Speech Separation

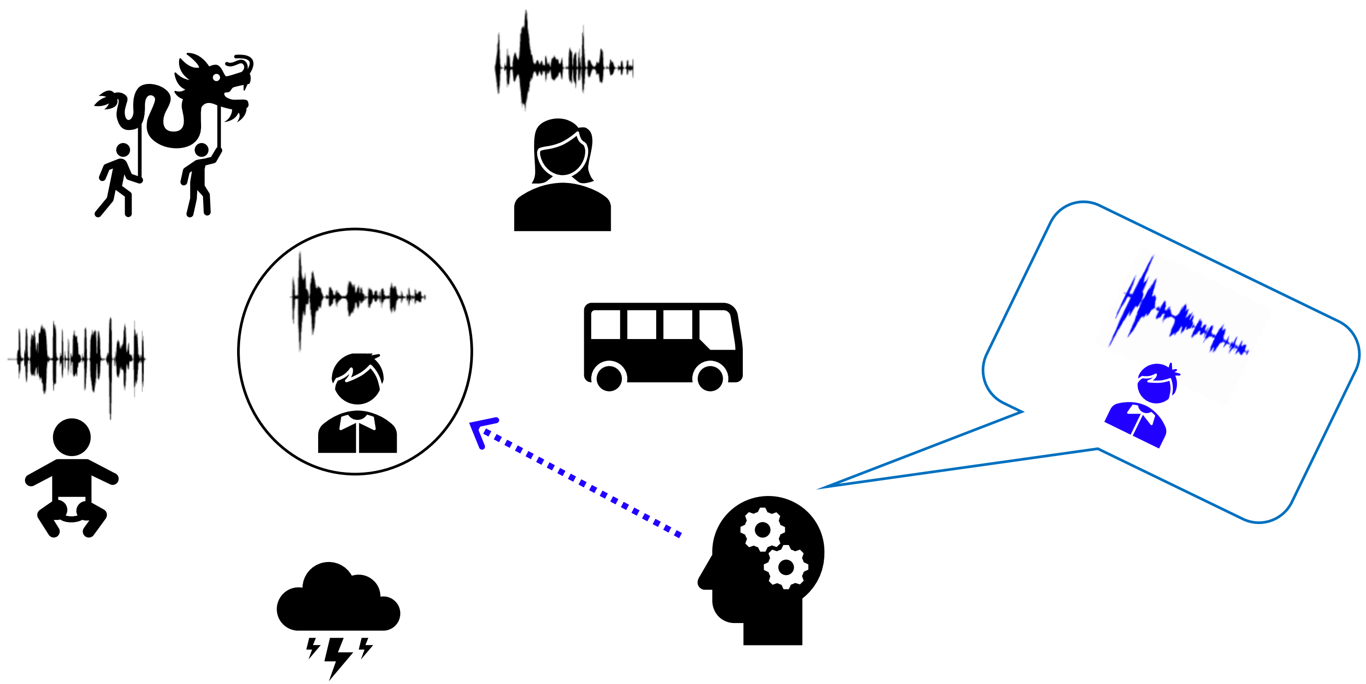
李宏毅

HUNG-YI LEE

Some slides are from 楊靖平

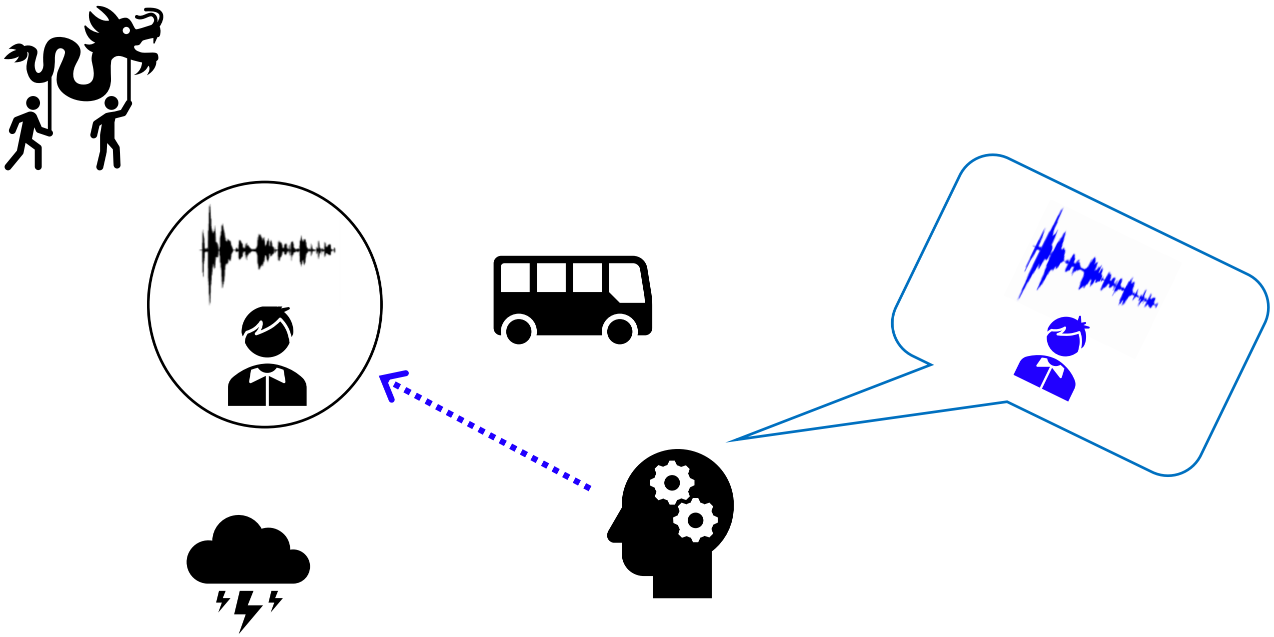
Speech Separation

- Humans can focus on the voice produced by a single speaker in a crowded and noisy environments.



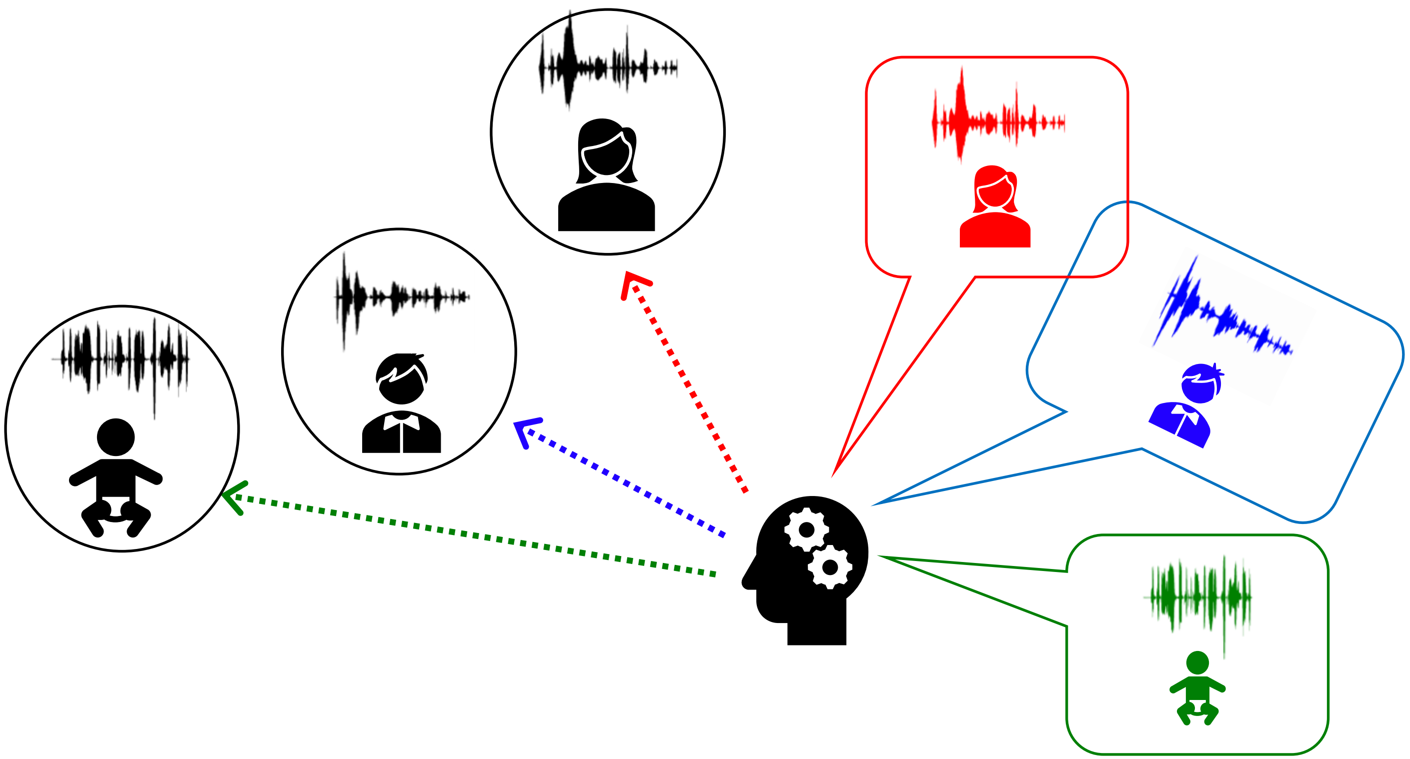
Speech Separation

- **Speech Enhancement**: speech-nonspeech separation



Speech Separation

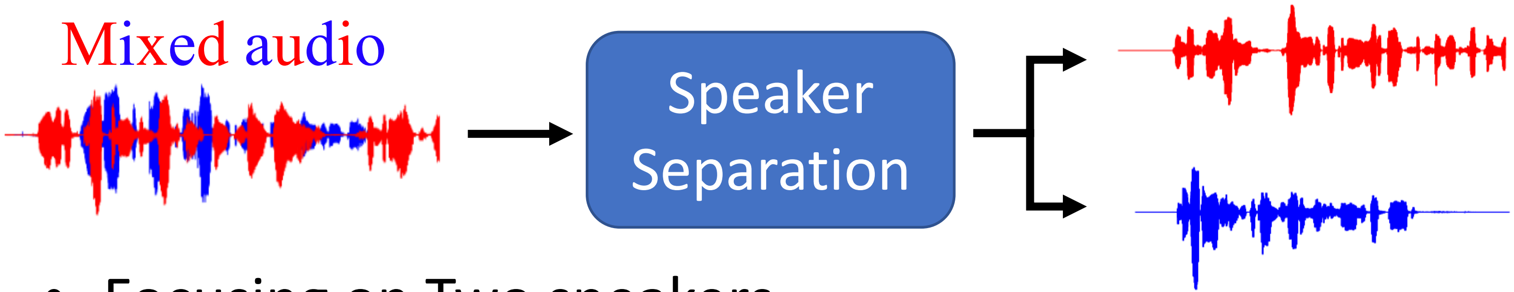
- **Speaker Separation**: multi-speaker talking



Speaker Separation

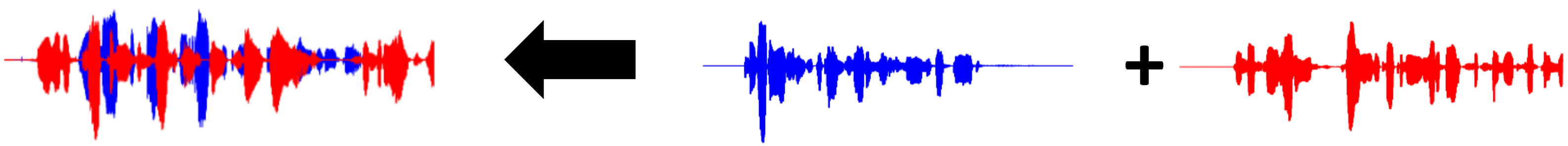
Input and output have the same length

Seq2seq is not needed



- Focusing on Two speakers
- Focusing on Single microphone
- **Speaker independent**: training and testing speakers are completely different

Training Data:



It is easy to generate training data.

Evaluation

X is speech signal (vector) here

- Signal-to-noise ratio (SNR)

$$SNR = 10\log_{10}\frac{\|\hat{X}\|^2}{\|E\|^2}$$

output of model

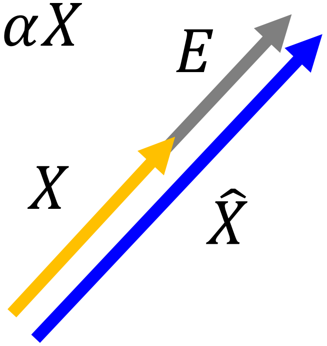
X

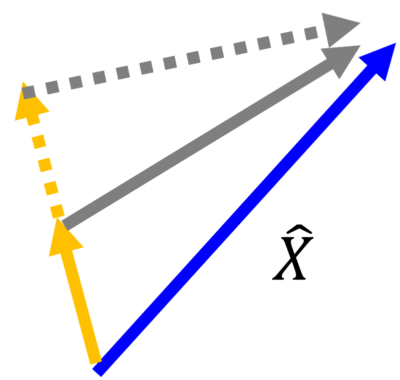
$E = \hat{X} - X$

ground truth

$\hat{X}$

$X = \alpha\hat{X}$



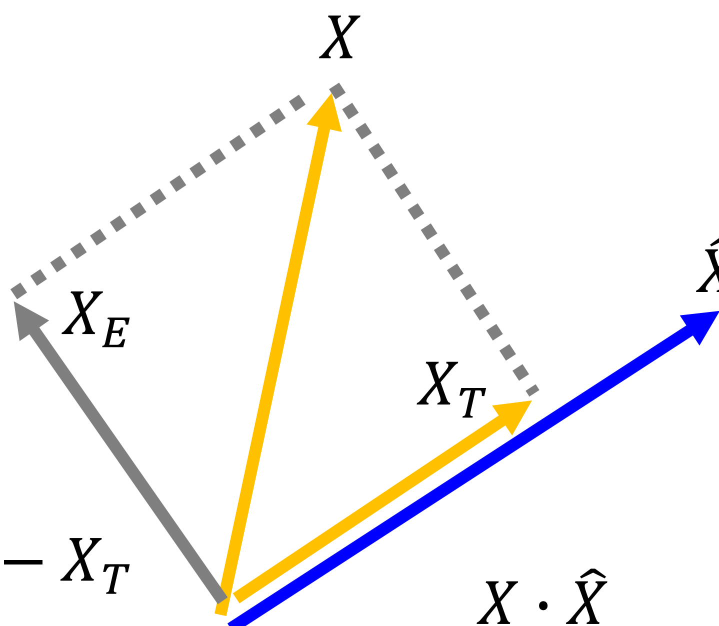


Simply larger the output can increase SNR?

Evaluation

X is speech signal (vector) here

- Scale invariant signal-to-distortion ration (SI-SDR)



$$X_E = X - X_T$$

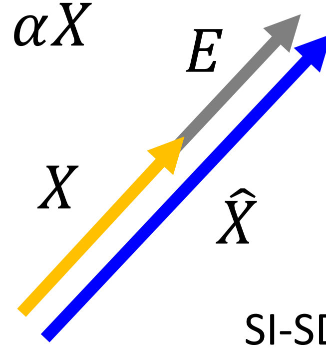
$$X_T = \frac{X \cdot \hat{X}}{\|\hat{X}\|^2} X$$

(請參見大學線代)

$$SISDR = 10 \log_{10} \frac{\|X_T\|^2}{\|X_E\|^2}$$

$$= \infty$$

$$X = \alpha \hat{X}$$

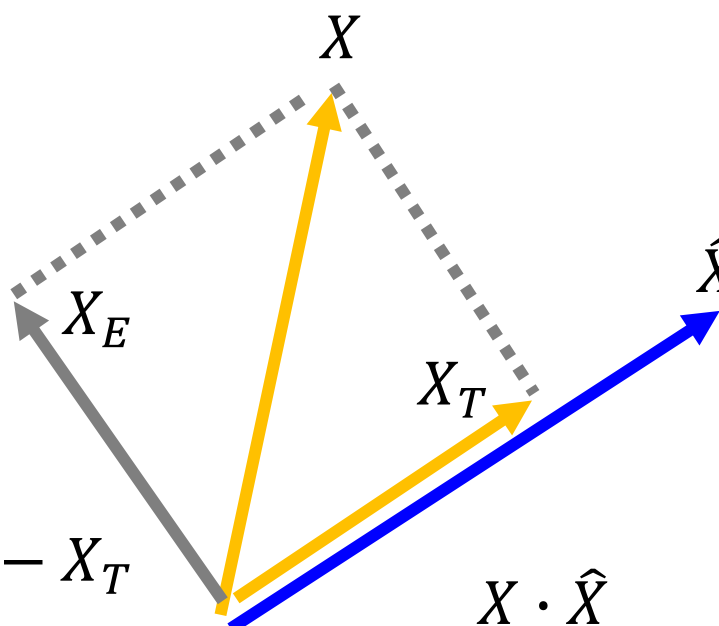


SI-SDR

Evaluation

X is speech signal (vector) here

- Scale invariant signal-to-distortion ration (SI-SDR)



$$X_E = X - X_T$$

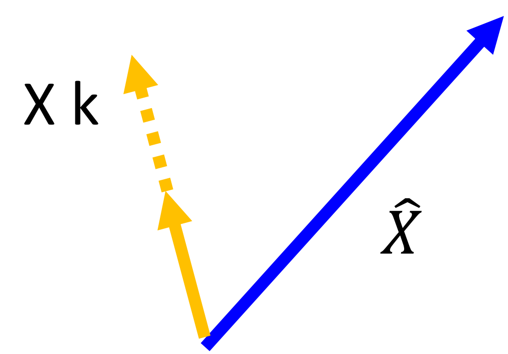
$$X_T = \frac{X \cdot \hat{X}}{\|\hat{X}\|^2} X$$

(請參見大學線代)

$$SISDR = 10 \log_{10} \frac{\|X_T\|^2}{\|X_E\|^2}$$

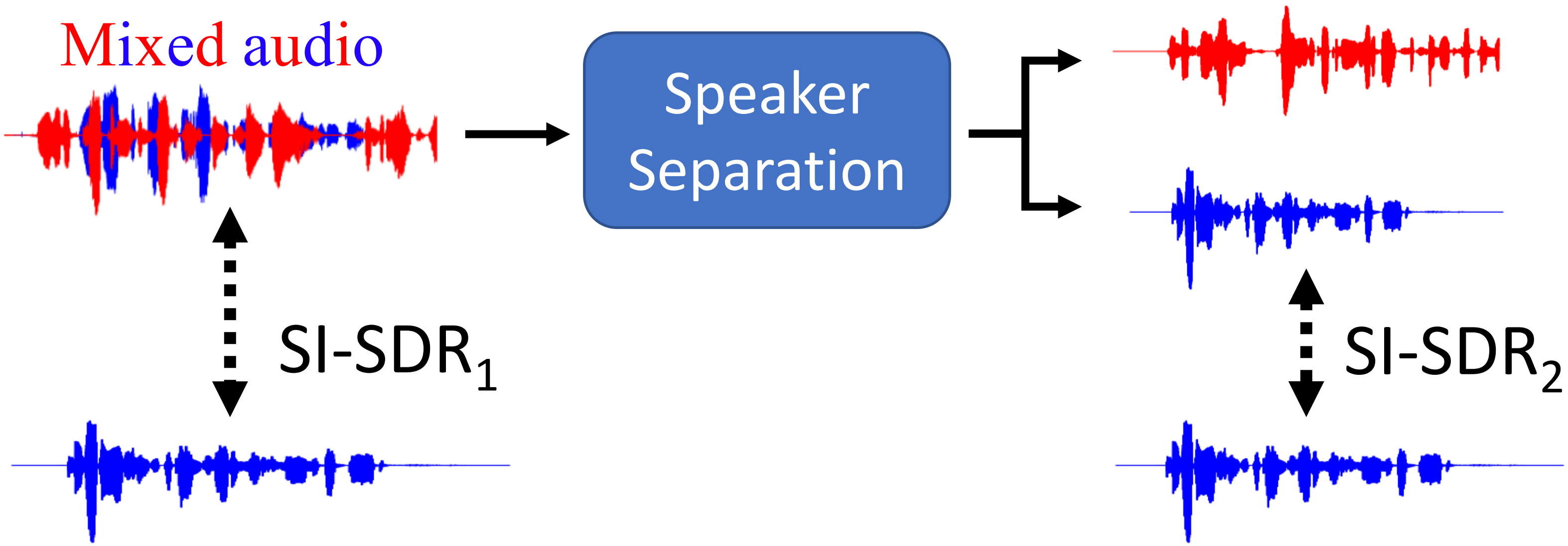
$$= 10 \log_{10} \frac{\|kX_T\|^2}{\|kX_E\|^2}$$

the same



Evaluation

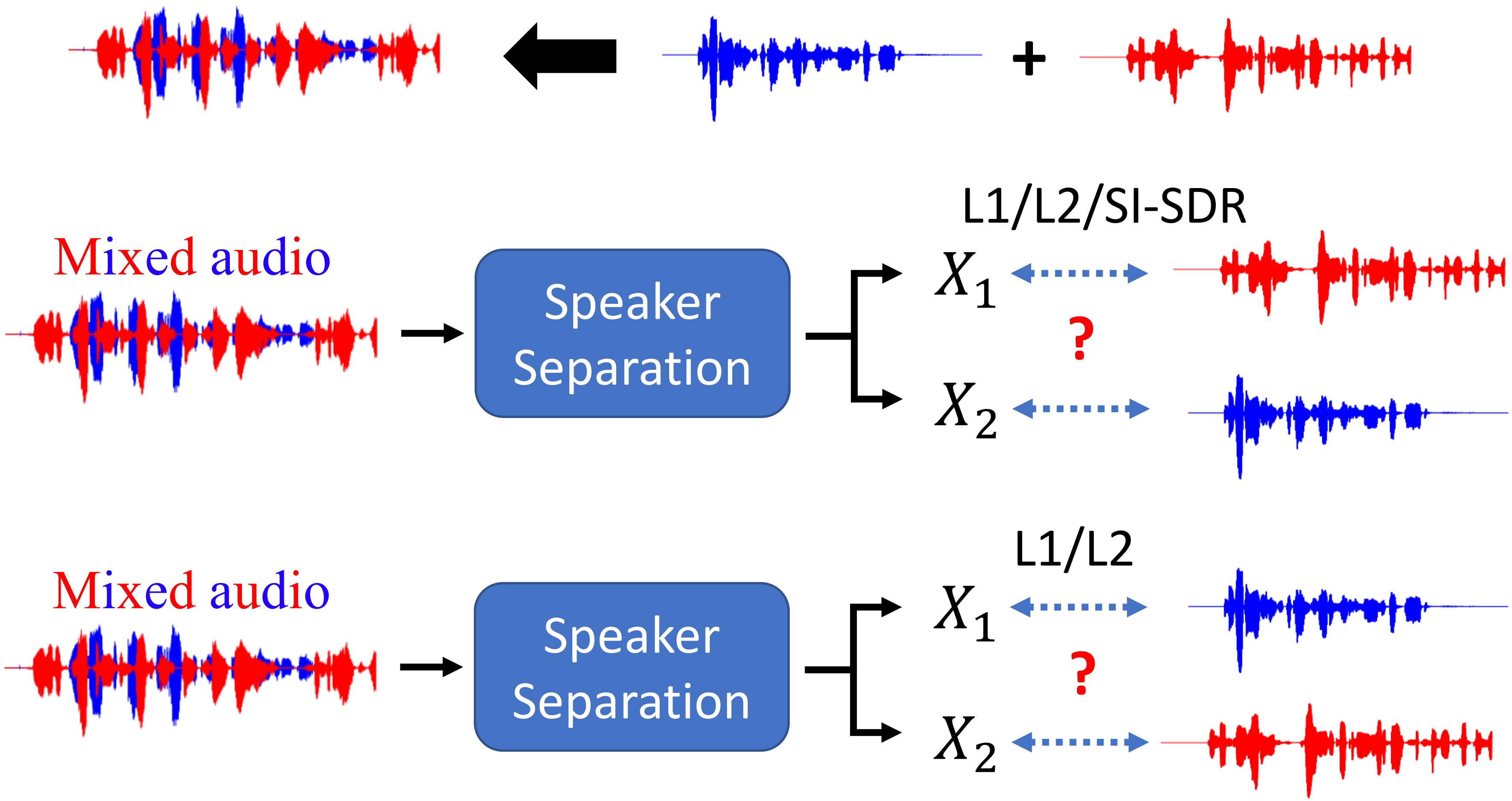
X is speech signal (vector) here



SI-SDR improvement: $\text{SI-SDR}_i = \text{SI-SDR}_2 - \text{SI-SDR}_2$

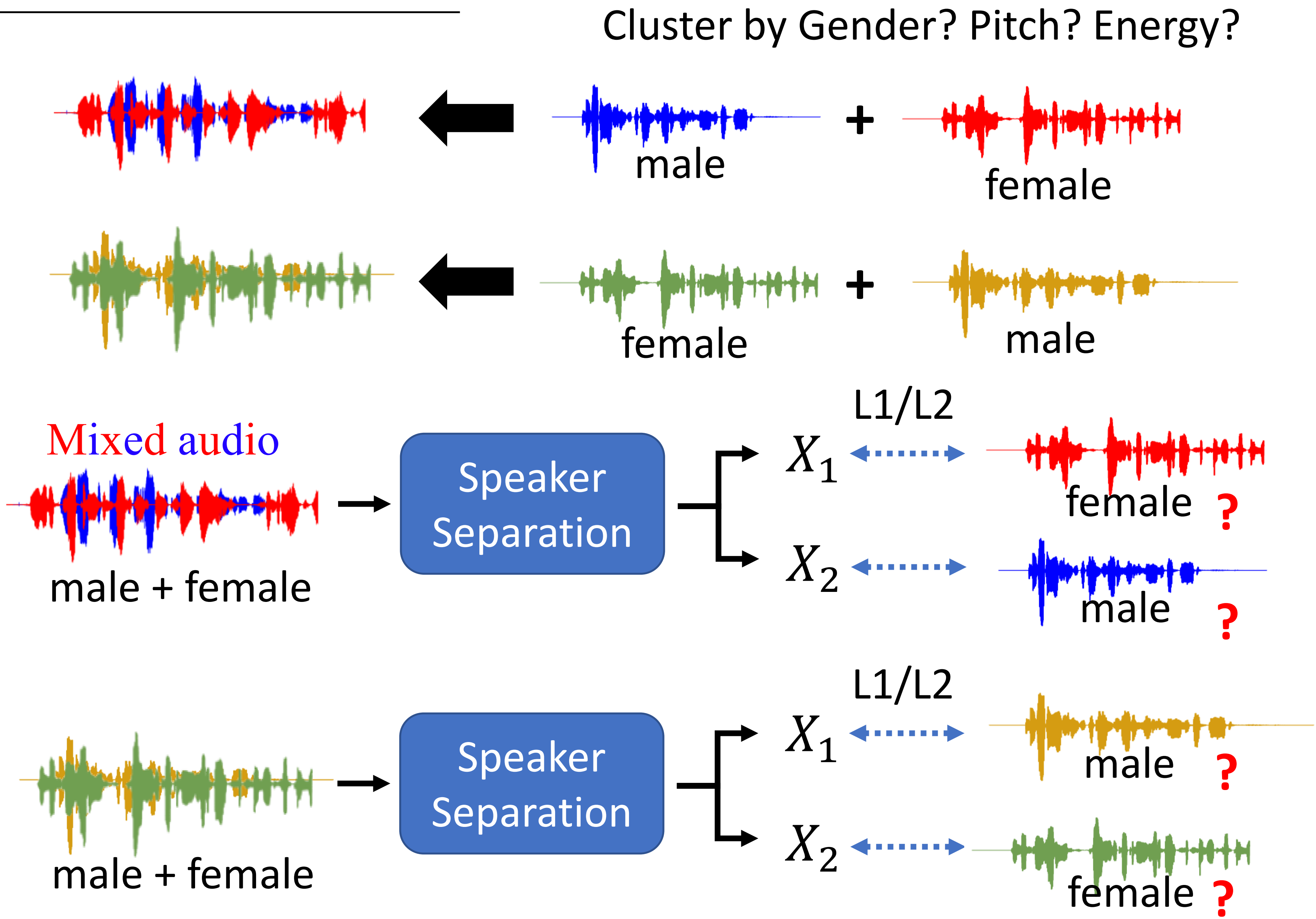
- Perceptual evaluation of speech quality (**PESQ**) was designed to evaluate the **quality**, and the score ranges from -0.5 to 4.5.
- Short-time objective intelligibility **(STOI)** was designed to compute **intelligibility**, and the score ranges from 0 to 1.

Permutation Issue



To achieve speaker independent training, the training data contain many different speakers.

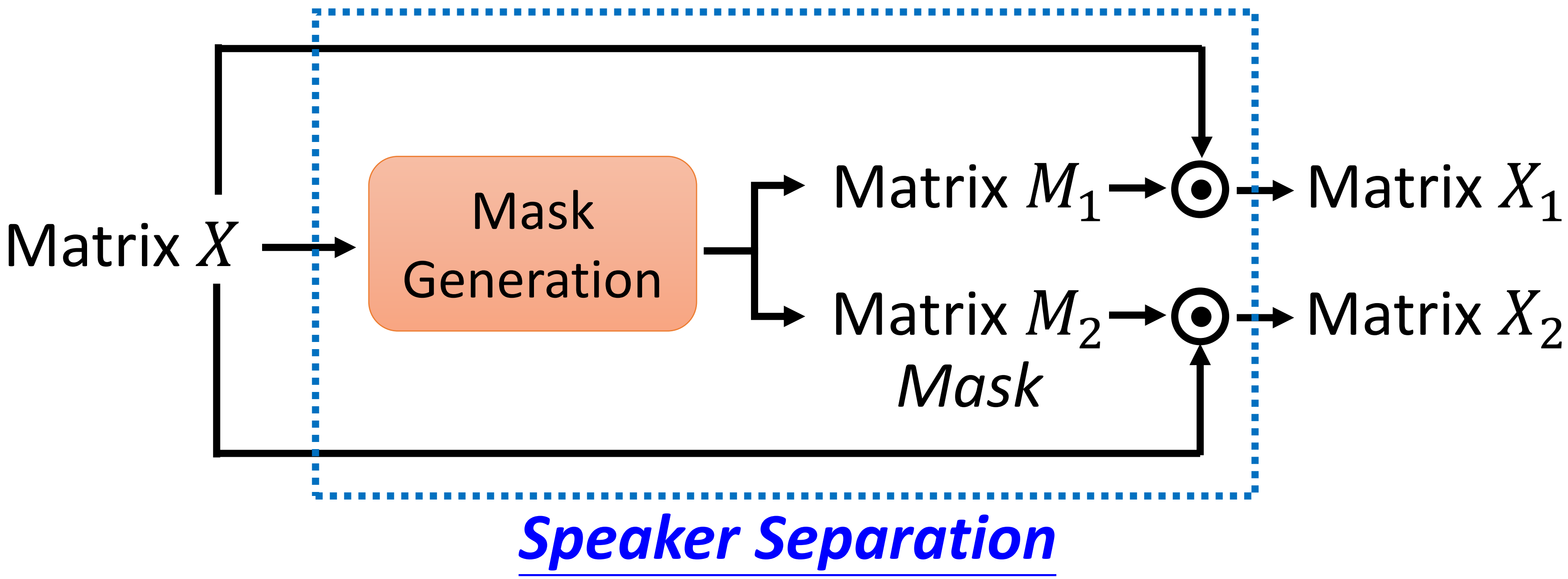
Permutation Issue



Deep Clustering

Masking

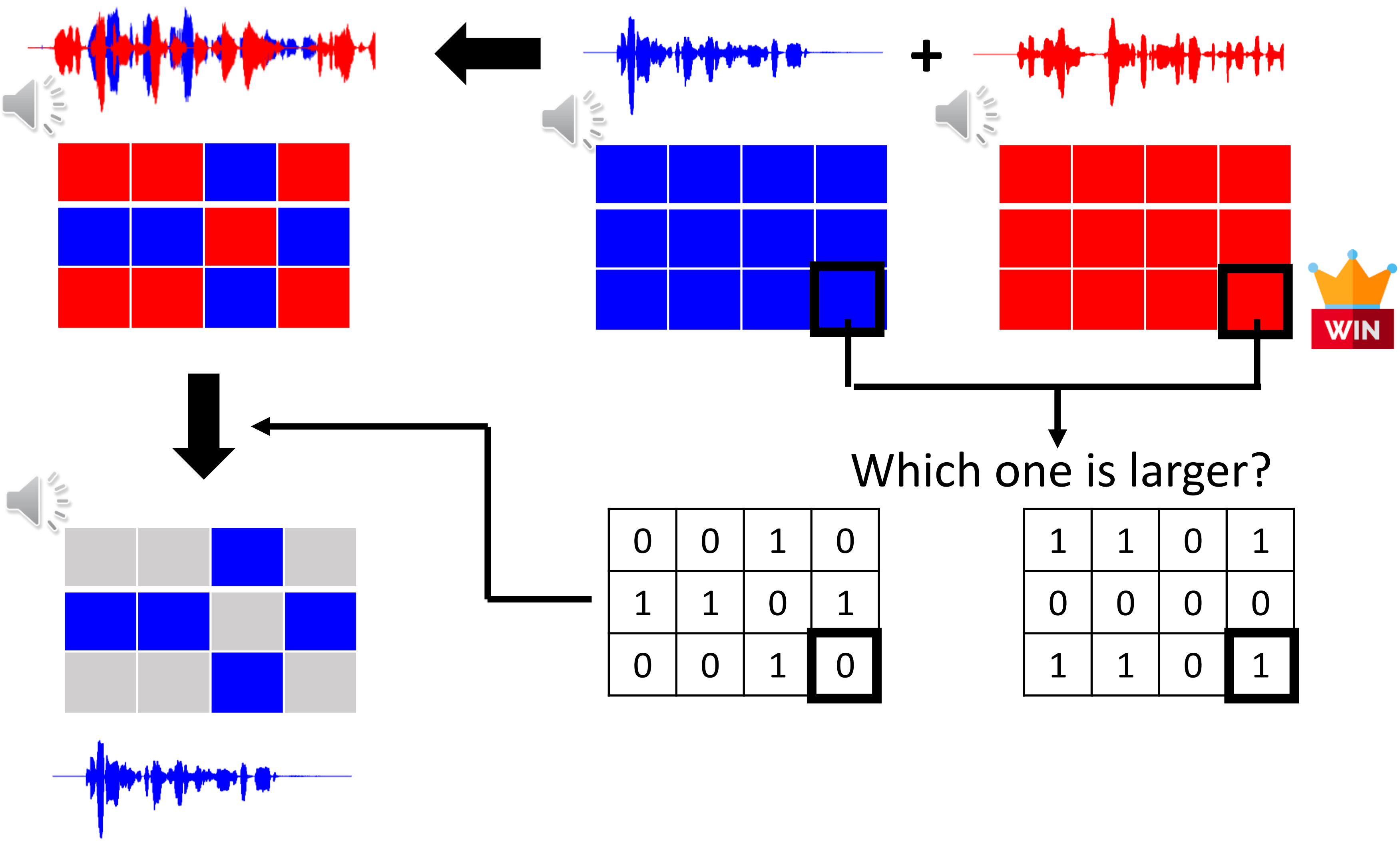
Mask can be binary or continuous.



Ideal Binary Mask (IBM)

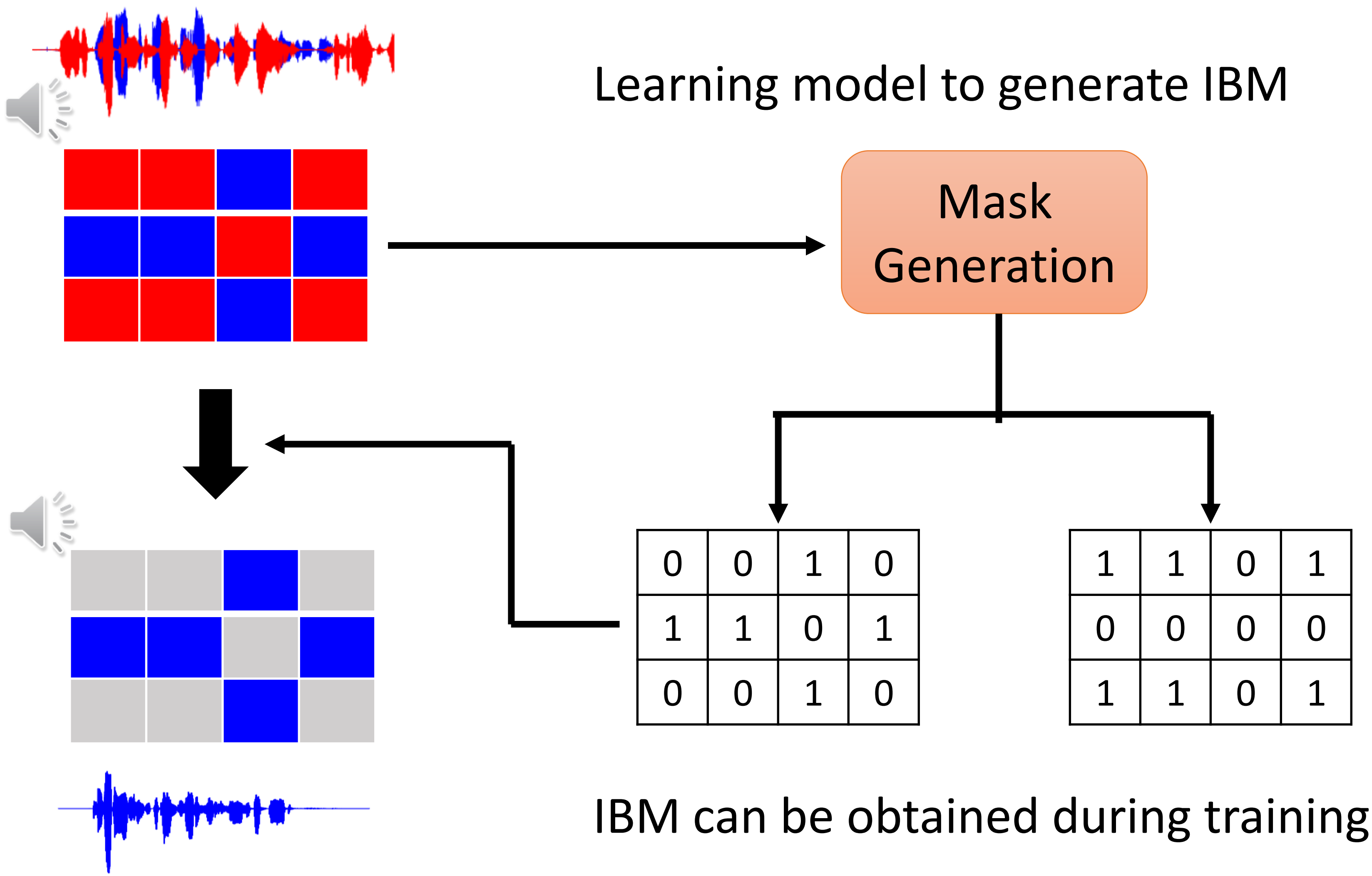
感謝吳元魁同學提供實驗結果

Each audio is represented by its spectrogram.



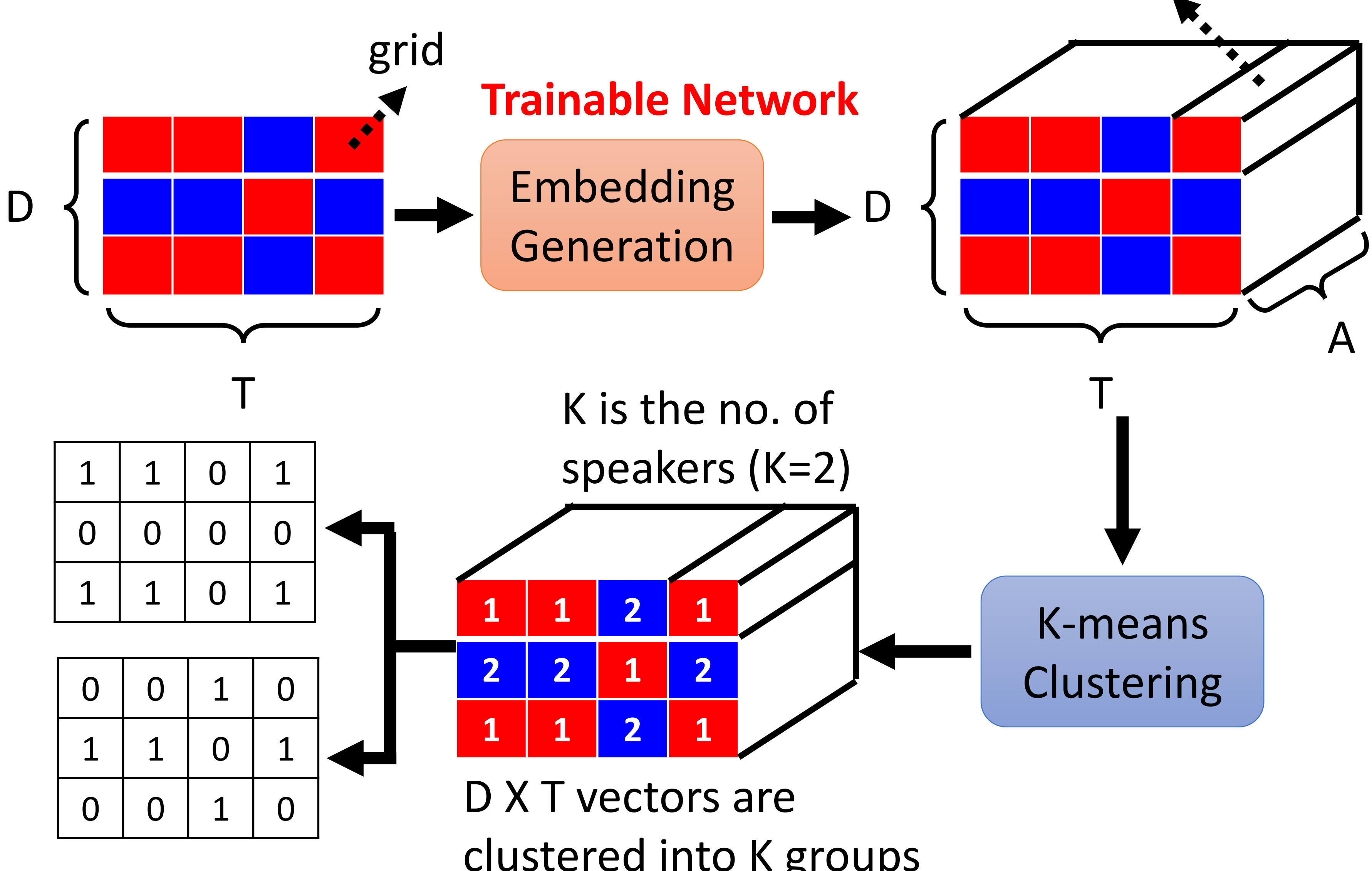
Ideal Binary Mask (IBM)

Each audio is represented by its spectrogram.

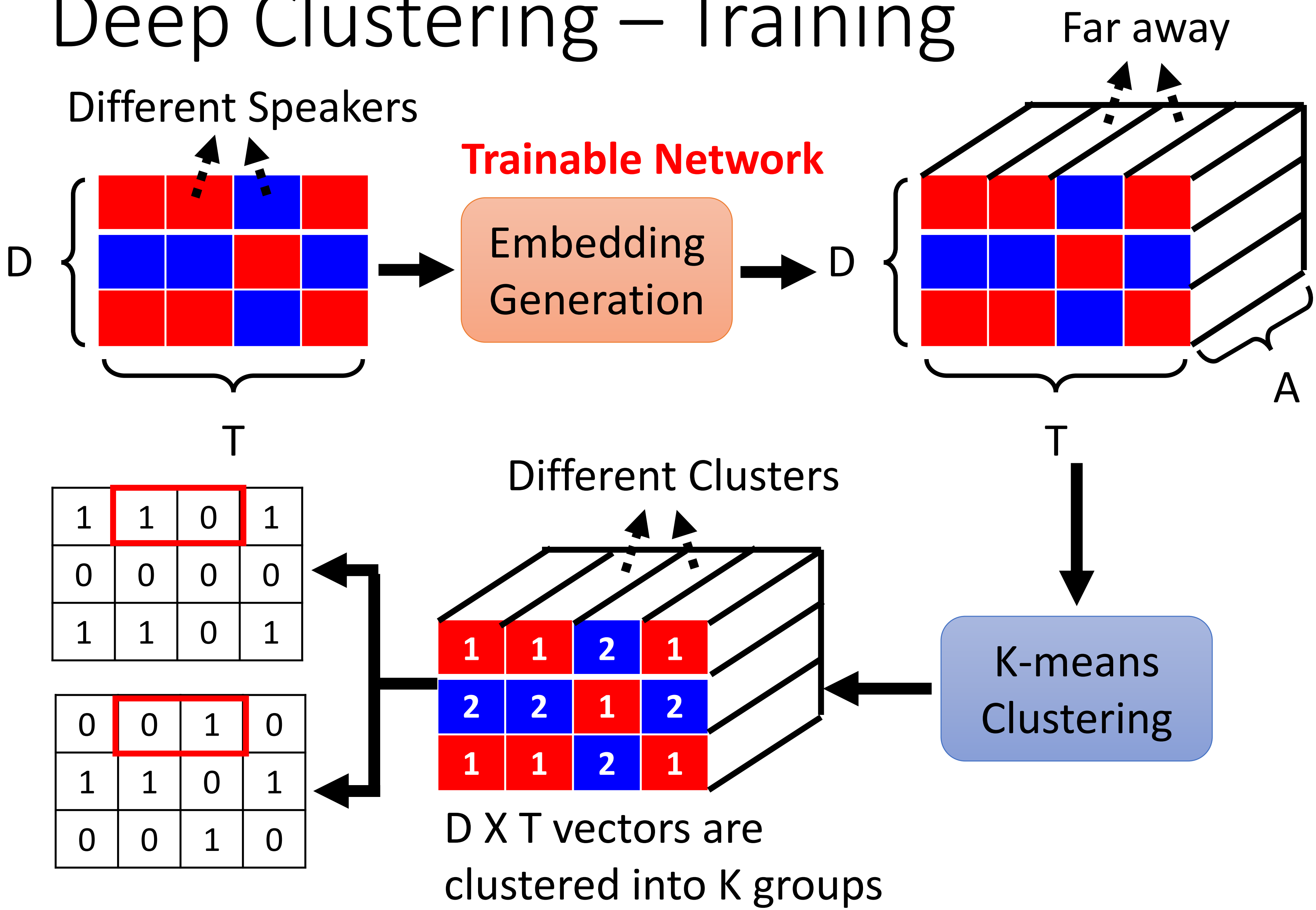


[Hershey, et al., ICASSP'16]

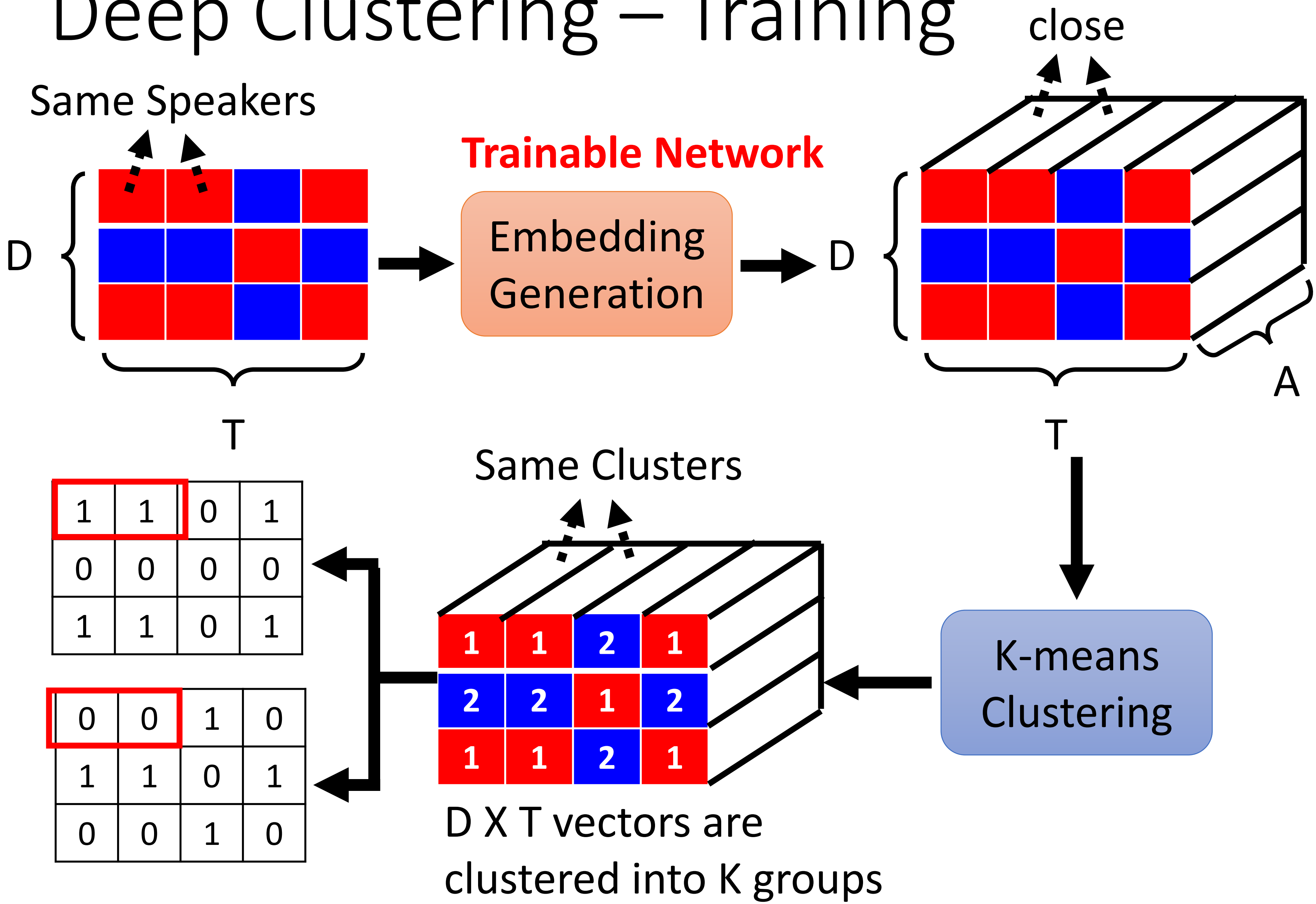
Deep Clustering



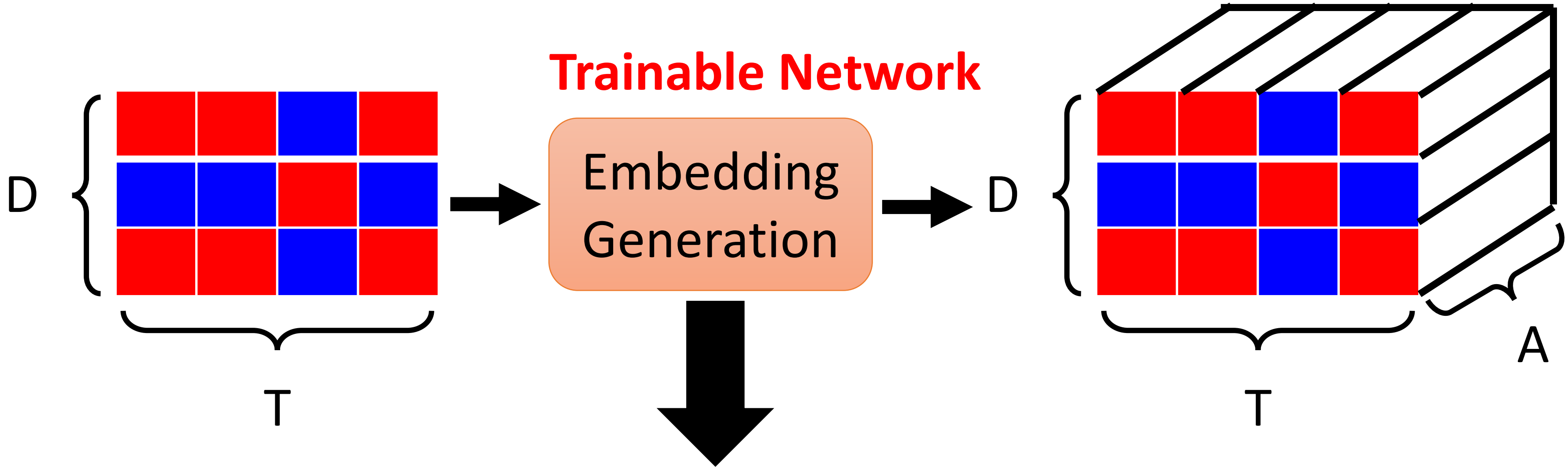
Deep Clustering – Training



Deep Clustering – Training



Deep Clustering – Training



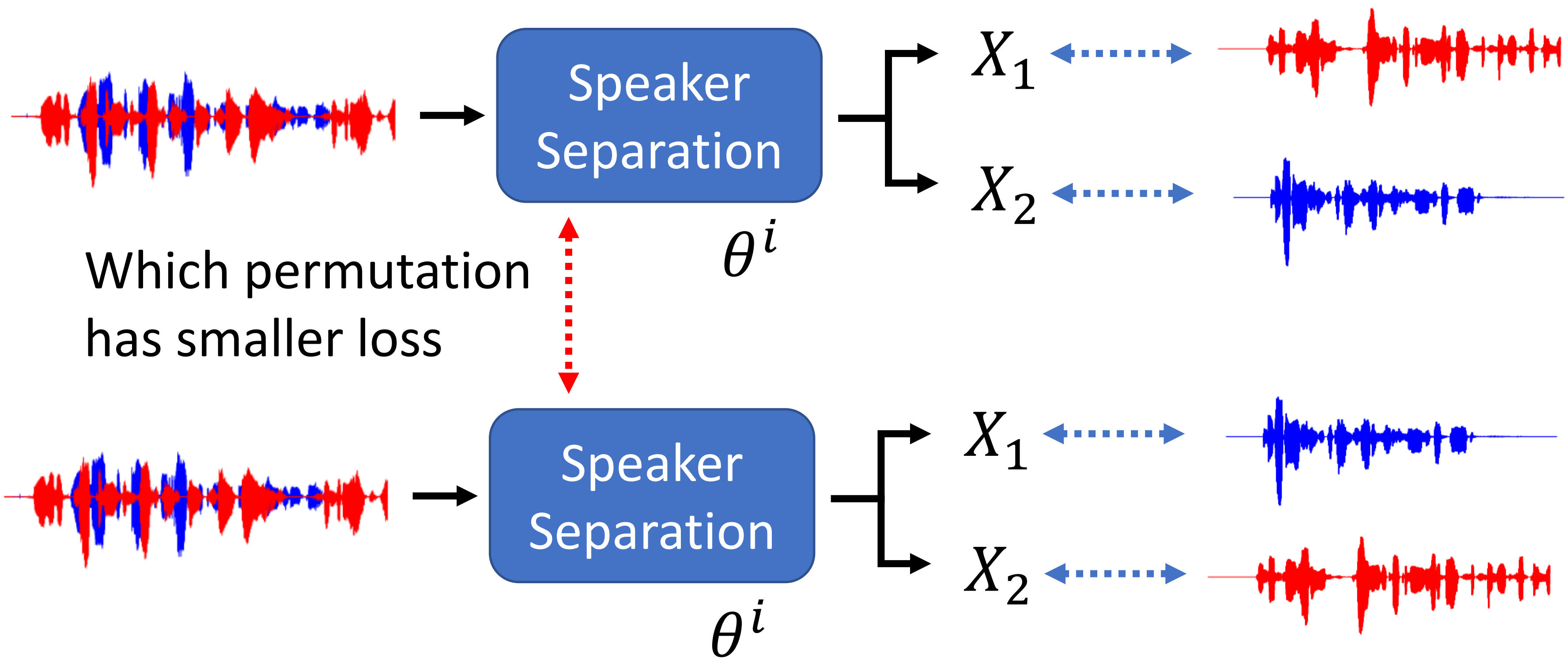
- The grids for different speakers are far way.
- The grids belonging to the same speakers are close to each other.

It is possible to train with two speakers, but test on three speakers (K=3 during k-means)! [Hershey, et al., ICASSP'16]

Permutation Invariant Training (PIT)

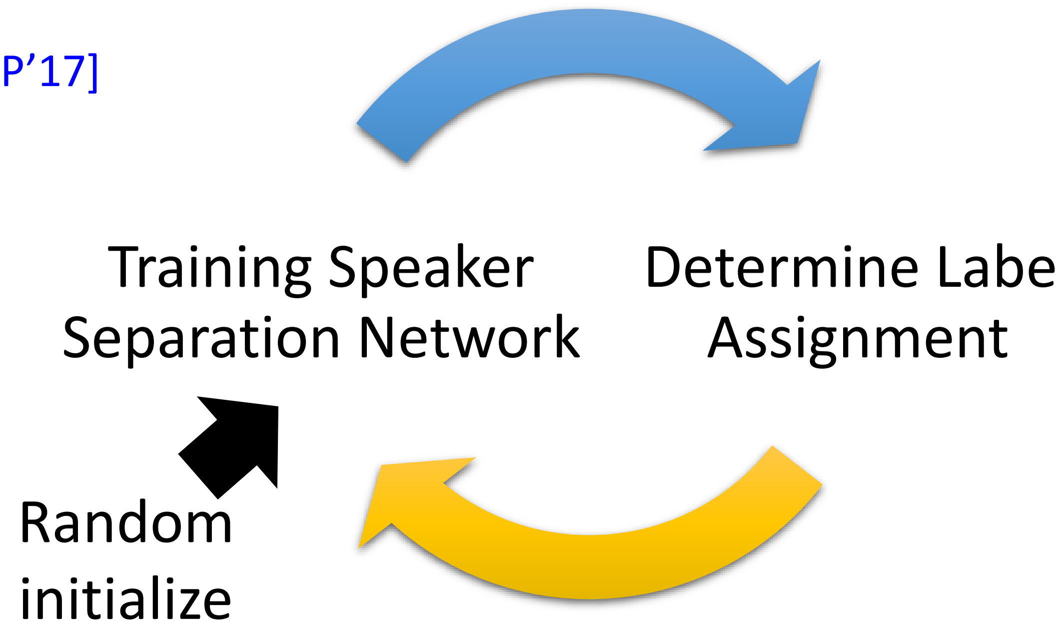
Permutation Invariant Training (PIT)

Given a speaker separation model θ^i, we can determine the permutation

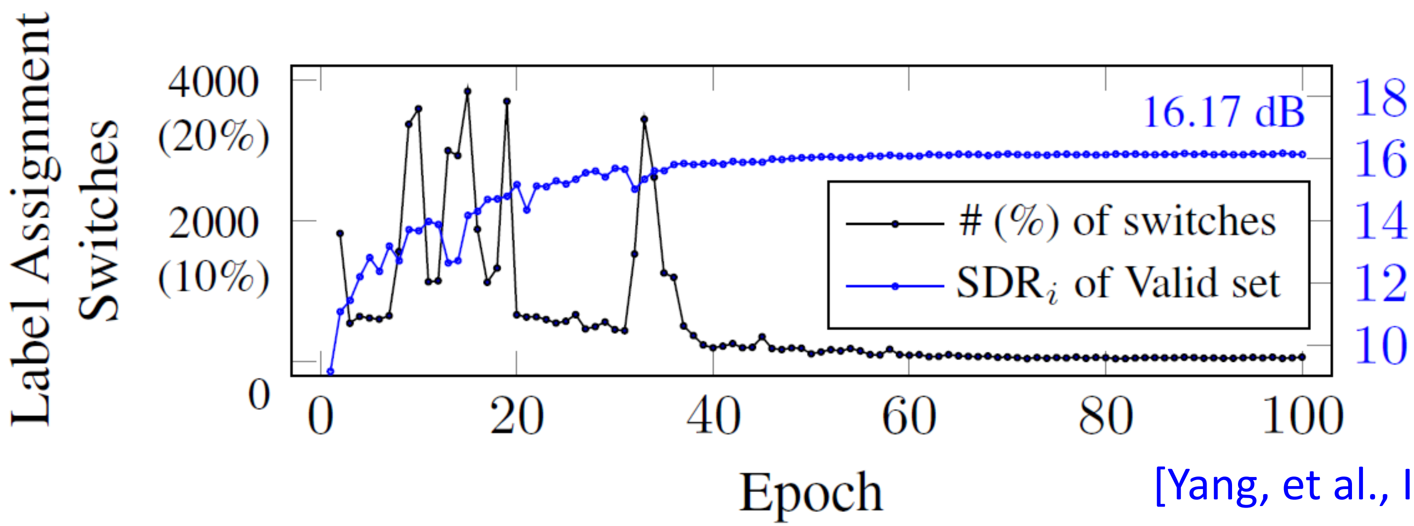


But we need permutation to train speaker separation model …

PIT [Kolbæk, et al., TASLP'17]

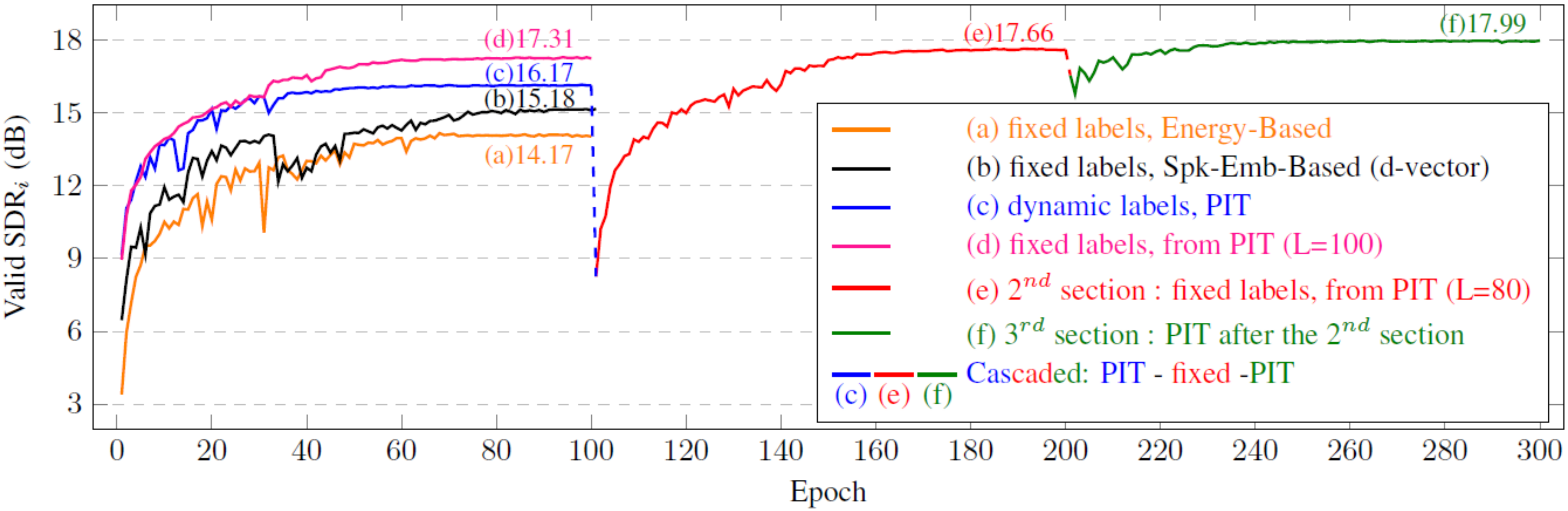


At the beginning, the assignment is not stable.



[Yang, et al., ICASSP'20]

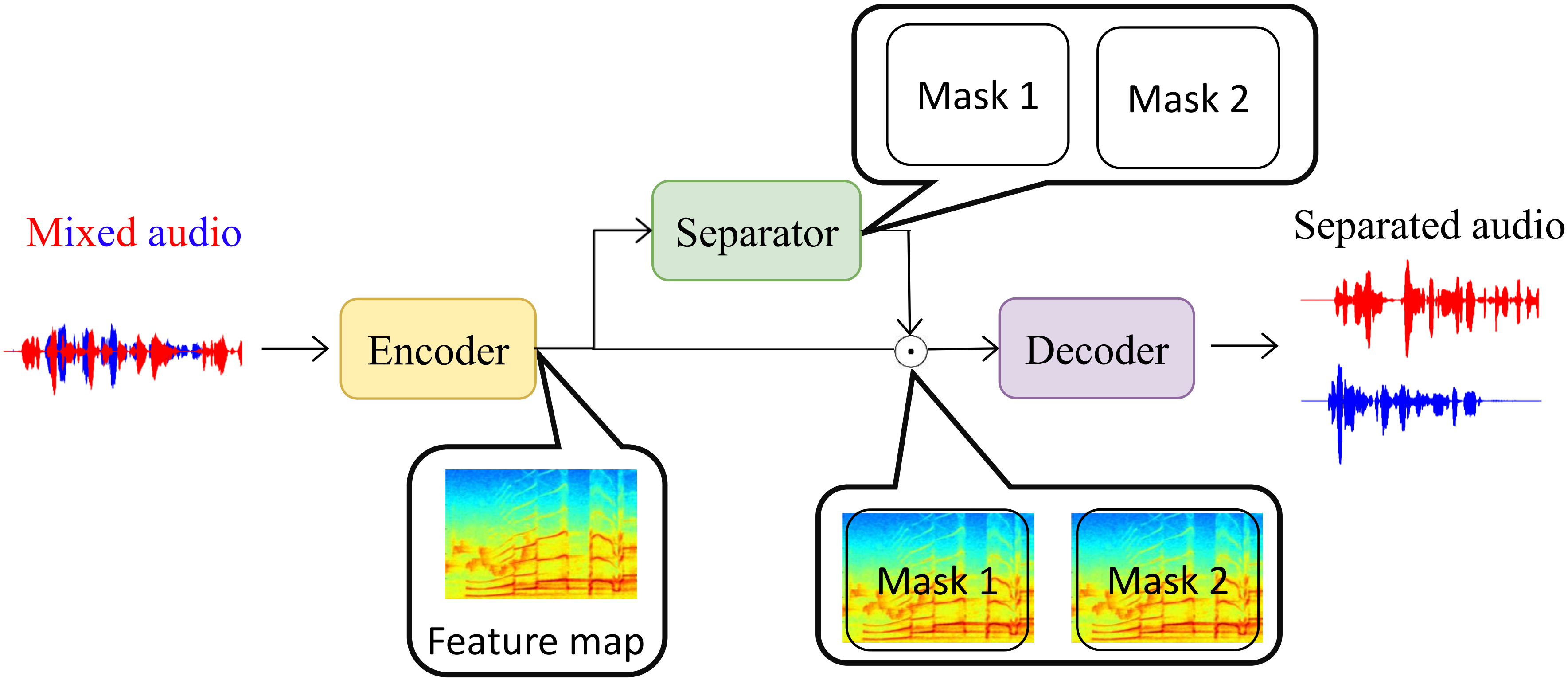
PIT [Kolbæk, et al., TASLP'17]

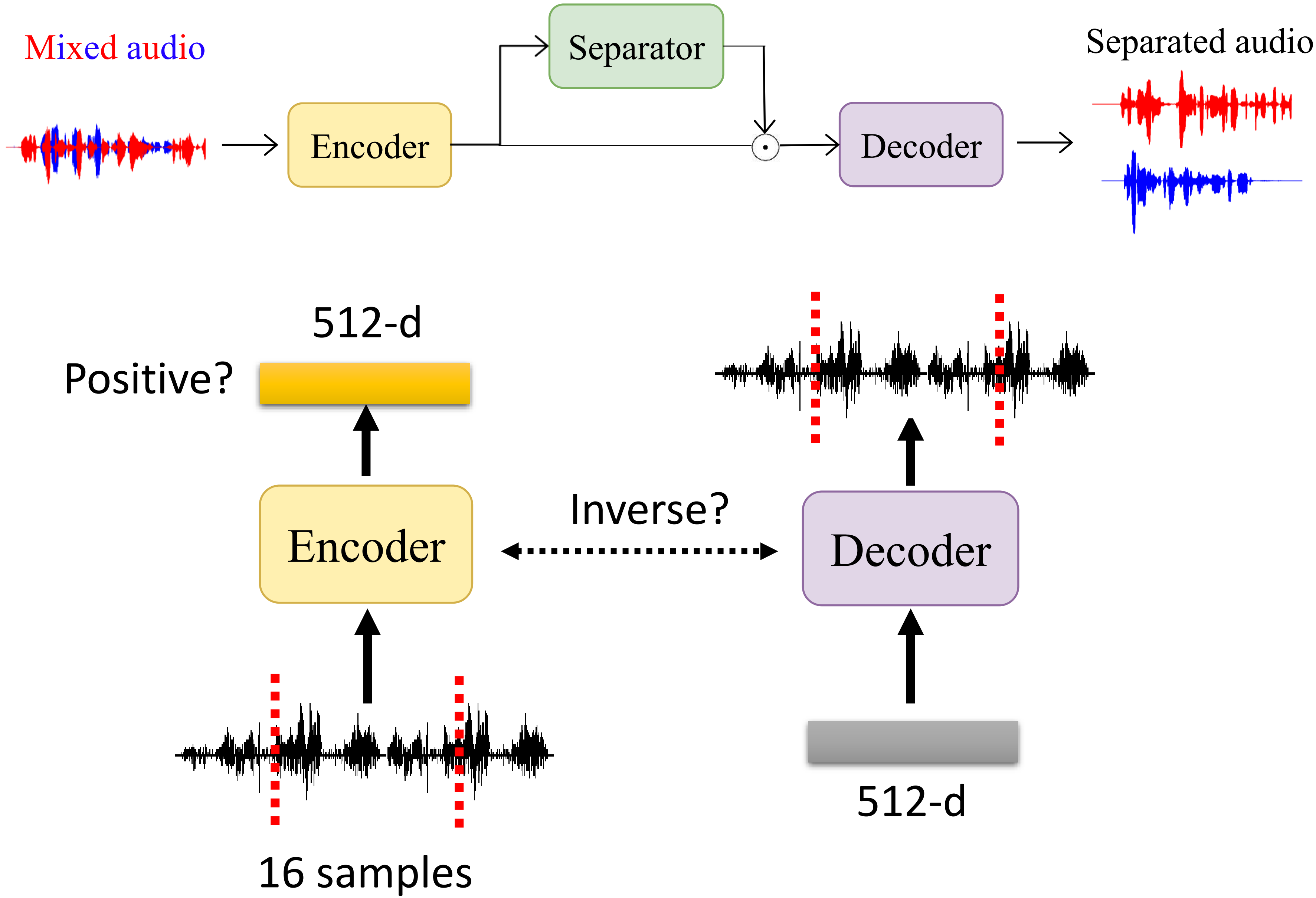


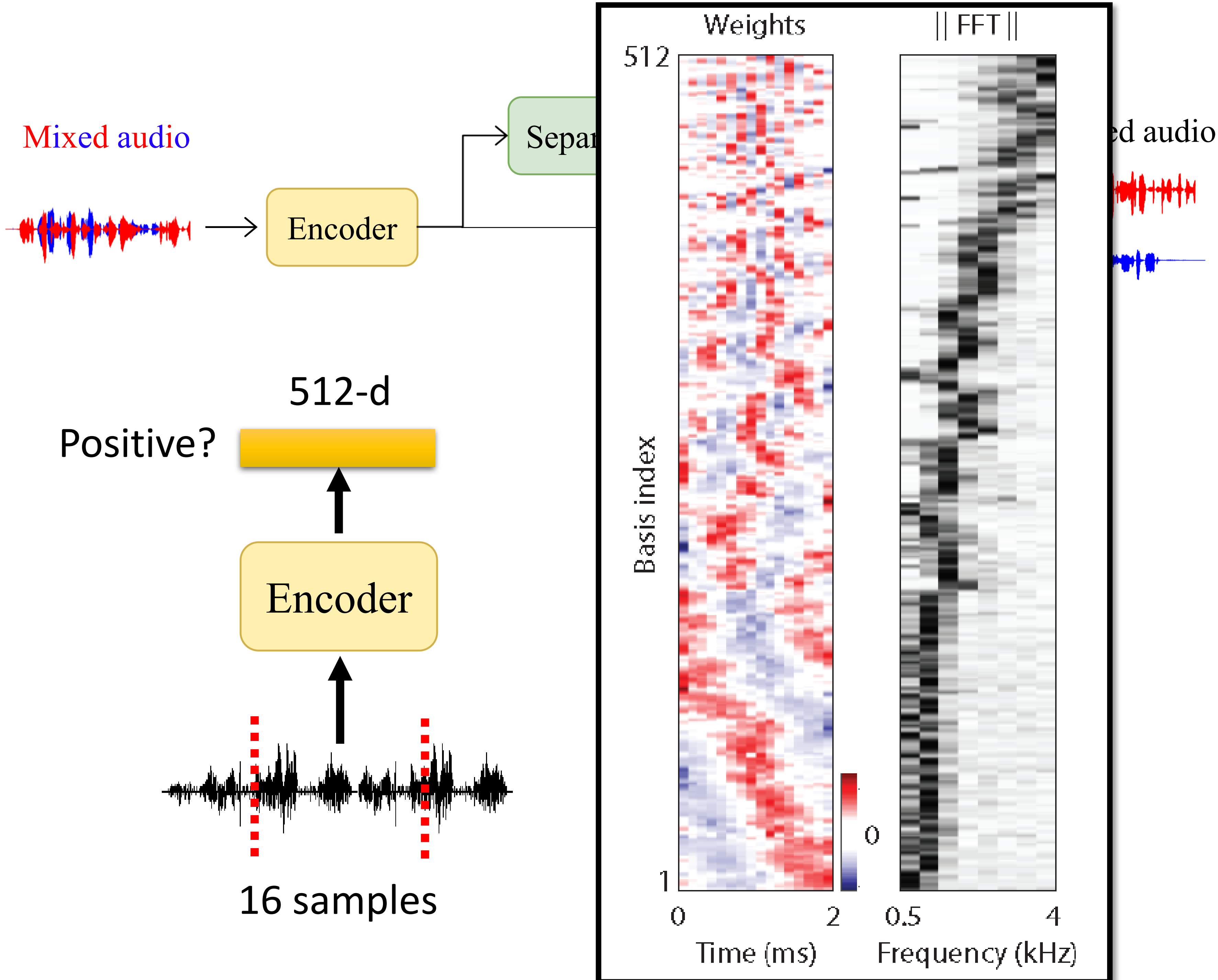
[Yang, et al., ICASSP'20]

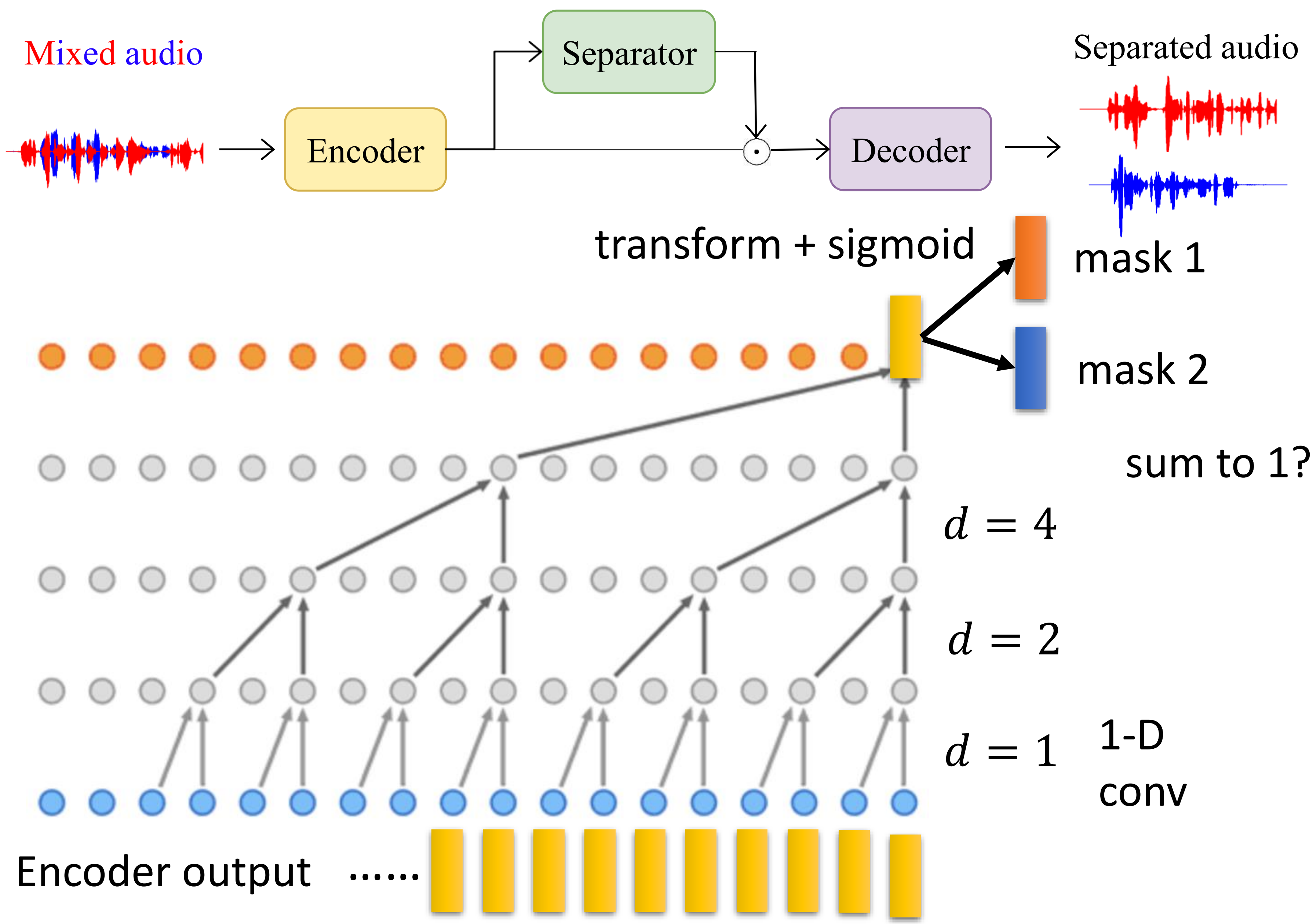
TasNet – Time-domain Audio Separation Network

[Luo, et al., TASLP'19]

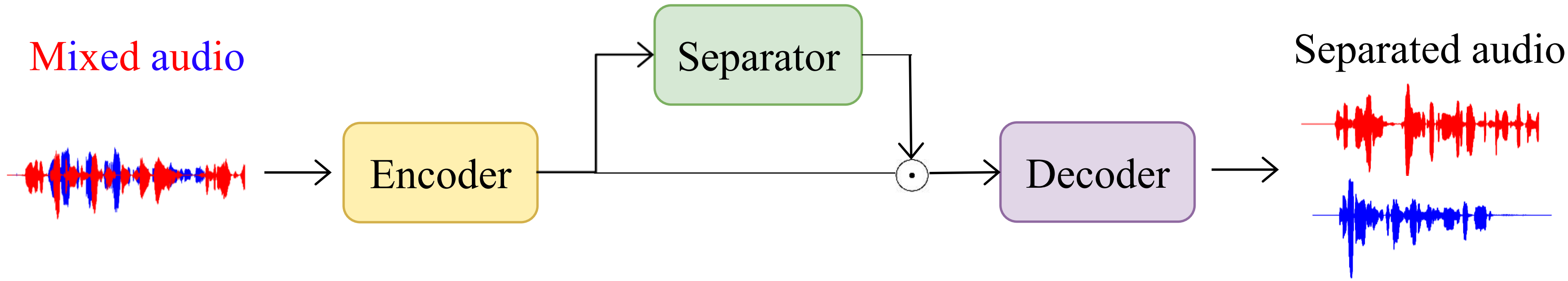


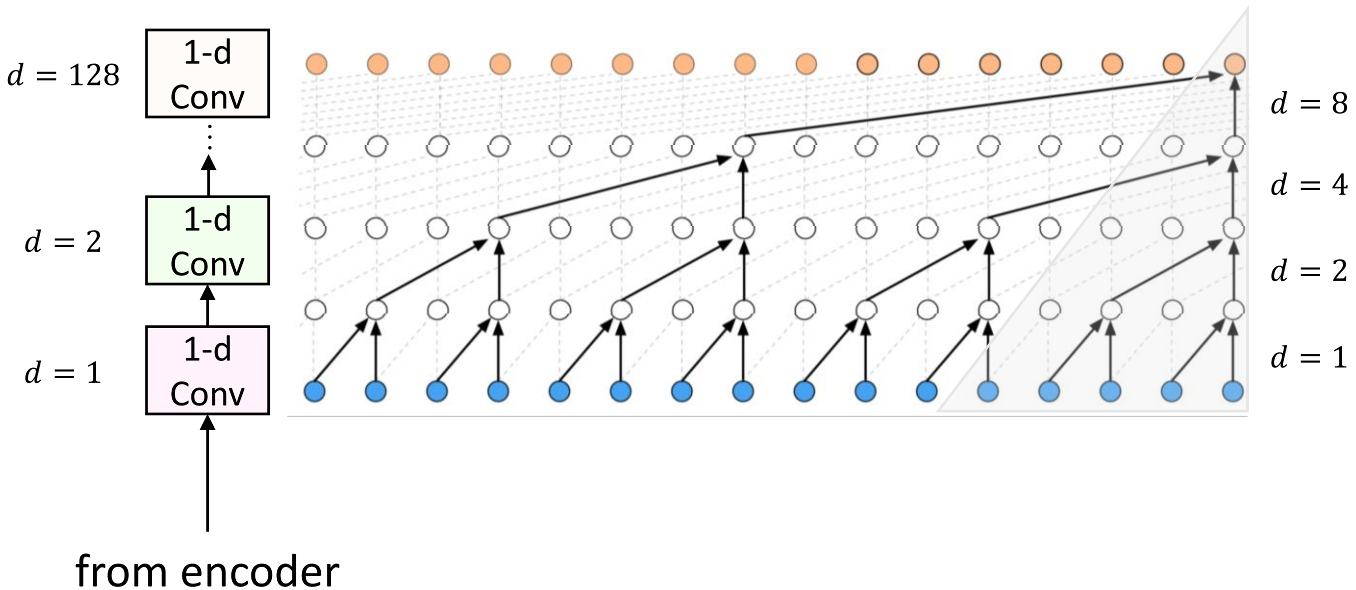

Mixed audio
Encoder
Separator
Decoder
Separated audio
512-d
Positive?
Encoder
Inverse?
Decoder
16 samples
512-d

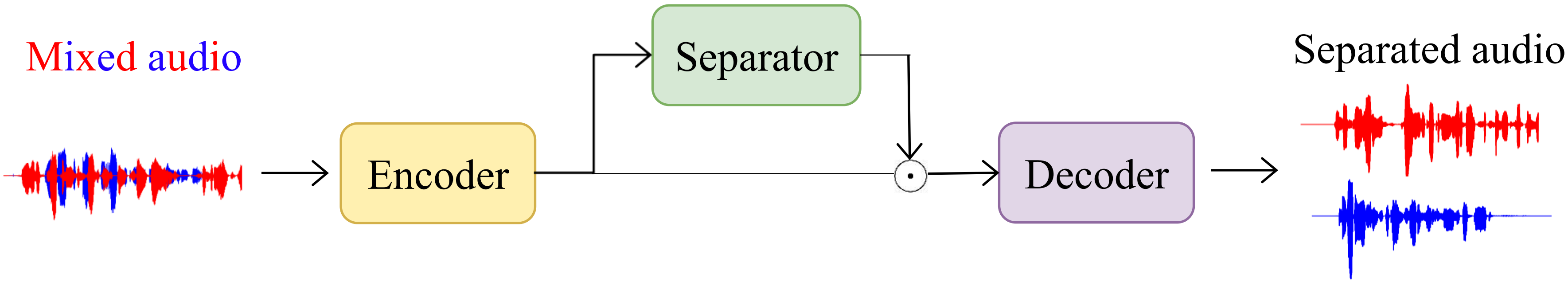

Mixed audio
Encoder
512-d
Positive?
Encoder
16 samples
Weights
|| FFT ||
512
Basis index
1
0
0
2
0.5
4
Time (ms)
Frequency (kHz)

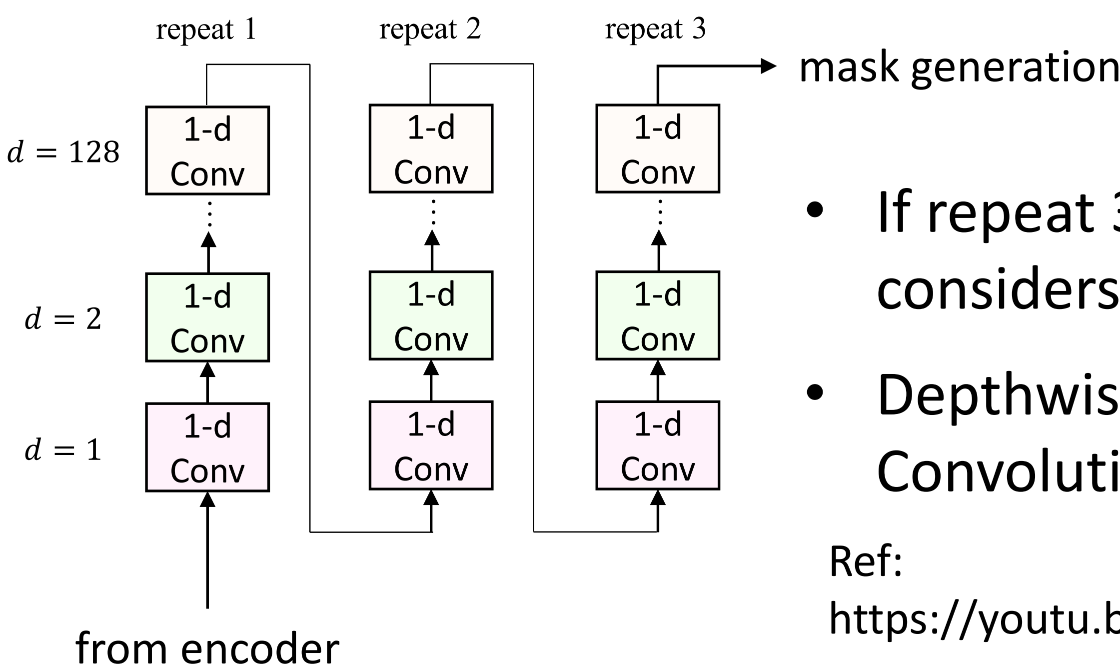


Mixed audio
Encoder
Separator
Decoder
Separated audio
transform + sigmoid
mask 1
mask 2
sum to 1?
d = 4
d = 2
d = 1
1-D conv
Encoder output ……


Mixed audio
Encoder
Separator
Decoder
Separated audio


d = 128
1-d Conv
d = 2
1-d Conv
d = 1
1-d Conv
from encoder
d = 8
d = 4
d = 2
d = 1





- If repeat 3, the model considers 1.53s
- Depthwise Separable Convolution

Ref: https://youtu.be/L0TOXlNpCJ8

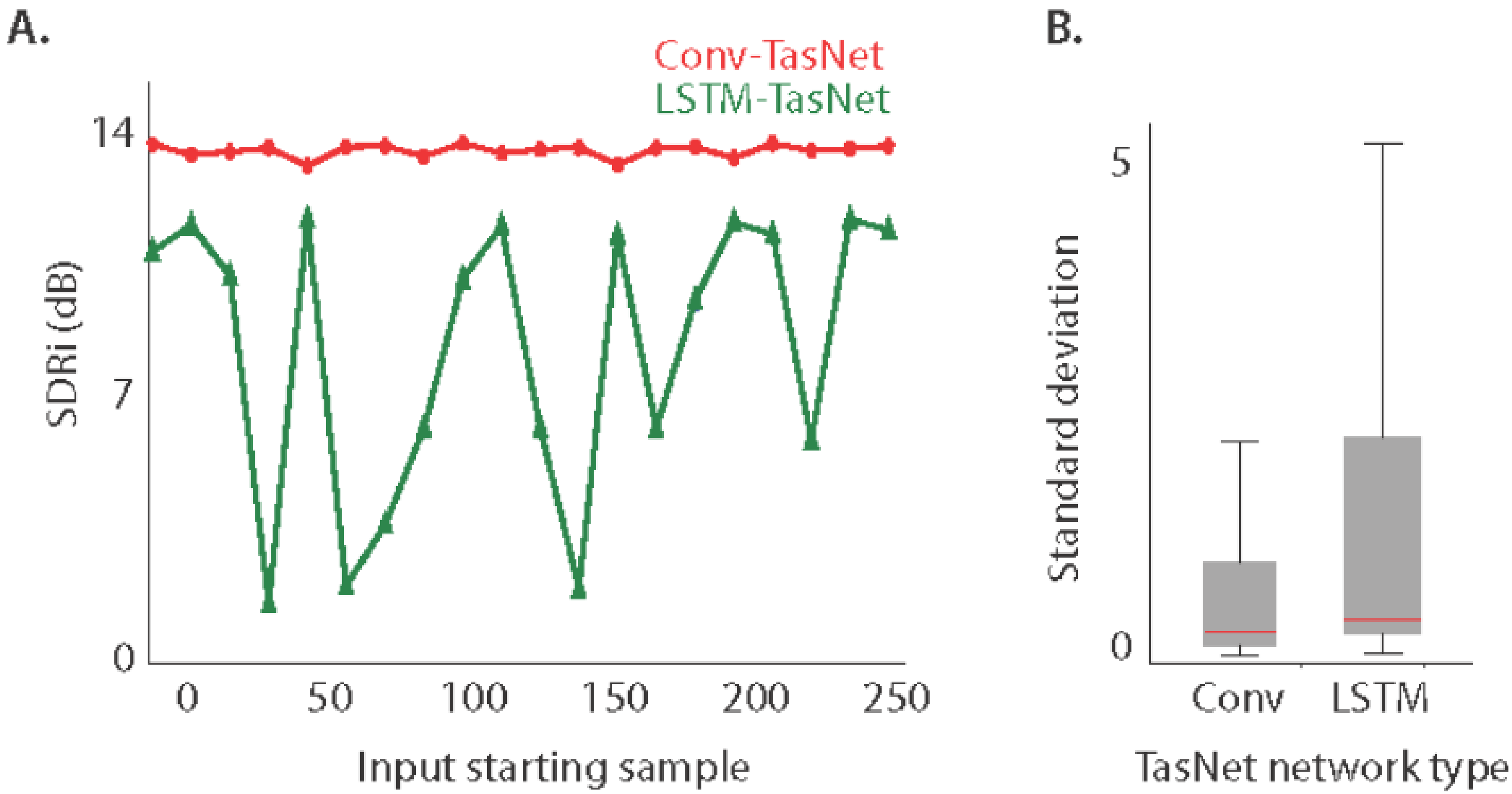
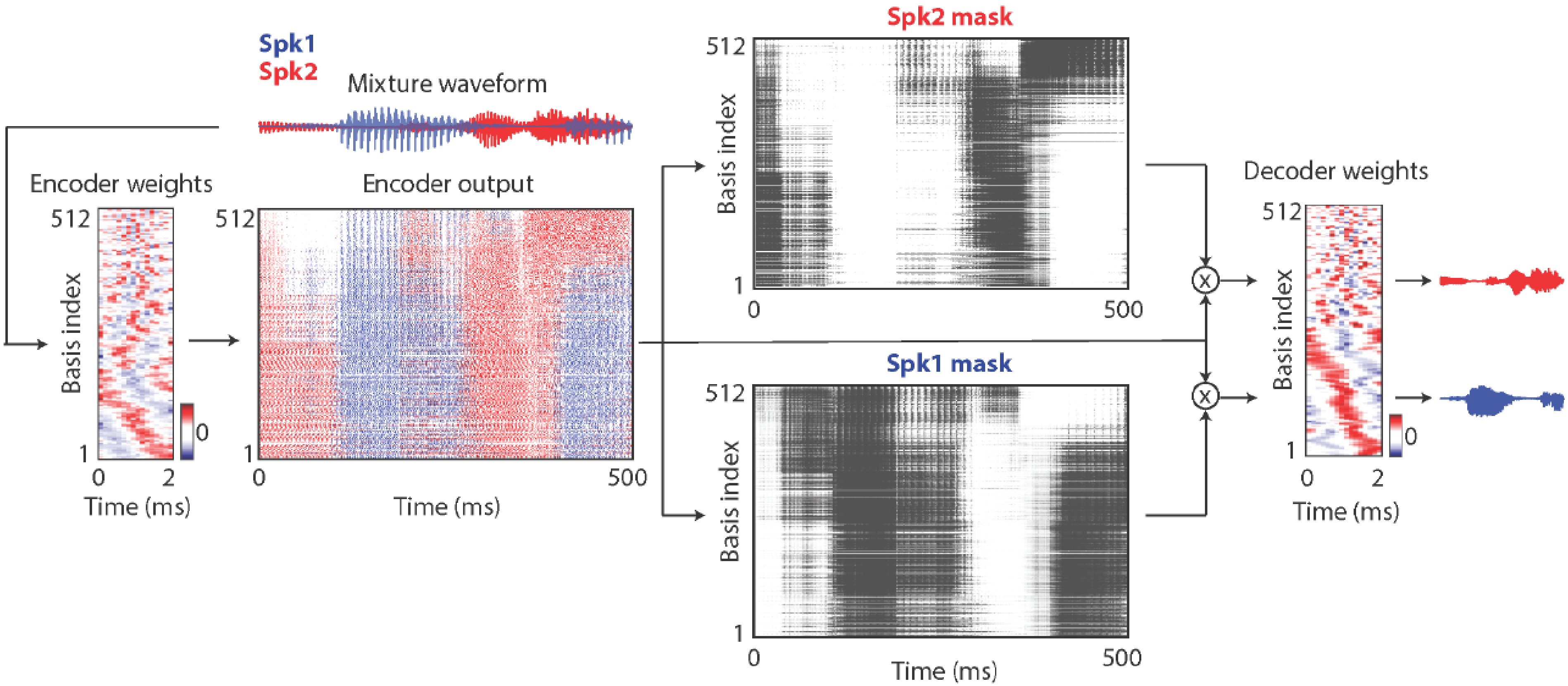


Fig. 4. (A): SDRi of an example mixture separated using LSTM-TasNet and causal Conv-TasNet as a function of the starting point in the mixture. The performance of Conv-TasNet is considerably more consistent and insensitive to the start point. (B): Standard deviation of SDRi across all the mixtures in the WSJ0-2mix test set with varying starting points.


Spk1
Spk2
Mixture waveform
Encoder weights
512
Basis index
1
0
2
Time (ms)
0
Encoder output
512
1
0
500
Time (ms)
Spk2 mask
512
Basis index
1
0
500
Spk1 mask
512
Basis index
1
0
Time (ms)
500
Decoder weights
512
Basis index
1
0
2
Time (ms)
0

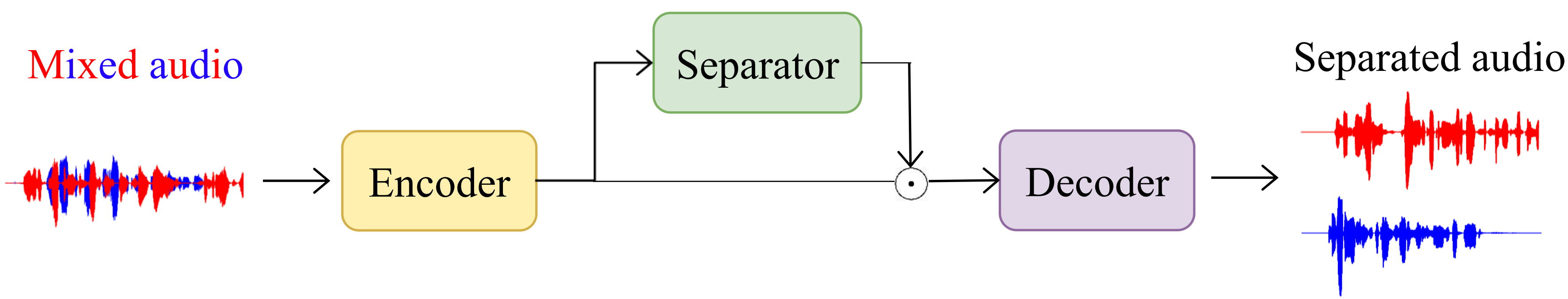

Mixed audio
Encoder
Separator
Decoder
Separated audio

Table 2. SI-SDR and SDR improvements (dB) on WSJ0-2mix.

Model	ΔSI-SDR	ΔSDR
Deep Clustering (Isik et al., 2016)	10.8	–
uPIT-blstm-st (Kolbaek et al., 2017)	–	10.0
Deep Attractor Net. (Chen et al., 2017)	10.5	–
Anchored Deep Attr. (Luo et al., 2018)	10.4	10.8
Grid LSTM PIT (Xu et al., 2018)	–	10.2
ConvLSTM-GAT (Li et al., 2018)	–	11.0
Chimera++ (Wang et al., 2018b)	11.5	12.0
WA-MISI-5 (Wang et al., 2018c)	12.6	13.1
blstm-TasNet (Luo & Mesgarani, 2018)	13.2	13.6
Conv-TasNet (Luo & Mesgarani, 2019)	15.3	15.6
Conv-TasNet+MBT (Lam et al., 2019)	15.5	15.9
DeepCASA (Liu & Wang, 2019)	17.7	18.0
FurcaNeXt (Zhang et al., 2020)	–	18.4
DualPathRNN (Luo et al., 2019)	18.8	19.0
Wavesplit	**19.0**	**19.2**
Wavesplit + Dynamic mixing	**20.4**	**20.6**

source of results: https://arxiv.org/pdf/2002.08933.pdf

Are all the problems solved?

Tasnet

mix

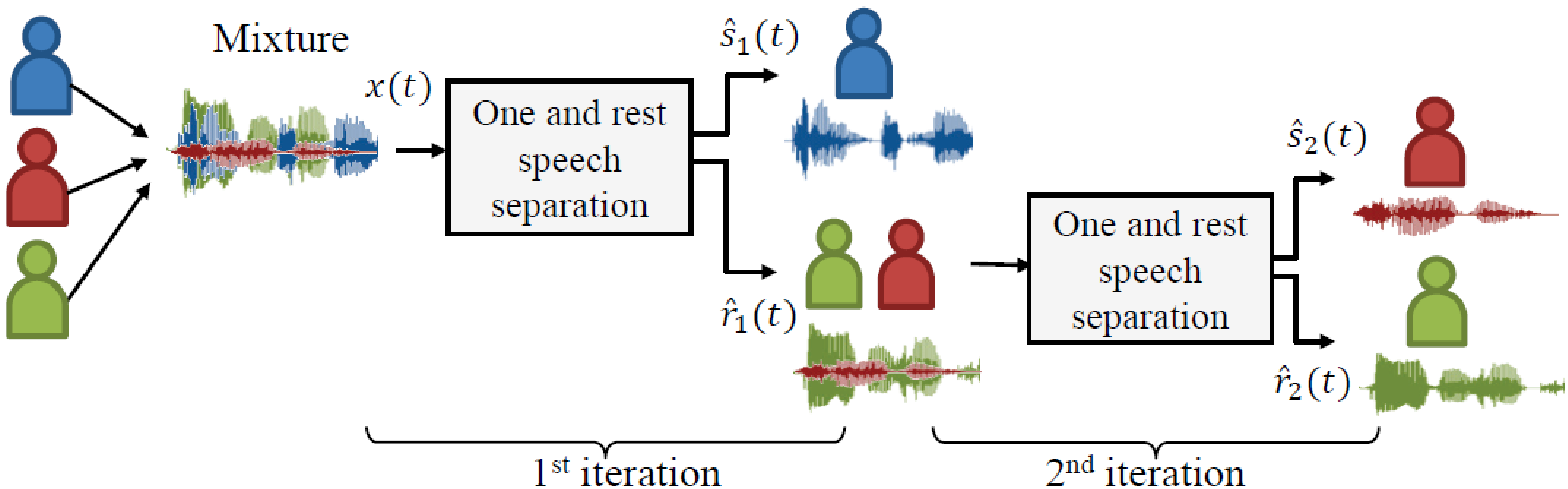
Deep Clustering

感謝 Taiwan AI Labs Machine Learning Engineer
林資偉 同學提供實驗結果

More …

Unknown number of speakers

[Takahashi, et al., INTERSPEECH'19]

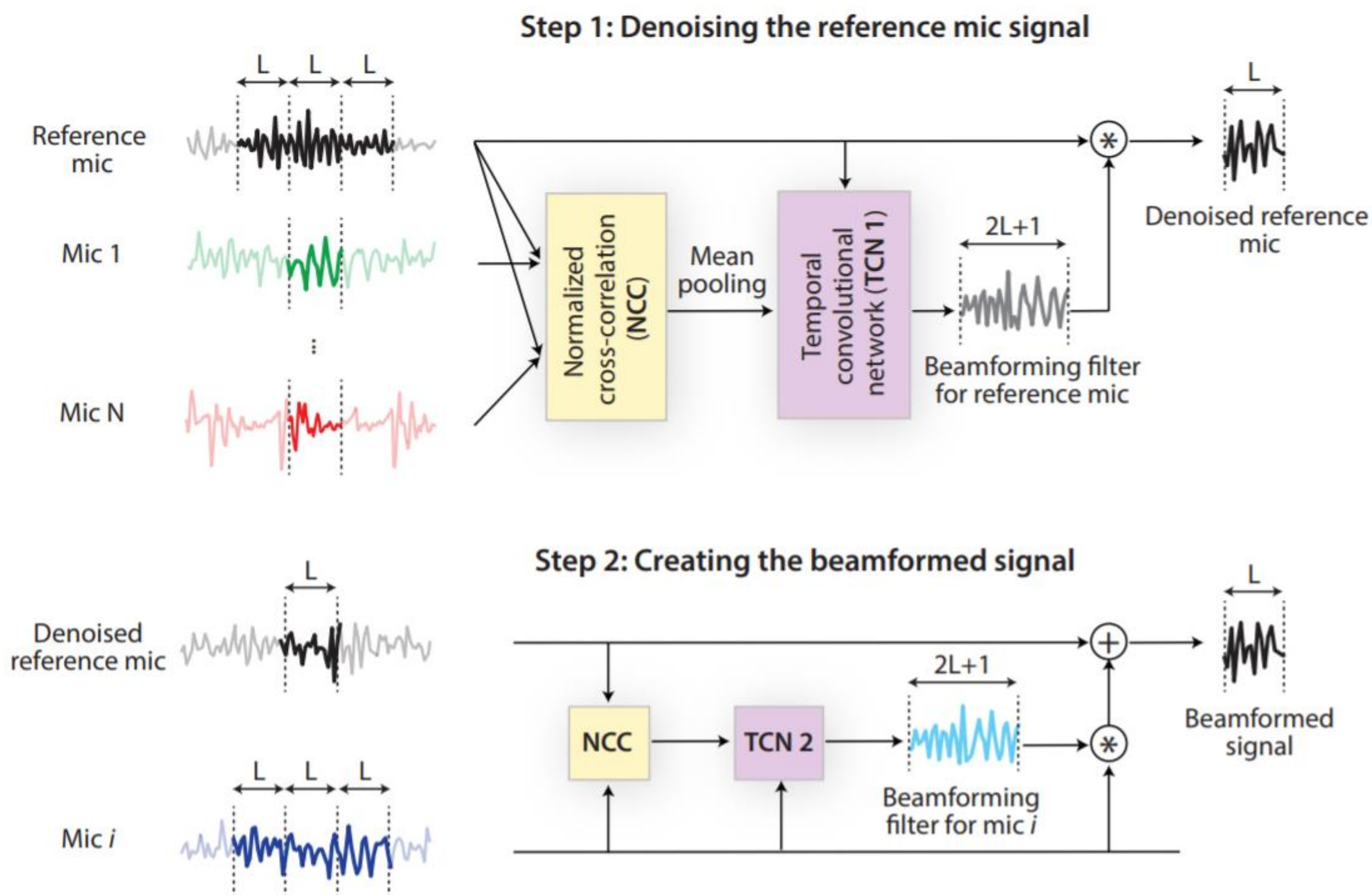


recursively separating a speaker

Source of image: https://arxiv.org/pdf/1904.03065.pdf

Multiple Microphones

[Luo, et al., ASRU'19]

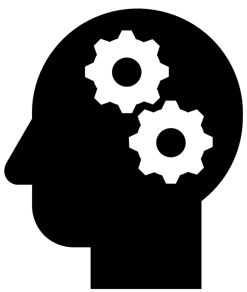


Source of image: https://arxiv.org/pdf/1909.13387.pdf

Task-oriented Optimization

Who would listen to the results of speech enhancement or speaker separation?

for human



Quality
Intelligibility

Optimizing STOI, PESQ

Non-differentiable

[Fu, et al., ICML'19]

for machine

ASR
Speaker Verification

Optimizing system performance

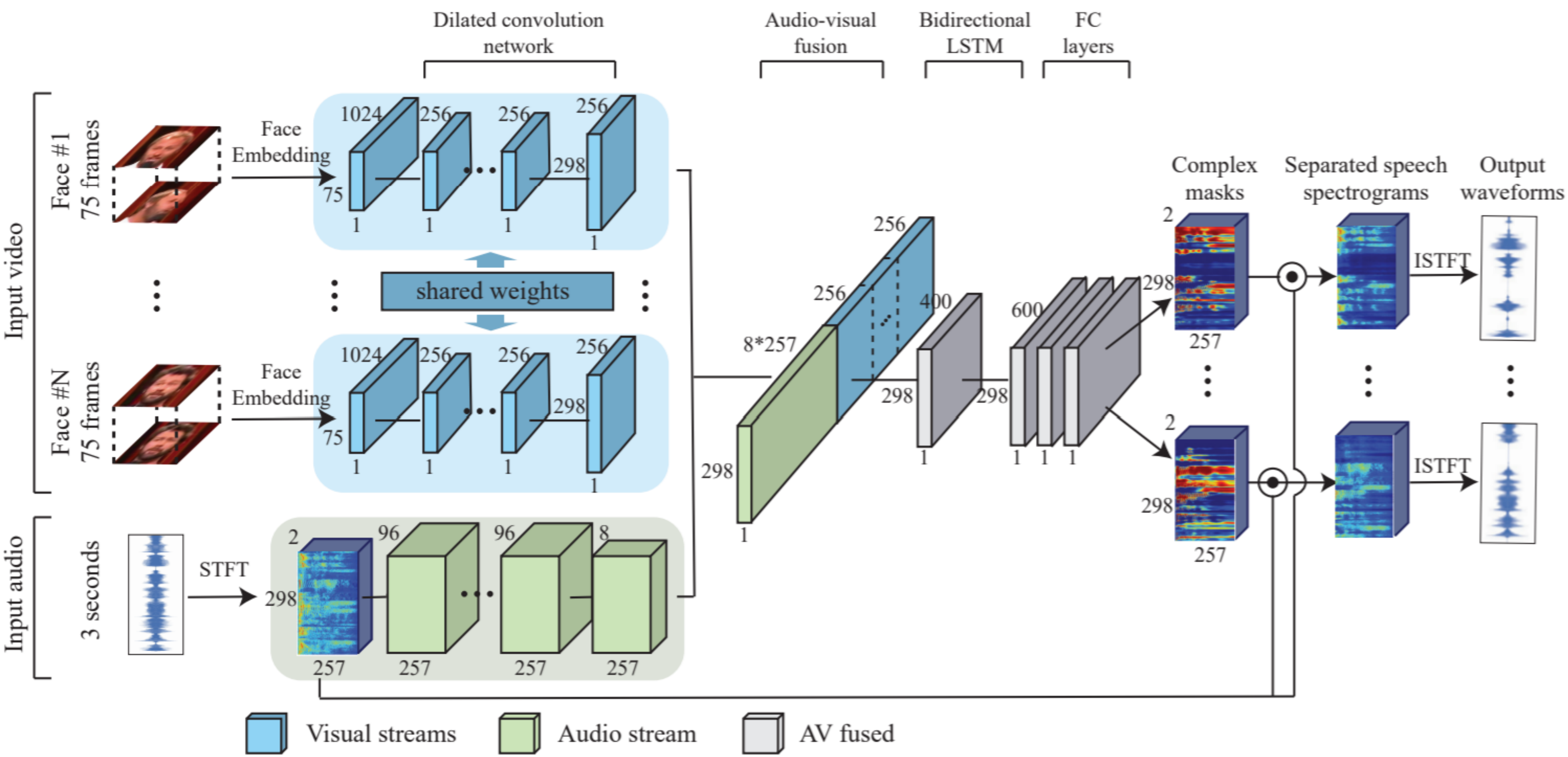
[Shon, et al., INTERSPEECH'19]

Visual Information

https://ai.googleblog.com/2018/04/looking-to-listen-audio-visual-speech.html

Visual Information

[Ephrat, et al., SIGGRAPH'18]



Source of image: https://arxiv.org/pdf/1804.03619.pdf

To learn more ……

- Denoise Wavnet [Rethage, et al., ICASSP’18]
- Chimera++ [Wang, et al., ICASSP’18]
- Phase Reconstruction Model [Wang, et al., ICASSP’19]
- Deep Complex U-Net: Complex masking [Choi, et al., ICLR’19]
- Deep CASA: Make CASA great again! [Liu, et al., TASLP’19]
- Wavesplit: state-of-the-art on benchmark corpus WSJ0-2mix [Zeghidour, et al., arXiv’20]

Reference

- [Hershey, et al., ICASSP’16] John R. Hershey, Zhuo Chen, Jonathan Le Roux, Shinji Watanabe, Deep clustering: Discriminative embeddings for segmentation and separation, ICASSP, 2016
- [Kolbæk, et al., TASLP’17] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, Jesper Jensen, Multi-talker Speech Separation with Utterance-level Permutation Invariant Training of Deep Recurrent Neural Networks, TASLP, 2017
- [Yang, et al., ICASSP’20] Gene-Ping Yang, Szu-Lin Wu, Yao-Wen Mao, Hung-yi Lee, Lin-shan Lee, Interrupted and cascaded permutation invariant training for speech separation, ICASSP, 2020
- [Ephrat, et al., SIGGRAPH’18] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, Michael Rubinstein, Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation, SIGGRAPH, 2018

Reference

- [Luo, et al., ASRU’19] Yi Luo, Enea Ceolini, Cong Han, Shih-Chii Liu, Nima Mesgarani, FaSNet: Low-latency Adaptive Beamforming for Multi-microphone Audio Processing, ASRU, 2019
- [Zeghidour, et al., arXiv’20] Neil Zeghidour, David Grangier, Wavesplit: End-to-End Speech Separation by Speaker Clustering, arXiv, 2020
- [Liu, et al., TASLP’19] Yuzhou Liu, DeLiang Wang, Divide and Conquer: A Deep CASA Approach to Talker-Independent Monaural Speaker Separation, in *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 2019
- [Choi, et al., ICLR’19] Hyeong-Seok Choi, Jang-Hyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, Kyogu Lee, Phase-aware Speech Enhancement with Deep Complex U-Net, ICLR, 2019
- [Luo, et al., TASLP’19] Yi Luo, Nima Mesgarani, Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation, TASLP, 2019
- [Rethage, et al., ICASSP’18] Dario Rethage, Jordi Pons, Xavier Serra, A Wavenet for Speech Denoising, ICASSP, 2018
- [Wang, et al., ICASSP’18] Zhong-Qiu Wang, Jonathan Le Roux, John R. Hershey, Alternative Objective Functions for Deep Clustering, ICASSP, 2018

Reference

- [Wang, et al., ICASSP'19] Zhong-Qiu Wang, Ke Tan, DeLiang Wang, Deep Learning Based Phase Reconstruction for Speaker Separation: A Trigonometric Perspective, ICASSP, 2019
- [Fu, et al., ICML'19] Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, Shou-De Lin, MetricGAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement, ICML, 2019
- [Shon, et al., INTERSPEECH'19] Suwon Shon, Hao Tang, James Glass, VoiceID Loss: Speech Enhancement for Speaker Verification, INTERSPEECH, 2019
- [Takahashi, et al., INTERSPEECH'19] Naoya Takahashi, Sudarsanam Parthasaarathy, Nabarun Goswami, Yuki Mitsufuji, Recursive speech separation for unknown number of speakers, INTERSPEECH, 2019