The USTC-NELSLIP Systems for CHiME-6 Challenge

Jun Du
National Engineering Lab for Speech and Language Information Processing (NELSLIP)
University of Science and Technology of China (USTC)
05/04/2020

Team

Jun Du (USTC)

Yan-Hui Tu (USTC)

Lei Sun (USTC)

Li Chai (USTC)

Xin Tang (USTC)

Mao-Kui He(USTC)

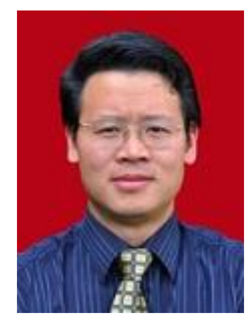
Feng Ma (USTC)

Jia Pan (USTC)

Jian-Qing Gao (USTC)

Dan Liu (USTC)

Chin-Hui Lee (GIT)



Jing-Dong Chen (NWPU)

CHiME-5 vs. CHiME-6

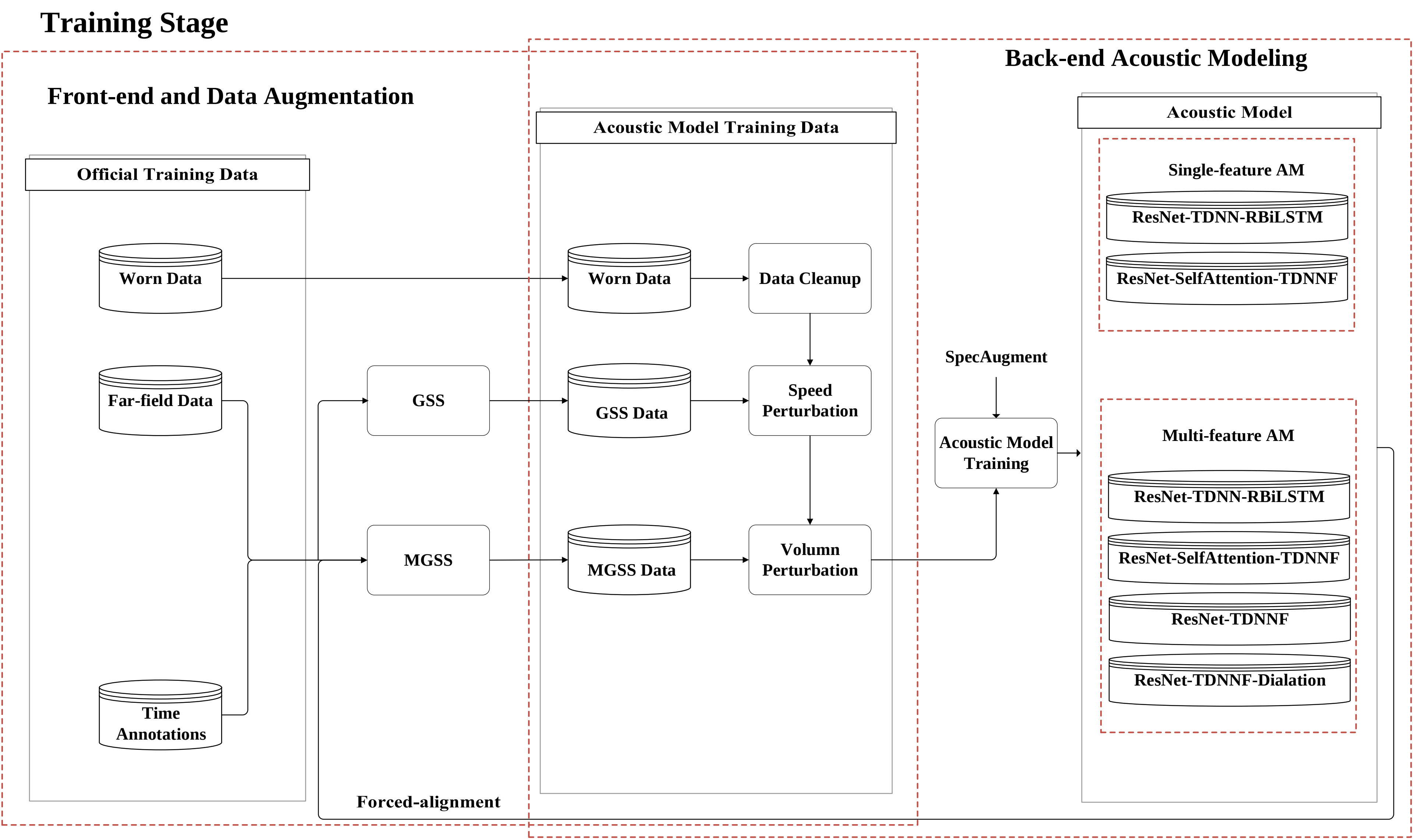
	CHiME-5	CHiME-6
Data/Label Quality	New Array Synchronization	
Front-End	Two-Stage SD-SS[1]	MGSS, BGSS
Acoustic Modeling	BLSTM, CNN-TDNN-LSTM, Improved CLDNN, DFCNNs [2]	ResNet, TDNNF, RBiLSTM, Self-attention, Dialation
System Fusion	State Posterior Average Lattice Fusion [2]	MBR Fusion
Speaker Diarization	N/A	NBSS ResNet Based x-vector Spectral Clustering

[1] Lei Sun, Jun Du, Tian Gao, Yi Fang, Feng Ma, Chin-Hui Lee, "A Speaker-Dependent Approach to Separation of Far-Field Multi-Talker Microphone Array Speech for Front-End Processing in the CHiME-5 Challenge," IEEE Journal of Selected Topics in Signal Processing, 2019, 13(4): 827-840.

[2] Feng Ma, Li Chai, Jun Du, Diyuan Liu, Zhongfu Ye and Chin-Hui Lee, "Acoustic Model Ensembling Using Effective Data Augmentation for CHiME-5 Challenge," INTERSPEECH 2019.

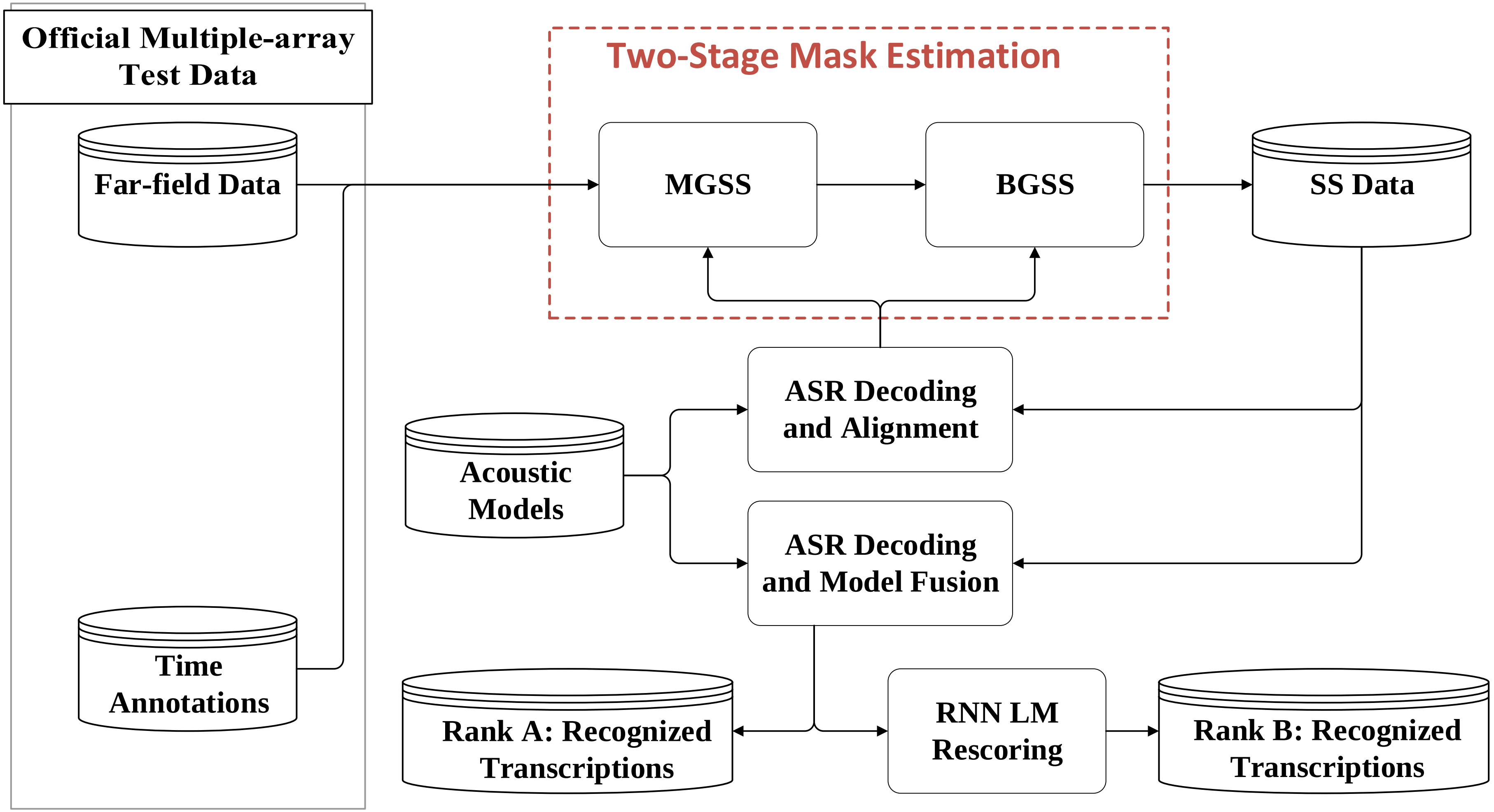
Track 1:
Multiple-array Speech Recognition

System Overview (I)



System Overview (II)

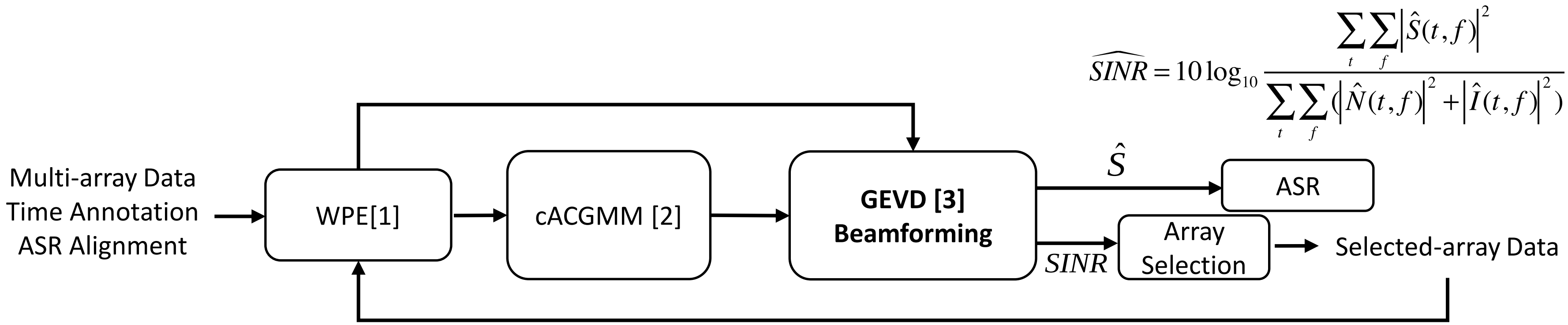
Recognition Stage



Implementation Platform

- The official Kaldi toolkit
  - Guided source separation (GSS)
  - Acoustic models
  - Language models
  - Model ensemble
- The Pytorch toolkit
  - Neural network based speech separation models
- Self-developed toolkit
  - cACGMM
  - Beamforming

Modified GSS (MGSS)



- cACGMM with 5 Gaussian mixtures
  - Corresponding to four speaker sources and one noise source
- The main difference between MGSS and GSS [4]
  - GEVD beamforming: offline + **online** [5]
  - Processing for **selected-array** data based on SINR

[1] L. Drude, J. Heymann, C. Boeddeker, R. Haeb-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," Speech Communication; 13th ITG-Symposium (pp. 1-5) , 2018.

[2] N. Ito, S. Araki , and T. Nakatani . "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing." EUSIPCO, 2016.

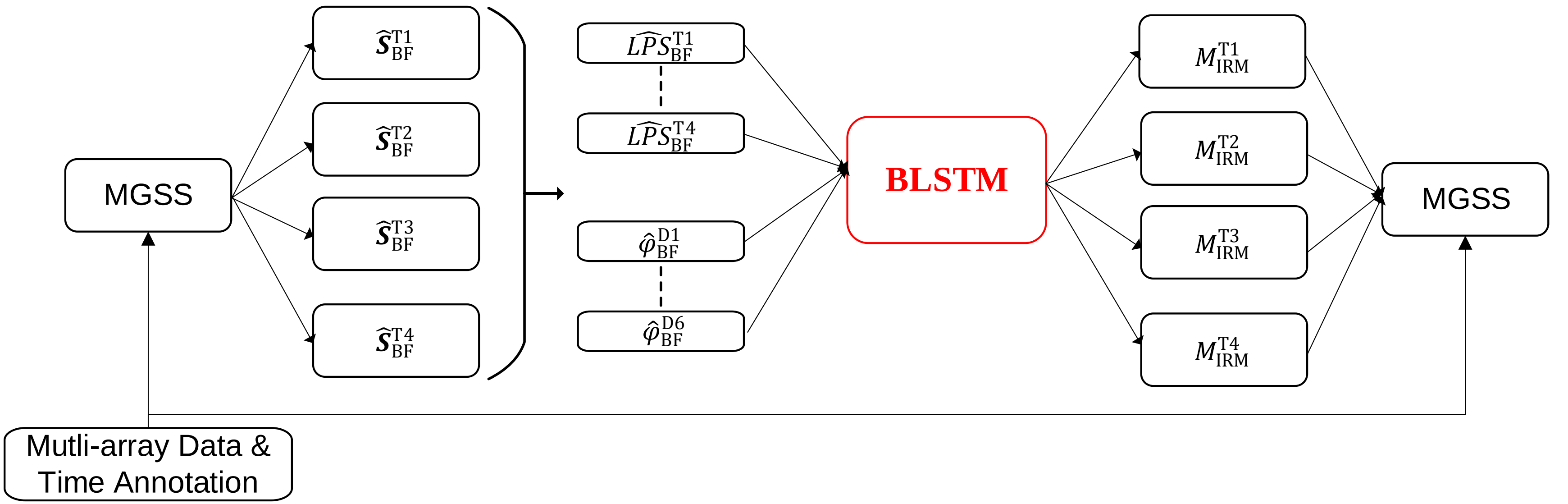
[3] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," IEEE TASLP, vol. 15, no. 5, pp.1529-1539, 2007.

[4] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in CHiME-5 Workshop, Hyderabad, India, 2018.

[5] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," ICASSP, 2016.

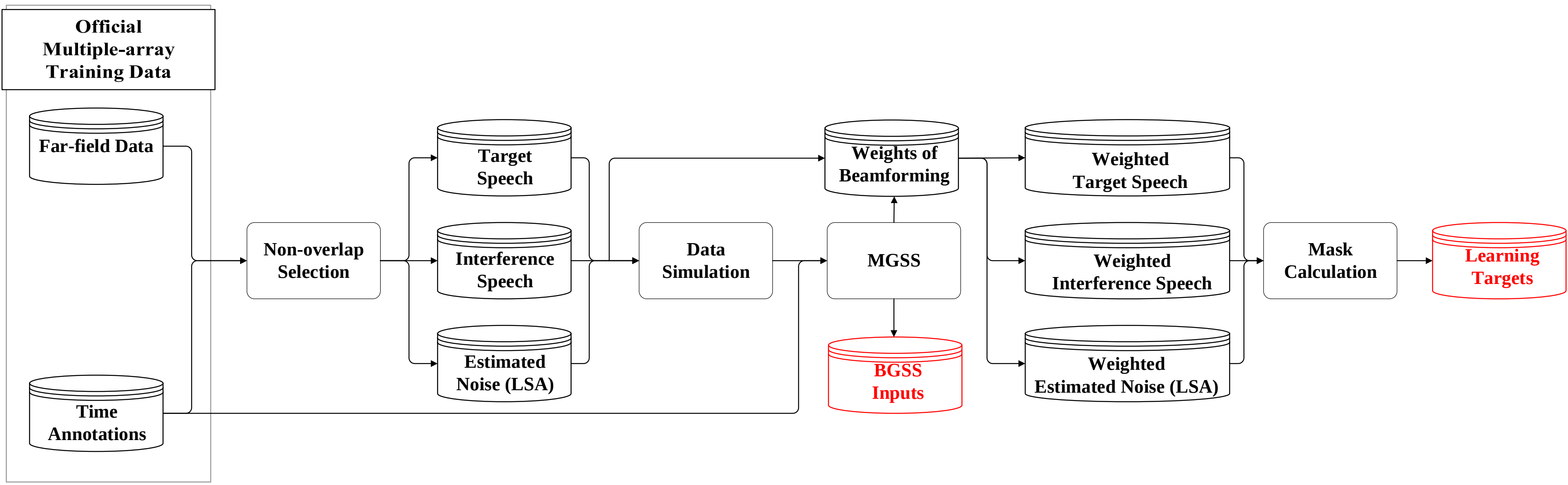
Beamforming GSS (BGSS)

- Motivation: improving the mask estimation of MGSS



$\hat{\boldsymbol{S}}^{T1}_{BF}$ denotes the beamformed STFT features of Target 1.
$\widehat{LPS}^{T1}_{BF}$ denotes the beamformed LPS features of Target 1.
M^{T1}_{IRM} denotes the learning mask of Target 1.
$\hat{\varphi}^{D1}_{BF}$ denotes the IPD (inter-phase difference) between $\hat{\boldsymbol{S}}^{T1}_{BF}$ and $\hat{\boldsymbol{S}}^{T2}_{BF}$

Training Data Generation for BGSS



- The above procedure generates the inputs/outputs for one target speaker
- The four speakers in one session are in turn considered as target speakers

Model Optimization for BGSS

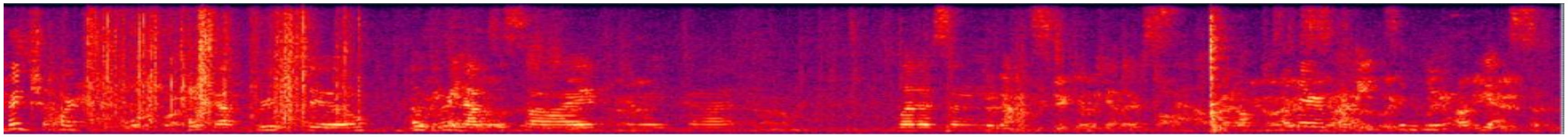
- Architecture
  - BLSTM
  - Input layer: 5130=513*10
  - Hidden layers: 1024*2
  - Output layer： 2052=513*4

- Objective function

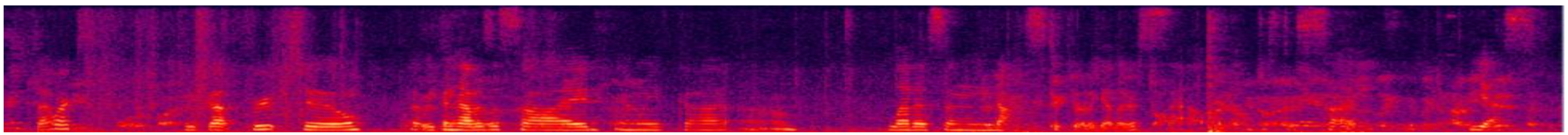
$$Err = (\widehat{M}_{\mathrm{BGSS}}^{\mathrm{T1}} - M_{\mathrm{IRM}}^{\mathrm{T1}})^2 + (\widehat{M}_{\mathrm{BGSS}}^{\mathrm{T2}} - M_{\mathrm{IRM}}^{\mathrm{T2}})^2 + (\widehat{M}_{\mathrm{BGSS}}^{\mathrm{T3}} - M_{\mathrm{IRM}}^{\mathrm{T3}})^2 + (\widehat{M}_{\mathrm{BGSS}}^{\mathrm{T4}} - M_{\mathrm{IRM}}^{\mathrm{T4}})^2$$

Speech Demo

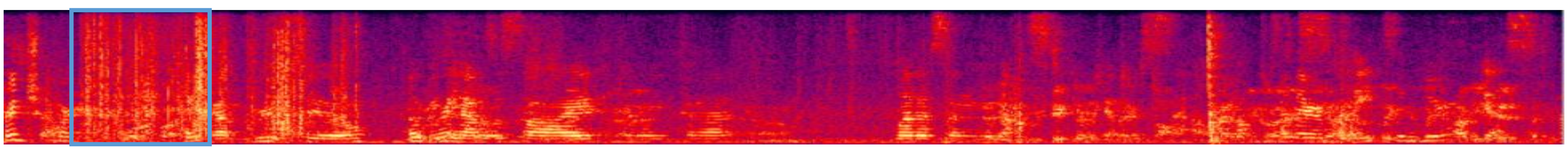
Original, channel-1



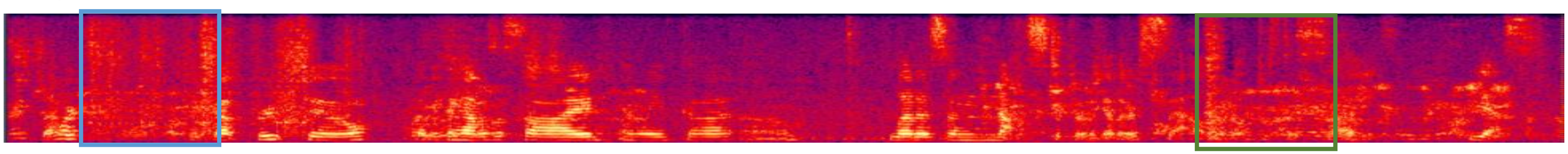
Worn



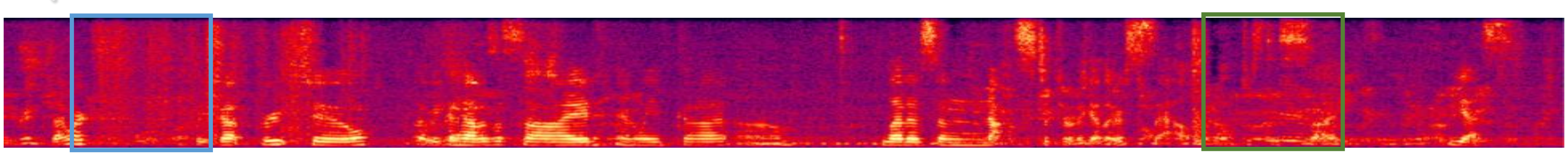
BeamformIt (interfering speaker is still existing)



GSS (Good suppression of interference, residual noises are still existing)

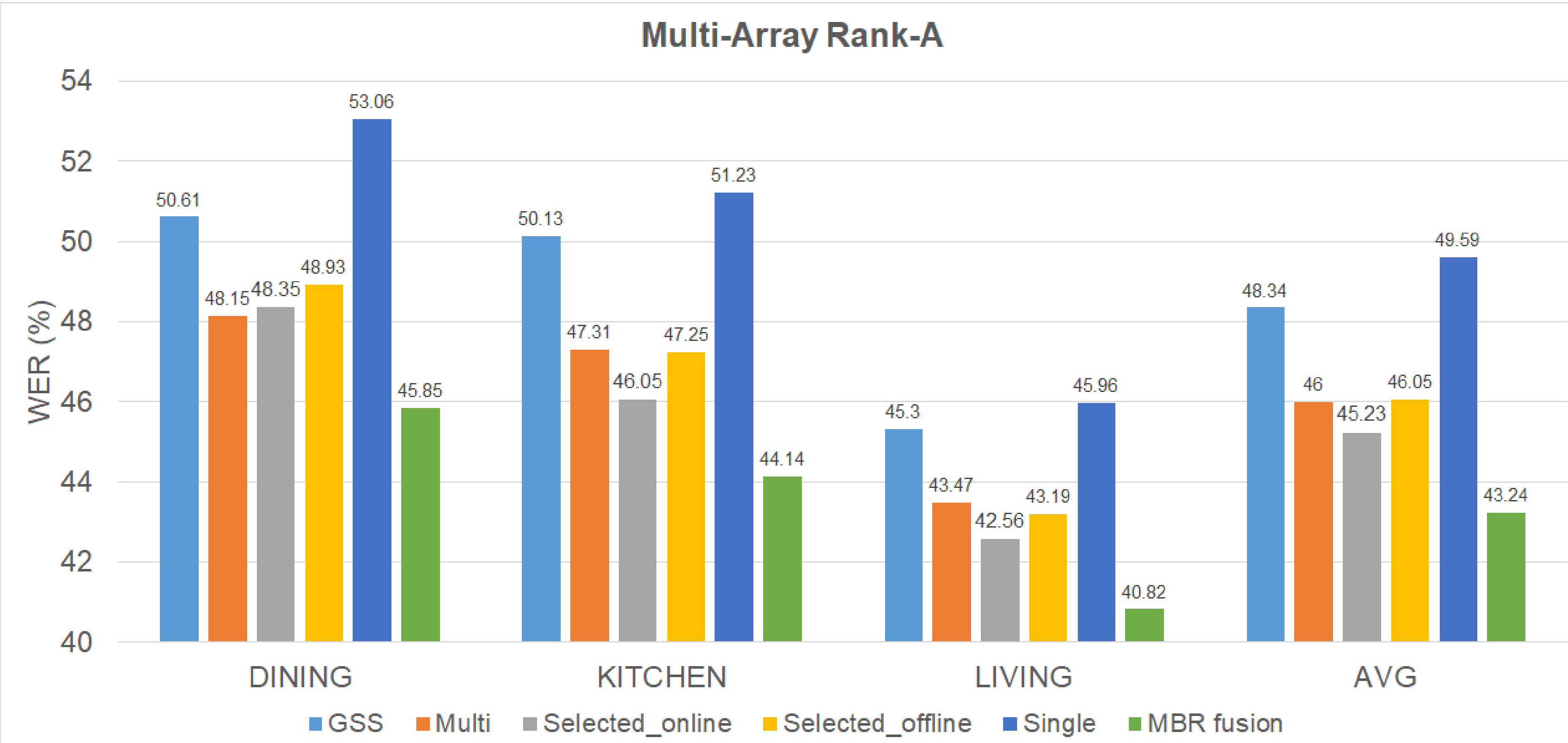


Our MGSS+BGSS (Good suppression of interference, better denoising)



Front-end (GSS vs. Ours)

Results on development sets using the official baseline AM



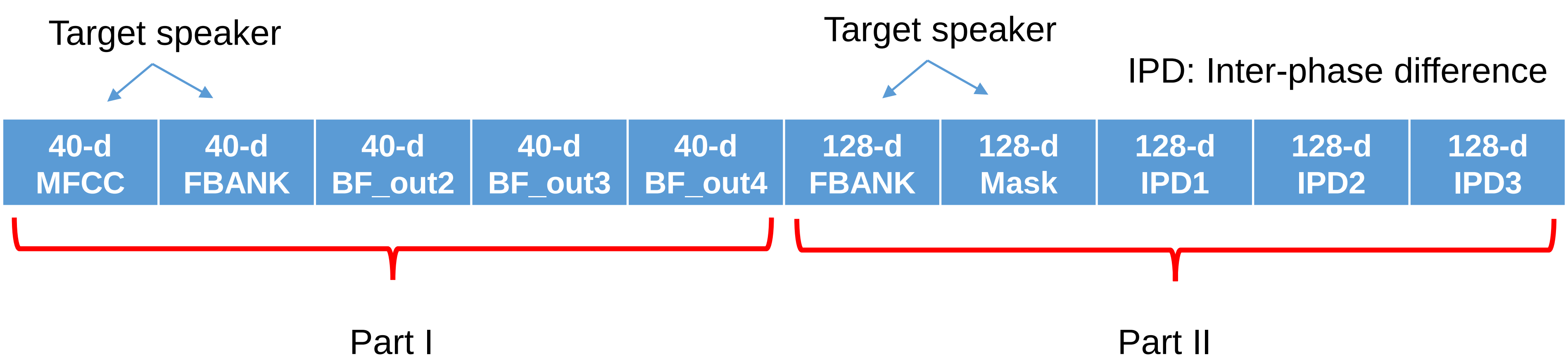
H. Xu, et al. "Minimum Bayes Risk decoding and system combination based on a recursion for edit distance." *Computer Speech & Language* 25.4 (2011): 802-828. (**MBR fusion**)

Acoustic Data Augmentation

- Worn data:
  - Left-channel and right-channel with data cleanup
  - Speed perturbation
  - Data size: (32+32)*3=192 hours
- GSS data:
  - Speed perturbation
  - Data size: 32*3=96 hours
- MGSS data:
  - Multi-array and selected-array
  - Data size: 32+32=64 hours
- Volume perturbation and SpecAugment for all data
- Total training data: 352 hours

Acoustic Models (AMs)

- Single-feature AMs
  - 40-dim MFCC with 100-dim i-vector
- Multi-feature (from 4 speakers) AMs
  - 100-dim i-vector

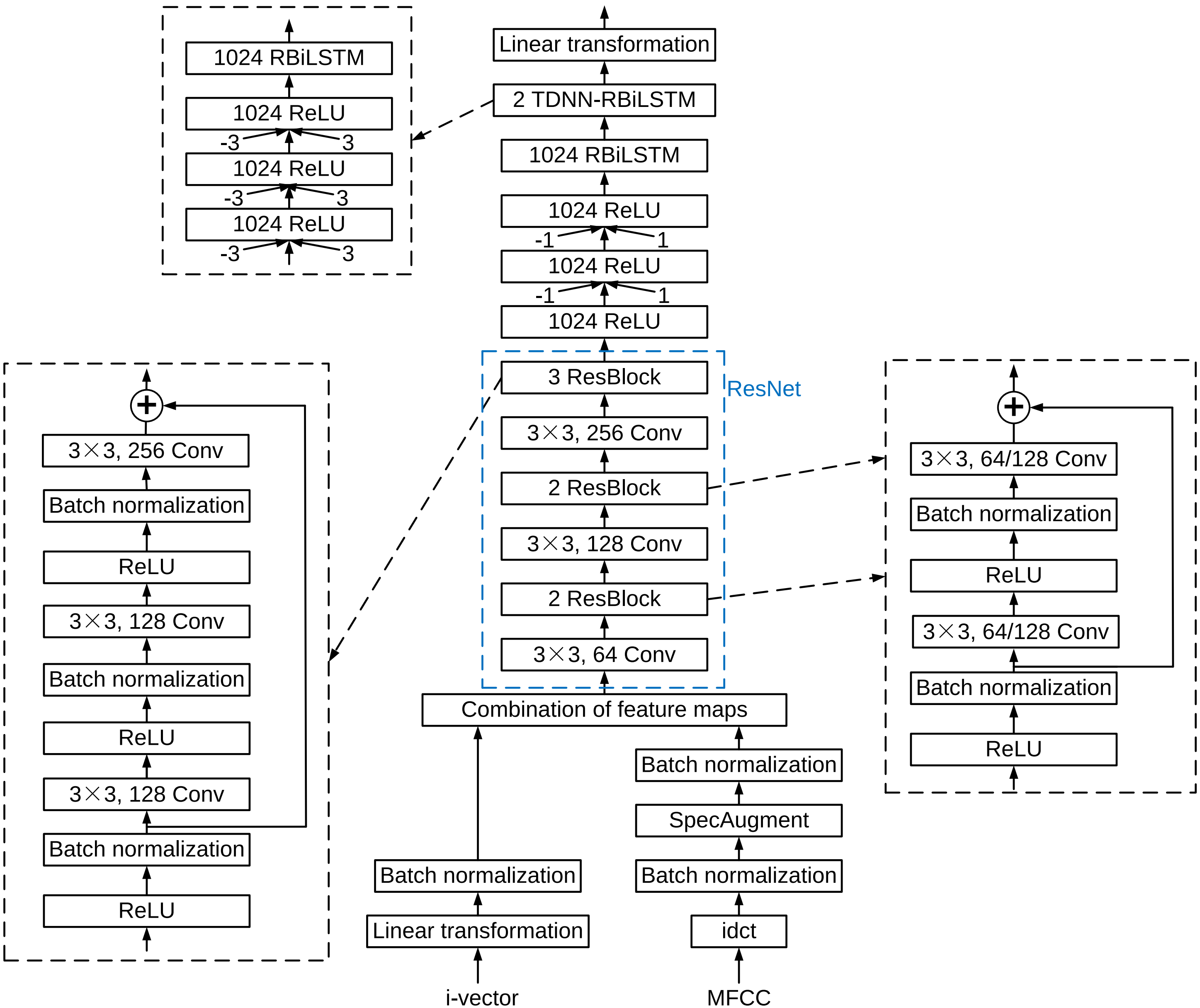


Architecture and Optimization

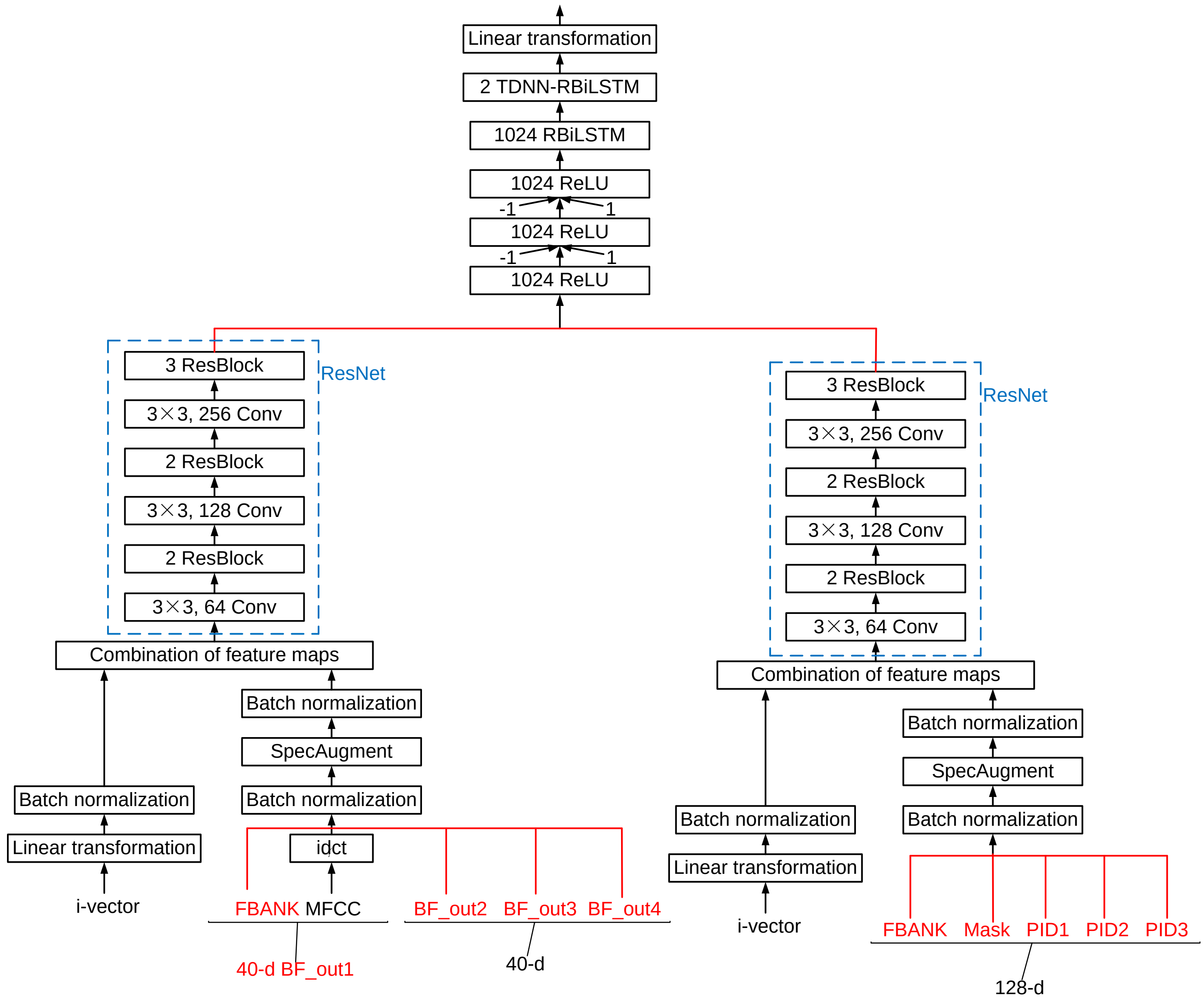
- Four types:
  - ResNet-TDNNF (Multi-feature AM)
  - ResNet-TDNNF-Dialation (Multi-feature AM)
  - ResNet-TDNN-RBiLSTM (Single/Multi-feature AMs)
  - ResNet-SelfAttention-TDNNF (Single/Multi-feature AMs)
- Lattice-Free MMI [1]

[1] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI", INTERSPEECH, 2016, pp.2751-2755.
[2] S. Zagoruyko, N. Komodakis, "Wide residual networks," in BMVC, 2016. (**ResNet**)
[3] N. Kanda, R. Ikeshita, S. Horiguchi, et al. "The Hitachi/JHU CHiME-5 system: Advances in speech recognition for everyday home environments using multiple microphone arrays," CHiME 2018. (**RBiLSTM**)
[4] D. Povey, H. Hadian, P. Ghahremani, et al. "A time-restricted self-attention layer for ASR," ICASSP 2018. (**SelfAttention**)
[5] D. Povey, G. Cheng, Y. Wang, et al. "Semi-orthogonal low-rank matrix factorization for deep neural networks," INTERSPEECH, 2018. (**TDNNF**)

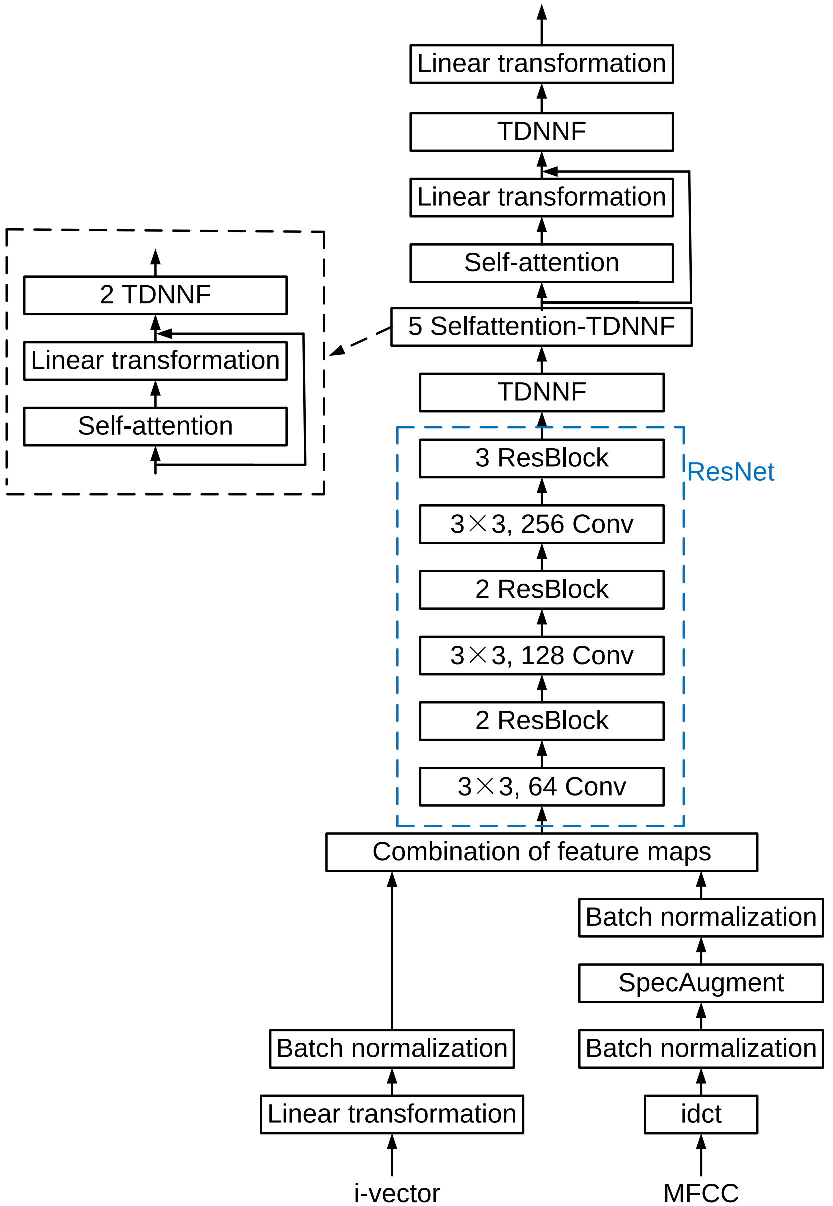
ResNet-TDNN-RBiLSTM (Single-feature)



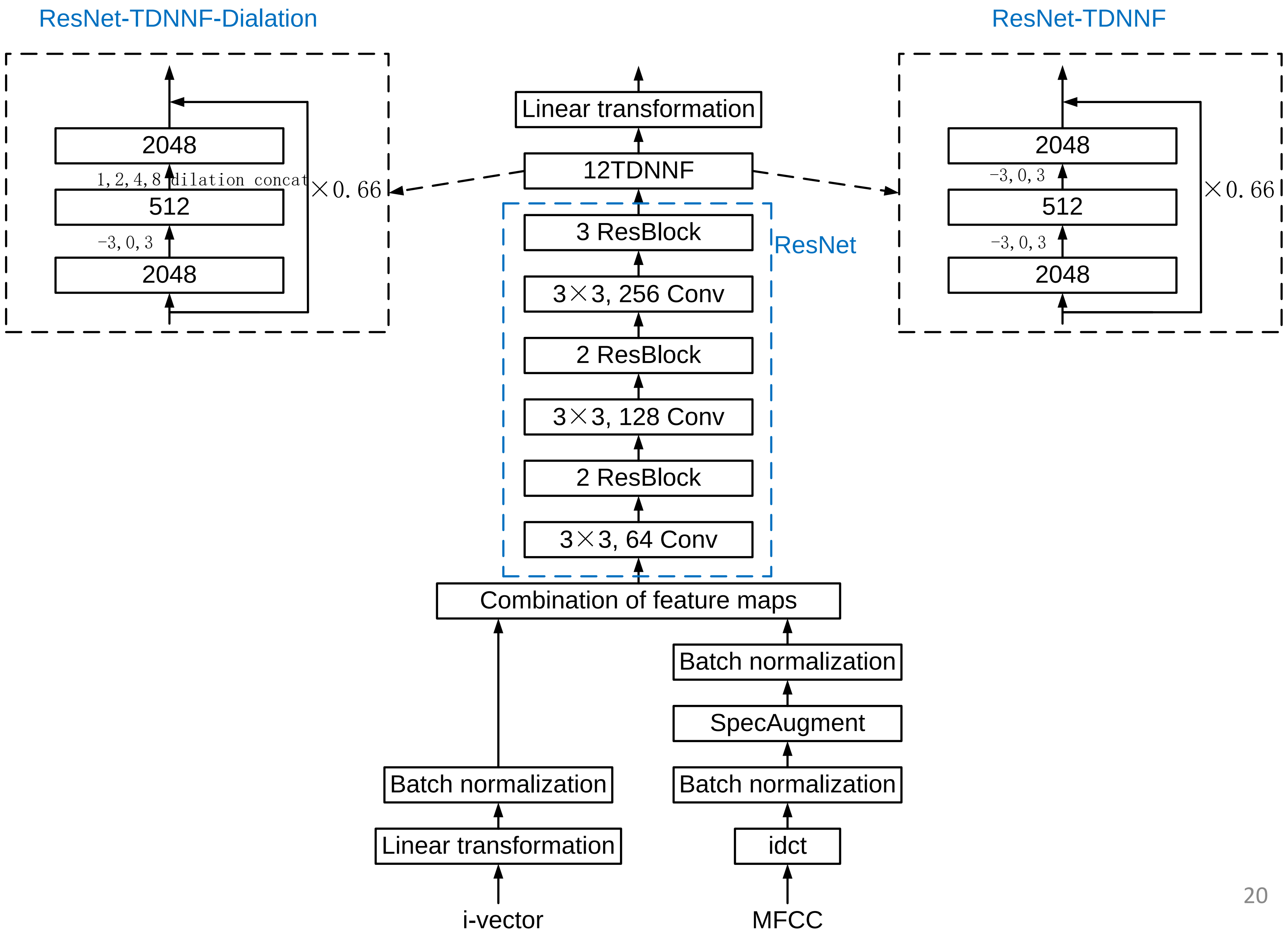
Resnet-TDNN-RBiLSTM (Multi-feature)



ResNet-SelfAttention-TDNNF



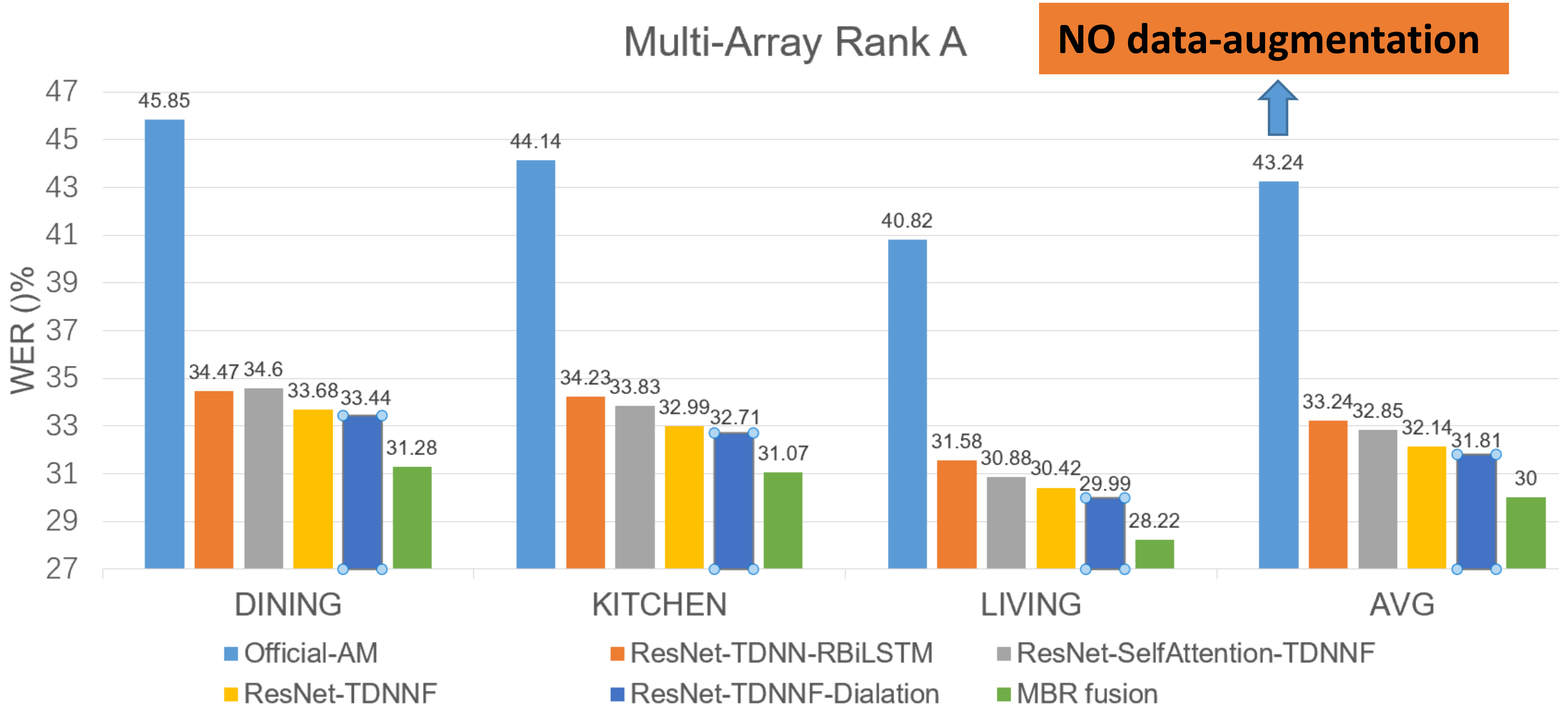
ResNet-TDNNF-Dialation & ResNet-TDNNF



AMs with Our Best Front-end

- Four multi-feature AMs with different architectures
- MBR fusion of 4 multi-feature AMs and 2 single-feature AMs

Results on development sets

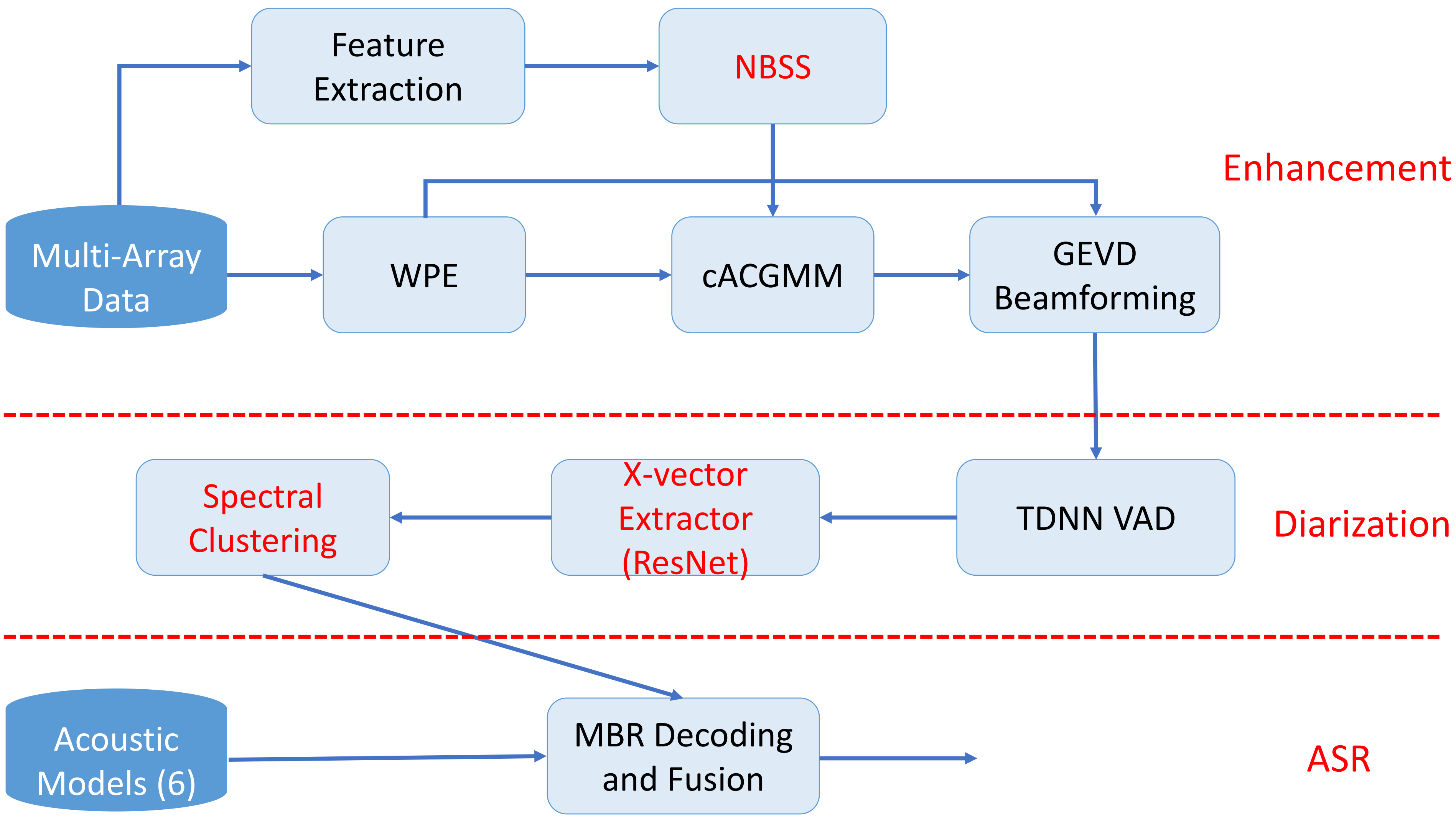


Submitted Results

Category	Session		Dining	Kitchen	Living	Overall
A	Dev	S02	34.95	35.13	29.77	31.11
		S09	29.60	28.10	27.65	
	Eval	S01	25.55	42.75	38.13	30.96
		S21	25.49	34.97	26.05	
B	Dev	S02	34.66	34.86	29.50	30.77
		S09	29.10	27.74	27.22	
	Eval	S01	25.01	42.66	37.44	30.50
		S21	25.14	34.84	25.34	

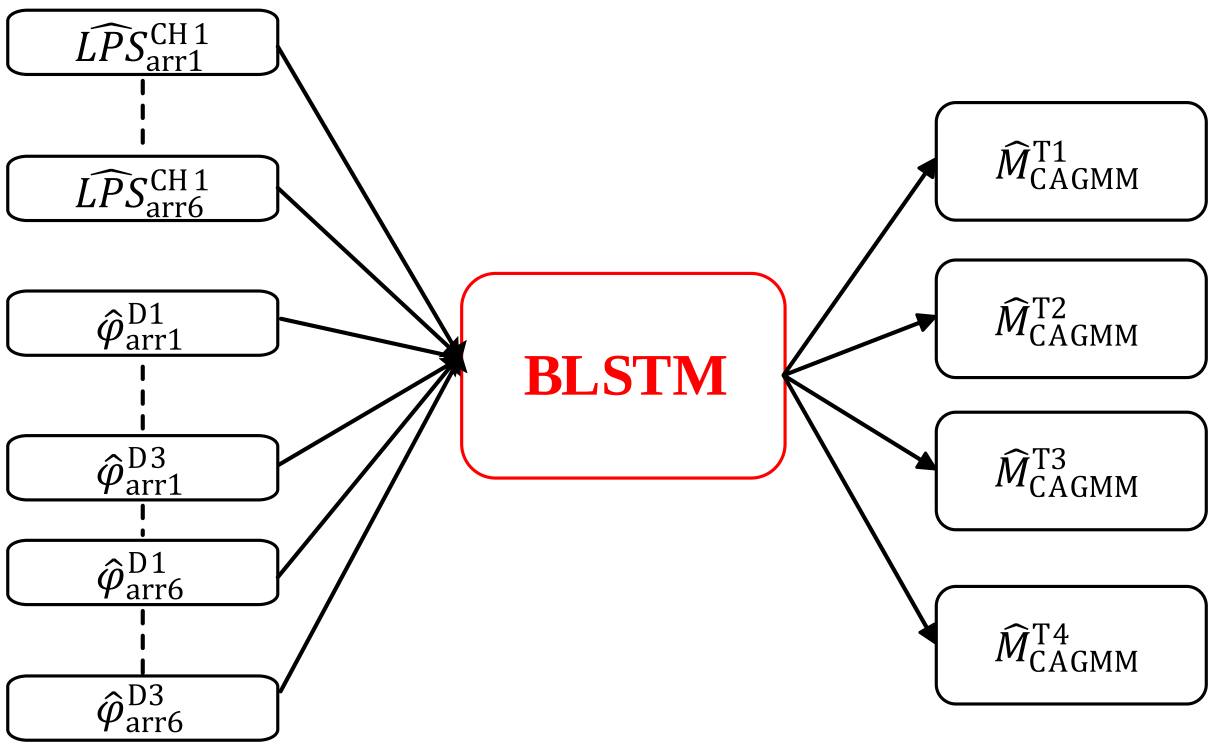
Track 2:
Multiple-array Diarization and Recognition

System Overview

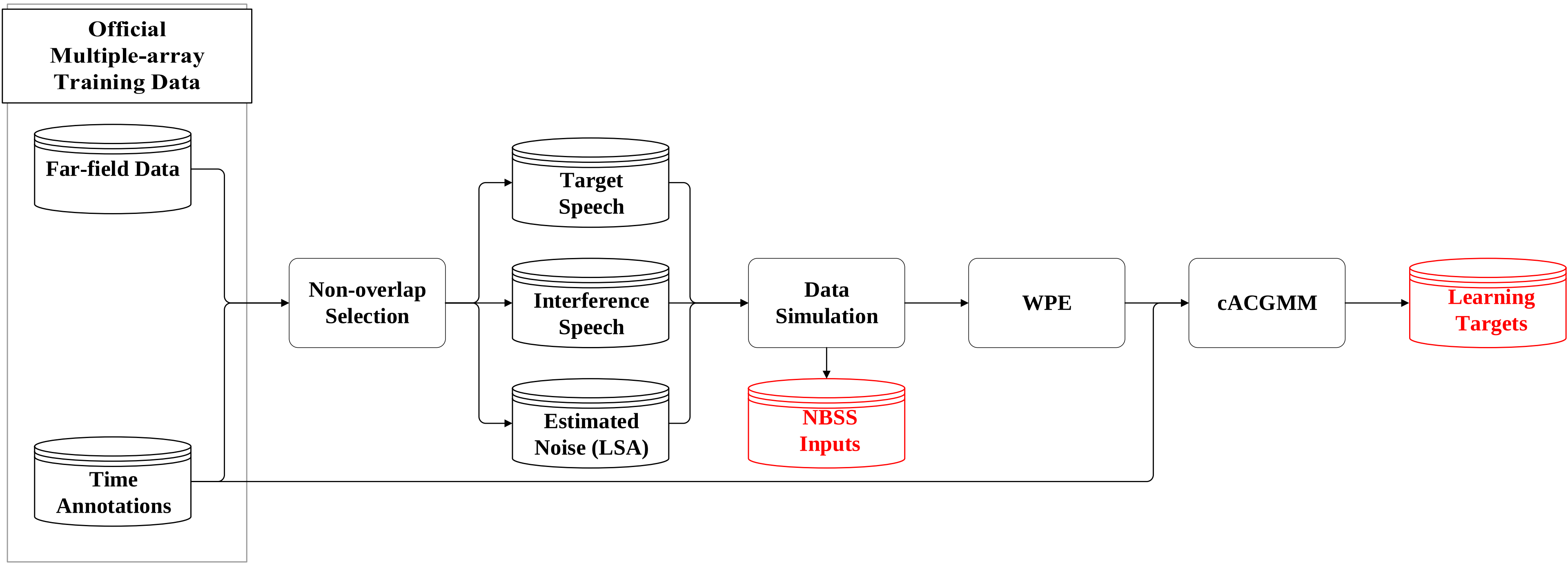


Neural Beamforming for SS (NBSS)

- Mask estimation for cACGMM with 5 Gaussian mixtures
- The “target” selection based on speech duration of beamformed data



Training Data Generation for NBSS



Speaker Diarization

- ResNet based x-vector extractor [1]

Layer name	Structure	Output
Input	–	$40 \times 200 \times 1$
Conv2D-1	3×3, Stride 1	$40 \times 200 \times 32$
ResNetBlock-1	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 3$, Stride 1	$40 \times 200 \times 32$
ResNetBlock-2	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 4$, Stride 2	$20 \times 100 \times 64$
ResNetBlock-3	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 6$, Stride 2	$10 \times 50 \times 128$
ResNetBlock-4	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 3$, Stride 2	$5 \times 25 \times 256$
StatsPooling	–	5×256
Flatten	–	2560
Dense1	–	256
Dense2 (Softmax)	–	N

- Spectral clustering [2]

[1] A. Gusev, et al. "Deep Speaker Embeddings for Far-Field Speaker Recognition on Short Utterances", arXiv:2002.06033v1.
[2] T. J. Park, et al. “Auto-Tuning Spectral Clustering for Speaker Diarization Using Normalized Maximum Eigengap”, arXiv:2003.02405v1

Submitted Results

Category	Development set			Evaluation set		
	DERs	JERs	WER	DERs	JERs	WER
A	56.69	58.49	68.22	65.37	64.15	68.48
B	56.69	58.49	68.15	65.37	64.15	68.42

Thanks
Q&A