# Multi-channel Source Separation 2

안인규 (Inkyu An)

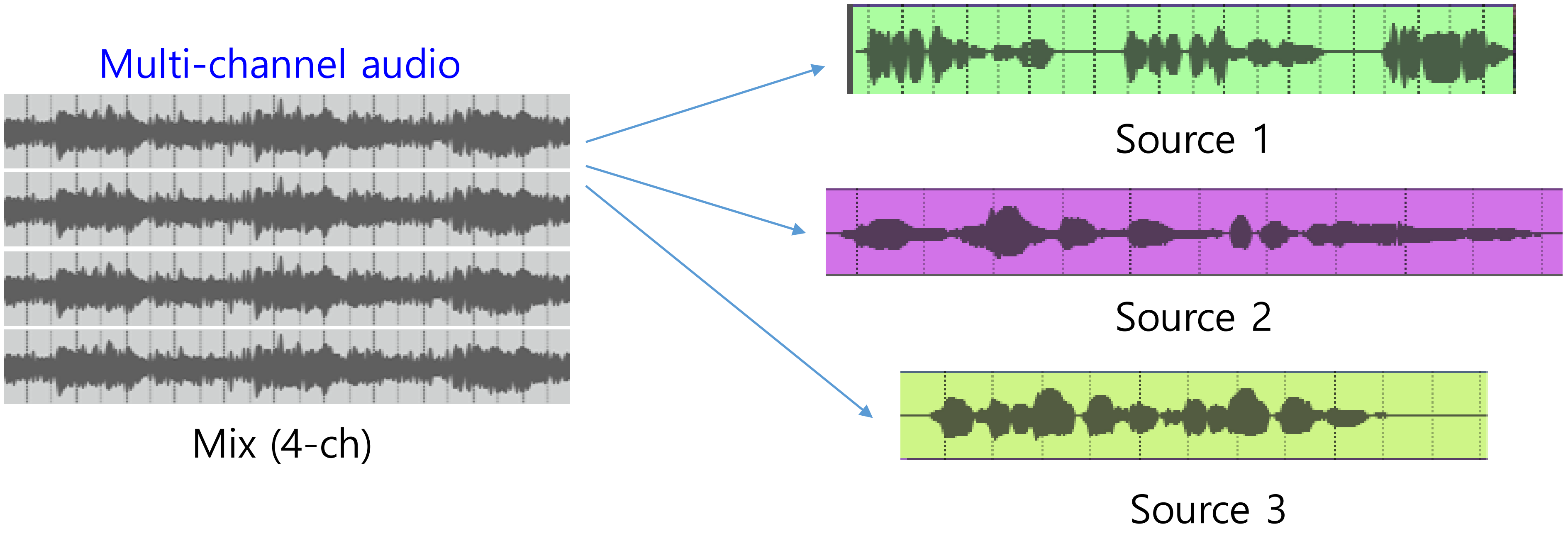
**Speech And Audio Recognition (오디오 음성인식)**

https://mairlab-km.github.io/

This lecture material refers to https://github.com/yandexdataschool/speech_course?tab=readme-ov-file and https://github.com/markovka17/dla

# Multi-channel Source Separation

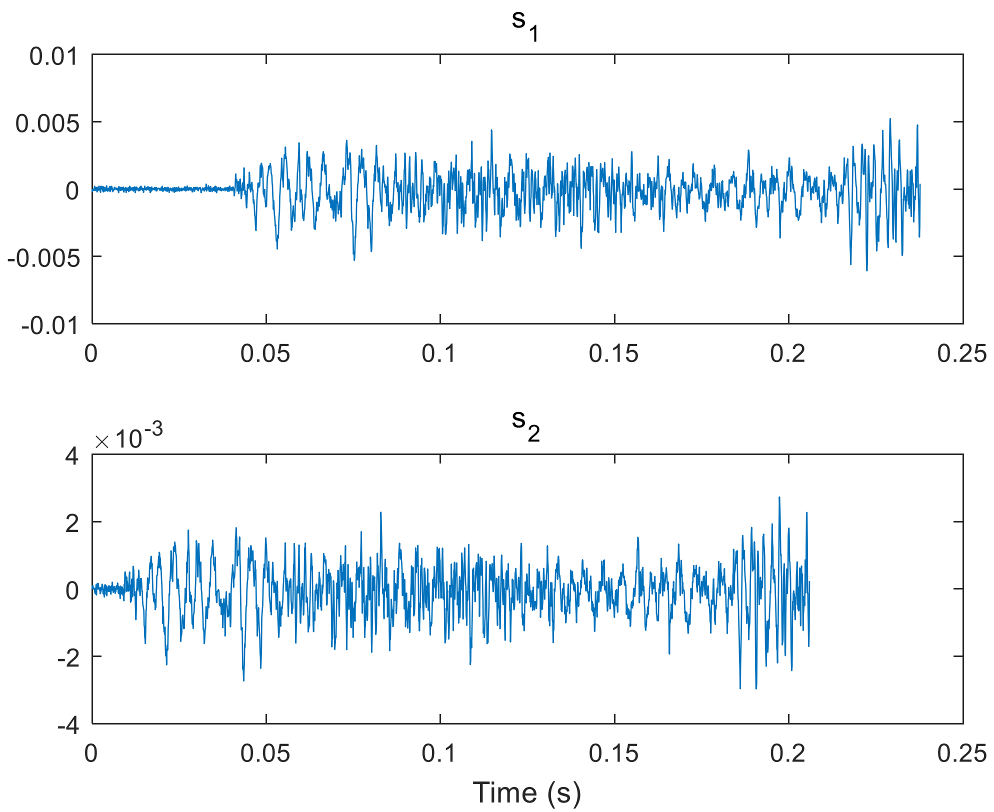
- **Goal**: extract K sources from the noisy mixture of <u>multi-channel audio</u>

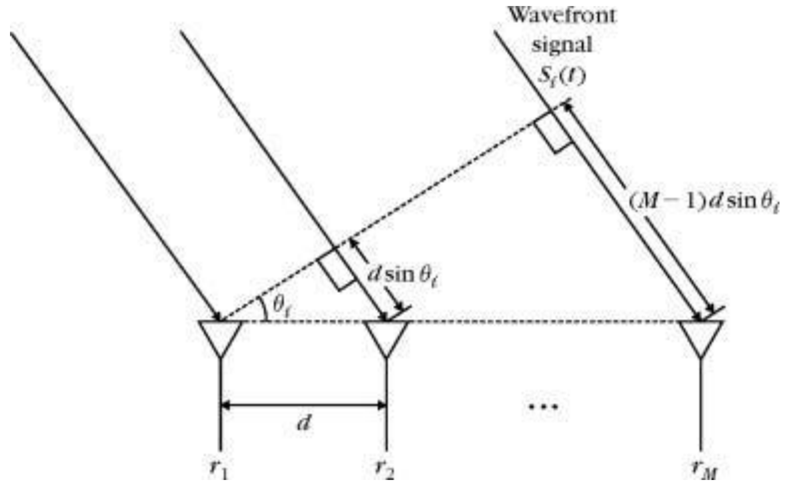



- How can we measure the multi-channel audio signal?

# Multi-channel Source Separation

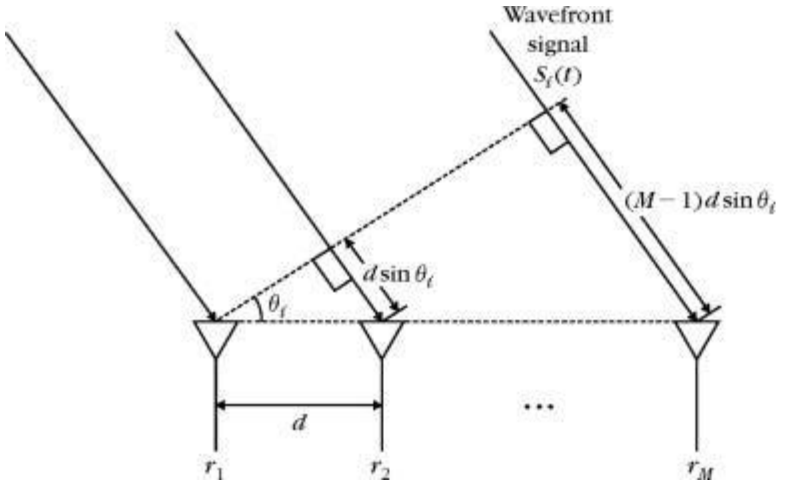
- Microphone Array
  - Consider the Uniform Linear Array
  - What's different about the audio collected at each microphone?

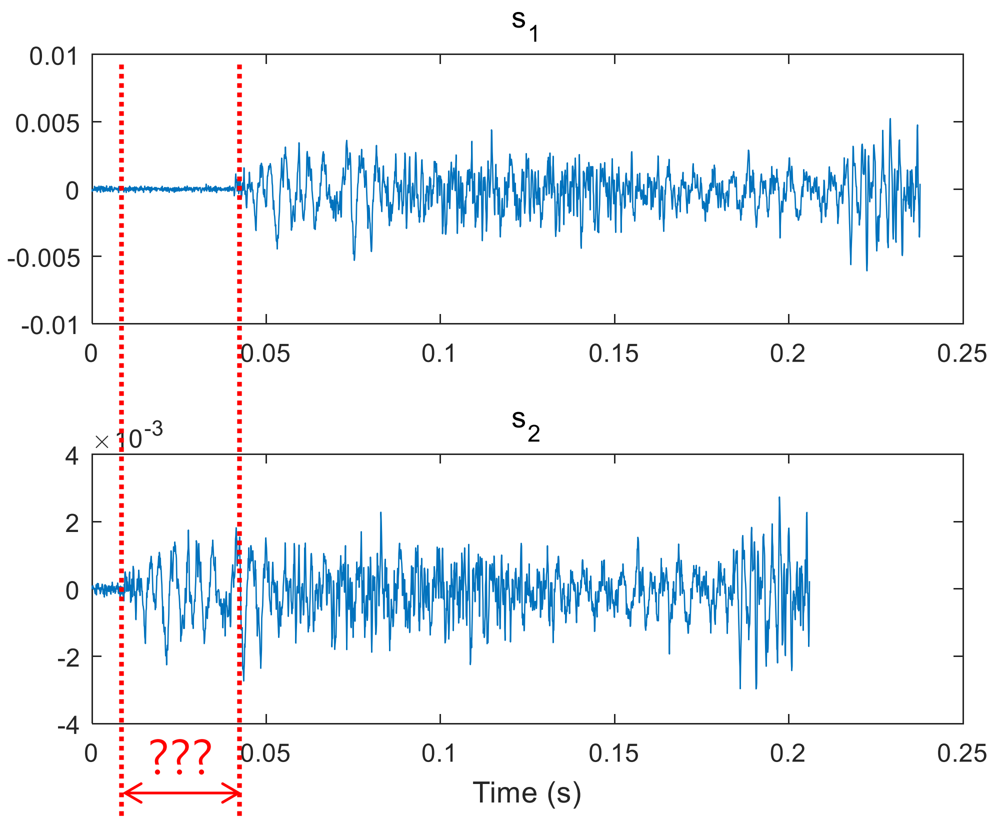


# Multi-channel Source Separation

- Microphone Array
  - Consider the Uniform Linear Array
  - What's different about the audio collected at each microphone?



# Multi-channel Source Separation

- We have to know the direction-of-arrival (DoA) of the sound

$$y(\boldsymbol{\theta}, \boldsymbol{t}) = \boldsymbol{w}(\boldsymbol{\theta}, -\boldsymbol{t})^{\boldsymbol{H}} \underset{\text{Convolution}}{*} \boldsymbol{x}(\boldsymbol{t}) \quad \overset{}{\underset{\text{FFT \& iFFT}}{\Longleftrightarrow}} \quad \underbrace{y(\boldsymbol{\theta}, \boldsymbol{f})}_{\text{Beamforming output}} = \underbrace{\boldsymbol{w}(\boldsymbol{\theta}, \boldsymbol{f})^{\boldsymbol{H}}}_{\text{Weight (Spatial filter)}} \underbrace{\boldsymbol{x}(\boldsymbol{f})}_{\text{Measured pressure signals}}$$

- However, It is have to find out the accurate DoA...

# Problem Definition

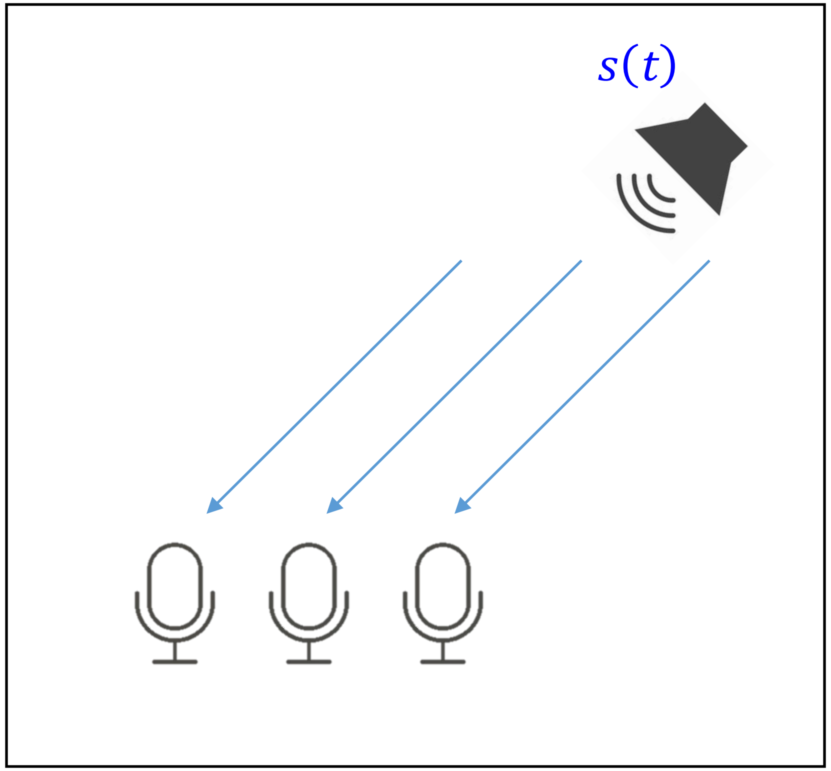
- Problem Definition
  - 마이크에서 녹음되는 신호는 어떻게 정의할 수 있을까?
  - Problem definition:

$$\boldsymbol{y}(t) = \begin{bmatrix} y_1(t) \\ \cdots \\ y_M(t) \end{bmatrix} = \boldsymbol{x}(t) + \boldsymbol{n}(t)$$

  - $\boldsymbol{y}(t)$: observation signals of M microphones
  - $\boldsymbol{x}(t)$: original signal of M microphones
  - $\boldsymbol{n}(t)$: noise of M microphones

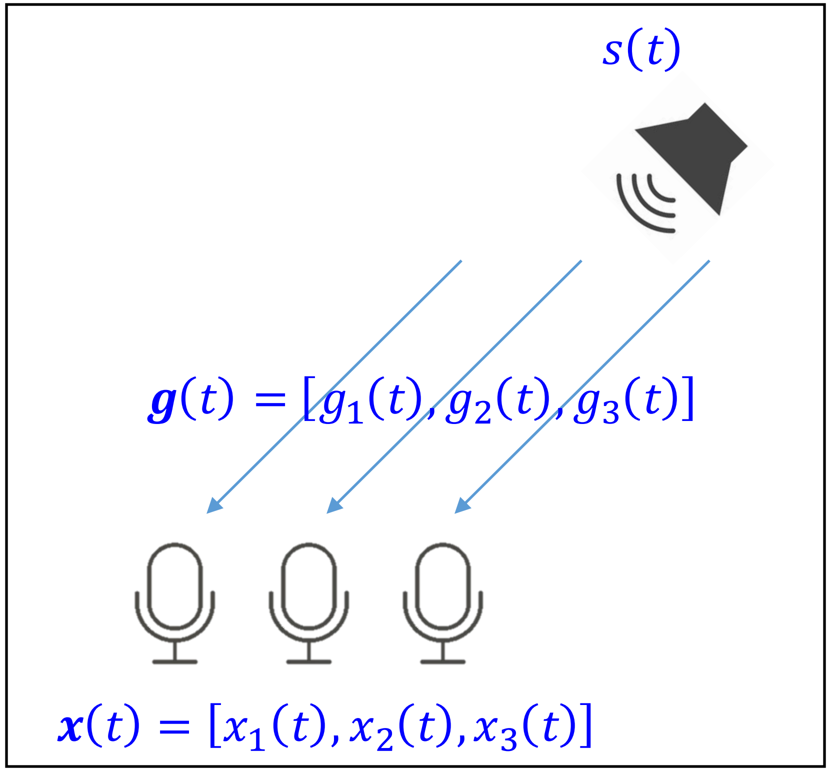
# Problem Definition

- Problem Definition
  - 마이크에서 녹음되는 신호는 어떻게 정의할 수 있을까?
  - Problem definition: $$\boldsymbol{y}(t) = \begin{bmatrix} y_1(t) \\ \cdots \\ y_M(t) \end{bmatrix} = \boldsymbol{x}(t) + \boldsymbol{n}(t)$$
  - $\boldsymbol{y}(t)$: observation signals at M microphones
  - $\boldsymbol{x}(t)$: original signal at M microphones
  - $\boldsymbol{n}(t)$: noise of M microphones



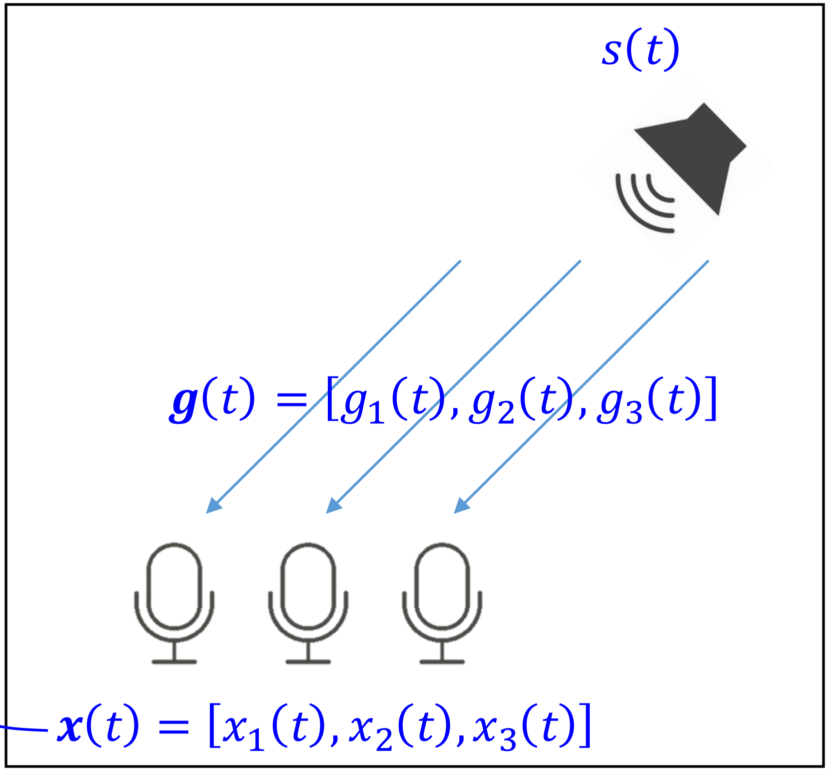
# Problem Definition

- Problem Definition
  - 마이크에서 녹음되는 신호는 어떻게 정의할 수 있을까?
  - Problem definition: $$\boldsymbol{y}(t) = \begin{bmatrix} y_1(t) \\ \cdots \\ y_M(t) \end{bmatrix} = \boldsymbol{x}(t) + \boldsymbol{n}(t)$$
  - $\boldsymbol{y}(t)$: observation signals at M microphones
  - $\boldsymbol{x}(t)$: original signal at M microphones
  - $\boldsymbol{n}(t)$: noise of M microphones



# Problem Definition

- Problem Definition
  - 마이크에서 녹음되는 신호는 어떻게 정의할 수 있을까?
  - Problem definition: $\boldsymbol{y}(t) = \begin{bmatrix} y_1(t) \\ \cdots \\ y_M(t) \end{bmatrix} = \underline{\boldsymbol{x}(t)} + \boldsymbol{n}(t)$




  - $\boldsymbol{y}(t)$: observation signals at M microphones
  - $\boldsymbol{x}(t)$: original signal at M microphones
  - $\boldsymbol{n}(t)$: noise of M microphones

# Problem Definition

- Wiener Filter
  - Linear filter: $\hat{x}(t) = \boldsymbol{w}^H * y(t)$
  - Given, $\boldsymbol{y}(t) = \begin{bmatrix} y_1(t) \\ \cdots \\ y_M(t) \end{bmatrix} = \boldsymbol{x}(t) + \boldsymbol{n}(t) \xrightarrow[\text{FFT}]{} \boldsymbol{Y}(f) = \boldsymbol{X}(f) + \boldsymbol{N}(f)$

  - Goal1: design a spatial filter $w(t)$ <u>to minimize the response for noise</u>
    - noise를 줄여서 $\boldsymbol{x}(t)$를 복원하도록 $w(t)$를 설계
    - 평균 제곱 오차 (Mean Square Error, MSE)를 최소화하는 $w(t)$를 계산

$$J(\boldsymbol{w}) = E[|y(t) - \boldsymbol{w}^H * \boldsymbol{y}(t)|^2] \xrightarrow[\text{FFT}]{} J(\boldsymbol{w}) = E[|Y(f) - \boldsymbol{W}^H * \boldsymbol{Y}(f)|^2]$$

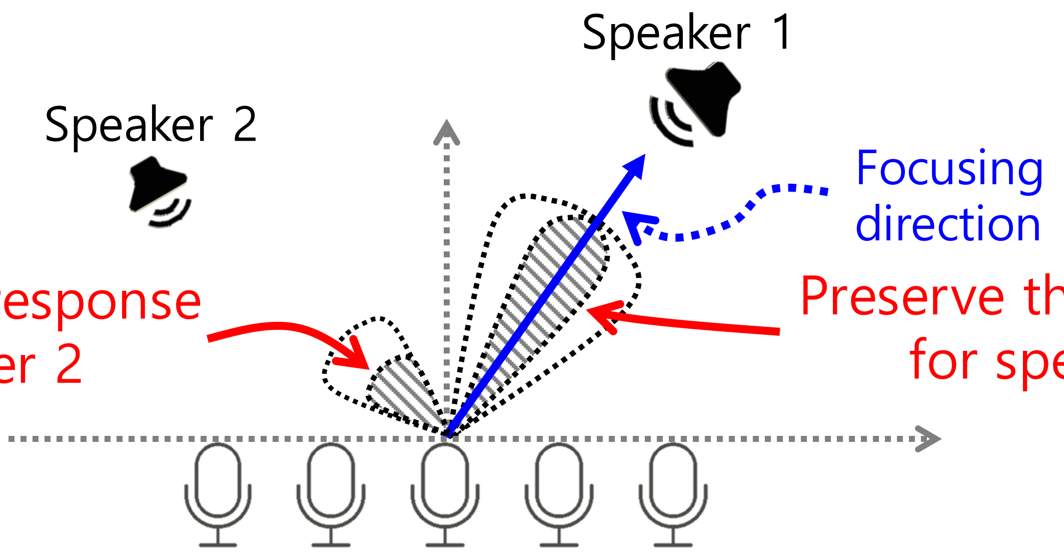
# Multi-channel Source Separation

- MVDR beamformer
  - Design a spatial filter $w$ that minimizes the total beamforming power but maintain the response in the aiming angle

$$\textbf{Minimize } \boldsymbol{\beta}(\boldsymbol{\theta}) = \boldsymbol{w}(\boldsymbol{\theta})^H \boldsymbol{R} \boldsymbol{w}(\boldsymbol{\theta}) \qquad \textbf{Subject to } \boldsymbol{w}(\boldsymbol{\theta})^H \mathbf{h}(\boldsymbol{\theta}) = \boldsymbol{b}_0$$

$$\xrightarrow{\text{solve}} \boldsymbol{w}(\boldsymbol{\theta}) = \frac{\boldsymbol{R}^{-1}\mathbf{h}(\boldsymbol{\theta})}{\mathbf{h}(\boldsymbol{\theta})^H \boldsymbol{R}^{-1}\mathbf{h}(\boldsymbol{\theta})} \boldsymbol{b}_0^*$$

Minimize the response for speaker 2




Preserve the response for speaker 1

# Problem Definition

- Wiener Filter-based MVDR beamformer
  - Linear filter: $\hat{x}(t) = \boldsymbol{w}^H * y(t)$
  - Given, $\boldsymbol{y}(t) = \begin{bmatrix} y_1(t) \\ \cdots \\ y_M(t) \end{bmatrix} = \boldsymbol{x}(t) + \boldsymbol{n}(t) \xrightarrow[\text{FFT}]{} \boldsymbol{Y}(f) = \boldsymbol{X}(f) + \boldsymbol{N}(f)$

  - Goal1: design a spatial filter $w(t)$ <u>to minimize the response for noise</u>
    - noise를 줄여서 $\boldsymbol{x}(t)$를 복원하도록 $w(t)$를 설계
    - 평균 제곱 오차 (Mean Square Error, MSE)를 최소화하도록 $w(t)$를 설계

$$J(\boldsymbol{w}) = E[|y(t) - \boldsymbol{w}^H * \boldsymbol{y}(t)|^2] \xrightarrow[\text{FFT}]{} J(\boldsymbol{w}) = E[|Y(f) - \boldsymbol{W}^H * \boldsymbol{Y}(f)|^2]$$

  - Goal2: Preserve the response for $x(t)$

$$E[|X(f) - \boldsymbol{W}^H * \boldsymbol{Y}(f)|^2] = G(f)$$

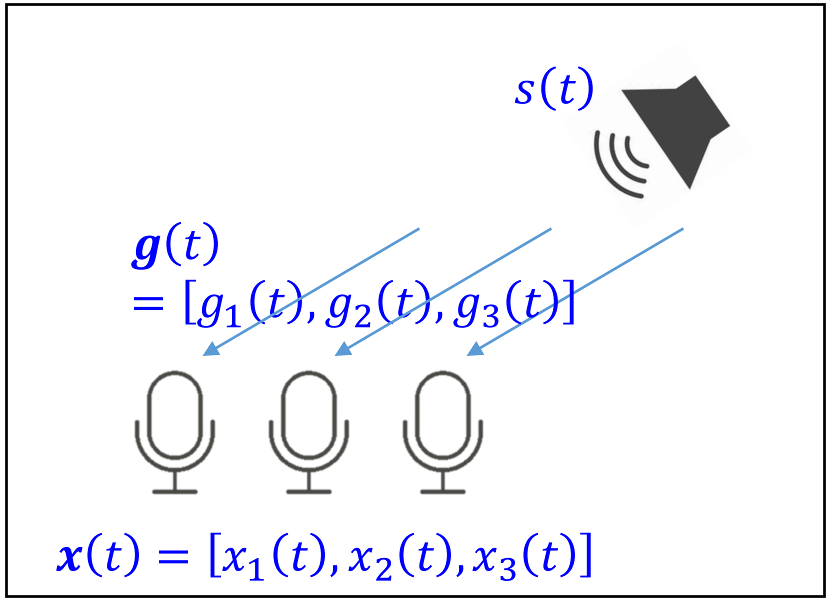
# Problem Definition

- Requirement of Wiener Filter-based MVDR beamformer
  - Goal1 + Goal2: design a spatial filter $w(t)$ to minimize the response for noise while preserving the response for $x(t)$

$$J(\boldsymbol{w}) = \min_{W} E[|Y(f) - \boldsymbol{W}^H * \boldsymbol{Y}(f)|^2]$$

subset to $E[|X(f) - \boldsymbol{W}^H * \boldsymbol{Y}(f)|^2] = G(f)$

$$\boldsymbol{w} = \frac{E[|\boldsymbol{N}(f)|^2]^{-1} \cdot E[|\boldsymbol{X}(f)|^2]}{\mathrm{Tr}(E[|\boldsymbol{N}(f)|^2]^{-1} \cdot E[|\boldsymbol{X}(f)|^2])} u_{n_0} = \frac{\Phi^N(f)^{-1} \cdot \Phi^s(f)}{\mathrm{Tr}\left(\Phi^N(f)^{-1} \cdot \Phi^s(f)\right)} u_{n_0}$$




$$\boldsymbol{Y}(f) = \begin{bmatrix} Y_1(f) \\ \cdots \\ Y_M(f) \end{bmatrix} = \boldsymbol{X}(f) + \boldsymbol{N}(f)$$

# Problem Definition

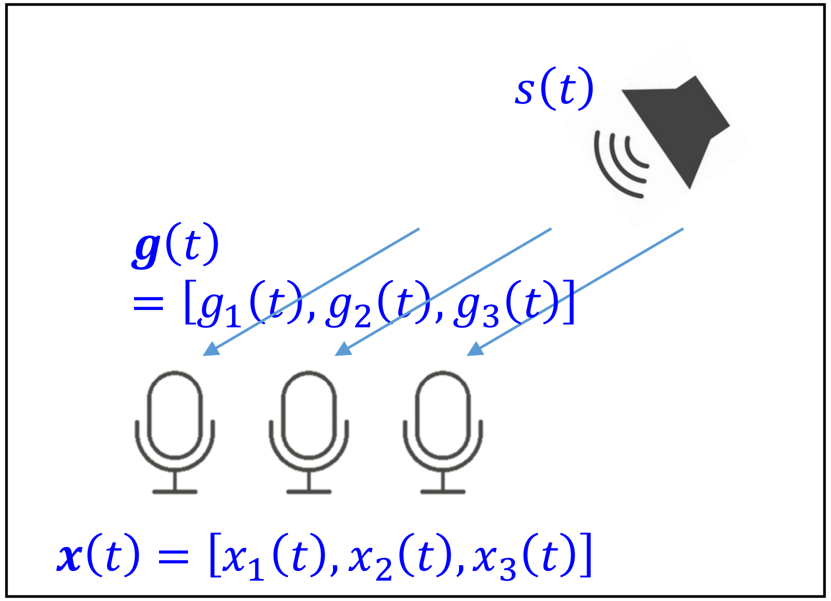
- Requirement of Wiener Filter-based MVDR beamformer
  - Goal1 + Goal2: design a spatial filter $w(t)$ to minimize the response for noise while preserving the response for $x(t)$

$$J(\boldsymbol{w}) = \min_{W} E[|Y(f) - \boldsymbol{W}^H * \boldsymbol{Y}(f)|^2]$$

$$\text{subset to } E[|X(f) - \boldsymbol{W}^H * \boldsymbol{Y}(f)|^2] = G(f)$$

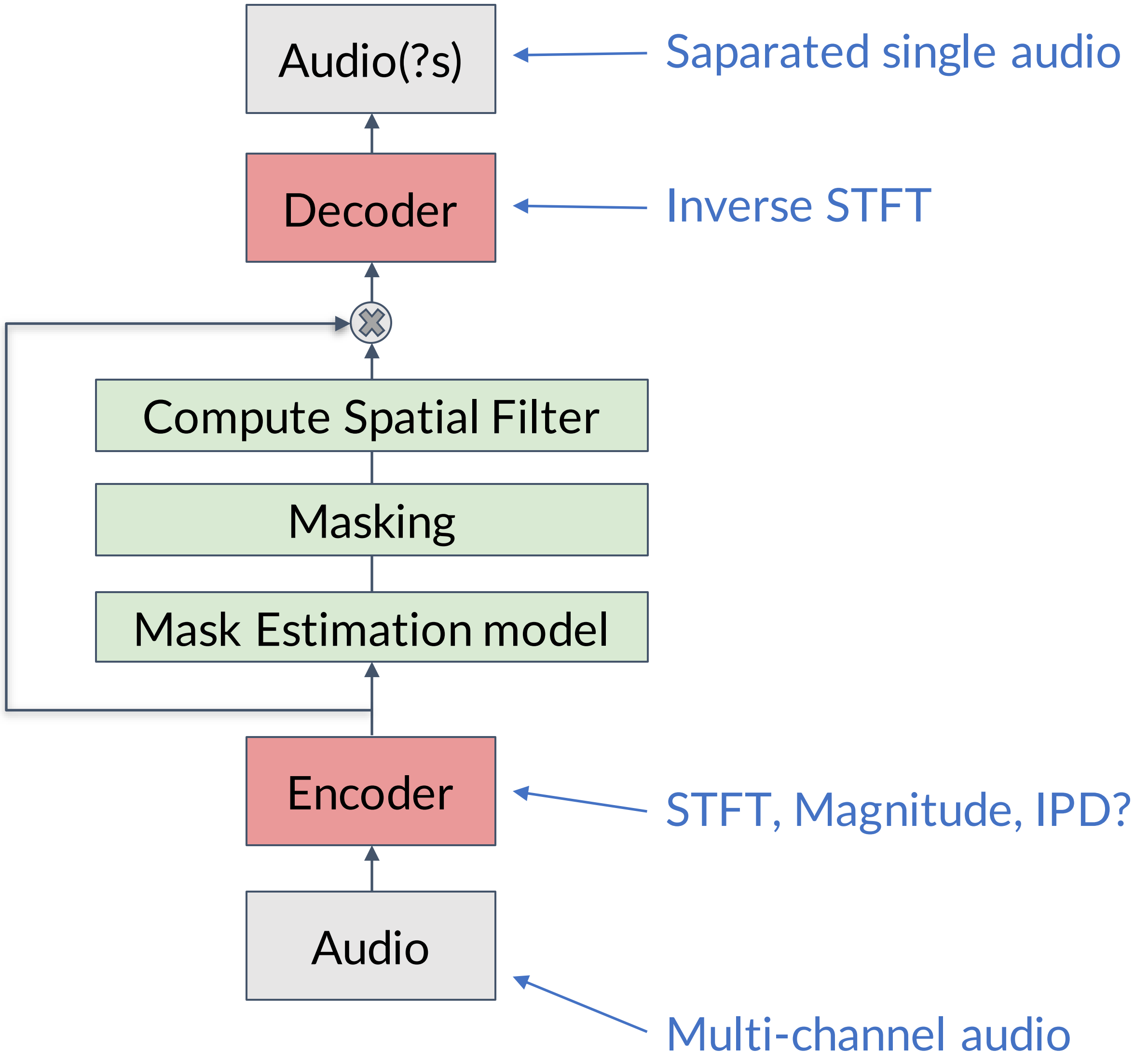
Original signal과 Noise signal을 알아야 한다.

$$\boldsymbol{w} = \frac{E[|\boldsymbol{N}(f)|^2]^{-1} \cdot E[|\boldsymbol{X}(f)|^2]}{\mathrm{Tr}(E[|\boldsymbol{N}(f)|^2]^{-1} \cdot E[|\boldsymbol{X}(f)|^2])} u_{n_0} = \frac{\Phi^N(f)^{-1} \cdot \Phi^s(f)}{\mathrm{Tr}\left(\Phi^N(f)^{-1} \cdot \Phi^s(f)\right)} u_{n_0}$$

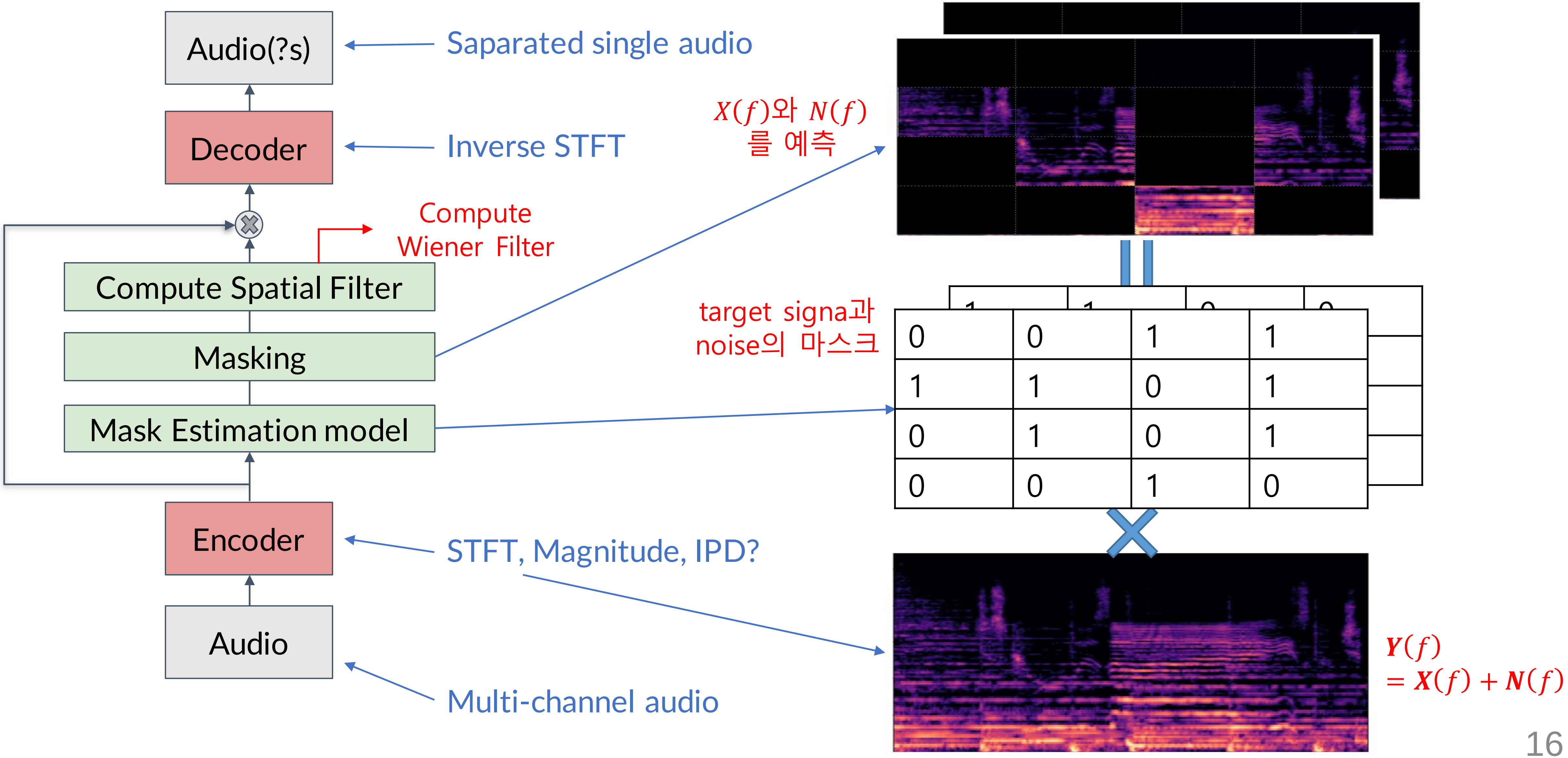



$$\boldsymbol{Y}(f) = \begin{bmatrix} Y_1(f) \\ \cdots \\ Y_M(f) \end{bmatrix} = \boldsymbol{X}(f) + \boldsymbol{N}(f)$$

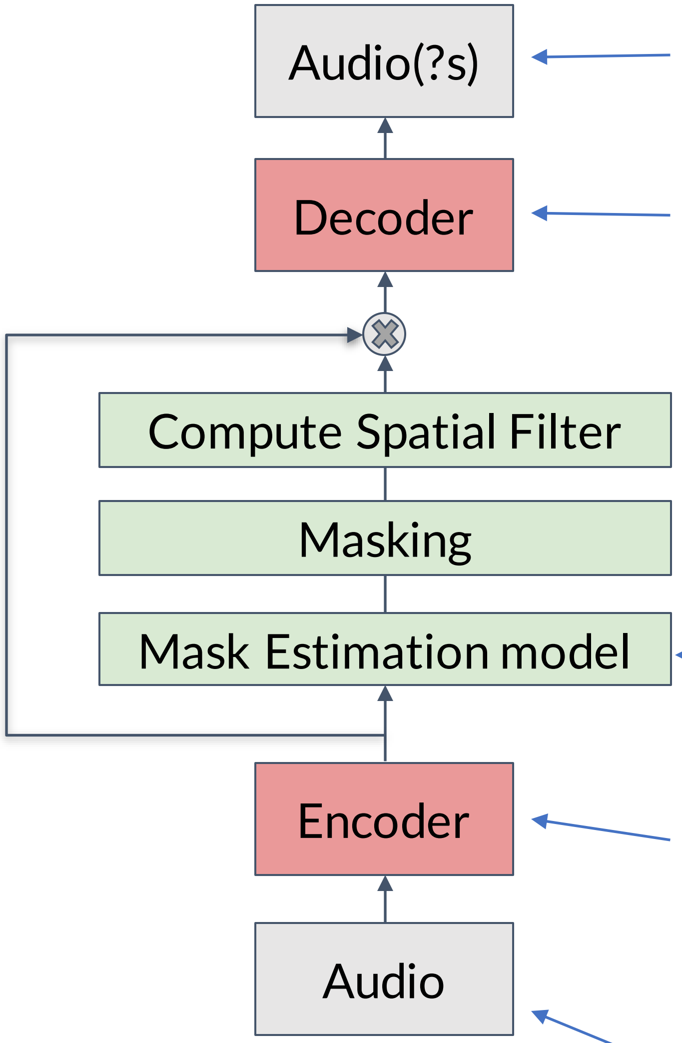
# Multi-Channel Encoder-Separation-Decoder (ESD)



# Multi-Channel Encoder-Separation-Decoder (ESD)



# Multi-Channel Encoder-Separation-Decoder (ESD)




Saparated single audio

Inverse STFT

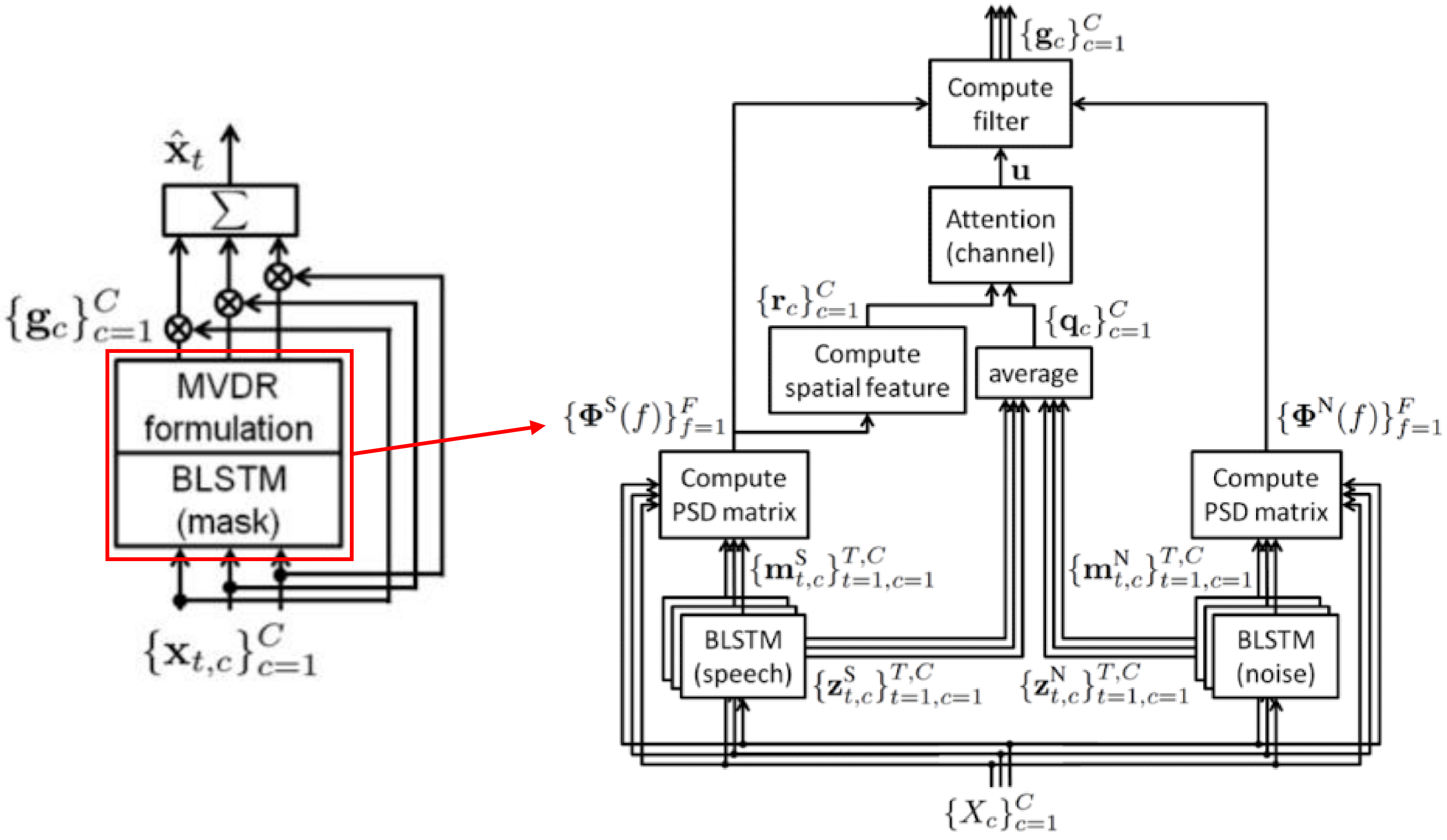
**How to compute masks?** → **Deep learning-based method**

STFT, Magnitude, IPD?

Multi-channel audio

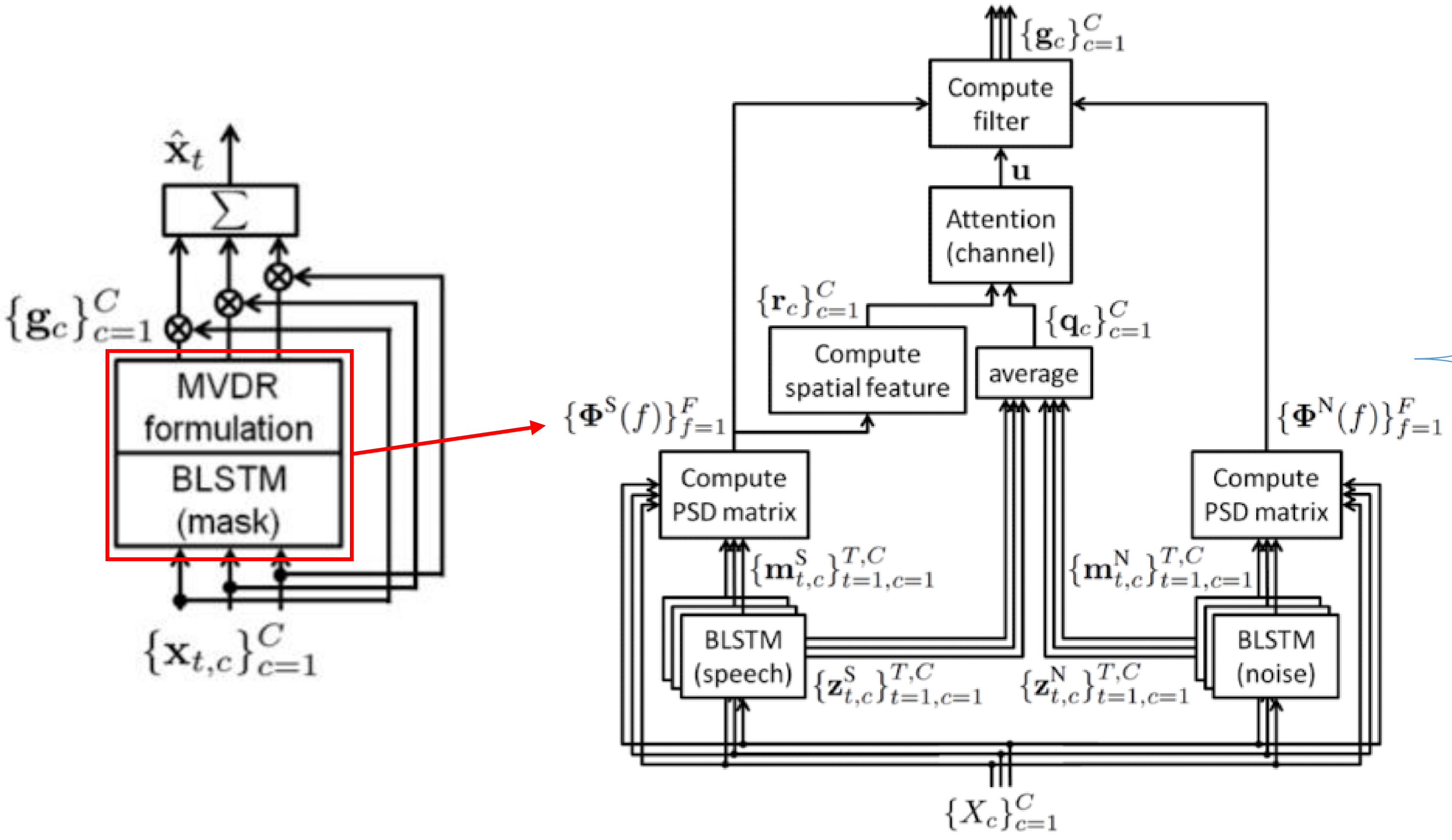
# Multi-Channel ASR

- Multichannel End-to-end Speech Recognition, ICML17
  - The neural Beamformer is based on the MVDR formalizations



# Multi-Channel ASR

- Multichannel End-to-end Speech Recognition, ICML17
    - The neural Beamformer is based on the MVDR formalizations




- The filter:

$$\mathbf{g}(f) = \frac{\boldsymbol{\Phi}^{\mathrm{N}}(f)^{-1}\boldsymbol{\Phi}^{\mathrm{S}}(f)}{\mathrm{Tr}(\boldsymbol{\Phi}^{\mathrm{N}}(f)^{-1}\boldsymbol{\Phi}^{\mathrm{S}}(f))}\mathbf{u},$$

- The cross-channel power spectral density (PSD) matrices :

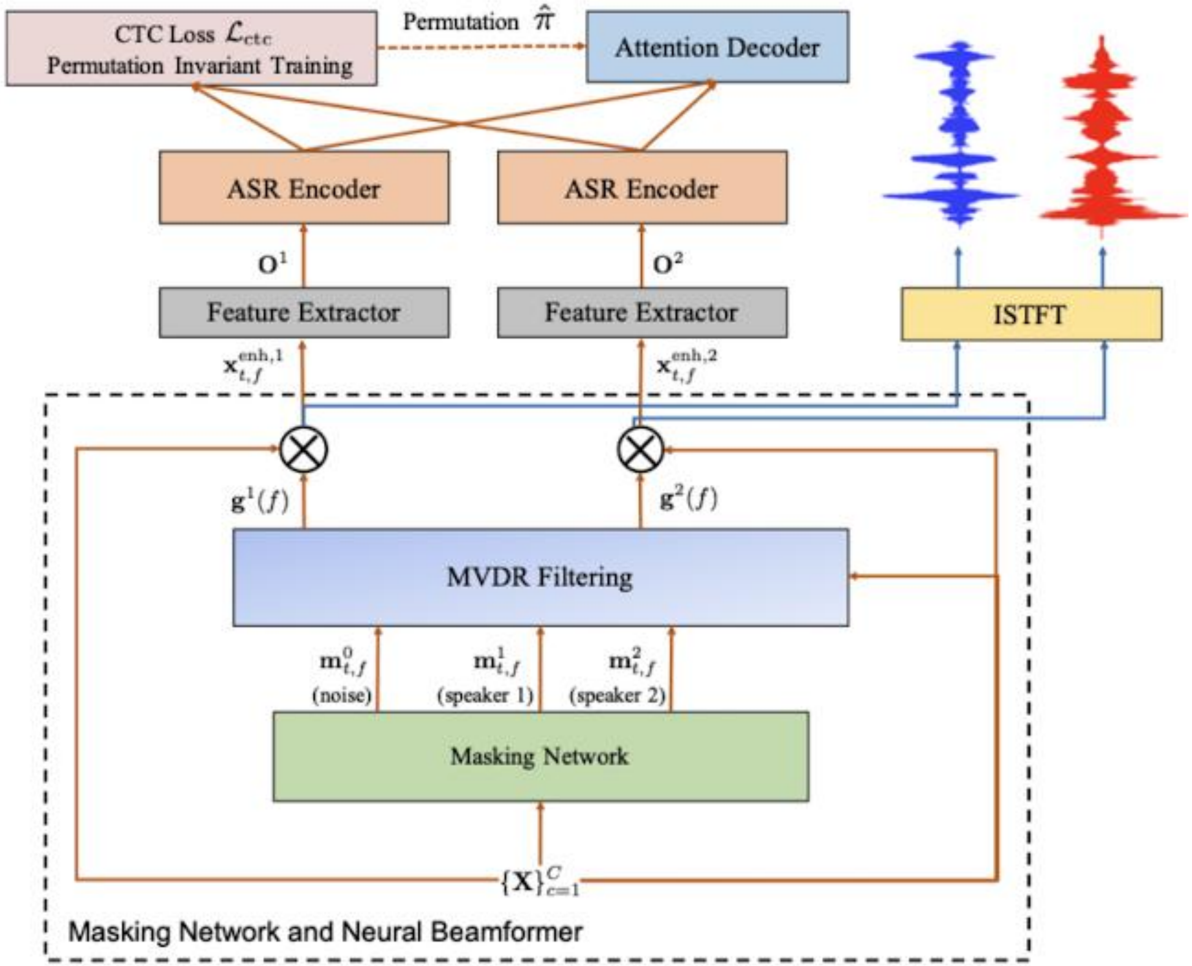
$$\boldsymbol{\Phi}^{\mathrm{S}}(f) = \frac{1}{\sum_{t=1}^{T} m_{t,f}^{\mathrm{S}}} \sum_{t=1}^{T} m_{t,f}^{\mathrm{S}} \mathbf{x}_{t,f}\mathbf{x}_{t,f}^{\dagger},$$

$$\boldsymbol{\Phi}^{\mathrm{N}}(f) = \frac{1}{\sum_{t=1}^{T} m_{t,f}^{\mathrm{N}}} \sum_{t=1}^{T} m_{t,f}^{\mathrm{N}} \mathbf{x}_{t,f}\mathbf{x}_{t,f}^{\dagger},$$

**Target** and **noise masks** can be obtained using a DL model

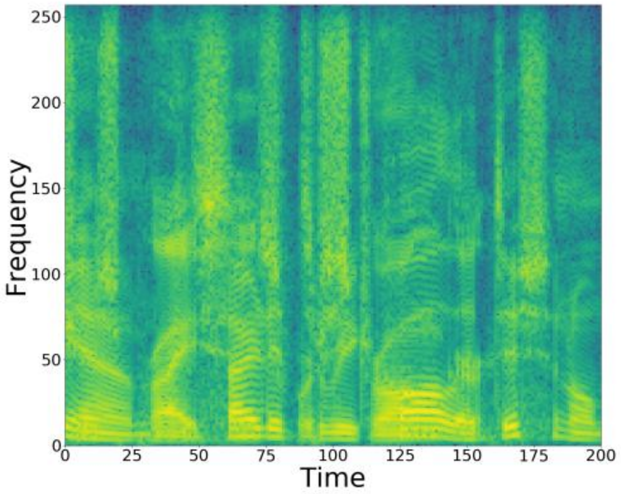
# Multi-Channel ASR

- MINO-Speech: End-To-End Multi-Channel **Multi-Speaker** Speech Recognition, ASRU19

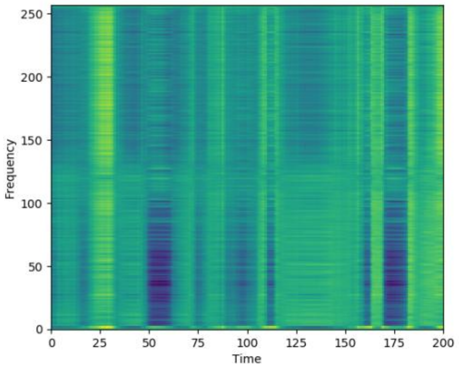


# Multi-Channel ASR

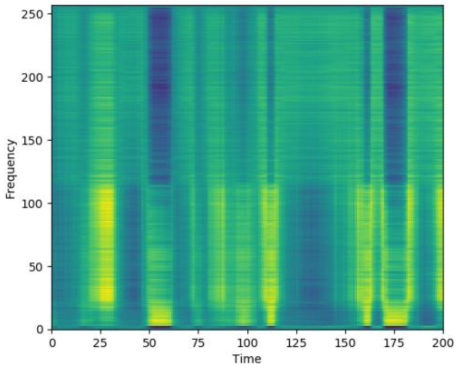
- MINO-Speech: End-To-End Multi-Channel **<u>Multi-Speaker</u>** Speech Recognition, ASRU19



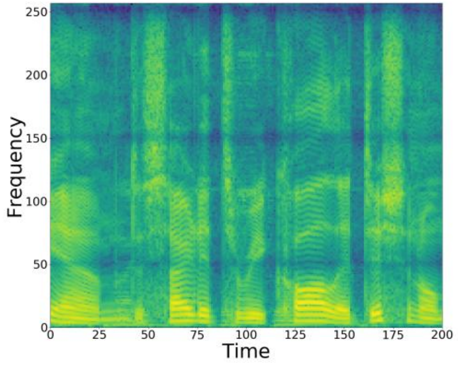

(e) Overlapped Speech



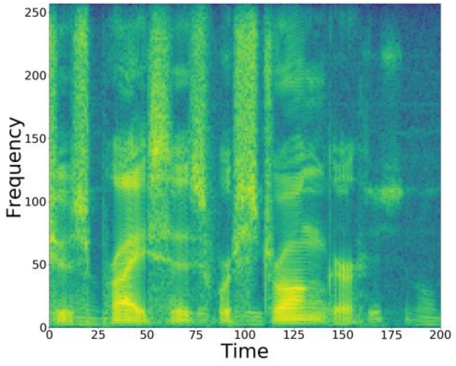

(a) Mask for Speaker 1




(b) Mask for Speaker 2




(c) Separated Speech for Speaker 1




(d) Separated Speech for Speaker 2

# Multi-Channel ASR

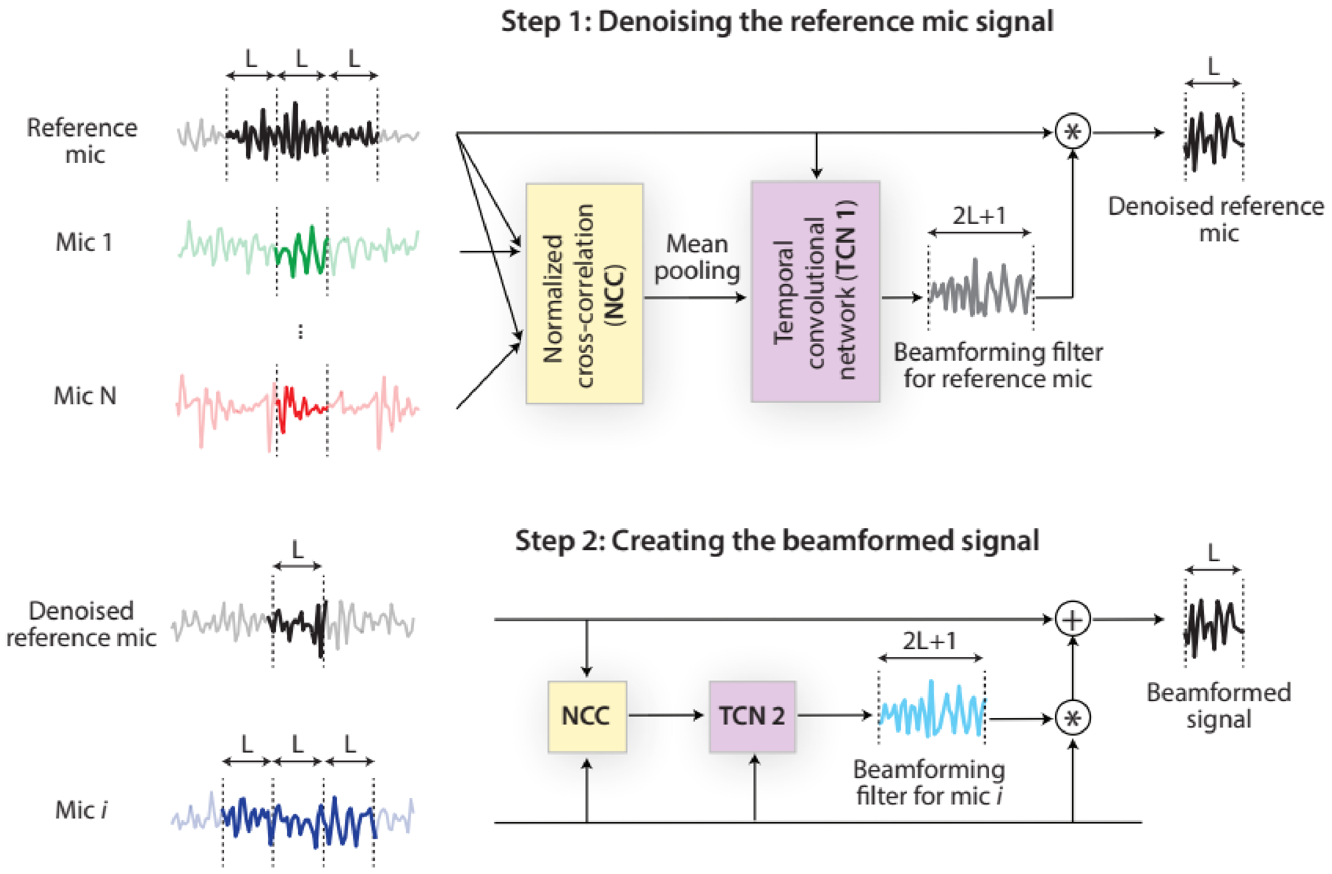
- MINO-Speech: End-To-End Multi-Channel **Multi-Speaker** Speech Recognition, ASRU19

**Table 1**. Performance in terms of average CER and WER [%] on the spatialized anechoic wsj1-2mix corpus.

| Model | dev CER | eval CER |
|---|---|---|
| 2-spkr ASR (1st channel) | 22.65 | 19.07 |
| BeamformIt Enhancement (2-spkr ASR) | 15.23 | 12.45 |
| BeamformIt Separation (1-spkr ASR) | 77.30 | 77.10 |
| MIMO-Speech | 7.29 | 4.51 |
| + Curriculum Learning (SNRs) | **6.34** | **3.75** |
| **Model** | **dev WER** | **eval WER** |
| 2-spkr ASR (1st channel) | 34.98 | 29.43 |
| BeamformIt Enhancement (2-spkr ASR) | 26.61 | 21.75 |
| BeamformIt Separation (1-spkr ASR) | 98.60 | 98.00 |
| MIMO-Speech | 13.54 | 8.62 |
| + Curriculum Learning (SNRs) | **12.59** | **7.55** |

# Multi-Channel ASR

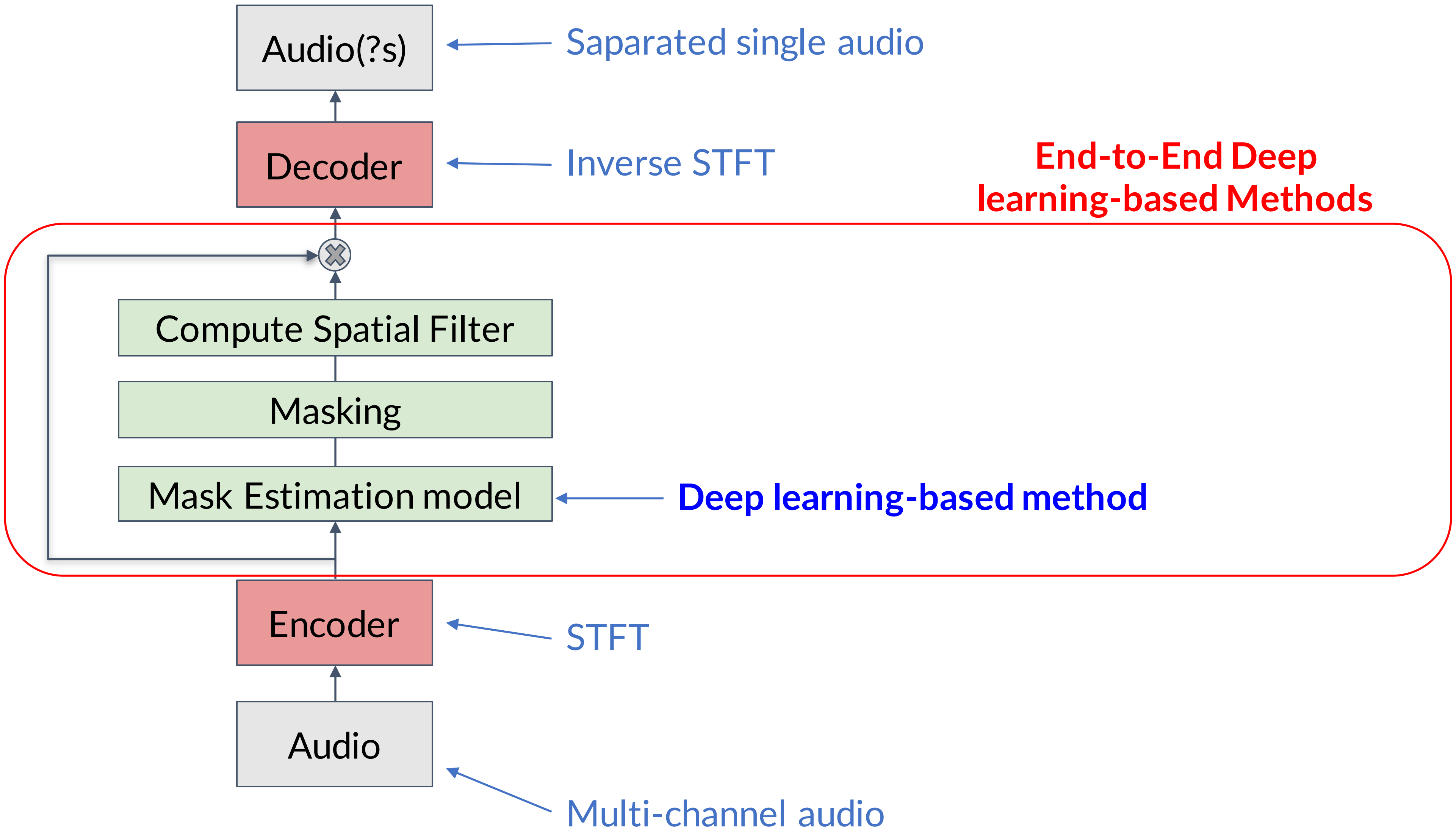
- FaSNet: Low-Latency Adaptive Beamforming for Multi-Microphone Audio Processing, ASRU 2019
  - Time domain neural beamforming $\hat{x}(t) = \boldsymbol{w}^H * y(t)$




Target signal의 clue를 찾은 후,

이를 활용해 spatial clue를 고려하도록 설계

# Multi-Channel Encoder-Separation-Decoder (ESD)



# Multi-Channel ASR

- A comprehensive study of speech separation: spectrogram vs waveform separation, Interspeech 2019

**IPD (Inter-Phase Difference):** $\mathrm{IPD}_{i,t,f} = \angle(\frac{Y_{i_1,t,f}}{Y_{i_2,t,f}}), i = 1:6$

Waveform 기반 Spectrum 기반

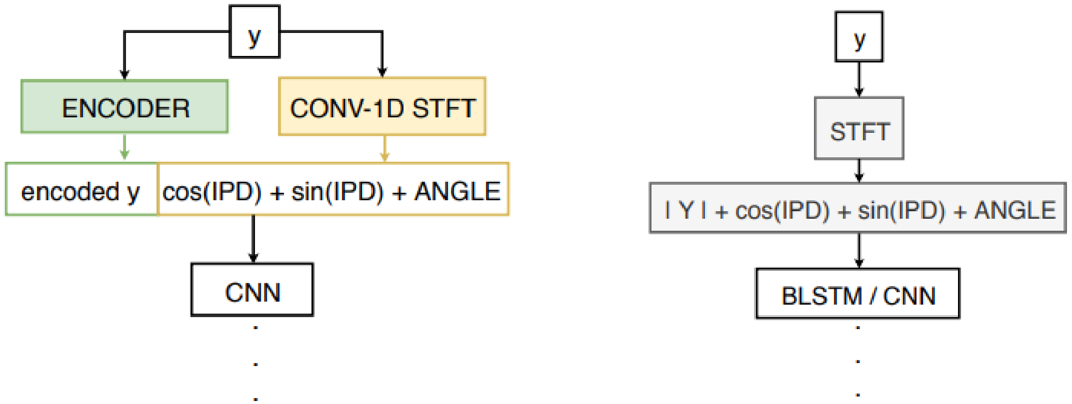



Figure 1: *Multi-channel speech separation; (left) waveform separation, (right) spectrogram separation.*

- Waveform / Spectrum 중 어떤 것이 더 효과적인지?
- STFT / IPD / Angle 중 어떤 input feature가 효과적인지?

# Multi-Channel ASR

- A comprehensive study of speech separation: spectrogram vs waveform separation, Interspeech 2019

Table 1: *Experimental setup for time/frequency-domain models.*

| Model | Input | # of Parameters | Setting | Normalization |
|---|---|---|---|---|
| F-CNN-1 | $\lvert Y_0 \rvert$ | 8.78M | N=257 | gLN |
| F-CNN-2 | $\lvert Y_0 \rvert$ + cos(IPD) + sin(IPD) | 9.58M | N=257×13 | gLN |
| F-CNN-3 | $\lvert Y_0 \rvert$ + cos(IPD) + sin(IPD) + Angle | 9.71M | N=257×15 | gLN |
| F-BLSTM-1 | $\lvert Y_0 \rvert$ | 67.05M | 4 × BLSTM-896 | - |
| F-BLSTM-2 | $\lvert Y_0 \rvert$ + cos(IPD) + sin(IPD) | 89.15M | 4 × BLSTM-896 | - |
| F-BLSTM-3 | $\lvert Y_0 \rvert$ + cos(IPD) + sin(IPD) + Angle | 92.84M | 4 × BLSTM-896 | - |
| T-CNN-1 | $y_0$ | 8.76M | L/N=40/256 | gLN |
| T-CNN-2 | $y_0$ + cos(IPD) + sin(IPD) | 8.83M | L/N=40/256 | BN |

Table 4: *Comparing spectrogram and waveform separation for both separation and ASR tasks.*

| # Channels | Domain | Model | Si-SNR 0-15° | Si-SNR 15-45° | Si-SNR 45-90° | Si-SNR 90-180° | Si-SNR AVG | SDR 0-15° | SDR 15-45° | SDR 45-90° | SDR 90-180° | SDR AVG | PESQ | WER Reduc. (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1-ch | time | T-BLSTM | - | - | - | - | - | - | - | - | - | - | - | - |
| | | T-CNN-1 | 9.02 | 9.33 | 9.59 | 9.71 | **9.47** | 9.57 | 9.83 | 10.09 | 10.2 | **9.97** | 1.95 | 45.53 |
| | freq | F-BLSTM-1 | 7.54 | 7.80 | 7.72 | 7.81 | **7.74** | 8.14 | 8.39 | 8.29 | 8.38 | **8.32** | 1.77 | 32.21 |
| | | F-CNN-1 | 7.08 | 7.48 | 7.45 | 7.48 | **7.42** | 7.70 | 8.06 | 8.02 | 8.06 | **8** | 1.77 | 35.17 |
| m-ch | time | T-BLSTM | - | - | - | - | - | - | - | - | - | - | - | - |
| | | T-CNN-2 | 7.70 | 11.63 | 12.33 | 12.62 | **11.55** | 8.31 | 12.07 | 12.74 | 13.03 | **11.99** | 2.10 | 59.11 |
| | freq | F-BLSTM-2 | 5.41 | 9.37 | 10.13 | 10.65 | **9.38** | 6.13 | 9.89 | 10.62 | 11.13 | **9.91** | 1.92 | 45.32 |
| | | F-CNN-2 | 6.88 | 10.27 | 11.02 | 11.54 | **10.36** | 7.5 | 10.75 | 11.47 | 11.99 | **10.84** | 2.00 | 51.73 |
| Oracle IBM | | | 11.56 | 11.51 | 11.53 | 11.53 | **11.53** | 11.93 | 11.86 | 11.88 | 11.88 | **11.88** | 2.01 | 50.37 |
| Oracle IAM | | | 11.05 | 11.03 | 11.05 | 11.03 | **11.04** | 11.33 | 11.3 | 11.31 | 11.29 | **11.30** | 2.23 | 71.45 |
| Oracle IRM | | | 11.01 | 10.96 | 10.98 | 10.97 | **10.98** | 11.45 | 11.39 | 11.39 | 11.39 | **11.40** | 2.22 | 70.28 |
| Oracle IPSM | | | 13.68 | 13.6 | 13.64 | 13.63 | **13.63** | 14.04 | 13.94 | 13.98 | 13.97 | **13.98** | 2.28 | 71.10 |
| Reference | | | | | | | | | | | | | 2.35 | 73.01 |

# Multi-Channel ASR

- A comprehensive study of speech separation: spectrogram vs waveform separation, Interspeech 2019

Table 1: *Experimental setup for time/frequency-domain models.*

| Model | Input | # of Parameters | Setting | Normalization |
|---|---|---|---|---|
| F-CNN-1 | $\lvert Y_0 \rvert$ | 8.78M | N=257 | gLN |
| F-CNN-2 | $\lvert Y_0 \rvert$ + cos(IPD) + sin(IPD) | 9.58M | N=257×13 | gLN |
| F-CNN-3 | $\lvert Y_0 \rvert$ + cos(IPD) + sin(IPD) + Angle | 9.71M | N=257×15 | gLN |
| F-BLSTM-1 | $\lvert Y_0 \rvert$ | 67.05M | 4 × BLSTM-896 | - |
| F-BLSTM-2 | $\lvert Y_0 \rvert$ + cos(IPD) + sin(IPD) | 89.15M | 4 × BLSTM-896 | - |
| F-BLSTM-3 | $\lvert Y_0 \rvert$ + cos(IPD) + sin(IPD) + Angle | 92.84M | 4 × BLSTM-896 | - |
| T-CNN-1 | $y_0$ | 8.76M | L/N=40/256 | gLN |
| T-CNN-2 | $y_0$ + cos(IPD) + sin(IPD) | 8.83M | L/N=40/256 | BN |

Table 4: *Comparing spectrogram and waveform separation for both separation and ASR tasks.*

| # Channels | Domain | Model | Si-SNR 0-15° | Si-SNR 15-45° | Si-SNR 45-90° | Si-SNR 90-180° | Si-SNR AVG | SDR 0-15° | SDR 15-45° | SDR 45-90° | SDR 90-180° | SDR AVG | PESQ | WER Reduc. (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1-ch | time | T-BLSTM | - | - | - | - | - | - | - | - | - | - | - | - |
| | | T-CNN-1 | 9.02 | 9.33 | 9.59 | 9.71 | **9.47** | 9.57 | 9.83 | 10.09 | 10.2 | **9.97** | 1.95 | 45.53 |
| | freq | F-BLSTM-1 | 7.54 | 7.80 | 7.72 | 7.81 | **7.74** | 8.14 | 8.39 | 8.29 | 8.38 | **8.32** | 1.77 | 32.21 |
| | | F-CNN-1 | 7.08 | 7.48 | 7.45 | 7.48 | **7.42** | 7.70 | 8.06 | 8.02 | 8.06 | **8** | 1.77 | 35.17 |
| m-ch | time | T-BLSTM | - | - | - | - | - | - | - | - | - | - | - | - |
| | | T-CNN-2 | 7.70 | 11.63 | 12.33 | 12.62 | **11.55** | 8.31 | 12.07 | 12.74 | 13.03 | **11.99** | 2.10 | 59.11 |
| | freq | F-BLSTM-2 | 5.41 | 9.37 | 10.13 | 10.65 | **9.38** | 6.13 | 9.89 | 10.62 | 11.13 | **9.91** | 1.92 | 45.32 |
| | | F-CNN-2 | 6.88 | 10.27 | 11.02 | 11.54 | **10.36** | 7.5 | 10.75 | 11.47 | 11.99 | **10.84** | 2.00 | 51.73 |
| Oracle IBM | | | 11.56 | 11.51 | 11.53 | 11.53 | **11.53** | 11.93 | 11.86 | 11.88 | 11.88 | **11.88** | 2.01 | 50.37 |
| Oracle IAM | | | 11.05 | 11.03 | 11.05 | 11.03 | **11.04** | 11.33 | 11.3 | 11.31 | 11.29 | **11.30** | 2.23 | 71.45 |
| Oracle IRM | | | 11.01 | 10.96 | 10.98 | 10.97 | **10.98** | 11.45 | 11.39 | 11.39 | 11.39 | **11.40** | 2.22 | 70.28 |
| Oracle IPSM | | | 13.68 | 13.6 | 13.64 | 13.63 | **13.63** | 14.04 | 13.94 | 13.98 | 13.97 | **13.98** | 2.28 | 71.10 |
| Reference | | | | | | | | | | | | | 2.35 | 73.01 |

- 1-ch보다 m-ch이 성능이 좋다
- CNN이 BLSTM 보다 성능이 좋다
- Time domain + IPD가 성능이 좋다 → Inter-channel relation을 잘 고려해야 한다.

# Multi-Channel ASR

- DASFORMER: DEEP ALTERNATING SPECTROGRAM TRANSFORMER FOR MULTI/SINGLE-CHANNEL SPEECH SEPARATION, ICASSP23

Spatial information을 포착할 수 있도록 학습

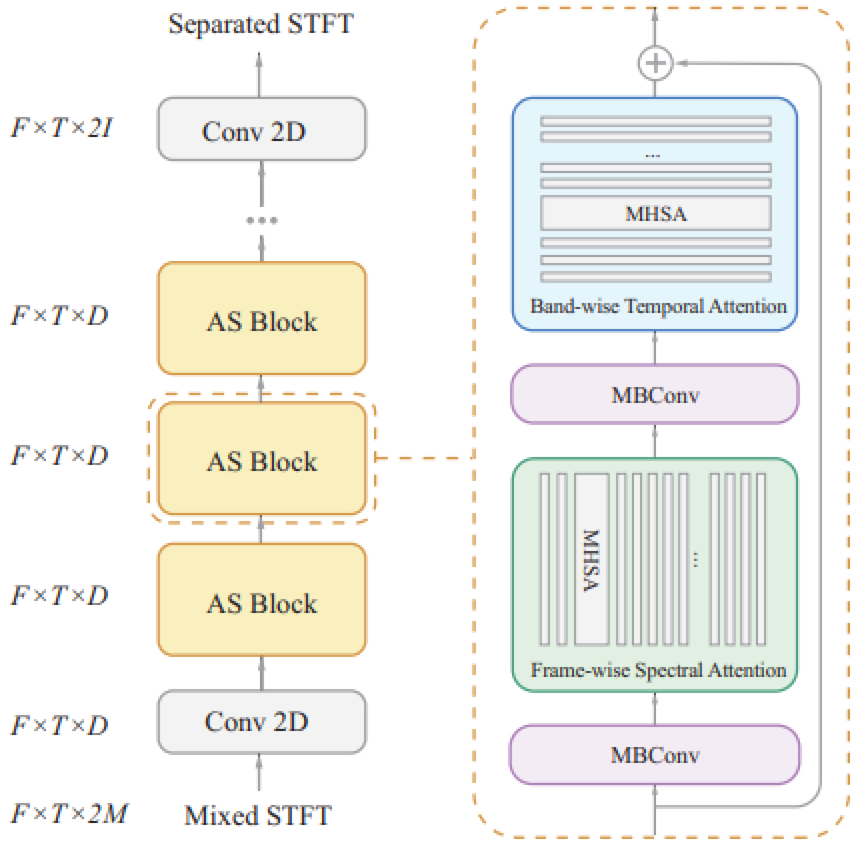



**Fig. 1**: The architecture of the proposed DasFormer.

| Model | Params. (M) | PESQ | SDRi (dB) |
|---|---|---|---|
| RIR settings as [16, 5] | | | |
| Mixture | – | 1.80 | 0.0 |
| FaSNet-TAC [21] | 2.8 | 2.90 | 11.7 |
| NBC [4] | 2.0 | 2.95 | 13.3 |
| Beam-TasNet [16] | – | – | 16.8 |
| BeamGuided-TasNet [5] | 5.4 | – | 21.5 |
| DasFormer (ours) | 2.2 | **4.33** | **25.9** |
| RIR settings as [4] | | | |
| Mixture | – | 1.80 | 0.0 |
| NBC [4] | 2.0 | 3.53 | 15.3 |
| DasFormer (ours) | 2.2 | **4.11** | **20.5** |

**Table 1**: Experiment results on spatialized WSJ0-2Mix.

# Multi-Channel ASR

- SpatialNet: Extensively Learning Spatial Information for Multichannel Joint Speech Separation, Denoising and Dereverberation, ASRU19

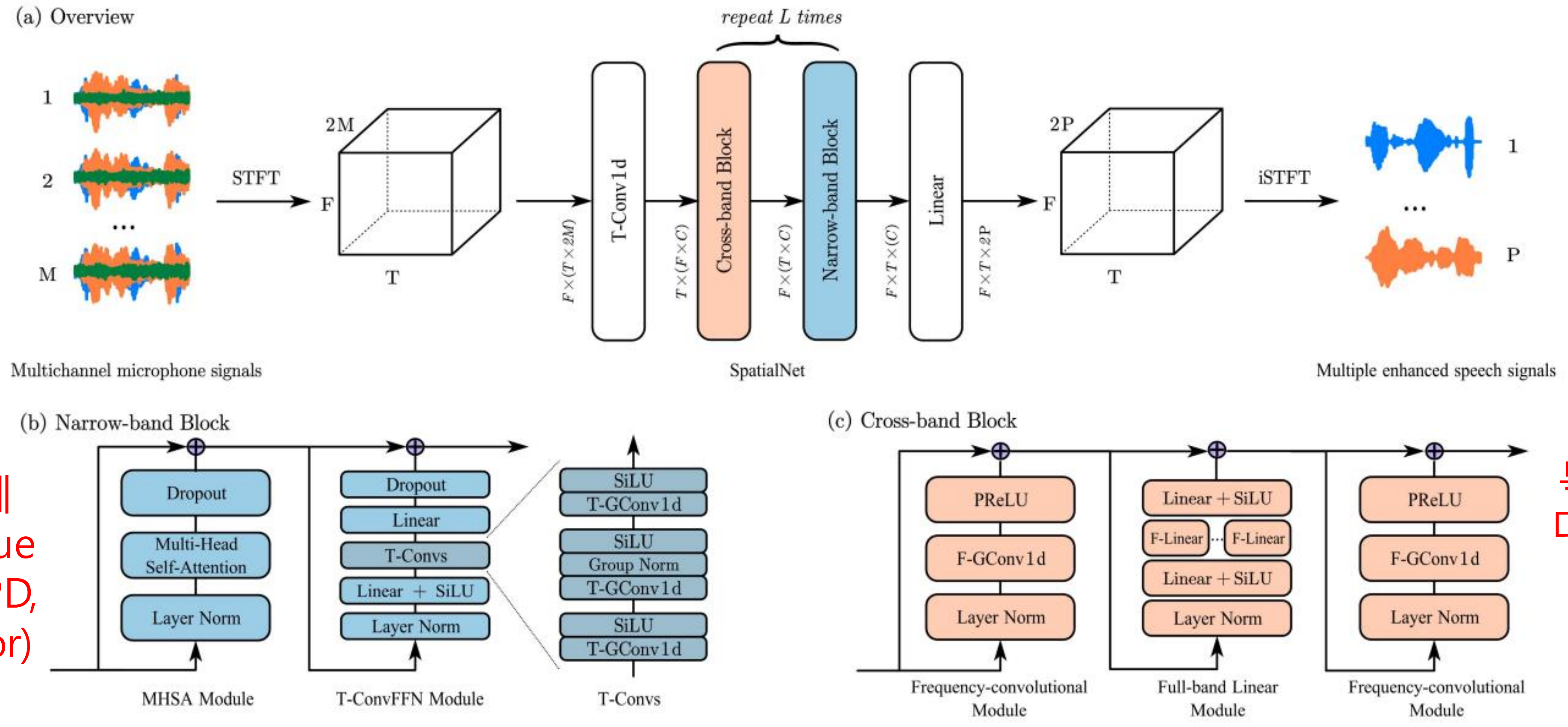



Fig. 1. The proposed SpatialNet. (a) The system overview. The input dimensions of neural blocks are presented before each of them in the form "batch dimension × (dimension of one sample in batch)". (b) The narrow-band block. (c) The cross-band block.

각 freq. bin에 대해 spatial cue를 포착 (예: IPD, steering vector)

특정 time bin에 대해 freq. bin 들의 관계를 학습

# Multi-Channel ASR

- SpatialNet: Extensively Learning Spatial Information for Multichannel Joint Speech Separation, Denoising and Dereverberation, ASRU19

TABLE IV
RESULTS ON 2-CHANNEL AND 6-CHANNEL SMS-WSJ DATASET. * DENOTE THAT THE SCORES ARE QUOTED FROM [17].

| **Method** | SISDR (dB) | SDR (dB) | NB-PESQ | eSTOI | WER (%) |
|---|---|---|---|---|---|
| unproc. | -5.45 | -0.38 | 1.50 | 0.441 | 78.7 |
| oracle direct-path | $\infty$ | $\infty$ | 4.5 | 1.0 | 6.31 |
| **Results for 2-channel SMS-WSJ** | | | | | |
| FasNet+TAC * [12] | 6.9 | - | 2.27 | 0.731 | 34.84 |
| Multi-TasNet * [61] | 5.8 | - | 2.16 | 0.720 | 45.72 |
| $\text{MISO}_1$-BF-$\text{MISO}_3$ [17] | 12.7 | - | 3.43 | 0.907 | 10.67 |
| Convolutional Prediction [62] | 15.8 | - | 3.71 | - | 8.60 |
| TFGridNet [26] | 20.3 | 22.0 | 3.81 | 0.967 | 7.41 |
| SpatialNet-small (prop.) | 19.4 | 21.0 | 3.80 | 0.957 | 8.10 |
| SpatialNet-large (prop.) | **23.3** | **24.6** | **4.03** | **0.975** | **7.20** |
| **Results for 6-channel SMS-WSJ** | | | | | |
| FasNet+TAC * [12] | 8.60 | - | 2.37 | 0.771 | 29.8 |
| Multi-TasNet * [61] | 10.8 | - | 2.78 | 0.844 | 23.1 |
| $\text{MISO}_1$-BF-$\text{MISO}_3$ [17] | 15.6 | - | 3.76 | 0.942 | 8.28 |
| MC-CSM with LBT [63] | 13.2 | 14.8 | 3.33 | 0.910 | 9.62 |
| TFGridNet [26] | 22.8 | 24.9 | 4.08 | 0.980 | 6.76 |
| SpatialNet-small (prop.) | 21.3 | 23.2 | 3.99 | 0.974 | 7.05 |
| SpatialNet-large (prop.) | **25.1** | **27.1** | **4.17** | **0.986** | **6.70** |