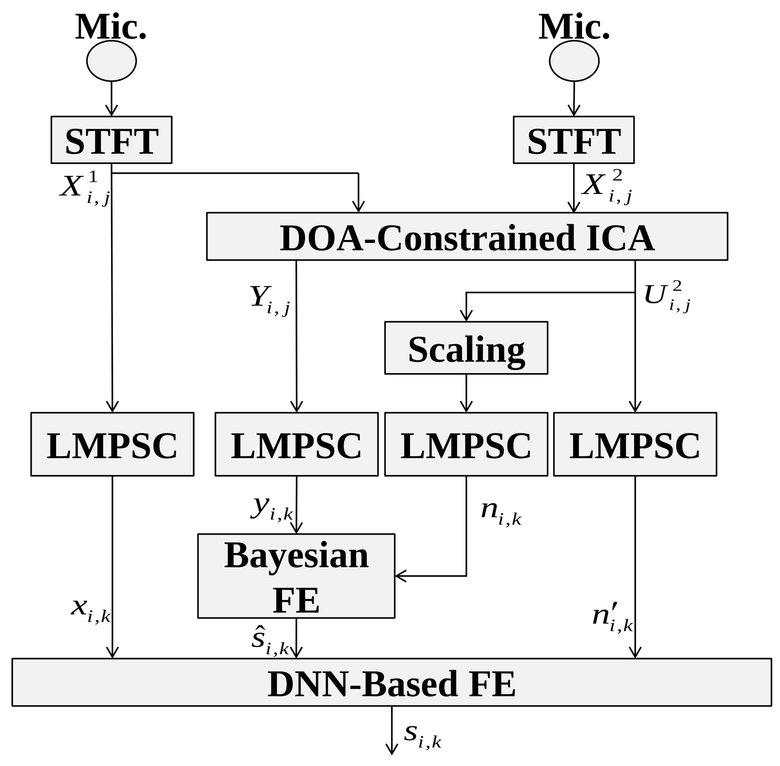
- Prior model
  - An LMPSC of the noise $n_{i,k}$ is assumed to be a Gaussian random process.
  - An LMPSC of clean speech $s_{i,k}$ is assumed to be described by an ergodic HMM with the single-Gaussian observation.

Spectral FE Based on DNN

DNN-based FE
- Recently used as a regression function for mapping noisy speech LMPSCs to clean ones.
- Highly useful because DNN can capture acoustic information along the time or frequency axis simultaneously by using a sequence of seven feature vectors of 24 LMPSCs.
- Tends to degrade in unseen noise environments even with multicondition training.
- Features enhanced by DCICA-FE may be helpful because DCICA-FE does not suffer from performance degradation due to unseen noise corruption.
- Noise spectral features computed from DCICA used as additional inputs to the DNN for further improvement.
- DNN
  - Three hidden layers with 1024 units per layer.
  - Activation functions: sigmoid for hidden units and linear functions for output units.



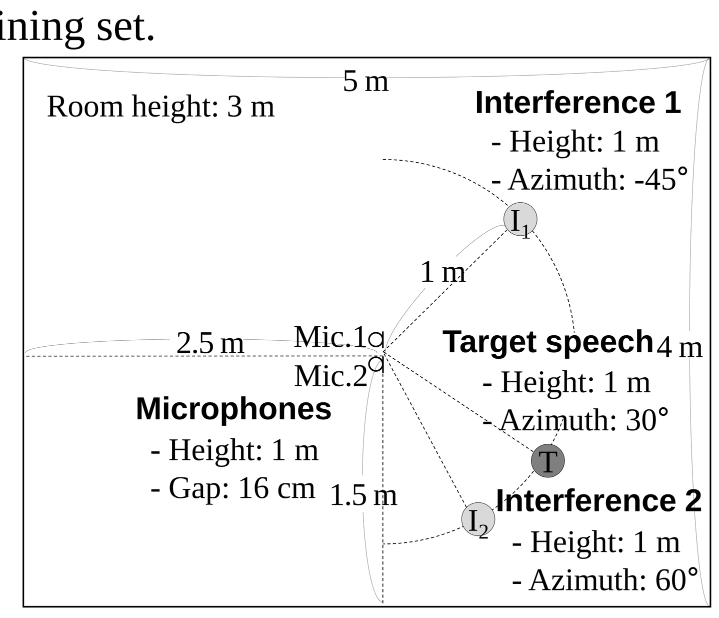
< Structure of the DNN training for FE >



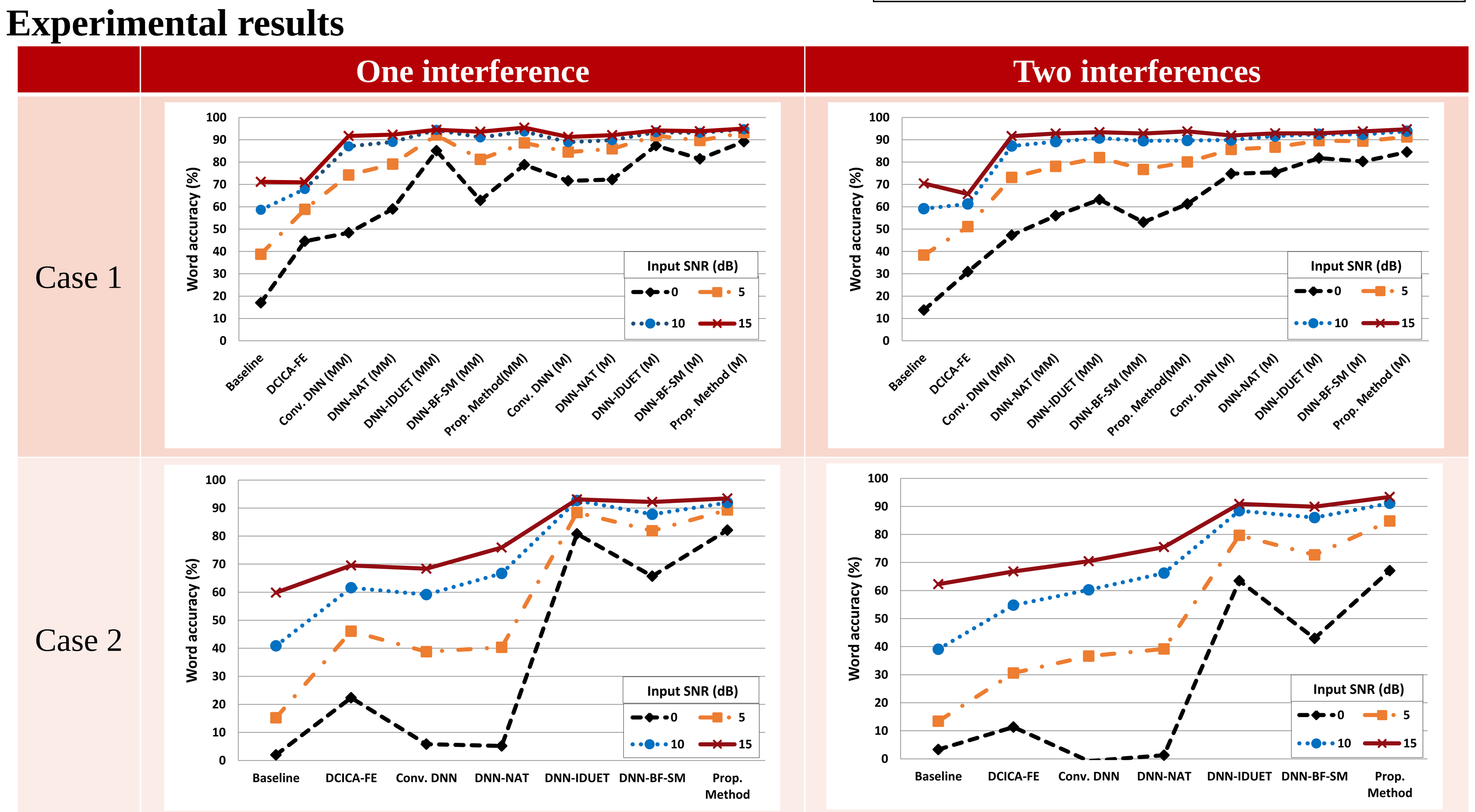
< Overall procedure of the proposed method >

Experimental Evaluation

Task and implementation
- DARPA resource management database (training set: 3990 sentences, test set: 300 sentences).
- Fully continuous HMM acoustic models and the 39th-order MFCCs.
- 128-state HMM prior model for Bayesian FE.
- Test utterance corrupted by (Case 1) babble noise or (Case 2) competing speech from the TIMIT database.
- Noisy speech samples to train DNNs for FE : the training set.
  - Case 1
    - Babble noise in matched noise condition.
    - Car, F16, factory, and operations room noises in mismatched noise condition.
  - Case 2
    - Randomly chosen from the resource management database.
- Two microphone signals simulated by the image method in a room with a RT_{60} of 0.3 s.



Experimental results



Conclusion

DNN-based FE method for robust ASR
- DNN inputs included LMPSCs preenhanced by DCICA-FE and noise LMPSCs.
- Significantly improved the recognition performance even in mismatched noise conditions.
- Evaluation on real data needs to be studied in the future.

Selected References

[1] J.-W. Cho and H.-M. Park, "Independent vector analysis followed by HMM-based feature enhancement for robust speech recognition," *Signal Process.*, vol. 120, pp. 200–208, 2016.

[2] M. Kim and H.-M. Park, "Efficient online target speech extraction using DOA-constrained independent component analysis of stereo data for robust speech recognition," *Signal Process.*, vol. 117, pp. 126–137, 2015.