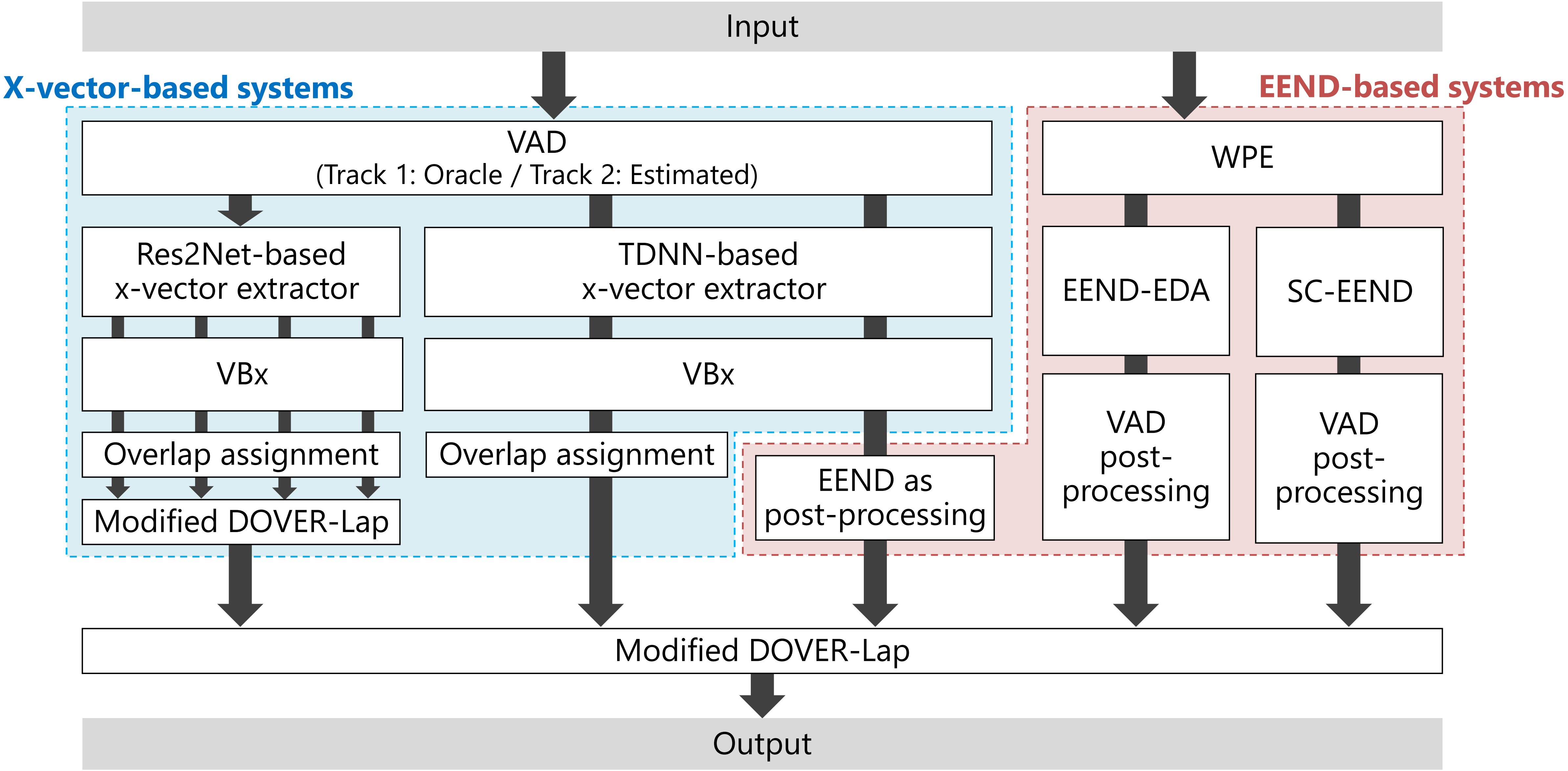
Hitachi-JHU System
for the Third DIHARD Speech Diarization Challenge

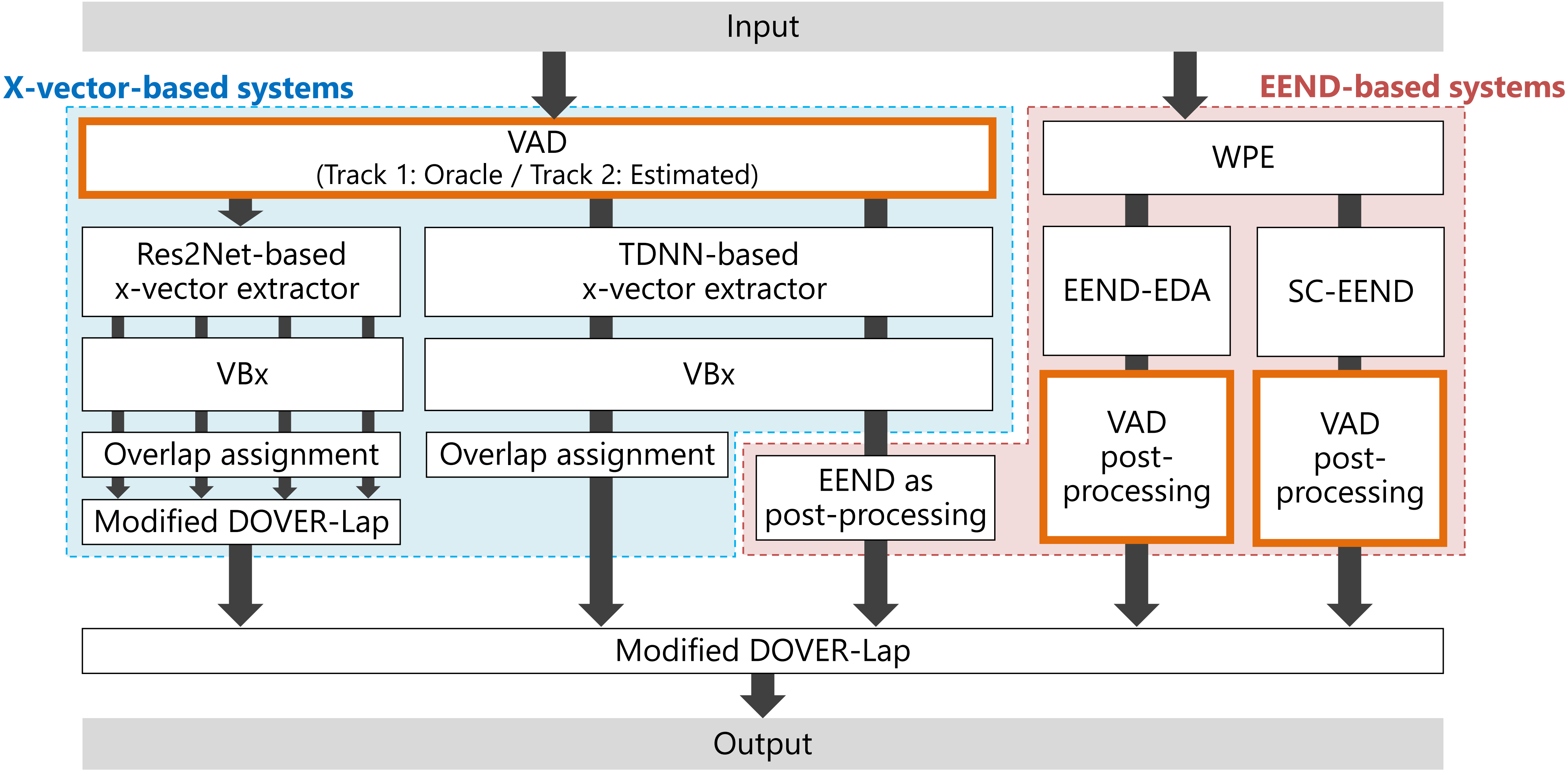
Shota Horiguchi[1*] Nelson Yalta[1*] Paola Garcia[2] Yuki Takashima[1] Yawen Xue[1]
Desh Raj[2] Zili Huang[2] Yusuke Fujita[1] Shinji Watanabe[2] Sanjeev Khudanpur[2]

1 HITACHI Inspire the Next
2 JOHNS HOPKINS UNIVERSITY

Overview of Hitachi-JHU System



VAD



Input
X-vector-based systems
EEND-based systems
VAD
(Track 1: Oracle / Track 2: Estimated)
WPE
Res2Net-based
x-vector extractor
TDNN-based
x-vector extractor
EEND-EDA
SC-EEND
VBx
VBx
VAD
post-
processing
VAD
post-
processing
Overlap assignment
Overlap assignment
EEND as
post-processing
Modified DOVER-Lap
Modified DOVER-Lap
Output

VAD

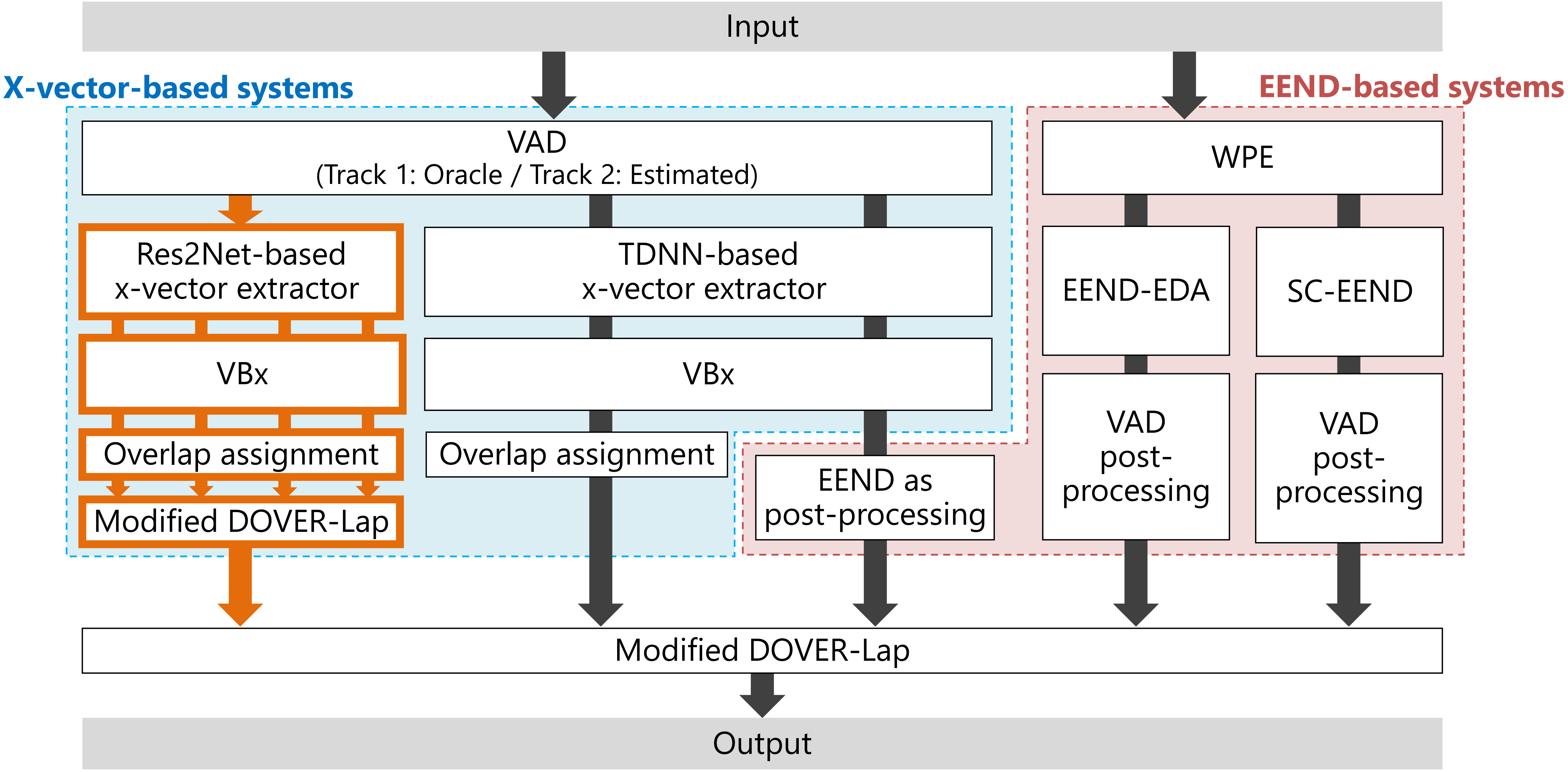
HITACHI
Inspire the Next

- Method: Posterior average of two models
  - **SincNet-based VAD** [Lavechin+, INTERSPEECH'20]
    - SincNet followed by BiLSTM layers and a fully-connected layer
    - Trained on DIHARD III DEV for 300 epochs
  - **TDNN-based VAD**
    - Five-layer TDNN using statistics pooling for long-context
    - Trained on DIHARD III DEV for 10 epochs
      with data augmentation using MUSAN corpus and simulated RIRs

- Results on DIHARD III DEV

Method	False alarm (%)	Missed (%)
SincNet-based VAD	2.78	2.51
TDNN-based VAD	2.85	2.80
Posterior average	2.58	2.55

(1) Res2Net-Based System



(1) Res2Net-Based System

HITACHI
Inspire the Next

- X-vector extractors trained on VoxCeleb

Model	# of layers	Normalization	Compression	SpecAugment
Res2Net-BN	23	BatchNorm	$\ln x$	
Res2Net-UN	23	UtteranceNorm	$\log_{10} x$	
Res2Net-BN-Large	50	BatchNorm	$\ln x$	
Res2Net-UN-Large	50	UtteranceNorm	$\log_{10} x$	✓

- VBx clustering
  - Initial clustering using AHC with PLDA, the interpolation of VoxCeleb PLDA and DIHARD III PLDA
  - Then, Bayesian HMM clustering with the LDA
- Overlap assignment
  - The same model as SincNet-based was trained to detect overlap using DIHARD III DEV
  - Assigned the closest other speaker in the time axis for each detected frame
- Modified DOVER-Lap to combine the results from the four models

(1) Res2Net-Based System: Results

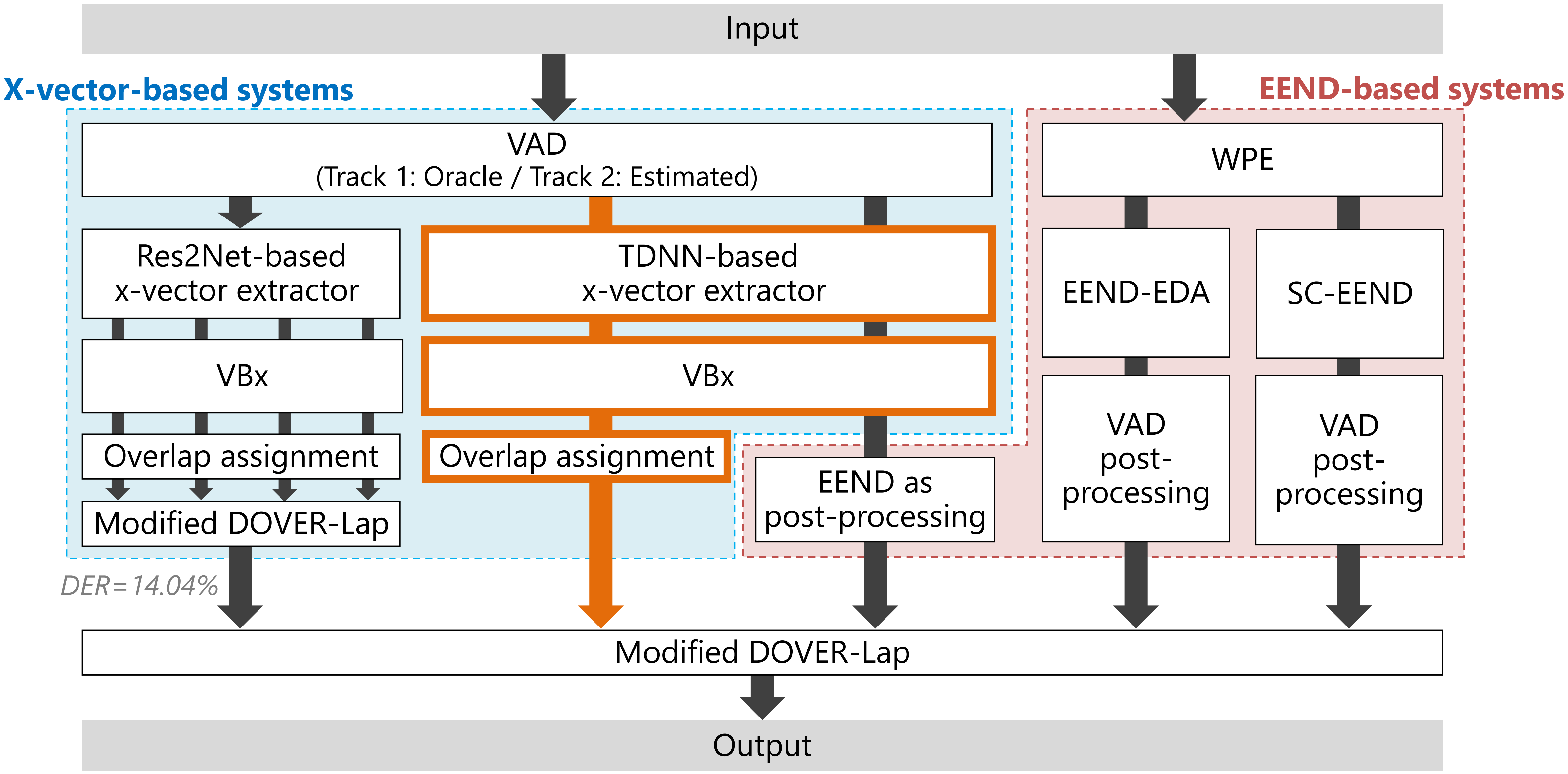
HITACHI
Inspire the Next

DERs/JERs (%) of DIHARD III Track 1 DEV

	Res2Net-BN	**Res2Net-UN**	**Res2Net-BN-Large**	**Res2Net-UN-Large**
	• 23 layers • BatchNorm • $\ln x$	• 23 layers • UttteranceNorm • $\log_{10} x$	• 50 layers • BatchNorm • $\ln x$	• 50 layers • UtteranceNorm • $\log_{10} x$ • SpecAugment
X-vector + Auto-tuning Spectral Clustering	17.09 / 35.69	17.53 / 37.15	16.96 / 35.77	17.55 / 36.78
X-vector + VBx	17.24 / 37.12	17.04 / 36.17	16.85 / 35.86	17.08 / 35.95
X-vector + VBx + OvlAssign	14.89 / 35.64	14.72 / 34.65	14.56 / 34.31	14.74 / 34.40
DOVER-Lap	14.04 / 34.29			

(2) TDNN-Based System

HITACHI
Inspire the Next



(2) TDNN-Based System

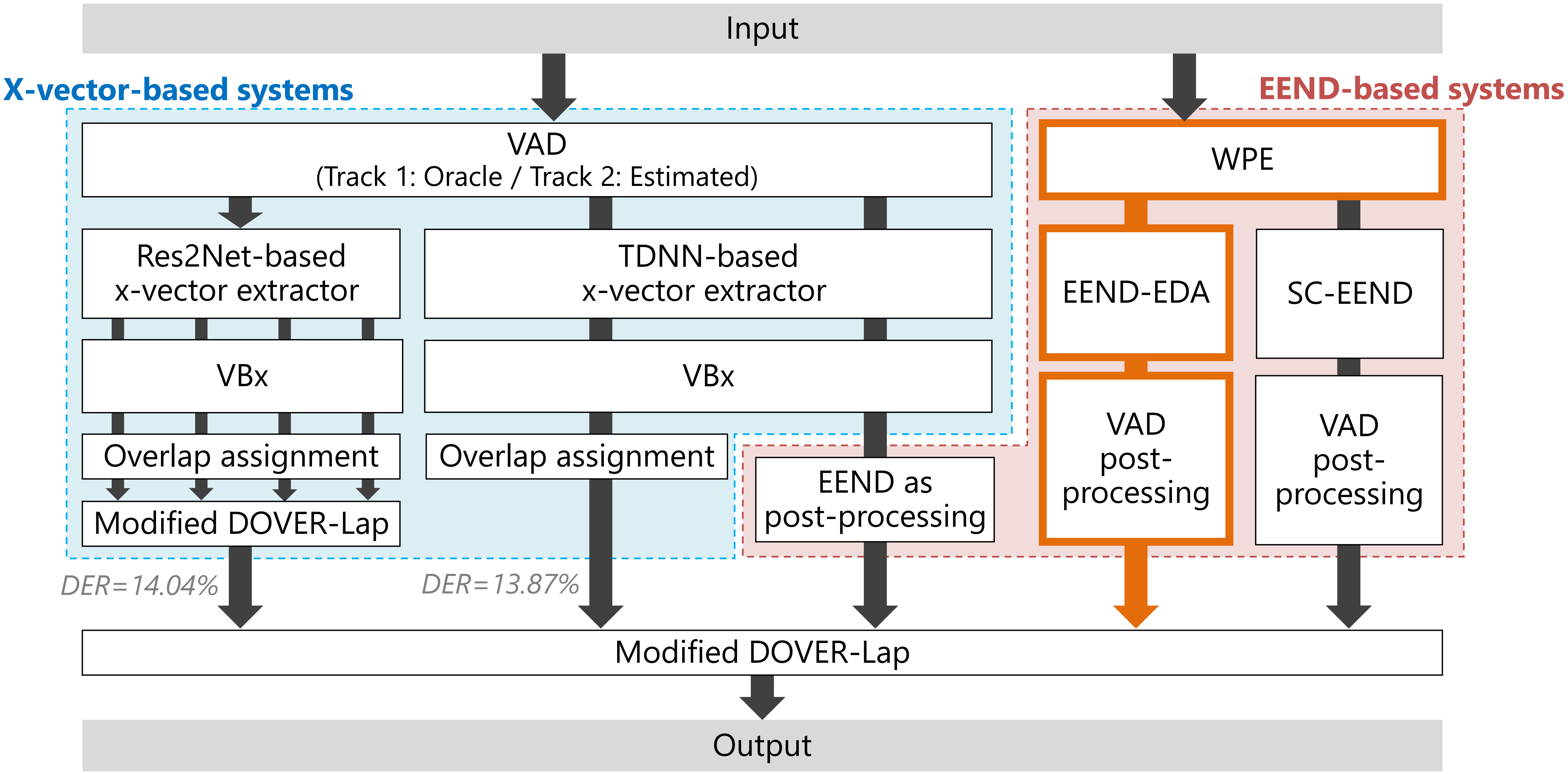
HITACHI
Inspire the Next

- X-vector extractor
  - TDNN-based model in the Kaldi VoxCeleb recipe [Snyder+, ICASSP'19]
    - Input: 40-dimensional filterbanks, with a 25 ms window and 15 ms shift
    - Output: 512-dimensional embeddings
- VBx clustering (Same as Res2Net-based system)
  - Initial clustering using AHC with PLDA, the interpolation of VoxCeleb PLDA and DIHARD III PLDA
  - Then, Bayesian HMM clustering with the LDA
- Overlap assignment (Same as Res2Net-based system)
  - The same model as SincNet-based was trained to detect overlap using DIHARD III DEV
  - Assigned the closest other speaker in the time axis for each detected frame

Results of DIHARD III Track 1 DEV

	DER (%)	JER (%)
X-vector + VBx	16.33	34.18
X-vector + VBx + OvlAssign	13.87	32.73

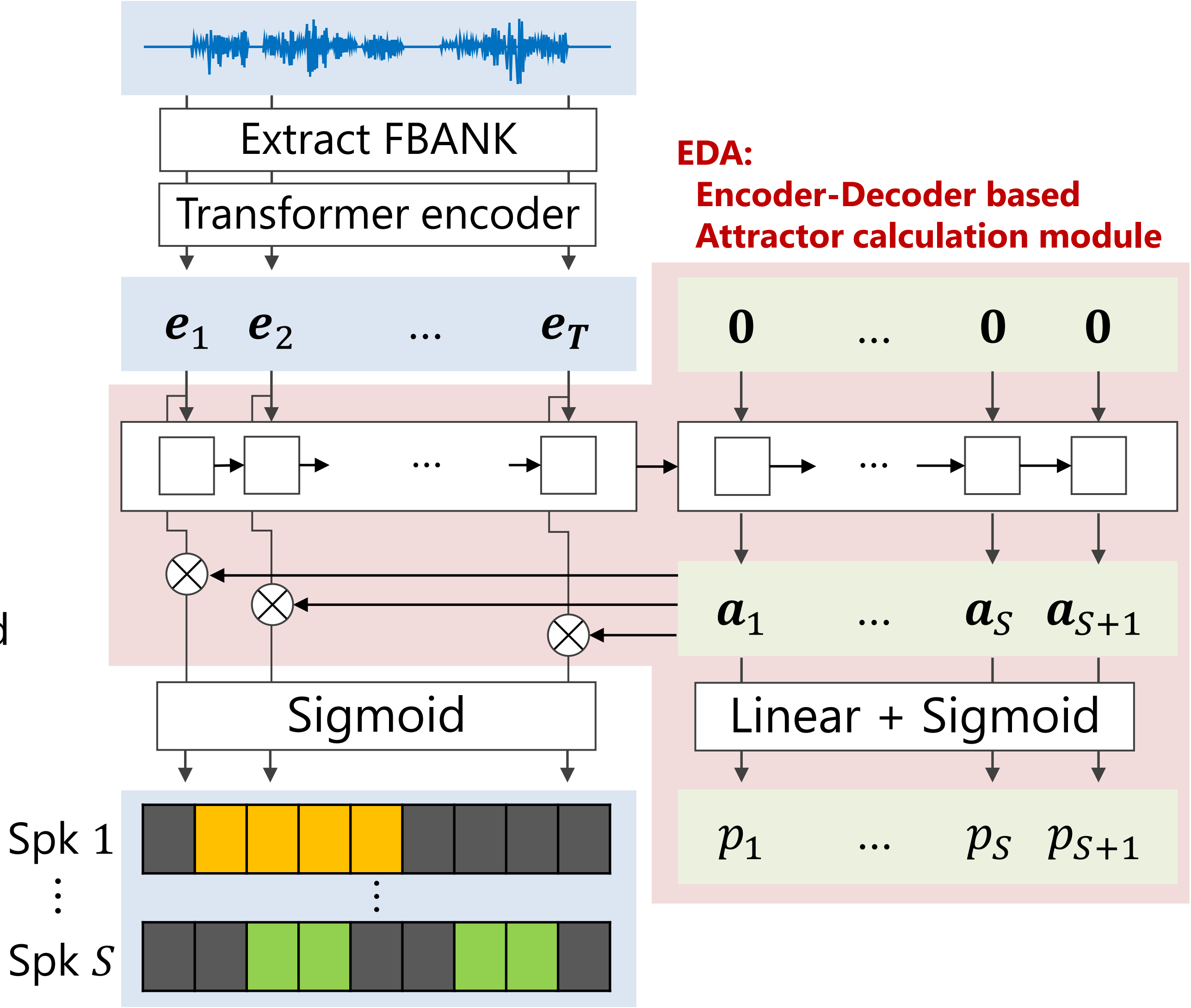
(3) EEND-EDA-Based System



(3) EEND-EDA-Based System

EEND-EDA [Horiguchi+, INTERSPEECH'20]

- Method
  - Calculate a flexible number of attractors from embeddings using an LSTM encoder-decoder
  - Then calculate diarization results based on the dot products of the attractors and embeddings
- Training
  - Train the model for 100 epochs using simulated two-speaker mixtures created from Switchboard and NIST SRE
  - Finetune the model for 75 epochs using simulated mixtures, each of which contains at most 5 speakers (instead of 4 in the IS paper)
  - Adapt the model for using the DIHARD III DEV



(3) EEND-EDA-Based System: Issues and Solutions

HITACHI
Inspire the Next

- Issue 1
  - EEND-EDA performs VAD and diarization simultaneously
    → Need to incorporate with external VAD if the oracle or more accurate VAD is given

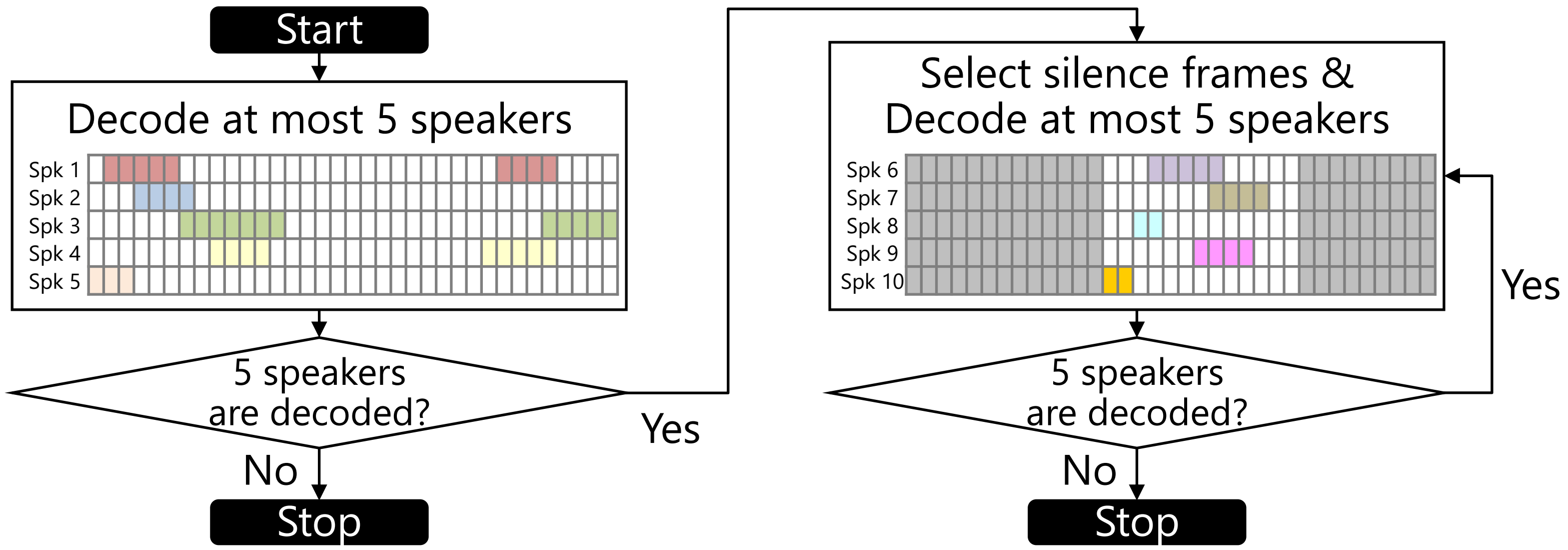
- Solution 1: **VAD post-processing**
  - Remove false alarms using VAD
  - Recover missed speech by assigning the speaker with the highest posteriors using VAD

(3) EEND-EDA-Based System: Issues and Solutions

HITACHI
Inspire the Next

- **Issue 2**
  - EEND-EDA cannot produce diarization results of large number of speakers (>5)

- **Solution 2-1: Iterative inference**
  - Decode 5 speakers repeatedly until EEND output less than 5 speakers



  - Problem: The 6th speaker's speech activities are never overlapped with the 1st-5th speakers

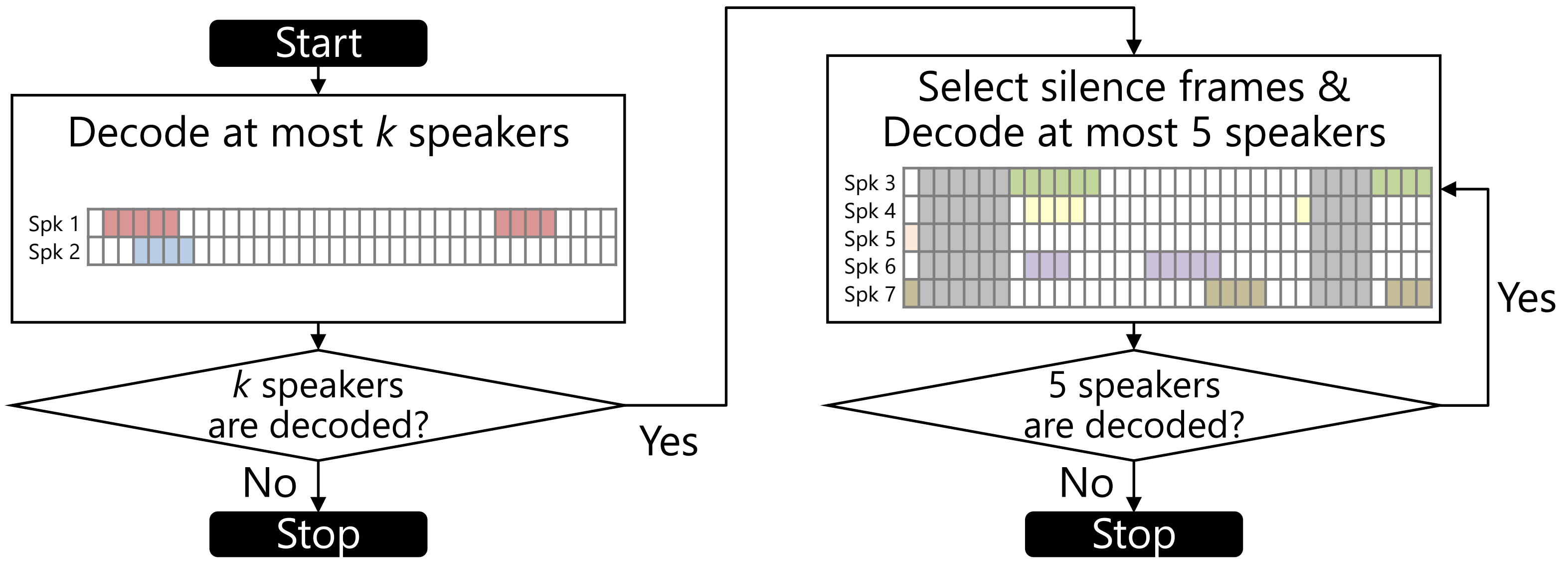
(3) EEND-EDA-Based System: Issues and Solutions

HITACHI
Inspire the Next

- Issue 2
  - EEND-EDA cannot produce diarization results of large number of speakers (>5)

- Solution 2-2: **Iterative inference + DOVER-Lap**
  - Decode at most *k* speakers at the first iteration (*k*=1,2,3,4,5)
  - Decode at most 5 speakers from the second iteration
  - Finally, the five estimated results are combined using DOVER-Lap

Ex) *k*=2



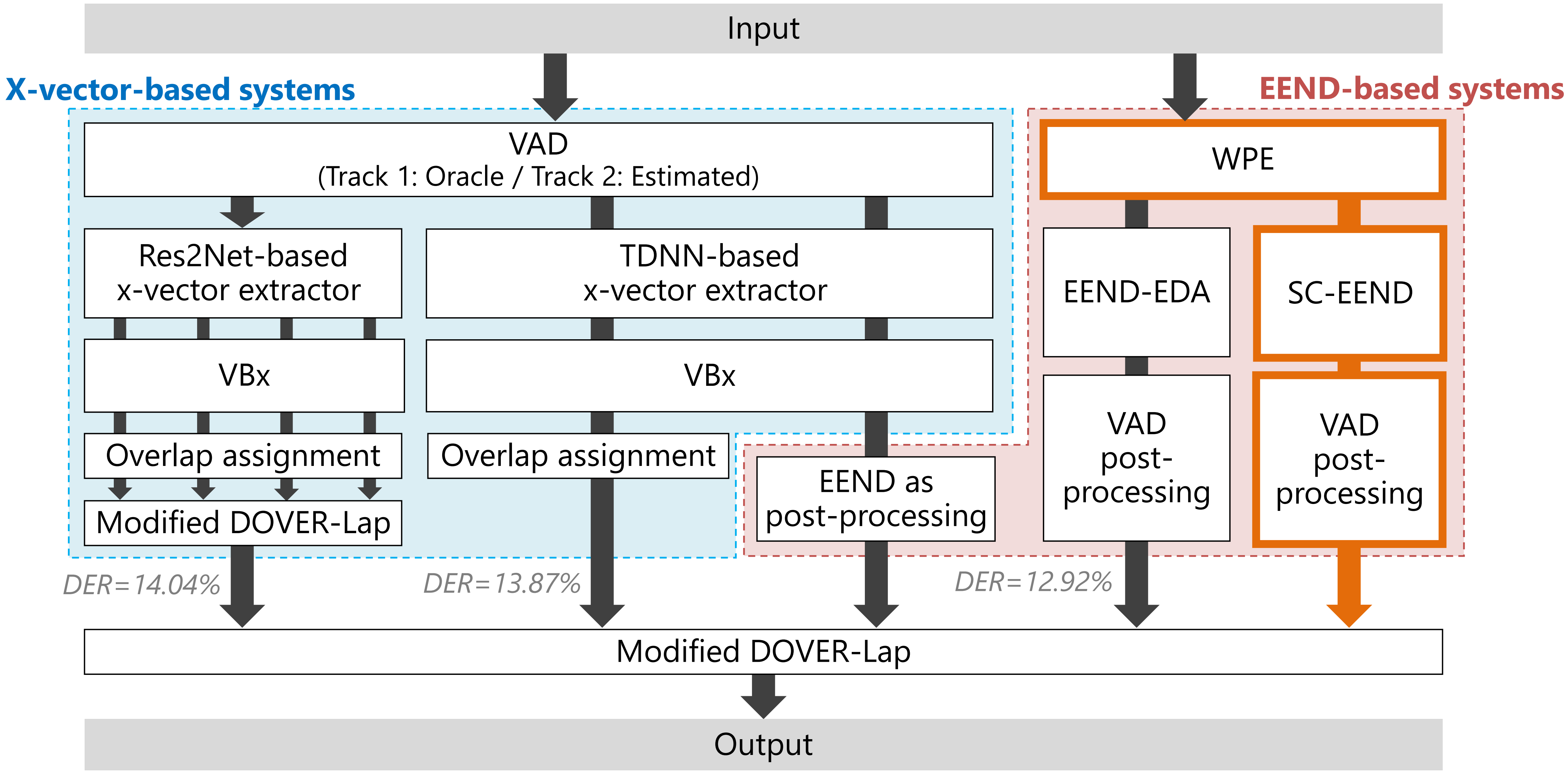
(3) EEND-EDA-Based System: Results of Track 1 DEV

HITACHI
Inspire the Next

Results of DIHARD III Track 1 DEV

Model	Remove false alarms	Recover missed speech	Iterative inference	Iterative inference +DOVER-Lap	DER	JER
4-speaker model [Horiguchi+, INTERSPEECH'20]					21.06	41.63
5-speaker model					18.77	38.98
	✓				17.33	37.92
	✓	✓			13.08	35.38
	✓	✓	✓		13.35	34.19
	✓	✓		✓	**12.92**	**33.85**

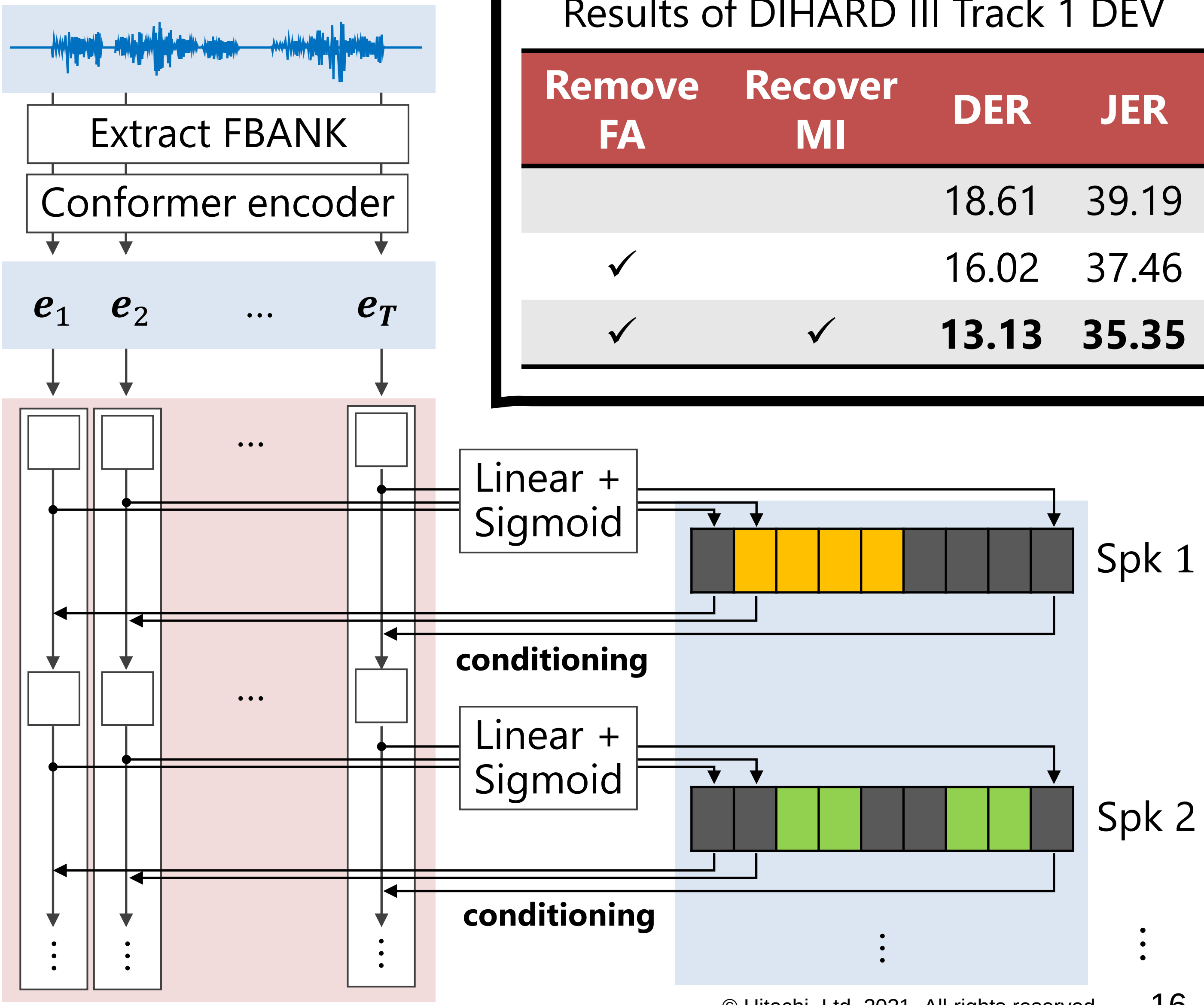
(4) SC-EEND-Based System



(4) SC-EEND-Based System

Speaker-wise conditional EEND (SC-EEND) [Fujita+, arXiv'20]

- Method
  - Estimate each speaker's speech activities sequentially, conditioned on previously estimated speaker's speech activities
  - We replaced Transformer encoders with Conformer [Gulati+, INTERSPEECH'20] encoders
  - VAD post-processing was also applied as in SA-EEND-based system
- Training
  - Train the model for 200 epochs using simulated mixtures, each of which contains at most 4 speakers
  - Adapt the model for another 100 epochs using the DIHARD III DEV set

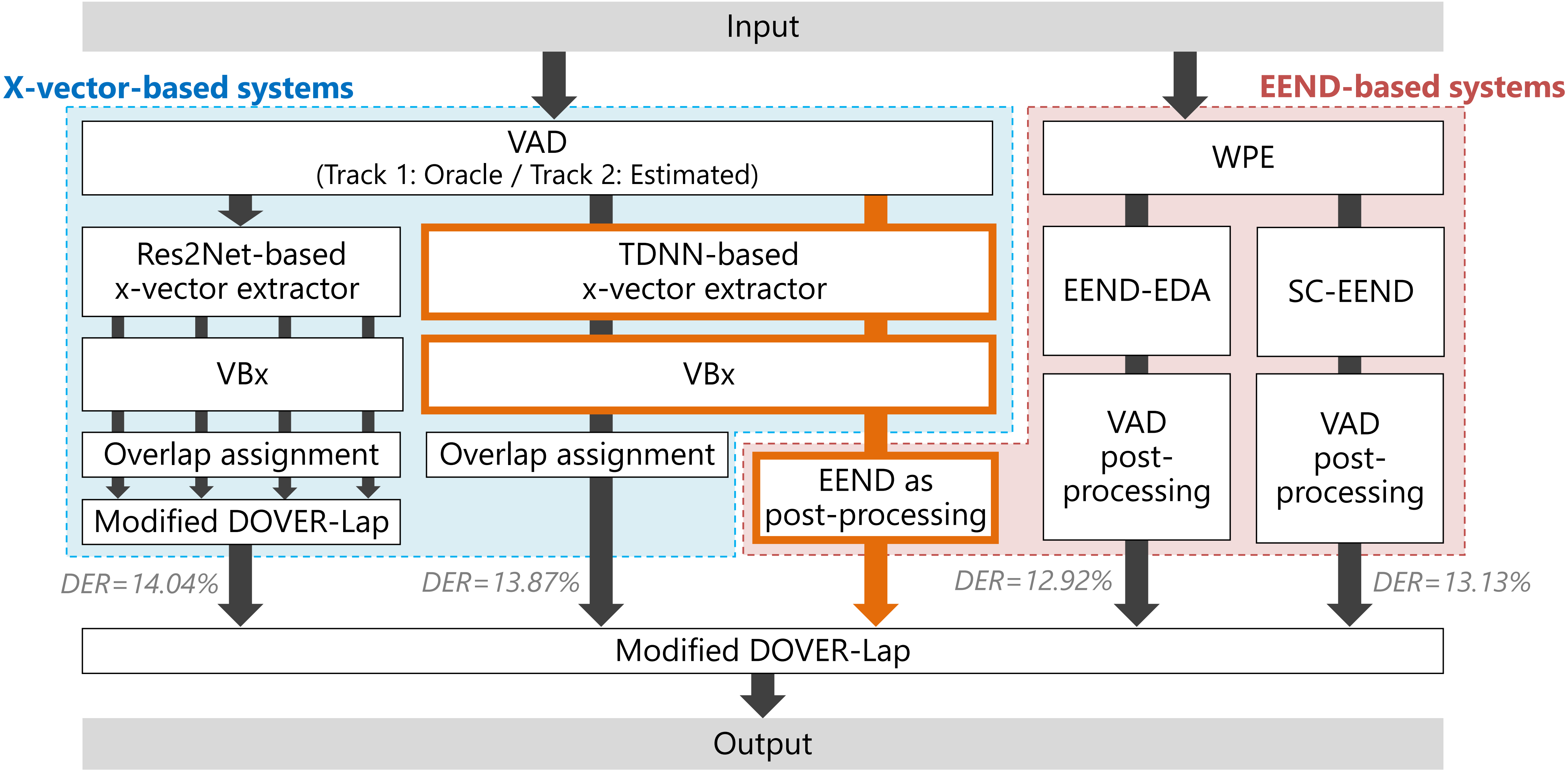


Results of DIHARD III Track 1 DEV

Remove FA	Recover MI	DER	JER
		18.61	39.19
✓		16.02	37.46
✓	✓	**13.13**	**35.35**

(5) TDNN-Based X-vectors + EEND as Post-Processing

HITACHI
Inspire the Next



(5) TDNN-Based X-vectors + EEND as Post-Processing

HITACHI
Inspire the Next

EEND as Post-Processing [Horiguchi+, arXiv'20]

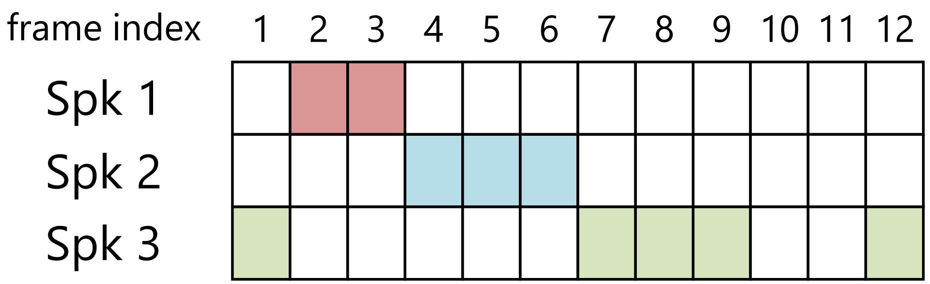
- Motivation
  - X-vector-based system
    - ✓ can deal with large number of speakers
    - × has difficulty on overlap processing
  - EEND-based system
    - ✓ Can handle overlapping speech
    - × cannot deal with large number of speakers

- Method
  - Update diarization results of x-vector-based system using EEND by applying the following steps iteratively
    1. Frame selection to contain only two speakers
    2. Overlap estimation using an EEND model

(5) TDNN-Based X-vectors + EEND as Post-Processing

HITACHI
Inspire the Next

Initial results (from x-vector clustering)



Processing order

Spks 2&3 (#Frames = |1,4,5,6,7,8,9,10,11,12| = 10)

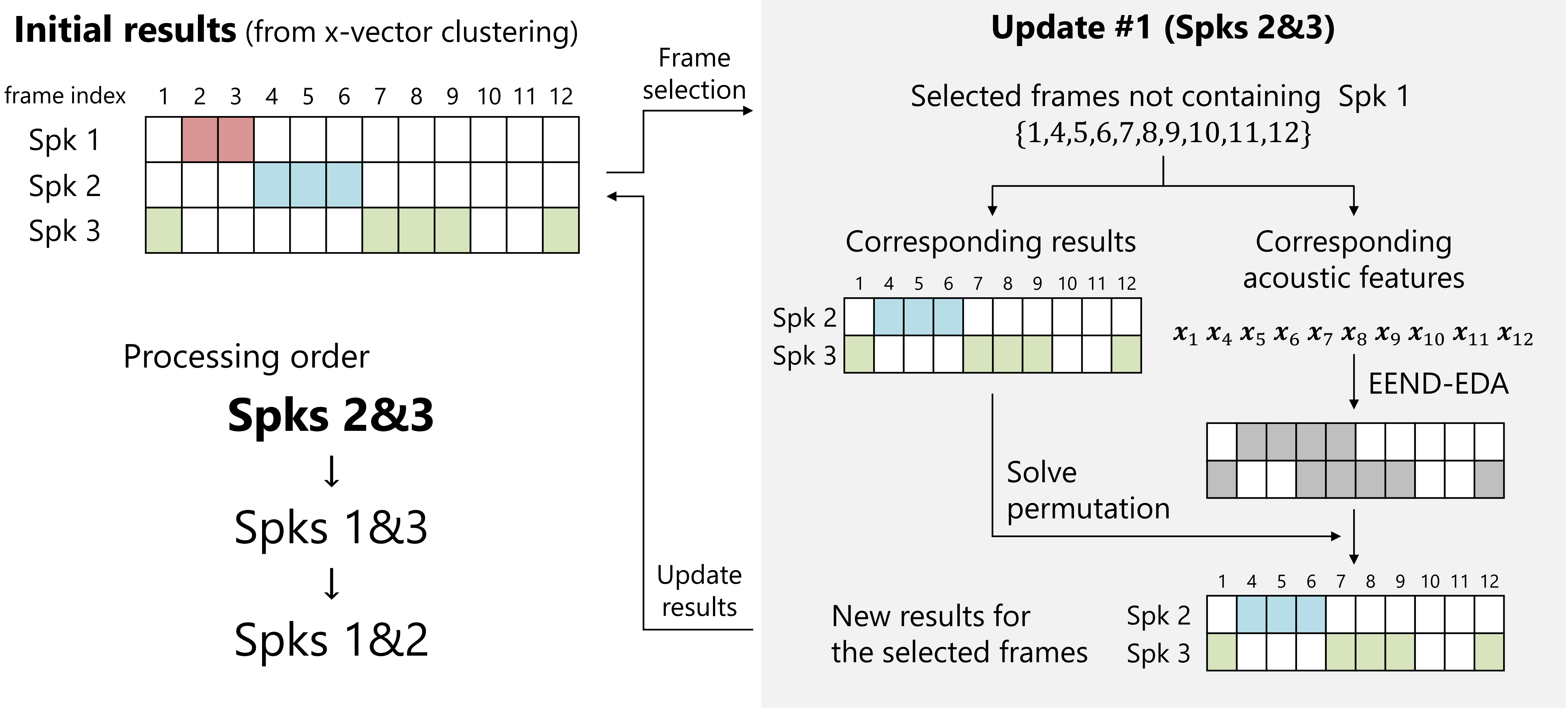
↓

Spks 1&3 (#Frames = |1,2,3,7,8,9,10,11,12| = 9)

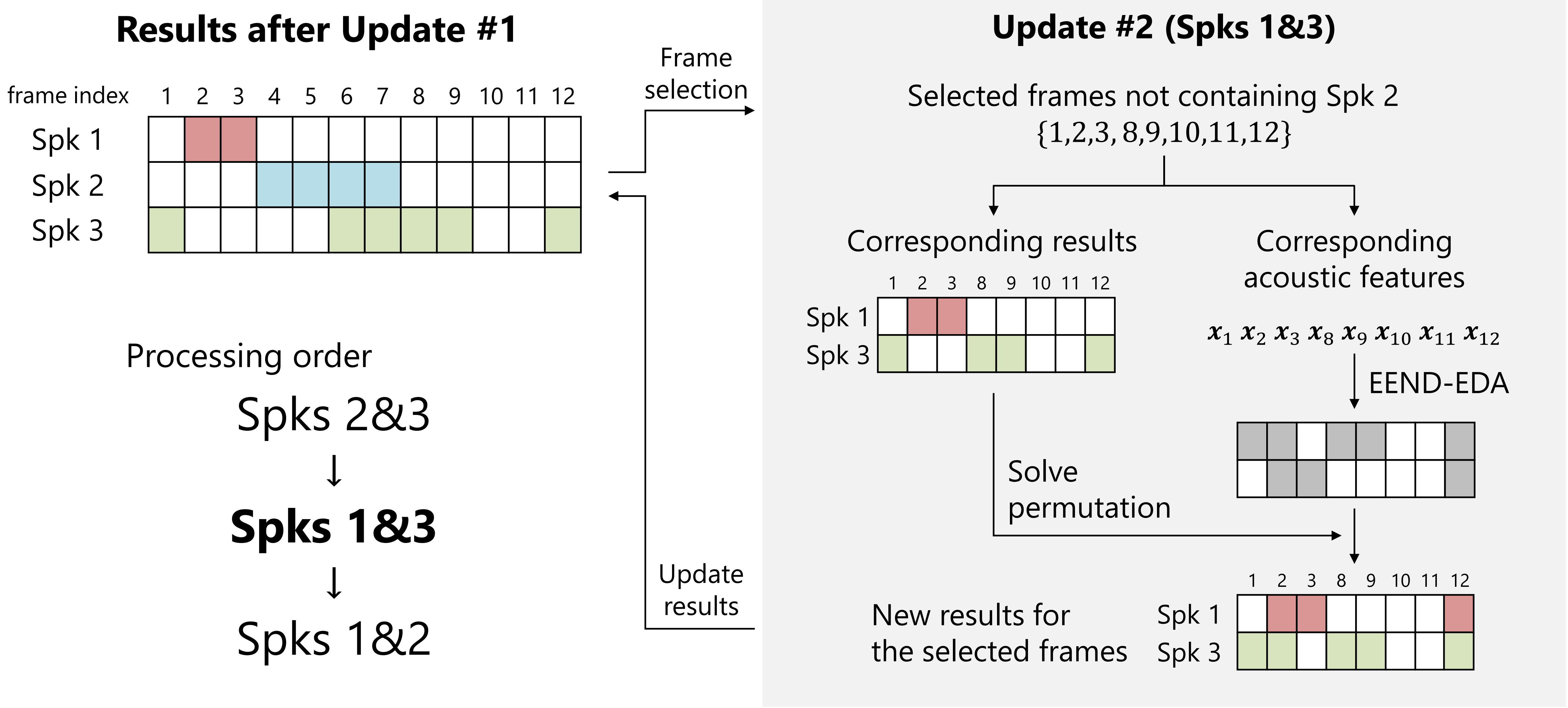
↓

Spks 1&2 (#Frames = |2,3,4,5,6,10,11| = 7)

(5) TDNN-Based X-vectors + EEND as Post-Processing

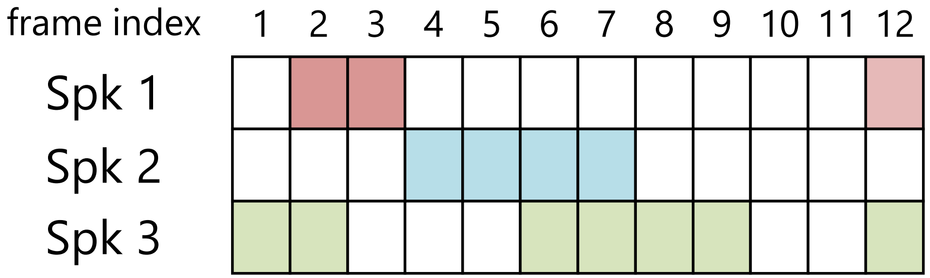


(5) TDNN-Based X-vectors + EEND as Post-Processing



(5) TDNN-Based X-vectors + EEND as Post-Processing

Results after Update #2



Processing order

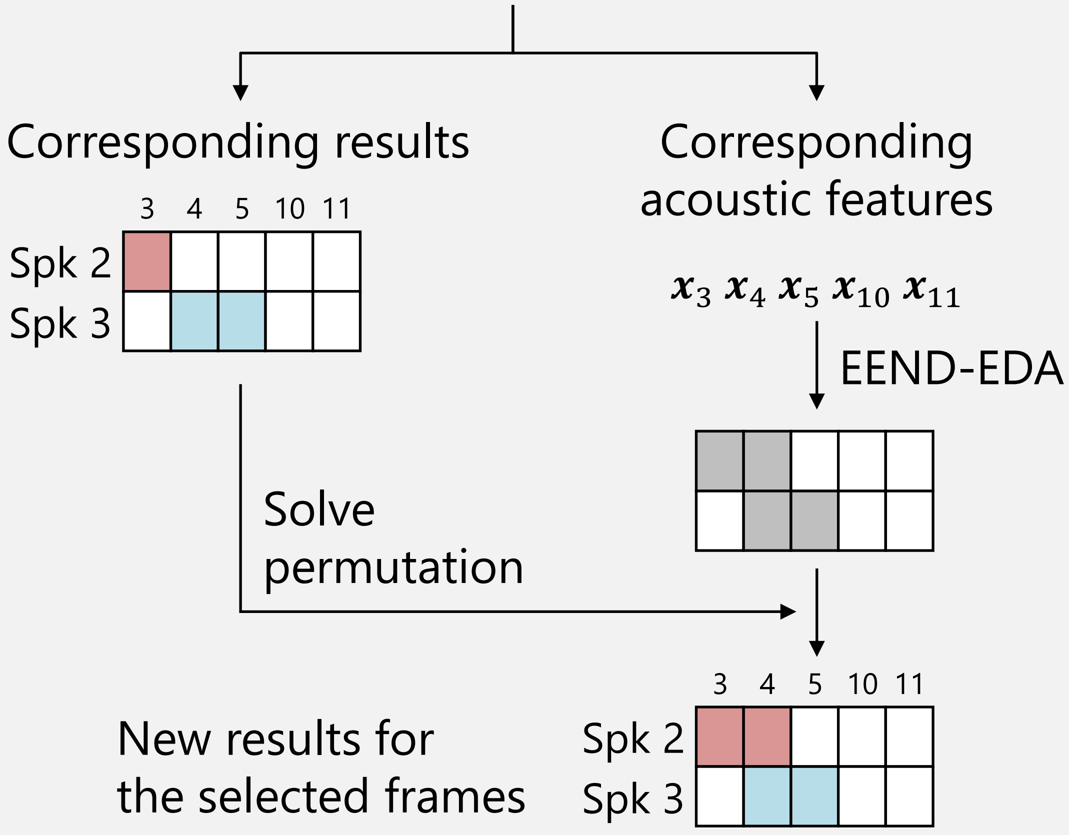
Spks 2&3
↓
Spks 1&3
↓
Spks 1&2

Frame selection

Update results

Update #3 (Spks 1&2)

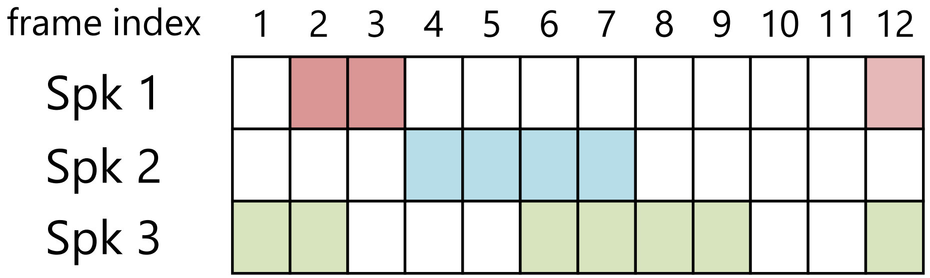
Selected frames not containing Spk 3
{3,4,5,10,11}



(5) TDNN-Based X-vectors + EEND as Post-Processing

HITACHI
Inspire the Next

Final results (Results after Update #3)



Processing order

Spks 2&3

↓

Spks 1&3

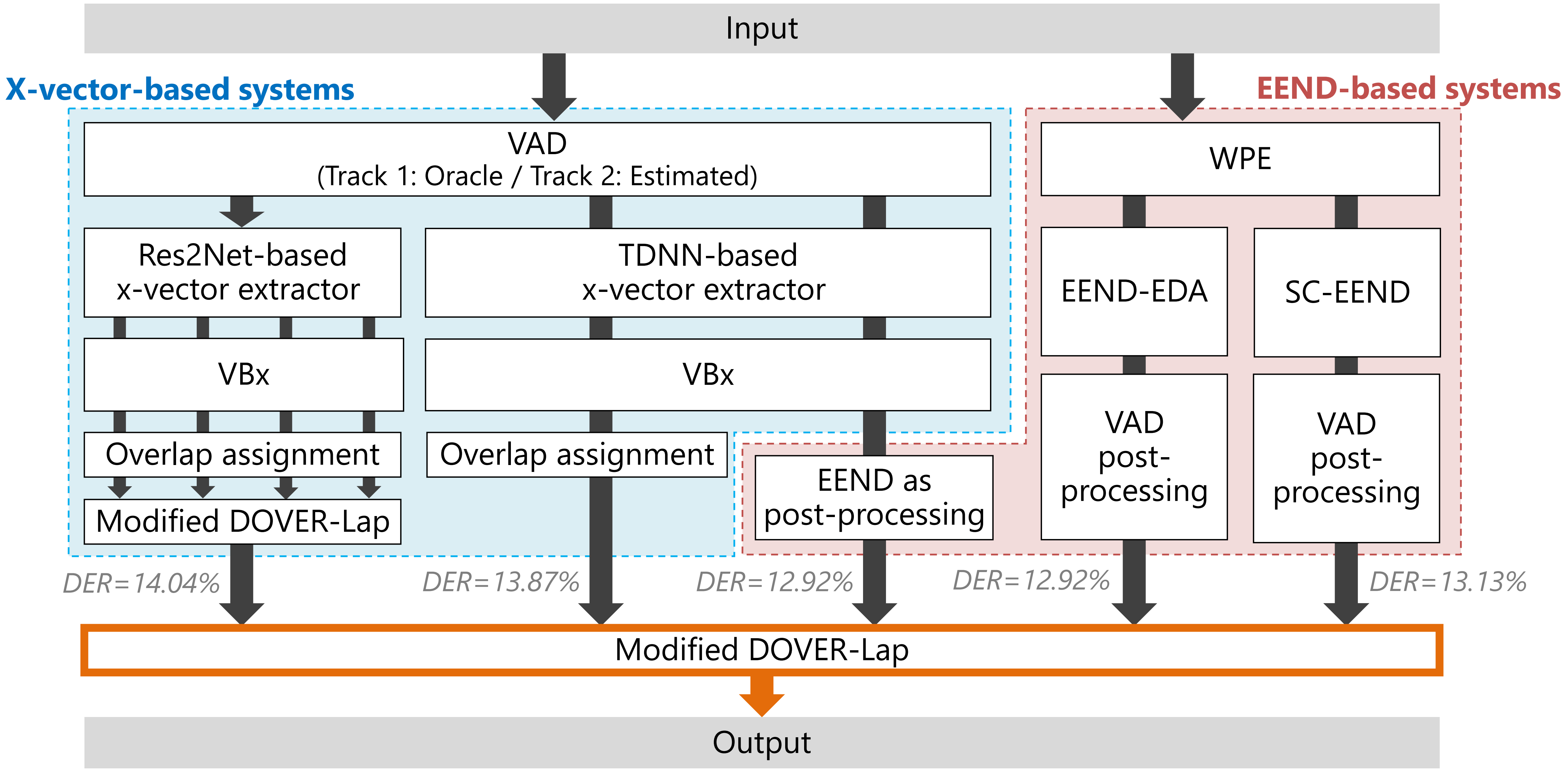
↓

Spks 1&2

Results of DIHARD III Track 1 DEV

	DER (%)	JER (%)
X-vector + VBx	16.33	34.18
X-vector + VBx + OvlAssign (System (2))	13.87	32.73
X-vector + VBx + EENDasP	12.63	31.52

(6) Modified DOVER-Lap



(6) Modified DOVER-Lap

DOVER-Lap [Raj+, SLT'21]

- Method to combine overlap-aware diarization results
- We modified the processing when multiple speakers have the same rank
  - **Original**: Assigns uniformly-divided regions for each speaker
  - **Modified**: Assigns the region for all the tied speakers without any division
- In addition, we introduced a weighting mechanism to change the importance of each system

Results of DIHARD III Track 1 DEV

Method	DER (%)	JER (%)
(1) Res2Net-based x-vector + VBx + OvlAssign	14.04	34.29
(2) TDNN-based x-vector + VBx + OvlAssign	13.87	32.73
(3) EEND-EDA	12.92	33.85
(4) SC-EEND	13.13	35.35
(5) TDNN-based x-vector + VBx + EENDasP	12.63	31.52
DOVER-Lap	12.07	32.29
Modified DOVER-Lap	10.73	31.39
Modified DOVER-Lap + manual weighting	10.68	31.01

DERs (So Far)

	Track 1				Track 2			
	DEV		EVAL		DEV		EVAL	
	full	core	full	core	full	core	full	core
Baseline	19.41	20.25	19.25	20.65	21.71	22.28	25.36	27.34
(1) Res2Net-based x-vector + VBx + OvlAssign	14.04	15.18	15.81	18.47	17.26	18.39	21.37	24.64
(2) TDNN-based x-vector + VBx + OvlAssign	13.87	14.88	15.65	18.20	17.61	18.64	21.47	24.58
(3) EEND-EDA	12.92	13.95	13.95	17.28	15.90	18.50	19.04	22.84
(4) SC-EEND	13.13	16.05	15.16	19.14	16.16	19.00	20.30	24.75
(5) TDNN-based x-vector + VBx + EENDasP	12.63	14.61	13.30	15.92	15.94	18.09	18.13	21.31
(6) DOVER-Lap of (1)(2)(3)(4)(5)	**10.73**	**12.56**	**11.83**	**14.41**	**14.13**	**16.06**	**17.21**	**20.34**

Use (6) for self-supervised adaptation (SSA) of EEND-EDA

- Created pseudo labels for the EVAL set and redid the adaptation
- We also tried SSA of SC-EEND, but the DOVER-Lap results became bad

DERs with Self-Supervised Adaptation of EEND-EDA

HITACHI
Inspire the Next

	Track 1				Track 2			
	Dev		Eval		Dev		Eval	
	full	core	full	core	full	core	full	core
Baseline	19.41	20.25	19.25	20.65	21.71	22.28	25.36	27.34
(1) Res2Net-based x-vector + VBx + OvlAssign	14.04	15.18	15.81	18.47	17.26	18.39	21.37	24.64
(2) TDNN-based x-vector + VBx + OvlAssign	13.87	14.88	15.65	18.20	17.61	18.64	21.47	24.58
(3) EEND-EDA	12.92	13.95	13.95	17.28	15.90	18.50	19.04	22.84
↓	↓	↓	↓	↓	↓	↓	↓	↓
(7) EEND-EDA (SSA)	12.95	15.69	12.74	15.86	15.03	17.52	17.81	21.31
(4) SC-EEND	13.13	16.05	15.16	19.14	16.16	19.00	20.30	24.75
(5) TDNN-based x-vector + VBx + EENDasP	12.63	14.61	13.30	15.92	15.94	18.09	18.13	21.31
↓	↓	↓	↓	↓	↓	↓	↓	↓
(8) TDNN-based x-vector + VBx + EENDasP (SSA)	12.54	14.55	12.74	15.34	15.45	17.77	17.60	20.84
(6) DOVER-Lap of (1)(2)(3)(4)(5)	10.73	12.56	11.83	14.41	14.13	16.06	17.21	20.34
↓	↓	↓	↓	↓	↓	↓	↓	↓
(9) DOVER-Lap of (1)(2)(7)(4)(8) → **Submitted**	**10.65**	**12.74**	**11.58**	**14.09**	**13.85**	**15.81**	**16.94**	**20.01**

Conclusion

System highlights

- SincNet-based and TDNN-based VAD
- Modified DOVER-Lap of five subsystems
  - Res2Net-based and TDNN-based x-vector systems
  - EEND-based systems with VAD post-processing and iterative inference
  - TDNN x-vectors + EEND as post-processing system
- Self-supervised adaptation of EEND

Results from the leaderboard

- Track 1
  - Full: DER=11.58 % (2nd place)
  - Core: DER=14.09 % (2nd place)
- Track 2
  - Full: DER=16.94 % (2nd place)
  - Core: DER=20.01 % (3rd place)