Deep Neural Networks based Speech Separation

Xiangang Li, Hui Song

Outline

1 Introduction

2 DNNs based speech separation

3 Monaural separation algorithms

- 3.1 Speech enhancement
- 3.2 Speech dereverberation
- 3.3 Speaker separation

Outline

1 Introduction

2 DNNs based speech separation

3 Monaural separation algorithms

- 3.1 Speech enhancement
- 3.2 Speech dereverberation
- 3.3 Speaker separation

1 Introduction

深蓝学院
shenlanxueyuan.com

- Better speech signal quality in everyday environment
  - For better ASR performance
  - For better human auditory perception
- The introduction of deep learning to speech processing
  - Have dramatically accelerated progress and boosted performance
  - Always everywhere in speech processing
  - Optimized independently, or combined in the ASR modeling framework

1 Introduction

- Speech separation
  - Separating target speech from background interference
  - A typical signal processing problem
  - Supervised learning problem
    - Leaning the discriminative patterns of speech, speakers, and background noise
- Speech separation in this talk
  - Monaural & microphone array
  - Speech enhancement (speech-nonspeech separation)
  - Speaker separation (multi-talker separation)
  - Speech dereverberation

Outline

2 DNN based speech separation

深蓝学院 shenlanxueyuan.com

- Speech separation in this talk

Mixture signal at M mics

Target speech + interfering speech + noise + reverberation

$$y^{(m)} = s_d * h^{(m)} + n^{(m)}$$

→

Clean target speech s_d

Or reverb clean target speech $s_d * h^{(m)}$

- Typical solution

Training data → Feature extraction → Neural network → Training target

2.1 Training data

Training data → Feature extraction → Neural network → Training target

- How to get the clean and noisy speech pairs as training data
  - Simulation
  - Collect real-world reverberation and noise

2.2 Feature extraction

Training data → **Feature extraction** → Neural network → Training target

- Features as input and learning machines play complementary roles in supervised learning
  - PLP/MFCC/PNCC/PITCH/GFCC
  - …

	Factory	Babble	Engine	Cockpit	Vehicle	Tank	Average
MRCG	**63** (7)	**49** (13)	**77** (4)	**73** (4)	**80** (10)	**77** (6)	**70** (7)
GF	61 (7)	45 (15)	75 (4)	71 (3)	80 (10)	76 (6)	68 (8)
GFCC	61 (6)	46 (14)	73 (4)	70 (3)	78 (11)	74 (6)	67 (7)
DSCC	56 (7)	42 (14)	70 (5)	66 (3)	77 (11)	73 (6)	64 (8)
MFCC	57 (7)	43 (14)	69 (5)	67 (4)	77 (11)	72 (7)	64 (8)
PNCC	56 (6)	44 (14)	69 (5)	66 (4)	77 (11)	71 (7)	64 (8)
PLP	56 (6)	41 (12)	68 (5)	66 (4)	77 (11)	71 (7)	63 (8)
AC-MFCC	56 (6)	42 (14)	67 (5)	65 (4)	77 (11)	71 (7)	63 (8)
RAS-MFCC	57 (6)	41 (14)	68 (5)	66 (4)	76 (11)	71 (7)	63 (8)
GFB	57 (7)	41 (18)	67 (5)	66 (4)	75 (12)	70 (7)	63 (9)
ZCPA	55 (8)	40 (16)	68 (5)	65 (4)	75 (13)	70 (8)	62 (9)
SSF	54 (7)	39 (15)	67 (5)	60 (4)	76 (11)	69 (7)	61 (8)
RASTA-PLP	52 (6)	38 (15)	64 (5)	61 (4)	76 (12)	67 (7)	60 (8)
GFMC	48 (7)	35 (15)	61 (6)	60 (5)	67 (17)	59 (9)	55 (10)
PITCH	46 (3)	29 (22)	50 (5)	50 (2)	59 (16)	53 (7)	48 (9)
AMS	40 (6)	27 (9)	49 (5)	52 (4)	50 (31)	45 (11)	44 (11)
PAC-MFCC	17 (5)	11 (8)	30 (9)	29 (7)	40 (48)	21 (17)	25 (16)

2.3 Neural networks

Training data → Feature extraction → **Neural network** → Training target

- Various neural network architectures
  - Full-connected/CNN/RNN/LSTM
- Neural networks loss function
  - For classification: softmax & cross entropy
  - For regression: linear(or other activations) & mean square error (MSE)
  - Generative adversarial networks (GANs)
    - A generative model G: e.g. mapping from noisy speech to clean counterparts
    - A discriminative model D: e.g. discriminate between generated samples and target samples from training data

2.3 Neural networks

深蓝学院 shenlanxueyuan.com

Training data → Feature extraction → **Neural network** → Training target

- Various neural network architectures
  - Full-connected/CNN/RNN/LSTM
- Neural networks loss function
  - For classification: softmax & cross entropy
  - For regression: linear(or other activations) & mean square error (MSE)
  - Generative adversarial networks (GANs)
  - Multi-task Joint training
    - Speech separation & ASR

2.4 Training targets

Training data → Feature extraction → Neural network → Training target

- Masking-based targets
  - Learning to describe the time-frequency relationships of clean speech to background interference
- Mapping-based targets
  - Learning to estimate the spectral representations of clean speech from noisy speech

2.4 Training targets

深蓝学院 shenlanxueyuan.com

Training data → Feature extraction → Neural network → Training target

- Masking-based targets
  - Ideal Binary Mask (IBM)
  - G. Hu and D. L.Wang, “Speech segregation based on pitch tracking and amplitude modulation,” in Proc. IEEE Workshop Appl. Signal Process. Audio Acoust., 2001, pp. 79–82
  - G. Hu and D. L. Wang, “Monaural speech segregation based on pitch tracking and amplitude modulation,” IEEE Trans. Neural Netw., vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

$$IBM = \begin{cases} 1, & \text{if } SNR(t,f) > LC \\ 0, & \text{otherwise} \end{cases}$$

2.4 Training targets

Training data → Feature extraction → Neural network → Training target

- Masking-based targets
  - Ideal Binary Mask (IBM)
  - Target Binary Mask (TBM)
    - S. Gonzalez and M. Brookes, “Mask-based enhancement for very low quality speech,” in Proc. Int. Conf. Acoust., Speech Signal Process., 2014, pp. 7029–7033.

2.4 Training targets

深蓝学院
shenlanxueyuan.com

Training data → Feature extraction → Neural network → **Training target**

- Masking-based targets
  - Ideal Binary Mask (IBM)
  - Target Binary Mask (TBM)
  - Ideal Ratio Mask (IRM)
    - A soft version of IBM

$$IRM = \left(\frac{S(t+f)^2}{S(t+f)^2 + N(t+f)^2}\right)^{\beta}$$

2.4 Training targets

Training data → Feature extraction → Neural network → **Training target**

- Masking-based targets
  - Ideal Binary Mask (IBM)
  - Target Binary Mask (TBM)
  - Ideal Ratio Mask (IRM)
    - A soft version of IBM

$$IRM = \left(\frac{S(t+f)^2}{S(t+f)^2 + N(t+f)^2}\right)^{\beta}$$

- Estimate the mask
  - Ideal Ratio Mask

$$IRM = \left(\frac{S(t+f)^2}{S(t+f)^2 + N(t+f)^2}\right)^{\beta} \in [0,1]$$

  - Minimize

$$Loss = MSE\left(\widehat{IRM}, IRM\right)$$

  - Recover reverb clean

$$S^2(t,f) = \widehat{IRM}(|S^2(t,f)| + |N^2(t,f)|)$$

2.4 Training targets

深蓝学院 shenlanxueyuan.com

Training data → Feature extraction → Neural network → **Training target**

- Masking-based targets
  - Ideal Binary Mask (IBM)
  - Target Binary Mask (TBM)
  - Ideal Ratio Mask (IRM)
  - Spectral Magnitude Mask (SMM)

$$SMM(t,f) = \frac{|S(t,f)|}{|Y(t,f)|}$$

- where $|S(t,f)|$ and $|Y(t,f)|$ represent spectral magnitudes of clean speech and noisy speech, respectively.

2.4 Training targets

深蓝学院
shenlanxueyuan.com

Training data → Feature extraction → Neural network → Training target

- Masking-based targets
  - Ideal Binary Mask (IdBM)
  - Target Binary Mask (TBM)
  - Ideal Ratio Mask (IRM)
  - Spectral Magnitude Mask (SMM)
  - Phase-Sensitive Mask (PSM)

$$SMM(t,f) = \frac{|S(t,f)|}{|Y(t,f)|} cos\theta$$

- where θ denotes the difference of the clean speech phase and the noisy speech phase within the T-F unit

2.4 Training targets

深蓝学院 shenlanxueyuan.com

Training data → Feature extraction → Neural network → **Training target**

- Masking-based targets
  - Ideal Binary Mask (IdBM)
  - Target Binary Mask (TBM)
  - Ideal Ratio Mask (IRM)
  - Spectral Magnitude Mask (SMM)
  - Phase-Sensitive Mask (PSM)
  - Complex Ideal Ratio Mask (cIRM)
    - An ideal mask in complex domain

2.4 Training targets

Training data → Feature extraction → Neural network → Training target

- Masking-based targets

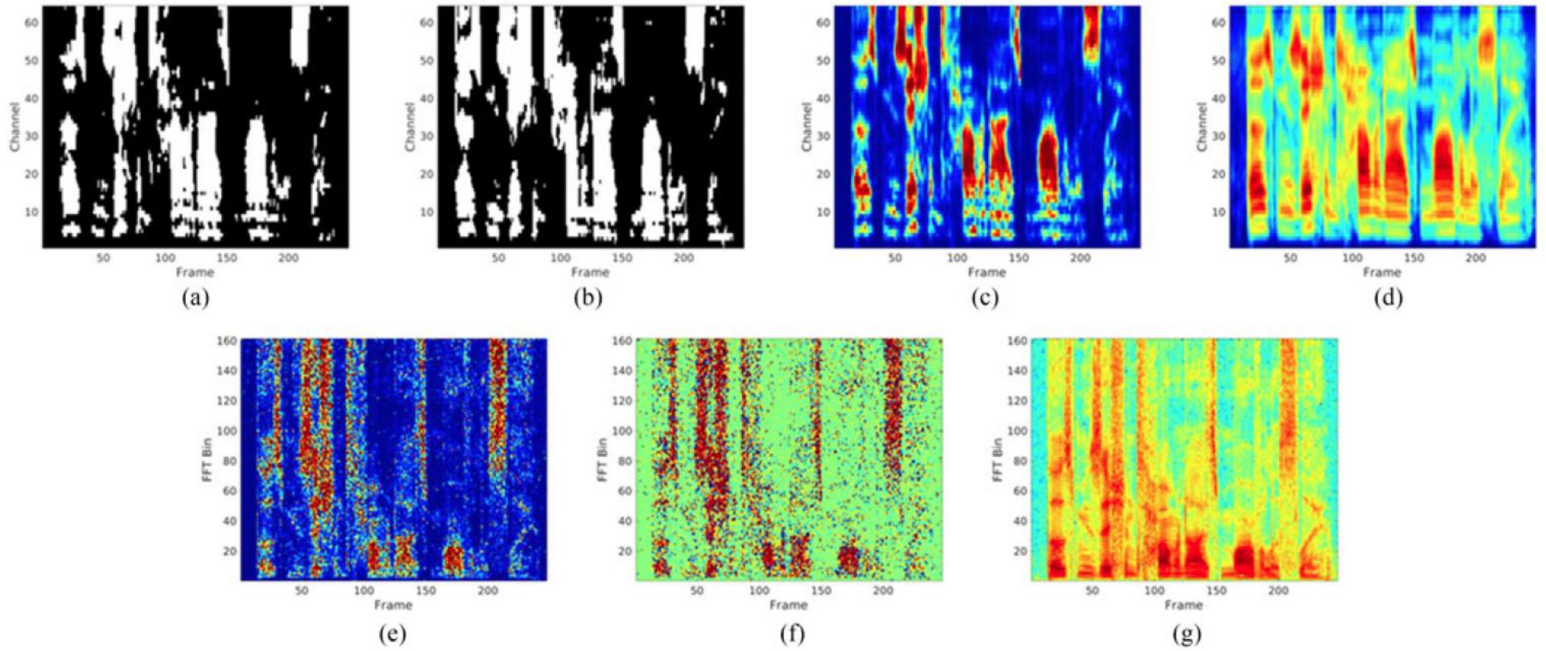


Fig. 2. Illustration of various training targets for a TIMIT utterance mixed with a factory noise at −5 dB SNR. (a) IBM. (b) TBM. (c) IRM. (d) GF-TPS. (e) SMM. (f) PSM. (g) TMS.

D.L. Wang, and J. Chen. Supervised Speech Separation Based on Deep Learning: An Overview. IEEE/ACM Trans. Audio Speech Lang. Process. vol. 26, no. 10, pp. 1702-1726, Oct. 2018

2.4 Training targets

Training data → Feature extraction → Neural network → Training target

- Masking-based targets
- Mapping-based targets
  - Target Magnitude Spectrum (TMS)
    - supervised learning aims to estimate the magnitude spectrogram of clean speech from that of noisy speech
    - Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Trans. Audio Speech Lang. Process., vol. 23, no. 1, pp. 7–19, Jan. 2015.

2.4 Training targets

深蓝学院 shenlanxueyuan.com

Training data → Feature extraction → Neural network → **Training target**

- Masking-based targets
- Mapping-based targets
  - Target Magnitude Spectrum (TMS)
  - Gammatone Frequency Target Power Spectrum (GT-TPS)
    - Y. Wang, A. Narayanan, and D. L. Wang, “On training targets for supervised speech separation,” IEEE/ACM Trans. Audio Speech Lang. Process., vol. 22, no. 12, pp. 1849–1858, Dec. 2014
    - Unlike the TMS defined on a spectrogram, this target is defined on a cochleagram based on a gammatone filterbank
  - …

2.4 Training targets

Training data → Feature extraction → Neural network → Training target

- Masking-based targets
- Mapping-based targets
  - Target Magnitude Spectrum (TMS)
  - Gammatone Frequency Target Power Spectrum (GT-TPS)
  - Signal Approximation

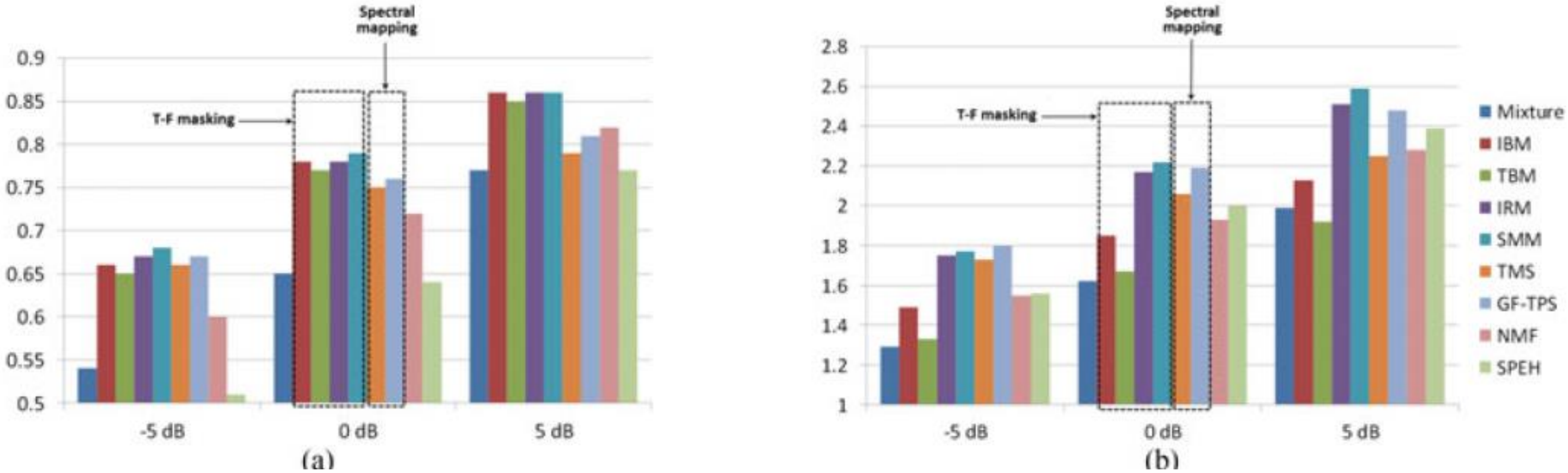
$$SA(f,t) = [RM(t,f)|Y(t,f)| - |S(t,f)|]^2$$

    - Spectral Magnitude Mask + MSE

2.4 Training targets

Training data → Feature extraction → Neural network → **Training target**

- Comparison of training targets



D.L. Wang, and J. Chen. Supervised Speech Separation Based on Deep Learning: An Overview. IEEE/ACM Trans. Audio Speech Lang. Process. vol. 26, no. 10, pp. 1702-1726, Oct. 2018

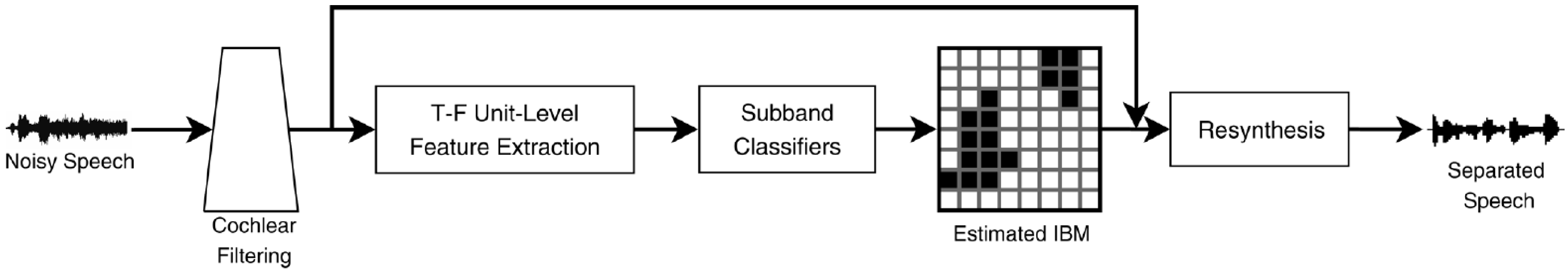
Outline

3 Monaural separation algorithms

- Monaural Speech separation
  - Speech enhancement (speech-nonspeech separation)
  - Speaker separation (multi-talker separation)
  - Speech dereverberation

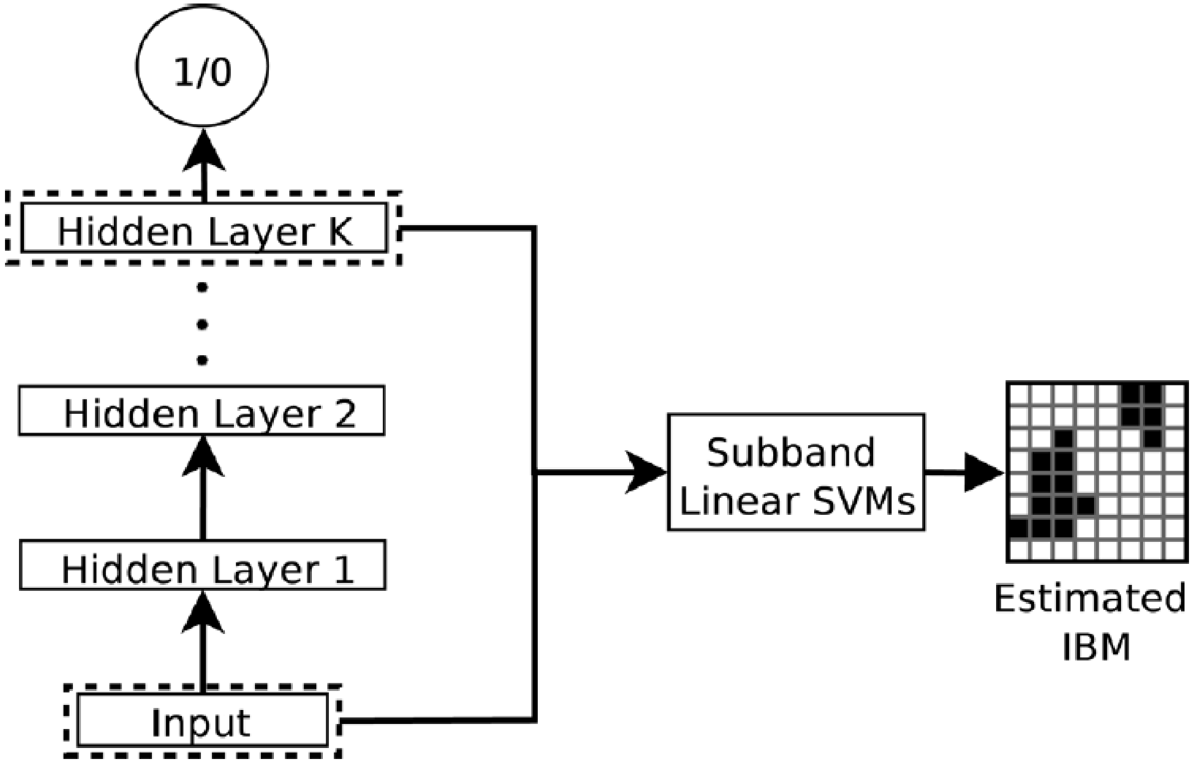
3.1 DNN based speech enhancement

- Masking-based targets
  - Y. Wang and D.L. Wang, “Towards scaling up classification-based speech separation,” IEEE Trans. Audio Speech Lang. Process., vol. 21, no. 7, pp. 1381–1390, Jul. 2013.



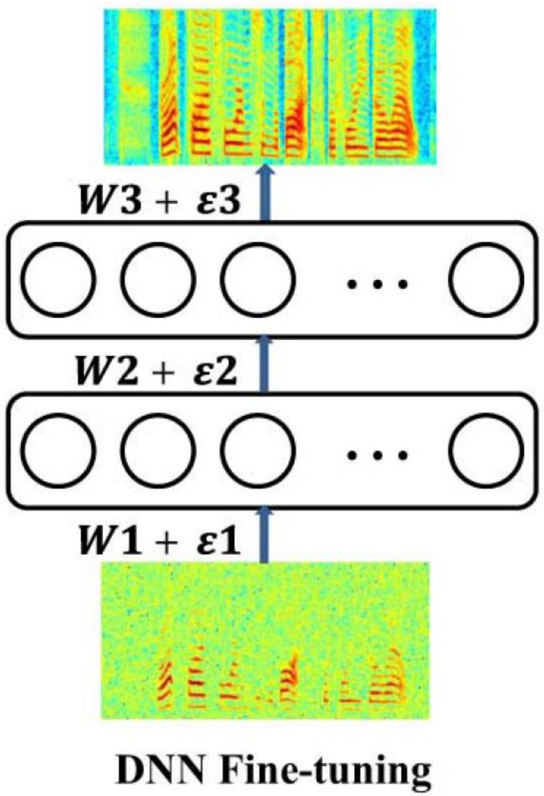
3.1 DNN based speech enhancement

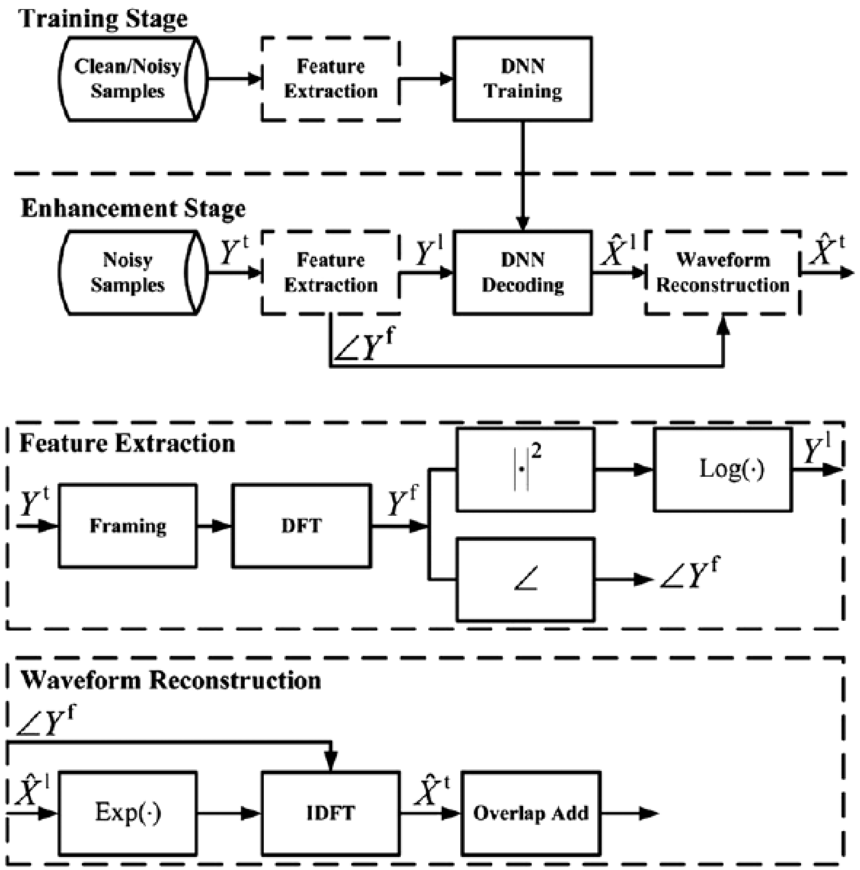
- Masking-based targets
  - Y. Wang and D.L. Wang, "Towards scaling up classification-based speech separation," IEEE Trans. Audio Speech Lang. Process., vol. 21, no. 7, pp. 1381–1390, Jul. 2013.



3.1 DNN based speech enhancement

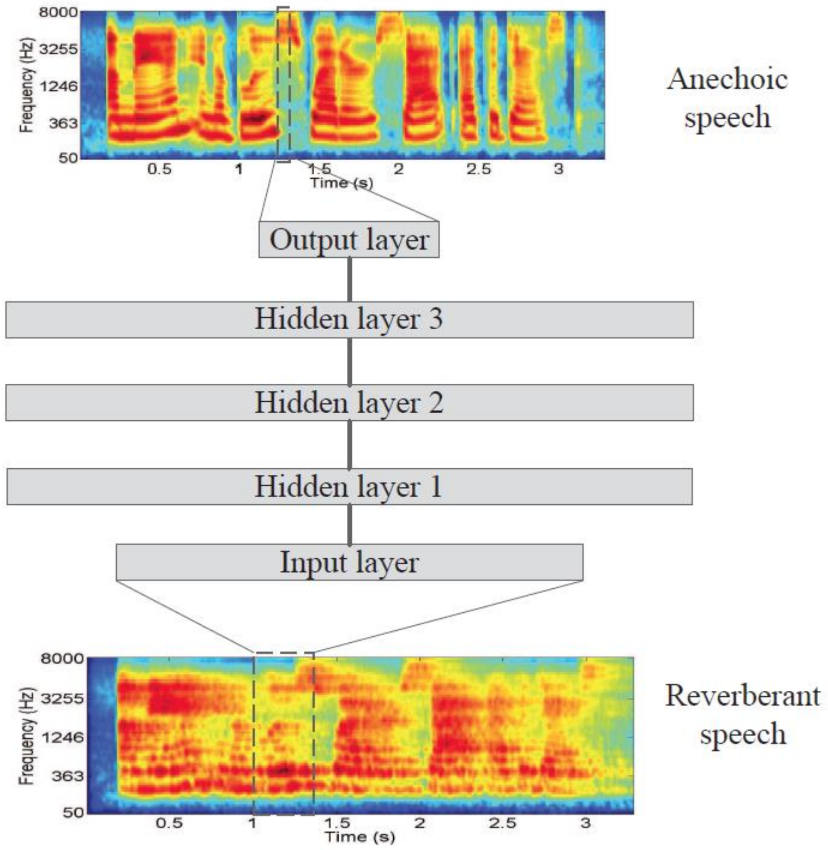
- Mapping-based targets
  - Y. Xu, J.Du, L.-R.Dai, and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks,” IEEE Signal Process. Lett., vol. 21, no. 1, pp. 65–68, Jan. 2014.





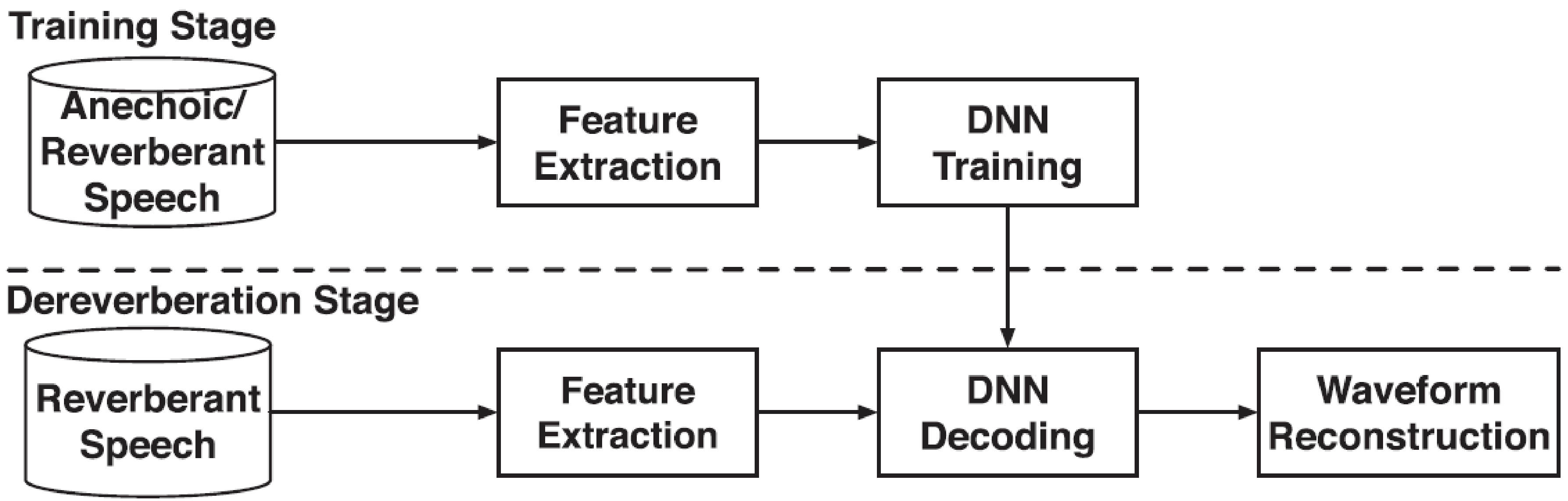
3.2 DNN based speech dereverberation

- Mapping-based targets
  - K. Han, Y. Wang, and D. L.Wang, “Learning spectral mapping for speech dereverberation,” in Proc. Int. Conf. Acoust., Speech Signal Process., 2014, pp. 4661–4665.



3.2 DNN based speech dereverberation

- Mapping-based targets
  - K. Han, Y. Wang, and D. L.Wang, "Learning spectral mapping for speech dereverberation," in Proc. Int. Conf. Acoust., Speech Signal Process., 2014, pp. 4661–4665.
  - B. Wu, K. Li, M. Yang, and C.-H. Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 25, no. 1, pp. 102–111, Jan. 2017.



3.3 DNN based speaker separation

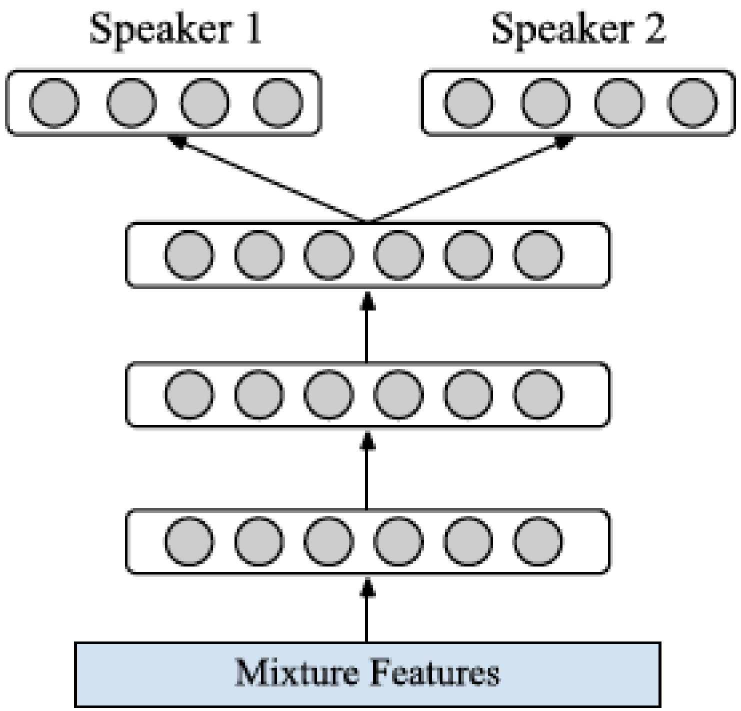
- Goal
  - Extract multiple speech signals, one for each speaker, from a mixture containing two or more voices
- Speaker separation
  - Speaker dependent: the underlying speakers are not allowed to change from training to testing
  - Target speaker dependent: interfering speakers are allowed to change, but the target speaker is fixed
  - Speaker independent: none of the speakers are required to be the same between the training and testing

3.3 DNN based speaker separation

- Speaker dependent
  - P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Deep learning for monaural speech separation,” in Proc. Int. Conf. Acoust., Speech Signal Process., 2014, pp. 1581–1585.

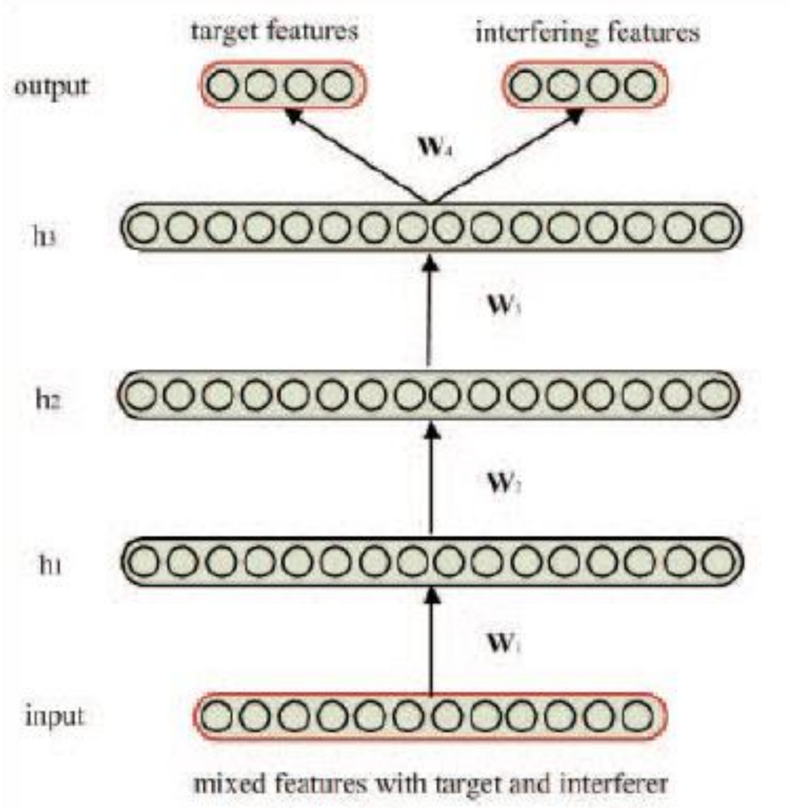
$$\tilde{S}_1(t) = \frac{\left|\hat{S}_1(t)\right|}{\left|\hat{S}_1(t)\right| + \left|\hat{S}_2(t)\right|} \odot \boldsymbol{Y}(t)$$

$$\tilde{S}_2(t) = \frac{\left|\hat{S}_2(t)\right|}{\left|\hat{S}_1(t)\right| + \left|\hat{S}_2(t)\right|} \odot \boldsymbol{Y}(t)$$



3.3 DNN based speaker separation

- Target speaker dependent
  - Y. Tu, J. Du, Y. Xu, L.-R. Dai, and C.-H. Lee, “Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers,” in Proc. 9th Int. Symp. Chinese Spoken Lang. Process., 2014, pp. 250–254.

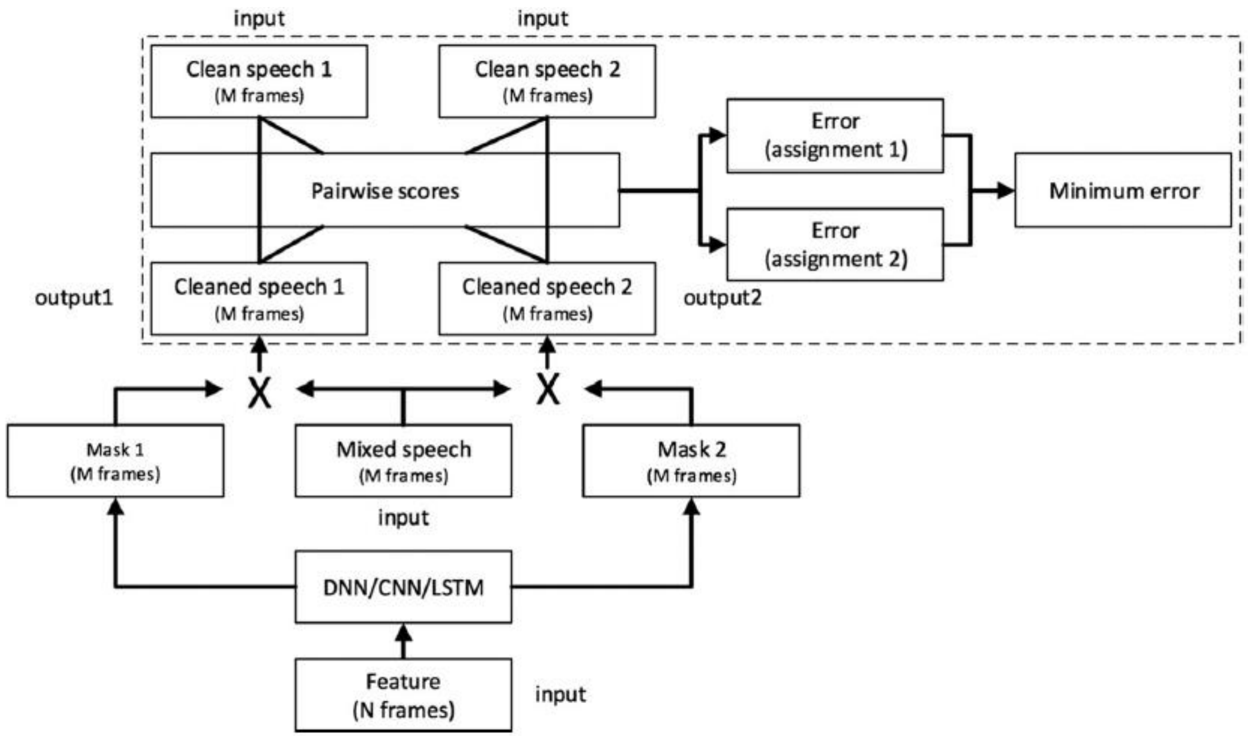


3.3 DNN based speaker separation

- Speaker independent
  - J. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in Proc. Int. Conf. Acoust., Speech Signal Process., 2016, pp. 31–35.
    - Treated as unsupervised clustering where T-F units are clustered into distinct classes dominated by individual speakers

3.3 DNN based speaker separation

- Speaker independent
  - J. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in Proc. Int. Conf. Acoust., Speech Signal Process., 2016, pp. 31–35.
  - D. Yu, M. Kolbak, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2017, pp. 241–245.



3.3 DNN based speaker separation

深蓝学院 shenlanxueyuan.com

- Speaker independent
  - J. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in Proc. Int. Conf. Acoust., Speech Signal Process., 2016, pp. 31–35.
  - D. Yu, M. Kolbak, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2017, pp. 241–245.
  - Z. Chen, Y. Luo, and N. Mesgarani, “Deep attractor network for singlemicrophone speaker separation,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2017, pp. 246–250.
  - Yi Luo, Nima Mesgarani. “Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation”. arXiv 1809.07454.

Thanks for your Attention