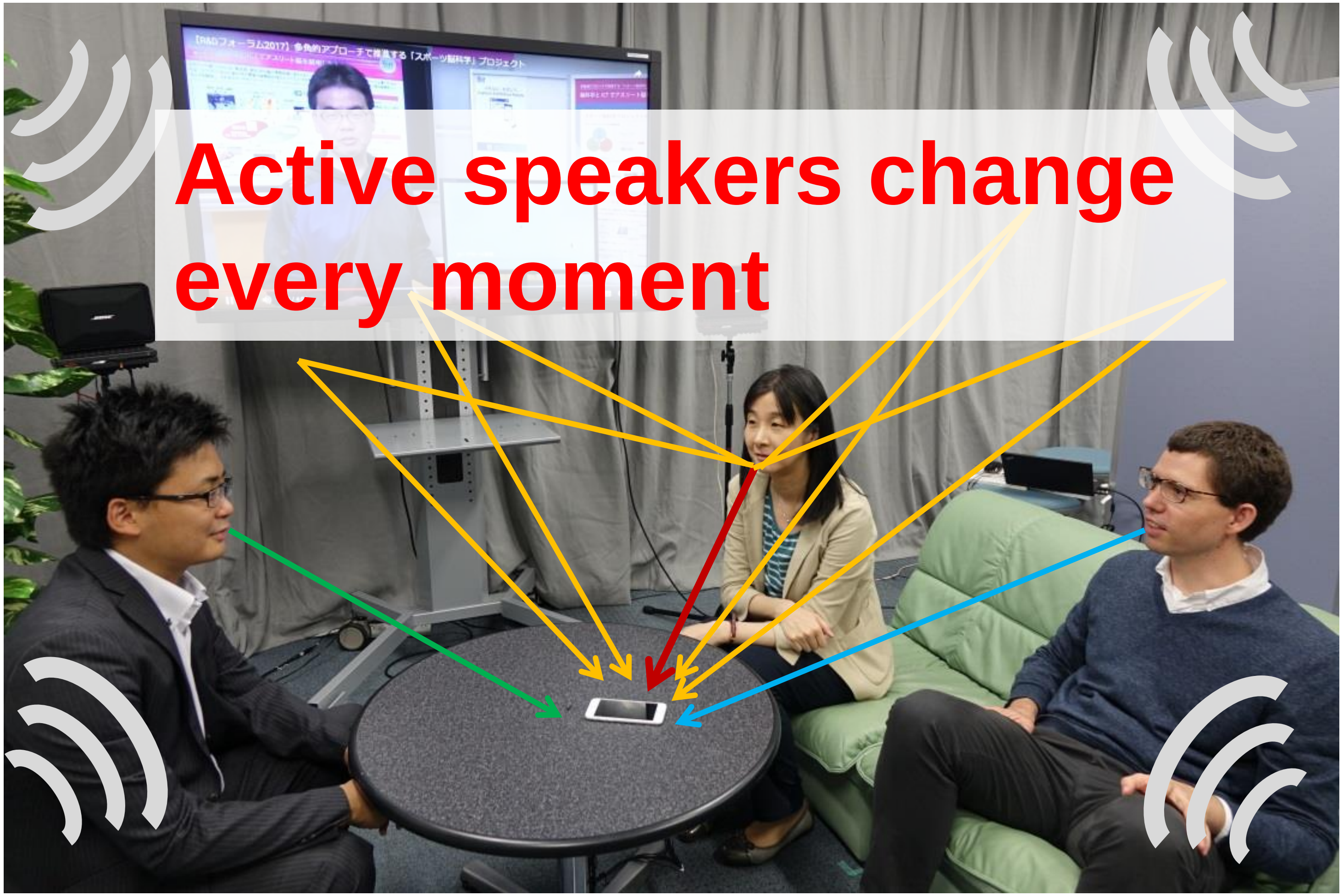
Part V.
Meeting Analysis

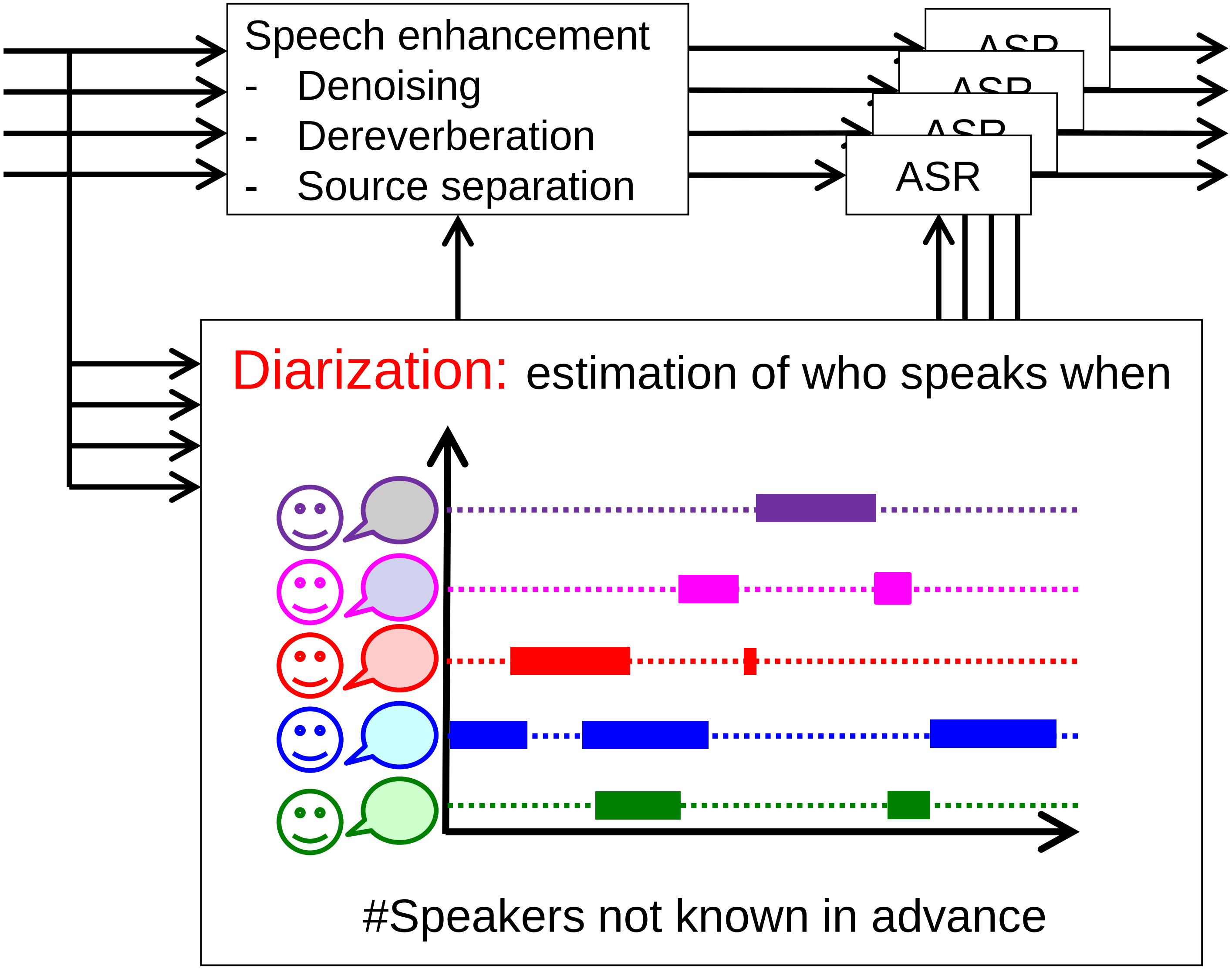
Tomohiro Nakatani

Speech recording in meeting situation



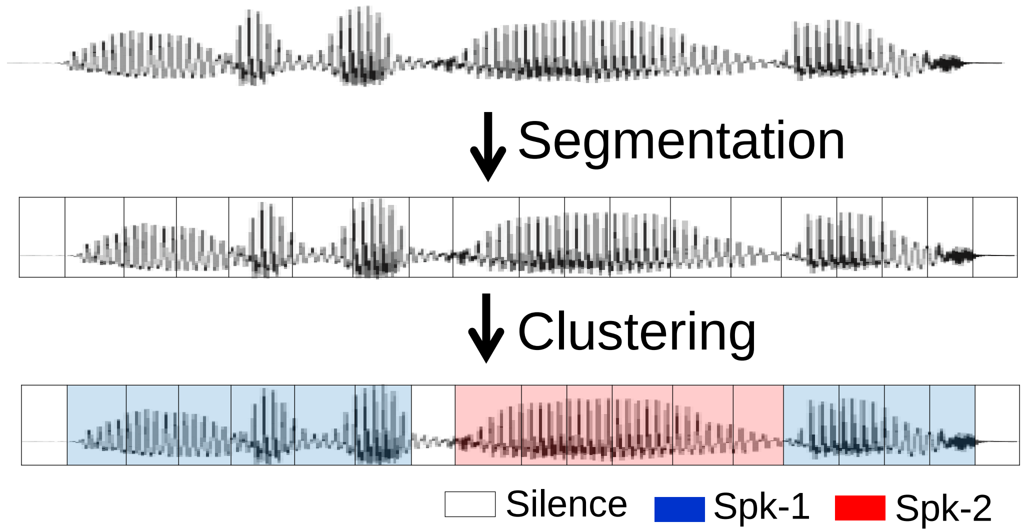
- Estimation of who speaks when (=**diarization**) is crucial for speech enhancement and ASR

Problems in meeting analysis

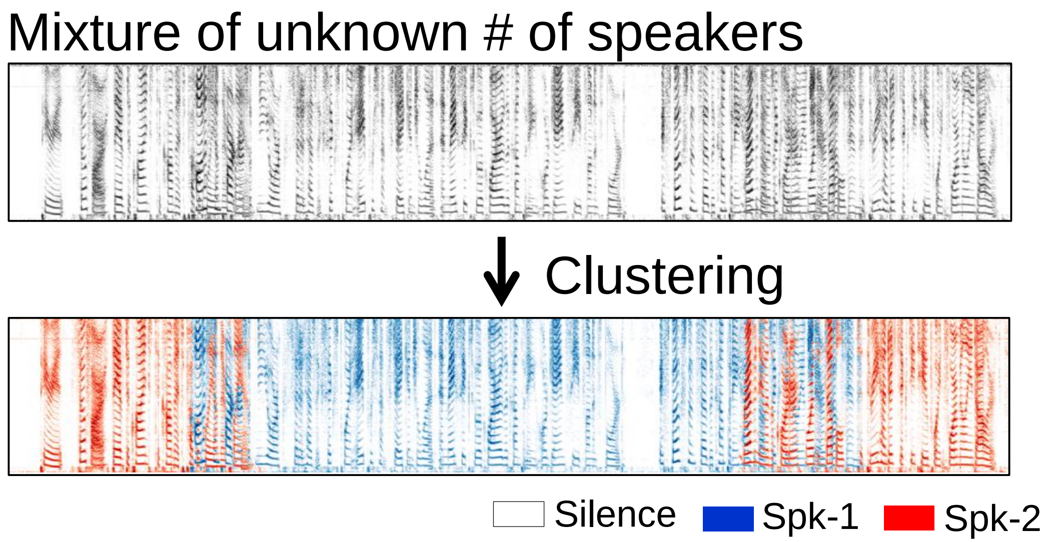


Two approaches to diarization

- Clustering of time segments
  - Based on spectral features
    - MFCC, i-vector, d-vector, x-vector, etc.
  - Speaker overlapping segments are disregarded
  - 1-ch processing

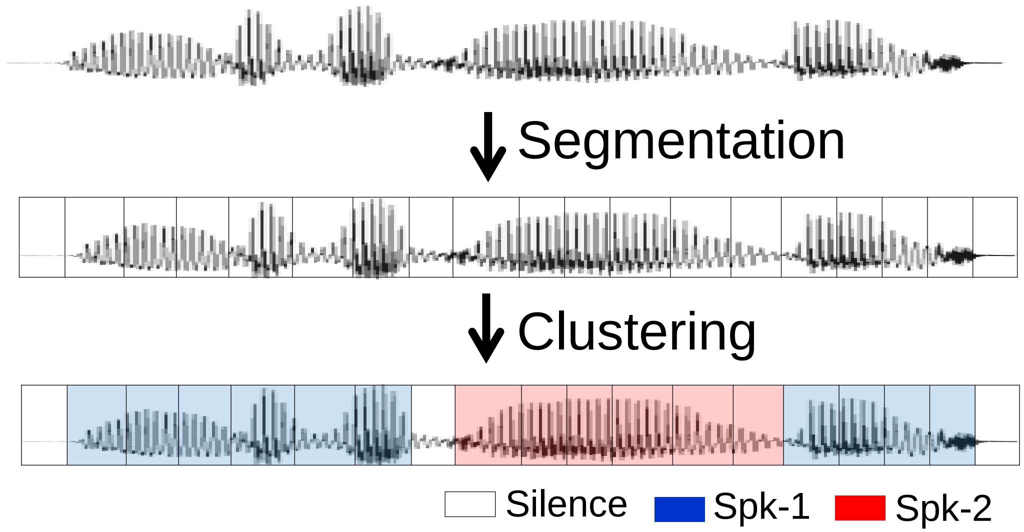


- Clustering of TF points
  - Mask-based source separation for unknown #sources
  - Speaker overlapping segments can be separated
  - 1-ch/multi-ch processings



Approaches to diarization

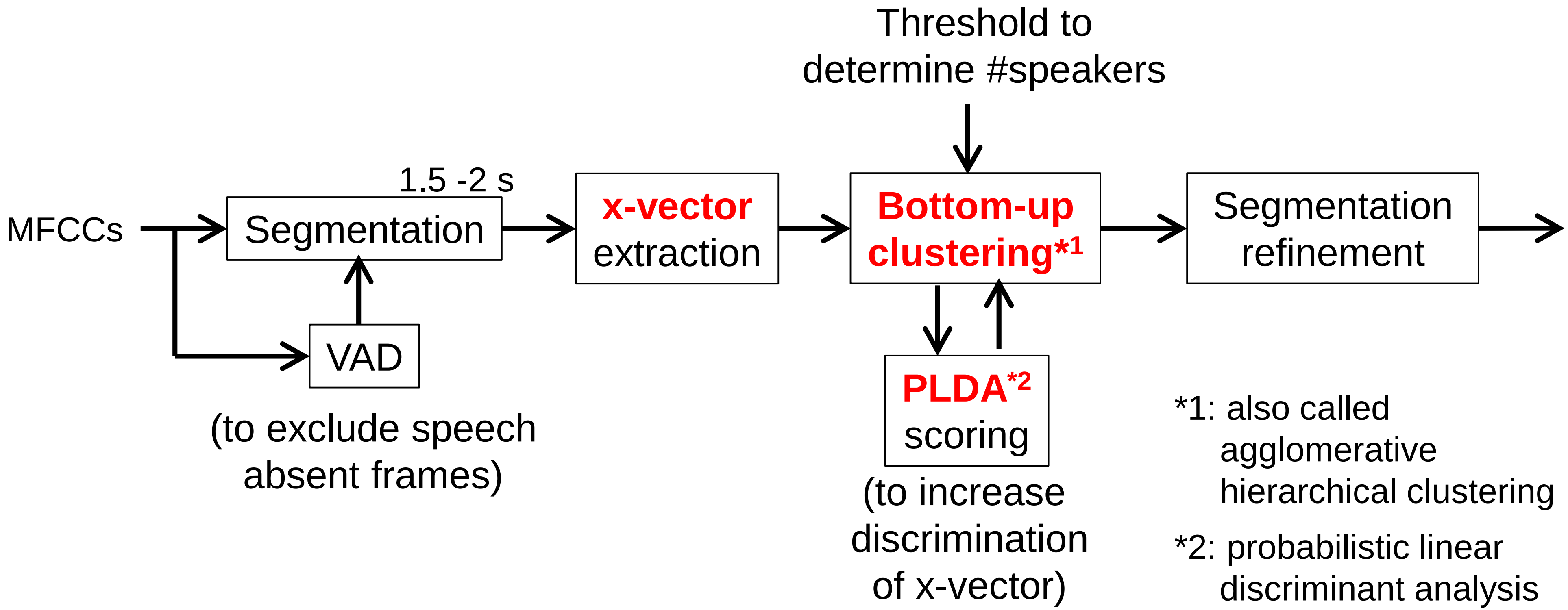
- Clustering of time segments
  - Based on spectral features
    - MFCC, i-vector, d-vector, x-vector, etc.
  - Speaker overlapping segments are disregarded
  - 1-ch processing



- Clustering of TF points
  - Mask-based source separation for unknown #sources
  - Speaker overlapping segments can be separated
  - 1-ch/multi-ch processings

JHU DIHARD challenge system [Sell et al., 2018]

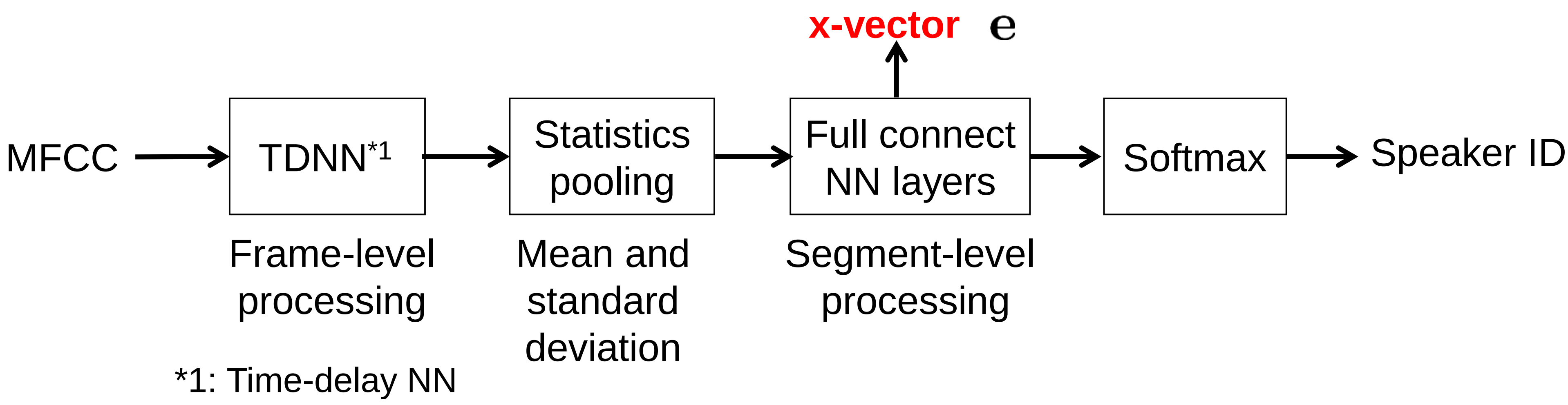
- Best score at Track 1 of DIHARD-I challenge
  - DIHARD-I,II: diarization challenges with HARD corpora [Ryant et al., 2019]



Robust speaker feature extraction and scoring are crucial

x-vector [Snyder et al., 2018]

- A bottleneck feature of speaker verification NN
  - Trained using data augmentation (noise, reverb)



A speaker characteristic essential for speaker verification

PLDA [Silovsky et al., 2011]

- Decompose an x-vector into different factors

$$\mathbf{e} = \underbrace{\mathbf{m}}_{\text{Speaker independent mean}} + \underbrace{\mathbf{F}\mathbf{h}_i}_{\text{Speaker inherent feature}} + \underbrace{\mathbf{G}\mathbf{w}_{i,j}}_{\text{Utterance dependent feature}} + \underbrace{\mathbf{n}_{i,j}}_{\text{noise}}$$

Speaker independent mean | **Speaker inherent feature** | Utterance dependent feature | noise

i: Speaker index
j: Utterance index

$\mathbf{m}, \mathbf{F}, \mathbf{G}$ and Σ : Model parameters determined in advance using training data

$$p(\mathbf{e} \mid \mathbf{h}_i, \mathbf{w}_{i,j}; \theta) = \mathcal{N}(\mathbf{m} + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{i,j}, \Sigma)$$

Cluster likelihood : $p(\mathbf{e}_1, \ldots, \mathbf{e}_J) = \mathcal{N}(\mathbf{m}', \mathbf{A}\mathbf{A}^\top + \Sigma')$

where $\mathbf{m}' = (\mathbf{m}, \ldots, \mathbf{m})^\top$ $\mathbf{A} = \begin{pmatrix} \mathbf{F} & \mathbf{G} & 0 & \ldots & 0 \\ \mathbf{F} & 0 & \mathbf{G} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \mathbf{F} & 0 & 0 & \ldots & \mathbf{G} \end{pmatrix}$ $\Sigma' = \begin{pmatrix} \Sigma & 0 & \ldots & 0 \\ 0 & \Sigma & \ldots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \ldots & \Sigma \end{pmatrix}$

Diarization can be performed with speaker inherent features

Evaluation metric for diarization

- Diarization error rates (DER) [NIST speech group, 2007]

$$\mathrm{DER} = \frac{\#\text{frames with wrongly estimated speaker}}{\text{total } \#\text{frames}}$$

  – Includes: missed speaker time (MST), false active time (FAT), and speaker error time (SET)

DERs with DIHARD-I challenge [Sell et al., 2018]

Dataset includes: clinical interviews, child language acquisition recordings, YouTube videos, speech in restaurants

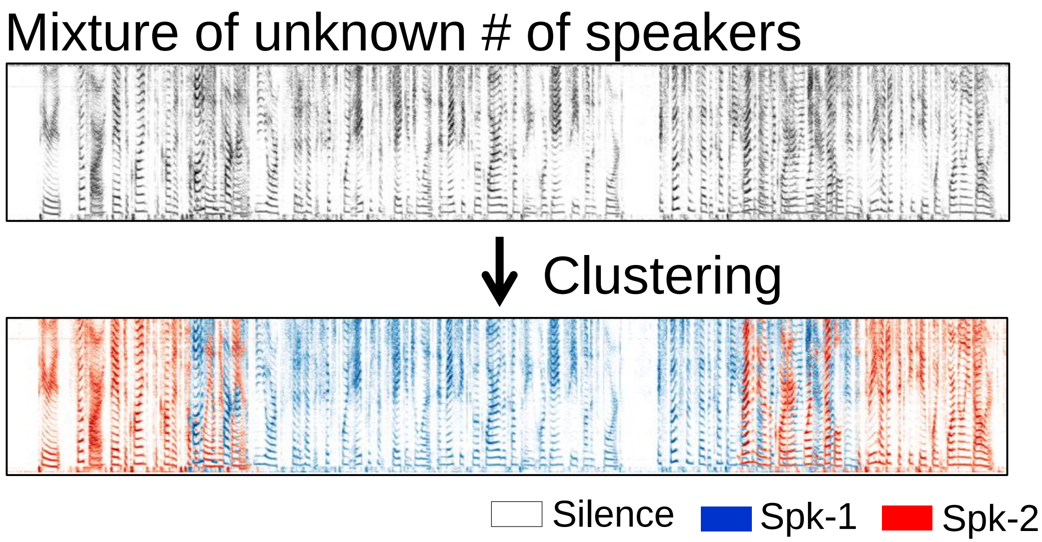
Track1: w/ oracle speech segmentation (Challenge top for Eval: 23.73 %)
Track2: w/o oracle speech segmentation (Challenge top for Eval: 35.51 %)

	Track1	Track2
All same speaker	39.01 %	55.93 %
i-vector + PLDA	28.06 %	40.42 %
x-vector + PLDA	25.94 %	39.43 %
x-vector + PLDA, with seg. refinement	**23.73 %**	**37.29 %**

Approaches to diarization

- Clustering of time segments
  - Based on spectral features
    - MFCC, i-vector, d-vector, x-vector, etc.
  - Speaker overlapping segments are disregarded
  - 1-ch processing

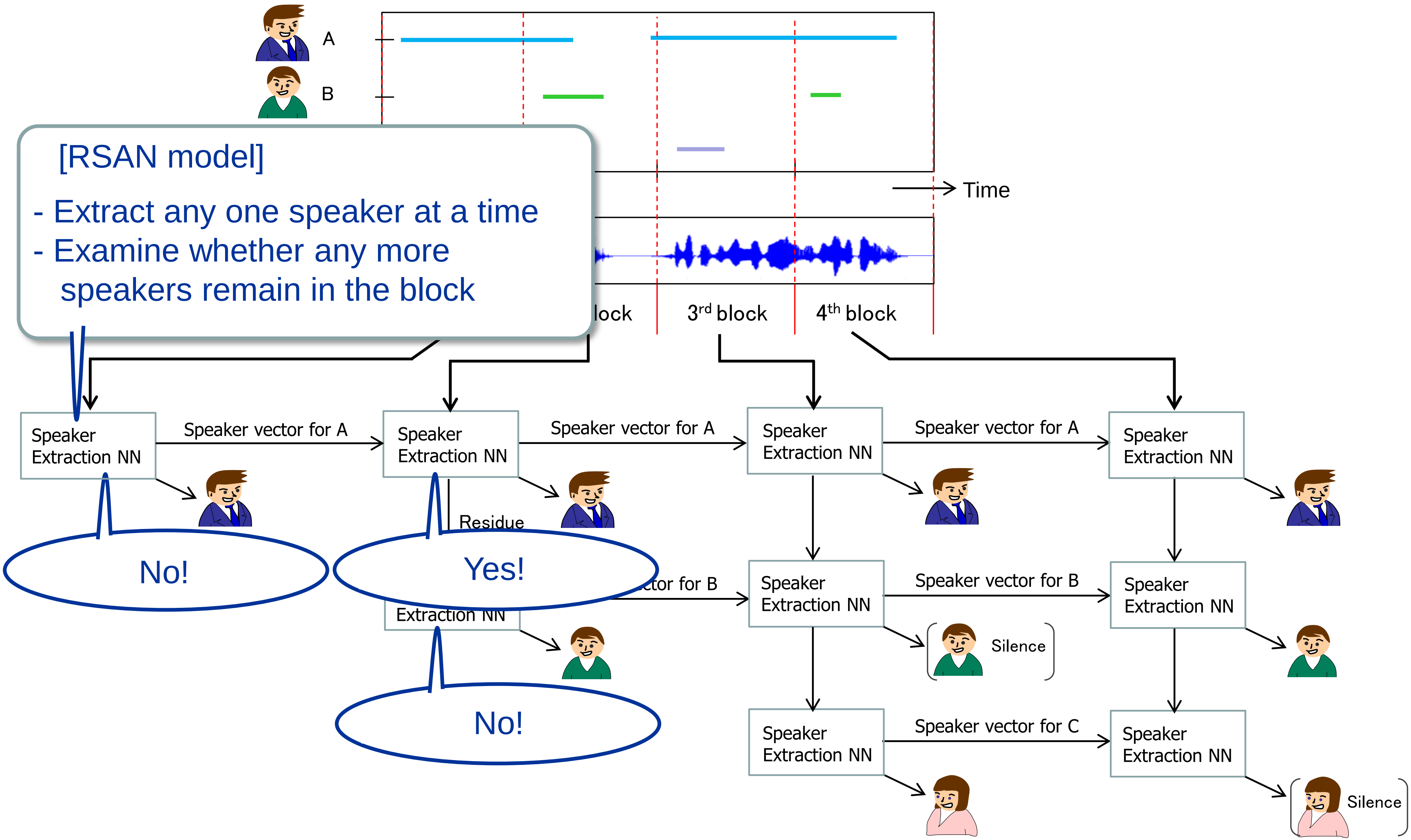
- Clustering of TF points
  - Mask-based source separation for unknown #sources
  - Speaker overlapping segments can be separated
  - 1-ch/multi-ch processings



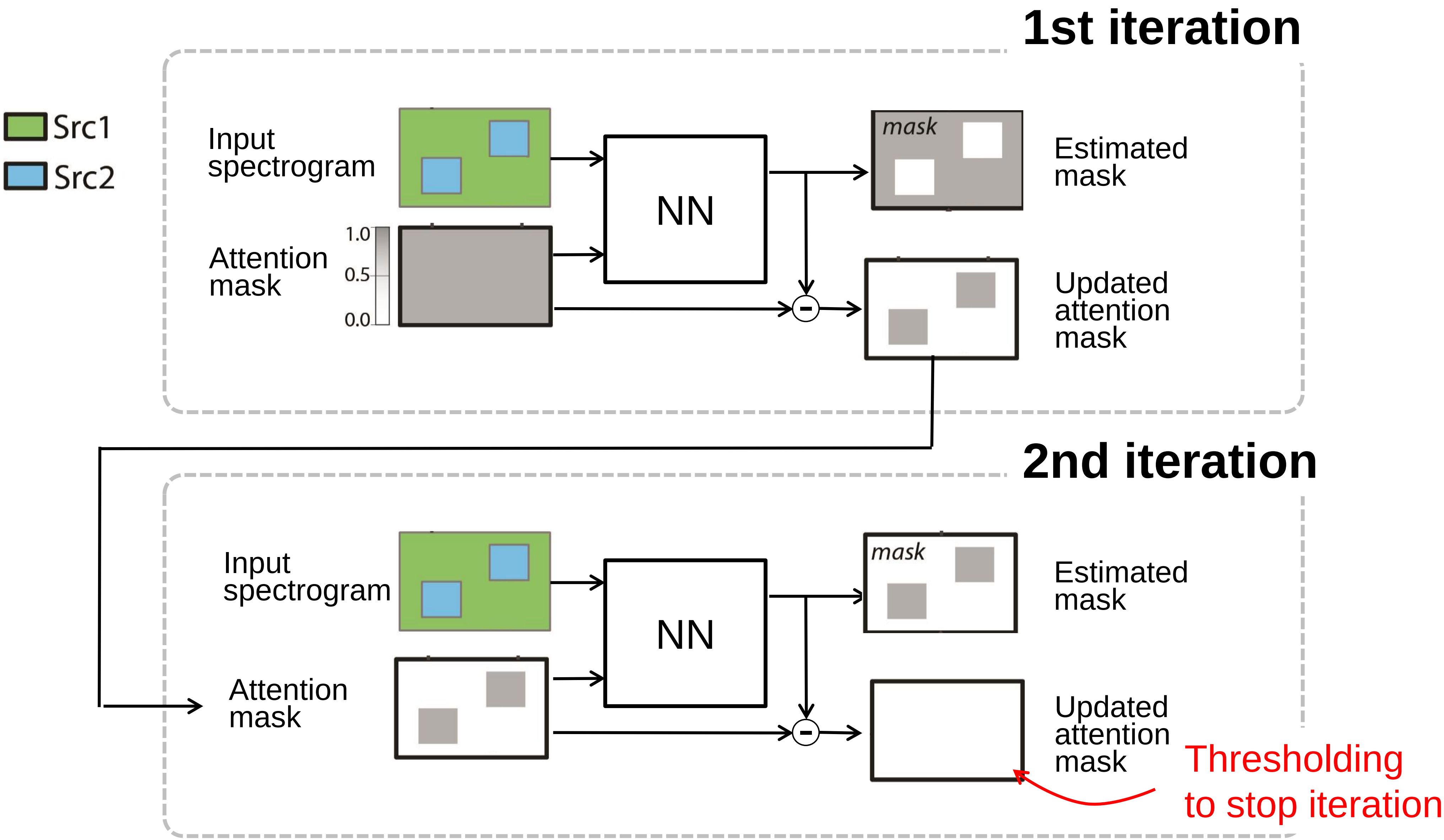
Recurrent Selective Attention Network (RSAN) [Kinoshita et al., 2018, von Neumann et al., 2019]

- Neural network based mask estimator for unknown #speakers
- Perform block online meeting analysis
  - By dynamically assigning a NN to extract a source every time it detects a new source,
- Can be optimized in an end-to-end manner for feature extraction, source counting, diarization, and source separation

Overall online processing flow by RSAN



How to control #iterations at each block



Training of RSAN : loss function

$$\mathcal{L} = \mathcal{L}^{\mathrm{Sep}} + \alpha \mathcal{L}^{\mathrm{Count}}$$

Loss for separation

$$\mathcal{L}^{\mathrm{Sep}} = \sum_i \|\hat{\mathbf{Y}}_i - \mathbf{Y}_i^{ref}\|_2^2$$

$\hat{\mathbf{Y}}_i, \mathbf{Y}_i^{ref}$: Estimated and clean speech spectra

Loss for source counting

$$\mathcal{L}^{\mathrm{Count}} = \max(\mathbf{R}, 0)$$

$$\mathbf{R} = \mathbf{1} - \sum_{\mathbf{i}} \mathbf{M}^{(\mathbf{i})}$$

: Attention mask after masks for all the sources are extracted

Source separation, counting, feature extraction, and diarization are jointly optimized in an end-to-end processing manner

Preliminary results with simulated conversation

Test data:
- Simulated conversation composed of utterances (WSJ)
- Average conversation length: 30 s

	DER	SCER	DER+ SCER
All same speaker	38.8 %	27.4 %	66.2 %
Bottom up clustering of RSAN speaker vectors (batch)	15.8 %	6.2 %	22.0 %
PIT based mask estimation (batch)	9.8 %	**4.4 %**	14.2 %
RSAN (online)	**6.6 %**	4.9 %	**11.5 %**

Speaker confusion error rate (SCER): [von Neumann et al., 2019]

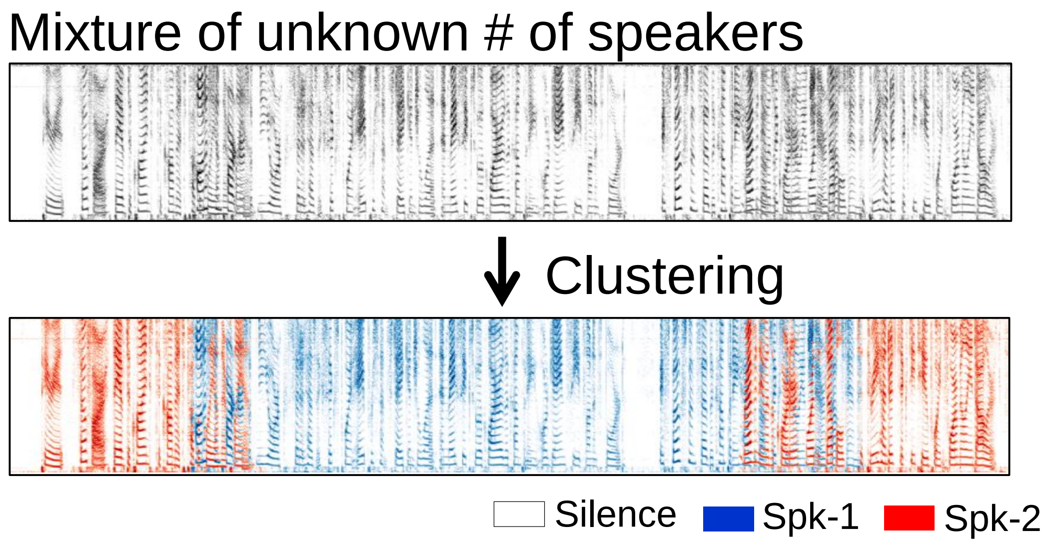
$$\mathrm{SCER} = \frac{\#\text{frames with confused speaker assignments}}{\text{total }\#\text{frames}}$$

- Confused assignments: speakers correctly detected but assigned to wrong clusters
- SCER is not counted by DER, and DER+SCER accounts for total errors

Approaches to diarization

- Clustering of time segments
  - Based on spectral features
    - MFCC, i-vector, d-vector, x-vector, etc.
  - Speaker overlapping segments are disregarded
  - 1-ch processing

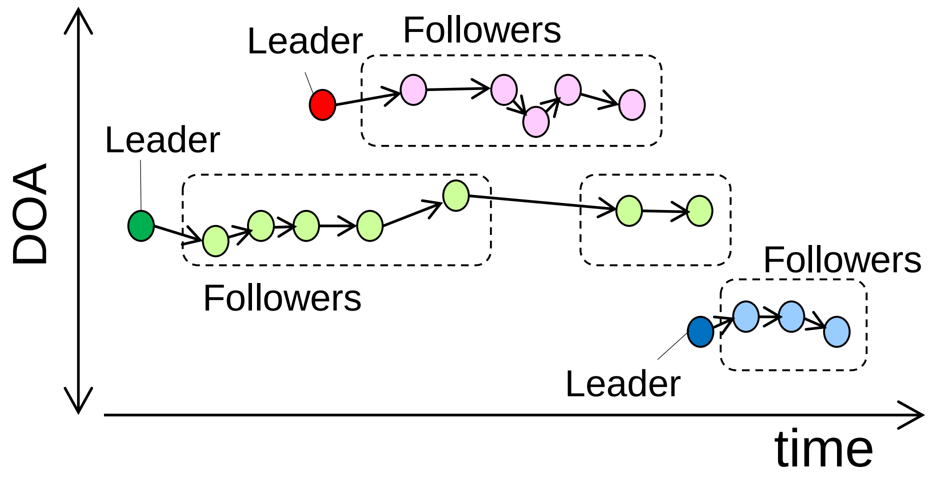
- Clustering of TF points
  - Mask-based source separation for unknown #sources
  - Speaker overlapping segments can be separated
  - 1-ch/multi-ch processings



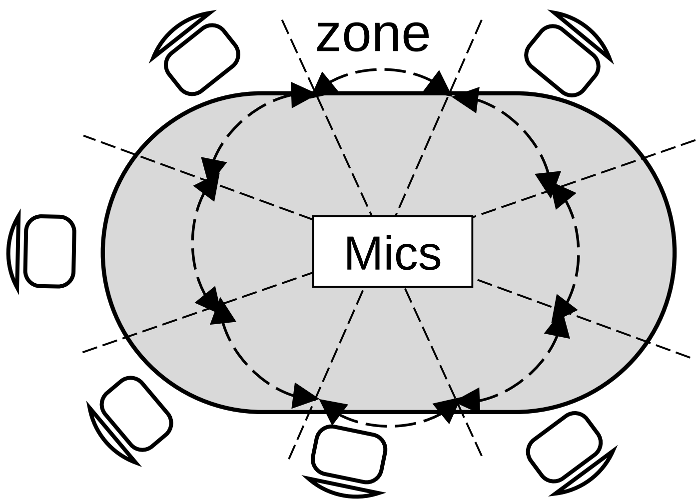
Clustering of TF bins (Multi-ch)

- Features for localization
  - DOAs, and many variants
- Online processing works
  - Multi-target tracking problem
    - Leader-follower clustering [Hori et al., 2012]
    - Probabilistic hypothesis density filter with random finite set [Evers and Naylor, 2018]
  - Zone-based speaker diarization [Fallon and Godsill, 2011, Ito et al., 2017]
    - Divides possible speaker locations into pre-determined zones
    - VAD at each zone results in diarization

Leader-follower clustering



Zones for speaker diarization



Probabilistic spatial dictionary based diarization [Ito et al., 2017]

- Model of signal from each possible speaker location
  - Complex Watson distribution

$$p(\tilde{\mathbf{y}}_{tf}^{(k)}) = \mathcal{W}(\tilde{\mathbf{y}}_{tf}^{(k)}; \kappa_f^{(k)}, \mathbf{a}_f^{(k)})$$

$\mathbf{a}_f^{(k)}$: parameter for RIR (dictionary, pretrained)

$\kappa_f^{(k)}$: parameter for variance (dictionary, pretrained)

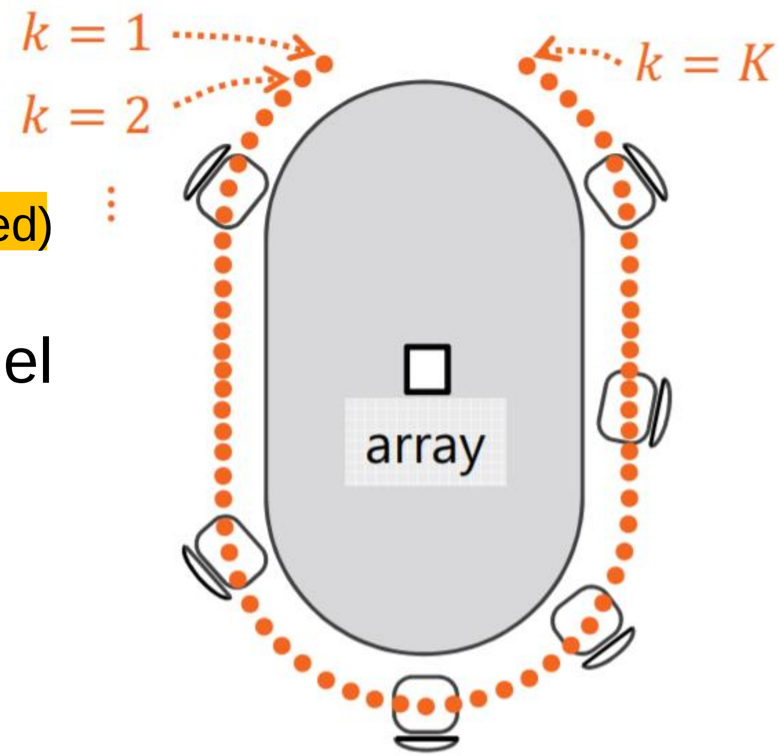
- Model of meeting recording: mixture model

$$p(\tilde{\mathbf{y}}_{tf}) = \sum_{k=1}^{K} \alpha_t^{(k)} \mathcal{W}(\tilde{\mathbf{y}}_{tf}; \kappa_f^{(k)}, \mathbf{a}_f^{(k)})$$

$\alpha_t^{(k)}$: mixture weight (estimated from test data) which indicates active speaker locations

Useful for online diarization

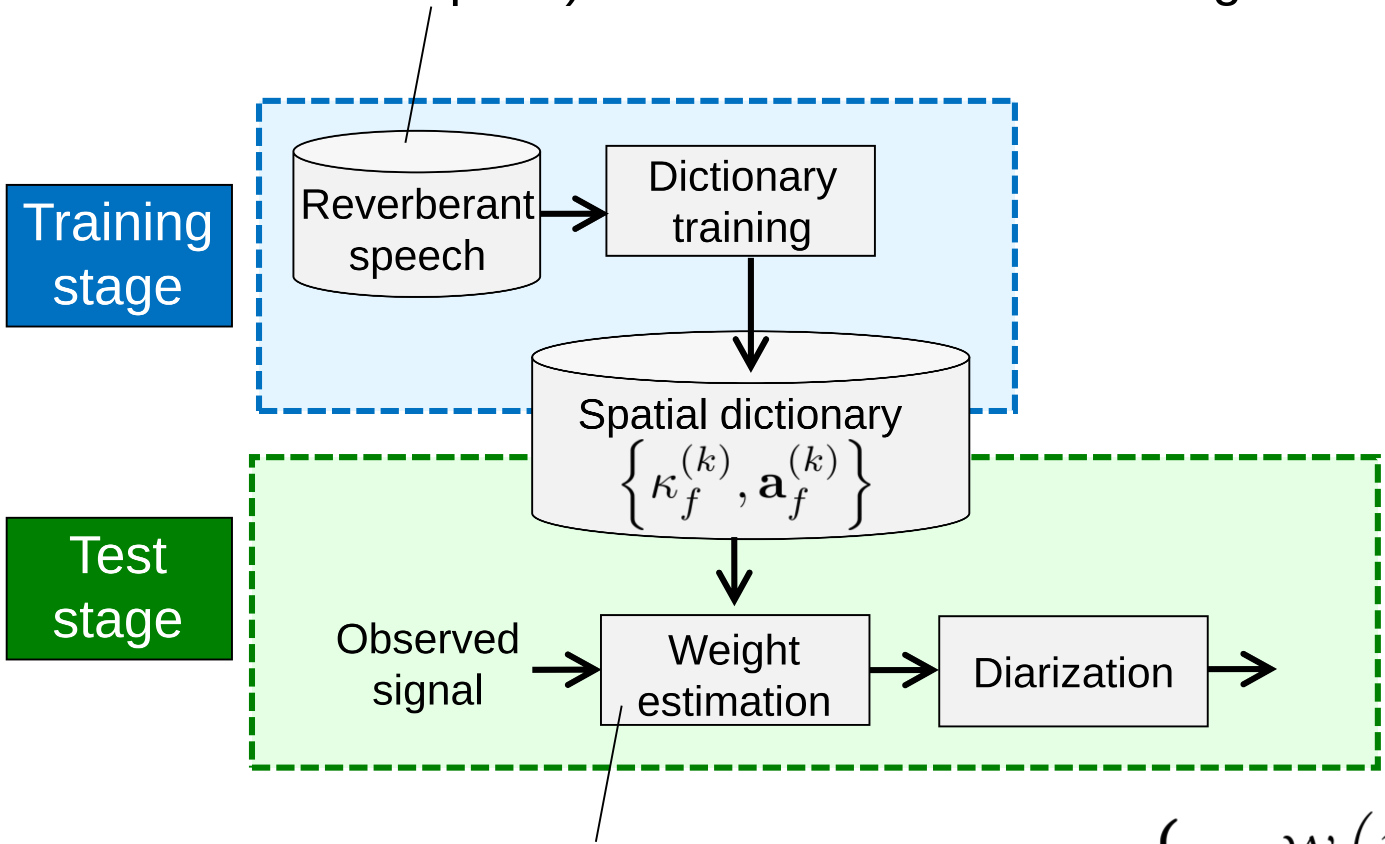
Recording condition



k : possible speaker location

Processing diagram of probabilistic spatial dictionary based diarization

Simulated microphone signals (with a plain wave assumption) can be used for the training



Posterior of source location: $$\alpha_t^{(k)} = \sum_f \left\{ \frac{\mathcal{W}\left(\tilde{\mathbf{y}}_{tf}; \kappa_f^{(k)}, \mathbf{a}_f^{(k)}\right)}{\sum_{k'=1}^{K} \mathcal{W}\left(\tilde{\mathbf{y}}_{tf}; \kappa_f^{(k')}, \mathbf{a}_f^{(k')}\right)} \right\}$$

DERs under reverberant babble noise condition

Reverberation time: 500 ms
Length of meeting: 15-20 min
SNR: 3-15 dB

#mics: 8
K=65
Information on chair locations is given

Session ID	#Speakers	Noise level (babble noise)	DER	
			Leader-follower clustering [Hori 2012]	Probabilistic spatial dictionary
1	6	No noise	46.8 %	**9.3 %**
2			64.6 %	**12.2 %**
3	5	Low	23.8 %	**17.2 %**
4			47.5 %	**18.9 %**
5	6		62.6 %	**15.6 %**
6	4	High	70.9 %	**27.7 %**
7			73.6 %	**24.8 %**
8	6		67.2 %	**18.9 %**

Discussion

- 1-ch processing
  - Use of neural network is a key to successful diarization
    - End-to-end neural processing is also investigated
  - Treatment of adverse noise conditions is still a challenging problem

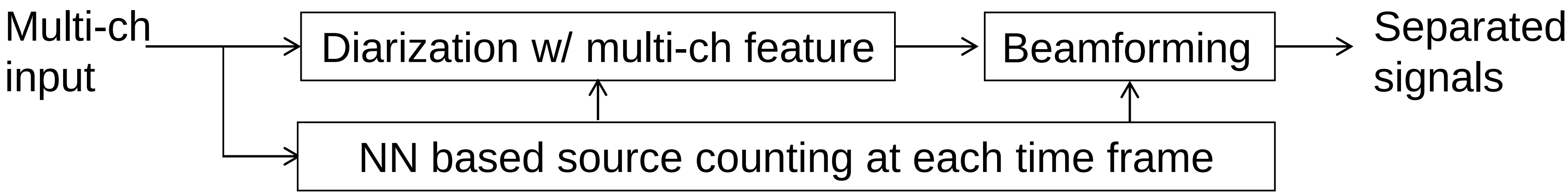
- Multi-ch processing
  - Spatial features work effectively even under noisy reverberant envs
  - Hard to track speakers who move with no utterance

Integration of 1-ch and multi-ch approaches should be explored

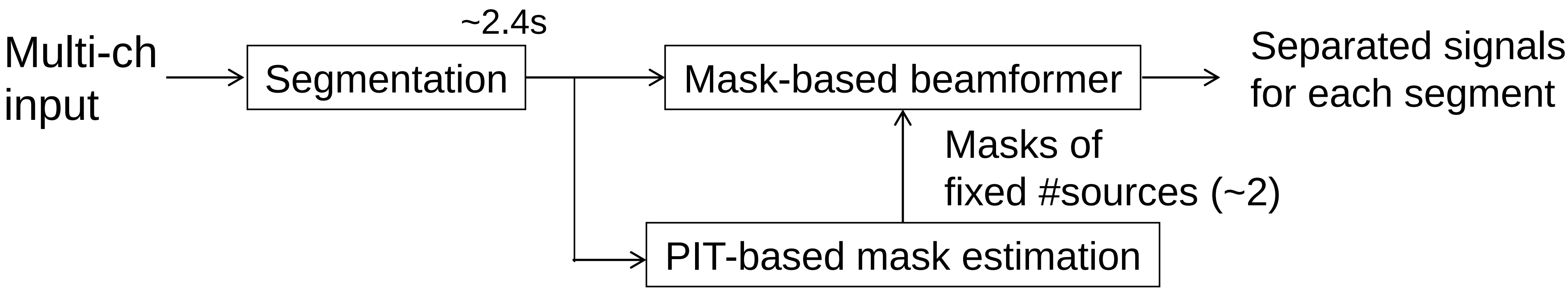
- only a few attempts made so far

Meeting analysis based on source separation with integration of NN and microphone array

- NN-based source counting is combined with beamforming [Chazan et al., 2018]



- Segment-wise separation of fixed #sources based on NN and beamforming [Yoshioka et al., 2018]
  - Applicable without performing source counting or diarization



Software

- JHU diarization system (DIHARD-II challenge baseline)
  - https://github.com/iiscleap/DIHARD_2019_baseline_alltracks
  - Based on JHU diarization system developed for the DIHARD-I challenge, and prepared for the DIHARD-II challenge by Ganapathy et al.
  - Segmentation refinement block is omitted

Table of contents