The Academia Sinica Systems of Speech Recognition and Speaker Diarization for the CHiME-6 Challenge

Hung-Shin Lee, Yu-Huai Peng, Pin-Tuan Huang, Ying-Chun Tseng, Chia-Hua Wu, Yu Tsao, Hsin-Min Wang

Academia Sinica, Taiwan

Outline

- Track 1: multiple-array ASR
  - Our contributions
  - Results

- Track 2: multiple-array diarization+ASR
  - Our contributions
  - Results

Track 1: Our Contributions

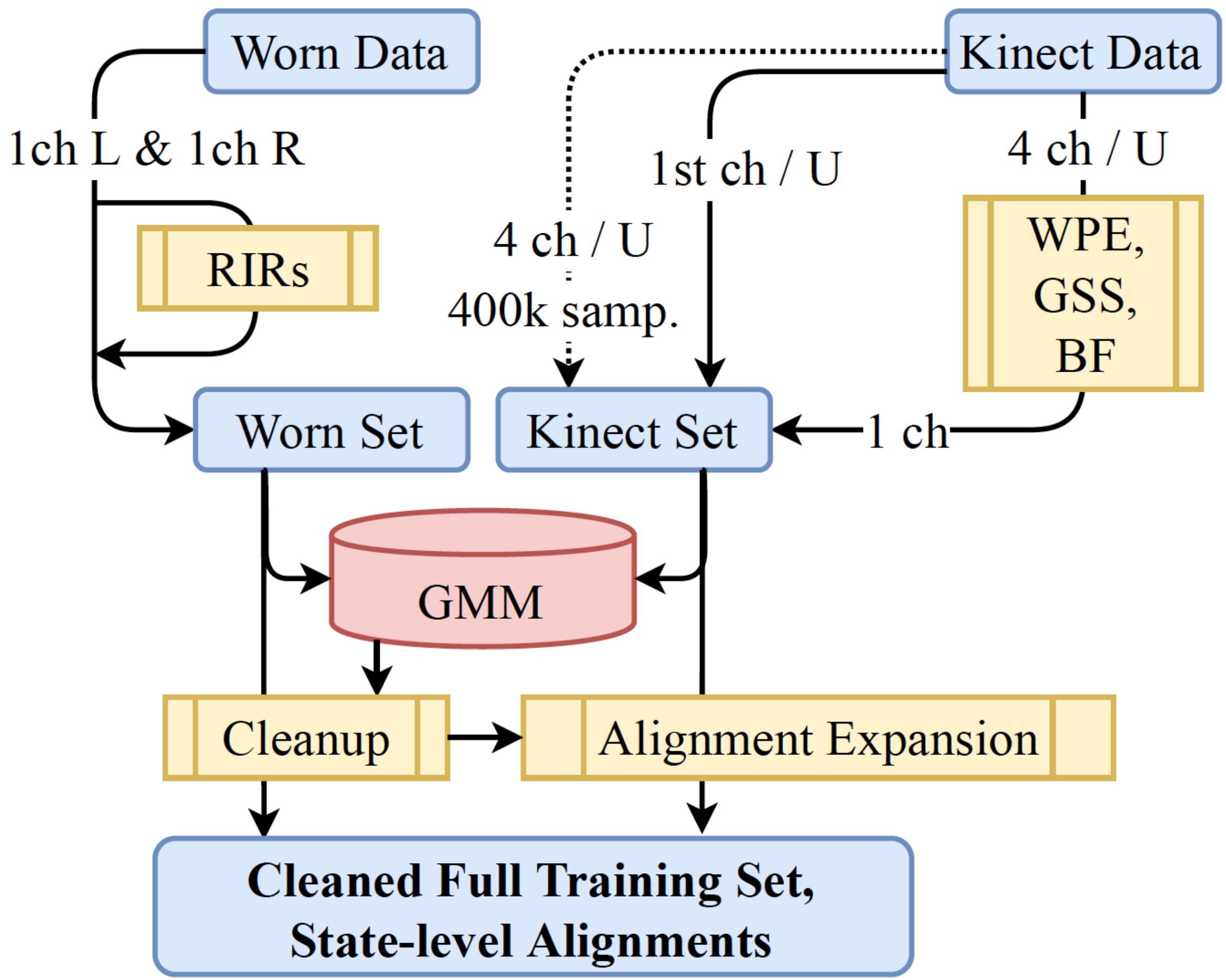
- Compared with the baseline system, we
  - We applied WPE, GSS, and BF to all the **Kinect data** in the training phase
  - Alignment expansion from the **Worn data** to the **Kinect data** was used
  - Four other kinds acoustic models, including our proposed **DcAE** and **FEAM**, were used

Track 1: Results

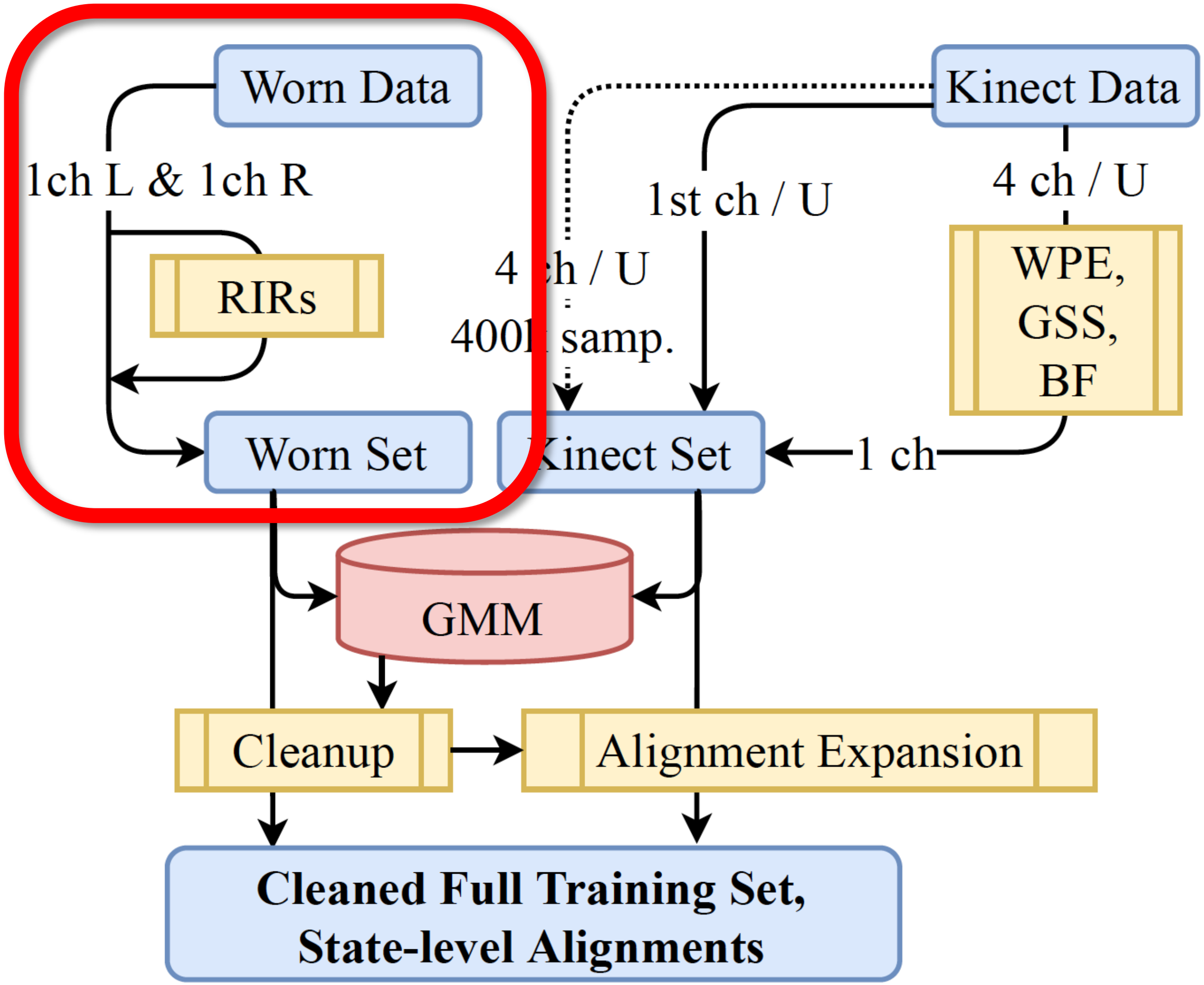
Table 1: *WERs (%) for Track 1 and Track 2 (Category A only).*

Model	Track 1		Track 2	
	Dev	**Eval**	**Dev**	**Eval**
Baseline	51.32	51.36	84.25	77.94
TDNN-F	50.12	49.36	75.89	73.68
RBiLSTM	52.43	50.26	76.90	73.39
DcAE-B	50.12	49.68	75.90	73.66
DcAE-U	49.86	49.63	75.78	73.54
FEAM-U	53.47	52.70	78.70	76.20
ROVER	**47.28**	**46.82**	**74.36**	**71.56**

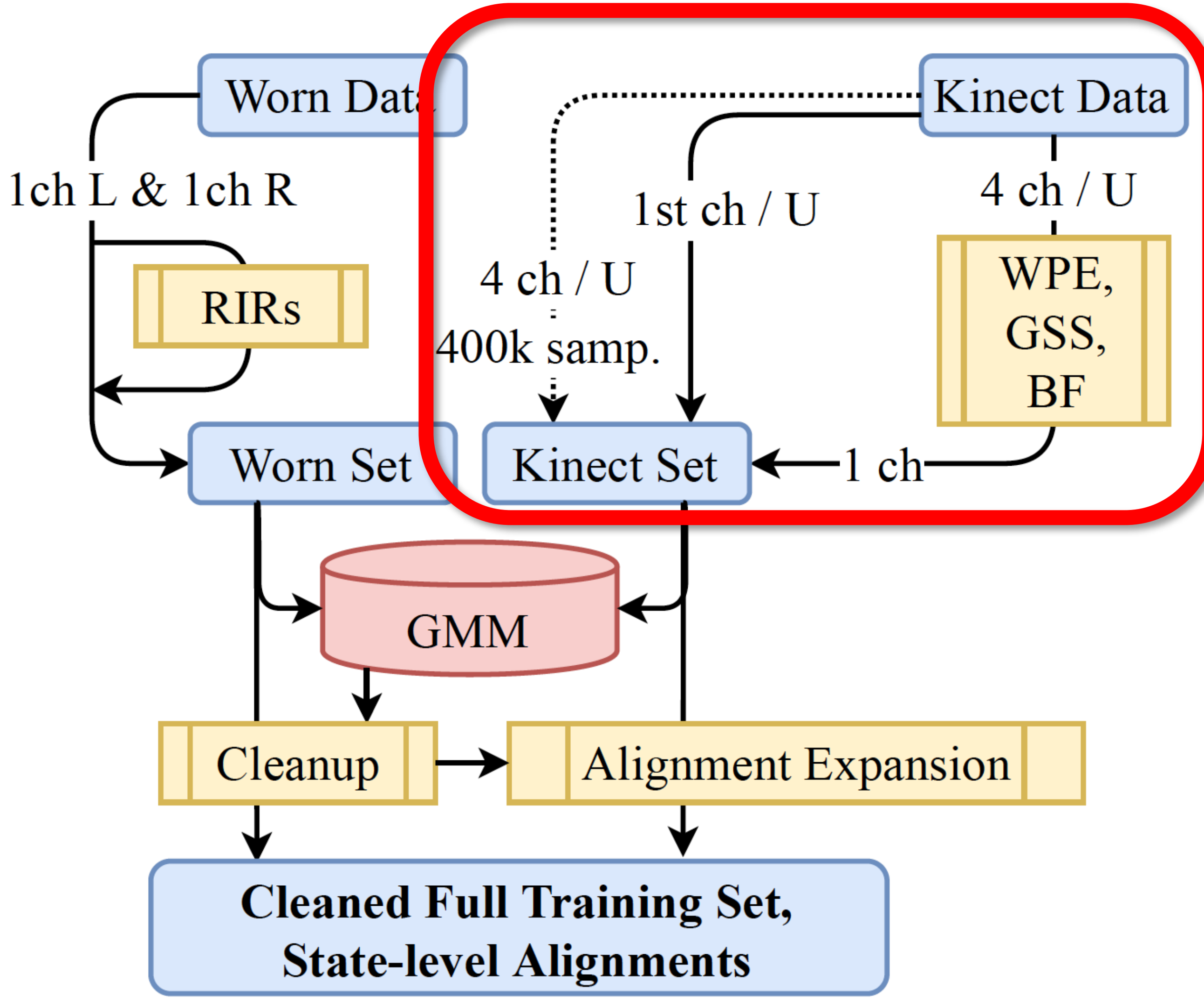
Front-end Data Processing



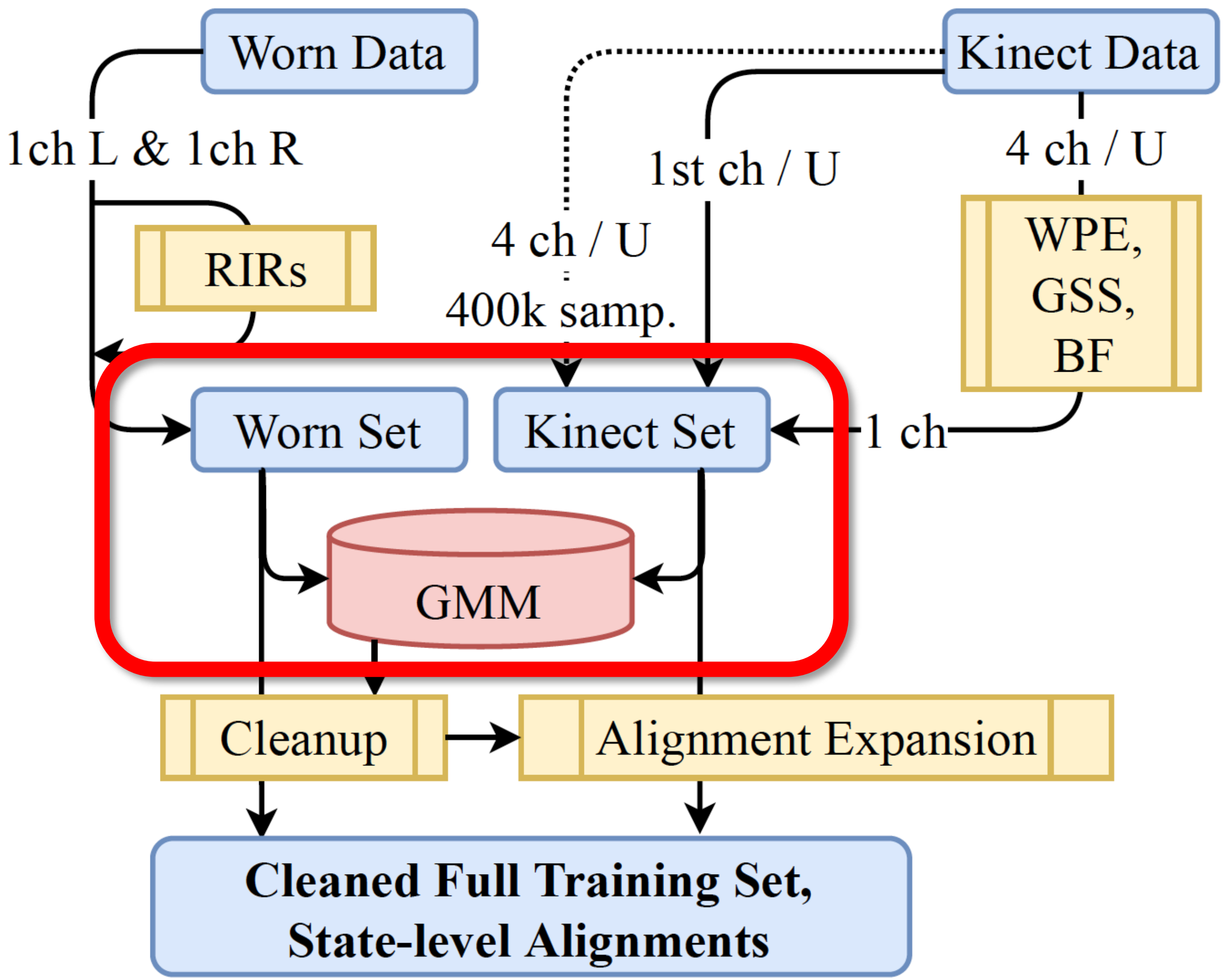
Front-end Data Processing



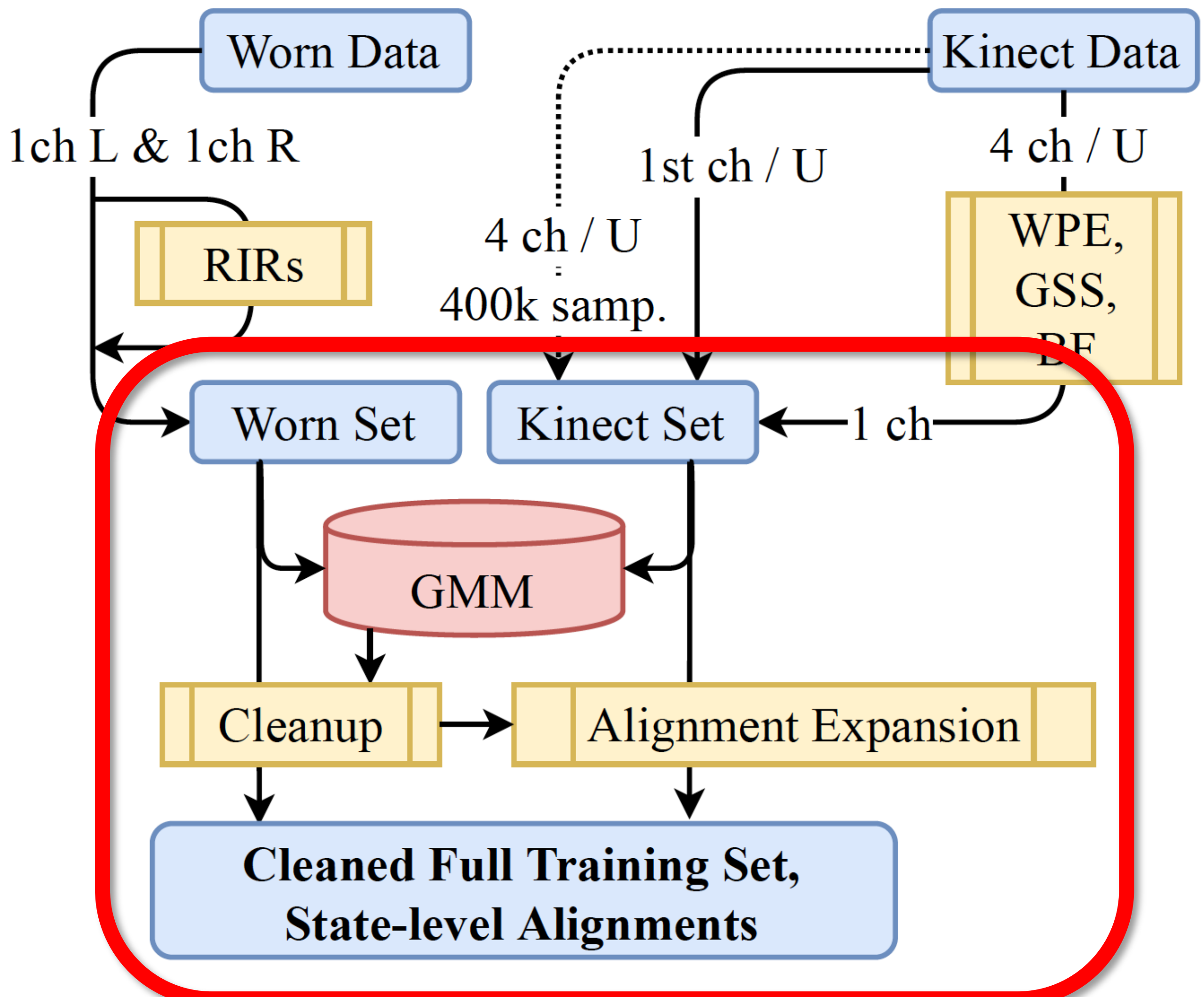
Front-end Data Processing



Front-end Data Processing



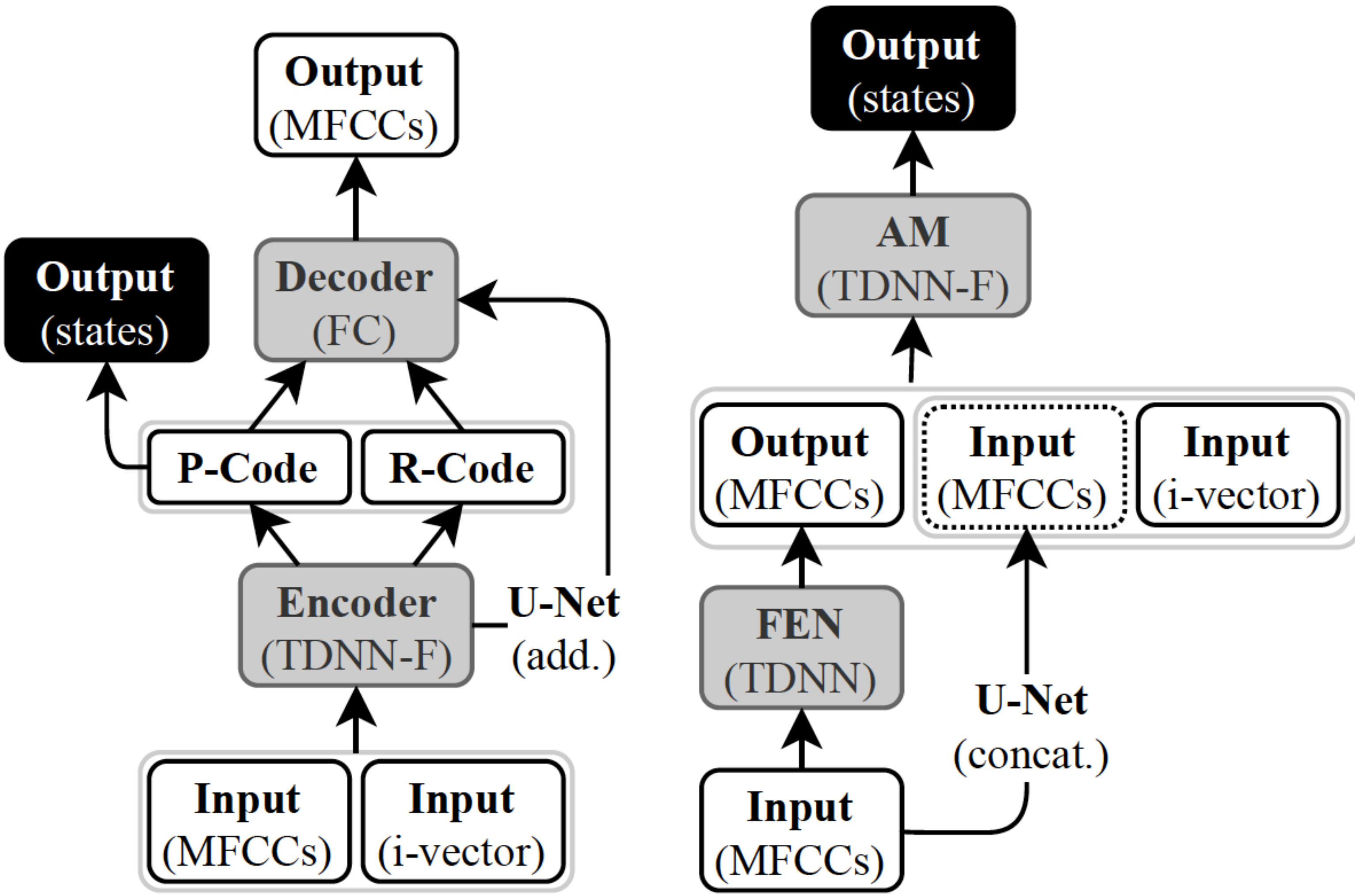
Front-end Data Processing



Back-end Acoustic Modeling

- Data augmentation
  - speed perturbation
  - volume perturbation
- 40-d MFCCs & 100-d i-vectors
- Five kinds of acoustic models
  - TDNN-F
  - RBiLSTM (1ch)
  - **DcAE-B**
  - **DcAE-U**
  - **FEAM-U**

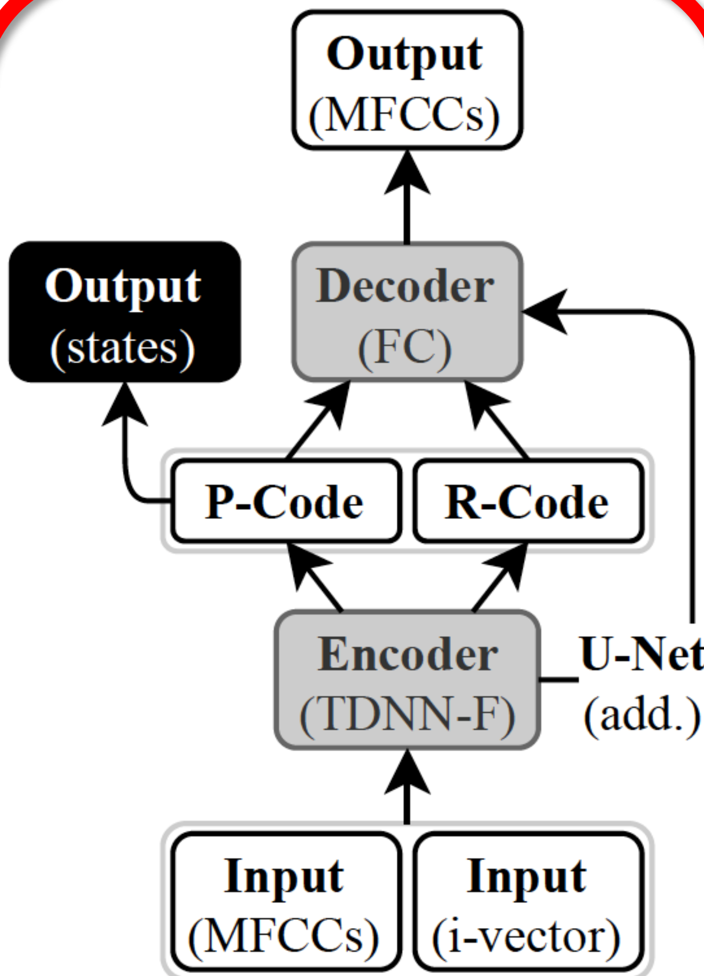
Back-end Acoustic Modeling



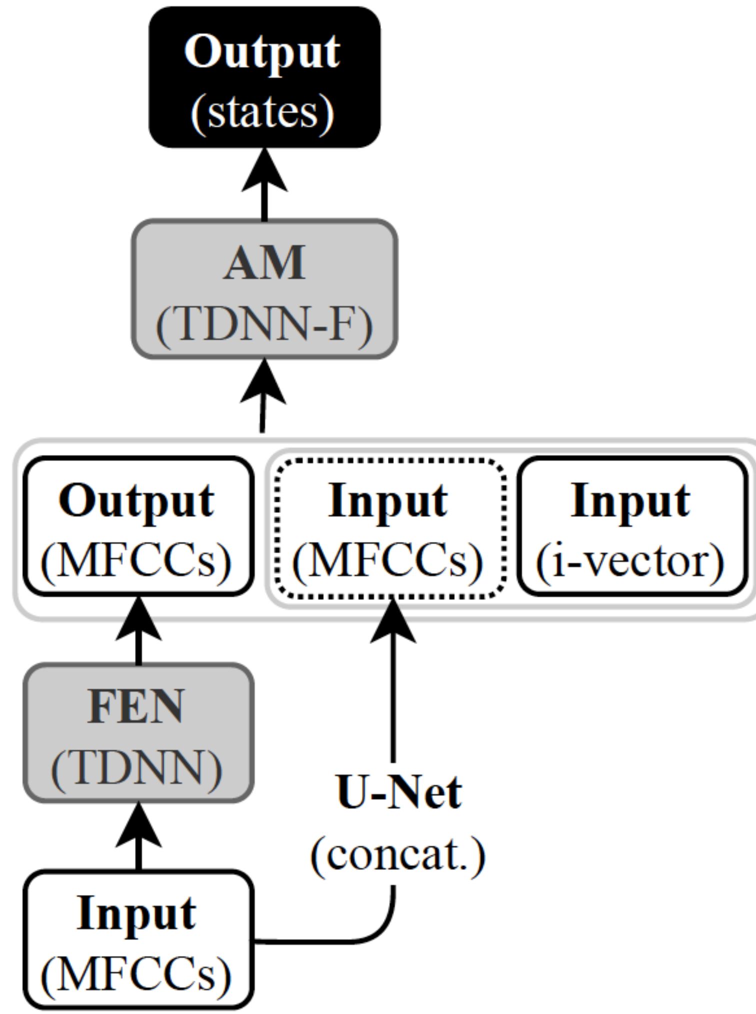
(a) DcAE-B & -U

(b) FEAM-U

Back-end Acoustic Modeling

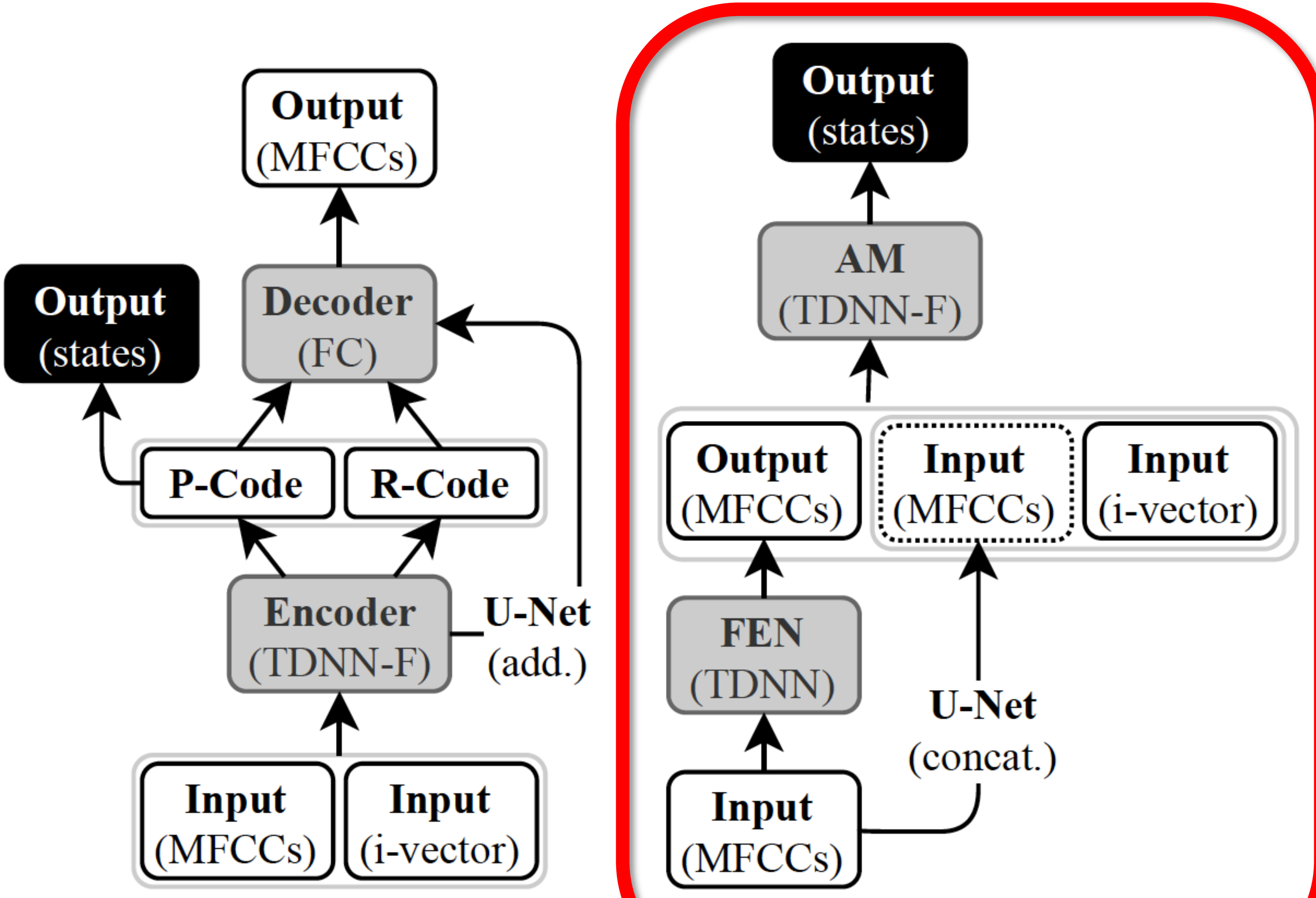


(a) DcAE-B & -U



(b) FEAM-U

Back-end Acoustic Modeling



(a) DcAE-B & -U

(b) FEAM-U

Track 2: Our Contributions

- Compared with the baseline system, we
  - Combined all channels of the Kinect data with BeamformIt (BF)
  - Developed a new training scheme for speaker representations using Speaker Change information and CNN-based ResNet-34
  - Performed re-segmentation with VB diarization

Track 2: Results

Table 1: *WERs (%) for Track 1 and Track 2 (Category A only).*

Model	Track 1 Dev	Track 1 Eval	Track 2 Dev	Track 2 Eval
Baseline	51.32	51.36	84.25	77.94
TDNN-F	50.12	49.36	75.89	73.68
RBiLSTM	52.43	50.26	76.90	73.39
DcAE-B	50.12	49.68	75.90	73.66
DcAE-U	49.86	49.63	75.78	73.54
FEAM-U	53.47	52.70	78.70	76.20
ROVER	**47.28**	**46.82**	**74.36**	**71.56**

Track 2: Results

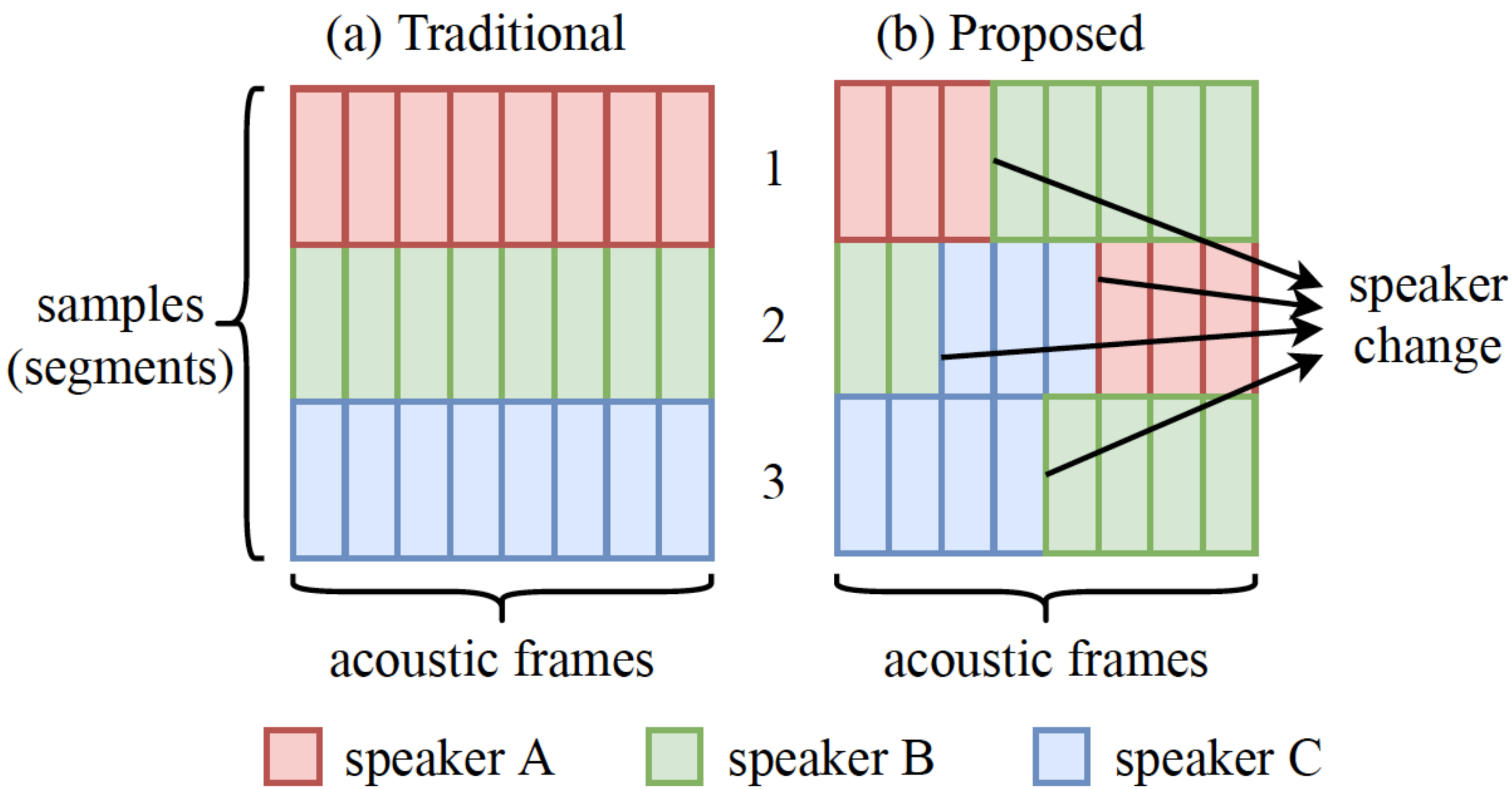
Table 2: *Results for Track 2. The acoustic models are the same.*

Model	**Dev**			**Eval**		
	DER	**JER**	**WER**	**DER**	**JER**	**WER**
Baseline	63.42	70.83	84.25	68.20	72.54	77.94
Proposed	**56.77**	**60.62**	**75.57**	**59.17**	**63.40**	**72.82**

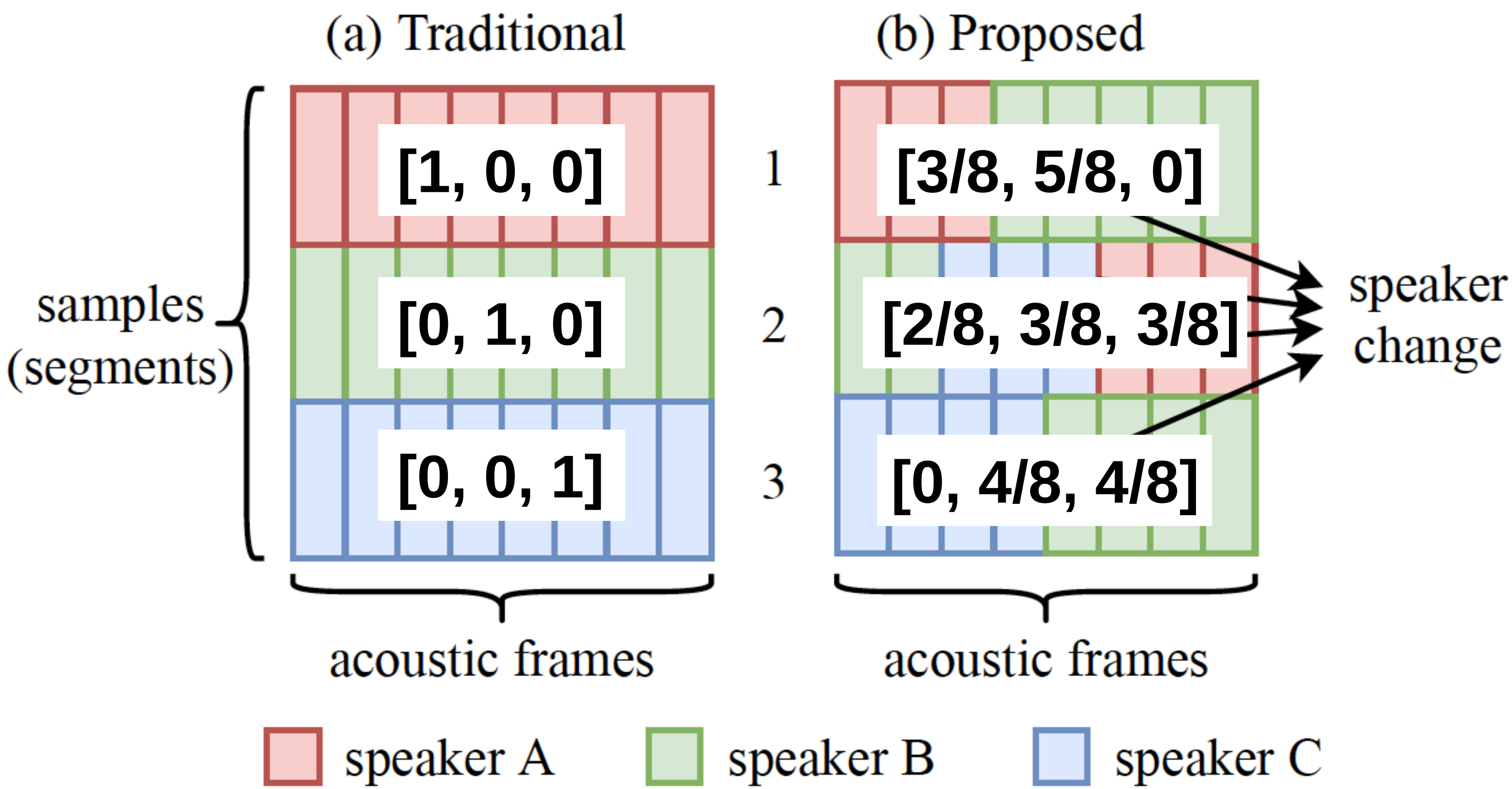
Track 2: Front-end Processing

- Our front-end data processing follows the baseline program, except that…
  - We used all channels in the Kinect data
  - (Only one specific Kinect was used in baseline)

Track 2: Speaker Modeling



Track 2: Speaker Modeling



Conclusions & Future Work

- In Track 1, we evaluated newly proposed acoustic models, namely DcAE and FEAM
  - DcAE outperforms TDNN-F
  - FEAM needs some modifications and fine tuning in the future
- In Track 2, our proposed speaker modeling method was proved useful for speaker diarization and downstream ASR