Study on Overlap-Aware Speaker Diarization and Its Applications

（発話の重複を考慮した話者ダイアライゼーションとその応用に関する研究）

Shota Horiguchi

Multimedia Laboratory
Graduate School of Systems and Information Engineering
University of Tsukuba

Multi-Talker ASR

- Automatic speech recognition (ASR) will contribute to:
  - Solve a labor shortage (if incorporated with dialog system, text mining, etc.)
  - Improve human well-being by freeing humans from simple labors (e.g., transcription, documentation)
  - Ease a language barrier (if incorporated with translation)

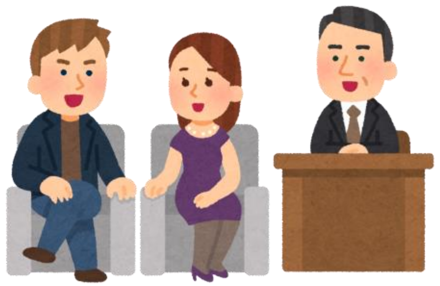
- Situations where we want to transcribe speech usually have multiple speakers

  → Not only "what was spoken" but also "**who spoke when**" is essential

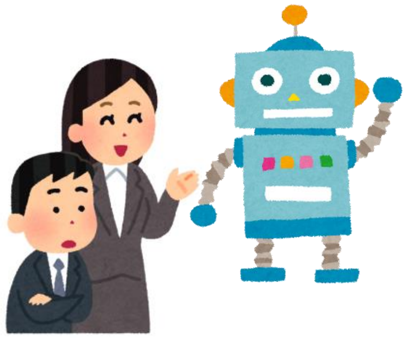
Call center

Doctor-patient conversation

Meeting



TV show

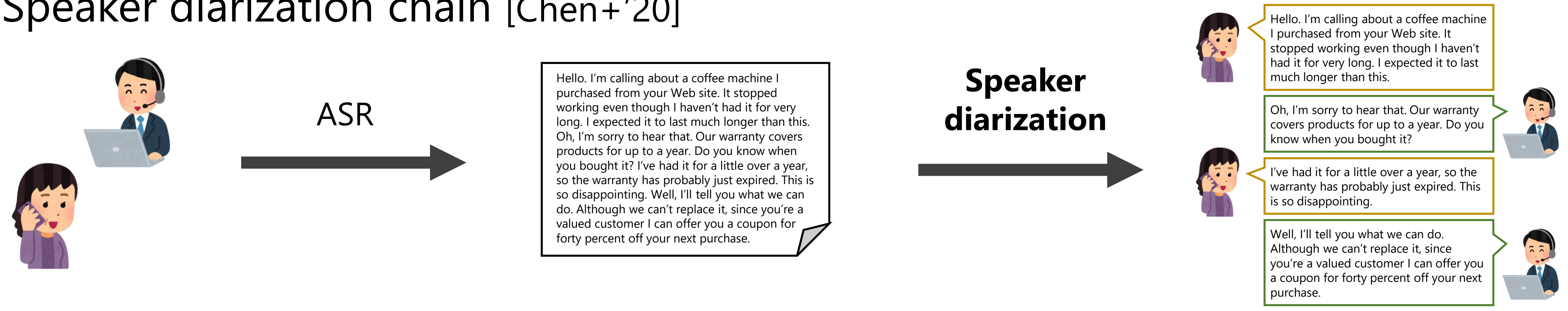


Human-robot interaction

Speaker Diarization in Multi-Talker ASR

Speaker diarization (who spoke when) plays an essential role in multi-talker ASR

- ASR → Speaker diarization chain [Chen+'20]



Hello. I'm calling about a coffee machine I purchased from your Web site. It stopped working even though I haven't had it for very long. I expected it to last much longer than this. Oh, I'm sorry to hear that. Our warranty covers products for up to a year. Do you know when you bought it? I've had it for a little over a year, so the warranty has probably just expired. This is so disappointing. Well, I'll tell you what we can do. Although we can't replace it, since you're a valued customer I can offer you a coupon for forty percent off your next purchase.

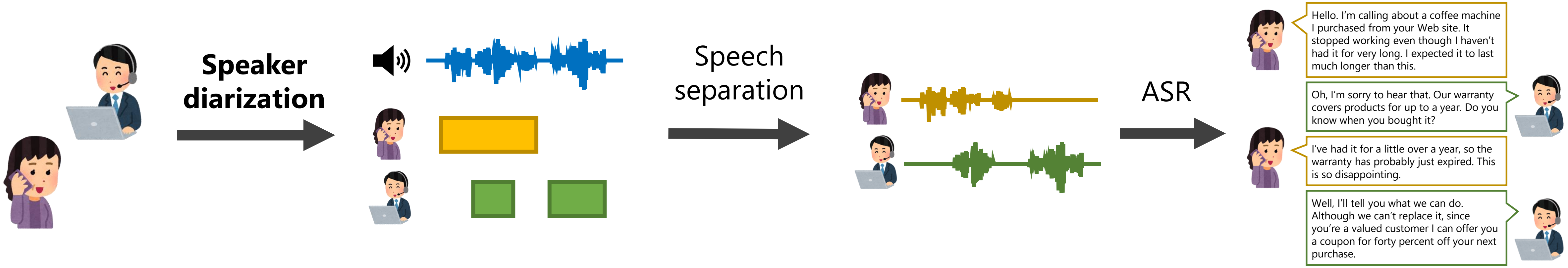
Hello. I'm calling about a coffee machine I purchased from your Web site. It stopped working even though I haven't had it for very long. I expected it to last much longer than this.

Oh, I'm sorry to hear that. Our warranty covers products for up to a year. Do you know when you bought it?

I've had it for a little over a year, so the warranty has probably just expired. This is so disappointing.

Well, I'll tell you what we can do. Although we can't replace it, since you're a valued customer I can offer you a coupon for forty percent off your next purchase.

- Speaker diarization → Speech separation → ASR chain [Watanabe+'20]



Hello. I'm calling about a coffee machine I purchased from your Web site. It stopped working even though I haven't had it for very long. I expected it to last much longer than this.

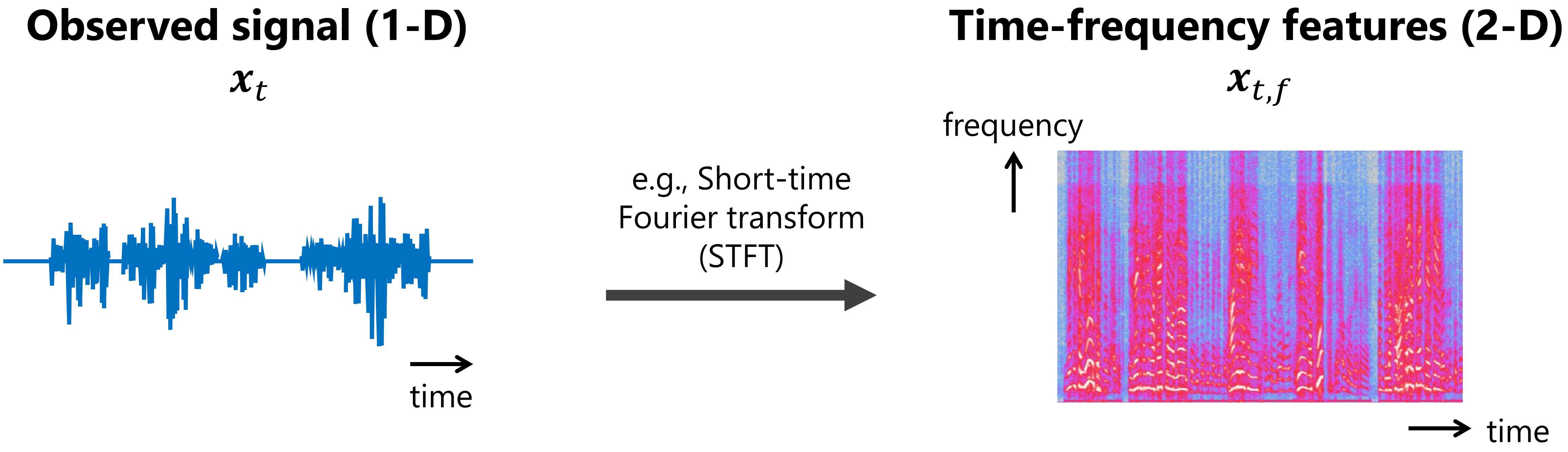
Oh, I'm sorry to hear that. Our warranty covers products for up to a year. Do you know when you bought it?

I've had it for a little over a year, so the warranty has probably just expired. This is so disappointing.

Well, I'll tell you what we can do. Although we can't replace it, since you're a valued customer I can offer you a coupon for forty percent off your next purchase.

* The conversational text was taken from TOEIC sample questions. URL: https://www.iibc-global.org/toeic/toeic_program/sample_all.html#L3

Preliminary



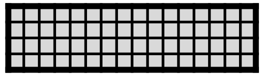
Observed signal (1-D)

$\boldsymbol{x}_t$

e.g., Short-time Fourier transform (STFT)

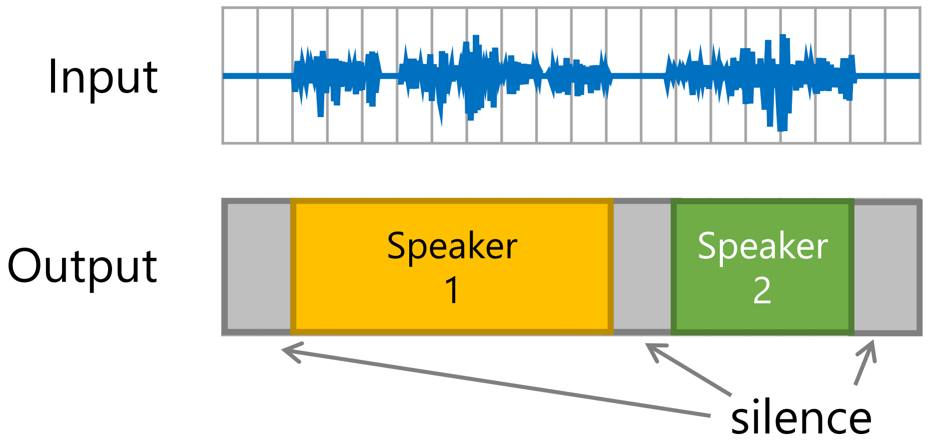
Time-frequency features (2-D)

$\boldsymbol{x}_{t,f}$

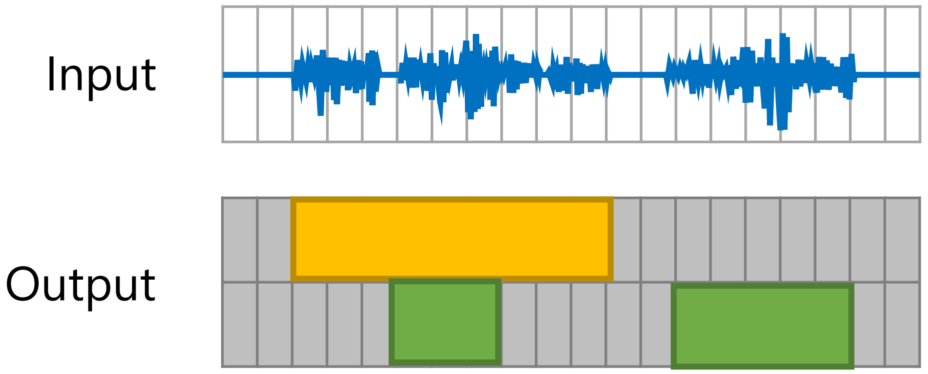
- Discussions in this presentation are based on **time-frequency features**
- We illustrates a sequence of features like : 

Two Problem Definitions for Speaker Diarization

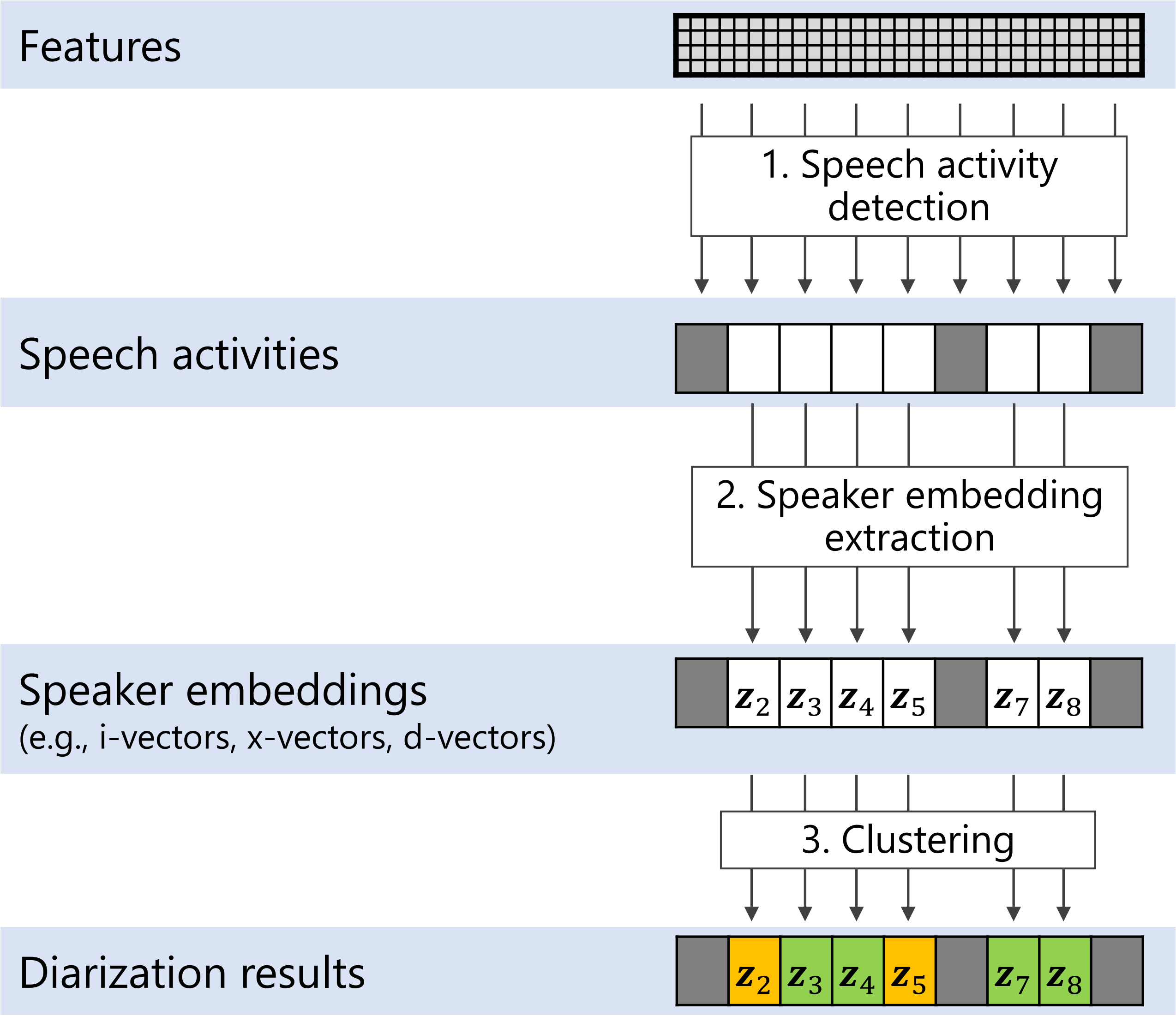
- **Set-partitioning problem**
  - Assign a single speaker or silence to each time frame
  - Adopted by most cascaded approaches



- **Multi-label classification problem**
  - Estimate all the active speakers for each time frame
  - Adopted by most end-to-end approaches



Cascaded Approach [Sell+'14] [Landini+'22]



- **Method**

  Cascade of the following:

  1. Speech activity detection
  2. Speaker embedding extraction
  3. Clustering
  4. (Optional) overlap detection and speaker assignment

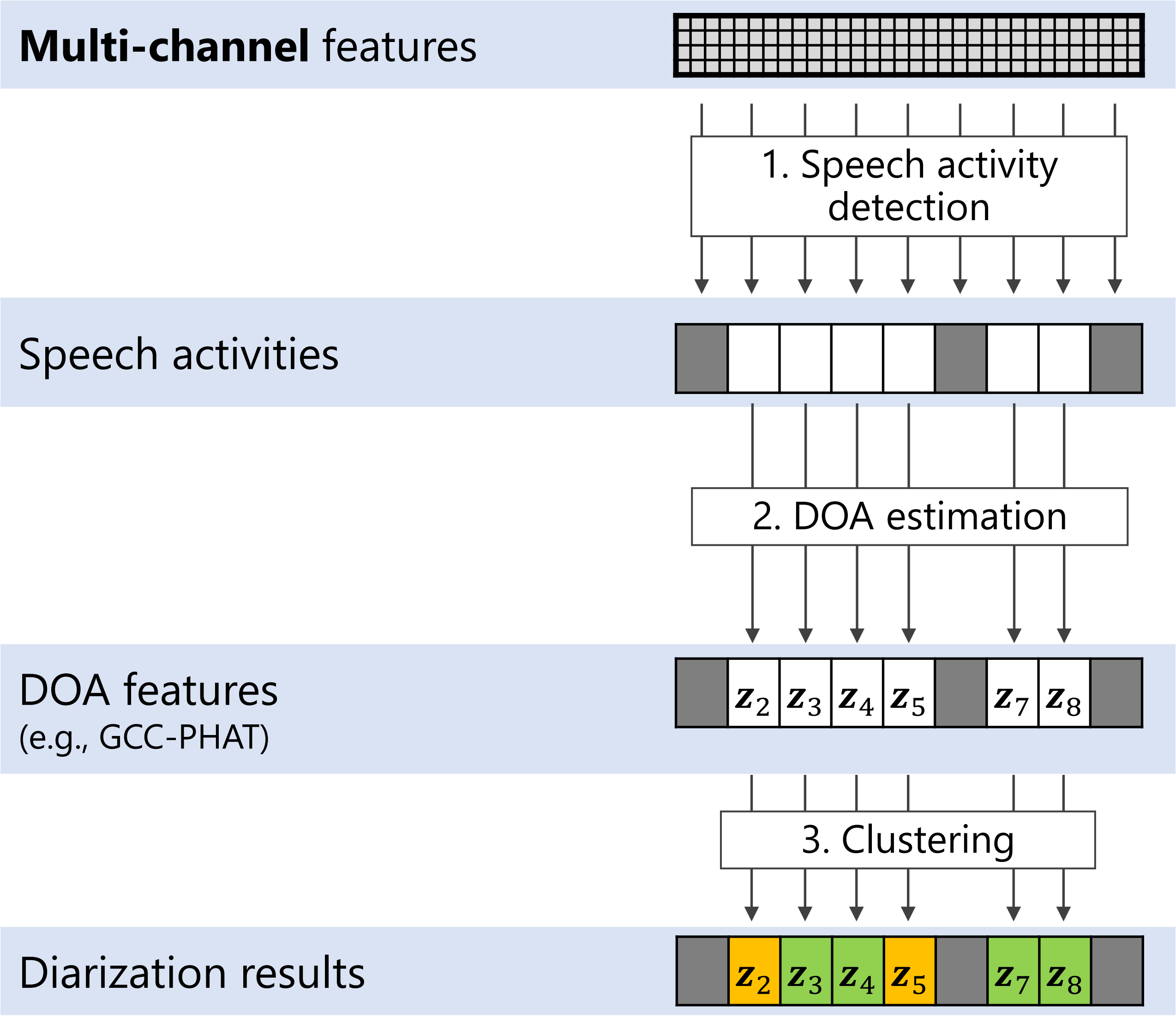
- **Pros & Cons**

  ✘ Complicated pipeline

  ✘ Cannot handle speaker overlap (without additional modules)

  ✓ The number of speakers can be set flexibly in the clustering step

Direction-of-Arrival-based Approach [Araki+’08] [Ishiguro+’11]



- **Method**

  Cascade of the following:

  1. Speech activity detection
  2. Direction-of-arrival (DOA) estimation
  3. Clustering

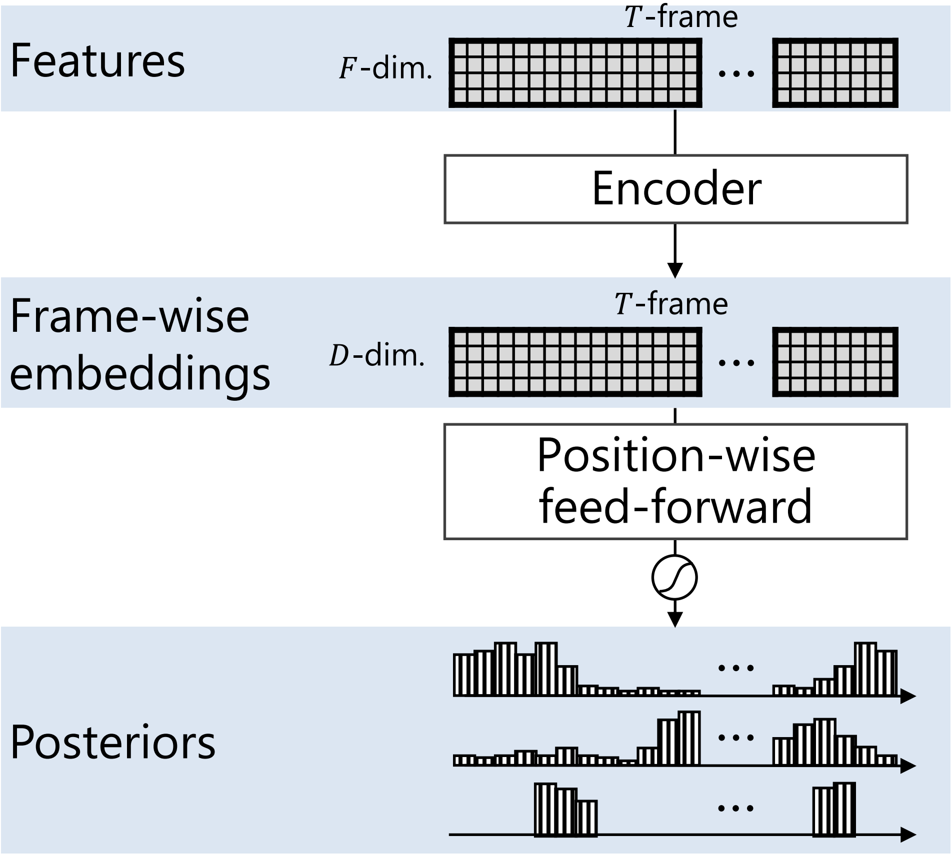
  - Variant of cascaded approach

- **Pros & Cons**

  ✓ Can benefit from spatial information

  ✘ Speakers from the same direction cannot be distinguished

End-to-End Approach [Fujita+'19]



End-to-End Neural Diarization (EEND)

- **Method**
  - Estimate multiple speakers' speech activities simultaneously from input acoustic features
- **Pros & Cons**
  - ✓ Simple pipeline (only a single neural network)
  - ✓ Can handle speaker overlap
  - ✗ The architecture fixes the number of speakers

The details will be introduced later

Comparison of Various Approaches

	Cascade-based approach	Direction-of-Arrival-based approach	End-to-end approach
Pipeline	✗ Complicated	✗ Complicated	✓ Simple
Speech overlap	✗ Cannot handle (without additional modules)	✓ Can handle (Based on methods)	✓ Can handle
Number of Speakers	✓ Flexible	✓ Flexible	✗ Fixed
Clue for diarization	Spectral information	Spatial information	Spectral information

- End-to-end approach is superior at many points, but has difficulty on the number of speakers
- End-to-end approach has room for improvement by leveraging spatial information
- If we ignore the complexity of the pipeline, it is hard to discard the cascaded approach as long as it can handle speech overlap

Purpose of This Thesis

1. To improve the utility of speaker diarization methods…

- *End-to-end approach is superior at many points, but have difficulty on the number of speakers*

→ Propose an end-to-end method that even works when **the number of speakers is unknown**

- *End-to-end approach has room for improvement by leveraging spatial information*

→ Propose an end-to-end method that accepts **multi-channel inputs**

- *It is hard to discard the cascaded approach as long as it can handle speech overlap*

→ Propose to use the end-to-end approach for **overlap handling of the cascaded approach**

2. To utilize speaker diarization for multi-talker ASR…

- *How speaker diarization contributes to multi-talker ASR is not well explored*

→ Propose a **diarization-driven meeting transcription system**

- *Speech separation conditioned on speaker diarization results is quite slow*

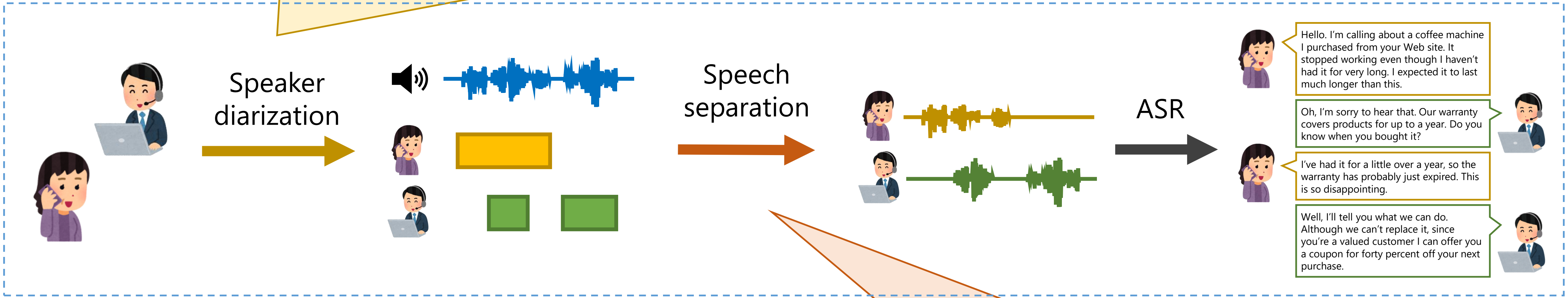
→ Propose a **block-online algorithm of diarization-conditioned speech separation**

Thesis Overview

Chapter 3
End-to-end speaker diarization for *unknown numbers of speakers*
[TASLP'22] [TASLP'23] [INTERSPEECH'20] [ASRU'21]

Chapter 4
Multi-channel end-to-end speaker diarization
[ICASSP'22] [SLT'22]

Chapter 5
End-to-end speaker diarization as *post-processing*
[ICASSP'21]



Hello. I'm calling about a coffee machine I purchased from your Web site. It stopped working even though I haven't had it for very long. I expected it to last much longer than this.

Oh, I'm sorry to hear that. Our warranty covers products for up to a year. Do you know when you bought it?

I've had it for a little over a year, so the warranty has probably just expired. This is so disappointing.

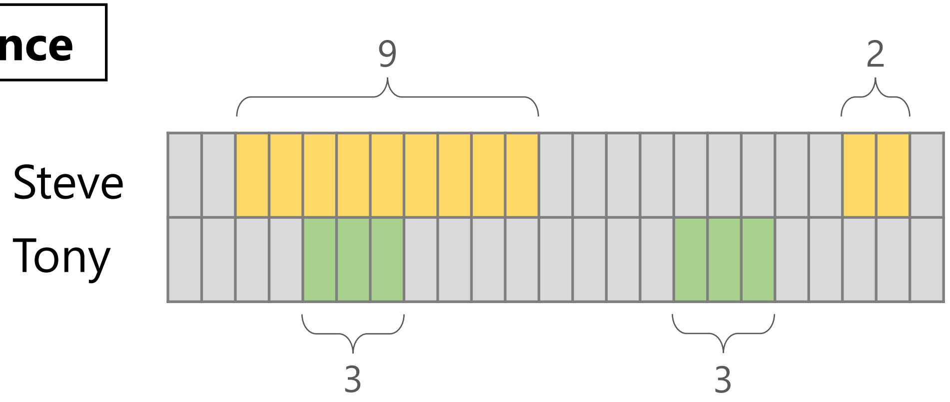
Well, I'll tell you what we can do. Although we can't replace it, since you're a valued customer I can offer you a coupon for forty percent off your next purchase.

Chapter 6
Speaker-diarization-driven meeting transcription
[INTERSPEECH'20]

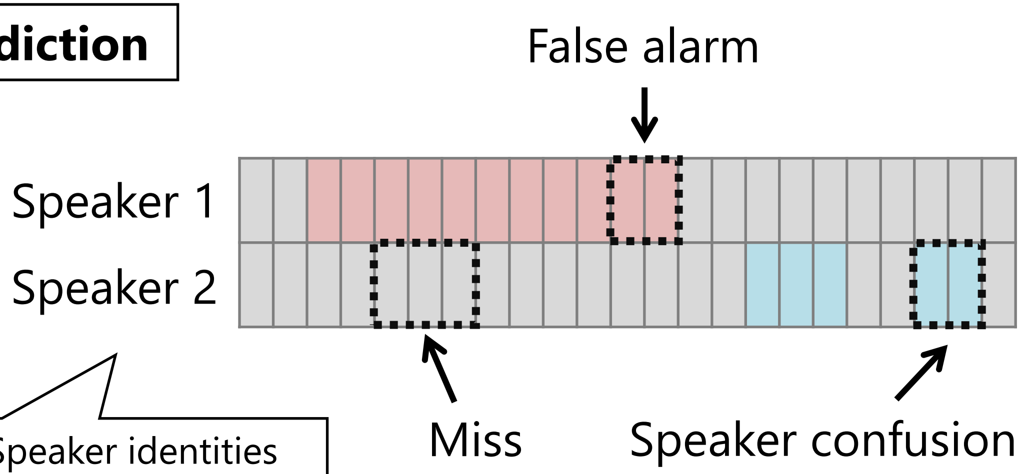
Chapter 7
Block online speech separation using speaker diarization results
[SLT'21]

Evaluation Metric: Diarization Error Rate (DER)

Reference



Prediction



Speaker identities are not considered in speaker diarization

- Definition

$$\mathrm{DER} = \frac{T_{\mathrm{MI}} + T_{\mathrm{FA}} + T_{\mathrm{CF}}}{T_{\mathrm{Speech}}} \quad \left(= \frac{3+2+2}{9+2+3+3} = 41.1\% \right)$$

T_{Speech} : Duration of speech (17=9+2+3+3)
T_{MI} : Duration of missed speech (3)
T_{FA} : Duration of false alarmed speech (2)
T_{CF} : Duration of speaker confusion (2)

- Common evaluation metric of speaker diarization
- The lower, the better
- Not upper-bounded by 100 %

Summary of Chapter 3

- **Problem**
  - The conventional EEND assumes that the number of speakers is known in advance

- **Solutions**
  - 3-1: End-to-end speaker diarization for **flexible** numbers of speakers
    - Core contribution: Encoder-decoder based attractors for EEND (EEND-EDA)
    - Related publications: [INTERSPEECH’20] [TASLP’22]

  - 3-2: End-to-end speaker diarization for **unlimited** numbers of speakers
    - Core contribution: Use of attractors from calculated from global and local contexts (EEND-GLA)
    - Related publication: [ASRU’21] [TASLP’23]

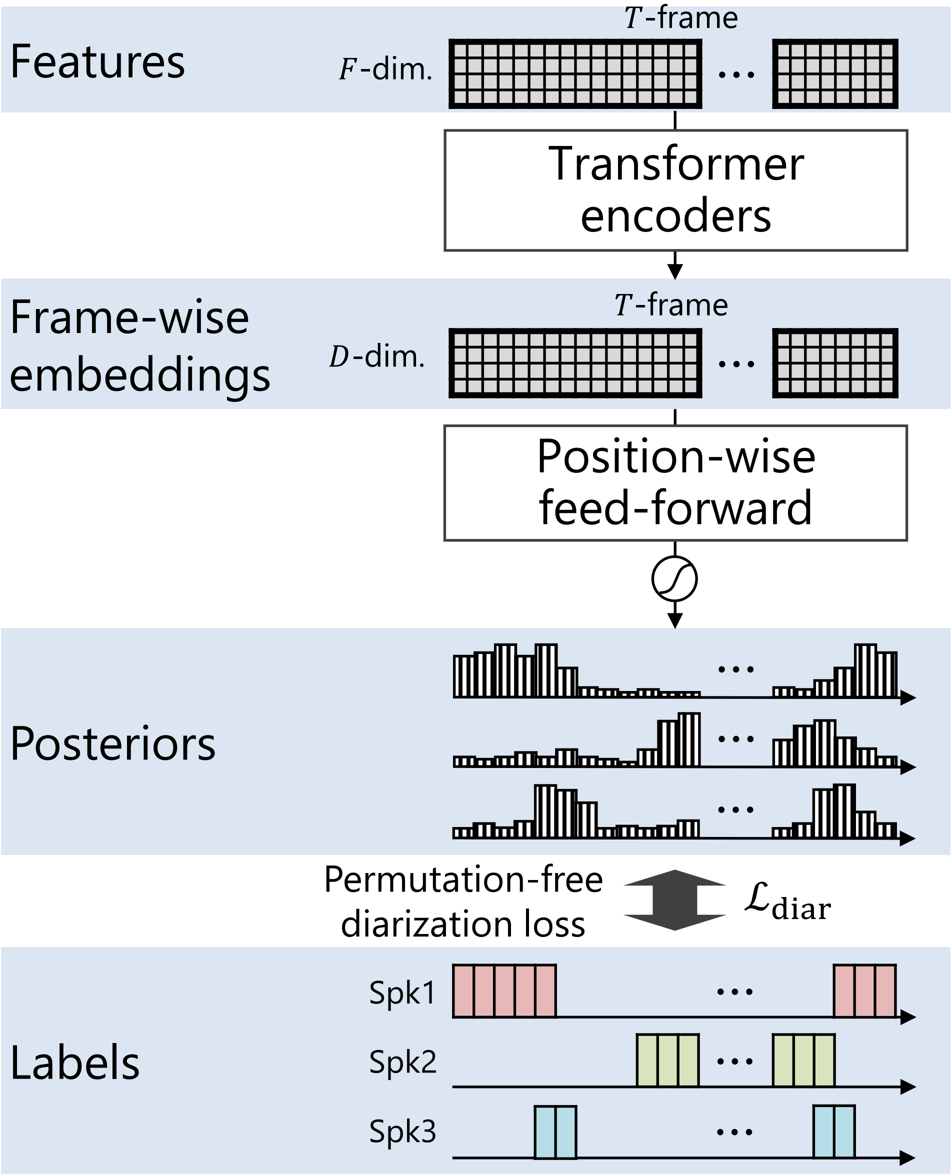
  - 3-3: **Online** end-to-end speaker diarization for unlimited numbers of speakers
    - Core contribution: An extension to speaker-tracing buffer to make it compatible with EEND-GLA
    - Related publication: [TASLP’23]

Summary of Chapter 3

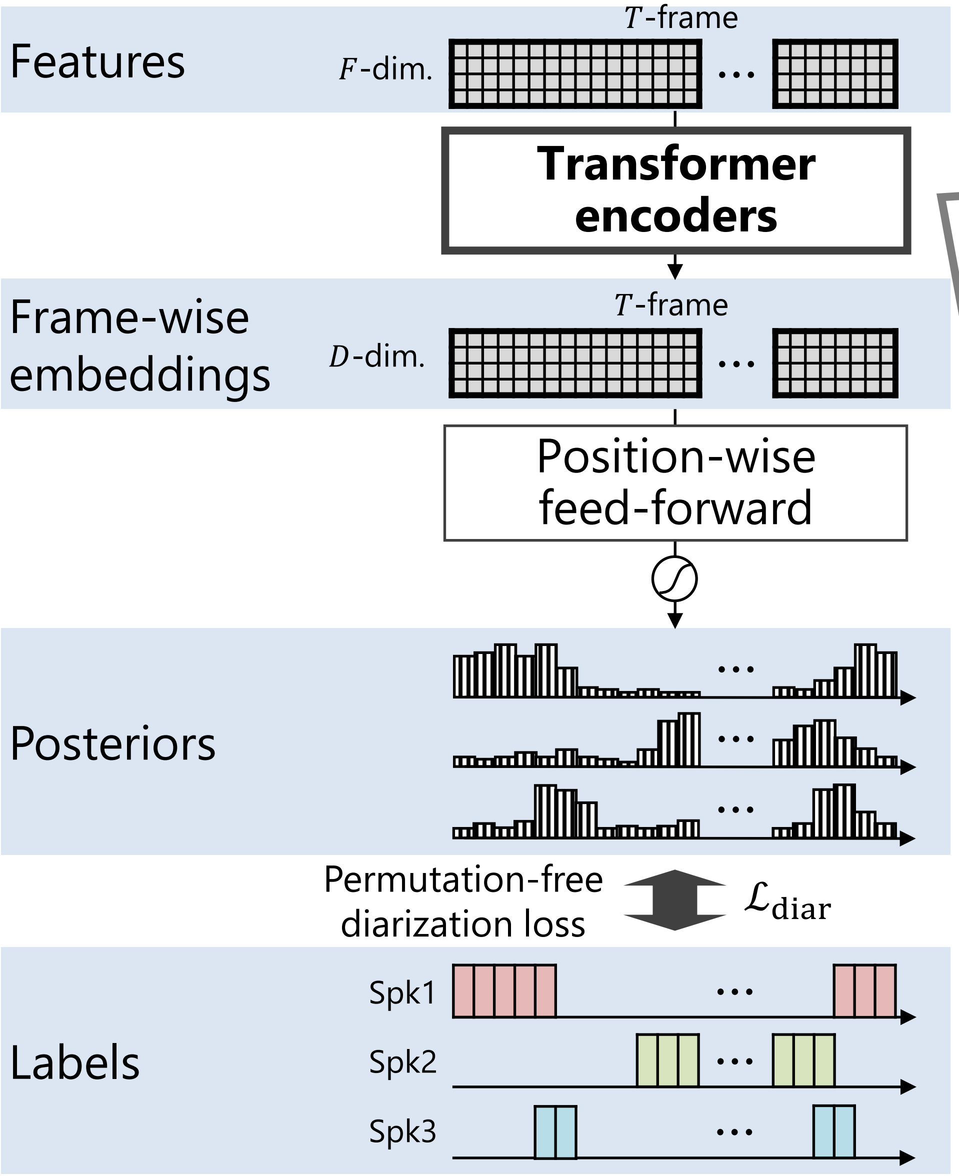
- **Problem**
  - The conventional EEND assumes that the number of speakers is known in advance

- **Solutions**
  - 3-1: End-to-end speaker diarization for **flexible** numbers of speakers
    - Core contribution: Encoder-decoder based attractors for EEND (EEND-EDA)
    - Related publications: [INTERSPEECH’20] [TASLP’22]
  - 3-2: End-to-end speaker diarization for **unlimited** numbers of speakers
    - Core contribution: Use of attractors from calculated from global and local contexts (EEND-GLA)
    - Related publication: [ASRU’21] [TASLP’23]
  - 3-3: **Online** end-to-end speaker diarization for unlimited numbers of speakers
    - Core contribution: An extension to speaker-tracing buffer to make it compatible with EEND-GLA
    - Related publication: [TASLP’23]

End-to-End Neural Diarization [Fujita+’19]



End-to-End Neural Diarization [Fujita+'19]



Features

T-frame

F-dim.

Transformer encoders

Frame-wise embeddings

T-frame

D-dim.

Position-wise feed-forward

Posteriors

Permutation-free diarization loss

$\mathcal{L}_{\text{diar}}$

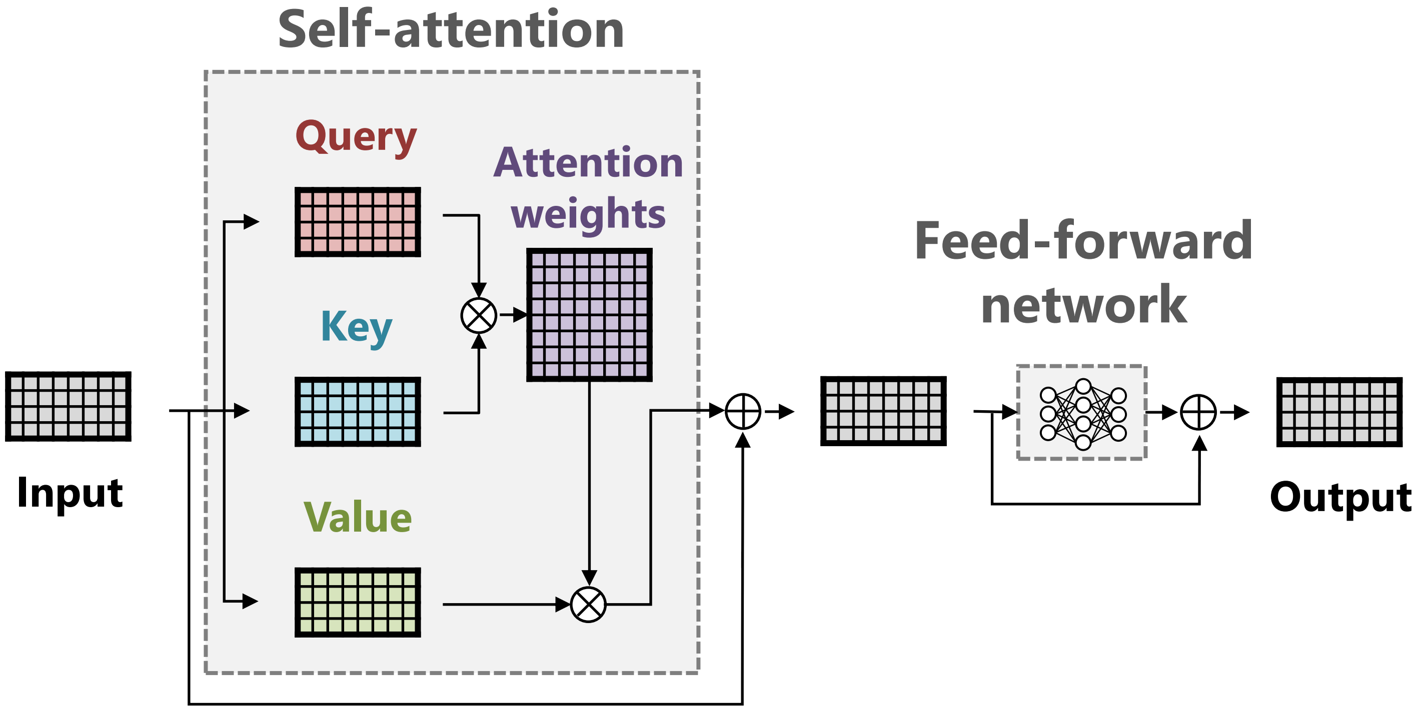
Labels

Spk1

Spk2

Spk3

Transformer encoder [Vaswani+'17]



Self-attention

Query

Key

Value

Attention weights

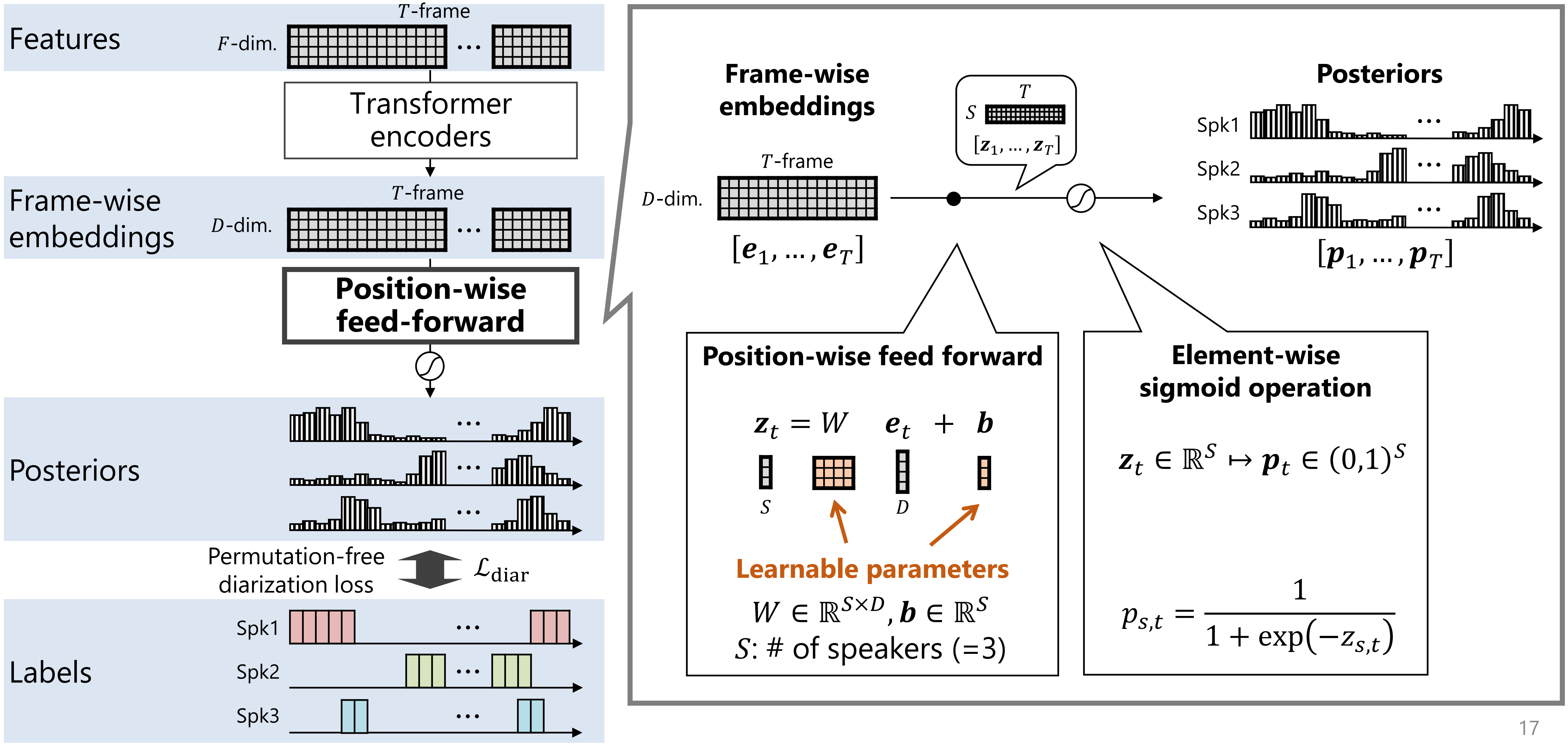
Input

Feed-forward network

Output

- Architecture to treat time-series data (Positional encodings are omitted in this study)
- High performance in various tasks (e.g., NLP, CV, audio)

End-to-End Neural Diarization [Fujita+'19]



Position-wise feed forward

$$\boldsymbol{z}_t = W \boldsymbol{e}_t + \boldsymbol{b}$$

Learnable parameters

$W \in \mathbb{R}^{S\times D}, \boldsymbol{b} \in \mathbb{R}^S$

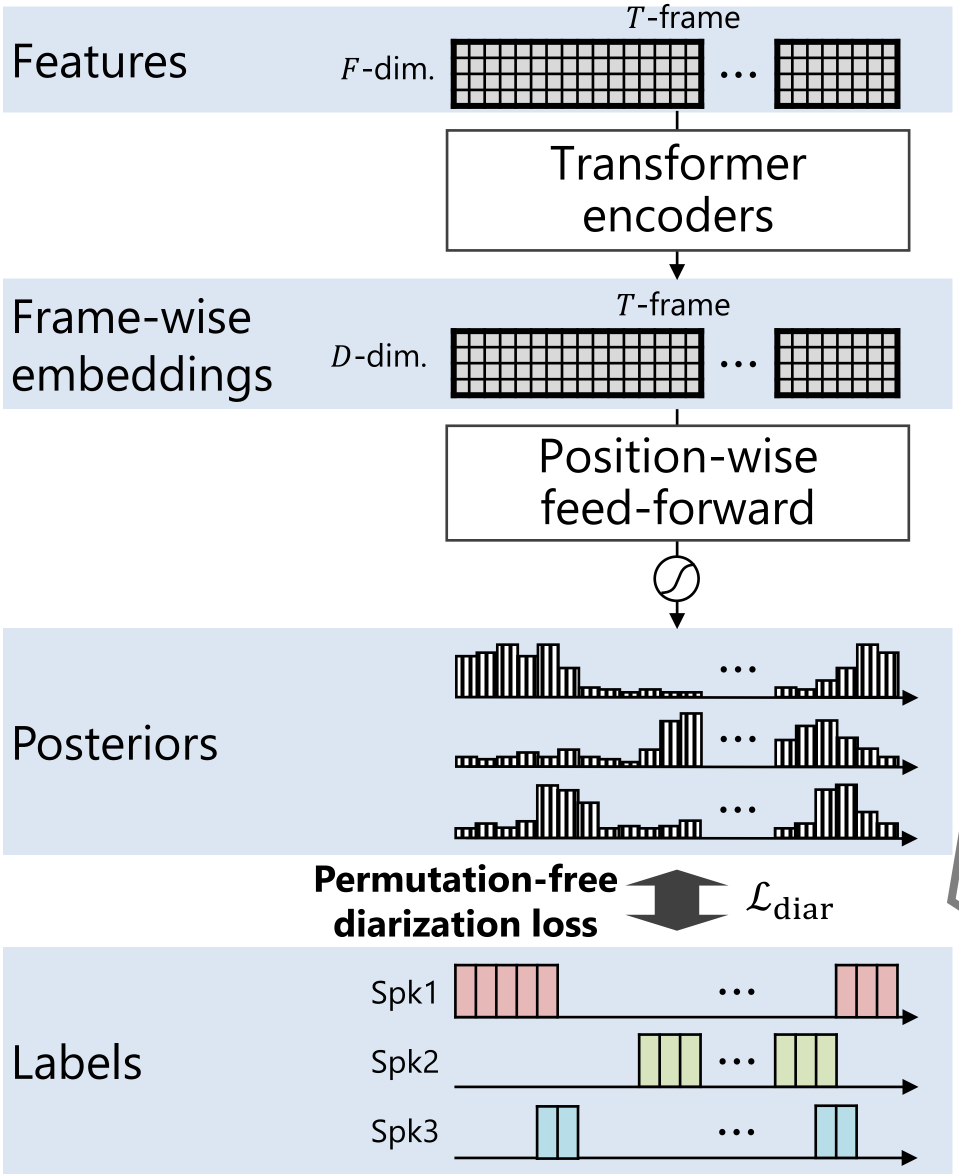
S: # of speakers (=3)

Element-wise sigmoid operation

$$\boldsymbol{z}_t \in \mathbb{R}^S \mapsto \boldsymbol{p}_t \in (0,1)^S$$

$$p_{s,t} = \frac{1}{1 + \exp\left(-z_{s,t}\right)}$$

End-to-End Neural Diarization [Fujita+'19]



Permutation-free diarization loss

$$\mathcal{L}_{\text{diar}} = \frac{1}{TS} \min_{P_\phi \in \Phi(S)} \sum_{t=1}^{T} \text{BCE}\left(\hat{\boldsymbol{y}}_t, P_\phi \boldsymbol{y}_t\right)$$

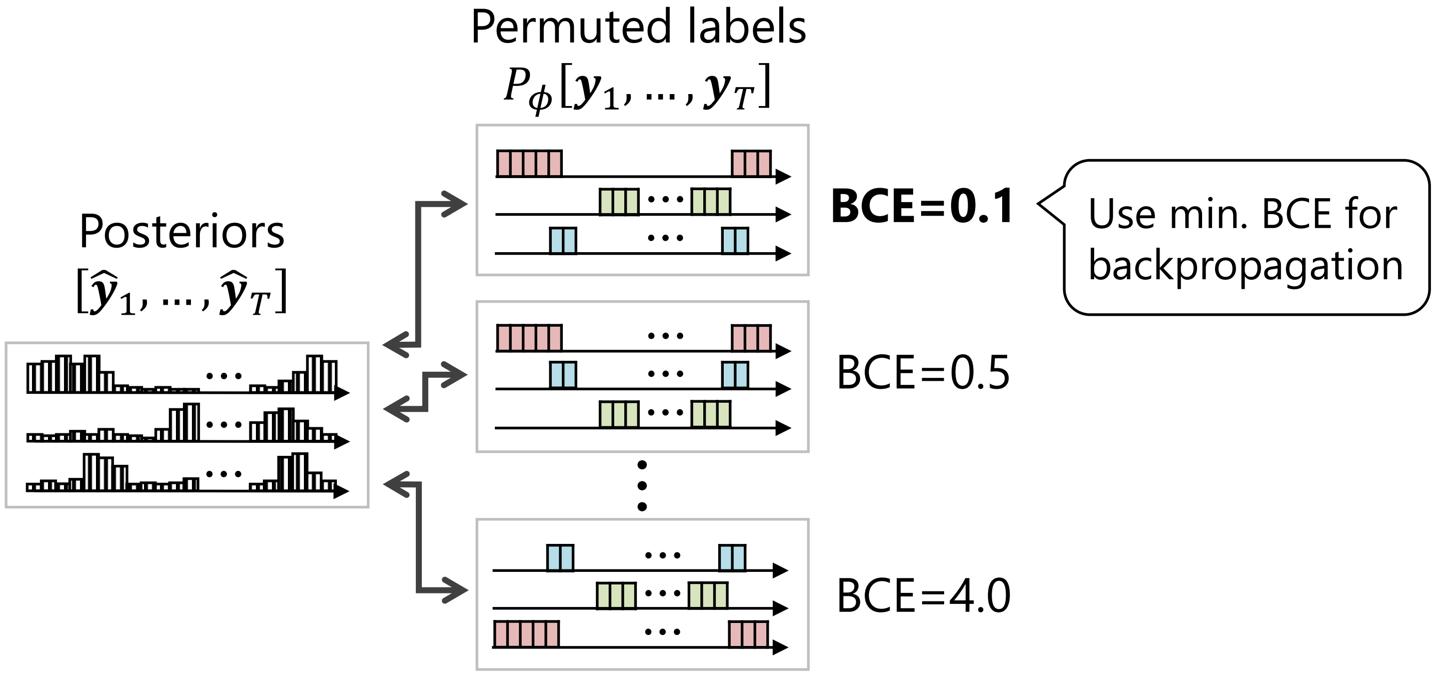
S: Number of speakers

T: Number of frames

$\Phi(S)$: Set of all the possible $S \times S$ permutation matrices

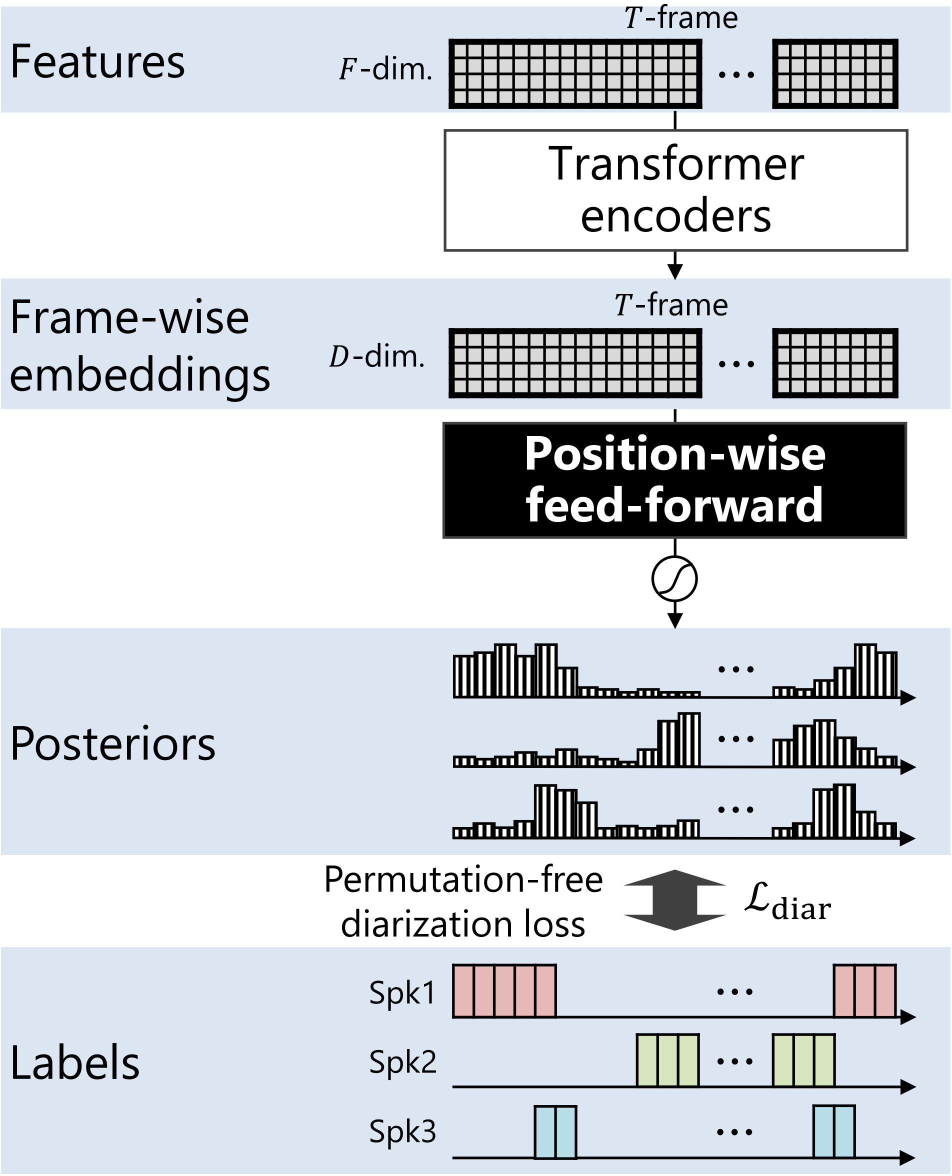
P_ϕ: Permutation matrix of ϕ

BCE: Binary cross entropy

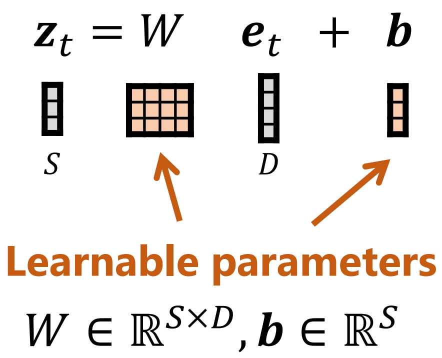


Use min. BCE for backpropagation

Problems of Conventional EEND

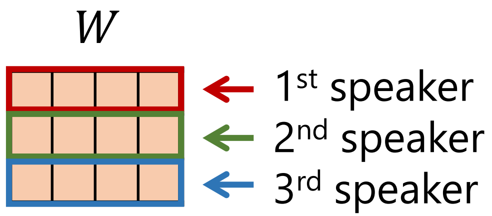


Problem 1: Fixed number of speakers



- ✘ The parameters W and $\boldsymbol{b}$ fix the number of speakers S
- ✘ Sufficiently large S eases the issue, but the performance degrades

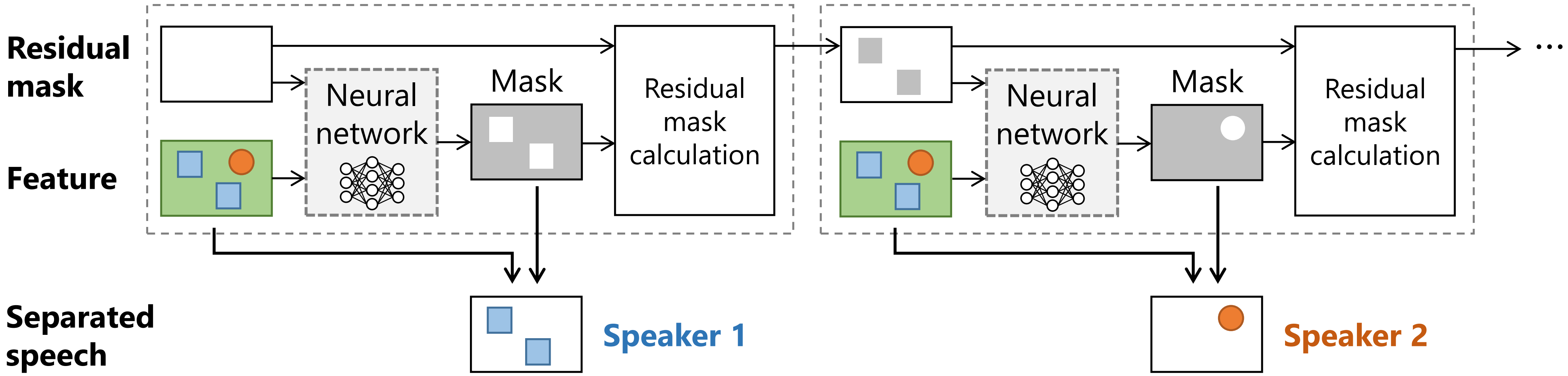
Problem 2: Not speaker adaptive



- ✘ Fixed parameters regardless of the speakers appeared in the recording

Related Work on *Speech Separation*

- **Recurrent selective attention network** [Kinoshita+'18]



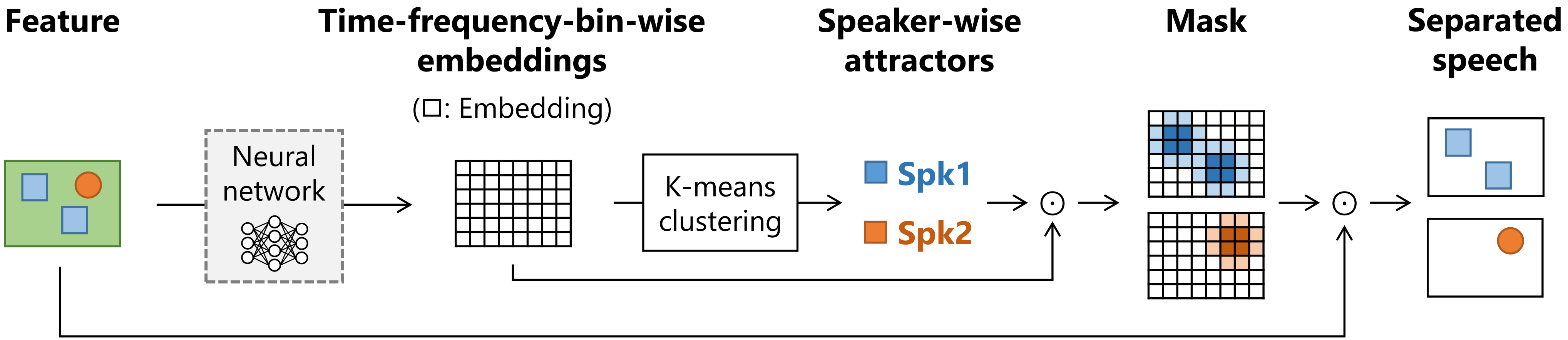
- Extract speakers one-by-one using residual masks

✓ Estimate the number of speakers simultaneously

✘ Residual mask cannot be determined for speaker diarization
  - Speech separation: 0 or 1 speaker at each time-frequency bin
  - Speaker diarization: No restriction of the number of speakers at each time frame

Related Work on *Speech Separation*

- **Deep attractor network** [Zhuo+'17]



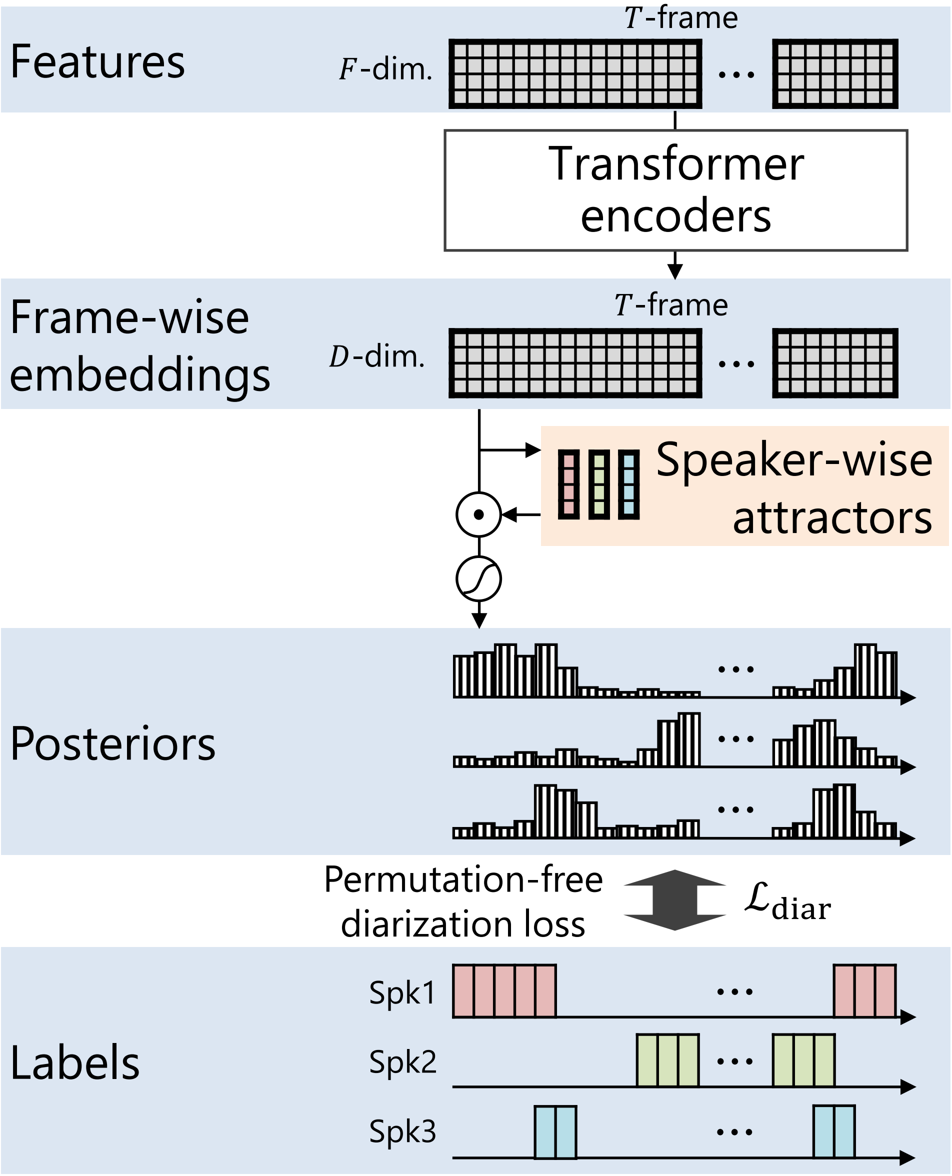
1. Calculate speaker-wise attractors (representative vectors) using K-means clustering of time-frequency-bin-wise embeddings
2. Estimate masks with the dot-products of the attractors and embeddings

✓ No need for residual masks

✓ Adaptive attractors for each speaker

✘ Need to set the number of speakers manually

Proposed Method: EEND-EDA

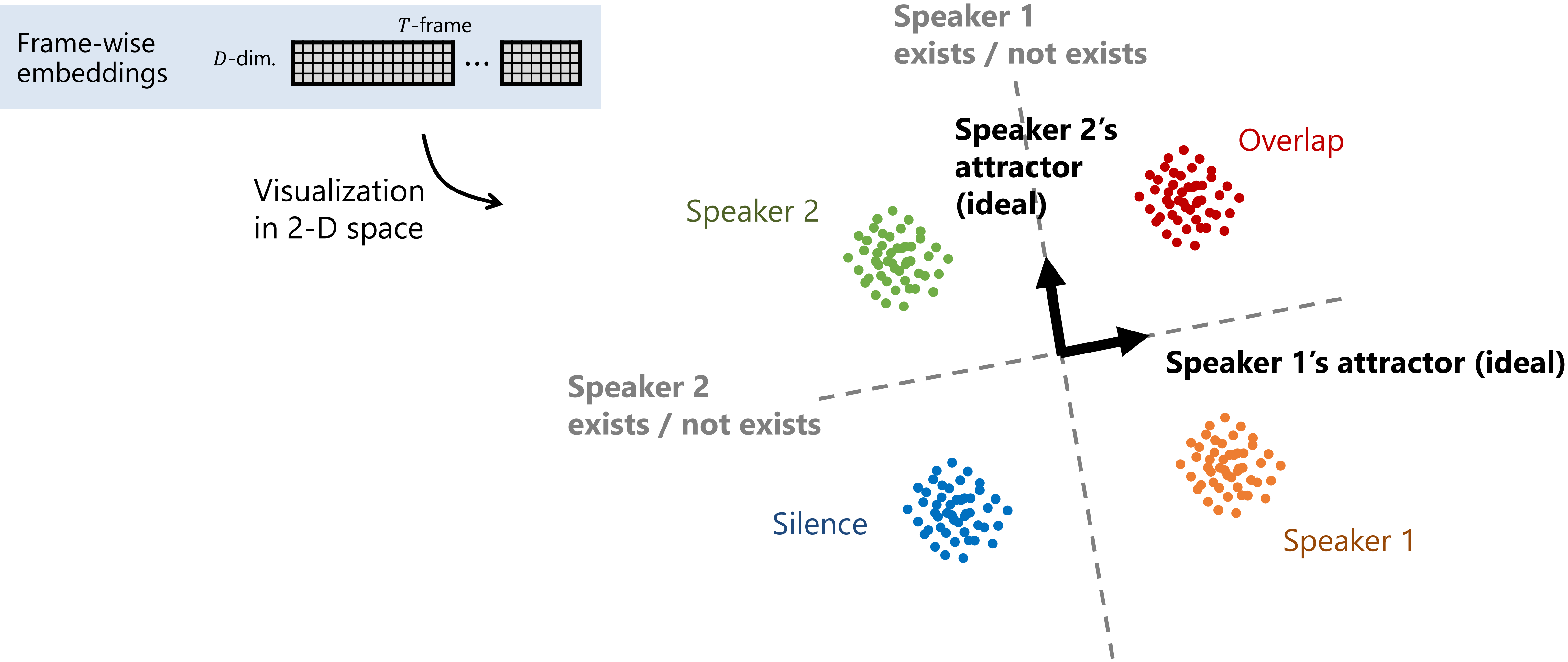


EEND with Encoder-Decoder Based Attractors (EEND-EDA)

Core idea:

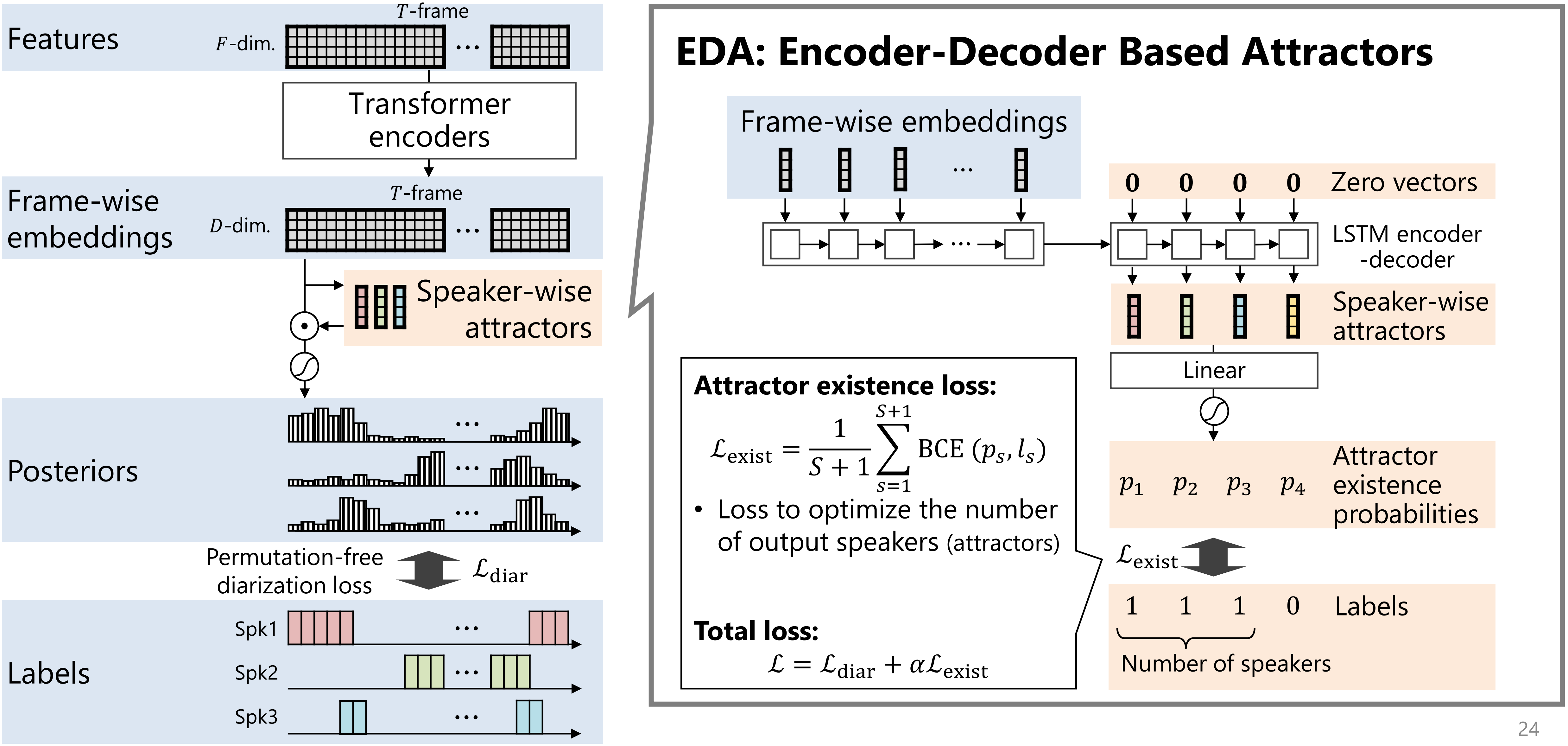
1. Calculate adaptive speaker-wise attractors in an autoregressive manner
2. Estimate the number of speakers simultaneously by evaluating the existence of each attractor

Adaptive Attractors (Ex. Two-Speaker Case)



- Attractors are never obtained via PCA / K-means clustering.

Proposed Method: EEND-EDA



Features

T-frame

F-dim.

Transformer encoders

Frame-wise embeddings

T-frame

D-dim.

Speaker-wise attractors

Posteriors

Permutation-free diarization loss

$\mathcal{L}_{\text{diar}}$

Labels

Spk1

Spk2

Spk3

EDA: Encoder-Decoder Based Attractors

Frame-wise embeddings

Zero vectors

0 **0** **0** **0**

LSTM encoder -decoder

Speaker-wise attractors

Linear

Attractor existence probabilities

p_1 p_2 p_3 p_4

$\mathcal{L}_{\text{exist}}$

Labels

1 1 1 0

Number of speakers

Attractor existence loss:

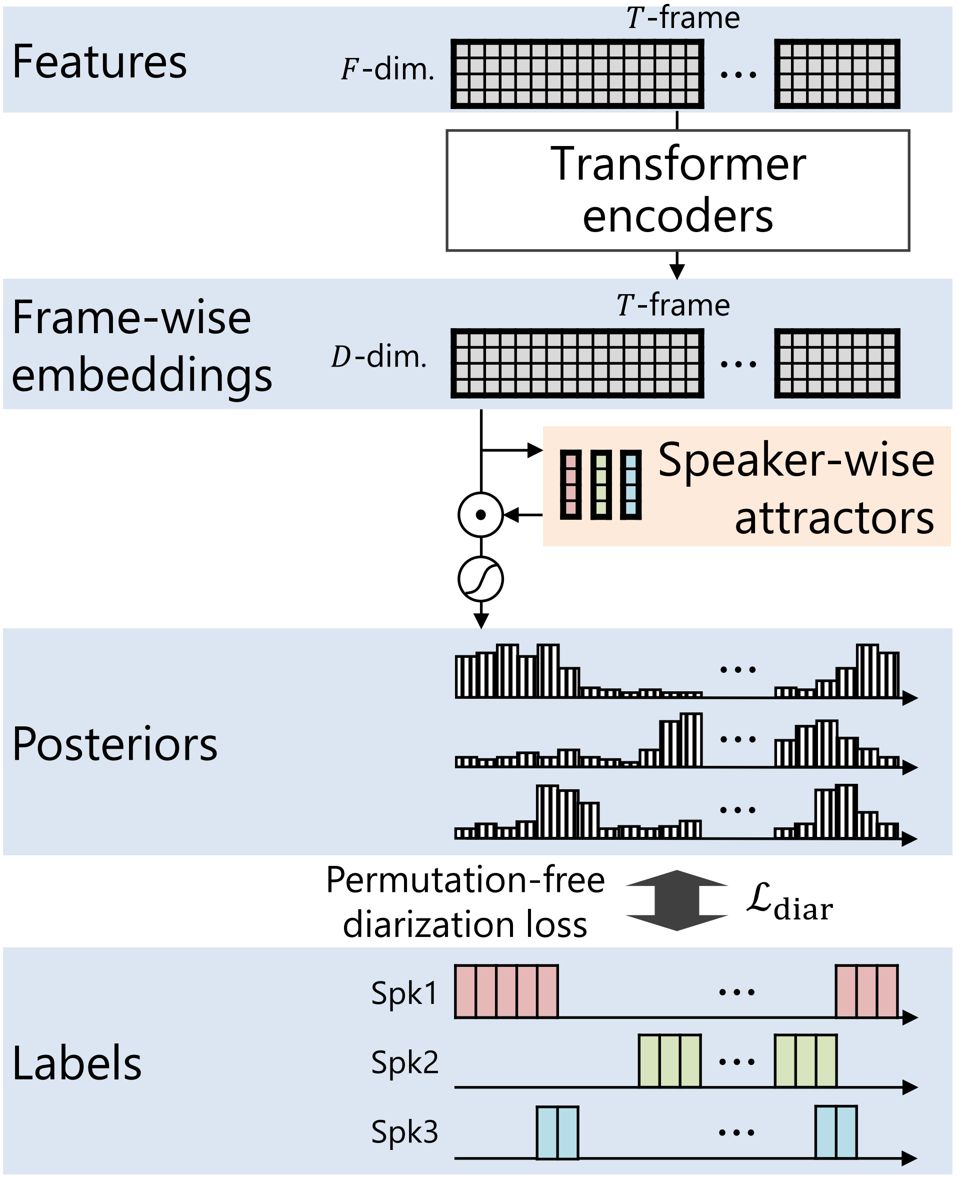
$$\mathcal{L}_{\text{exist}} = \frac{1}{S+1} \sum_{s=1}^{S+1} \mathrm{BCE}\left(p_s, l_s\right)$$

- Loss to optimize the number of output speakers (attractors)

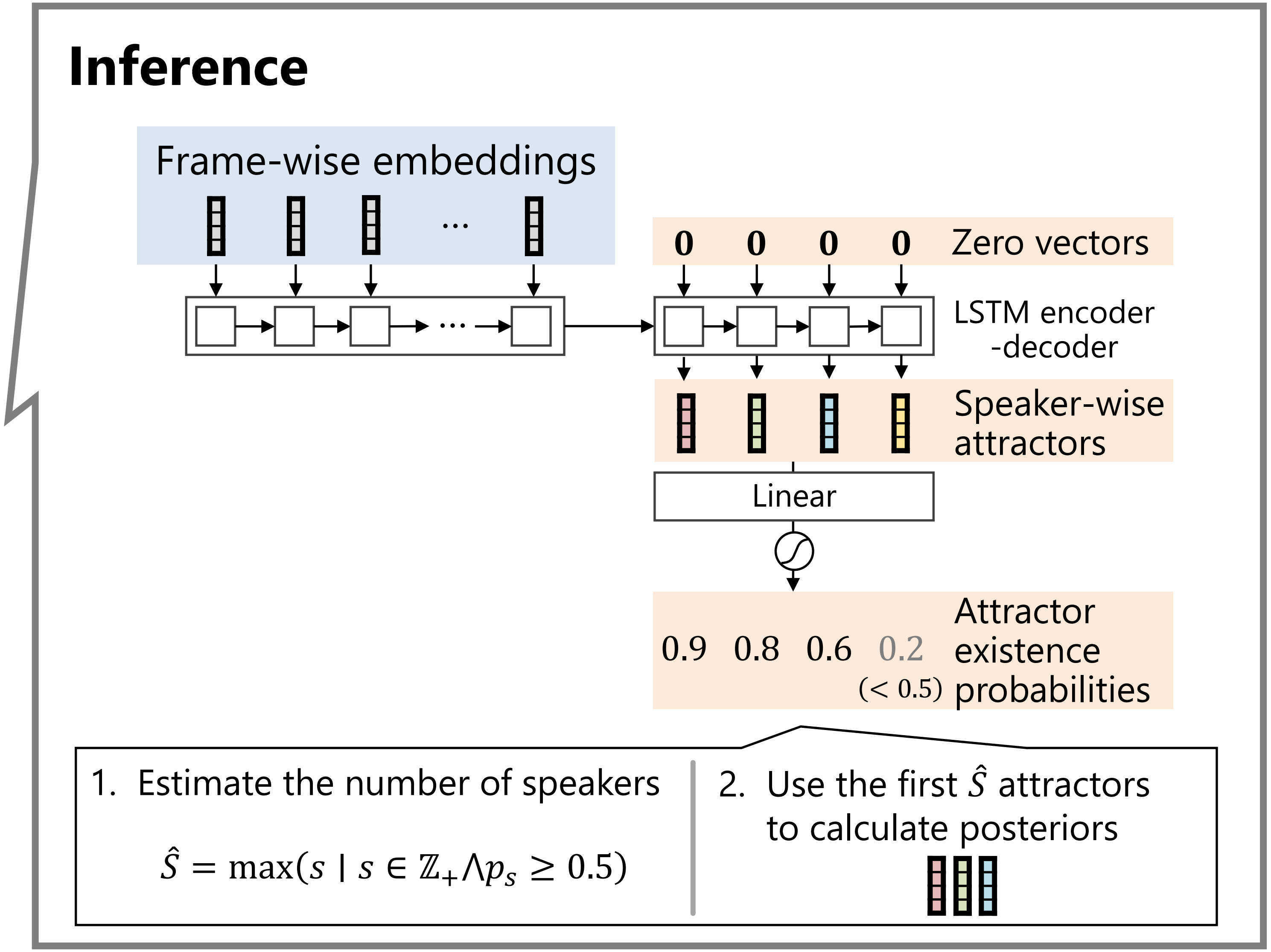
Total loss:

$$\mathcal{L} = \mathcal{L}_{\text{diar}} + \alpha \mathcal{L}_{\text{exist}}$$

Proposed Method: EEND-EDA



Inference



1. Estimate the number of speakers

$$\hat{S} = \max(s \mid s \in \mathbb{Z}_+ \land p_s \geq 0.5)$$

2. Use the first $\hat{S}$ attractors to calculate posteriors

Experimental Settings

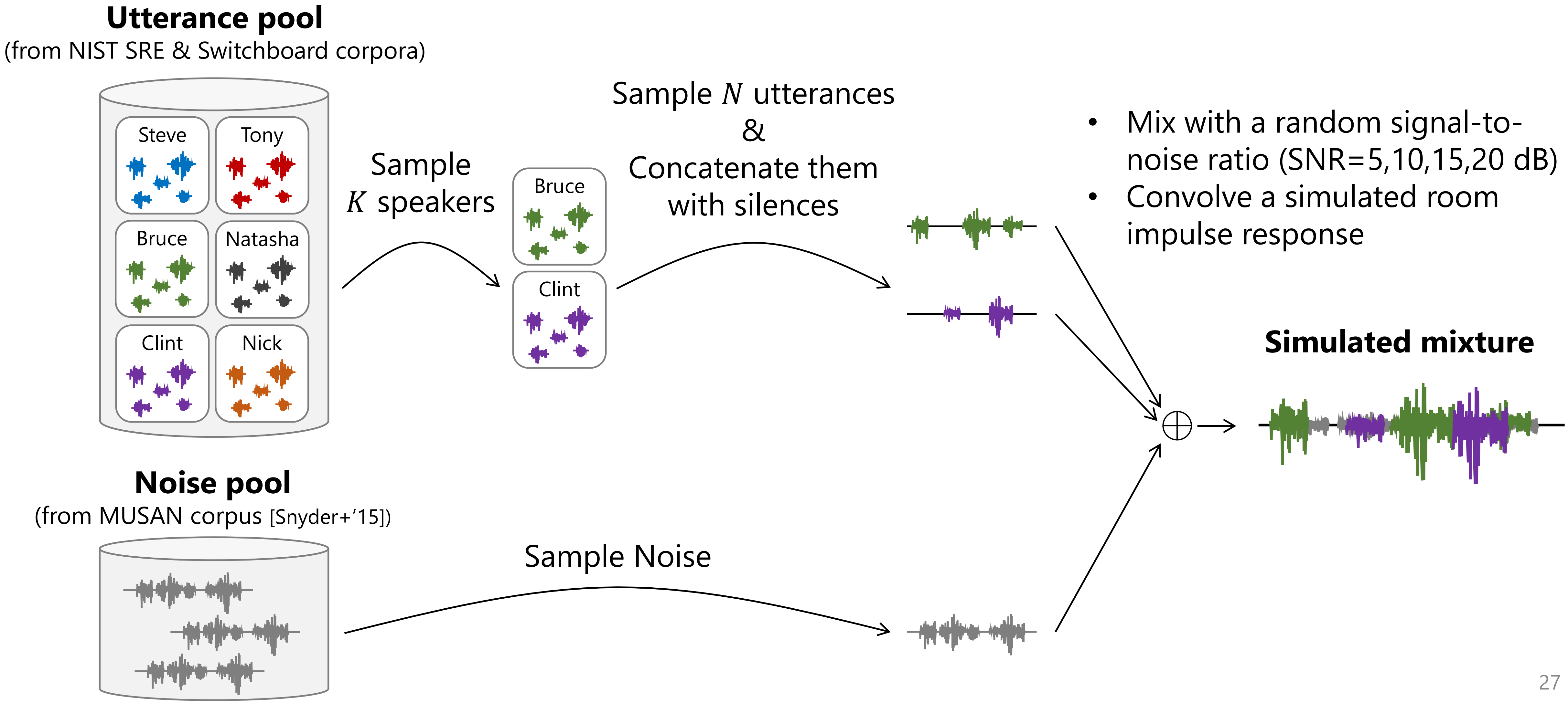
- Model configuration
  - 4-stacked Transformer encoders for the embedding extractor

- Experiments
  - Fixed-number-of-speaker experiments
    - Purpose: To see if the proposed speaker adaptive attractors improve the performance
    1. Train & evaluate a model using the **simulated 2 (or 3)-speaker dataset**
    2. Finetune & evaluate the model using the **real 2 (or 3)-speaker dataset**

  - Flexible-number-of-speaker experiments
    - Purpose: To see if the proposed method can treat flexible numbers of speakers
    1. Train a model using **simulated the 2-speaker dataset**
    2. Finetune & evaluate the model using the **simulated {1,2,3,4,5}-speaker datasets**
    3. Finetune & evaluate the model using the **real multi-speaker datasets**

Experimental Settings – Simulated Datasets

- Simulation protocol



Utterance pool
(from NIST SRE & Switchboard corpora)

Sample
K speakers

Sample N utterances
&
Concatenate them
with silences

- Mix with a random signal-to-noise ratio (SNR=5,10,15,20 dB)
- Convolve a simulated room impulse response

Simulated mixture

Noise pool
(from MUSAN corpus [Snyder+’15])

Sample Noise

Experimental Settings – Simulated Datasets

- Data sources
  - Utterance
    - Switchboard-2 (Phase I & II & III)
    - Switchboard Cellular (Part 1 & 2)
    - NIST SRE (2004, 2005, 2006, 2008)
  - Noise
    - MUSAN [Snyder+'15]

	Dataset	#Spk	#Mixtures	Overlap ratio (%)
Train	Sim1spk	1	100,000	0.0
	Sim2spk	2	100,000	34.1
	Sim3spk	3	100,000	34.2
	Sim4spk	4	100,000	31.5
	Sim5spk	5	100,000	30.3
Test	Sim1spk	1	500	0.0
	Sim2spk	2	500	34.4 / 27.3 / 19.1
	Sim3spk	3	500	34.7 / 27.4 / 19.2
	Sim4spk	4	500	32.0
	Sim5spk	5	500	30.7

- Overlap ratios are controlled by changing the silence length between utterances

- The set of speakers in the train/test sets are not overlapped (Open-set setting)

Experimental Settings – Real Datasets

- CALLHOME
  - Telephone conversation (mostly in English but not limited to)
  - CALLHOME-*k*spk is a *k*-speaker portion of this dataset
- CSJ
  - Face-to-face conversation in Japanese
- DIHARD II & III
  - Various domains, various languages
    - Audiobook, broadcast, clinical, court, meeting, restaurant, web video, etc.

	Dataset	**#Spk**	**#Mixtures**	**Overlap ratio (%)**
Adapt	CALLHOME-2spk	2	155	14.0
	CALLHOME-3spk	3	61	19.6
	CALLHOME	2-7	249	17.0
	DIHARD II dev [Ryant+'19]	1-10	192	9.8
	DIHARD III dev [Ryant+'21]	1-10	254	10.7
Test	CALLHOME-2spk	2	148	13.1
	CSJ [Maekawa'03]	2	54	20.1
	CALLHOME-3spk	3	74	17.0
	CALLHOME	2-6	250	16.7
	DIHARD II eval [Ryant+'19]	1-9	194	8.9
	DIHARD III eval [Ryant+'21]	1-9	259	9.2

Results of Two-Speaker Experiments

- **Diarization error rates (DERs) (%) on two-speaker mixtures**

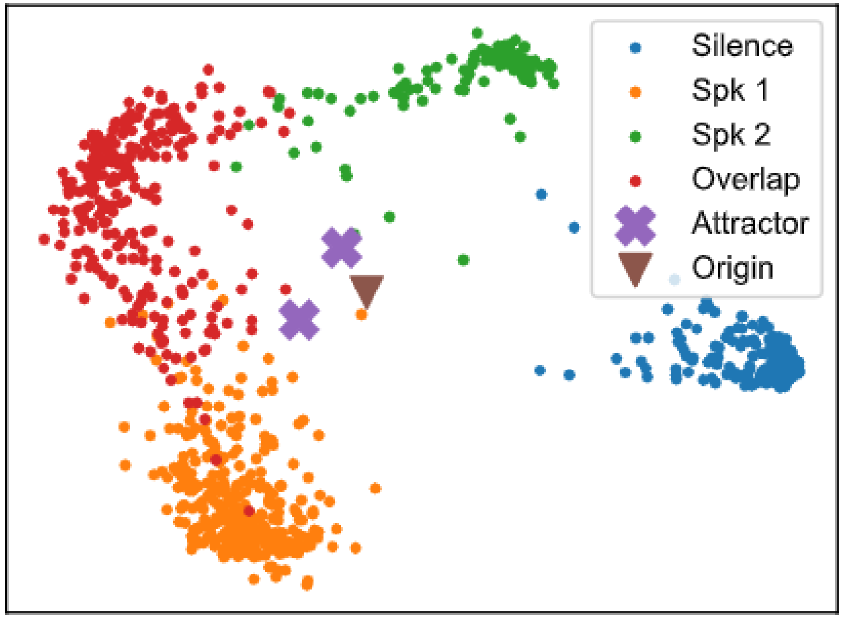
ρ: overlap ratio

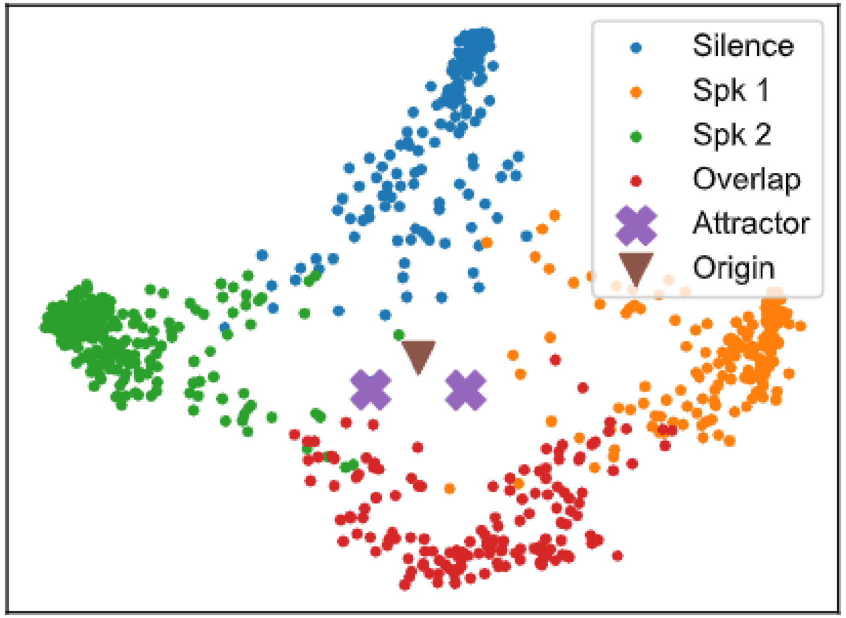
		Simulated			Real	
	Method	**Sim2spk** ρ=34.4%	**Sim2spk** ρ=27.3%	**Sim2spk** ρ=19.1%	**CALLHOME-2spk** ρ=13.1%	**CSJ** ρ=20.1%
Cascaded	i-vector clustering	33.74	30.93	25.96	12.10	27.99
	x-vector clustering	28.77	24.46	19.78	11.53	22.96
End-to-end	EEND [Fujita+'19]	4.56	4.50	3.85	9.54	20.48
	EEND-EDA	**2.69**	**2.44**	**2.60**	**8.07**	**16.27**

- EEND-based methods outperformed cascaded approach methods
- EDA improved the performance of EEND even when the number of speakers is fixed to two

Visualization

- Frame-wise embeddings and speaker-wise attractors visualized using PCA





- Embeddings of **Silence**, **Speaker 1**, and **Speaker 2** are well separated
- Embeddings of **Overlap** are distributed between those of **Speaker 1** and **Speaker 2**
- **Attractors** are successfully calculated for each of two speakers

Results of Three-Speaker Experiments

- **Diarization error rates (DERs) (%) on three-speaker mixtures**

ρ: overlap ratio

		Simulated			Real
	Method	Sim3spk ρ=34.4%	Sim3spk ρ=27.4%	Sim3spk ρ=19.2%	CALLHOME-3spk ρ=17.0%
Cascaded	x-vector clustering	31.78	26.06	19.55	19.01
End-to-end	EEND [Fujita+'19]	8.69	7.64	6.92	14.00
	EEND-EDA	**8.38**	**7.06**	**6.21**	**13.92**

- Similar to the results on two-speaker mixtures, EEND-EDA outperformed cascaded and conventional end-to-end approaches

Results of Flexible-Number-of-Speaker Experiments

- **DERs (%) on the simulated datasets**
  - EEND was trained to output null speech activities for absent speakers
  - EEND-EDA outperformed conventional EEND in every conditions

Method	**Simulated dataset** **Sim1spk** ρ=0.0%	 **Sim2spk** ρ=34.4%	 **Sim3spk** ρ=34.7%	 **Sim4spk** ρ=32.0%	 **Sim5spk** ρ=30.7%
EEND [Fujita+'19]	0.50	3.95	9.18	12.24	17.42
EEND-EDA	**0.36**	**3.65**	**7.70**	**9.97**	**11.95**

- **DERs (%) on CALLHOME** (with oracle speech activity detection)
  - EEND-EDA outperformed x-vector clustering and conventional EEND
  - EEND-EDA is better when #Speakers≤4, while x-vector clustering is better when #Speakers>4

Method	**#Speakers** **2**	 **3**	 **4**	 **5**	 **6**	**All**
X-vector clustering	9.44	13.89	16.05	**13.87**	**24.73**	13.28
EEND [Fujita+'19]	6.51	15.07	26.09	36.47	46.93	16.79
EEND-EDA	**5.85**	**9.97**	**12.61**	24.04	26.06	**10.46**

Results of Flexible-Number-of-Speaker Experiments

- **DERs (%) on DIHARD II and DIHARD III** (with oracle speech activity detection)

	Datasets	
Method	**DIHARD II** ρ=8.9%	**DIHARD III** ρ=9.2%
X-vector (TDNN) clustering [Landini+'20] [Horiguchi+'21]	**18.21**	15.83
EEND [Fujita+'19]	23.25	16.19
EEND-EDA	20.54	**14.91**

- **Breakdown of the DERs on DIHARD III**

	#Speakers								
Method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
X-vector (TDNN) clustering	**1.30**	11.43	16.76	**23.09**	**44.99**	**26.43**	**25.61**	**35.57**	**2.03**
EEND-EDA	2.80	**7.52**	**15.79**	25.63	47.66	31.73	35.47	38.19	18.73

  - Limitation: EEND-EDA performed worse when the number of speaker was large

Summary of Chapter 3

- **Problem**
  - The conventional EEND assumes that the number of speakers is known in advance

- **Solutions**
  - 3-1: End-to-end speaker diarization for **flexible** numbers of speakers
    - Core contribution: Encoder-decoder based attractors for EEND (EEND-EDA)
    - Related publications: [INTERSPEECH'20] [TASLP'22]

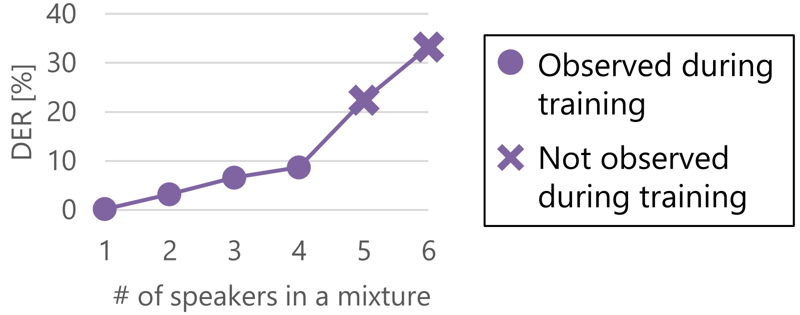
  - 3-2: End-to-end speaker diarization for **unlimited** numbers of speakers
    - Core contribution: Use of attractors from calculated from global and local contexts (EEND-GLA)
    - Related publication: [ASRU'21] [TASLP'23]

  - 3-3: **Online** end-to-end speaker diarization for unlimited numbers of speakers
    - Core contribution: An extension to speaker-tracing buffer to make it compatible with EEND-GLA
    - Related publication: [TASLP'23]

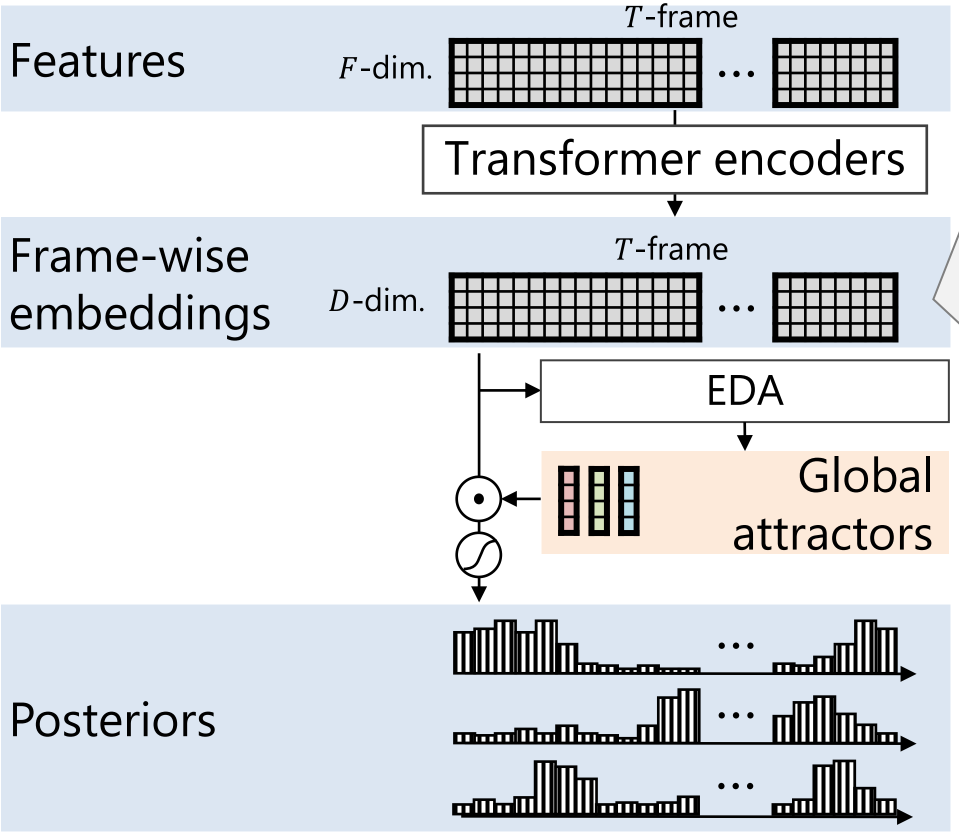
Limitation of EEND-EDA

- **Limitation**
  - The maximum number of speakers to be output from EEND-EDA is empirically limited by the training dataset

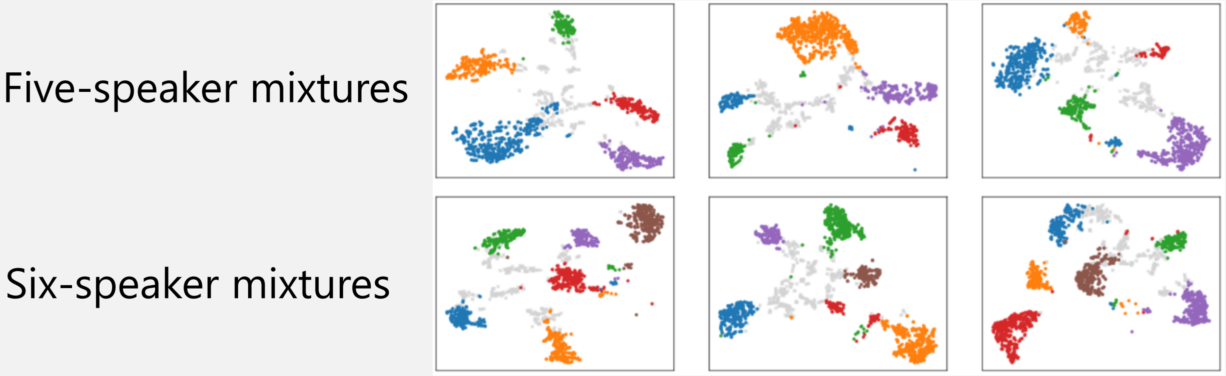
ex) When EEND-EDA was trained using {1,2,3,4}-speaker mixtures



- **Which part of EEND-EDA causes this limitation?**



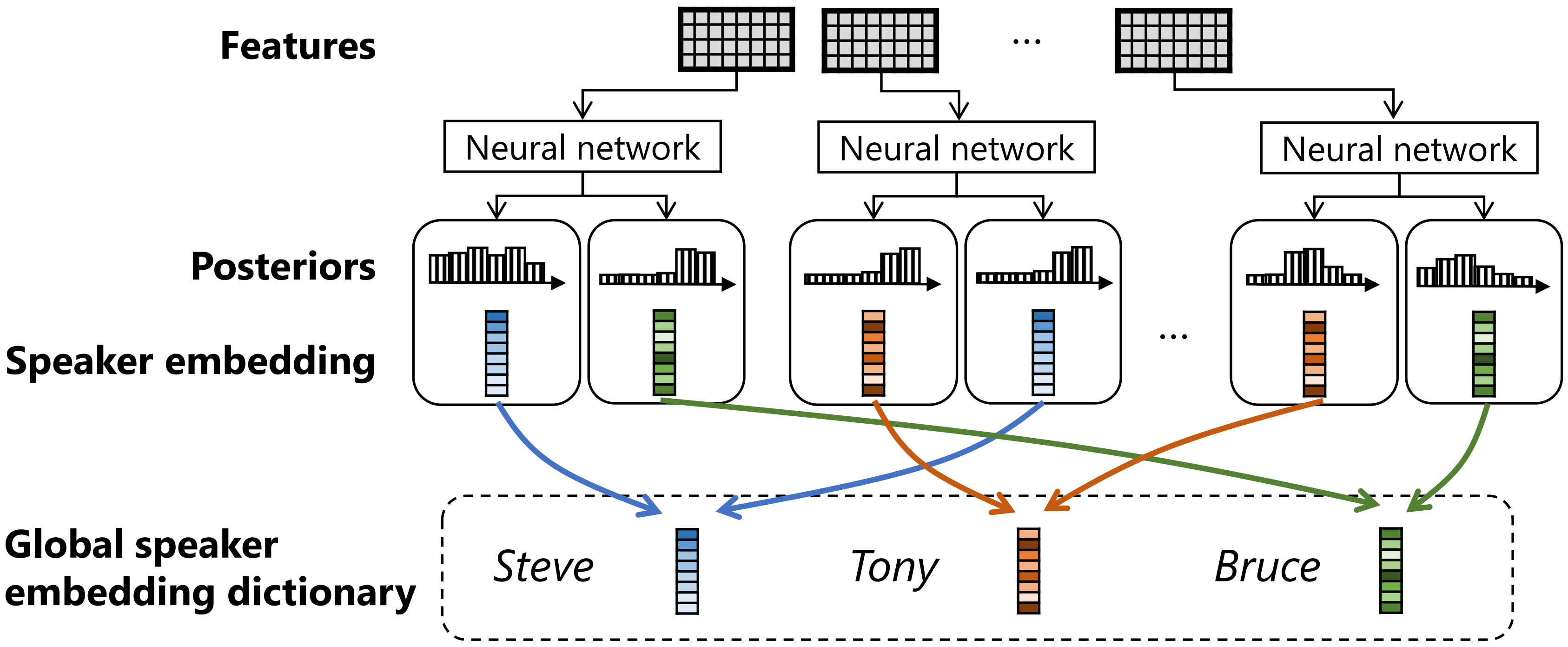
Visualization of the embeddings using t-SNE



Speakers are well-separated in the embedding space even when the input contains more than four speakers

→ <u>EDA limits the number of output speakers</u>

Related Work: EEND-vector Clustering [Kinoshita+'21]



1. Estimate diarization results as well as speaker embeddings from each short block-wise features
2. Clustering the speaker embeddings to solve inter-block speaker permutation

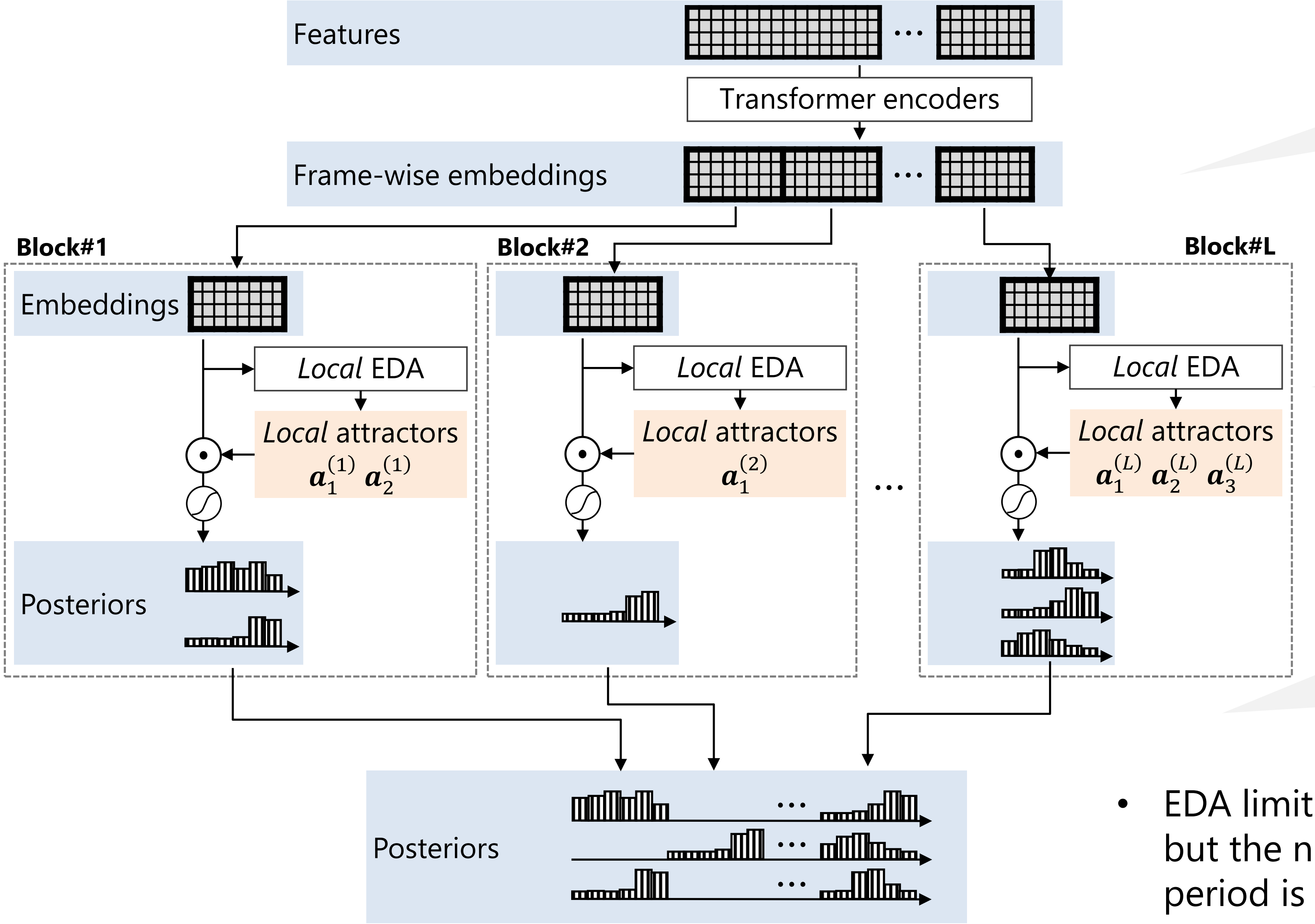
✓ The number of speakers is no longer limited

✗ Not speaker-adaptive posterior estimation

✗ Require speaker identities across recordings to construct the global speaker embedding dictionary

✗ Require somewhat long block (e.g., 30 sec) to obtain reliable speaker embeddings

EEND with Local Attractors – Basic Idea



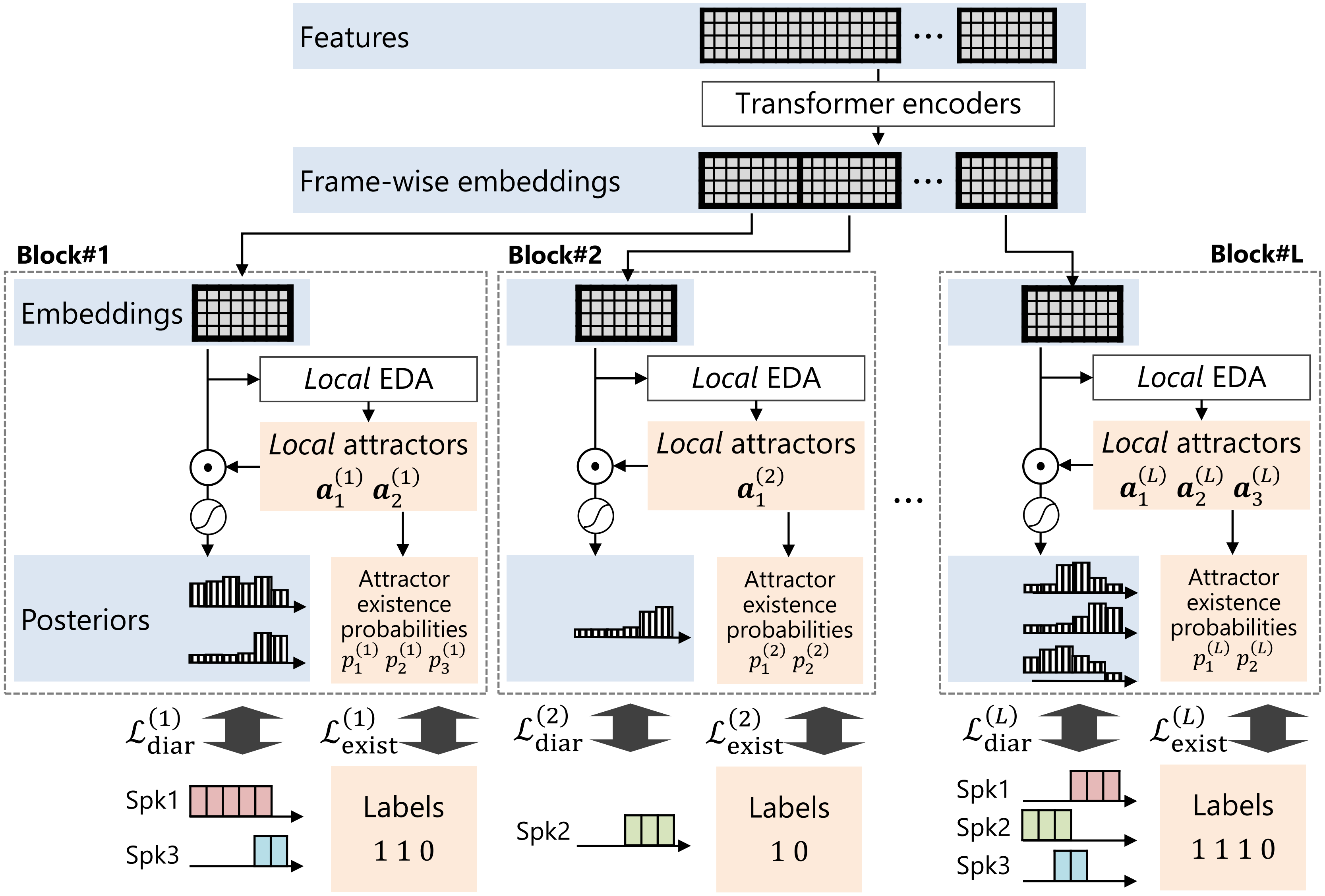
1 Split the embeddings into short blocks (5 sec in this study)

2 Perform diarization using EDA for each block

3 Find the optimal inter-block speaker correspondence based on the similarity of local attractors

- EDA limits the number of speakers, but the number of speakers in a short period is small so it is no more a problem

EEND with Local Attractors – Training



Total loss:

$$\mathcal{L}_{\text{local}} = \underline{\frac{1}{L}\sum_{l=1}^{L}\left(\mathcal{L}_{\text{diar}}^{(l)} + \alpha\mathcal{L}_{\text{exist}}^{(l)}\right)} + \gamma\mathcal{L}_{\text{pair}}$$

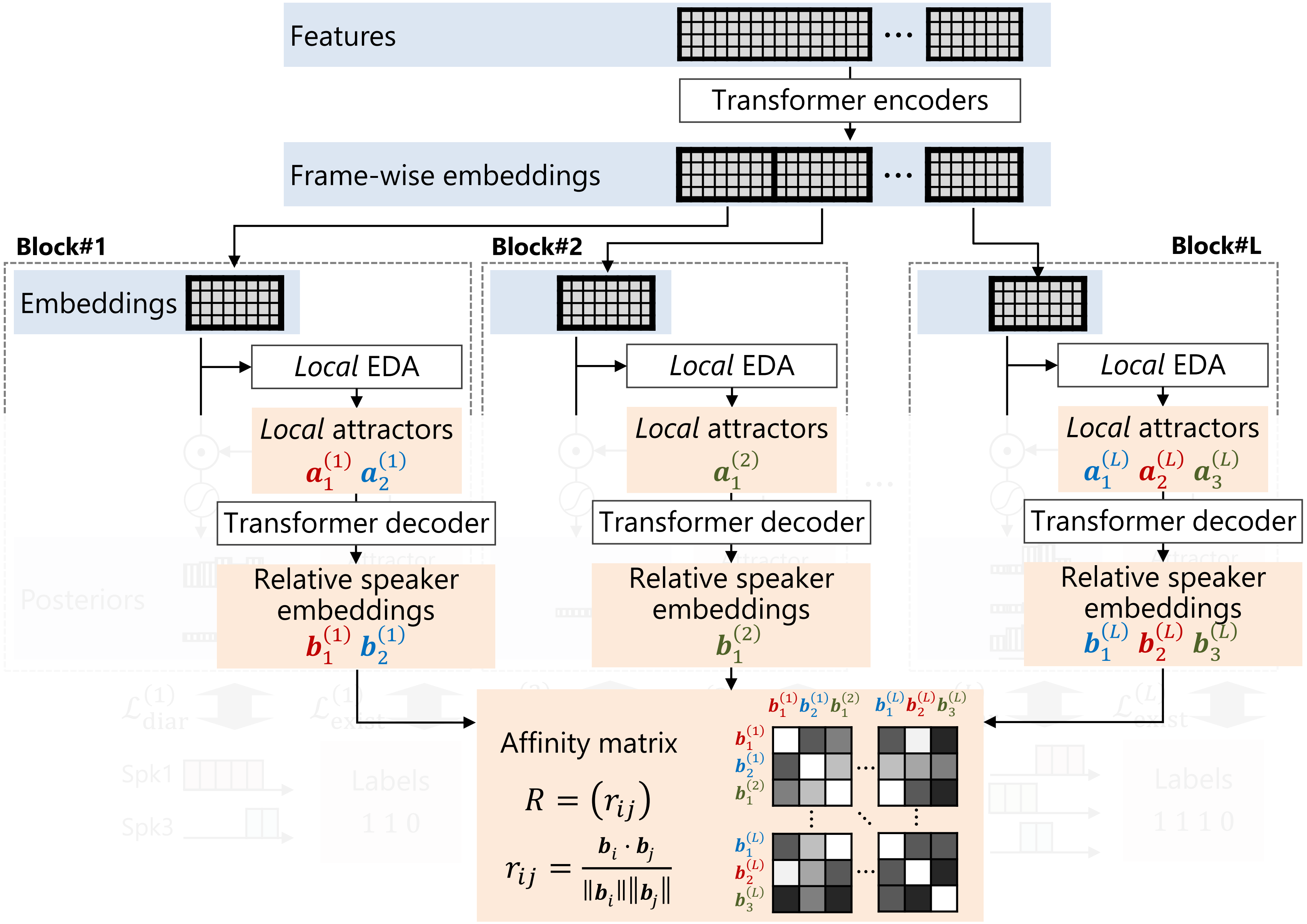
1st term:

Average of block-wise diarization loss and attractor existence loss

L: Number of blocks

- By calculating $\mathcal{L}_{\text{diar}}^{(l)}$, the optimal correspondence between local attractors and speakers are obtained

EEND with Local Attractors – Training



Features

Transformer encoders

Frame-wise embeddings

Block#1

Embeddings

Local EDA

Local attractors $a_1^{(1)}$ $a_2^{(1)}$

Transformer decoder

Relative speaker embeddings $b_1^{(1)}$ $b_2^{(1)}$

Block#2

Local EDA

Local attractors $a_1^{(2)}$

Transformer decoder

Relative speaker embeddings $b_1^{(2)}$

Block#L

Local EDA

Local attractors $a_1^{(L)}$ $a_2^{(L)}$ $a_3^{(L)}$

Transformer decoder

Relative speaker embeddings $b_1^{(L)}$ $b_2^{(L)}$ $b_3^{(L)}$

Affinity matrix

$$R = (r_{ij})$$

$$r_{ij} = \frac{\boldsymbol{b}_i \cdot \boldsymbol{b}_j}{\|\boldsymbol{b}_i\|\|\boldsymbol{b}_j\|}$$

Total loss:

$$\mathcal{L}_{\text{local}} = \frac{1}{L}\sum_{l=1}^{L}\left(\mathcal{L}_{\text{diar}}^{(l)} + \alpha\mathcal{L}_{\text{exist}}^{(l)}\right) + \underline{\gamma\mathcal{L}_{\text{pair}}}$$

2nd term:

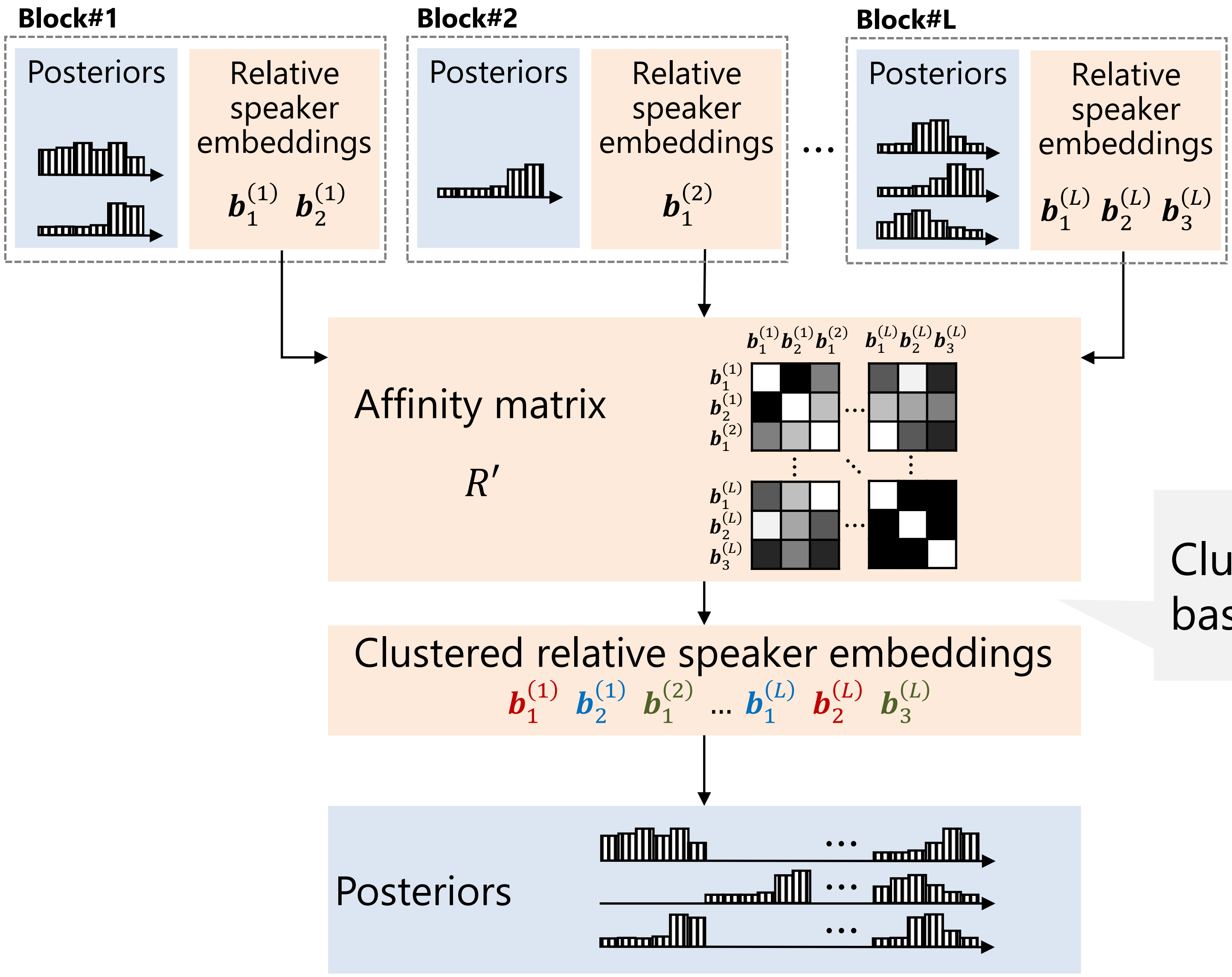
Pairwise loss to make the angle between relative speaker embeddings of the same speaker be zero and those of different speakers be at least arccos δ apart

$$\mathcal{L}_{\text{pair}} = \sum_{i,j\in\{1,\ldots,S^*\}} \frac{1}{S^2 c_i c_j} f(\boldsymbol{b}_i, \boldsymbol{b}_j)$$

$$f(\boldsymbol{b}_i, \boldsymbol{b}_j) = \begin{cases} 1 - r_{ij} & \text{(i-th and j-th local attractors correspond to the same speaker)} \\ \max(0, r_{ij} - \delta) & \text{(i-th and j-th local attractors correspond to the different speakers)} \end{cases}$$

S: # of speakers, S^*: # of local attractors
c_i: # of local attractors assigned to the i-th speaker

EEND with Local Attractors – Inference



Block#1

Posteriors

Relative speaker embeddings

$\boldsymbol{b}_1^{(1)}$ $\boldsymbol{b}_2^{(1)}$

Block#2

Posteriors

Relative speaker embeddings

$\boldsymbol{b}_1^{(2)}$

…

Block#L

Posteriors

Relative speaker embeddings

$\boldsymbol{b}_1^{(L)}$ $\boldsymbol{b}_2^{(L)}$ $\boldsymbol{b}_3^{(L)}$

Affinity matrix

R'

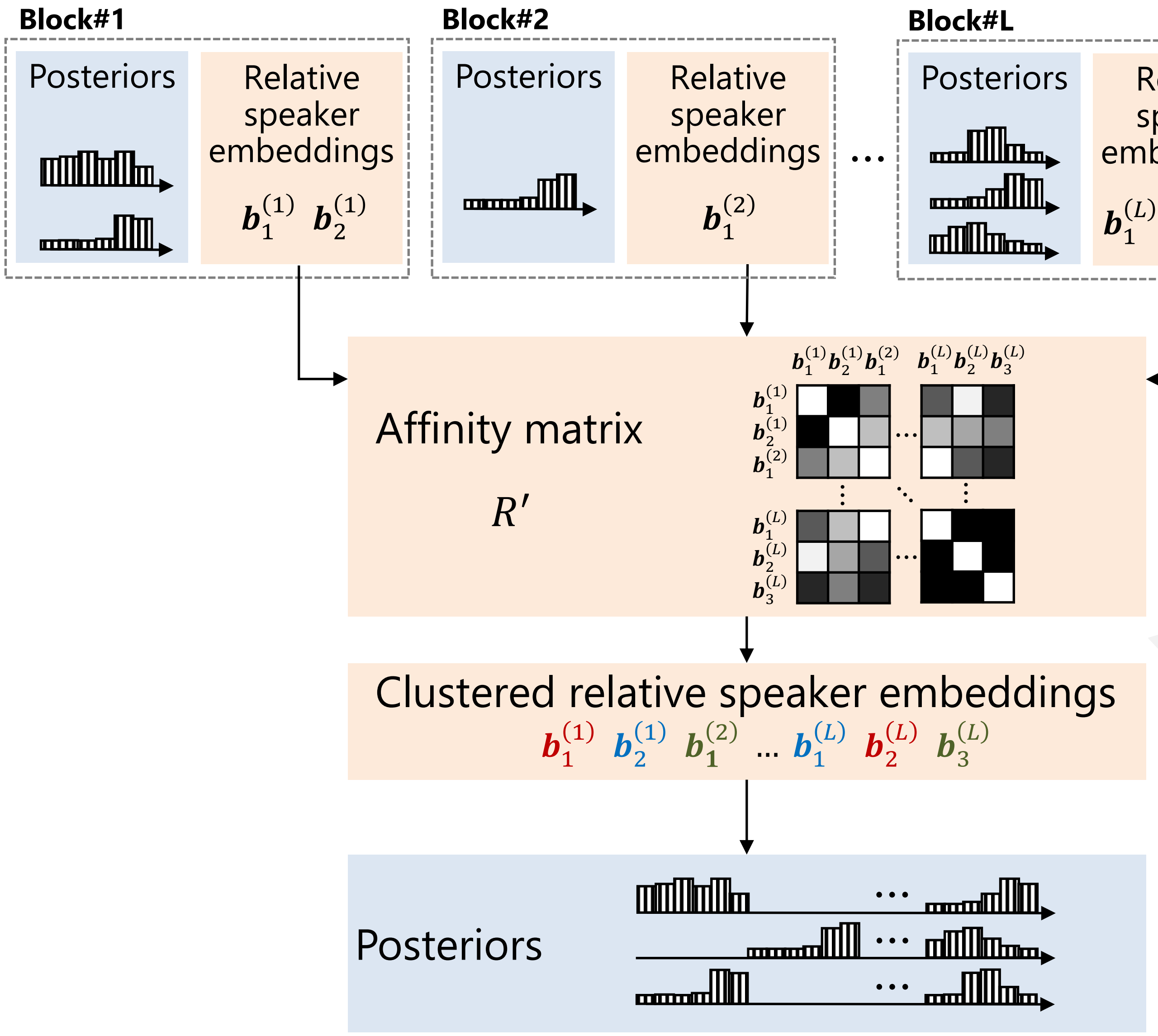
Clustered relative speaker embeddings

$\boldsymbol{b}_1^{(1)}$ $\boldsymbol{b}_2^{(1)}$ $\boldsymbol{b}_1^{(2)}$ … $\boldsymbol{b}_1^{(L)}$ $\boldsymbol{b}_2^{(L)}$ $\boldsymbol{b}_3^{(L)}$

Posteriors

Clustering the relative speaker embeddings based on the affinity matrix

EEND with Local Attractors – Inference



Block#1

Posteriors

Relative speaker embeddings

$\boldsymbol{b}_1^{(1)}$ $\boldsymbol{b}_2^{(1)}$

Block#2

Posteriors

Relative speaker embeddings

$\boldsymbol{b}_1^{(2)}$

…

Block#L

Posteriors

Relative speaker embeddings

$\boldsymbol{b}_1^{(L)}$ $\boldsymbol{b}_2^{(L)}$ $\boldsymbol{b}_3^{(L)}$

Affinity matrix

R'

$\boldsymbol{b}_1^{(1)} \boldsymbol{b}_2^{(1)} \boldsymbol{b}_1^{(2)}$ $\boldsymbol{b}_1^{(L)} \boldsymbol{b}_2^{(L)} \boldsymbol{b}_3^{(L)}$

Clustered relative speaker embeddings

$\boldsymbol{b}_1^{(1)}$ $\boldsymbol{b}_2^{(1)}$ $\boldsymbol{b}_1^{(2)}$ … $\boldsymbol{b}_1^{(L)}$ $\boldsymbol{b}_2^{(L)}$ $\boldsymbol{b}_3^{(L)}$

Posteriors

1. Estimate the number of speakers

1-1. Apply matrix decomposition

$$R' = V \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{S^*} \end{bmatrix} V^{-1}$$

$\lambda_1 \geq \cdots \geq \lambda_{S^*}$: Eigenvalues
V: Eigenvectors

Eigenvalues are good indicators of the size of clusters

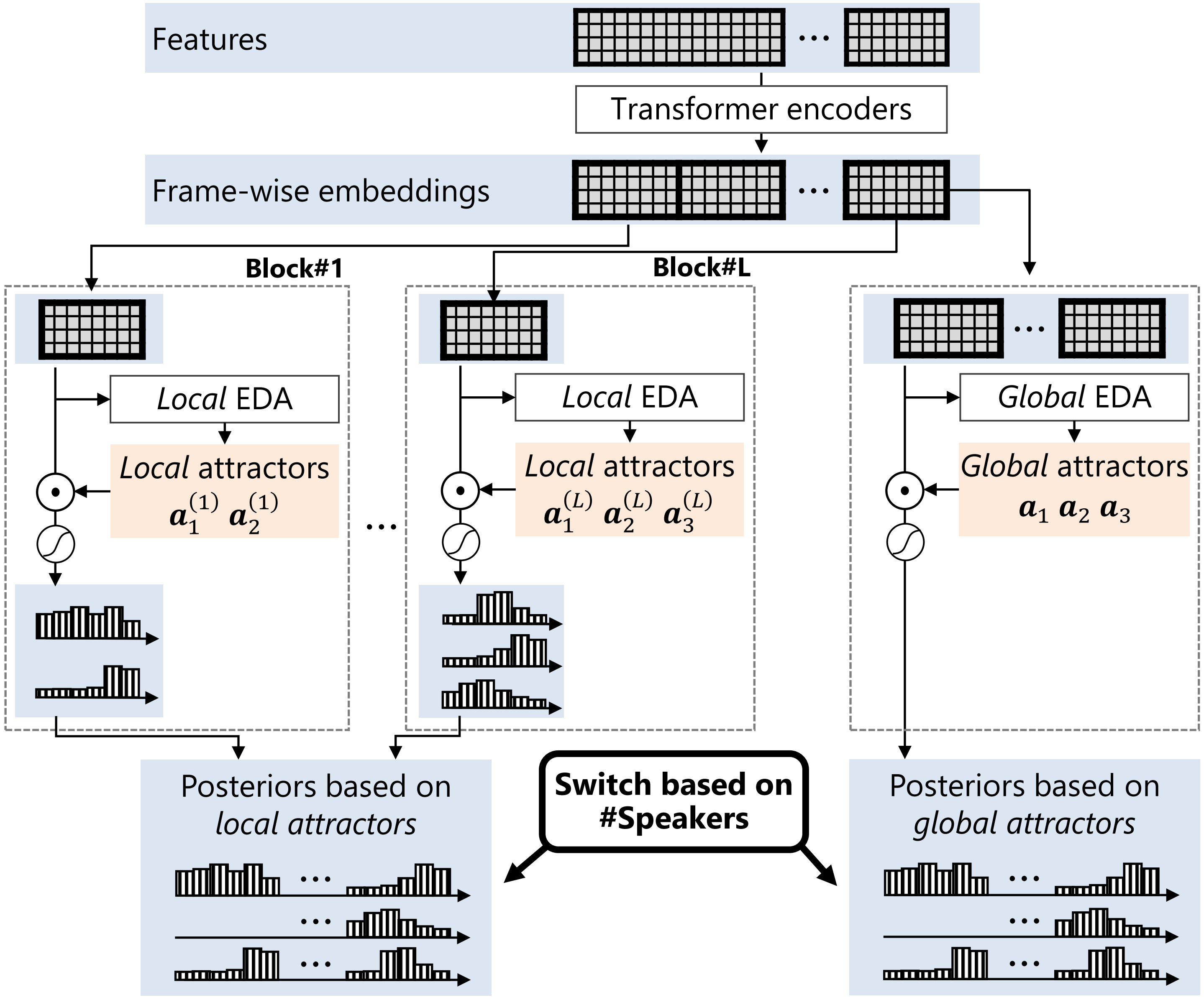
1-2. Estimate the number of speakers

$$\hat{S} = \min_{\substack{1 \leq s \leq S^*-1 \\ \lambda_s \geq 1}} \frac{\lambda_{s+1}}{\lambda_s}$$

2. Apply CLC-Kmeans clustering [Yang+’13]

To satisfy cannot-link constraints
(The attractors from the same block must be assigned to different clusters)

EEND-GLA: EEND with Globa and Local Attractors



Features

Transformer encoders

Frame-wise embeddings

Block#1

Local EDA

Local attractors

$\boldsymbol{a}_1^{(1)}\ \boldsymbol{a}_2^{(1)}$

Block#L

Local EDA

Local attractors

$\boldsymbol{a}_1^{(L)}\ \boldsymbol{a}_2^{(L)}\ \boldsymbol{a}_3^{(L)}$

Global EDA

Global attractors

$\boldsymbol{a}_1\ \boldsymbol{a}_2\ \boldsymbol{a}_3$

Posteriors based on *local attractors*

Switch based on #Speakers

Posteriors based on *global attractors*

Global-attractor-based diarization is still powerful when #Speakers is small

→ Use both global and local attractors

Training:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{global}}}_{\text{Loss of EEND-EDA}} + \mathcal{L}_{\text{local}}$$

Inference:

When EEND-GLA is trained using at most N-speaker mixtures

- If the estimated #Speakers ≥ *N*
  → Use the results based on *global* attractors
- If the estimated #Speakers < *N*
  → Use the results based on *local* attractors

Experimental Settings

- **Model configuration**
  - EEND-GLA-Small: The proposed method with 4-layer Transformer encoders
  - EEND-GLA-Large: The proposed method with 6-layer Transformer encoders

- **Datasets**

Simulated conversational datasets

	Dataset	#Spk	#Mixtures	Overlap ratio
Train	Sim1spk	1	100,000	0.0 %
	Sim2spk	2	100,000	34.1 %
	Sim3spk	3	100,000	34.2 %
	Sim4spk	4	100,000	31.5 %
Test	Sim1spk	1	500	0.0 %
	Sim2spk	2	500	34.4 %
	Sim3spk	3	500	34.7 %
	Sim4spk	4	500	32.0 %
	Sim5spk	5	500	30.7 %
	Sim6spk	6	500	29.9 %

} Not observed using training

Real conversational datasets

	Dataset	#Spk	#Mixtures	Overlap ratio
Adapt	CALLHOME	2-7	249	17.0 %
	DIHARD II dev	1-10	192	9.8 %
	DIHARD III dev	1-10	254	10.7 %
Test	CALLHOME	2-6	250	16.7 %
	DIHARD II eval	1-9	194	8.9 %
	DIHARD III eval	1-9	259	9.2 %

Results on the Simulated Datasets

- **DERs (%)**
  - EEND-GLA significantly reduced DER of unseen numbers of speakers

	#Speakers					
	1	2	3	4	5	6
X-vector clustering	37.42	7.74	11.46	22.45	31.00	38.62
EEND-EDA	0.15	**3.19**	6.60	8.68	22.43	33.28
EEND-GLA-Small	0.25	3.53	6.79	8.98	**12.44**	**17.98**
EEND-GLA-Large	**0.09**	3.54	**5.74**	**6.79**	12.51	20.42

Observed during training (1–4)

Not observed during training (5–6)

- **Speaker counting accuracy**
  - The number of speakers was predicted more accurately when the number of speakers is five or larger

EEND-EDA: Acc. 71.9%

Reference #Speakers

Predicted #Speakers

	1	2	3	4	5	6
1	**500**	0	0	0	0	0
2	0	**482**	0	0	0	0
3	0	17	**435**	5	1	0
4	0	1	65	**447**	224	139
5	0	0	0	48	**268**	337
6	0	0	0	0	7	**24**
7+	0	0	0	0	0	0

Observed (1–4)

Not observed (5–6)

EEND-GLA-Small: Acc. 80.8%

Reference #Speakers

	1	2	3	4	5	6
1	**498**	0	0	0	0	0
2	2	**474**	0	0	0	0
3	0	25	**451**	17	2	0
4	0	1	33	**412**	78	30
5	0	0	10	62	**361**	183
6	0	0	6	7	47	**229**
7+	0	0	0	2	12	57

Observed (1–4)

Not observed (5–6)

Results on the Real Recordings

- EEND-GLA performed better than EEND-EDA when the number of speakers is large

CALLHOME	#Speakers					
	2	3	4	5	6	All
X-vector clustering [Landini+'21]*	9.44	13.89	16.05	**13.87**	24.73	13.28
EEND-EDA	7.83	12.29	17.59	27.66	37.17	13.65
EEND-vector clustering [Kinoshita+'21]	7.96	11.93	16.38	21.21	23.10	12.49
EEND-GLA-Small	**6.94**	**11.42**	14.49	29.76	24.09	11.92
EEND-GLA-Large	7.11	11.88	**14.37**	25.95	**21.95**	**11.84**

* Oracle speech segments were used for x-vector clustering

DIHARD II	#Speakers		
	≤4	≥5	All
X-vector clustering + overlap handling [Landini+'20]	**21.34**	39.85	27.11
X-vector clustering + overlap-aware resegmentation [Bredin+'21]	21.41	**36.93**	**26.25**
EEND-EDA	22.09	47.66	30.07
EEND-GLA-Small	22.24	44.92	29.31
EEND-GLA-Large	21.40	43.62	28.33

DIHARD III	#Speakers		
	≤4	≥5	All
X-vector clustering + overlap handling [Horiguchi+'21]	16.38	42.51	21.47
X-vector clustering + overlap-aware resegmentation [Coria+'21]	15.32	**35.87**	**19.33**
EEND-EDA	15.55	48.30	21.94
EEND-GLA-Small	14.39	44.32	20.23
EEND-GLA-Large	**13.64**	43.67	19.49

Summary of Chapter 3

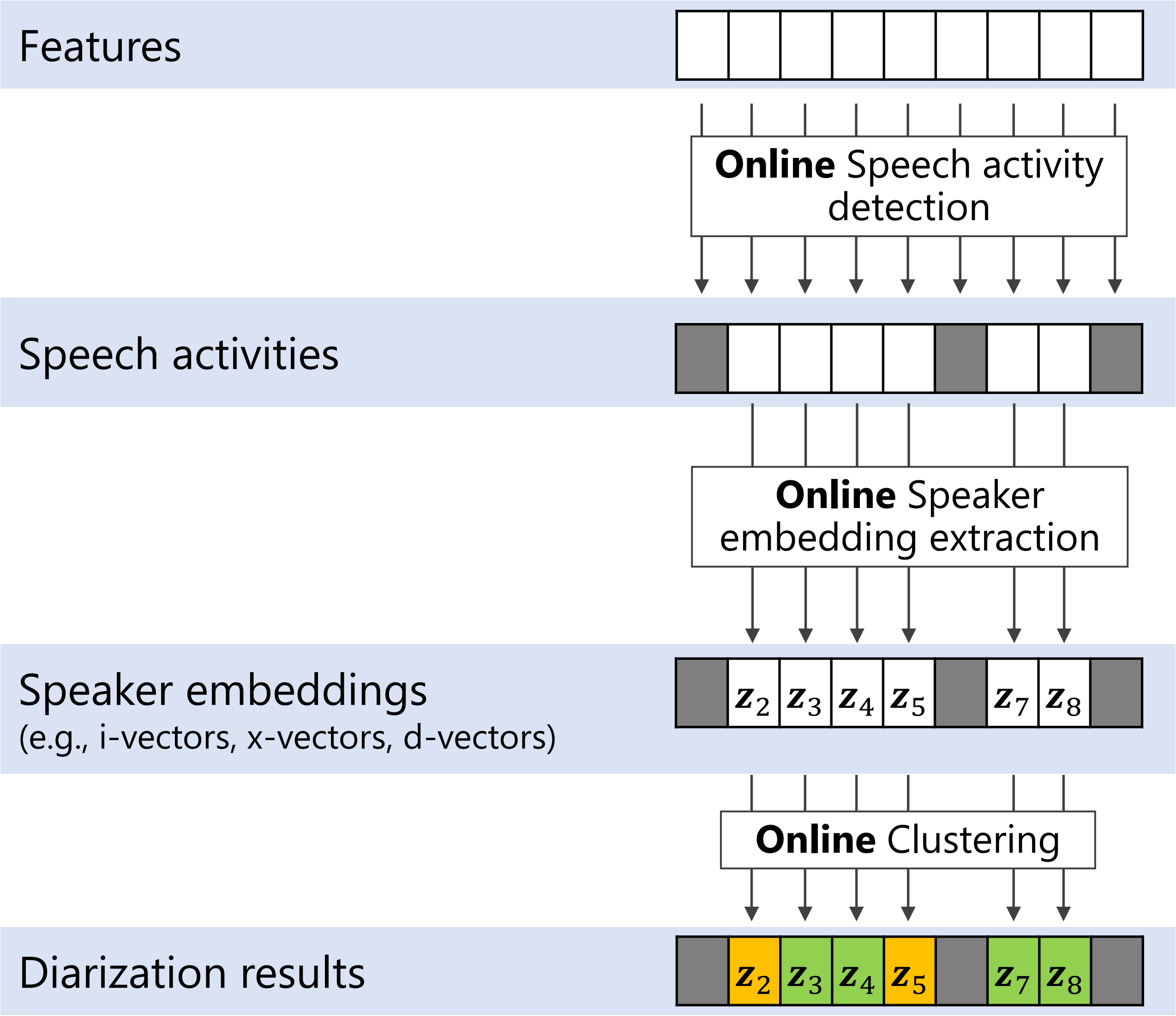
- **Problem**
  - The conventional EEND assumes that the number of speakers is known in advance

- **Solutions**
  - 3-1: End-to-end speaker diarization for **flexible** numbers of speakers
    - Core contribution: Encoder-decoder based attractors for EEND (EEND-EDA)
    - Related publications: [INTERSPEECH’20] [TASLP’22]

  - 3-2: End-to-end speaker diarization for **unlimited** numbers of speakers
    - Core contribution: Use of attractors from calculated from global and local contexts (EEND-GLA)
    - Related publication: [ASRU’21] [TASLP’23]

  - 3-3: **Online** end-to-end speaker diarization for unlimited numbers of speakers
    - Core contribution: An extension to speaker-tracing buffer to make it compatible with EEND-GLA
    - Related publication: [TASLP’23]

Related Work: Online Cascaded Speaker Diarization



Features

Online Speech activity detection

Speech activities

Online Speaker embedding extraction

Speaker embeddings
(e.g., i-vectors, x-vectors, d-vectors)

z_2 z_3 z_4 z_5 z_7 z_8

Online Clustering

Diarization results

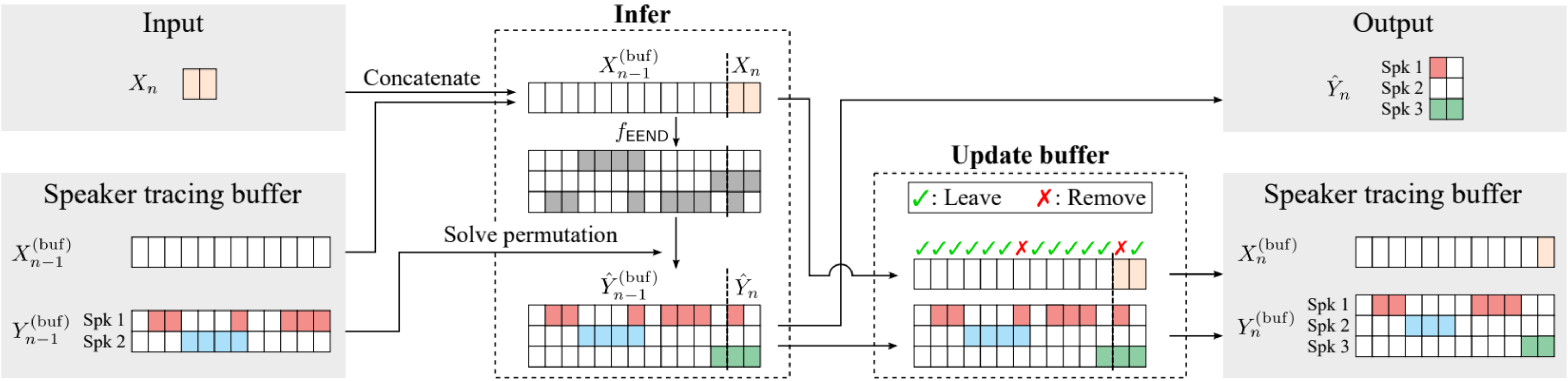
z_2 z_3 z_4 z_5 z_7 z_8

- Each module needs to be replaced to the online one
- Especially, online clustering causes a severe performance drop
  - High performance offline clustering methods are often two-staged [Landini+'22] [Bredin+'21]
  - Therefore, it is not applicable to online processing, i.e., we need completely different clustering algorithm for online processing

	DIHARD II	**DIHARD III**
Offline x-vector clustering [Bredin+'21]	26.25	19.33
Online x-vector clustering [Coria+'21]	34.99	27.55

Related Work: Online EEND

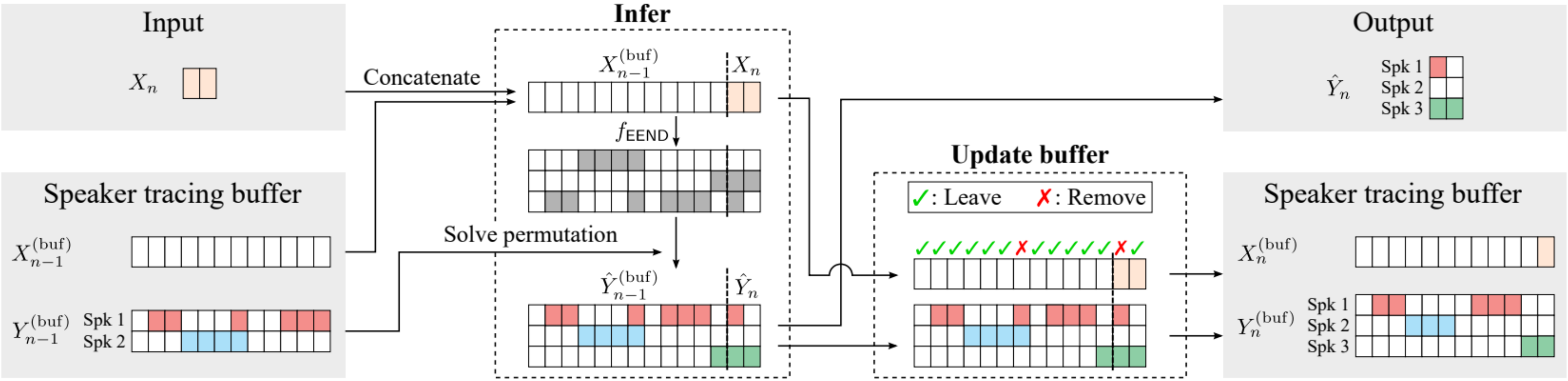
- **Speaker tracing buffer (Frame-wise speaker tracing buffer; FW-STB)** [Xue+'21]



- Assume block-wise input features X_n $(n = 1, 2, \dots)$
- FW-STB stores the features and corresponding estimated results
  - $X_{n-1}^{(\mathrm{buf})}$: Features
  - $Y_{n-1}^{(\mathrm{buf})}$: Previously estimated diarization results

Related Work: Online EEND

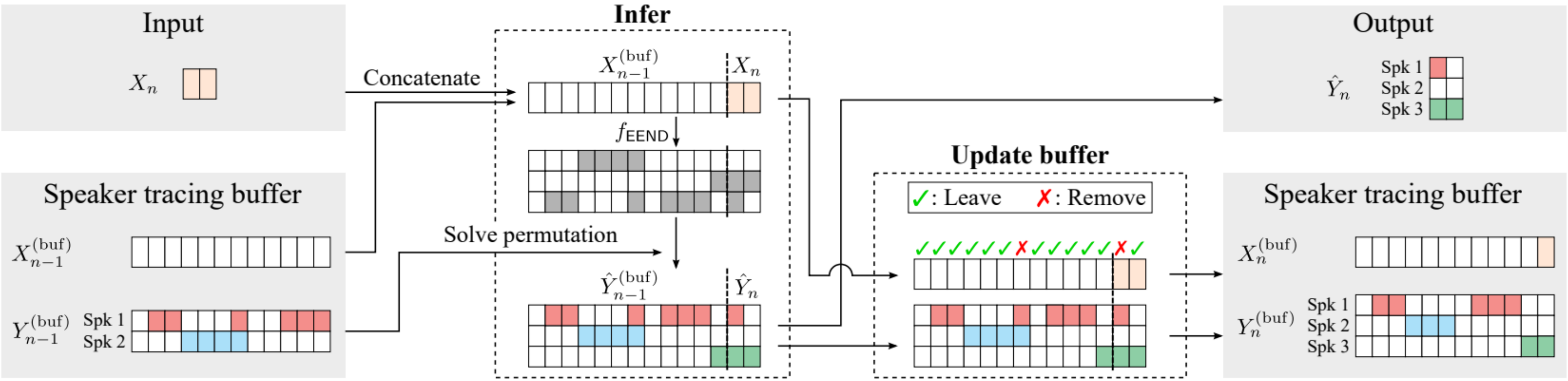
- **Speaker tracing buffer (Frame-wise speaker tracing buffer; FW-STB)** [Xue+'21]



1. Estimate diarization results from $\left[X_{n-1}^{(\mathrm{buf})} \quad X_n\right]$
2. Align the order of speakers to maximize the correlation between the new results and $Y_{n-1}^{(\mathrm{buf})}$
3. Output the estimated results $\hat{Y}_n$

Related Work: Online EEND

- **Speaker tracing buffer (Frame-wise speaker tracing buffer; FW-STB)** [Xue+'21]



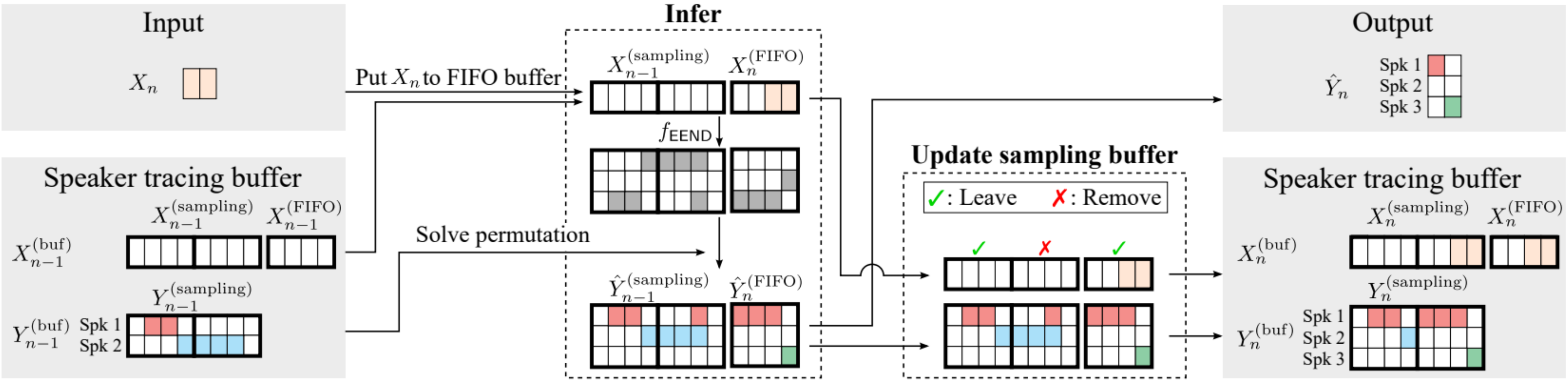
FW-STB is updated by sampling informative frames

- The frames where only one speaker is dominant are selected
- The features and corresponding estimation of those frames are stored in the buffer

✘ The frames are not consecutive → Not compatible with EEND-GLA

Proposed Method: Online Extension for EEND-GLA

- **Block-wise speaker tracing buffer (BW-STB)**

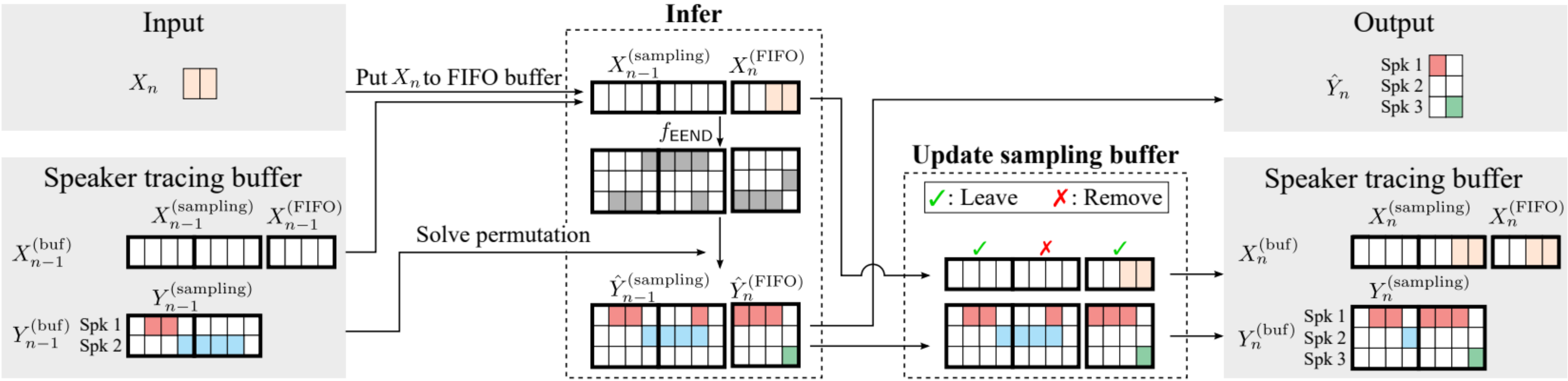


Block-wise speaker tracing buffer consists of two types of buffers

- **Block-wise sampling buffer:**
  - Consists of multiple blocks
  - Each block stores features and corresponding results of consecutive frames
- **First-in-first-out (FIFO) buffer:**
  - Consists of a single block
  - Stores recent features

Proposed Method: Online Extension for EEND-GLA

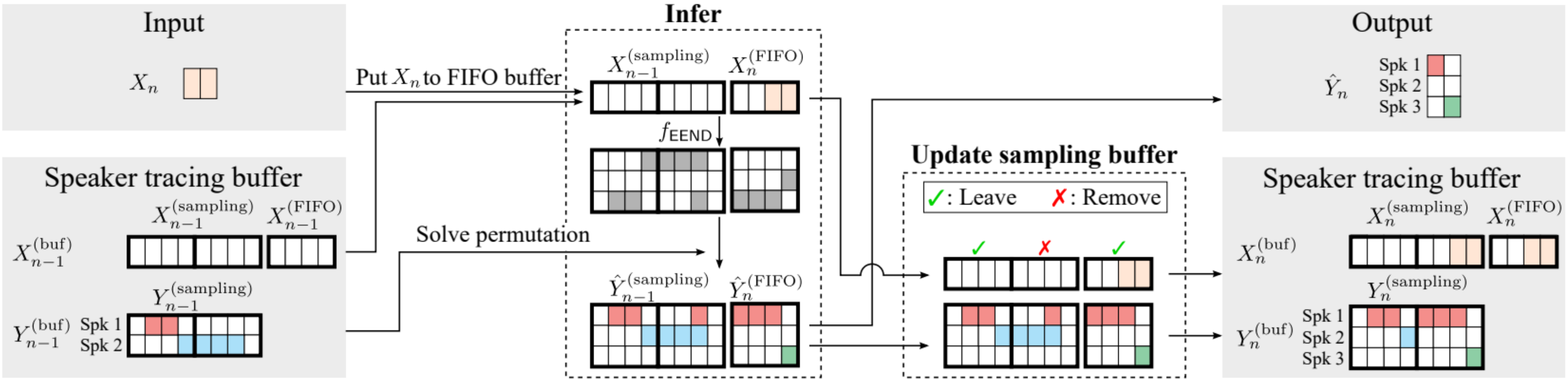
- **Block-wise speaker tracing buffer (BW-STB)**



1. Put the input feature X_n to the FIFO buffer
2. Estimate diarization results and solve speaker permutation in the same manner as FW-STB
3. Finally, output the results that correspond to X_n

Proposed Method: Online Extension for EEND-GLA

- **Block-wise speaker tracing buffer (BW-STB)**



Block-wise sampling to update the buffer

✓ BW-STB can be used with EEND-GLA because each block in BW-STB stores features and the corresponding results of consecutive frames

✓ Use of the FIFO buffer together enables low-latency processing

Experimental Settings for Online Experiments

- **Model configuration**
  - EEND-GLA-Small: The proposed method with 4-layer Transformer encoders
  - EEND-GLA-Large: The proposed method with 6-layer Transformer encoders

- **Datasets (same as the offline experiments)**
  - Simulated datasets
    - Training set: Sim{1,2,3,4}spk
    - Evaluation set: Sim{1,2,3,4,5,6}spk
  - Real datasets
    - CALLHOME
    - DIHARD II
    - DIHARD III

- **Settings for online experiments**
  - Features are input every second (=10 features)
  - Set the buffer length 100 seconds
    - BW-STB: Block-wise sampling buffer of 95 seconds length (5 seconds * 19 blocks) & FIFO buffer of 5 seconds

Results on the Simulated Datasets

- DERs (%) on the simulated mixtures
  - EEND-GLA with BW-STB improved DERs of unseen numbers of speakers compared to EEND-EDA with FW-STB

	#Speakers					
	1	2	3	4	5	6
BW-EDA-EEND [Han+'21]	**1.03**	6.10	12.58	19.17	N/A	N/A
EEND-EDA + FW-STB [Xue+'21]	1.50	5.91	9.79	11.85	26.63	37.25
EEND-GLA-Small + BW-STB	1.19	5.18	9.41	13.19	**16.95**	**22.55**
EEND-GLA-Large + BW-STB	1.12	**4.61**	**8.14**	**11.38**	17.27	25.77

Observed during training

Not observed during training

- Speaker counting accuracy
  - The number of speakers was predicted more accurately when the number of speakers is five or larger

EEND-EDA + FW-STB

Reference #Speakers

Predicted #Speakers

	1	2	3	4	5	6
1	**376**	0	0	0	0	0
2	120	**244**	0	0	0	0
3	4	249	**252**	1	1	0
4	0	7	245	**449**	271	172
5	0	0	3	50	**222**	314
6	0	0	0	0	7	**14**
7+	0	0	0	0	0	0

Observed

Not observed

EEND-GLA-Small + BW-STB

Reference #Speakers

	1	2	3	4	5	6
1	**411**	0	0	0	0	0
2	84	**343**	0	0	0	0
3	5	156	**370**	3	0	0
4	0	1	109	**302**	16	0
5	0	0	20	181	**364**	38
6	0	0	1	13	114	**385**
7+	0	0	0	1	6	77

Observed

Not observed

Results on the Real Recordings

- EEND-GLA + BW-STB improved the DERs from EEND-EDA + FW-STB when # of speakers is large (≥5)
- EEND-based methods suppressed the degradation due to online processing

DIHARD II dataset

Offline	**#Speakers**		
	≤4	**≥5**	**All**
X-vector clustering [Bredin+'21]	21.41	**36.93**	**26.25**
EEND-EDA	22.09	47.66	30.07
EEND-GLA-Small	22.24	44.92	29.31
EEND-GLA-Large	**21.40**	43.62	28.33

+8.74% →

+3.30% →

+2.16% →

+1.91% →

Online	**#Speakers**		
	≤4	**≥5**	**All**
X-vector clustering [Coria+21]	27.00	52.62	34.99
EEND-EDA + FW-STB [Xue+'21]	25.63	50.45	33.37
EEND-GLA-Small + BW-STB	23.96	48.06	31.47
EEND-GLA-Large + BW-STB	**22.62**	**47.06**	**30.24**

Results on the Real Recordings

- EEND-GLA + BW-STB improved the DERs from EEND-EDA + FW-STB when # of speakers is large (≥5)
- EEND-based methods suppressed the degradation due to online processing

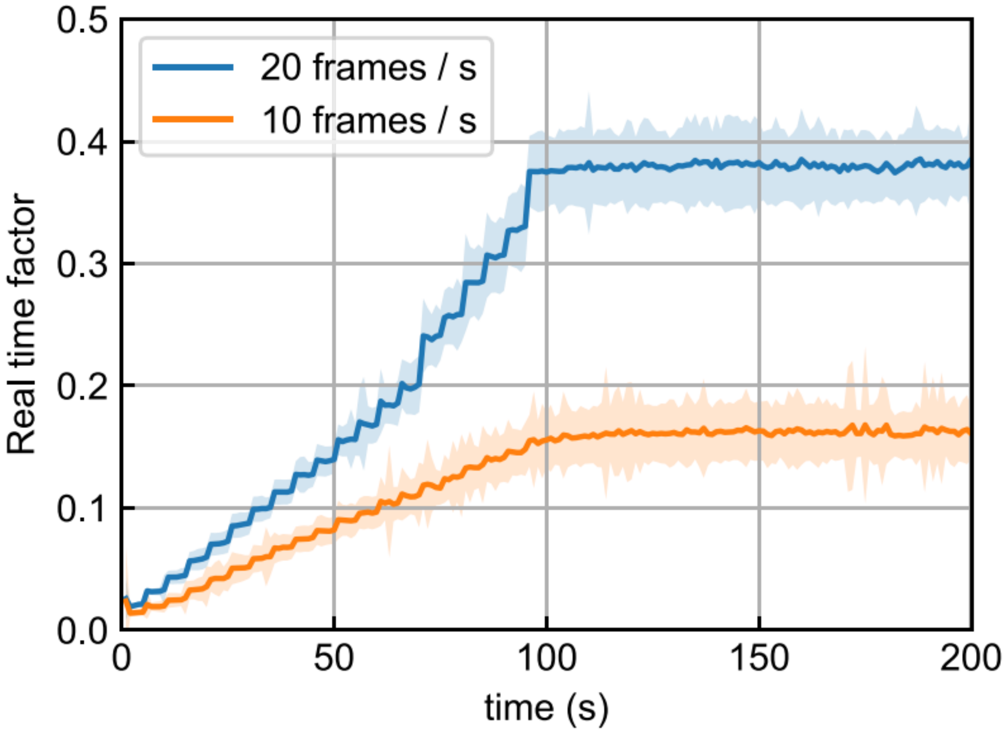
DIHARD III dataset

Offline	**#Speakers**				**Online**	**#Speakers**		
	≤4	**≥5**	**All**			**≤4**	**≥5**	**All**
X-vector clustering [Bredin+'21]	15.32	**35.87**	**19.33**	+8.22% →	X-vector clustering [Coria+'21]	21.07	54.28	27.55
EEND-EDA	15.55	48.30	21.94	+3.15% →	EEND-EDA + FW-STB [Xue+'21]	19.00	50.21	25.09
EEND-GLA-Small	14.39	44.32	20.23	+1.77% →	EEND-GLA-Small + BW-STB	15.87	47.24	22.00
EEND-GLA-Large	**13.64**	43.67	19.49	+1.24% →	EEND-GLA-Large + BW-STB	**14.81**	**45.17**	**20.73**

Real Time Factor

- Computing environment:
  - Intel Xeon Gold 6132 CPU @ 2.60 GHz using 7 threads
  - No GPU was used

- Dataset
  - Sim5spk



- Real time factor increased linearly until the buffer was filled
  - The time complexity of EEND-GLA is $O(n^3)$, but not constrained by it at least for buffer length of 100 sec.
- After the convergence, the real time factors were 0.16 (10 frames / sec) and 0.38 (20 frames / sec)

Summary of Chapter 3

- **Problem**
  - The conventional EEND assumes that the number of speakers is known in advance

- **Solutions**
  - 3-1: End-to-end speaker diarization for **flexible** numbers of speakers
    - Core contribution: Encoder-decoder based attractors for EEND (EEND-EDA)
    - Related publications: [INTERSPEECH’20] [TASLP’22]

  - 3-2: End-to-end speaker diarization for **unlimited** numbers of speakers
    - Core contribution: Use of attractors from calculated from global and local contexts (EEND-GLA)
    - Related publication: [ASRU’21] [TASLP’23]

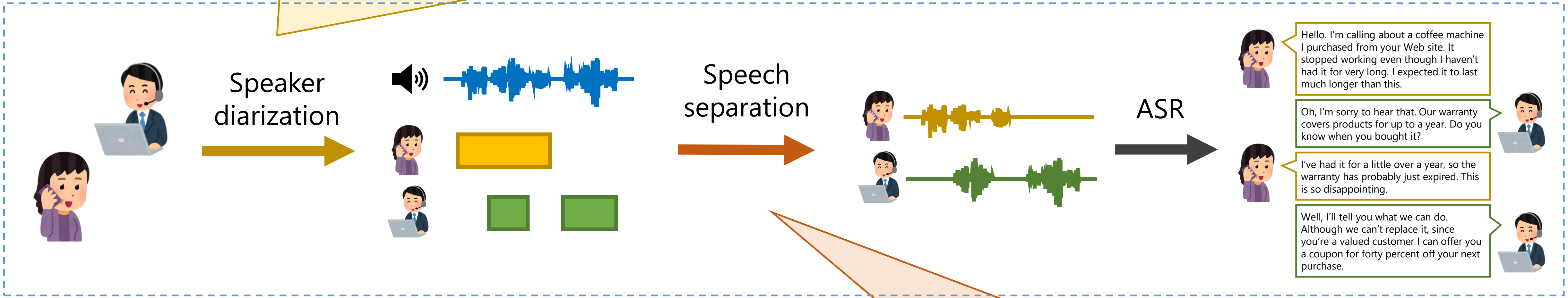
  - 3-3: **Online** end-to-end speaker diarization for unlimited numbers of speakers
    - Core contribution: An extension to speaker-tracing buffer to make it compatible with EEND-GLA
    - Related publication: [TASLP’23]

Thesis Overview

Chapter 3
End-to-end speaker diarization for *unknown numbers of speakers*
[TASLP'22] [TASLP'23] [INTERSPEECH'20] [ASRU'21]

Chapter 4
Multi-channel
end-to-end speaker diarization
[ICASSP'22] [SLT'22]

Chapter 5
End-to-end speaker diarization as *post-processing*
[ICASSP'21]



Hello. I'm calling about a coffee machine I purchased from your Web site. It stopped working even though I haven't had it for very long. I expected it to last much longer than this.

Oh, I'm sorry to hear that. Our warranty covers products for up to a year. Do you know when you bought it?

I've had it for a little over a year, so the warranty has probably just expired. This is so disappointing.

Well, I'll tell you what we can do. Although we can't replace it, since you're a valued customer I can offer you a coupon for forty percent off your next purchase.

Chapter 6
Speaker-diarization-driven meeting transcription
[INTERSPEECH'20]

Chapter 7
Block online speech separation using speaker diarization results
[SLT'21]

Chapter 4: Multi-Channel End-to-End Speaker Diarization

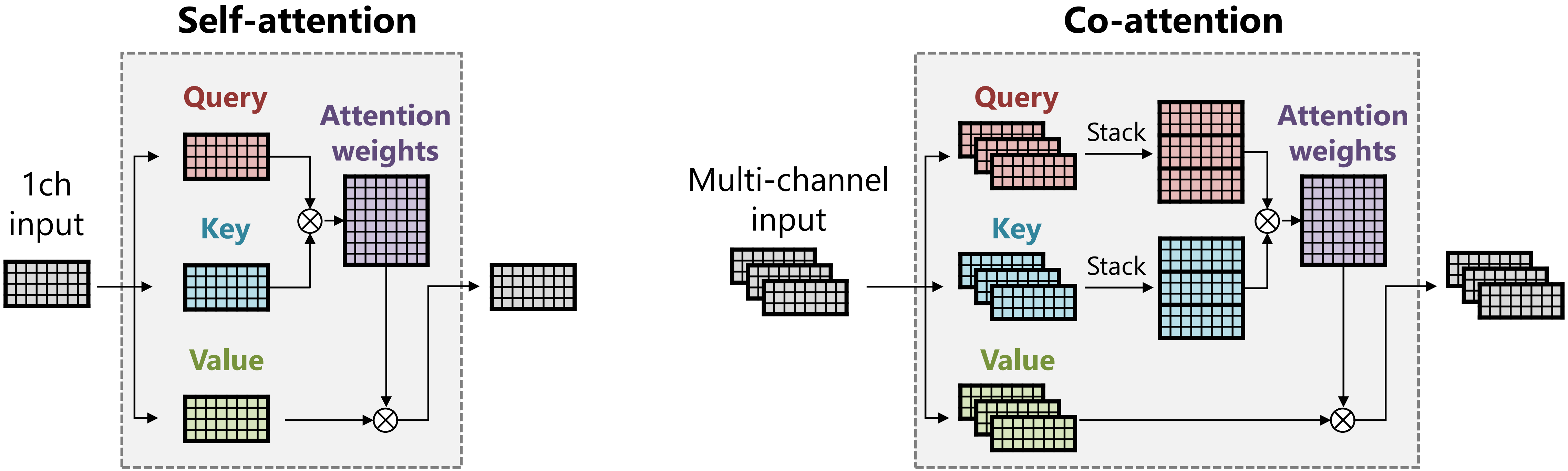
- **Problem**
  - EEND / EEND-EDA / EEND-GLA only utilize spectral information from single-channel inputs
  - Existing multi-channel methods highly depend on spatial information

- **Solutions**
  - 4-1: **Multi-channel** end-to-end speaker diarization that also handles single-channel inputs
    - Core contribution: Co-attention encoder that not rely on cross-channel attention
    - Related publication: [ICASSP’22]

  - 4-2: **Training method** of single- and multi-channel end-to-end speaker diarization
    - Core contribution: Iterative operation of transfer learning and knowledge distillation between two
    - Related publication: [SLT’22]

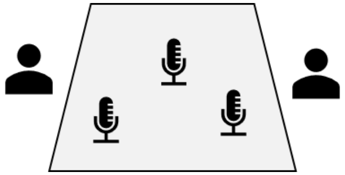
4-1: Co-Attention-Based Multi-Channel EEND

- **Method: Co-attention**
  - Process multi-channel input
  - Equivalent to the conventional self-attention when the number of channel is one
    → Not heavily rely on spatial information

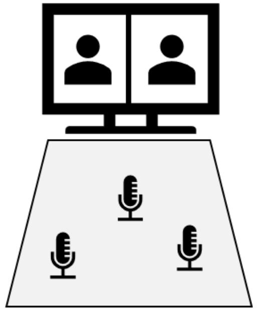


4-1: Co-Attention-Based Multi-Channel EEND

- **Datasets**
  - Two types of simulated 10-channel two-speaker datasets
    - Sim2spk-multi: Two speakers are at the different positions
    - Sim2spk-multi-hybrid: Two speakers are at the same position



Sim2spk-multi



Sim2spk-multi-hybrid

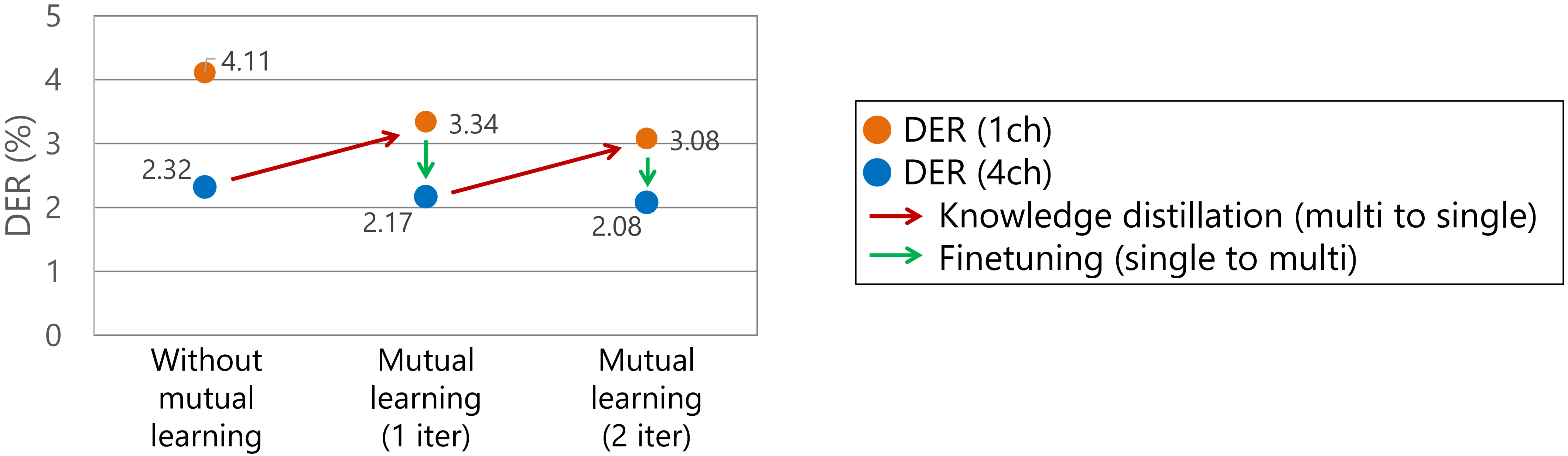
- **Results**
  - Co-attention-based model improved DER by utilizing spatial information
  - Co-attention-based model did not degrade even when spatial information is not available
    - Sim2spk-multi (1ch) & Sim2spk-multi-hybrid (1, 2, 4, 6, 10ch)

	Sim2spk-multi					**Sim2spk-multi-hybrid**				
Algorithm	**1ch**	**2ch**	**4ch**	**6ch**	**10ch**	**1ch**	**2ch**	**4ch**	**6ch**	**10ch**
1ch + posterior avg.	5.13	4.60	4.31	4.19	4.10	6.07	5.68	5.42	5.38	**5.33**
Spatio-temporal [Wang+'21]	6.34	3.02	**1.56**	**1.28**	**1.07**	8.11	8.23	6.98	6.72	6.40
Co-attention (proposed)	**4.68**	**2.52**	1.71	1.40	1.23	**5.73**	**5.34**	**5.05**	**5.18**	5.35

4-2: Mutual Learning of Single and Multi-Channel EEND

- **Method: Mutual learning**
  - Iteratively conduct the following:
    - Knowledge distillation from **multi-channel** to **single-channel EEND**
    - Finetuning from **single-channel** to **multi-channel EEND**

- **Results**
  - Proposed method improved DERs of both single- and multi-channel EEND

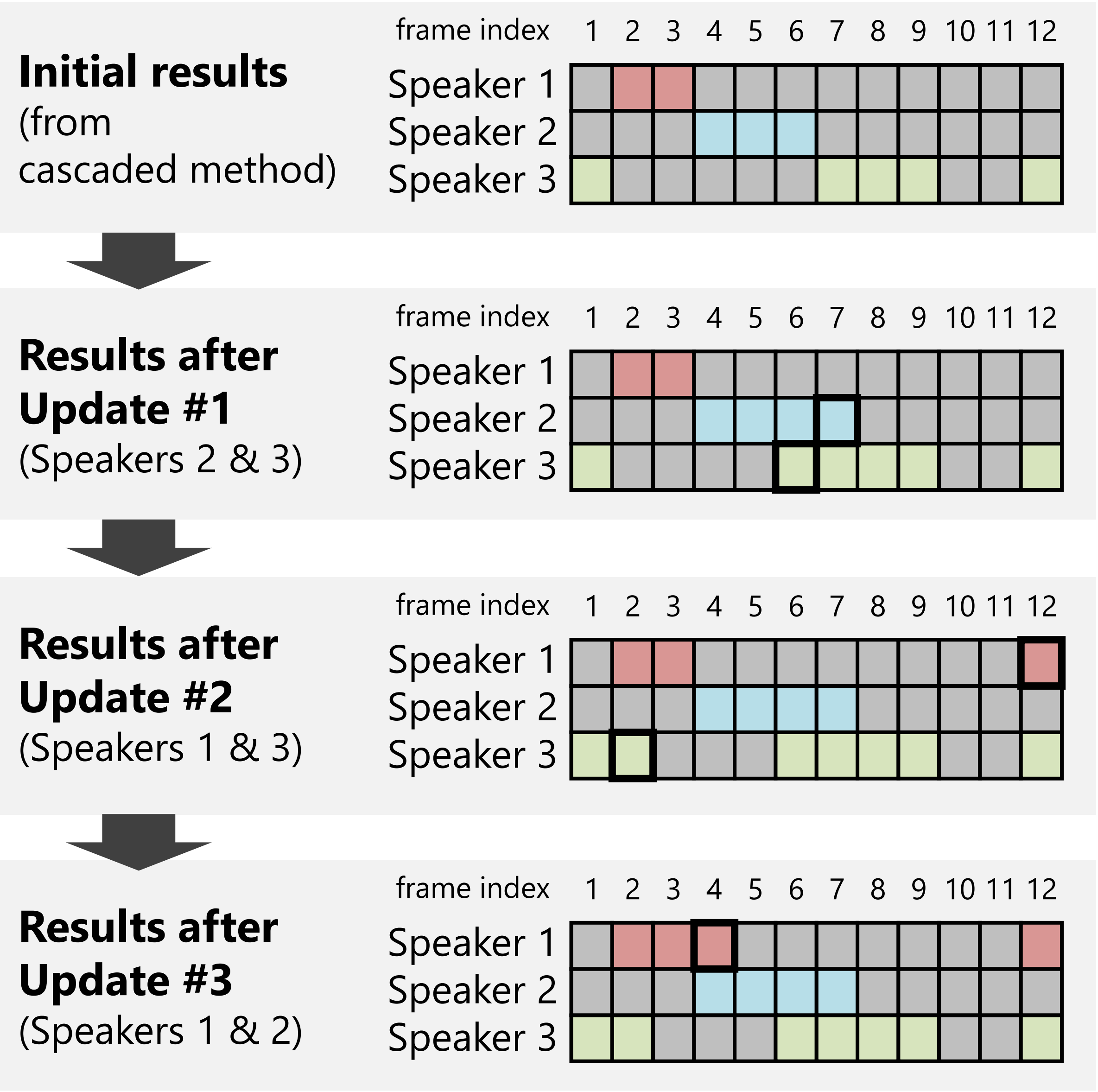


Summary of Chapter 5

- **Problem**
  - While the end-to-end approach is promising, cascaded approaches are still powerful
  - Cascaded approaches require overlap detection and speaker assignment as the last step

- **Solutions**
  - Use end-to-end speaker diarization for overlap detection and speaker assignment of cascaded approaches
    - Core contribution: An algorithm to use EEND to refine the results from cascaded approaches (EEND as post-processing)
    - Related publication: [ICASSP'21]

Chapter 5: End-to-End Speaker Diarization as Post-Processing



Initial results (from cascaded method)

Results after Update #1 (Speakers 2 & 3)

Results after Update #2 (Speakers 1 & 3)

Results after Update #3 (Speakers 1 & 2)

- **Method: EEND as post-processing (EENDasP)**

  For each speaker pair:

  1. Select frames not containing other speakers
  2. Process the frames using two-speaker EEND
  3. Update the results of the frames

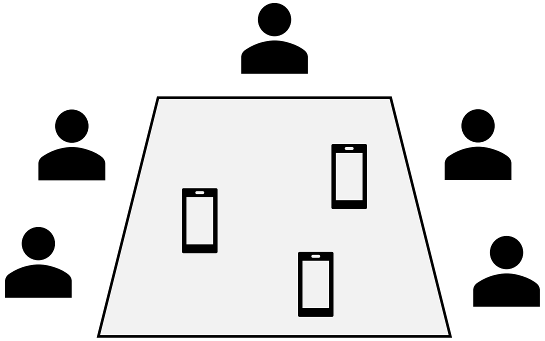
- **Results on the DIHARD II dataset**
  - ✓ Consistently improved DERs on both datasets
  - ✓ Can be used with other overlap detection and assignment methods

Model	DER (%)
DIHARD II baseline [Sell+'20]	40.86
DIHARD II baseline + EENDasP	37.90
BUT system (w/o OVL) [Landini+'20]	27.26
BUT system (w/o OVL) + EENDasP	26.91
BUT system (w/ OVL) [Landini+'20]	27.11
BUT system (w/ OVL) + EENDasP	**26.88**

OVL: Heuristic-based speaker assignment

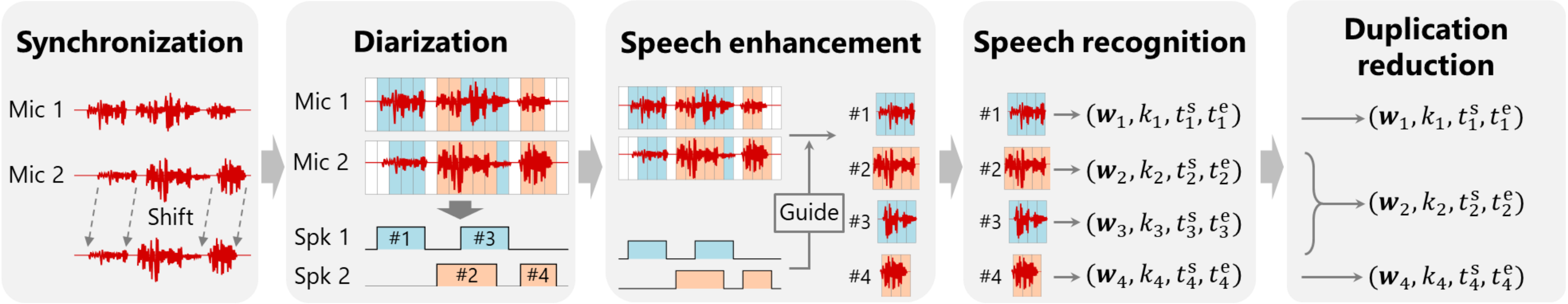
Summary of Chapter 6

- **Purpose**
  - To show how speaker diarization is important for meeting transcription using distributed microphones (e.g., smartphone / tablet device) without any special devices (microphone arrays, omnidirectional cameras)



- **Solutions**
  - Speaker-diarization-driven meeting transcription using distributed microphones
    - Core contribution: Demonstration of the effectiveness of the diarization-driven ASR on the realistic data
    - Related publication: [INTERSPEECH’20]

Chapter 6: System Overview



Synchronization using signal correlation

Cascaded diarization with oracle number of speakers

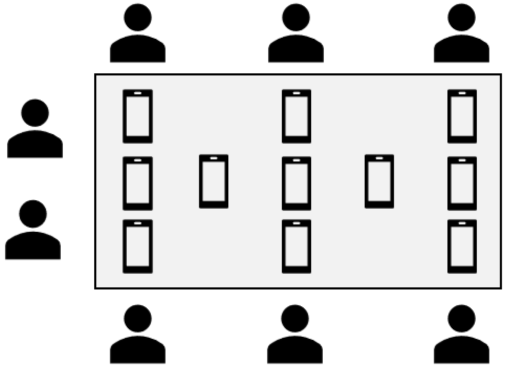
Diarization-guided speech enhancement [Boeddeker+'18]

ASR system [Kanda+'18]

Similar results with highly overlapped segments are merged

Chapter 6: Result Summary

- **Dataset**
  - ~2 hours of meeting consists of 8 sessions
  - 5-8 participants
  - 11 smartphones for recording
  - Each participant wore a headset microphone



- **Results**
  - Using multiple microphones successfully reduced the ASR performance measured using the character error rates (CERs)
  - If the oracle diarization results were given,
    the system achieved nearly headset-level CER
    → **Highly accurate speaker diarization is important**
  - Limitation: Diarization-guided speech separation is super slow

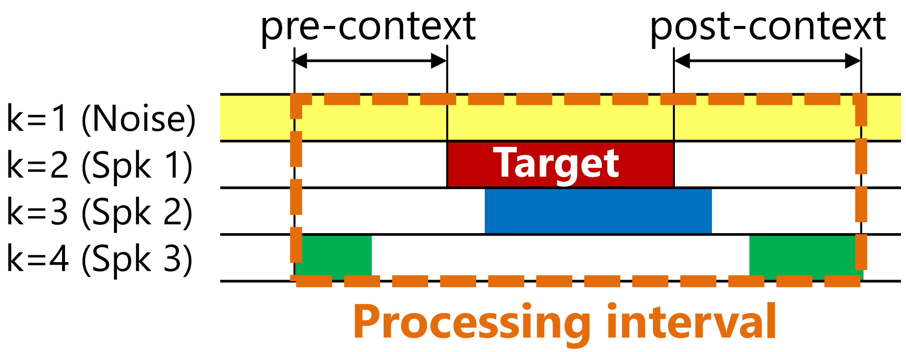
# of mics	CER
1	38.2
2	31.4
3	33.7
6	30.2
11	28.7
11 with oracle diarization	21.8
(Headset)	(19.2)

Summary of Chapter 7

- **Problem**
  - Diarization-guided speech enhancement (guided source separation [Boeddeker+’18]) is too slow for real-time applications
- **Solution**
  - Block-online algorithm of guided source separation
    - Core contribution: Real-time operation of guided source separation without performance degradation
    - Related publications: [SLT’21]

Chapter 7: Related Work

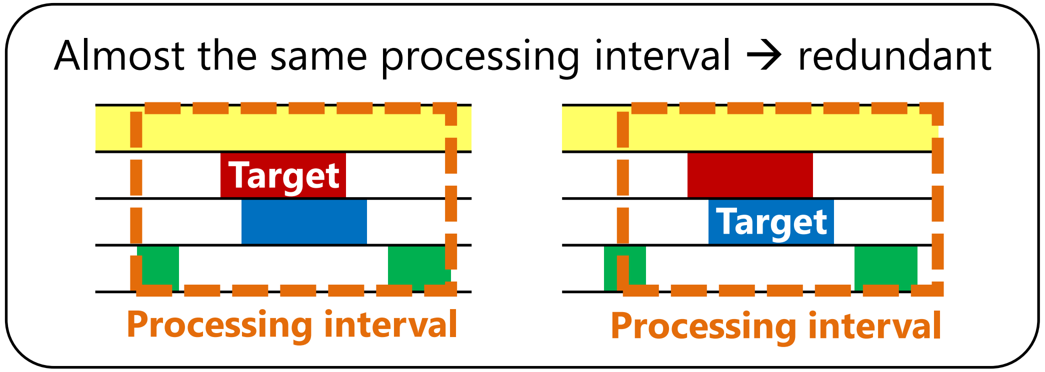
- **Guided source separation (GSS)**
  - Utterance-wise separation that utilize pre-context and post-context of the target utterance (~15 sec for each)
  - Use diarization results for conditioning in the separation step



✓ Perform well under unstable conditions (e.g., distributed microphones, moving speakers)

✘ High computational cost due to redundant calculation (85.44 hours to process 4.46 hours of CHiME-6 data)

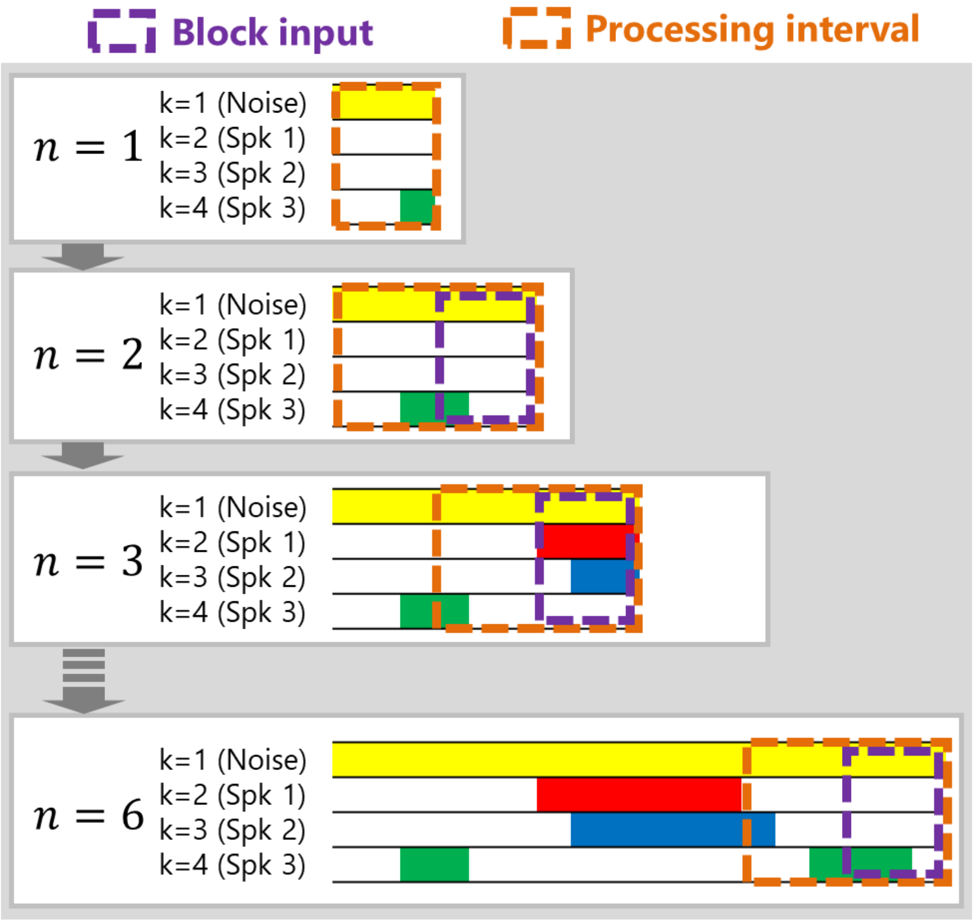
✘ Latency depending on utterance/post-context length


Almost the same processing interval → redundant

Chapter 7: Proposed Method

- **Proposed method: Block-online GSS**
  - Process block-wise inputs with their pre-context only

✓ Avoid redundancy of utterance-wise processing
✓ Latency depending on the block length



- **Experiments**
  - Dataset
    - Two sessions (S02 & S09) of the CHiME-6 dataset
      - S02: 8,902 sec
      - S09: 7,160 sec
  - Computational environment
    - Intel Xeon Gold 6132 CPU @ 2.60 GHz with 1 thread
  - Similar ASR performance in word error rate (WER)
  - 32x faster calculation, which is fast enough for real-time operation

	WER		Execution time (sec)	
Session	S02	S09	S02	S09
Offline GSS [Boeddeker+’18]	52.2	51.1	183529 ± 9567	124054 ± 7114
Block-online GSS	50.6	53.3	**6135 ± 93**	**3418 ± 66**

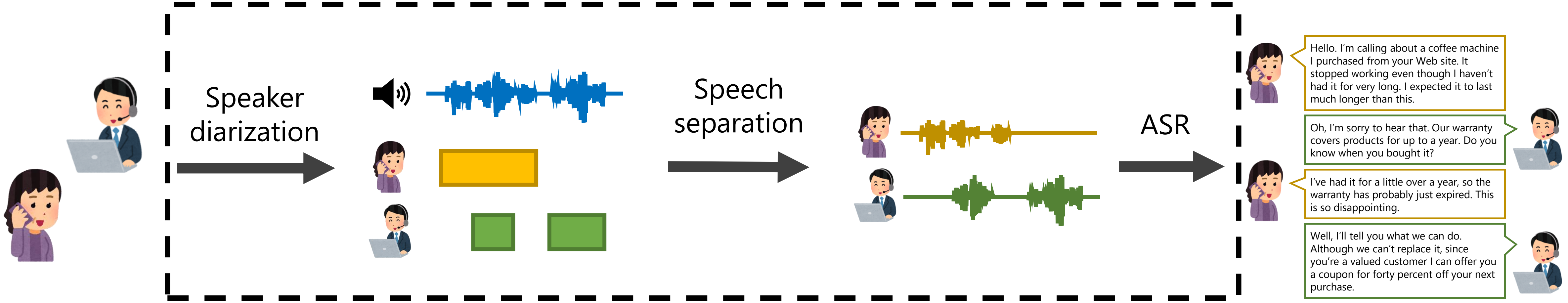
Conclusion

- **Part 1: Study on speaker diarization**
  - End-to-end speaker diarization for unknown numbers of speakers (Chapter 3)
    - **EEND-EDA:** A method of overlap-aware diarization of flexible numbers of speakers
    - **EEND-GLA:** A method of overlap-aware diarization of unlimited numbers of speakers
    - **Block-wise speaker-tracing buffer:** A method to enable online decoding of EEND-GLA
  - Multi-channel end-to-end speaker diarization (Chapter 4)
    - **Co-attention encoder:** An encoder that can treat any numbers of channels
    - **Mutual learning:** A training method to improve both single- and multi-channel diarization
  - End-to-end speaker diarization as post-processing (Chapter 5)
    - **EEND as post-processing:** A method to use EEND for overlap detection of cascaded approaches
- **Part 2: Study on applications of speaker diarization**
  - Speaker-diarization-driven meeting transcription (Chapter 6)
    - **Meeting transcription system based on distributed microphones**
  - Block-online speech separation conditioned on speaker diarization results (Chapter 7)
    - **Block-online guided source separation:** Fast and accurate speech separation method

Future Work

- Joint modeling of speaker diarization, speech separation, and ASR
  - Speaker diarization is informative for speech separation and ASR
  - Speech separation / ASR is also informative for speaker diarization [Xiao+'21] [India+'17]

One model / Co-training



Hello. I'm calling about a coffee machine I purchased from your Web site. It stopped working even though I haven't had it for very long. I expected it to last much longer than this.

Oh, I'm sorry to hear that. Our warranty covers products for up to a year. Do you know when you bought it?

I've had it for a little over a year, so the warranty has probably just expired. This is so disappointing.

Well, I'll tell you what we can do. Although we can't replace it, since you're a valued customer I can offer you a coupon for forty percent off your next purchase.

Related Publications

- **Journal articles**

[1] **Shota Horiguchi**, Shinji Watanabe, Paola Garcia, Yuki Takashima, and Yohei Kawaguchi, “Online neural diarization of unlimited numbers of speakers,” IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 31, pp. 704-720, 2023.

[2] **Shota Horiguchi**, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Paola Garcia, “Encoder-decoder based attractors for end-to-end neural diarization,” IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 30, pp. 1493-1507, 2022.

- **Peer-reviewed International Conference Paper**

[3] **Shota Horiguchi**, Yuki Takashima, Shinji Watanabe, and Paola Garcia, “Mutual learning of single- and multi-channel end-to-end neural diarization,” in SLT 2022, pp. 620-625.

[4] **Shota Horiguchi**, Yuki Takashima, Paola Garcia, Shinji Watanabe, and Yohei Kawaguchi, “Multi-channel end-to-end neural diarization with distributed microphones,” in ICASSP 2022, pp. 7332-7336.

[5] **Shota Horiguchi**, Paola Garcia, Shinji Watanabe, Yawen Xue, Yuki Takashima, and Yohei Kawaguchi, “Towards neural diarization for unlimited numbers of speakers using global and local attractors,” in ASRU 2021, pp. 98-105.

[6] **Shota Horiguchi**, Paola Garcia, Yusuke Fujita, Shinji Watanabe, and Kenji Nagamatsu, “End-to-end speaker diarization as post-processing,” in ICASSP 2021, pp. 7188-7192.

[7] **Shota Horiguchi**, Yusuke Fujita, and Kenji Nagamatsu, “Block-online guided source separation,” in SLT 2021, pp.236-242.

[8] **Shota Horiguchi**, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Kenji Nagamatsu, “End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors,” in INTERSPEECH 2020, 269-273.

[9] **Shota Horiguchi**, Yusuke Fujita, and Kenji Nagamatsu, “Utterance-wise meeting transcription system using asynchronous distributed microphones,” in INTERSPEECH 2020, pp. 344-348.

Other Publications (1st author)

- **Journal articles**
  - **Shota Horiguchi**, Daiki Ikami, and Kiyoharu Aizawa, "Significance of softmax-based features in comparison to distance metric learning-based features," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 5, pp. 1279-1285, 2020.
  - **Shota Horiguchi**, Sosuke Amano, Makoto Ogawa, and Kiyoharu Aizawa, "Personalized classifier for food image recognition," IEEE Transactions on Multimedia, vol. 20, no. 10, pp. 1497-1507, 2018.

- **Peer-reviewed International Conference Papers**
  - **Shota Horiguchi**, Naoyuki Kanda, and Kenji Nagamatsu, "Multimodal response obligation detection with unsupervised online domain adaptation," in INTERSPEECH, pp. 4180-4184, 2019.
  - **Shota Horiguchi**, Naoyuki Kanda, and Kenji Nagamatsu, "Face-voice matching using cross-modal embeddings," in ACMMM, pp. 1011-1019, 2018.
  - **Shota Horiguchi**, Kiyoharu Aizawa, and Makoto Ogawa, "The log-normal distribution of the size of objects in daily meal images and its application to the efficient reduction of object proposals," in ICIP, pp. 3668-3672, 2016.

Awards

- Itakura Prize Innovative Young Researcher Award, Acoustical Society of Japan, 2023
  - For the research on overlap-aware speaker diarization for unknown numbers of speakers

- 2nd prize in The Third DIHARD Speech Diarization Challenge (DIHARD III), 2021.
  - As the Hitachi-JHU team (lead author)

- 2nd prize in The 5th CHiME Speech Separation and Recognition Challenge (CHiME-5), 2018.
  - As the Hitachi-JHU team

- ITE Outstanding Research Presentation Award, 2015.

References (1/2)

- [Araki+’08] “A DOA based speaker diarization system for real meetings,” in Proc. HSCMA, 2008.
- [Boeddeker+’18] “Front-end processing for the CHiME-5 dinner party scenario,” in Proc. CHiME-5, 2018.
- [Bullock+] “Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection,” in Proc. ICASSP, 2020.
- [Chen+’17] “Deep attractor network for single-microphone speaker separation,” in Proc. ICASSP, 2017.
- [Chen+’20] “Continuous speech separation: Dataset and analysis,” in Proc. ICASSP, 2020.
- [Diez+’18] “Speaker diarization based on Bayesian HMM with eigenvoice priors,” in Proc. Odyssey, 2018.
- [Fujita+’19] “End-to-end neural speaker diarization with self-attention,” in Proc. ASRU, 2019.
- [Garcia+’20] “Speaker detection in the wild: Lessons learned from JSALT 2019,” in Proc. Odyssey, 2020.
- [Han+’21] “BW-EDA-EEND: Streaming end-to-end neural speaker diarization for a variable number of speakers,” in Proc. ICASSP, 2021.
- [India+’17] “LSTM neural network-based speaker segmentation using acoustic and language modeling,” in Proc. INTERSPEECH, 2017.
- [Ishiguro+’11] “Probabilistic speaker diarization with bag-of-words representations of speaker angle information,” IEEE TASLP, vol. 20, no. 2, pp. 447-460, 2011.
- [Ito+’16] “Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing,” in Proc. EUSIPCO, 2016.
- [Kanda+’17] “Investigation of lattice-free maximum mutual information-based acoustic models with sequence-level Kullback-Leibler divergence,” in Proc. ASRU, 2017.
- [Kanda+’18] “Lattice-free state-level minimum Bayes risk training of acoustic models,” in Proc. INTERSPEECH, 2018.

References (2/2)

- [Kinoshita+’18] “Listening to each speaker one by one with recurrent selective hearing networks,” in Proc. ICASSP, 2018.
- [Kinoshita+’21] “Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech,” in INTERSPEECH, 2021.
- [Landini+’20] “BUT system for the second DIHARD speech diarization challenge,” in Proc. ICASSP, 2020.
- [Landini+’22] “Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks,” Computer Speech and Language, vol. 71, p. 101254, 2022.
- [Maekawa’03] “Corpus of Spontaneous Japanese: Its design and evaluation,” in Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing Recognition, 2003.
- [Ryant+’19] “The second DIHARD diarization challenge: Dataset, task, and baselines,” in Proc. INTERSPEECH, 2019.
- [Ryant+’21] “The third DIHARD diarization challenge,” in Proc. INTERSPEECH, 2021.
- [Sell+’18] “Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge,” in Proc. INTERSPEECH, 2019.
- [Wang+’20] “Neural speech separation using spatially distributed microphones,” in Proc. INTERSPEECH, 2020.
- [Watanabe’17] “Student-teacher network learning with enhanced features,” in Proc. ICASSP, 2017.
- [Watanabe’20] “CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings,” in Proc. CHiME-6, 2020.
- [Xiao+’21] “Microsoft speaker diarization system for the VoxCeleb speaker recognition challenge 2020,” in Proc. ICASSP, 2021.
- [Xue+’21] “Online streaming end-to-end neural diarization handling overlapping speech and flexible numbers of speakers,” in Proc. INTERSPEECH, 2021.