Source separation

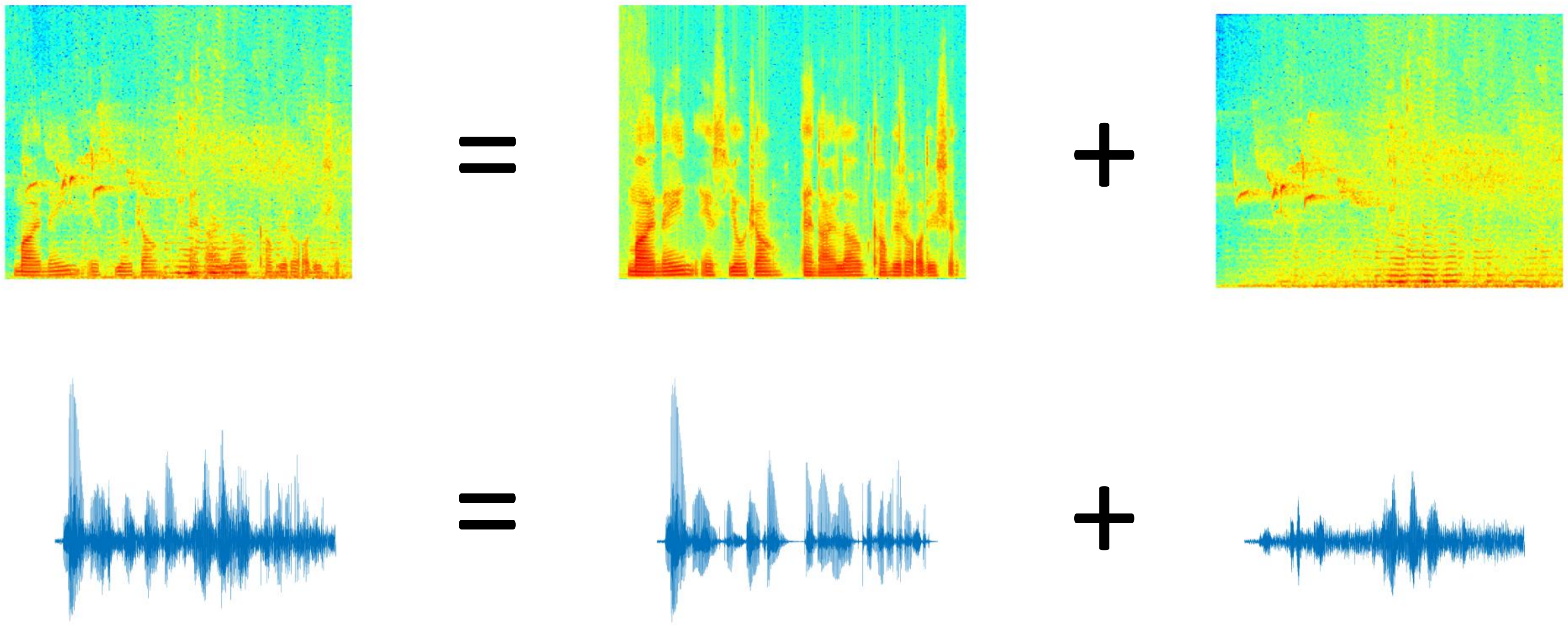
Yuxiang Wang & Neil Zhang

Outline

- Introduction
- Spectrogram-based approaches
- Waveform-based approaches
- Future directions

Introduction

- Source separation: mixed signals ⟶ a set of sources



Introduction

Several topics in audio source separation:

By source type:

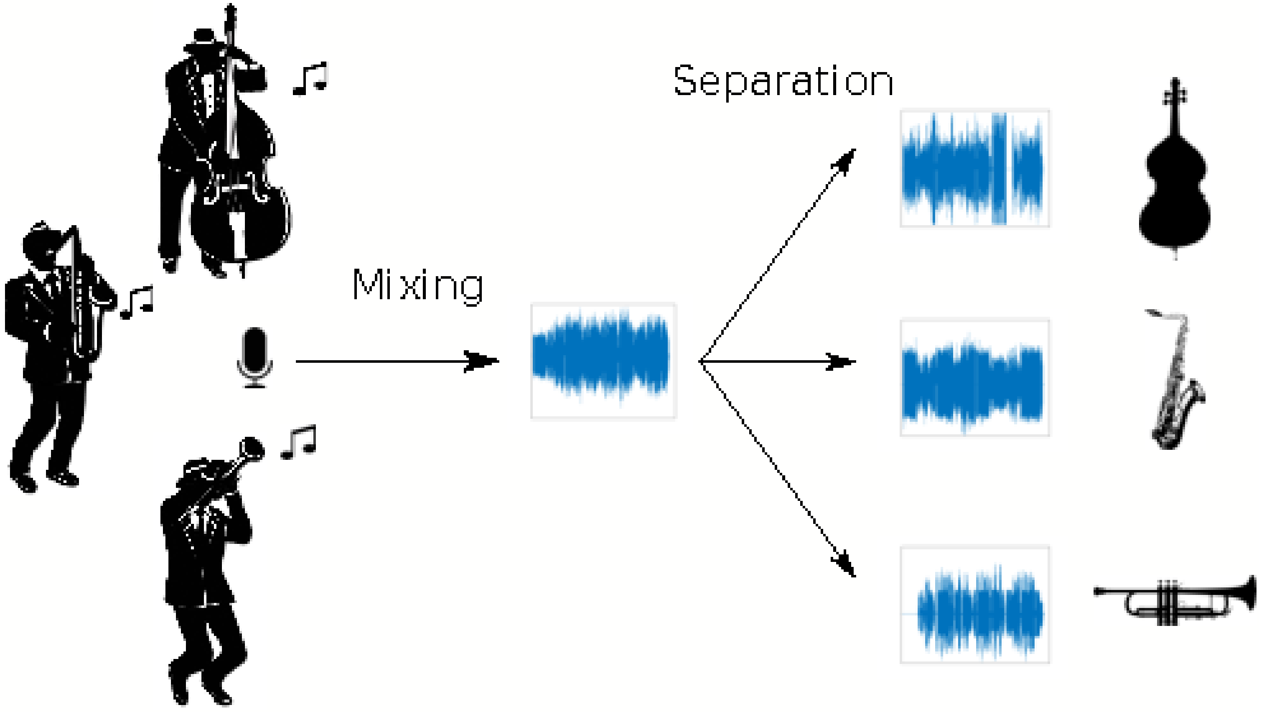
- Music source separation (music instruments)
- Speech separation (speech voices)
- Speech enhancement (speech + noise)
- Singing voice separation (singing voice + music)

By channel:

- Monaural source separation
- Multi-channel source separation
- Audio-visual source separation (separation tasks with visual information)

Introduction

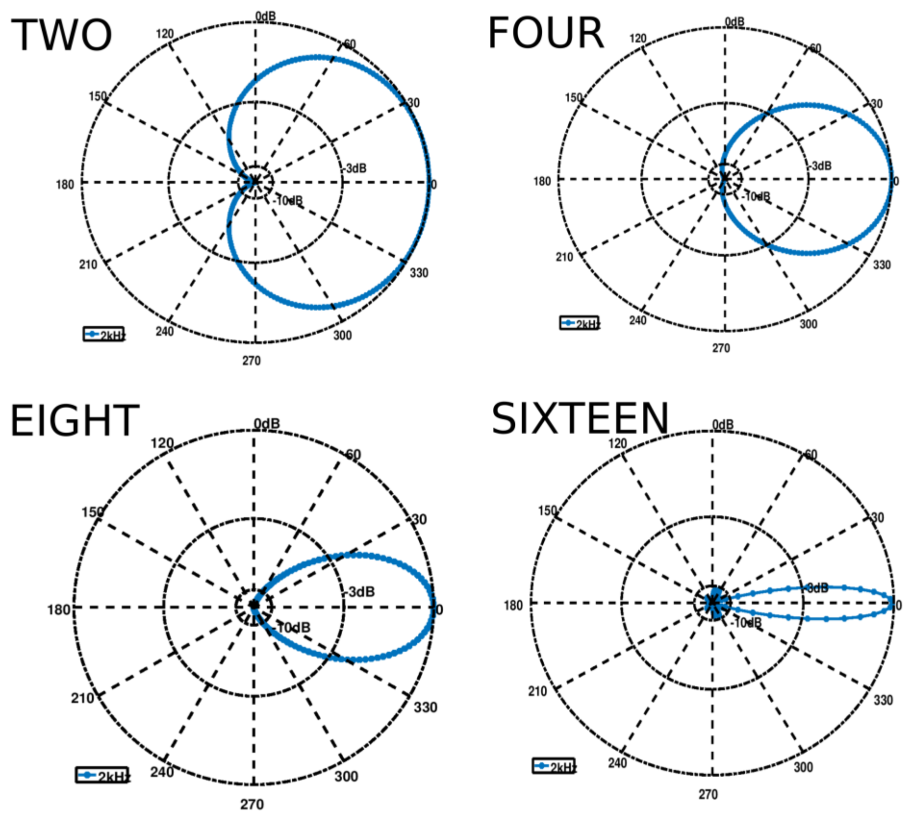
- Application
  - Front-end for speech recognition
  - Human computer interaction
  - Remote meeting system
  - Singer voice evaluation
  - Spatial sound reproduction



Introduction

Acoustic based approaches:

- Directional microphone
- Multiple microphone arrays
    - Beam-forming

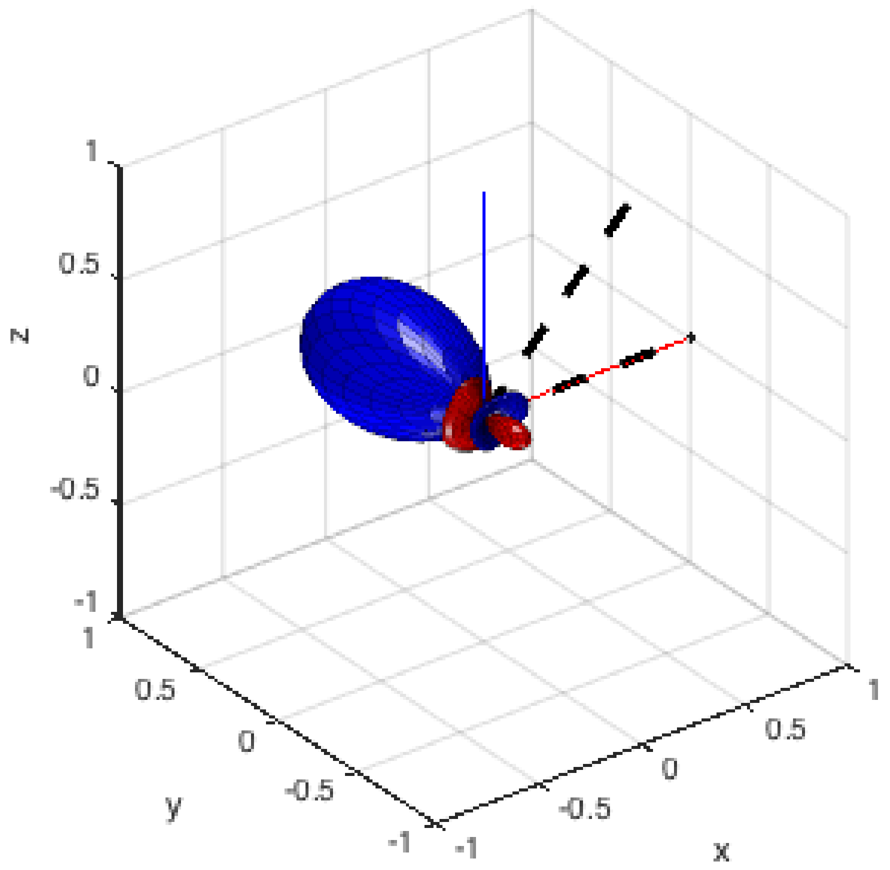


2-d beam patterns using linear equal spaced microphone arrays,
Adapted from *VOCAL Technologies, Ltd.*

Introduction

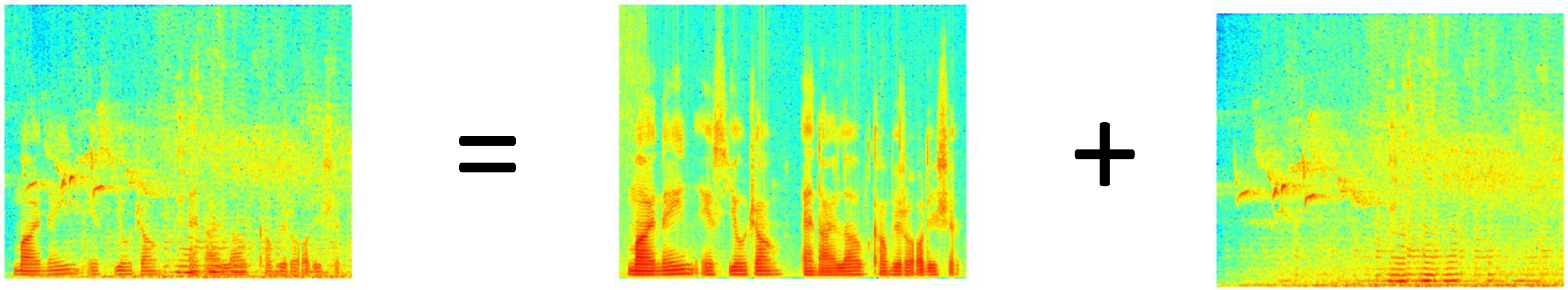
Acoustic based approaches:

- 2D & 3D Soundfield Beamforming
  - the sound field on the horizontal plane & spherical surface is sampled discretely
  - the samples are weighted and combined smartly to keep the sound from desired directions, but suppress the sounds from all other directions



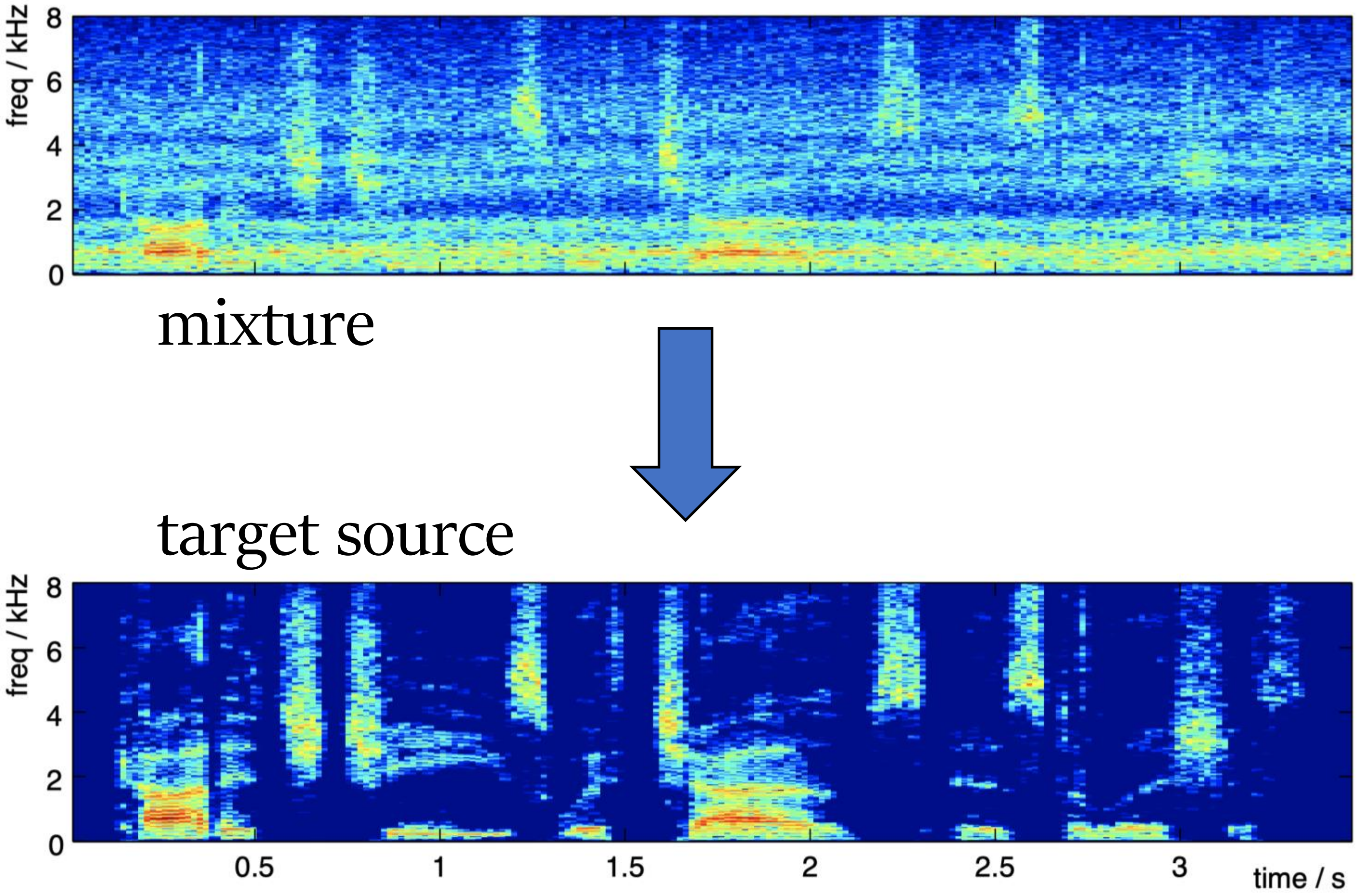
3-d beam patterns using 32-unit spherical microphone array, Adapted from Archontis Politis et al.

Spectrogram-based methods

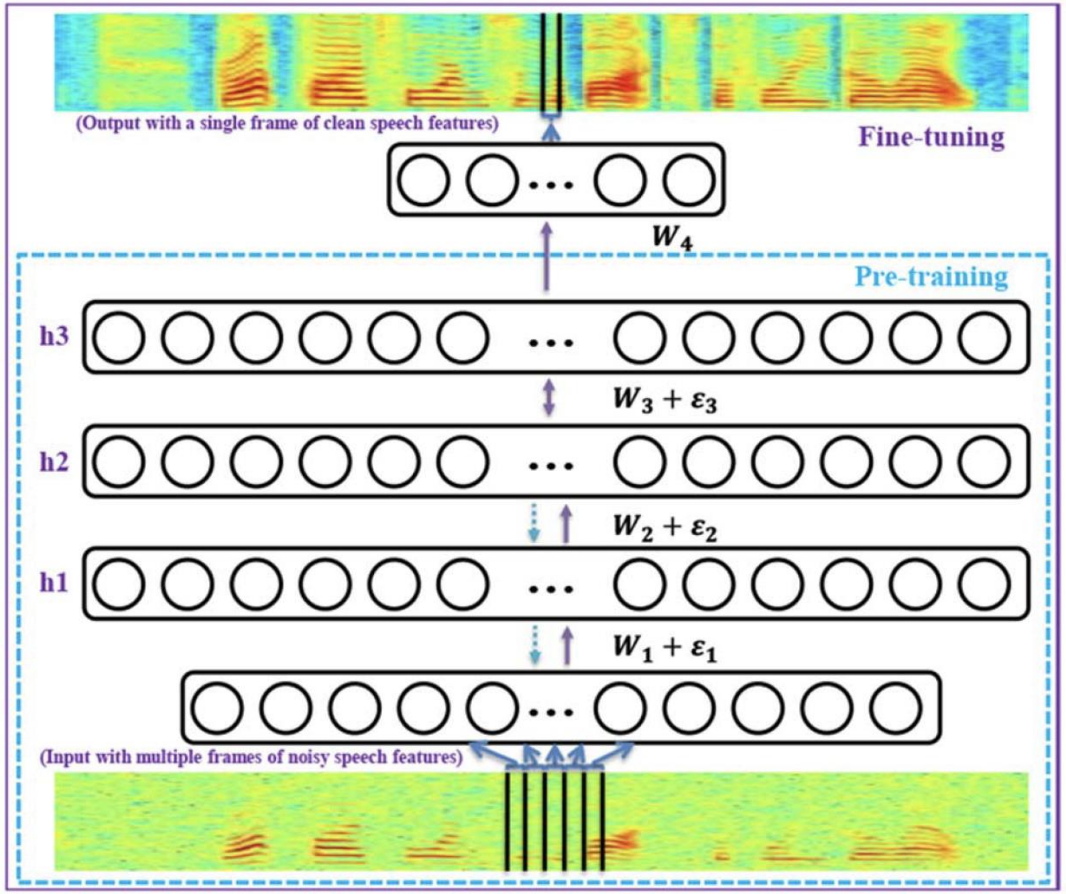


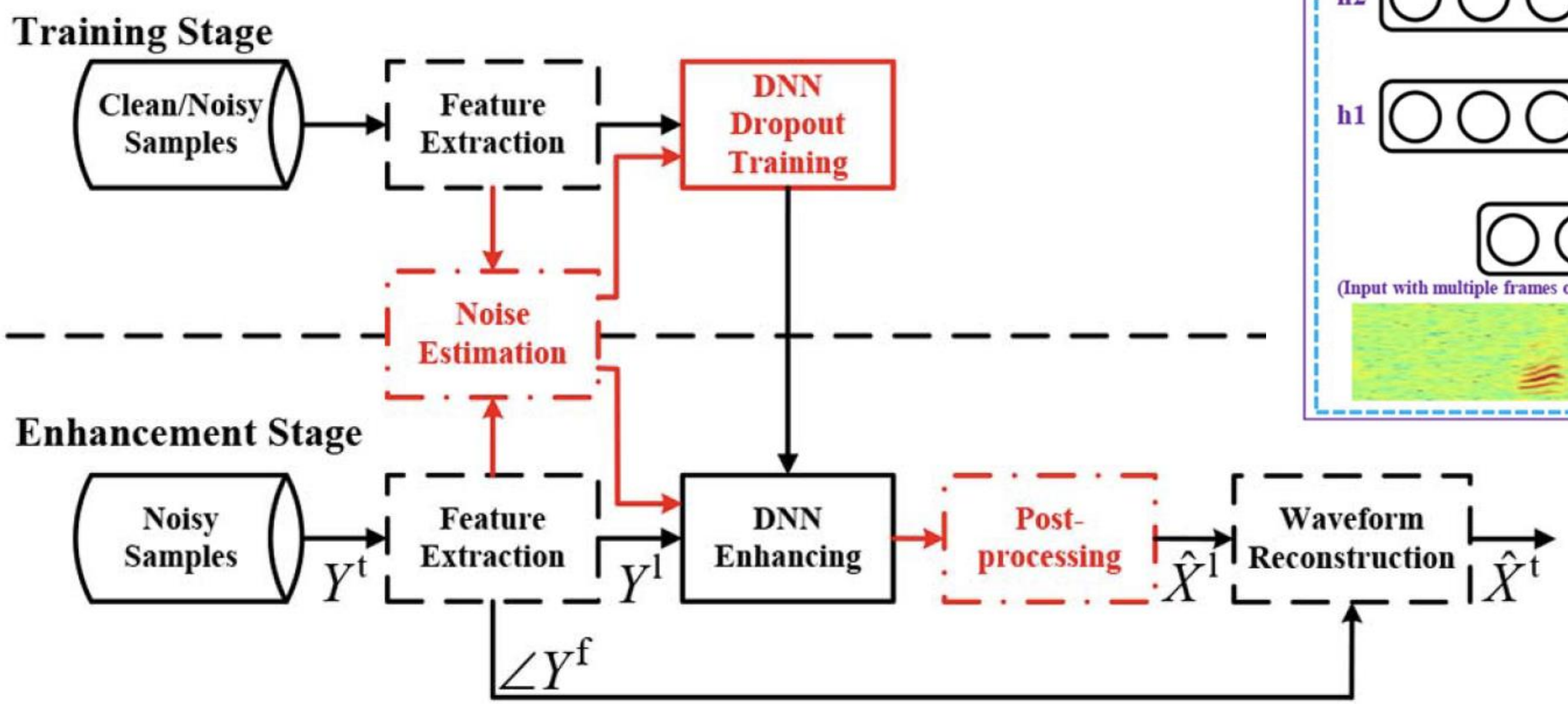
Mapping-based Methods

- learn a **regression function** from a mixed signal to target source directly



Speech Enhancement DNN [Xu et at., 2014]

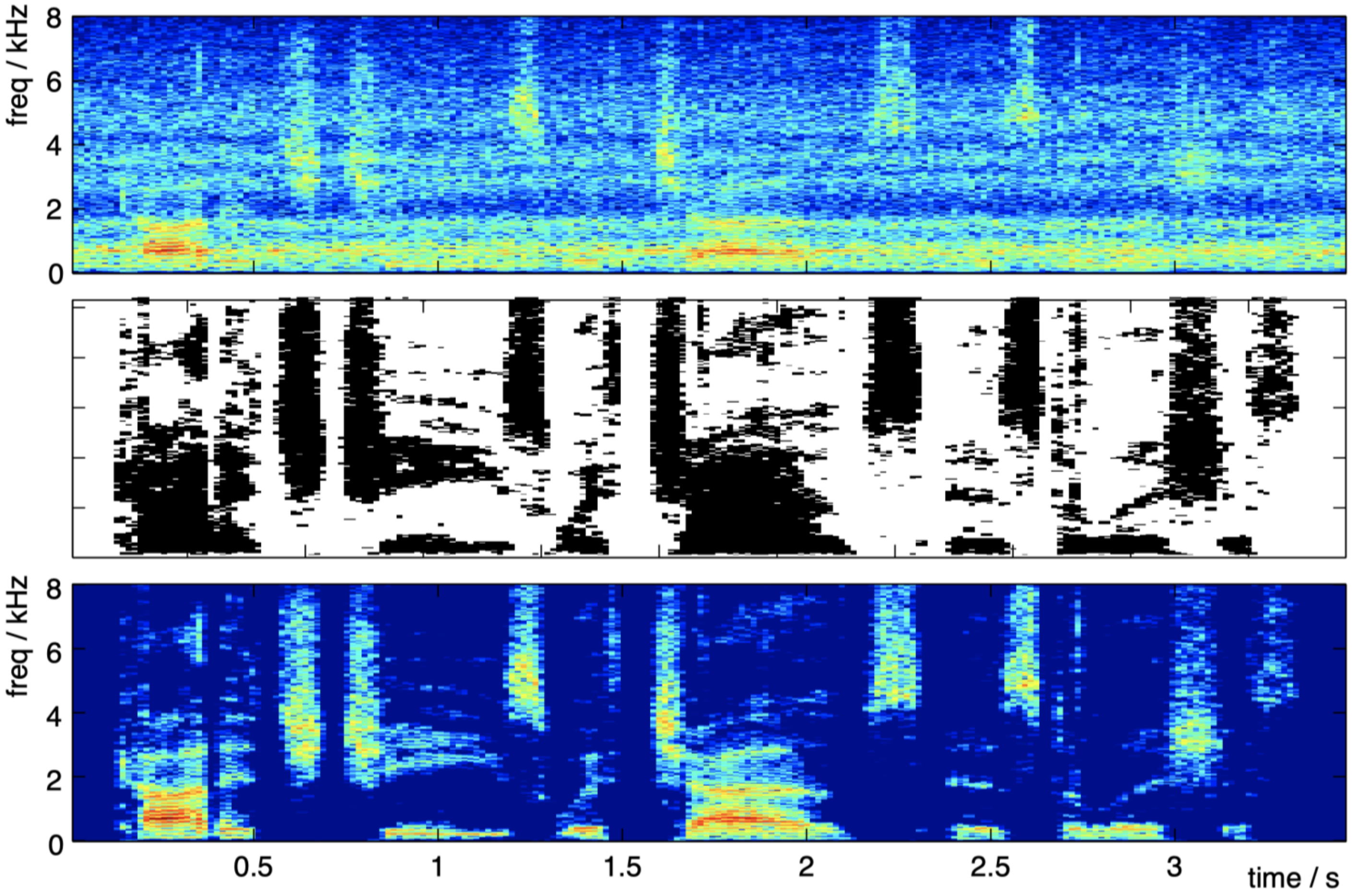




Mask-based Methods

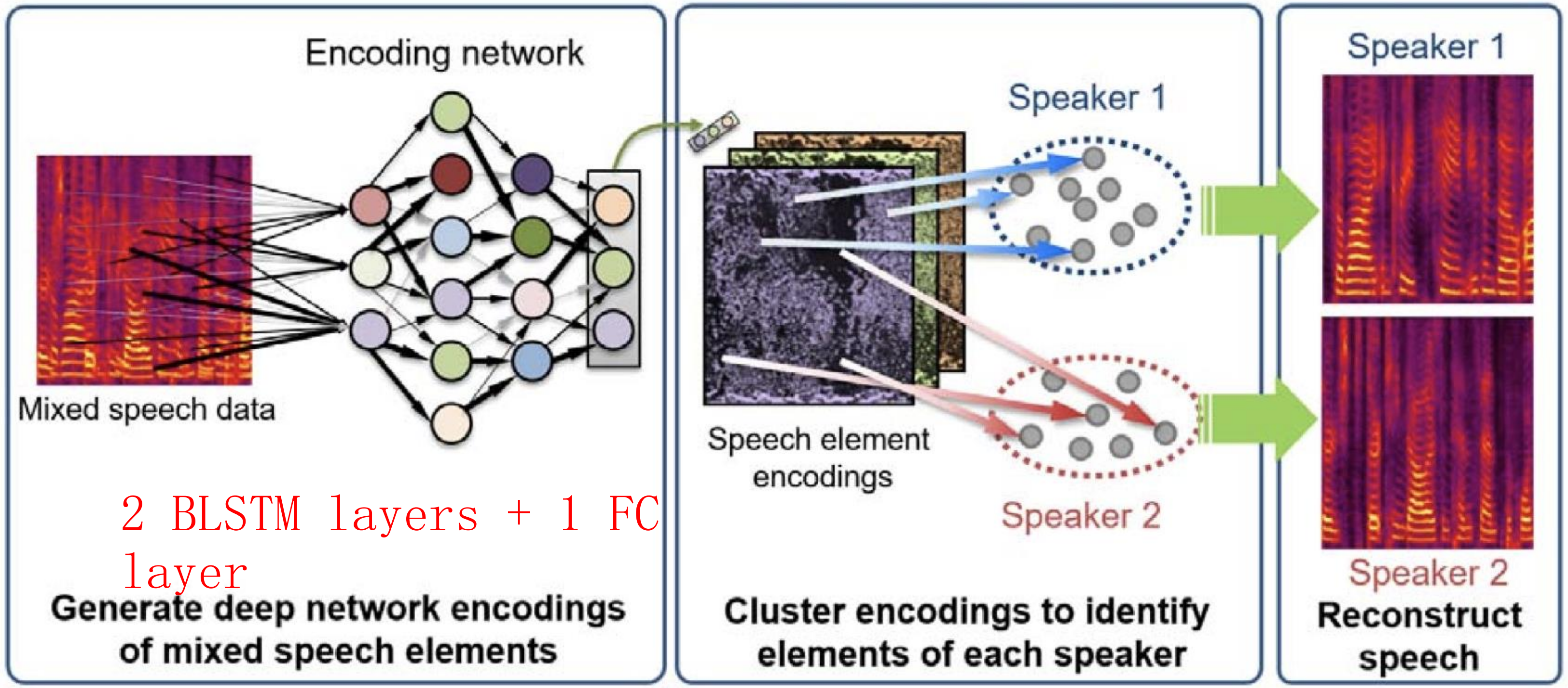
- ideal binary mask (**IBM**) : a T-F unit is assigned 1, if the signal-to-noise ratio (SNR) within the unit exceeds a local criterion, indicating **target dominance**

- ideal ratio mask (**IRM**): a T-F unit is assigned some **ratio of target** energy and mixture energy



Deep Clustering (DPCL) [Hershey et al., 2016]

For T-F units dominated by the same speaker, their embedding vectors are close to one another.



2 BLSTM layers + 1 FC layer

Train

Embedding network

$V = f_\theta(x)$ $\mathbf{V} \in \mathbb{R}^{TF \times D}$

Label indicator mat

$\mathbf{Y} \in \mathbb{R}^{TF \times C}$

Train to minimize $\mathcal{L}_{\mathrm{DC}} = ||\hat{\mathbf{A}} - \mathbf{A}||_F^2 = ||\mathbf{V}\mathbf{V}^T - \mathbf{Y}\mathbf{Y}^T||_F^2$

[1] [Availble Online] https://www.mitsubishielectric.com/news/2017/0524-e.html

Deep Clustering (DPCL) [Hershey et al., 2016]

Test

$V = f_\theta(x)$ Embedding matrix

Cluster the rows of *V* by k-means

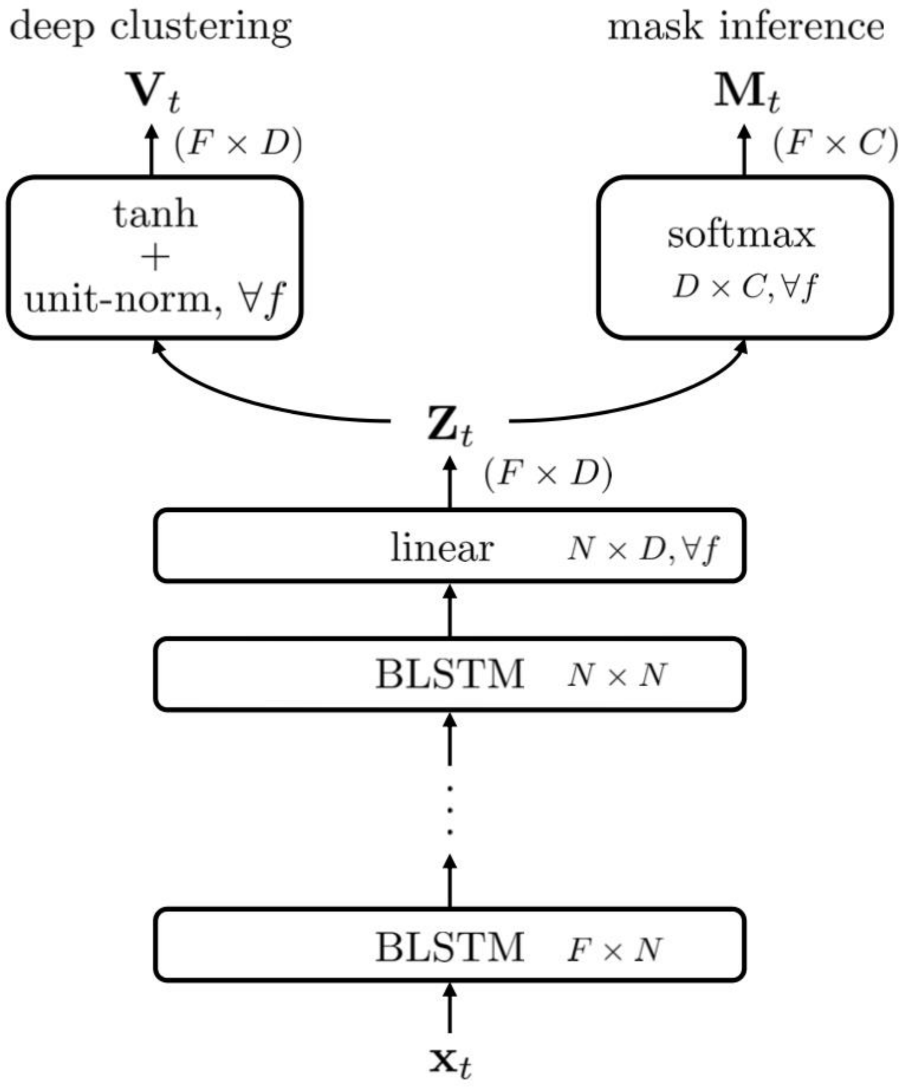
Binary masks $\bar{Y}$

X Complex spectrogram of the mixture

Estimated sources iSTFT

$\boldsymbol{s_1, s_2, \ldots}$

Chimera [Luo et al., 2017]



$$\mathcal{L}_{\text{DC}} = ||\hat{\mathbf{A}} - \mathbf{A}||_F^2 = ||\mathbf{V}\mathbf{V}^T - \mathbf{Y}\mathbf{Y}^T||_F^2.$$

$$\mathcal{L}_{\text{MSA}} = \sum_c ||\mathbf{R}^{(c)} - \mathbf{M}^{(c)} \odot \mathbf{S}||_2^2$$

masked magnitude spectrum approximation (mMSA)

$$\mathcal{L}_{\text{mMSA}} = \sum_c ||(\mathbf{O}^{(c)} - \mathbf{M}^{(c)}) \odot \mathbf{S}||_2^2$$

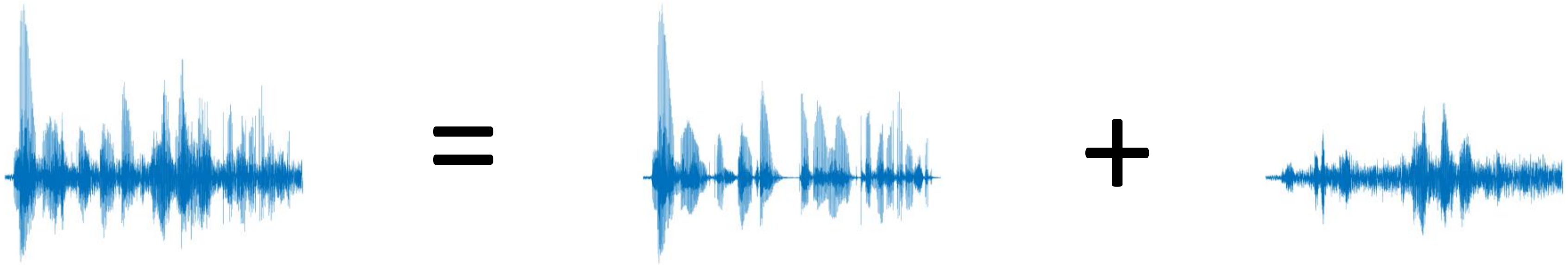
Multi-task learning

$$\mathcal{L}_{\text{CHI}} = \alpha \frac{\mathcal{L}_{\text{DC}}}{TF} + (1-\alpha)\mathcal{L}_{\text{MI}}$$

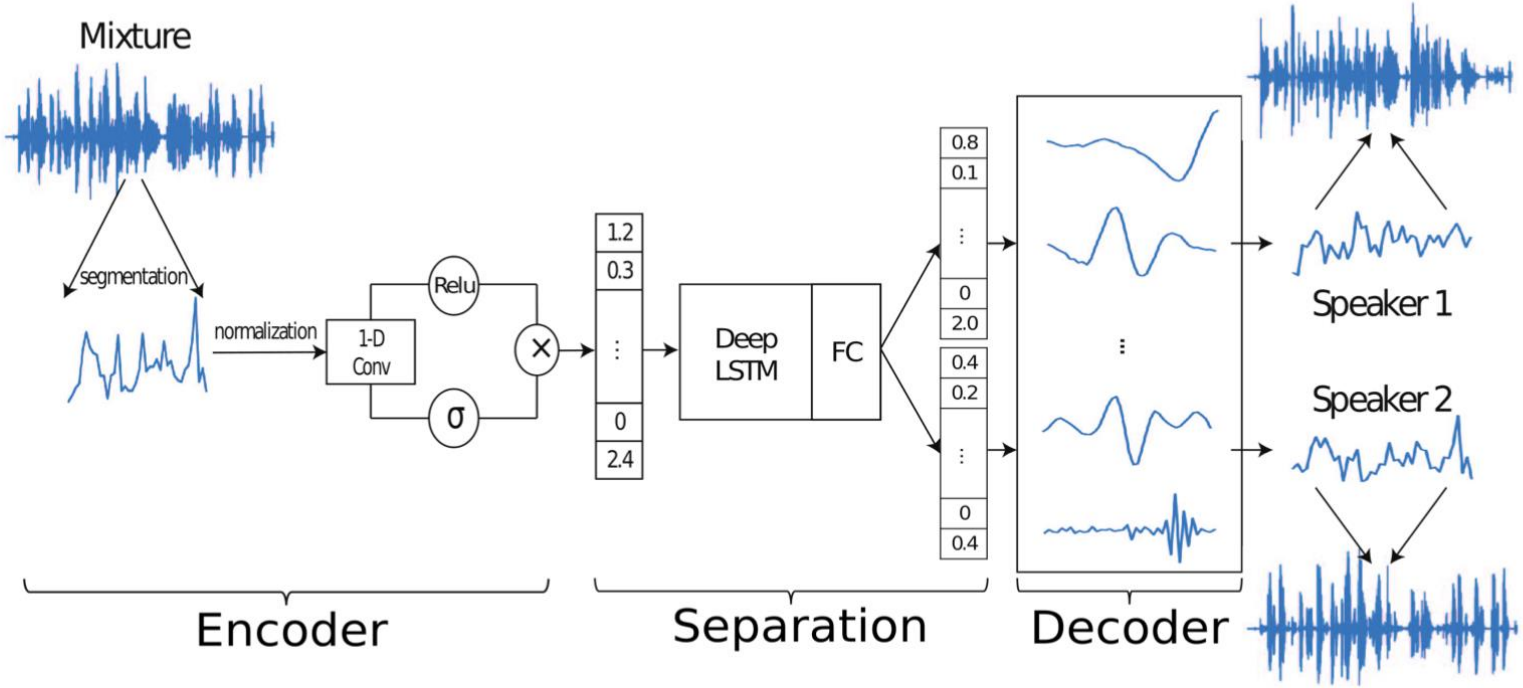
Shortcomings of spectrogram-based methods

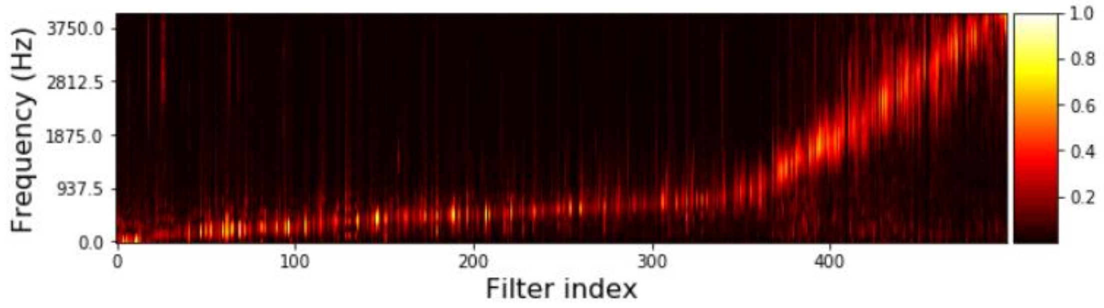
- STFT is **not** necessarily an **optimal transform**.
- Accurate reconstruction of the **phase**.
- Requires a high-resolution frequency decomposition of the mixture signal, **increases the minimum latency** of the system.

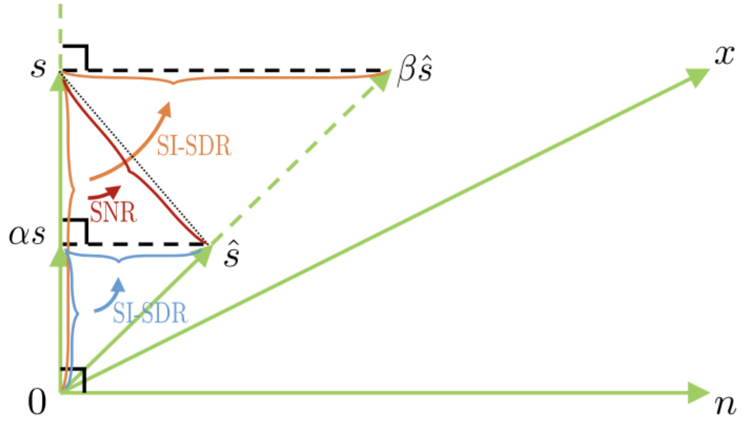
Waveform-based methods



Time-domain audio separation network (TasNet) [Luo & Mesgarani, 2018]



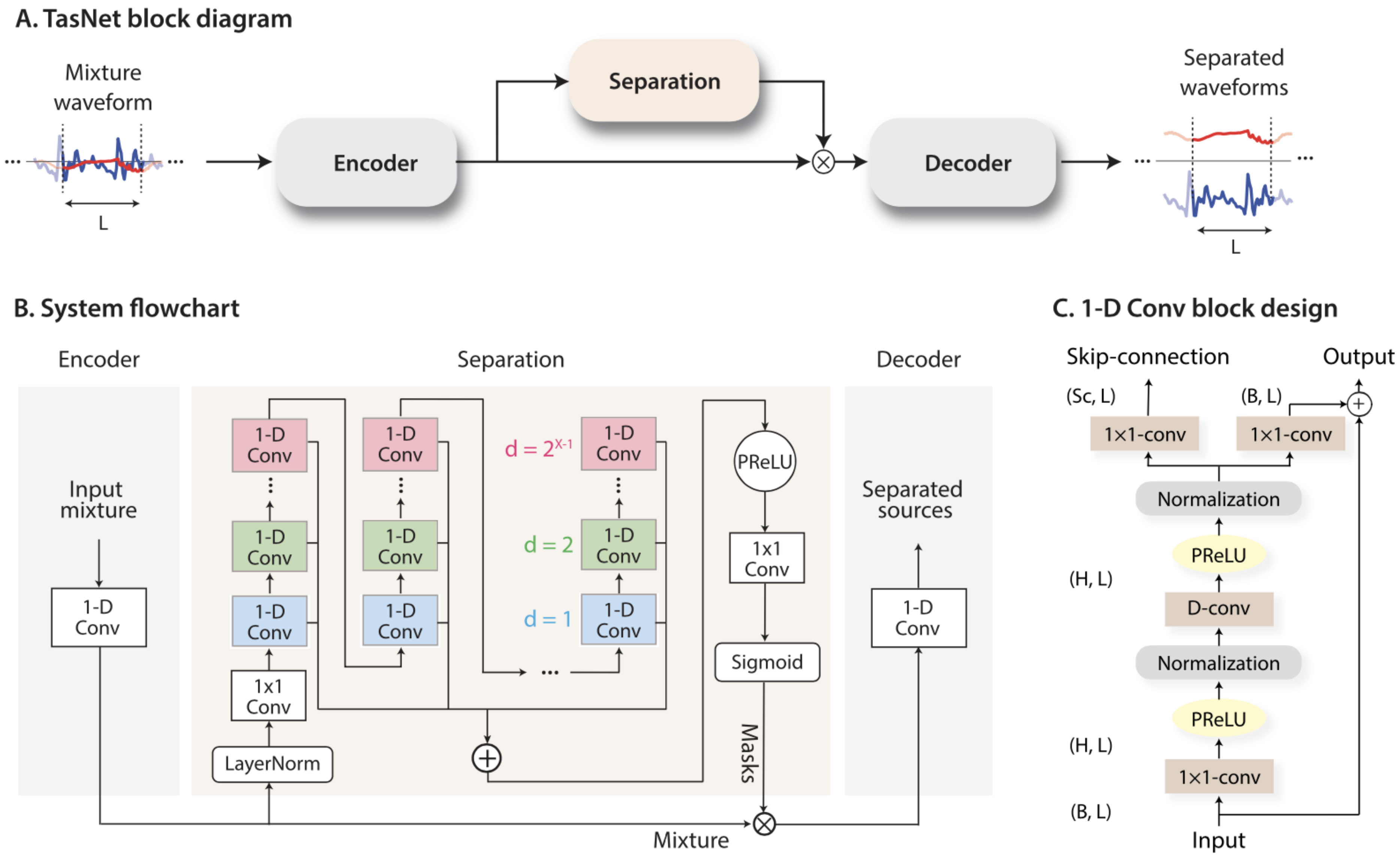




$$\begin{cases} \mathbf{s}_{target} := \frac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s}}{\|\mathbf{s}\|^2} \\ \mathbf{e}_{noise} := \hat{\mathbf{s}} - \mathbf{s}_{target} \\ \text{SI-SNR} := 10 \, log_{10} \frac{\|\mathbf{s}_{target}\|^2}{\|\mathbf{e}_{noise}\|^2} \end{cases}$$

Training objective : scale-invariant SNR (SI-SNR / SI-SDR)

Conv-TasNet [Luo & Mesgarani, 2019]



Pros & Cons

- Pros: high accuracy, short latency, small model size
  - Compared to STFT approach, replaces it with convolutional encoder-decoder architecture
  - Conv-TasNet has a smaller model size and a shorter minimum latency
- Limitation:
  - Long term tracking of the speakers may be compromised
  - Performance in lower SNR & distortion cases may fail
  - Importance of information at difference frequency regions remains un-explored.

Future directions

- Multi channel process

- Noise robustness architecture

- Speaker of interest is large(>3)

Thank you!

References:

[1] Y. Xu, J. Du, L. Dai, and C. Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 23, no. 1, pp. 7-19, 2015, doi: 10.1109/TASLP.2014.2364452.

[2] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016: IEEE, pp. 31-35.

[3] X.-L. Zhang and D. Wang, "A Deep Ensemble Learning Method for Monaural Speech Separation," (in eng), *IEEE/ACM transactions on audio, speech, and language processing,* vol. 24, no. 5, pp. 967-977, 2016, doi: 10.1109/TASLP.2016.2536478.

[4] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017: IEEE, pp. 61-65.

[5] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018: IEEE, pp. 696-700.

[6] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185,* 2018.

[7] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 26, no. 10, pp. 1702-1726, 2018.

[8] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 27, no. 8, pp. 1256-1266, 2019.