MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Single- and Multi-Channel Speech Enhancement and Separation for Far-Field Conversation Recognition

Masuyama, Yoshiki

TR2025-097 June 28, 2025

Abstract

While ASR achieves superhuman performance on clean benchmarks, it struggles in real-world scenarios like meeting transcription, where word error rates exceed 35% versus under 3% on clean data. This lecture examines the challenges of robust ASR for conversational speech, including noise, reverberation, multiple speakers, and overlapped speech (>15% of meeting duration). The lecture covers evaluation methodologies for long-form multi-speaker audio, including concatenated minimum permutation WER (cpWER), and surveys key datasets from AMI to current benchmarks like CHiME-7/8 and NOTSOFAR1. Technical approaches are categorized into front-end methods (speech separation, beamforming, target speaker extraction) and back-end methods (self-supervised features, serialized output training, target-speaker ASR). Robust ASR remains an active research area with significant opportunities, particularly as large language models enable new applications like automated meeting summarization. Key challenges include speaker tracking, training-inference mismatches, and integrating speech separation, diarization, and recognition components.

Jelinek Summer Workshop on Speech and Language Technology (JSALT) 2025

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Single- and Multi-Channel Speech Enhancement and Separation for Far-Field Conversation Recognition

Yoshiki Masuyama

Jelinek Summer Workshop on Speech and Language Technology

June 17, 2025

MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)
Cambridge, Massachusetts, USA
http://www.merl.com

Collaborators on today’s topics

Jonathan
Le Roux

Chiori
Hori

Gordon
Wichern

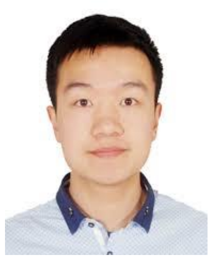
Francois
Germain

Ryo
Aihara

Current MERL
SA team

Shinji
Watanabe

Zhong-Qiu
Wang



Xuankai
Chang

Matthew
Maciejewski

Kohei
Saijo

Samuele
Cornell

Yoshiaki
Bando

Agenda

- **<u>Overview of speech separation and enhancement (SSE)</u>**
- Single-channel SSE addressing permutation issue
- Signal-processing-based multi-channel SSE and dereverberation
- DNN-based multi-channel SSE
- Advanced topics

Challenges in Far-Field Conversational Speech

- Multiple utterances are overlapped and contaminated by noise and reverberation.

Challenges in Far-Field Conversational Speech

- Multiple **utterances are overlapped** and contaminated by **noise** and **reverberation**.

Speech Separation and Enhancement (SSE)

- We aim to **isolate desired speech** signals from mixtures.

Speech Separation and Enhancement (SSE)

- We aim to **isolate desired speech** signals from mixtures.

Acoustic Propagation

- Acoustic propagation from the source position to a microphone can be characterized by a **room impulse response (RIR)** [Kuttruff2016].
  - We typically assume the acoustic propagation is linear and time-invariant.
  - RIR depends not only on the source and microphone positions but also room settings.

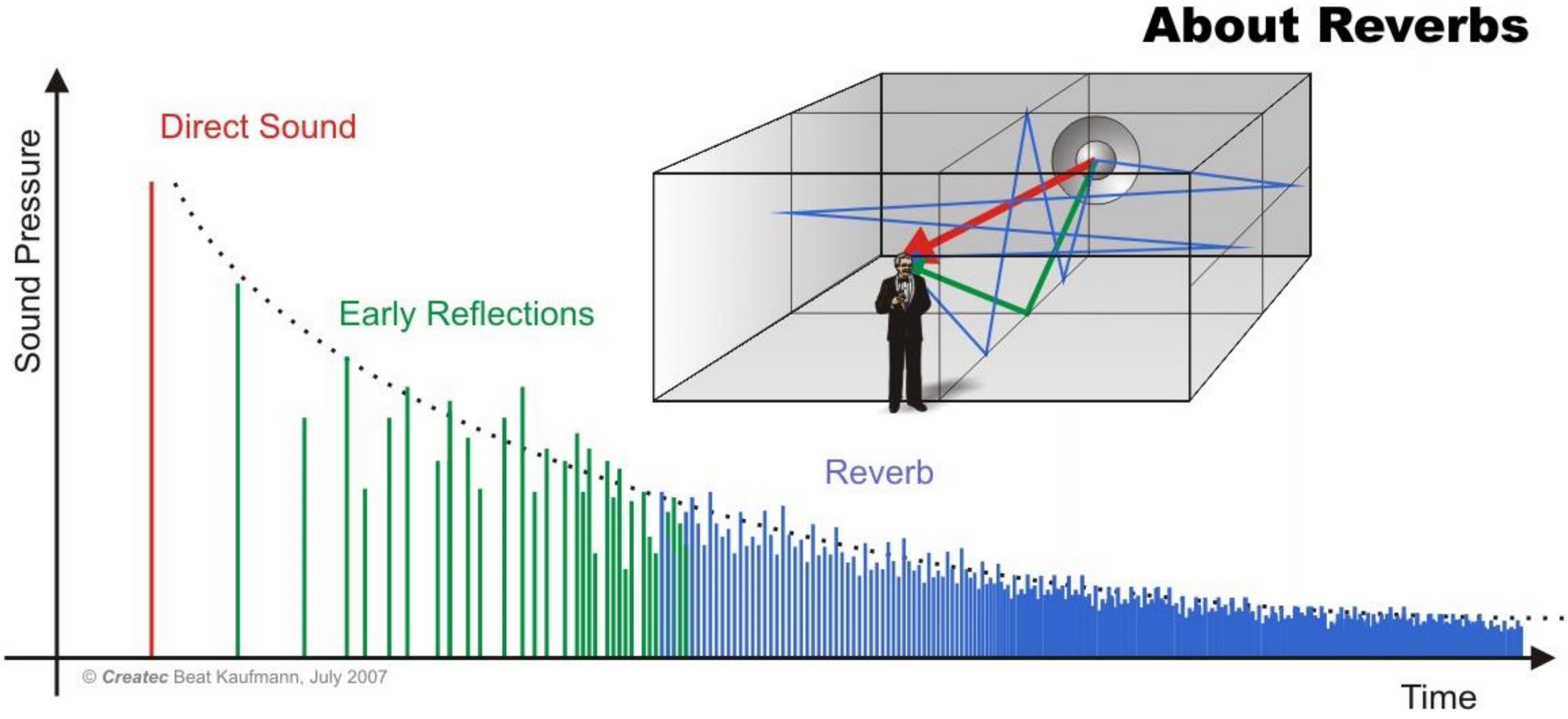


Image is from https://www.beat-kaufmann.com

Mathematical Notation of Mixing Process

- Acoustic propagation from the source position to a microphone can be characterized by a **room impulse response (RIR)** [Kuttruff2016].
  - We typically assume the acoustic propagation is linear and time-invariant.
  - RIR depends not only on the speaker and microphone positions but also room settings.

$$\mathbf{y}_{k,m} = \underline{\mathbf{h}_{k,m}} \circledast \mathbf{s}_k \in \mathbb{R}^L$$

$\mathbf{y}_{k,m}$: “image” of source k at microphone m

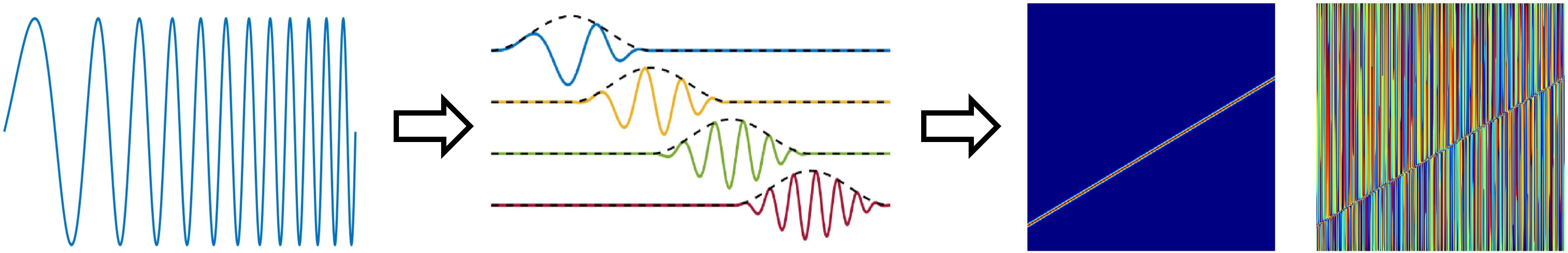
$\mathbf{s}_k$: kth dry source signal

- Microphone records the superposition of K source images and noise.

$$\underbrace{\mathbf{x}_m}_{\text{Mixture}} = \sum_{k=1}^{K} \mathbf{y}_{k,m} + \underbrace{\mathbf{n}_m}_{\text{Noise}}$$

Time-Frequency (TF) Analysis as Encoding

- Audio signal is typically encoded to the TF domain by short-time Fourier transform (STFT).
  - STFT Magnitude is easy to interpret.
  - We can perform both single- and multi-channel processing efficiently in the STFT domain.



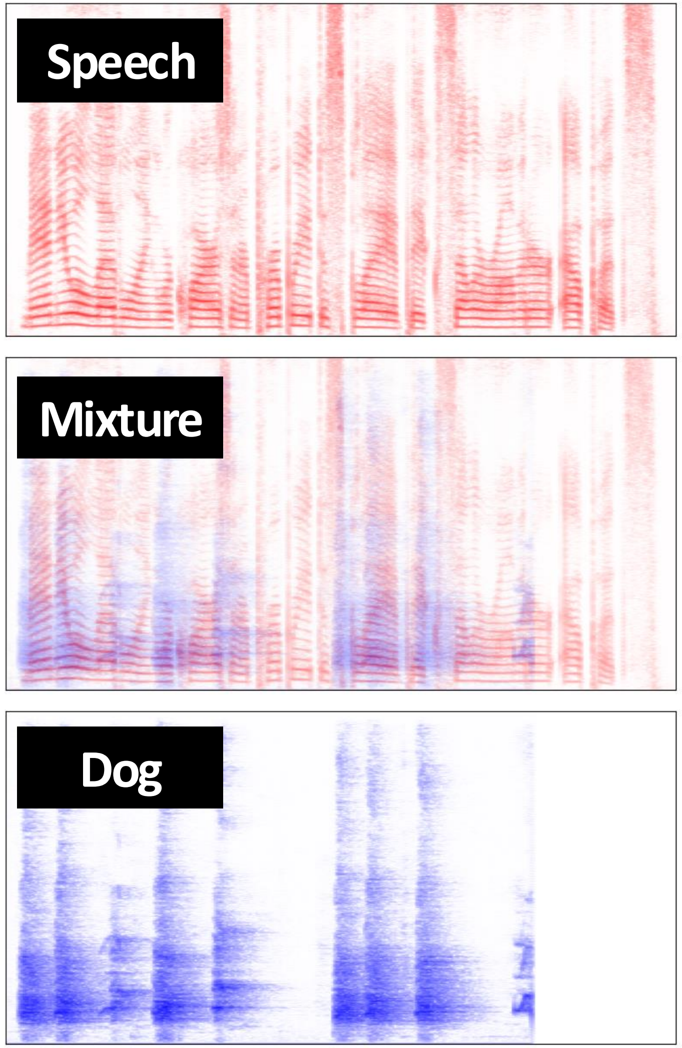
Time-domain signal — Sliding window — STFT magnitude — STFT phase

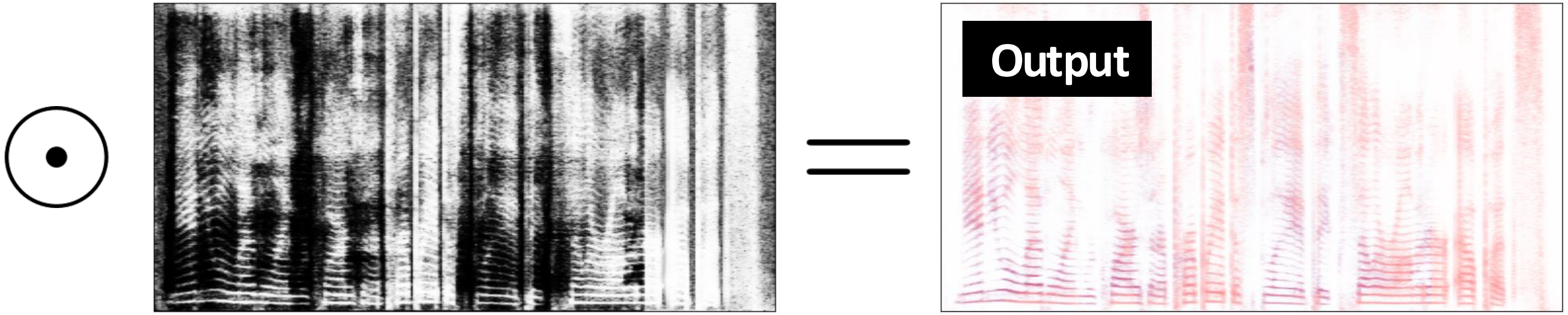
Recording → **STFT** → **Speech enhancement** → **Inverse STFT** → Play back

$$\mathbf{X}_m = \mathrm{STFT}(\mathbf{x}_m) \in \mathbb{C}^{T \times F}$$

TF Masking [Lyon1983, Weintraub1985, Wang+2006]

- Non-negative TF mask suppresses interference signals at each TF bin.



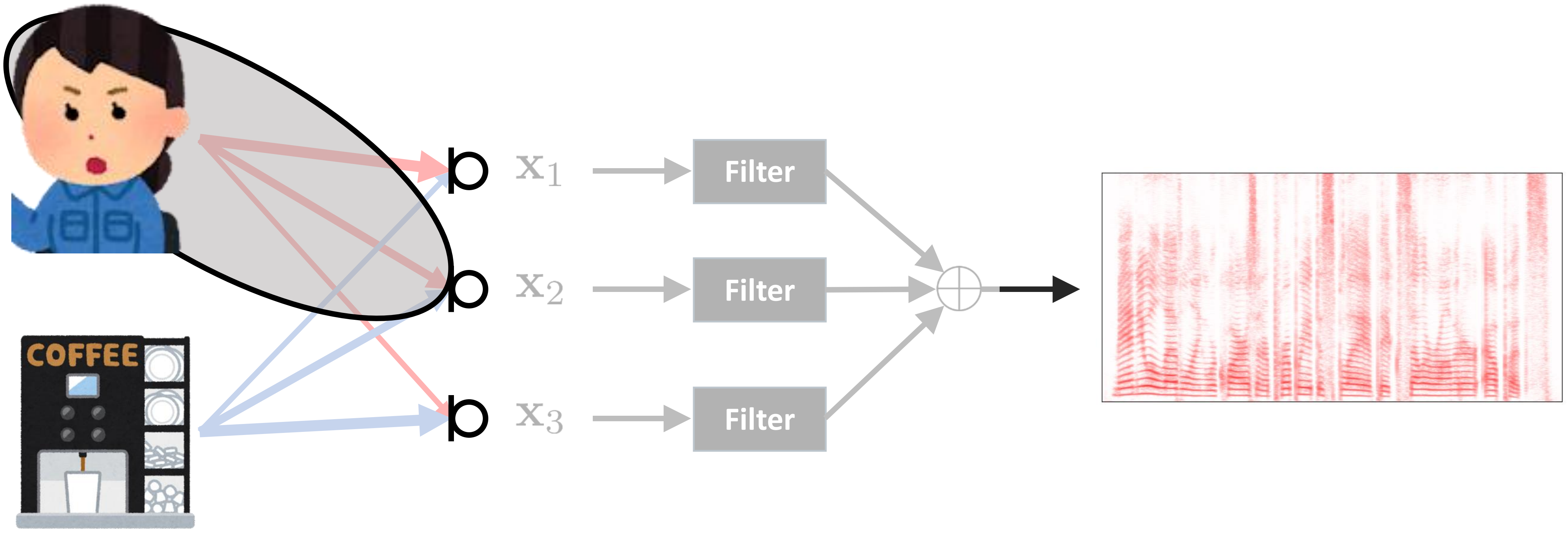


TF mask for enhancing speech
(**White part will be emphasized!**)

TF masks have been extended from non-negative value to complex value, i.e., more processing freedom.

Beamforming (Linear Spatial Filtering) [Trees2004, Benesty+2008]

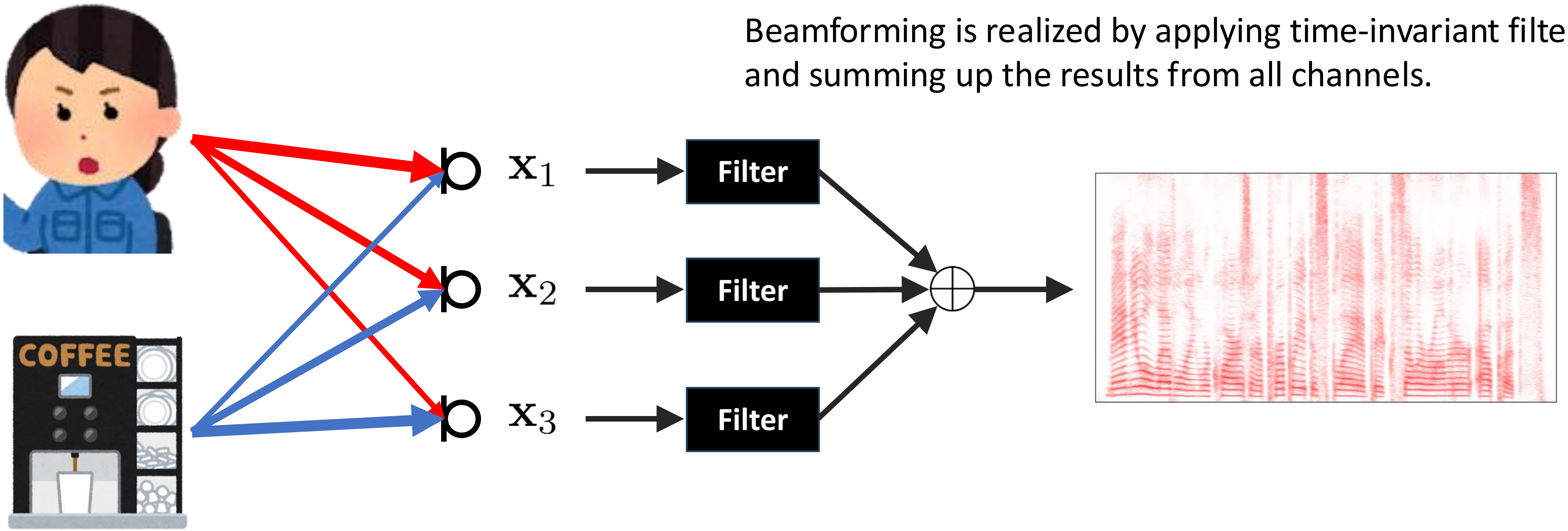
- Beamformer suppresses interference signals using spatial information.
  - Popular beamformers retain signals coming from a specific direction (i.e., target speaker's direction).
  - Interference signals from other directions will be suppressed.



Beamforming (Linear Spatial Filtering) [Trees2004, Benesty+2008]

- Beamformer suppresses interference signals using spatial information.
  - Popular beamformers retain signals coming from a specific direction (i.e., target speaker's direction).
  - Interference signals from other directions will be suppressed.

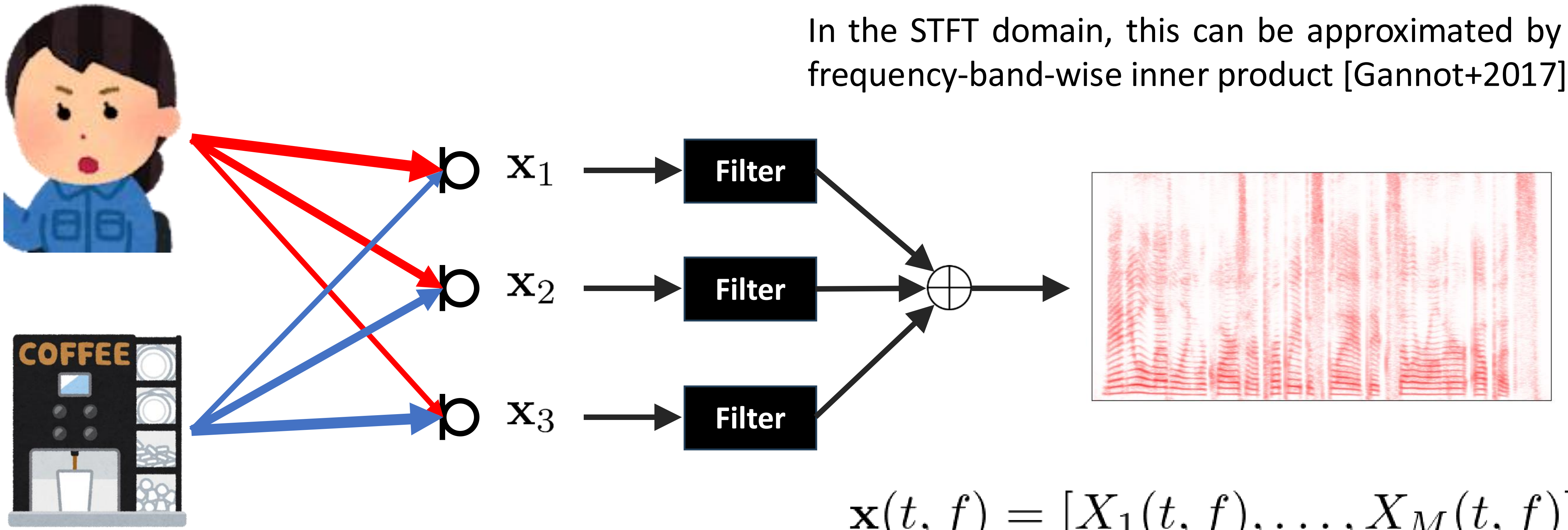
Beamforming is realized by applying time-invariant filters and summing up the results from all channels.



Beamforming (Linear Spatial Filtering) [Trees2004, Benesty+2008]

- Beamformer suppresses interference signals using spatial information.
  - Popular beamformers retain signals coming from a specific direction (i.e., target speaker's direction).
  - Interference signals from other directions will be suppressed.

In the STFT domain, this can be approximated by a frequency-band-wise inner product [Gannot+2017].

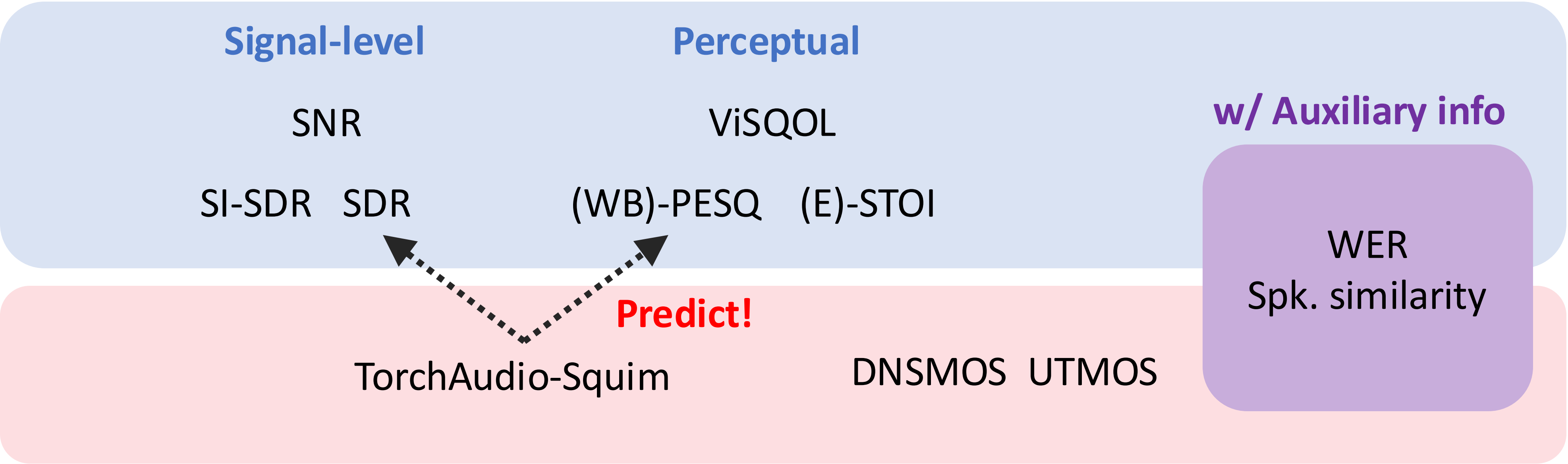


$$\mathbf{x}(t, f) = [X_1(t, f), \ldots, X_M(t, f)]^\mathsf{T}$$

$$\hat{Y}_1(t, f) = \mathbf{w}^\mathsf{H}(f)\mathbf{x}(t, f)$$

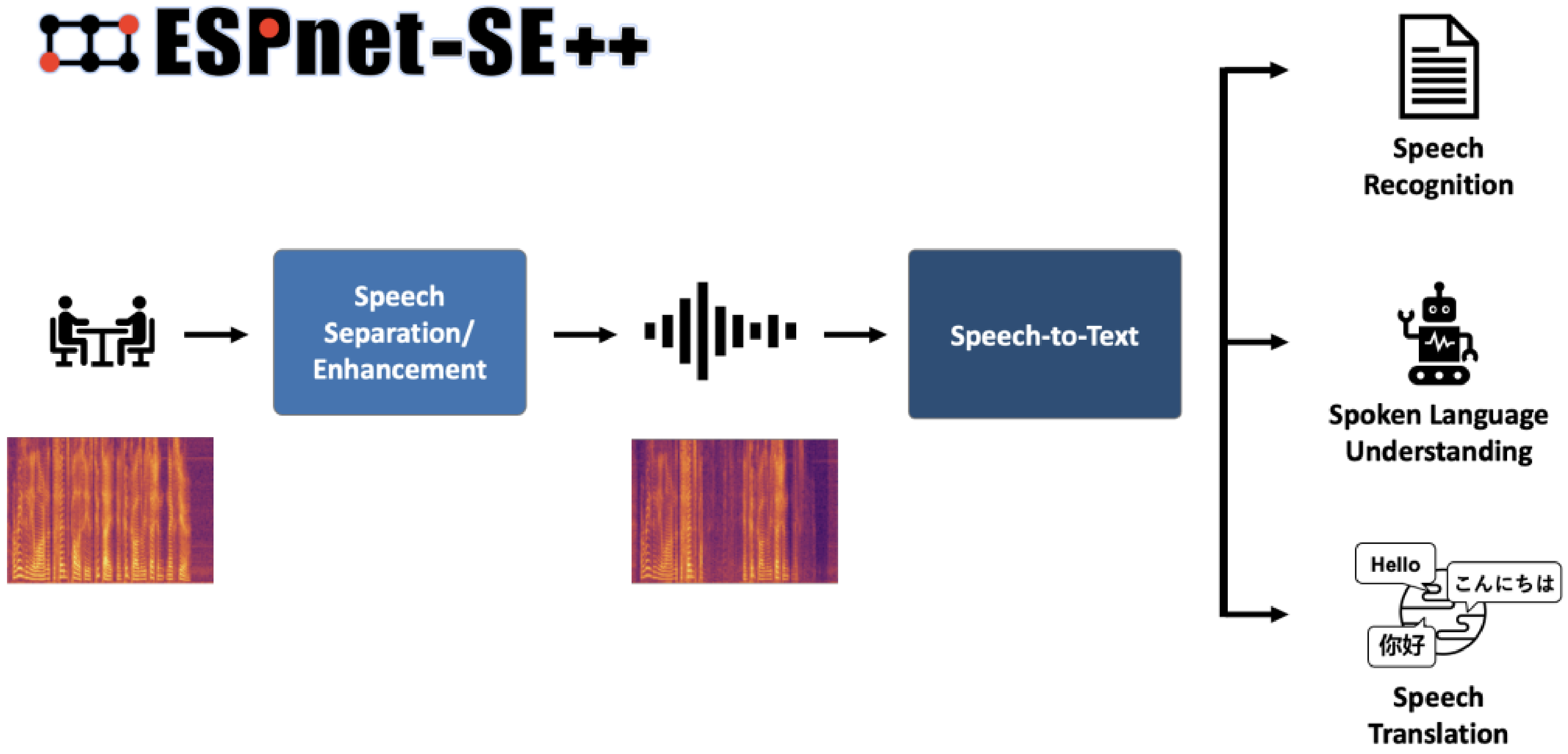
Performance Evaluation Metrics

- **Intrusive** metrics require a ground-truth signal
  - These metrics have been widely used for benchmarking SSE methods.
  - Ground-truth signals are accessible, when simulating mixtures by artificially summing up sources.
- **Non-intrusive** metrics are computed only from the enhanced/separated signals.
  - These metrics are easy to use with the recordings under realistic situations.



Open Tools for SSE

- Asteroid [Pariente+2020]: Focusing on SSE and is easy to use
- SpeechBrain [Ravanelli+2021]: Providing easy-to-start tutorials*
- ESPnet-SE [Li+2020, Lu+2022]: Supporting the end-to-end training of SSE and ASR modules



- Pyroomacoustics [Scheibler+2018]: Supporting RIR simulation via the image source method
  – Several array signal processing techniques are also implemented.

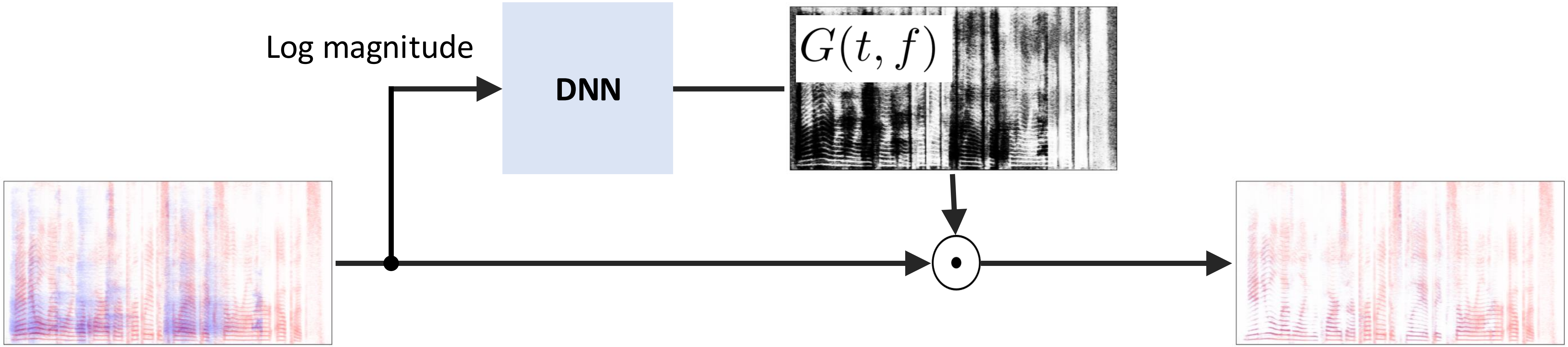
* https://github.com/speechbrain/speechbrain/blob/develop/docs/tutorials/tasks/source-separation.ipynb

Agenda

- Overview of speech separation and enhancement (SSE)
- **Single-channel SSE addressing permutation issue**
- Signal-processing-based multi-channel SSE and dereverberation
- DNN-based multi-channel SSE
- Advanced topics

DNN-Based Mask Estimation [Wang+2018]

- DNN predicts a mask for each speaker whose range is typically in [0, 1].



- Various ideal masks have been explored as targets.

$$G(t, f) = \frac{|Y(t, f)|^2}{|Y(t, f)|^2 + |N(t, f)|^2}$$

Wiener mask

$$G(t, f) = \mathtt{Real}\left(\frac{Y(t, f)}{X(t, f)}\right)$$

Phase-sensitive mask

(maximum SNR in real-valued masks)

- You can also use a loss function defined in the time domain.

$$\mathcal{L} = -\mathtt{SI\text{-}SDR}(\mathbf{y}, \mathtt{iSTFT}(\mathbf{G} \odot \mathbf{X}))$$

Trainable Encoder/Decoder and End-to-End Training

- STFT/inverse STFT are replaced by trainable 1D convolution/deconvolution.
  - These trainable encoder/decoder have a potential to improve the upper bound of masking.
  - This direction was dominant from 2018, Conv-TasNet [Luo+2018] to 2022.

- Dual-path modeling [Luo+2020]
  - The Encoded sequence is segmented to efficiently handle huge number of time frames.

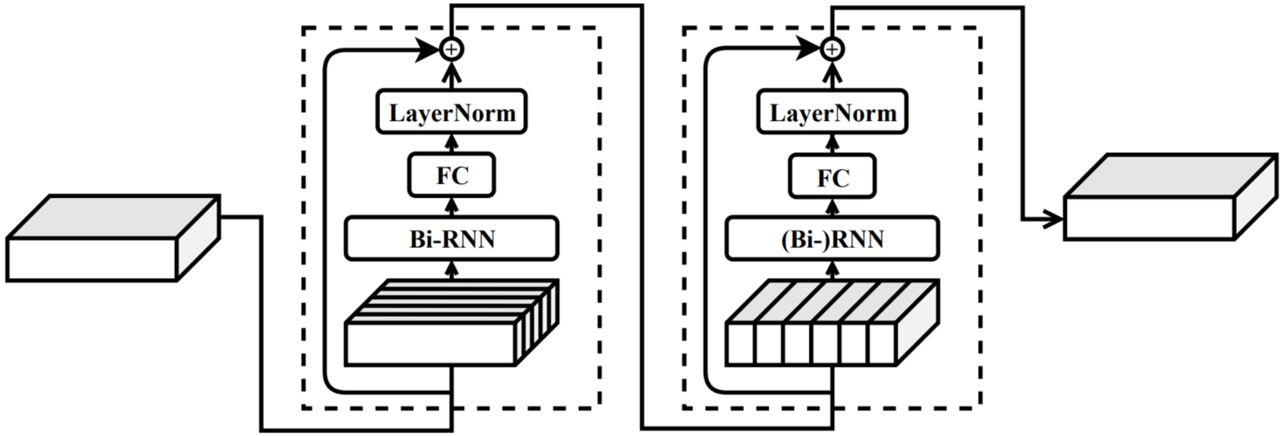


TABLE IV
COMPARISON WITH OTHER METHODS ON WSJ0-2MIX DATASET.

Method	Model size	Causal	SI-SNRi (dB)	SDRi (dB)
Conv-TasNet-gLN	**5.1M**	×	**15.3**	**15.6**
IRM	–	–	12.2	12.6
IBM	–	–	13.0	13.5
WFM	–	–	13.4	13.8

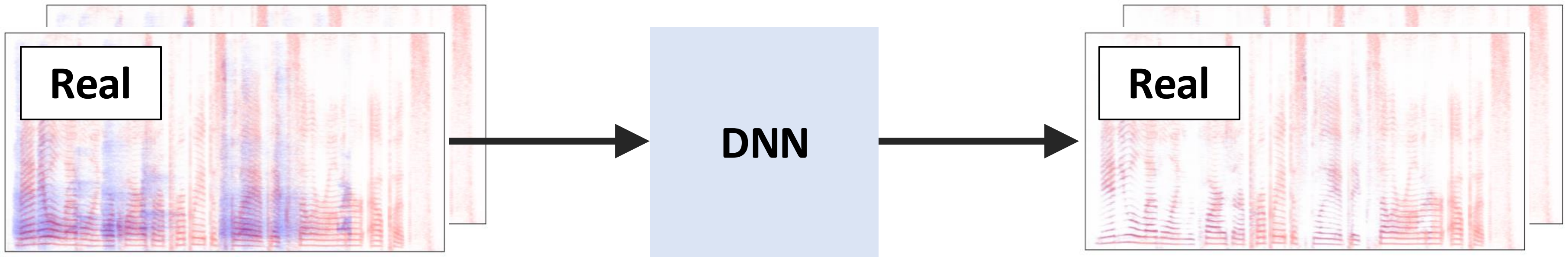
Progress in Network Architectures

- Temporal convolutional network (TCN):
  - TCN typically considers the frequency axis of STFT as the "channel" of the 1D convolution.
  - Dilation is doubled in each layer to increase receptive fields.

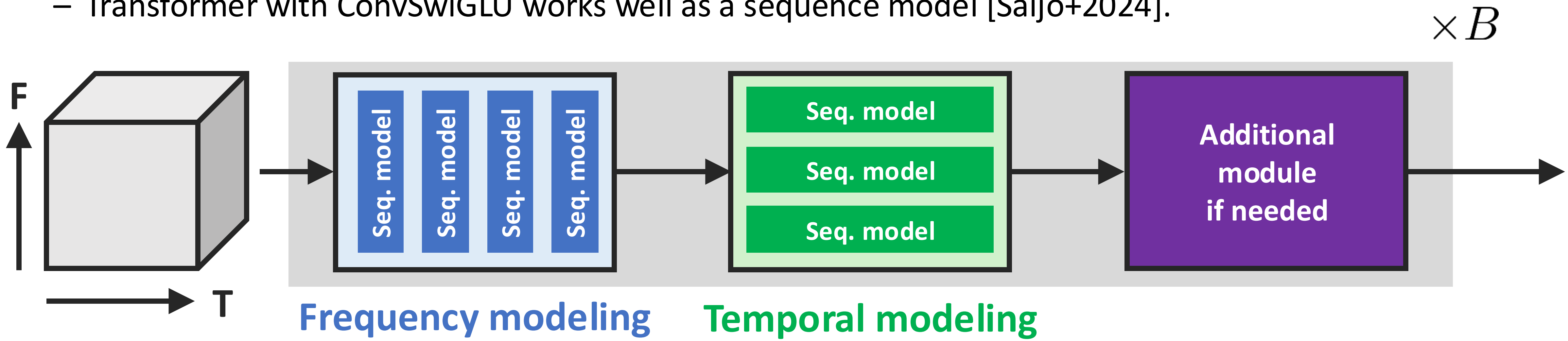
- LSTM/Transformer/Mamba:
  - LSTM has been widely used and is still strong compared with other speech tasks.
  - Transformer shows promising results when combined with local processing, but not as essential as in other speech tasks.
  - Mamba's efficiency with respect to the sequence length is suitable for SSE.

SOTA Approach in Single-Channel SSE

- DNN directly predicts complex STFT coefficients for each speaker [Wang+2020, 2021].

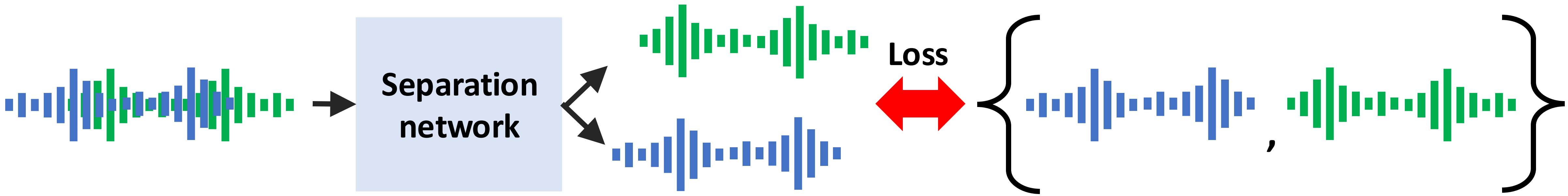


- TF dual-path modeling is widely used for complex spectral mapping [Yang+2022].
  - Each time frame (or frequency band) is handled separately.
  - Transformer with ConvSwiGLU works well as a sequence model [Saijo+2024].



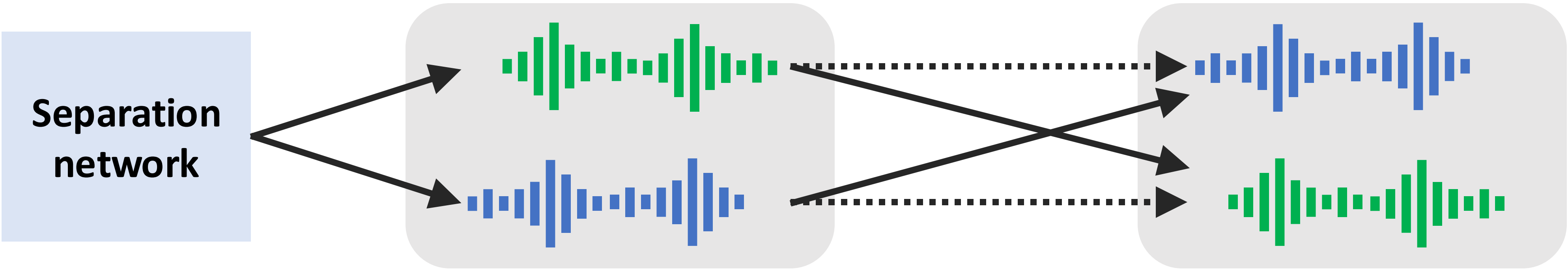
Permutation Issue in Separation

- The ground-truth signals are given as a set, and their order is not well-defined.
- Alignment between the ground-truth and estimates is required for computing the errors.



Ground-truth

- Permutation-invariant-training (PIT) calculates the losses for all possible permutations and backpropagates the smallest loss [Kolbæk+2017].

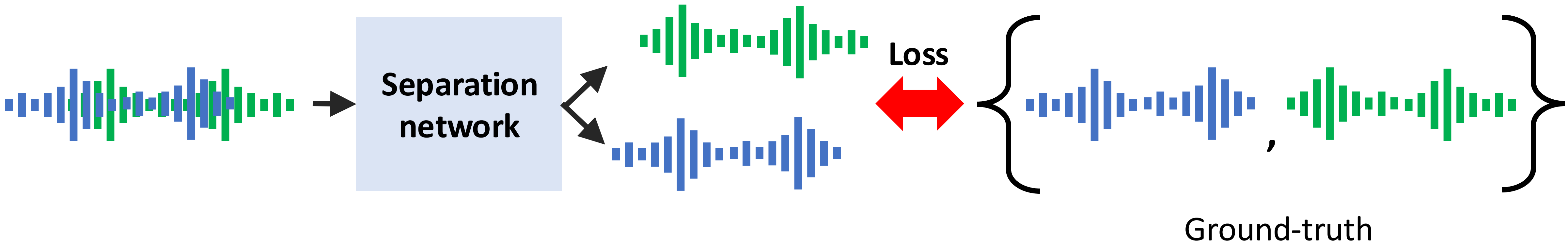


Estimates

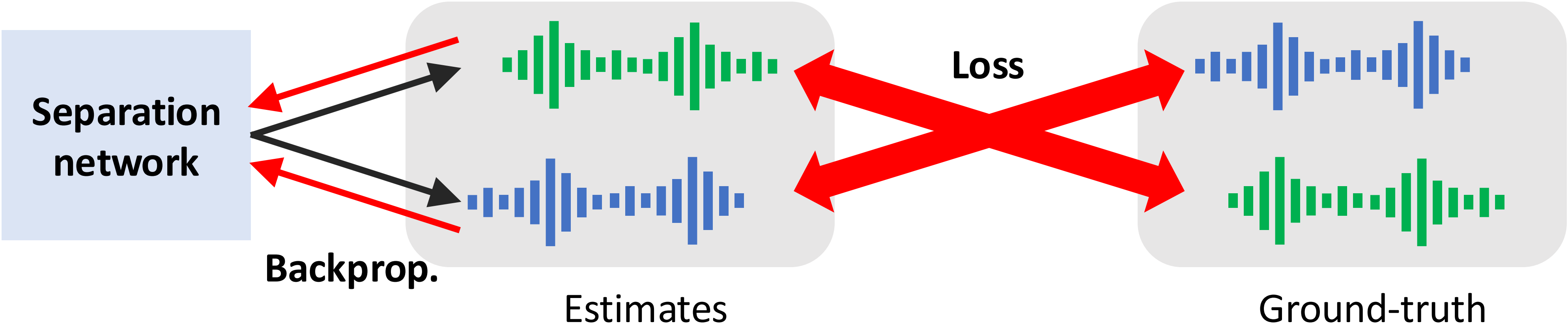
Ground-truth

Permutation Issue in Separation

- The ground-truth signals are given as a set, and their order is not well-defined.
- Alignment between the ground-truth and estimates is required for computing the errors.

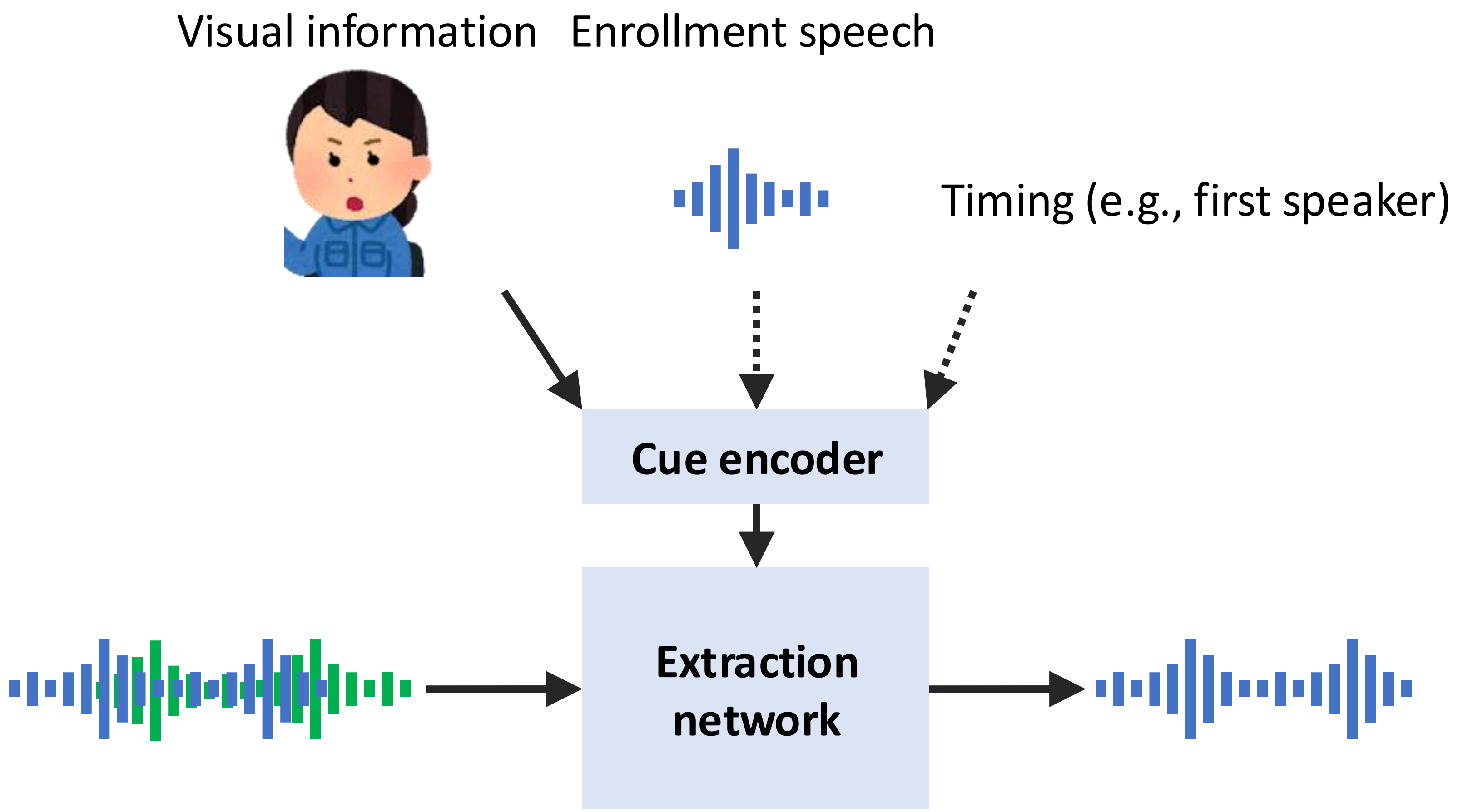


- Permutation-invariant-training (PIT) calculates the losses for all possible permutations and backpropagates the smallest loss [Kolbæk+2017].



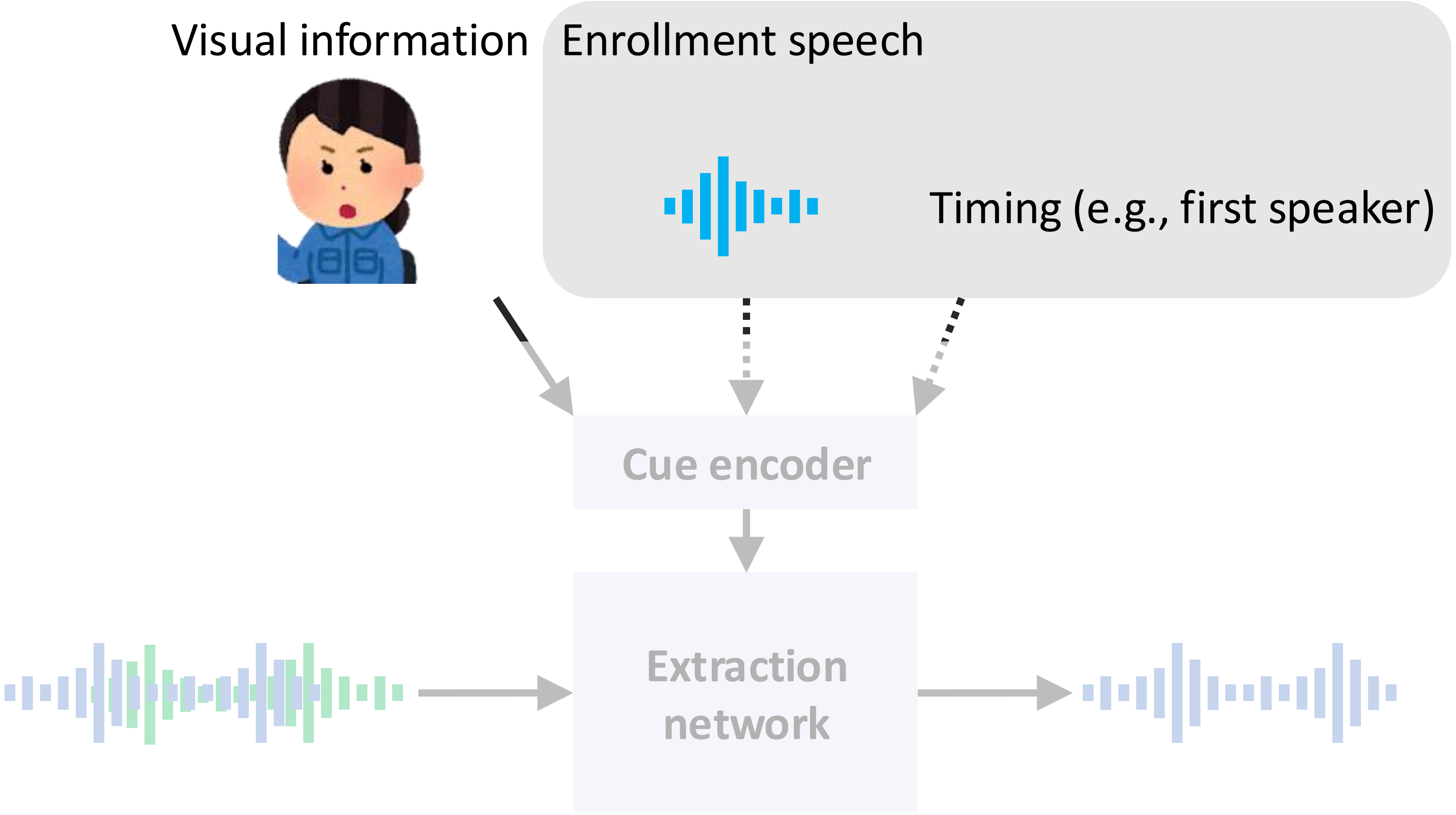
Another Approach: Target Speaker Extraction (TSE)

- DNN extracts the target speaker specified by a given cue [Delcroix+2018].
  - There is no permutation issue during training.
  - The DNN output is always one stream regardless of the number of speakers in a mixture.

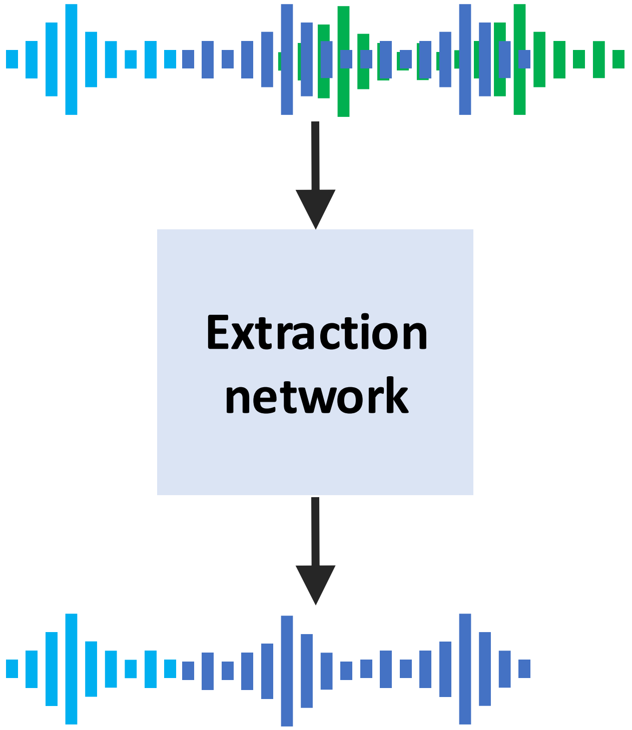


Another Approach: Target Speaker Extraction (TSE)

- DNN extracts the target speaker specified by a given cue [Delcroix+2018].
  - There is no permutation issue during training.
  - The DNN output is always one stream regardless of the number of speakers in a mixture.



Concatenate the enrollment and mixture [Shen+2025]



Agenda

- Overview of speech separation and enhancement (SSE)
- Single-channel SSE addressing permutation issue
- **Signal-processing-based multi-channel SSE and dereverberation**
- DNN-based multi-channel SSE
- Advanced topics

Mixing Process of Multi-Channel Audio

- **Room impulse responses (RIRs)** are convoluted with **dry sources** in the time domain.

$$\mathbf{y}_{k,m} = \underline{\mathbf{h}_{k,m}} \circledast \underline{\mathbf{s}_k} \in \mathbb{R}^L$$

$$\mathbf{x}_m = \sum_{k=1}^{K} \mathbf{y}_{k,m} + \mathbf{n}_m$$

- This convolutive process is approximated by an instantaneous one with STFT.

$$\mathbf{y}_k(t, f) = \underline{\tilde{\mathbf{h}}_k(f)} S_k(t, f) \in \mathbb{C}^M$$

Transfer function in the frequency domain

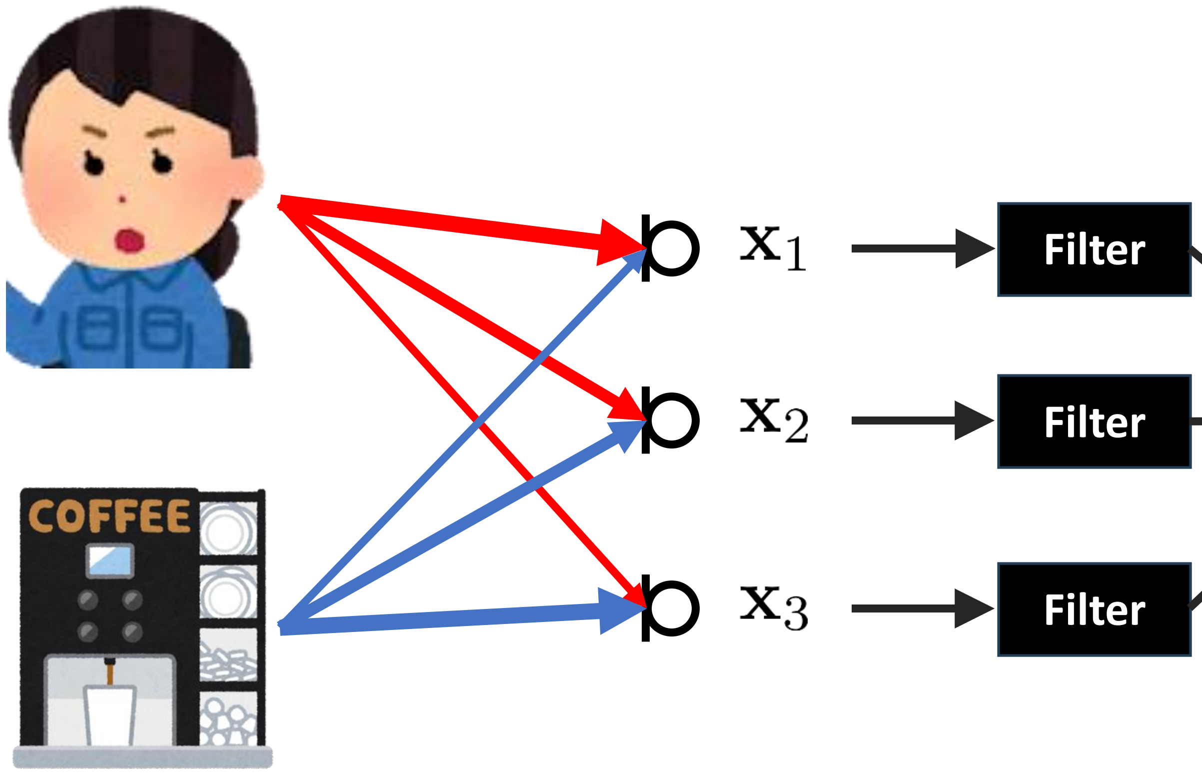
$$\mathbf{x}(t, f) = \sum_{k=1}^{K} \mathbf{y}_k(t, f) + \mathbf{n}(t, f)$$

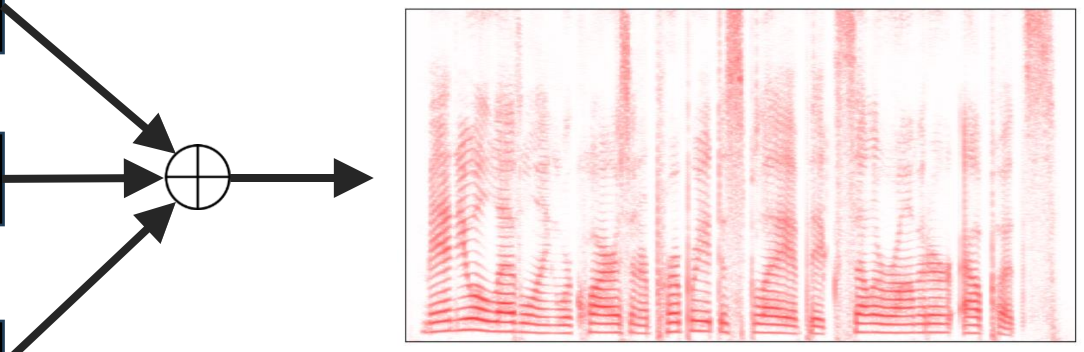
- We typically aim to predict the source image at the reference channel $Y_{k,1}(t, f)$.

Beamforming (Linear Spatial Filtering) [Trees2004, Benesty+2008]

- Beamformer suppresses interference signals using spatial information.
  - Popular beamformers retain signals coming from the direction of the target speaker.
  - Interference signals from other directions will be suppressed.

In the STFT domain, this can be approximated by a frequency band-wise inner product.

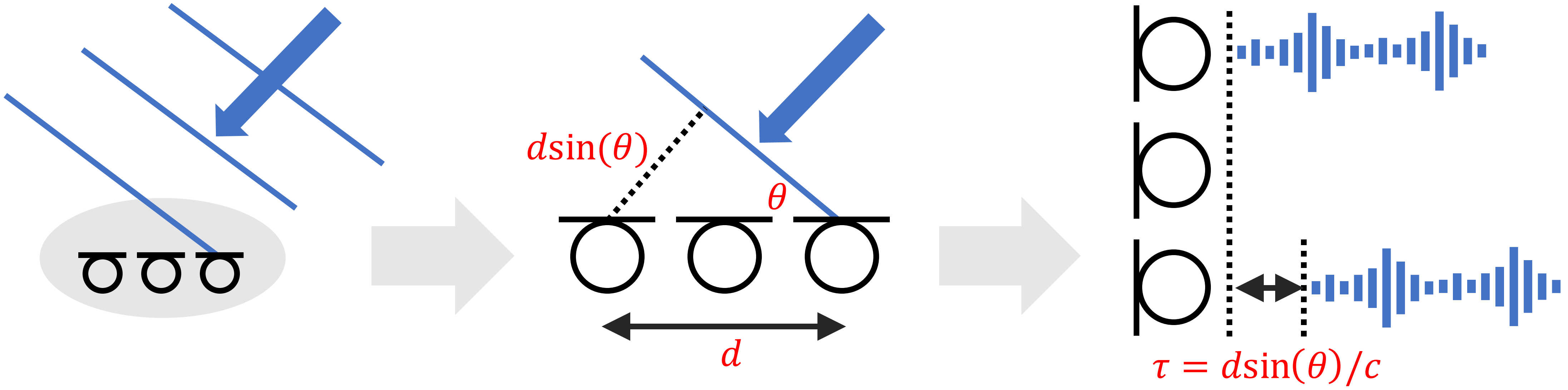




$$\hat{Y}_{k,1}(t,f) = \underbrace{\boxed{\mathbf{w}_k^{\mathsf{H}}(f)}}_{\text{Beamformer}}\mathbf{x}(t,f)$$

Delay-and-Sum (DS) Beamformer (1/2)

- DS beamformer relies on a **time-difference-of-arrival (TDoA)**.
  - Sound from the right side reaches a microphone on the right side faster.
  - TDoA can be calculated from the array geometry and sound source direction.



- Relative transfer function (RTF) is calculated from the TDoA.
  - The RTF describes the difference in sound propagation relative to the reference channel.

$$\mathbf{a}_k(f) = [1, \mathrm{e}^{-j\omega(f)\underline{\tau_{k,2}}/F}, \ldots, \mathrm{e}^{-j\omega(f)\underline{\tau_{k,M}}/F}]^\mathsf{T}$$

Delay-and-Sum (DS) Beamformer (2/2)

- DS beamformer compensates for TDoA of the target signal and takes the average.

$$\begin{aligned}\hat{Y}_{k,1}(t,f) &= \mathbf{w}_k^{\mathsf{H}}(f)\mathbf{x}(t,f) \\ &= \frac{1}{M}\mathbf{a}_k^{\mathsf{H}}(f)\mathbf{x}(t,f)\end{aligned}$$

- The target signal is preserved while the signals from other directions are suppressed.
  - Interference signals are averaged with phase differences and cancel each other.

$$\begin{aligned}\mathbf{w}_k^{\mathsf{H}}(f)\mathbf{y}_k(t,f) &= \frac{1}{M}\mathbf{a}_k^{\mathsf{H}}(f)\mathbf{y}_k(t,f) \\ &= \frac{1}{M}\mathbf{a}_k^{\mathsf{H}}(f)\left[\mathbf{a}_k(f)Y_{k,1}(t,f)\right] \\ &= Y_{k,1}(t,f)\end{aligned}$$

DS beamformer depends only on RTF (or steering vector) and **is independent of the observed signals**.
→ Its performance is typically insufficient.

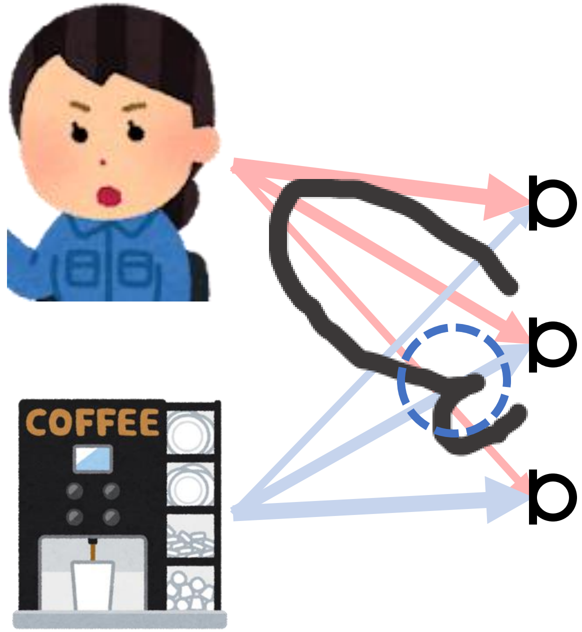
MVDR beamformer

- Minimum variance distortionless response (MVDR) beamformer minimizes the power of interference signals in the beamforming output while preserving the target signal.
  - **MVDR beamformer adaptively steers null toward other sources.**
  - It has been widely used as a front-end for ASR due to its distortionless property.

$$\min_{\mathbf{w}_k} \mathbf{w}_k^{\mathsf{H}} \mathbf{V}_{\backslash k} \mathbf{w}_k$$
$$\text{s.t. } \mathbf{w}_k^{\mathsf{H}} \mathbf{a}_k = 1,$$

$$\mathbf{w}_k = \frac{\mathbf{V}_{\backslash k}^{-1} \mathbf{a}_k}{\mathbf{a}_k^{\mathsf{H}} \mathbf{V}_{\backslash k}^{-1} \mathbf{a}_k}$$

Analytic solution



$$\mathbf{V}_{\backslash k}(f) = \sum_{k' \neq k} \mathbf{V}_{k'}(f) = \sum_{k' \neq k} \left(\underbrace{\frac{1}{T} \sum_{t=1}^{T} \mathbf{y}_{k'}(t,f) \mathbf{y}_{k'}^{\mathsf{H}}(t,f)}_{\text{Called spatial covariance matrix (SCM)}} \right)$$

Another Derivation of MVDR beamformer

- The SCM of the target signal is easier to estimate than RTF in some cases.
- The MVDR beamformer has been reformulated with SCMs.

$$\mathbf{w}_k = \frac{\mathbf{V}_{\backslash k}^{-1}\mathbf{V}_k}{\mathrm{trace}\left(\mathbf{V}_{\backslash k}^{-1}\mathbf{V}_k\right)}\mathbf{u}$$

One-hot vector indicating the reference channel

Another Derivation of MVDR beamformer

- The SCM of the target signal is easier to estimate than RTF in some cases.
- The MVDR beamformer has been reformulated with SCMs.

$$\mathbf{w}_k = \frac{\mathbf{V}_{\backslash k}^{-1}\mathbf{V}_k}{\mathrm{trace}\left(\mathbf{V}_{\backslash k}^{-1}\mathbf{V}_k\right)}\mathbf{u}$$

One-hot vector indicating the reference channel

How to estimate the SCMs and/or RTFs?

1. We can compute SCMs and RTFs from single-source segments.
2. Blind source separation (BSS) aims to estimate the spatial and source information only from the observed mixtures

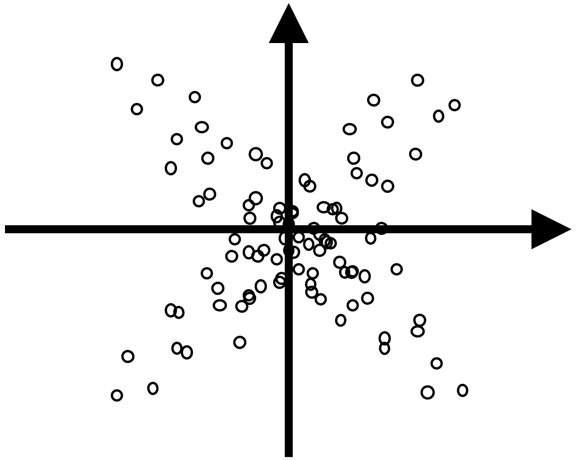
BSS Based on Spatial Probabilistic Models

- Complex Gaussian mixture model (cGMM) [Ito+2014,Otsuka+2014]
  - STFT coefficients of the observed signal are represented as the "mixture" of Gaussians.
  - cGMM assigns one source to each TF bin motivated by the sparseness of speech in the STFT domain.

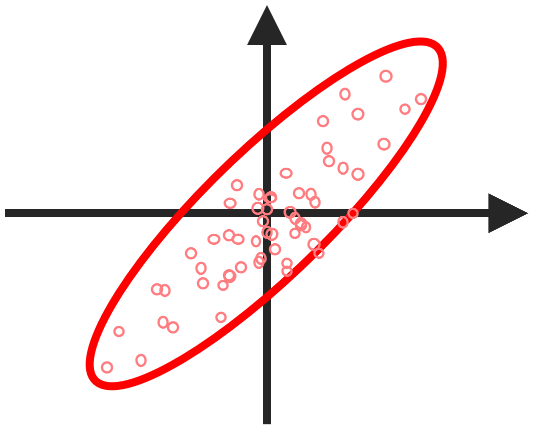
$$\mathbf{x}(t, f) \sim \sum_{k=1}^{K} \pi_k(f) \underbrace{\mathcal{N}_{\mathbb{C}}(\mathbf{0}, \lambda_k(t, f)\mathbf{V}_k(f))}_{\text{Source-wise complex Gaussian distribution with time-varying SCM}}$$

Mixing weight

- The distribution can be seen like this (not rigorous).



$$\sim \sum_{k=1}^{K} \pi_k(f)$$



Overview of cGMM Algorithm

- cGMM's data generation process is as follows:

$$\underbrace{\mathbf{z}(t,f)}_{\text{Latent speaker indicator}} \sim \mathrm{Categorical}(\boldsymbol{\pi}(f))$$

$$\mathbf{x}(t,f) \sim \prod_{k=1}^{K} \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \lambda_k(t,f)\mathbf{V}_k(f))^{z_k(t,f)}$$

- EM algorithm has been used for maximum-likelihood estimation of the parameters.
- EM algorithm alternately updates the posterior of the indicator and the parameters to maximize the evidence lower bound (ELBO).

$$\mathrm{ELBO}(q(\mathbf{Z}), \underbrace{\Theta)}_{\text{Parameters}} = \log p(\mathbf{X} \mid \Theta) - \mathrm{KL}(q(\mathbf{Z}), p(\mathbf{Z} \mid \mathbf{X}, \Theta))$$

E-step
Updating the variational posterior of the indicator

M-step
Updating the parameters

Details of CGMM Algorithm

- E-step maximizes the posterior of the indicator with the current parameters.

$$\mathrm{ELBO}(q(\mathbf{Z}), \boldsymbol{\Theta}) = \log p(\mathbf{X} \mid \boldsymbol{\Theta}) - \mathrm{KL}(q(\mathbf{Z}), p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta}))$$

$$q(\mathbf{z}(t,f))_k \leftarrow \underline{p(\mathbf{z}(t,f) \mid \mathbf{x}(t,f), \boldsymbol{\pi}(f), \lambda_k(t,f), \mathbf{V}_k(f))_k}$$

$$= \frac{\pi_k(f)\mathcal{N}(\mathbf{x}(t,f) \mid \mathbf{0}, \lambda_k(t,f)\mathbf{V}_k(f))}{\sum_{k'=1}^{K} \pi_{k'}(f)\mathcal{N}(\mathbf{x}(t,f) \mid \mathbf{0}, \lambda_{k'}(t,f)\mathbf{V}_{k'}(f))}$$

- M-step maximizes the ELBO for the parameters with the given variational posterior.

$$\lambda_k(t,f) \leftarrow \frac{1}{M}\mathrm{trace}(\mathbf{x}(t,f)\mathbf{x}^{\mathsf{H}}(t,f)\mathbf{V}_k^{-1}(f))$$

$$\mathbf{V}_k(f) \leftarrow \frac{1}{\sum_{t=1}^{T} q(\mathbf{z}(t,f))_k} \sum_{t=1}^{T} q(\mathbf{z}(t,f))_k \frac{1}{\lambda_k(t,f)} \mathbf{x}(t,f)\mathbf{x}^{\mathsf{H}}(t,f)$$

$$\pi_k(f) \leftarrow \frac{1}{T}\sum_{t=1}^{T} q(\mathbf{z}(t,f))_k$$

Guided Source Separation (GSS) [Boeddeker+2018]

- GSS guides a variant of cGMM, called cACGMM, by using **diarization results**.
  - The diarization results provide the number of speakers in a mixture.
  - The results tie together all the frequency bands, which is helpful in avoiding the permutation problem.

$$\pi_k(t,f) = \frac{\pi_k(f)\beta_k(t)}{\sum_{k'=1}^{K} \pi_{k'}(f)\underline{\beta_{k'}(t)}}$$

1 when the k'th speaker is active, 0 otherwise

- GSS has been widely used in the recent CHiME challenges with its accelerated version.

System	Diarization				Front-End	
	Segmentation	Spk-id Emb. Extr. & Clustering	Refinement	Multi-Channel Mechanism	Channel Selection	Separation
ESPnet Baseline	Pyannote Segmentation	Pyannote Diarization 2.1	- -	Top-1 Channel VAD Selection	EV	GSS
NeMo Baseline	MarbleNet VAD	Multi-scale TitaNet & NME-SC	Transformer MSDD	logit max pooling & multi-channel TitaNet embeddings	EV	GSS
STCON	TDNN stats-based	Ecapa-TDNN & wav2vec 2.0 AED multi-speaker OSD with UMAP+DBSCAN+GMM clustering	NSD-MS2S	DOVER-Lap and TS-VAD posterior averaging	EV	GSS with neural refinement (G-TSE)
NTT	EEND-VC	EEND-VC & ECAPA-TDNN with NME-SC	NSD-MS2S	DOVER-Lap & channel clustering for speaker counting	EV & Brouhaha C_{50}	GSS with SP-MWF
IACAS-Thinkit	Pyannote Segmentation	Pyannote Diarization 2.1	TS-VAD	DOVER-Lap	-	neural TSE for GSS init
USTC-NERCSLIP	Pyannote Segmentation	ECAPA-TDNN + NME-SC	NSD-MS2S	NSD-MS2S posterior averaging	EV + virtual subarray SINR	GSS

		EEND-VC	-	DOVER-Lap	-	GSS
		APA-TDNN + NME-SC	TS-VAD	TS-VAD posterior averaging	EV	GSS
		Pyannote arization 2.1	-	Top-3 Channel VAD Selection + DOVER-Lap	EV	GSS
		i-scale TitaNet & NME-SC	Transformer MSDD	logit max pooling & Multi-channel TitaNet embeddings	EV	GSS
		Pyannote arization 2.1	- -	Top-1 Channel VAD Selection	EV	GSS
		Pyannote arization 2.1	TS-VAD + d-vector refinement	TS-VAD posterior averaging	-	GSS
BUT FIT	Pyannote Segmentation	Pyannote Diarization 2.1	- -	Top-1 Channel VAD Selection	EV	GSS

Other BSS Techniques

- Independent vector analysis (IVA) [Kim+2006, Hiroe+2006]:
  - IVA assumes the determined scenario, i.e., # of spks. = # of mics., and aims to estimate the inverse of the mixing matrix (demixing matrix).

$$\mathbf{x}(t,f) = \mathbf{H}(f)[S_1(t,f),\ldots,S_K(t,f)]^{\mathsf{T}}$$

  - IVA jointly estimates the demixing matrix and the source activity based on the statistical independence between sources and that the source followed time-varying Gaussian distribution.

$$\min_{\mathbf{W}(f),\lambda_k(f)} \sum_{f=1}^{F}\left[-2\log|\det(\mathbf{W}(f))| + \sum_{k=1}^{K}\mathbf{w}_k^{\mathsf{H}}(f)\Big(\sum_{t=1}^{T}\frac{\mathbf{x}(t,f)\mathbf{x}^{\mathsf{H}}(t,f)}{2T\lambda_k(f)}\Big)\mathbf{w}_k(f)\right]$$

Ideally, the demixing matrix will converge to the inverse of the mixing matrix.

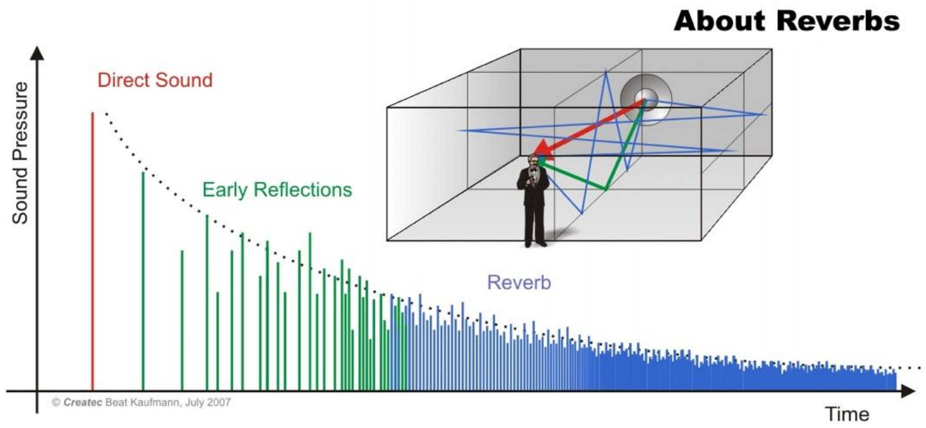
  - Majorization minimization algorithms have been widely used to solve the optimization problems of IVA and its variants.

Weighted Prediction Error (WPE) for Dereverberation [Nakatani+2010]

- Reverberation degrades speech intelligibility and makes speech recognition harder.

$$\mathbf{x}(t,f) = \mathbf{y}(t,f) + \mathbf{r}(t,f)$$

$$= \mathbf{y}(t,f) + \underline{\sum_{\tau=\Delta_{\min}}^{\Delta_{\max}} \tilde{\mathbf{h}}(\tau,f) S(t-\tau,f)}$$



- WPE suppresses the late reverberation by predicting it from the past observation.

$$\min_{\mathbf{W}(\tau,f),\lambda(t,f)} \frac{\|\hat{\mathbf{y}}(t,f)\|_2^2}{\lambda(t,f)} - \log \lambda(t,f)$$

$$\text{s.t.} \quad \hat{\mathbf{y}}(t,f) = \mathbf{x}(t,f) - \sum_{\tau=\Delta_{\min}}^{\Delta_{\max}} \mathbf{W}(\tau,f)\mathbf{x}(t-\tau,f)$$

WPE has been widely used in the CHiME challenges.

Pros and Cons in Signal-Processing-Based Methods

- Pros: We do not have to care about the train-test mismatch.
  - Signal-Processing-Based methods adapt the model to each scene.
  - Spatial probabilistic models work well as a prior.

- Cons: The assumption in spatial models might not be satisfied in complex situations.
  - Many models assume the scene is static, i.e., speakers do not move around.
  - Manually-designed speech models, e.g., sparsity, have a gap from real speech characteristics.

- Cons: Signal-Processing-Based methods rely too much on the spatial information.
  - Their performance is limited when the number of microphones are limited.
  - IVA variants are not applicable to the underdetermined situation.

Agenda

- Overview of speech separation and enhancement (SSE)
- Single-channel SSE addressing permutation issue
- Signal-processing-based multi-channel SSE and dereverberation
- **DNN-based multi-channel SSE**
- Advanced topics

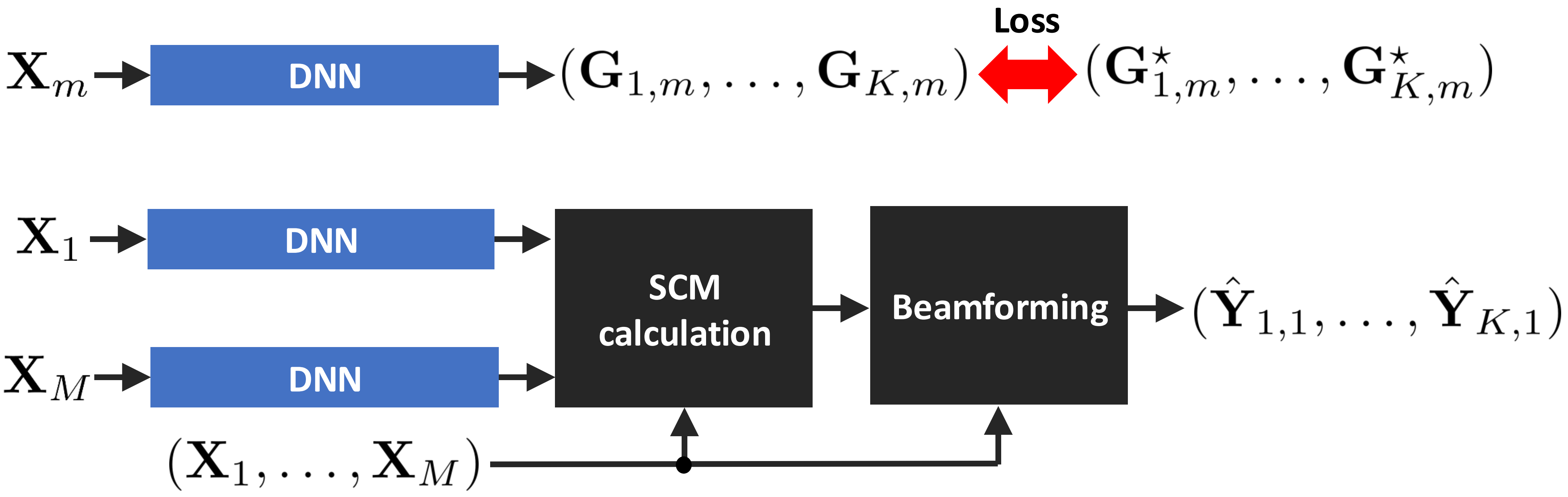
Application of Mask Estimation Network [Heymann+2016, Erdogan+2016]

- TF masks estimated by a DNN have been used to compute SCMs.
  - TF masks indicate the TF bins dominated by the target source.

$$\hat{\mathbf{V}}_k(f) = \frac{1}{\sum_{t=1}^{T} \underline{\mathcal{G}}_k(t,f)} \sum_{t=1}^{T} \underline{\mathcal{G}}_k(t,f)\mathbf{x}(t,f)\mathbf{x}(t,f)^{\mathsf{H}}$$

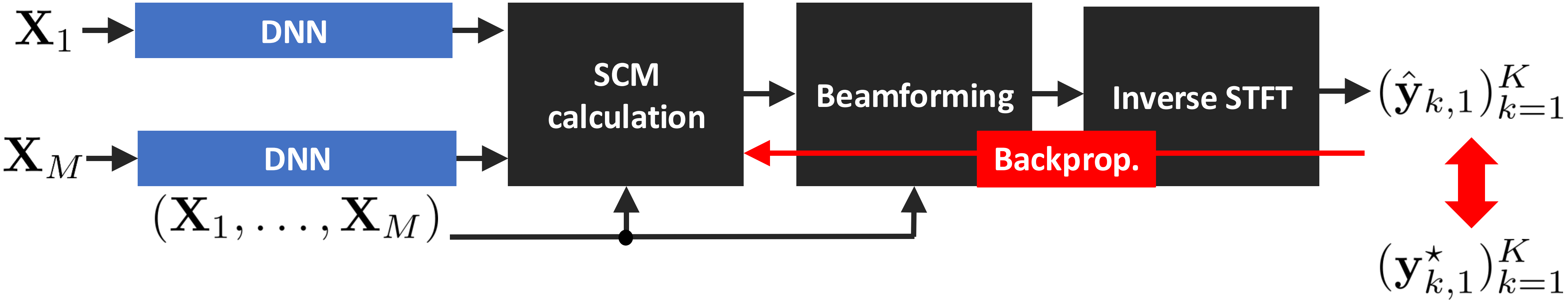
TF masks averaged over channels

- We can use DNNs pre-trained on single-channel data with a mask-level loss.



End-to-End Training of Mask-based beamforming

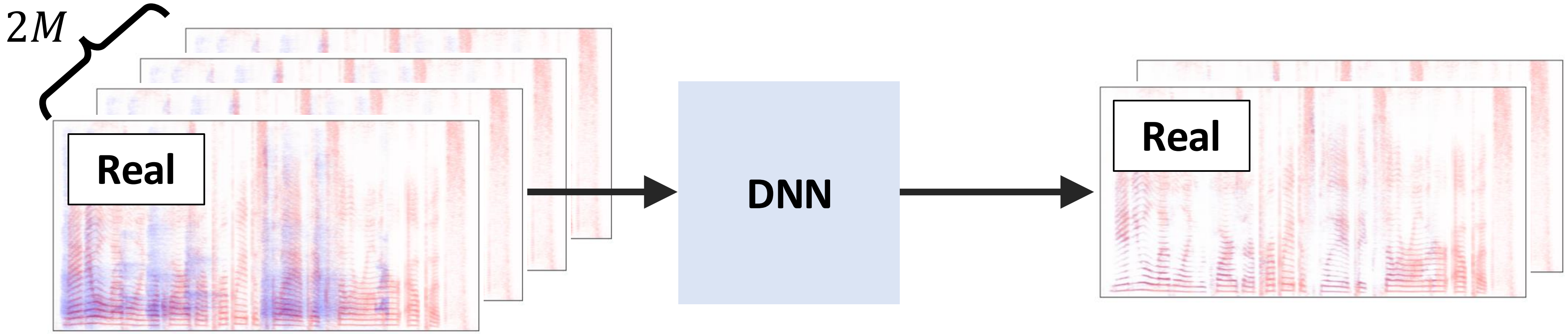
- Ideal masks for single-channel TF masking are not optimal for estimating SCMs.
- We can backprop a signal-level loss through beamforming.
  - The DNNs will be optimized for the SCM estimation.
  - In my experience, TF masks become more sparse (selecting TF bins not contaminated by other sources).



- Mask-based beamforming inherits the pros and cons of beamforming.
  - Pros: It is robust to the domain mismatch and compatible with different microphone arrays.
  - Cons: Its performance is still limited in underdetermined situations with diffuse noise.

Multi-Channel Complex Spectral Mapping [Wang+2020,2021]

- A DNN directly estimates complex STFT coefficients.
  - Input: Real and imaginary part of **the mixture at all the channels**.
  - Output: Real and imaginary part of each source image at the reference channel.

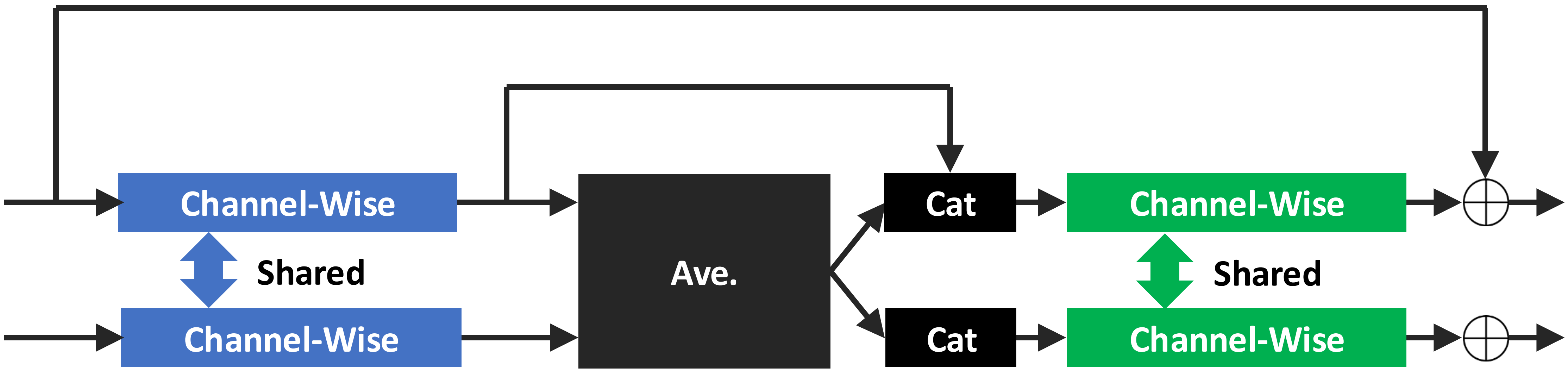


- DNN performs time-varying non-linear spatial processing.
  - Pros: It can suppress the interference sources more aggressively than classical beamforming.
  - Cons: It introduces processing artifacts that are harmful for ASR.
  - Cons: It is less robust to the domain mismatch, e.g., array-geometry, speaker-microphone distances, …

Transform-Average-Concatenate (TAC) [Luo+2020]

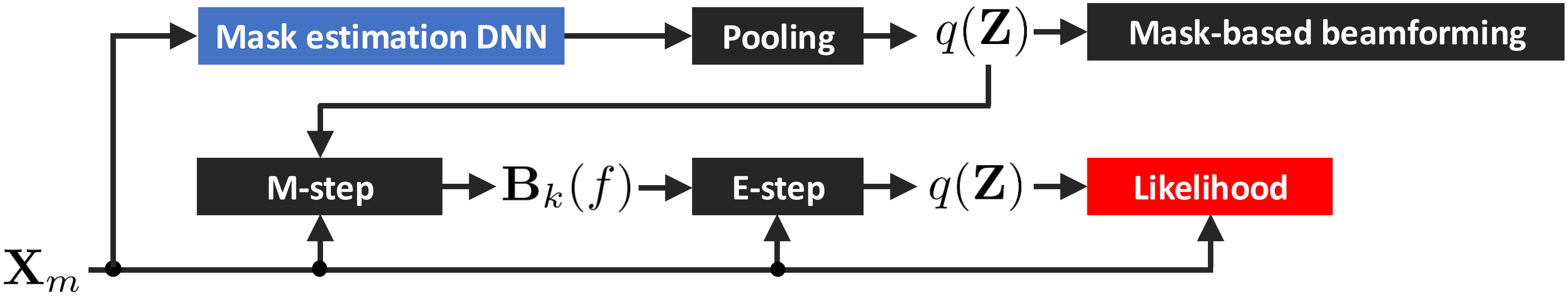
- We would like to handle various array configurations with a single model.
  - Concatenation of STFTs for each channel can not generalize to different numbers of microphones.

- TAC uses average pooling across channels to handle arbitrary numbers of microphones.
  - TAC transforms features in a channel-wise manner and takes the average of them.
  - The averaged feature is further processed and concatenated with the channel-wise feature.



Unsupervised Training with The cACGMM objective [Drude+2019]

- A DNN for mask estimation is trained based on the likelihood of cACGMM.
  - We need only multi-channel mixtures as training data, i.e., unsupervised training.



- The DNN is expected to learn speech characteristics and overlapping patterns.
  - This approach has a potential to outperform pure signal-processing-based methods.

		CHiME-4 WER (%)
cACGMM		13.06
DNN	Supervised	7.71
DNN	Unsupervised	7.80

Unsupervised Training with Full-Rank Component Analysis (FCA)

- Neural FCA [Bando+2021] trains a large VAE in which the decoder is based on a well-developed spatial probabilistic model (FCA).
  - Its objective is reconstruction of the input mixture like VAE (the maximization of ELBO).
  - Each source is estimated by time-varying multi-channel Wiener filter.

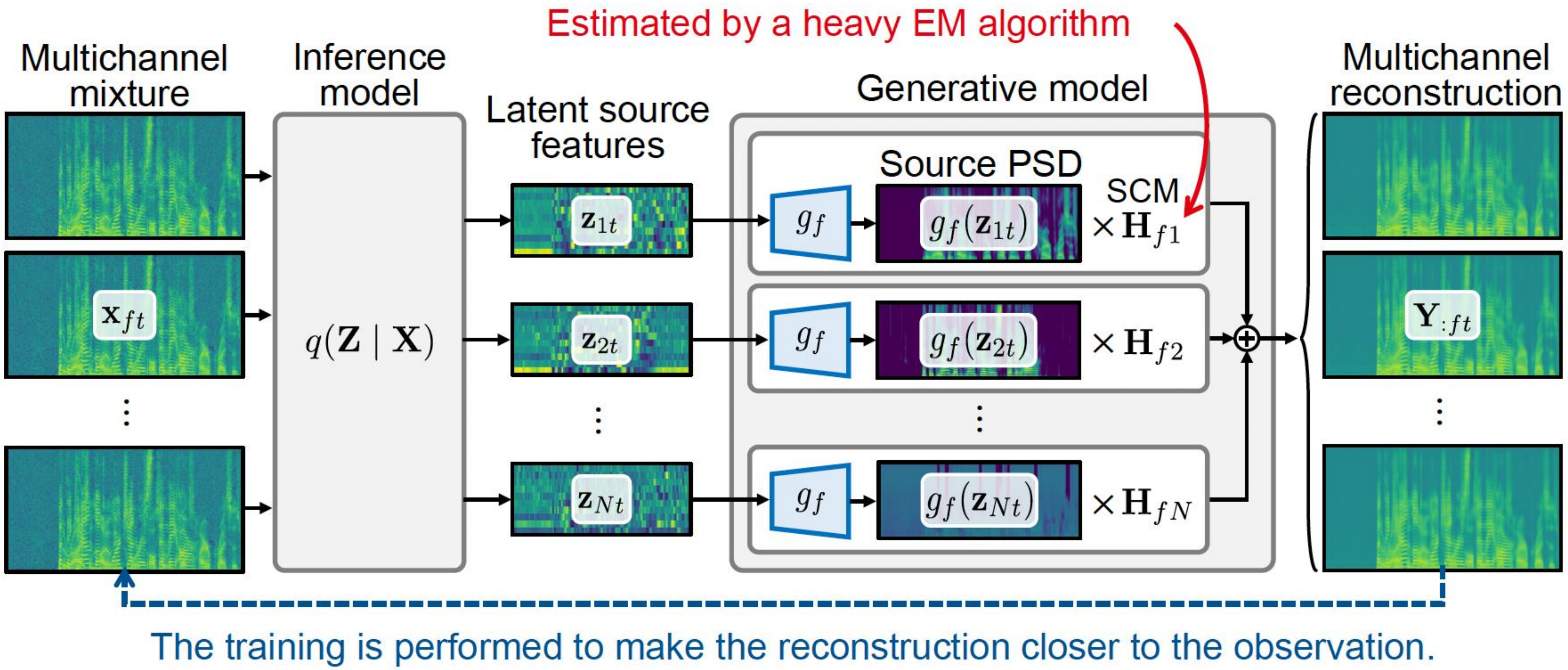
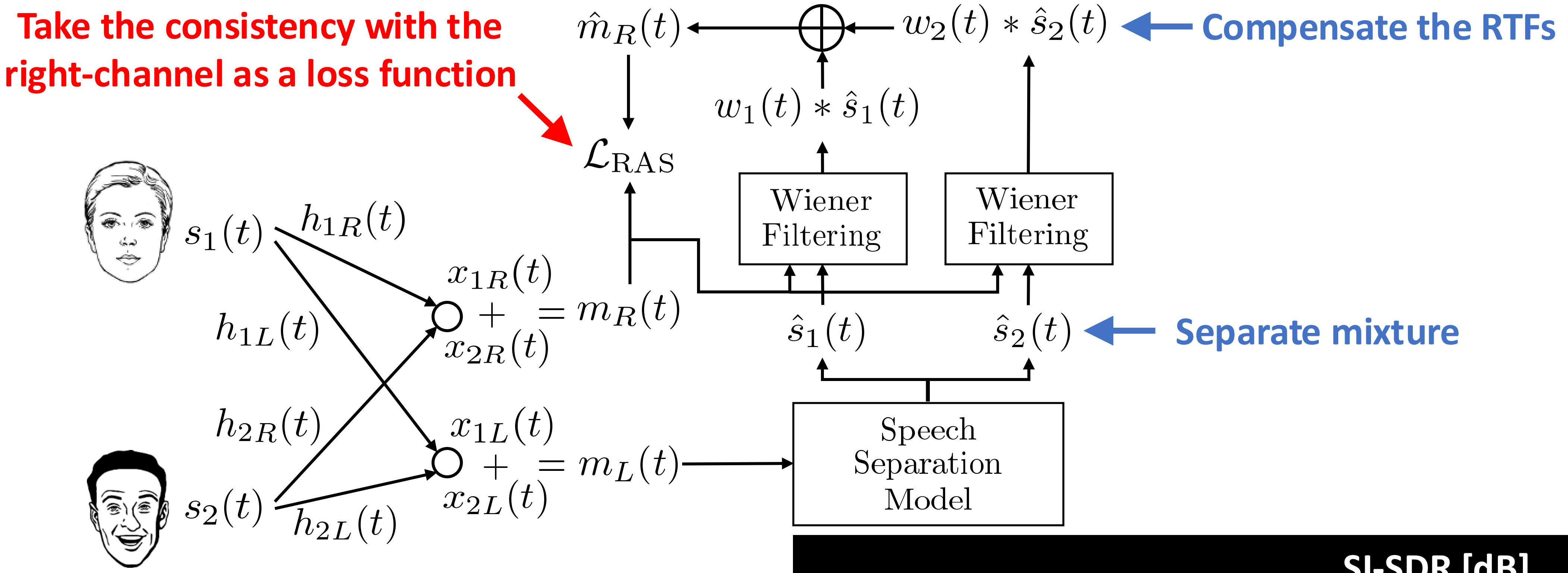


Figure is from https://speakerdeck.com/yoshipon/eusipco-2023-neural-fast-full-rank-spatial-covariance-analysis-for-blind-source-separation

Unsupervised Training Using Reverberation as Supervision (RAS)

- RAS leverages two-channel mixtures to train a monaural separation model [Aralikatti+2023].
- The DNN is trained to separate the left-channel mixture so that the right-channel mixture can be reconstructed from the separated sources.



- RAS and its variants can train any SSE models (complex spectral mapping).

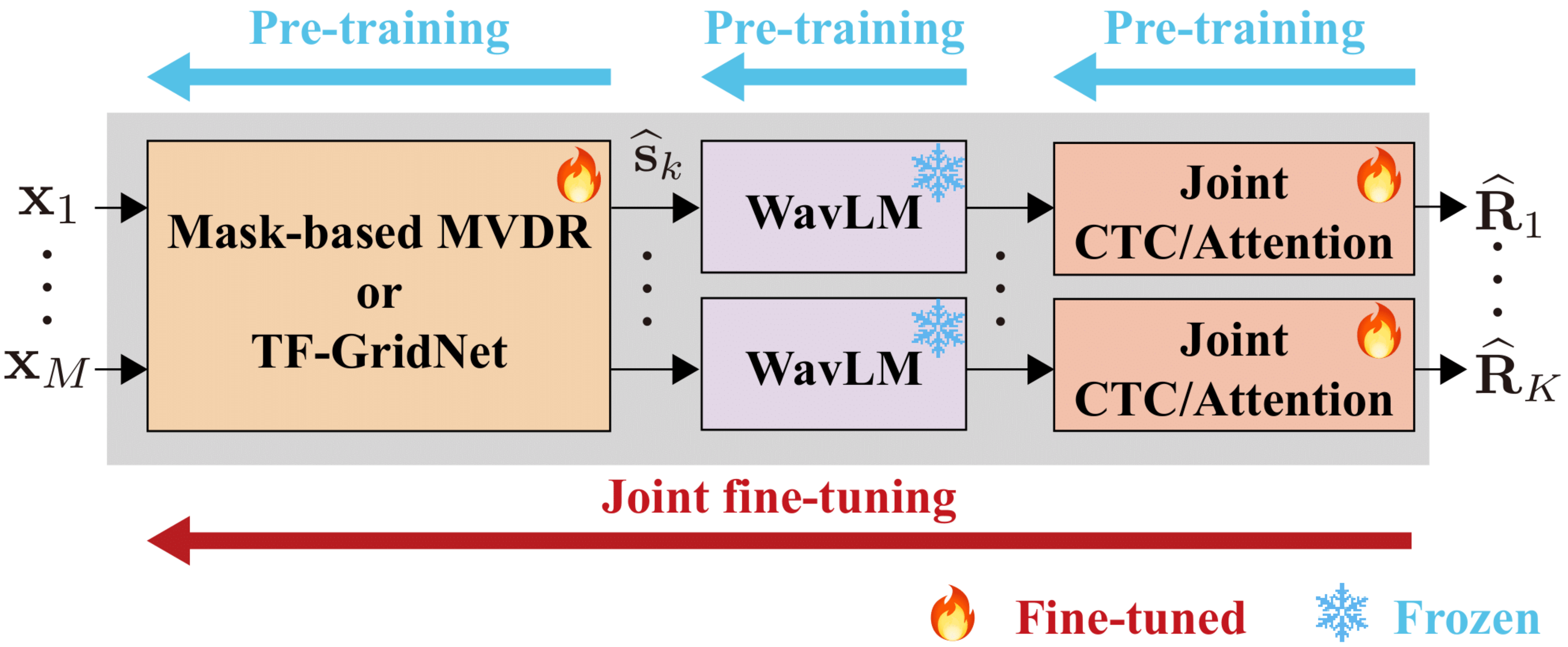
	SI-SDR [dB]	PESQ
Enhanced RAS [Saijo+2024]	13.9	3.55
Supervised	15.8	3.89

Agenda

- Overview of speech separation and enhancement (SSE)
- Single-channel SSE addressing permutation issue
- Signal-processing-based multi-channel SSE and dereverberation
- DNN-based multi-channel SSE
- **Advanced topics**

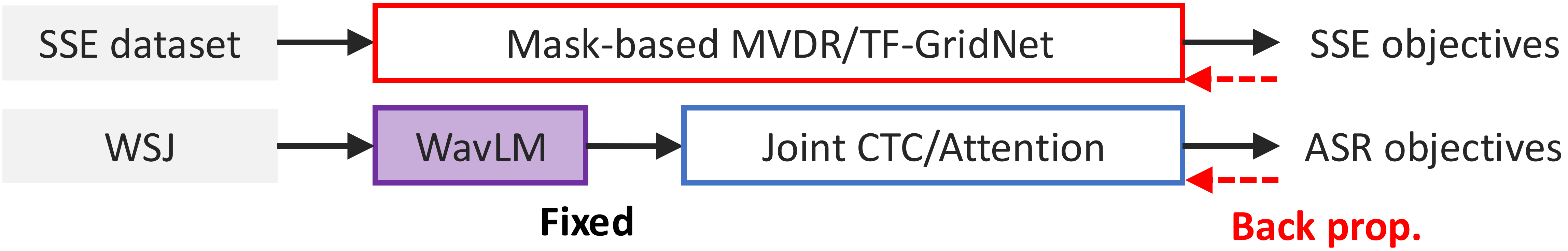
Overview of SIMO- and MIMO-IRIS [Masuyama+2025]

- SSE models trained with popular signal-level loss is not optimal as a frontend for ASR.
  - Artifacts caused by SSE is very harmful for ASR [Iwamoto+2022].
- Integrating SSE and ASR models into an end-to-end system [Ochiai+2017, von neumann+2020].
  - The SSE model will be optimized as a front-end for ASR.
  - The ASR model will be aware of artifacts from imperfect separation.

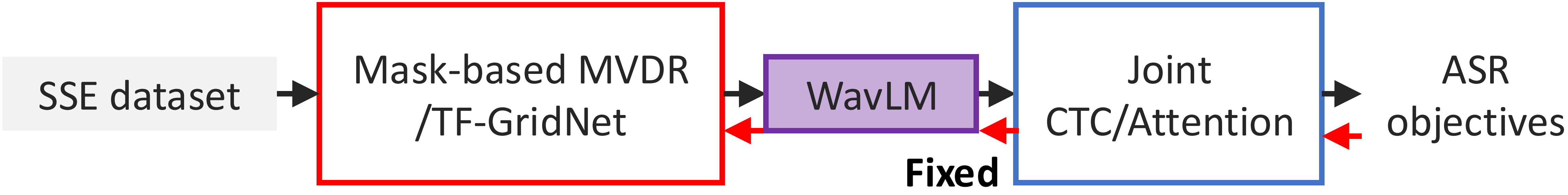


Training of SIMO- and MIMO-IRIS

- The modularity allows us to leverage pre-trained models.
- We first pre-train the SSE and ASR models with popular loss functions.



- Then, the SSE and ASR models are integrated and fine-tuned in an end-to-end manner.



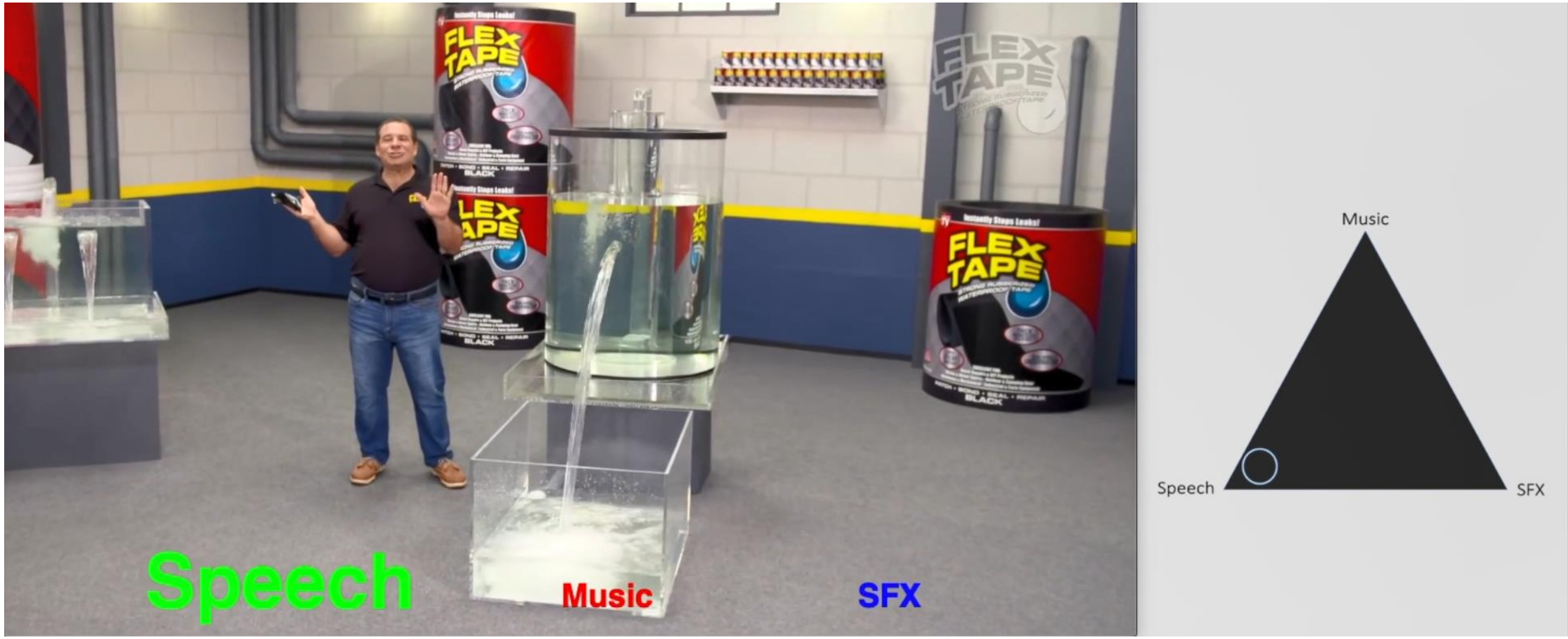
Results of SIMO-/MIMO-IRIS on WHAMR! dataset

- SDR and WER under a noisy reverberant condition.
  - The integration of the SOTA models works well, at least on static in-domain data.
  - **Joint fine-tuning improved WER** further but **degraded SDR** (a signal-level SSE metric).
  - WavLM fine-tuning easily overfitted to training data compared with SSE fine-tuning in my experience.

	Fine-tuned modules	SDR [dB]	WER (%)
Monaural TF-GridNet / WavLM	No	9.0	11.6
Monaural TF-GridNet / WavLM	ASR	9.0	5.7
Monaural TF-GridNet / WavLM	SSE, ASR	**4.0**	**3.1**
Two-channel TF-GridNet / WavLM	No	11.1	8.3
Two-channel TF-GridNet / WavLM	ASR	11.1	3.9
Two-channel TF-GridNet / WavLM	SSE, ASR	**7.9**	**2.3**
Mask-based beamforming / Fbank [Zhang+2022]	SSE, ASR	-2.27	28.9

General Audio Source Separation

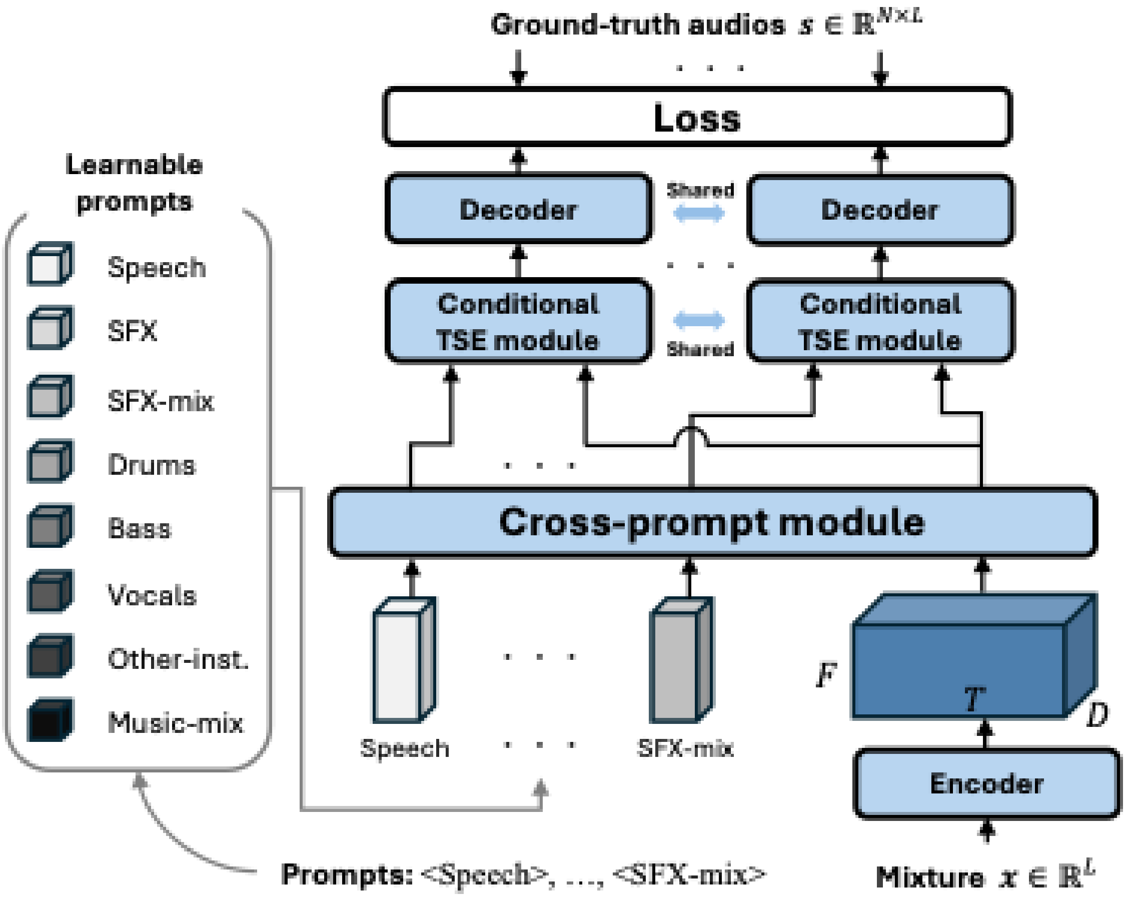
- We would like to separate any types of sound of interest.
  - Music source separation: Vocal, Bass, Drums, Other instruments
  - Universal source separation: Sound effects (dog barking, wind noise, …)
  - Cinematic audio source separation: Speech, Music, Mixture of sound effects/events



This demo is provided by Gordon Wichern

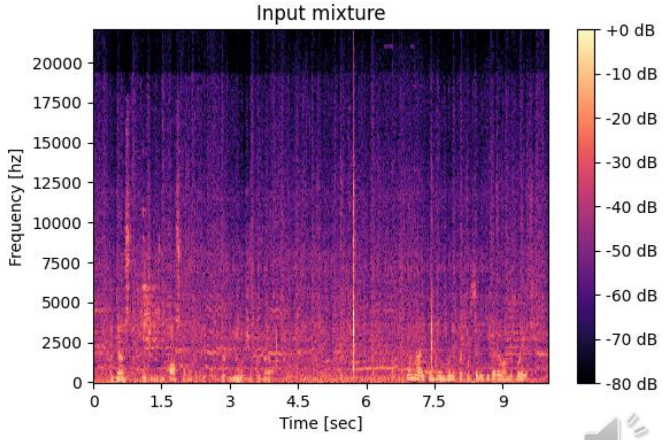
Task-Aware Unified Source Separation (TUSS) [Saijo+2025]

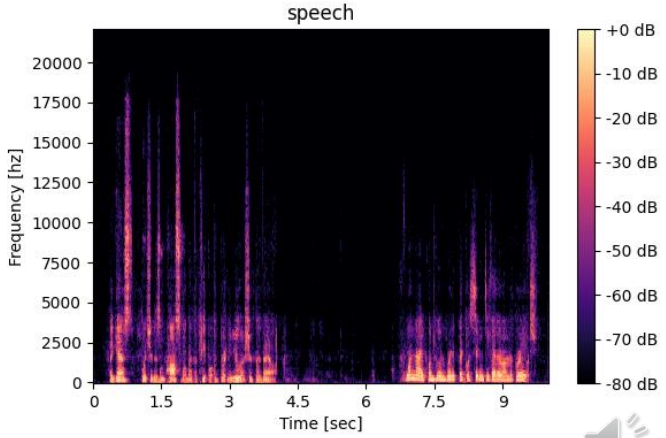
- TUSS controls separation outputs with learnable prompts.
  - Various separation tasks are covered by combinations of prompts.
  - The cross-prompt module performs MHSA across prompts and embeddings of the mixture.

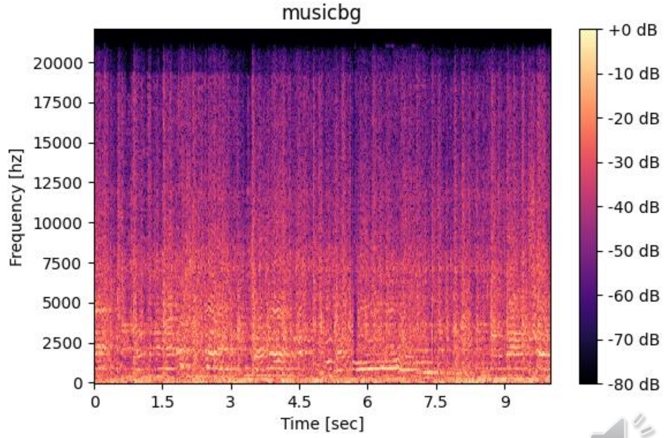


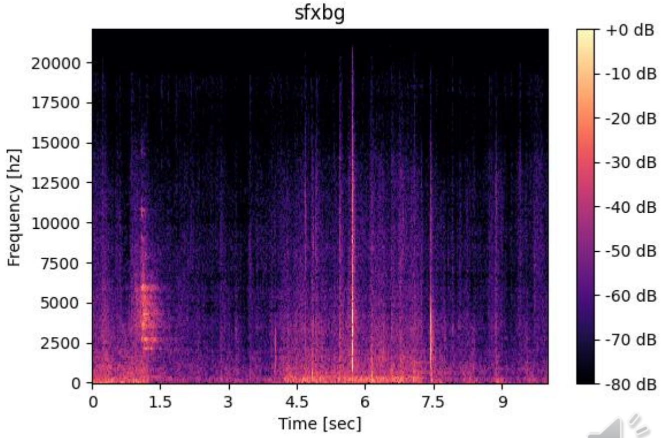
TUSS Demo (1/3)

- **TUSS output with prompts: <Speech>, <Music-mix>, <SFX-mix>**



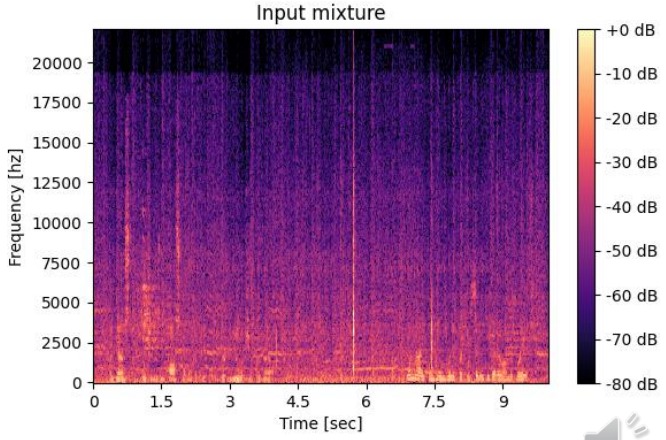


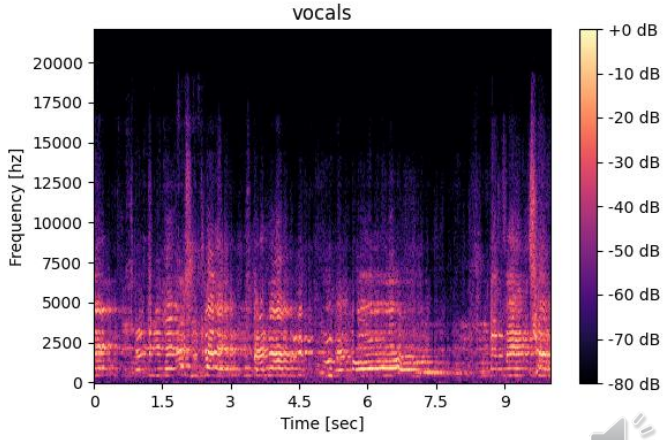


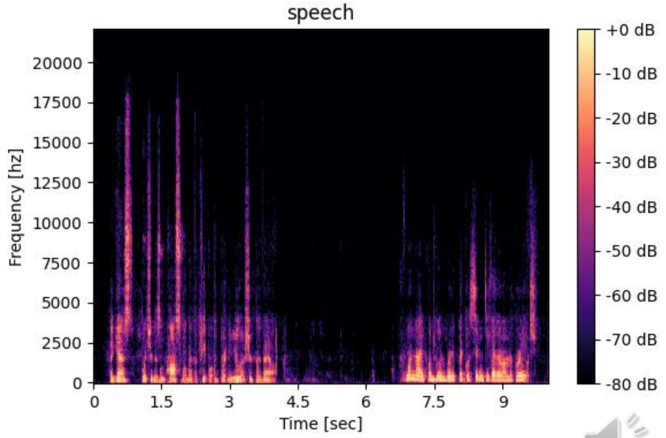


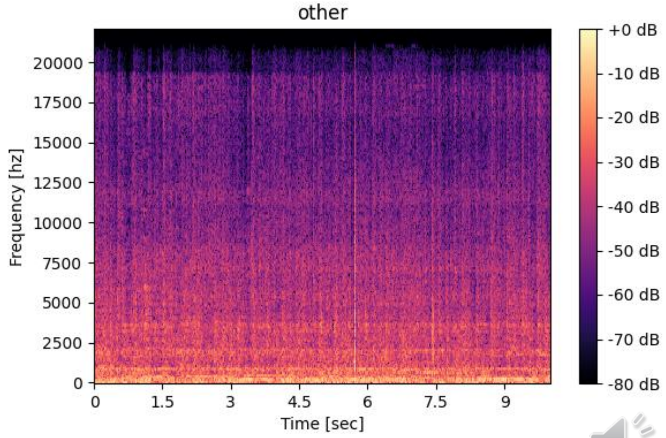
TUSS Demo (2/3)

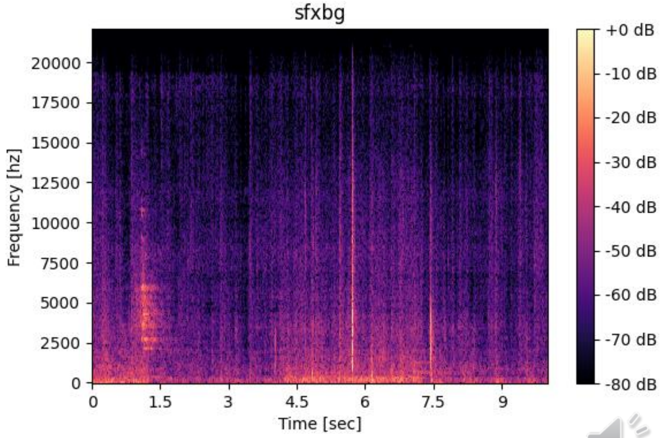
- **TUSS output with prompts: <Speech>, <Vocals>, <Other inst.>, <SFX-mix>**





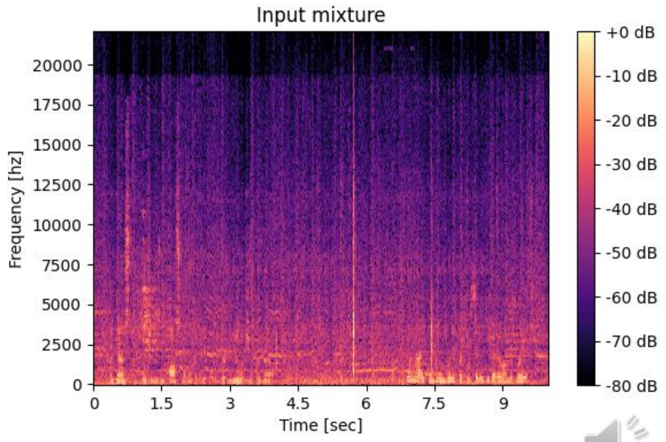


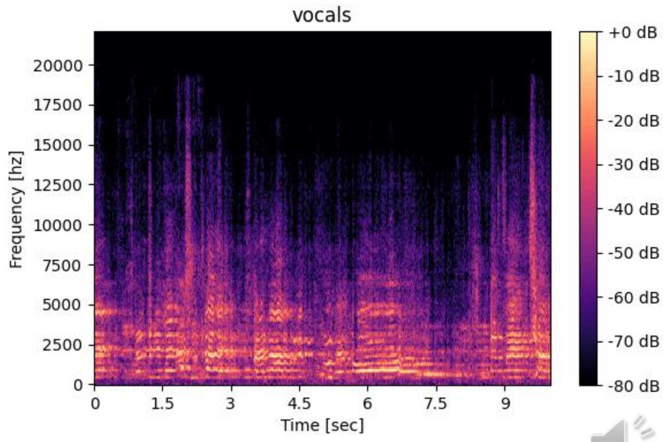


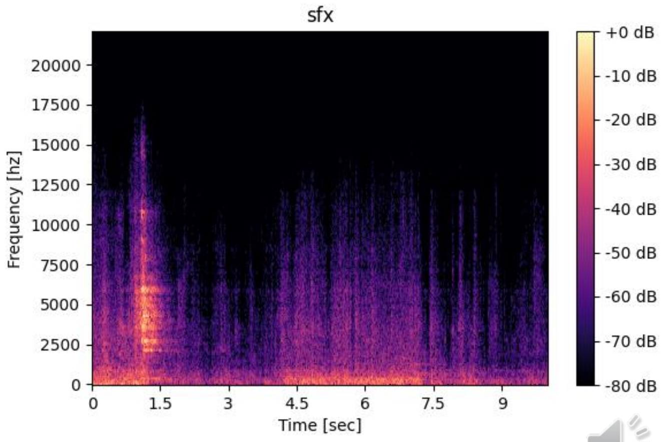


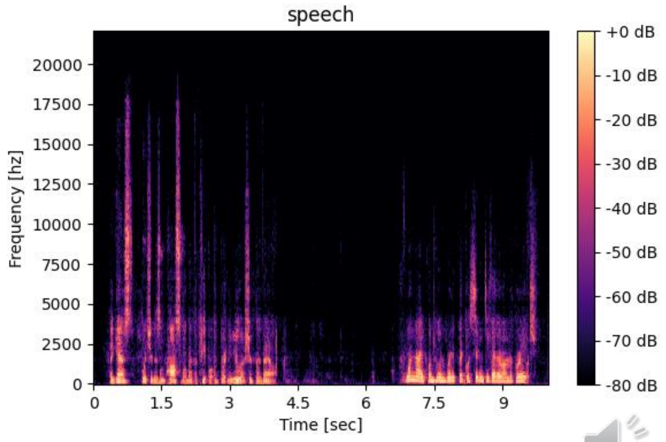
TUSS Demo (3/3)

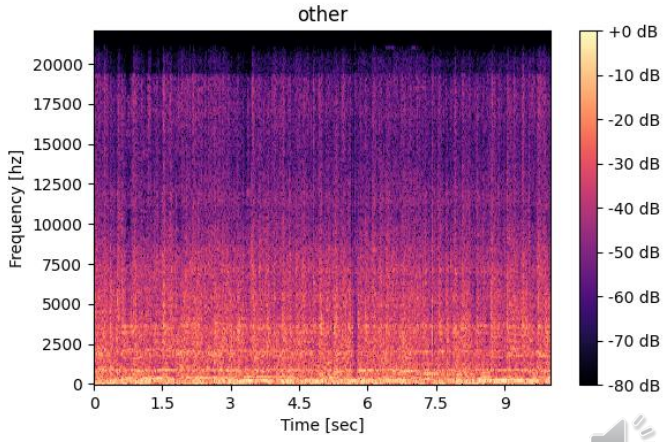
- **TUSS output with prompts: <Speech>, <Vocals>, <Other inst.>, <SFX>, <SFX>**

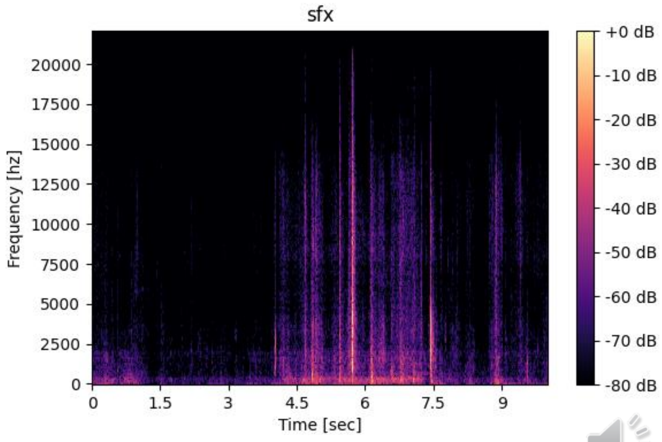












Conclusion

- Performance of SSE has been dramatically improved.
  - PIT enables us to train speaker separation networks in a supervised manner.
  - Complex spectral mapping with TF dual-path modeling has shown promising results on benchmarks.

- Hybrids of DNNs and signal processing are still preferred for separating real conversations.
  - GSS, cACGMM conditioned by (neural) diarization, is a standard in the recent CHiME challenge series.

- Unsupervised training based on spatial information is an active research topic.
  - Spatial probabilistic models (cGMM and FCA) have been leveraged to derive loss functions.
  - RAS and its variants are based on a weaker non-probabilistic model.
  - Unsupervised training with single-channel data is also an active topic, e.g., MixIt [Wisdom+2020].

- Real-world data is still challenging!
  - Domain mismatch between artificially generated mixtures and real far-field conversation recordings
  - Dynamic situation (moving speakers and/or microphones, variable number of speakers)

MITSUBISHI
ELECTRIC

Changes for the Better