The 5th International Workshop on Speech Processing in Everyday Environments

The USTC-iFlytek System for CHiME-5 Challenge

Jun Du

2018.09.07

@Hyderabad, India

Team

Jun Du (USTC)

Tian Gao (USTC)

Lei Sun (USTC)

Chin-Hui Lee (GIT)

Feng Ma (iFlytek)

Yi Fang (iFlytek)

Di-Yuan Liu (iFlytek)

Qiang Zhang (iFlytek)

Xiang Zhang (iFlytek)

Hai-Kun Wang (iFlytek)

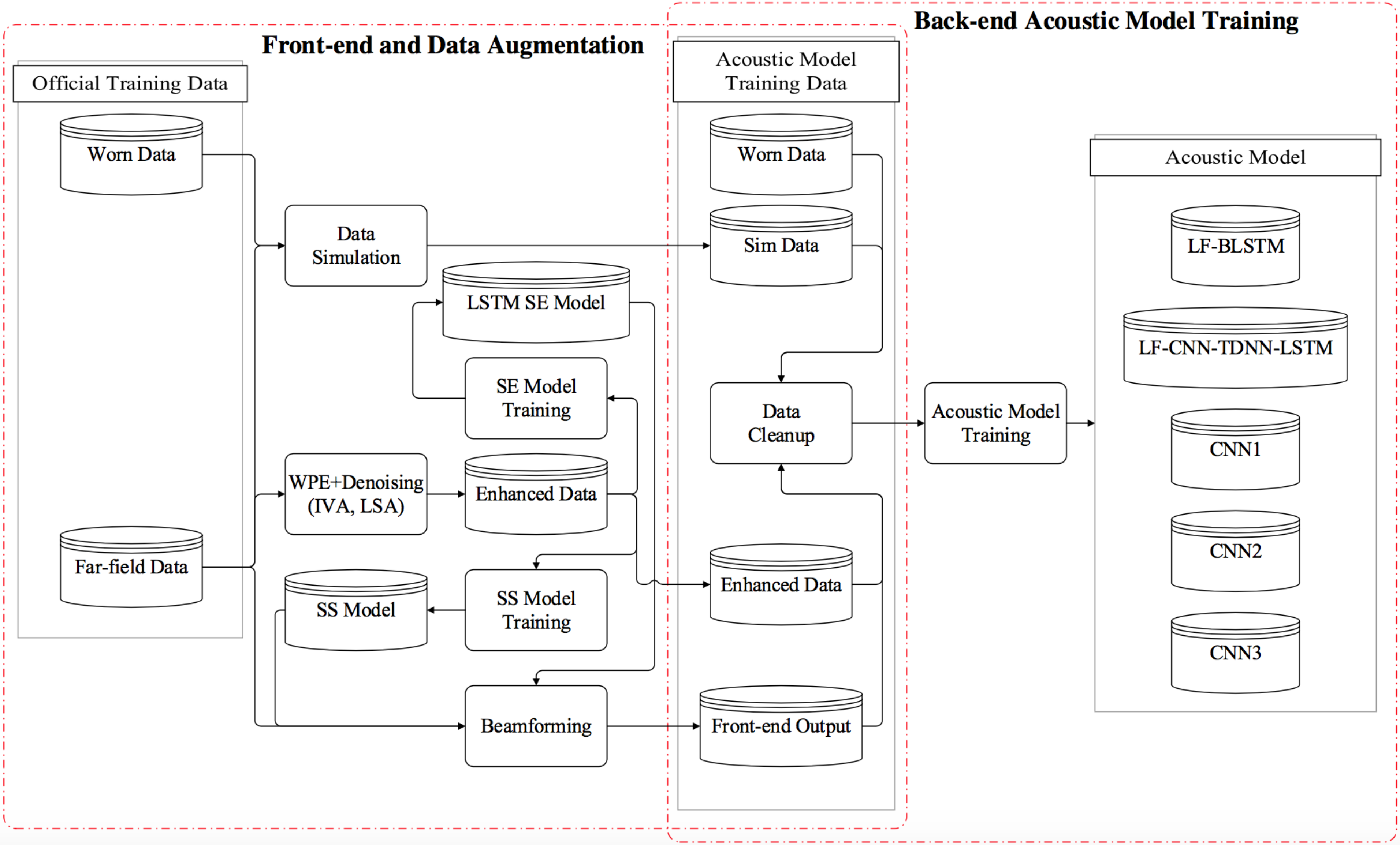
Jia Pan (iFlytek)

Jian-Qing Gao (iFlytek)

Jing-Dong Chen (NWPU)

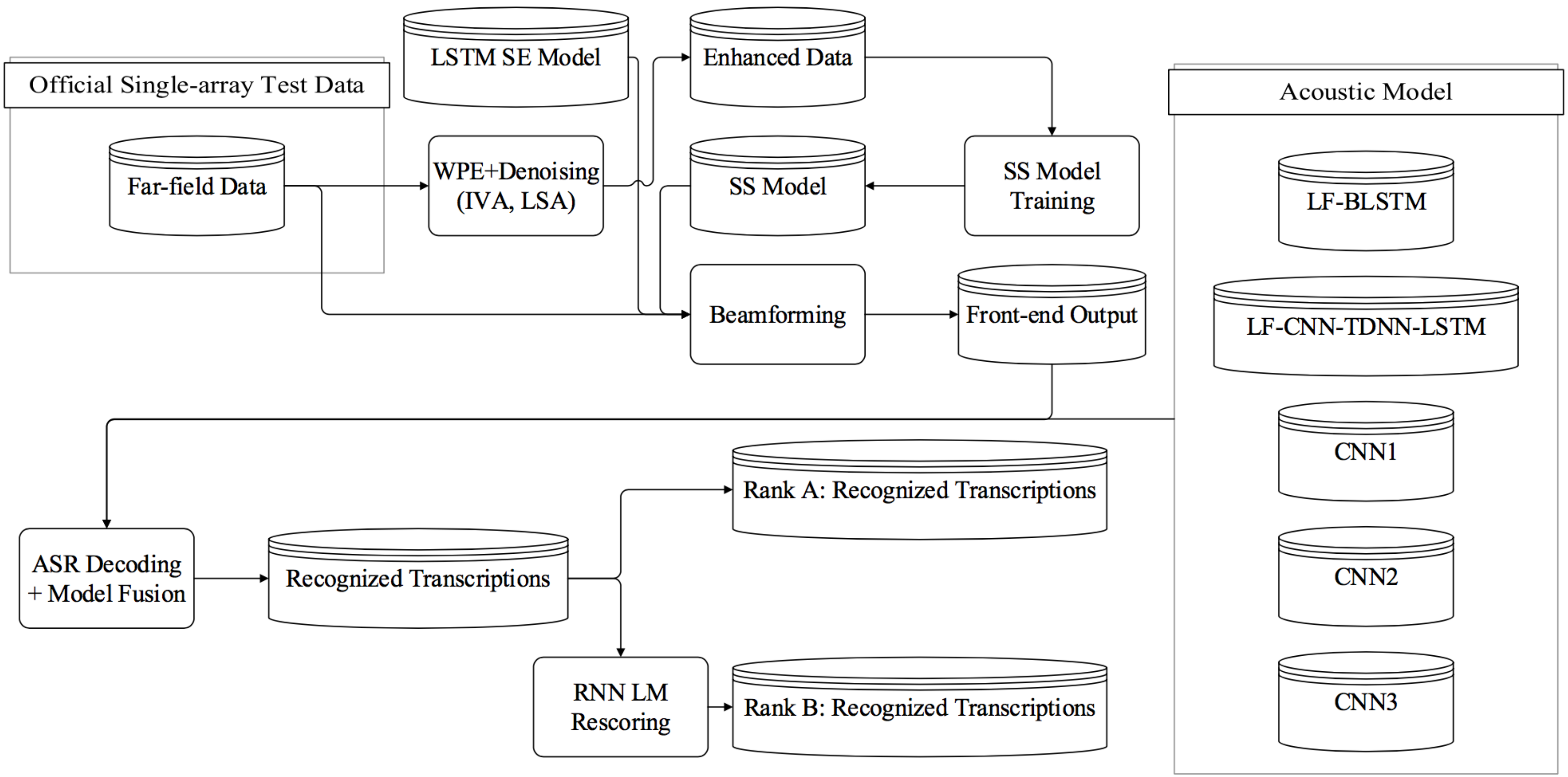
System Overview (I)

Training Stage



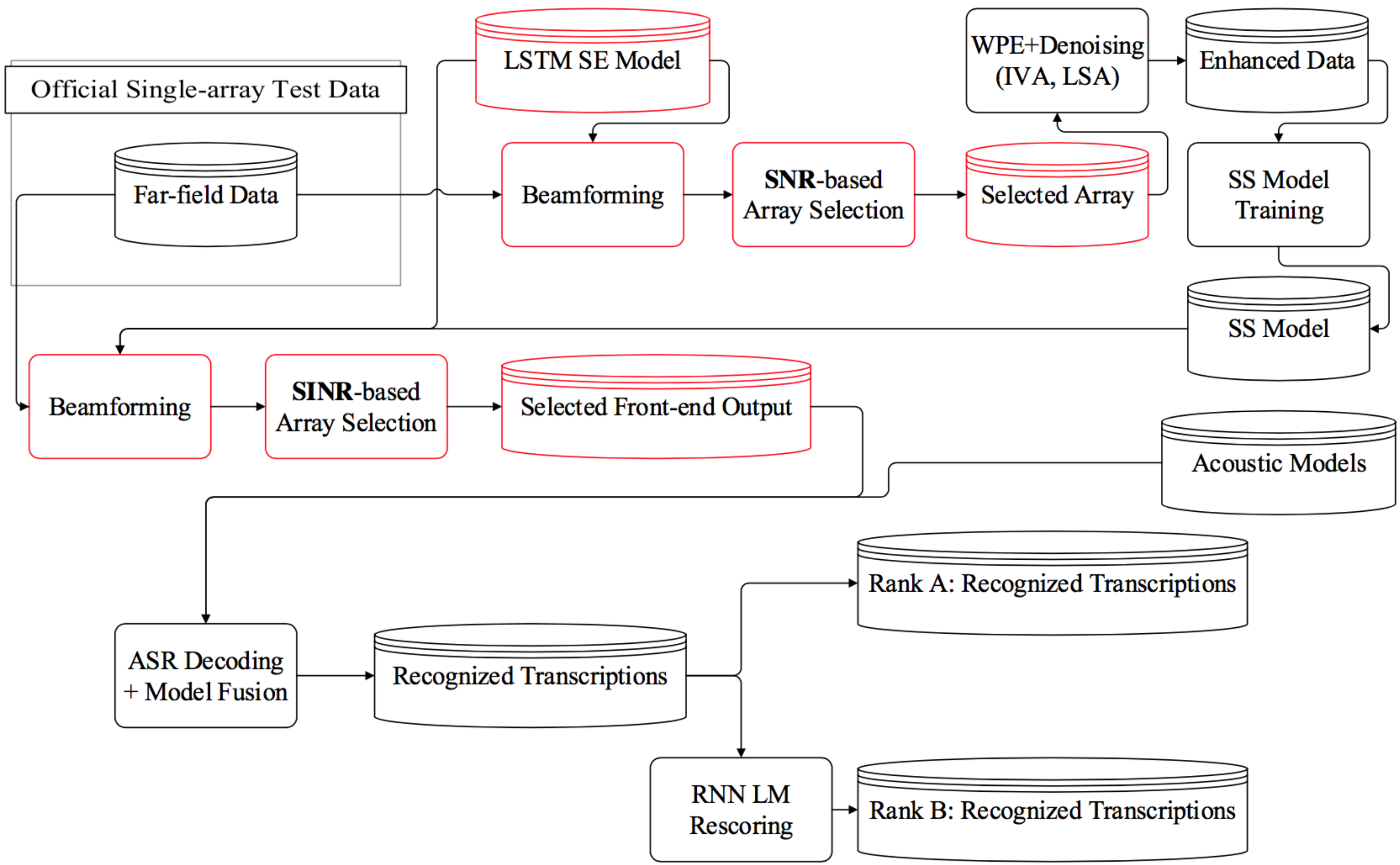
System Overview (II)

Single-array Track Testing Stage



System Overview (III)

Multiple-array Track Testing Stage

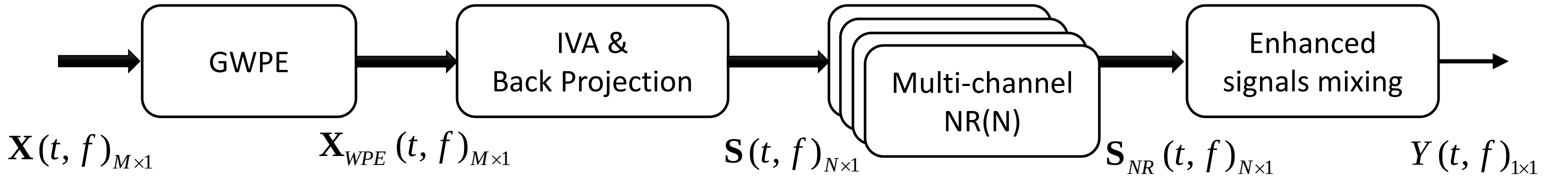


Implementation Platform

- The official Kaldi toolkit
  - Features: MFCC features
  - GMM-HMM acoustic model
  - LF-BLSTM HMM acoustic model
  - LF-CNN-TDNN-LSTM HMM acoustic model
  - Model ensemble
- The CNTK toolkit
  - LSTM-based single-channel speech separation models
  - LSTM-based single-channel speech enhancement models
- Self-developed toolkit
  - Beamforming
  - CNN-HMM acoustic models
  - LSTM language models

WPE + Denoising (IVA, LSA)

- Blind background noise reduction as preprocessing
  - Important to make the subsequent separation/beamforming working



$$Y(t,f) = \left[G_1, G_2 \ldots G_N(t,f)\right] \mathbf{W}_{IVA-BP}(f)\mathbf{X}_{WPE}(t,f)$$

- Step 1: Generalized Weighted Prediction Error (GWPE) [1]
- Step 2: Independent Vector Analysis & Back Projection (IVA-BP, N=M=4) [2]
- Step 3: Multichannel noise reduction using log-spectral amplitude (LSA) [3]

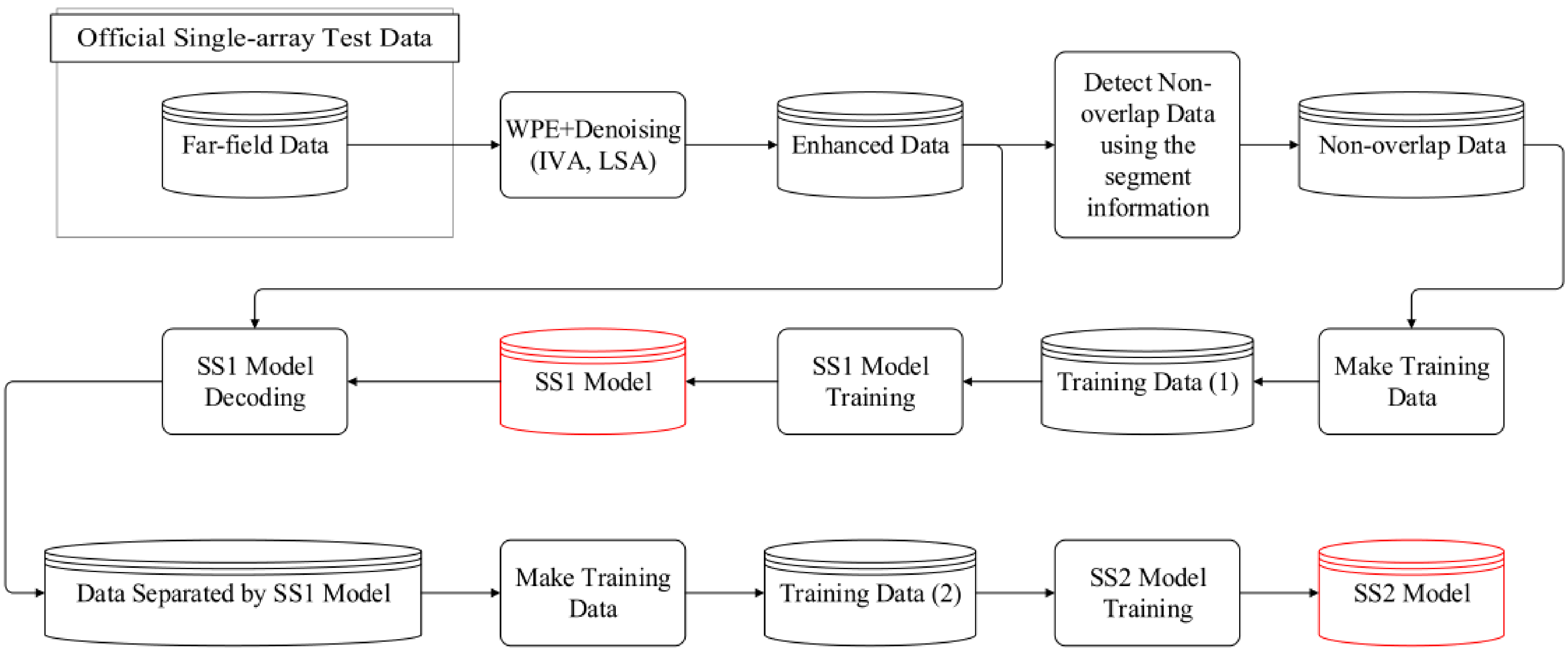
[1] T. Yoshioka and T. Nakatani , “Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening “, IEEE TASLP, vol. 20, no. 10, pp.2707-2720, 2012.

[2] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique”, IEEE WASPAA, 2011, pp.189-192.

[3] I. Cohen, “Multichannel post-filtering in non-stationary noise environments,” IEEE TSP, vol. 52, no. 5, pp.1149-1160, 2004.

Single-channel speech separation

Speaker-dependent Two-Stage Neural Network Based Speech Separation



SS1

- Motivation of SS1
  - Using non-overlap data from oracle diarization information
  - Problems of “non-overlap” data
    - Insufficient
    - Not pure
  - More pure non-overlap data is necessary
    - SS1 aiming at removing interference clearly with potential target distortions
    - Separated target data by SS1 as the new non-overlap data
- Objective function of SS1

$$Err = (\log(\widehat{IRM}) + \log(|Y|^2) - \log(|X|^2))^2$$

Y denotes input feature, X denotes learning target, IRM denotes network output

SS2

- Motivation of SS2
  - Large speech distortions of target introduced by SS1 models
  - Aiming at better speech preservation for ASR task
- Target training data of SS2
  - Using the separated target data by SS1
  - More target data for SS2 training than that for SS1 training
- Objective function of SS2

$$Err = (\widehat{IRM} - IRM)^2$$

Setup of SS1 and SS2

- Neural network architecture
  - 2-layer BLSTM network
  - 512 cells per LSTM layer
- Input features
  - Log-power spectral features
  - Frame-length: 32ms
  - Frame-shift: 16ms
- Training data for each speaker-dependent model
  - Interfering speakers: other 3 speakers
  - Simulated mixing data size: about 50 hours
  - Input SNR: -5dB, 0dB, 5dB, 10dB

Single-channel speech enhancement

- Densely connected progressive learning for LSTM [1]
  - Target data (or “clean data”)
    - 40-hour preprocessed data by WPE+Denoising
  - Noise data
    - unlabeled segments of channel-1 in training sets filtered using ASR model
  - Input SNR of training data: -5 dB, 0 dB, 5 dB
  - Simulated training data size: 120 hours
  - Architecture: the best configuration in [1]
  - Objective function of the output layer in progressive learning

$$Err = (\widehat{IRM} - IRM)^2$$

- Testing stage: channel-1 as the input

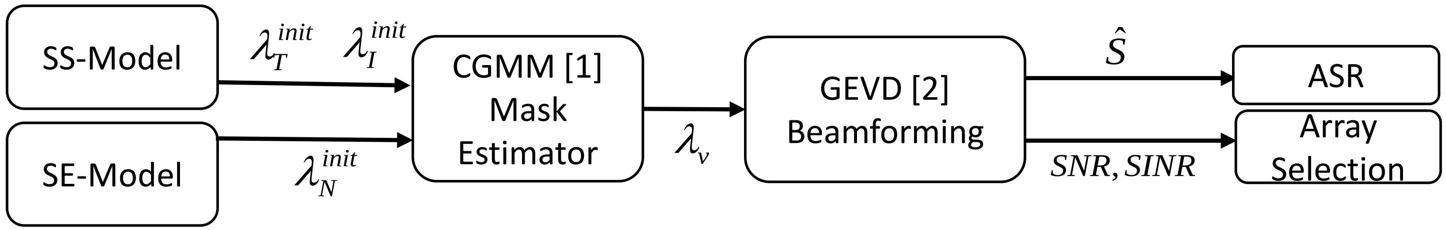
[1] T. Gao, J. Du, L.-R. Dai and C.-H. Lee, “Densely connected progressive learning for LSTM-based speech enhancement,” ICASSP 2018.

Beamforming

$$\widehat{SNR} = 10\log_{10}\frac{\sum_t\sum_f\left|\hat{S}(t,f)\right|^2}{\sum_t\sum_f\left|\hat{N}(t,f)\right|^2}$$

$$\widehat{SINR} = 10\log_{10}\frac{\sum_t\sum_f\left|\hat{S}(t,f)\right|^2}{\sum_t\sum_f\left(\left|\hat{N}(t,f)\right|^2+\left|\hat{I}(t,f)\right|^2\right)}$$

- DL-CGMM-GEVD source separation



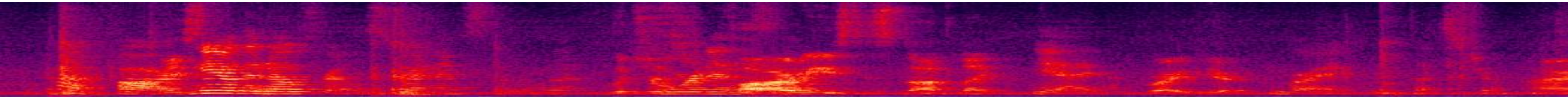
  - $\lambda_v(t, f)$: the posterior probability of TF bin belong to source v
  - Three sources: target speech, interfering speech, background noise
  - Using mask outputs of SS/SE deep models to initialize CGMM parameters
    - Extension of CGMM in [1] from 2 Gaussian mixtures to 3 Gaussian mixtures
    - Well addressing the source order permutation problem
  - Two-pass array selection for multi-array track
    - Selecting 3 arrays using SNR for SS model training (1-pass) and SINR for ASR tasks (2-pass)
    - Fusing the recognition results of time-aligned 3 arrays via acoustic model ensemble

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

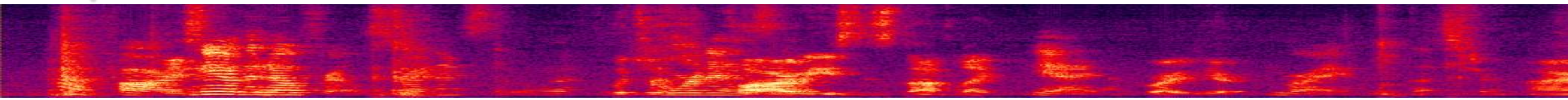
[2] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," IEEE TASLP, vol. 15, no. 5, pp.1529-1539, 2007.

Speech Demo

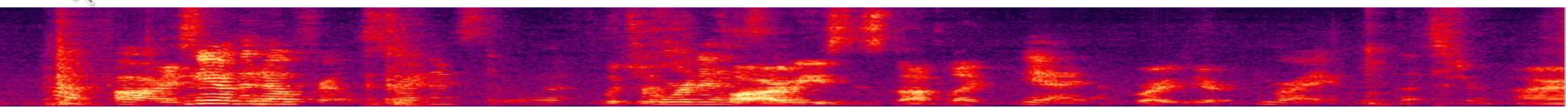
Original, channel-1



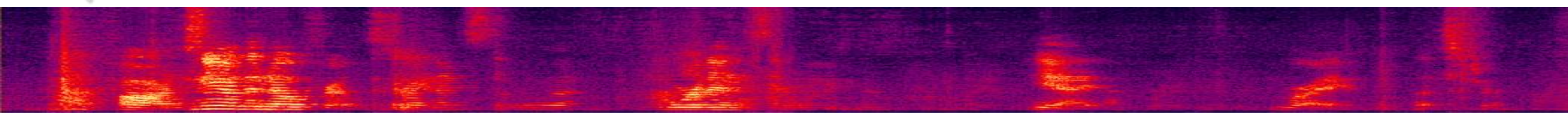
Official Beamforming (interfering male speaker is still existing)



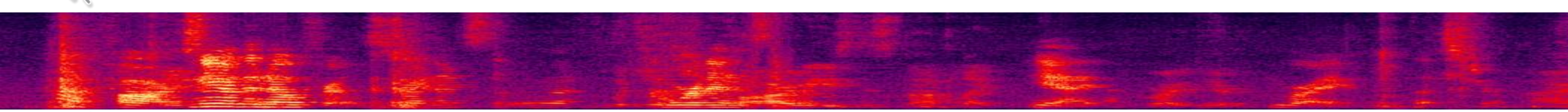
WPE, IVA, LSA (dereverberation and denoising as preprocessing)



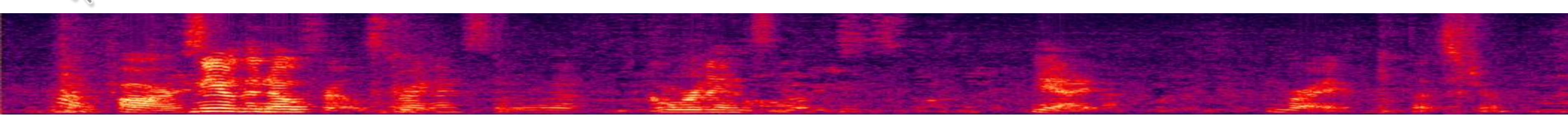
Single-channel SS1 (good suppression of interference, large distortions of target)



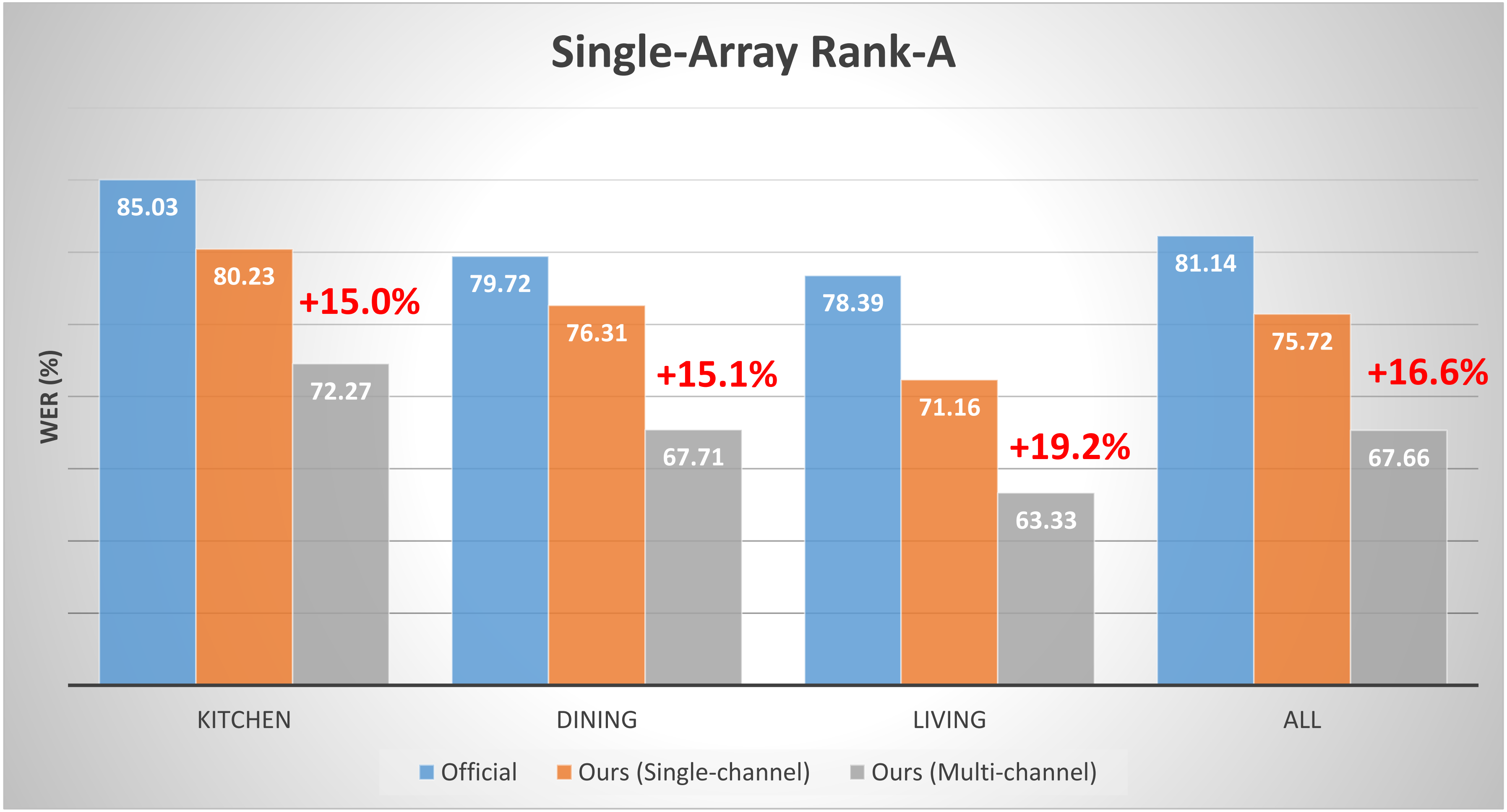
Single-channel SS2 (worse suppression of interference, smaller distortions of target)



Beamforming with SS2 (the best trade-off)



Front-end (Official vs. Ours)



Results on development sets using the official baseline LF-TDNN system

Acoustic Data Augmentation

- Worn data:
  - Left-channel + right-channel
  - Data cleanup (as used in baseline system)
  - Data size: 64 hours
- Far-field data:
  - Preprocessed data (WPE+IVA+LSA) of all arrays
  - Data cleanup
  - Data size: 110 hours + 110 hours (after front-end)
- Simulated far-field Data:
  - Calculating 1000+ RIRs using the recording pairs of worn and far-field
  - Using RIRs and noise segments to simulate far-field data from worn data
  - Data size: 250 hours
- Total training data: 534 hours

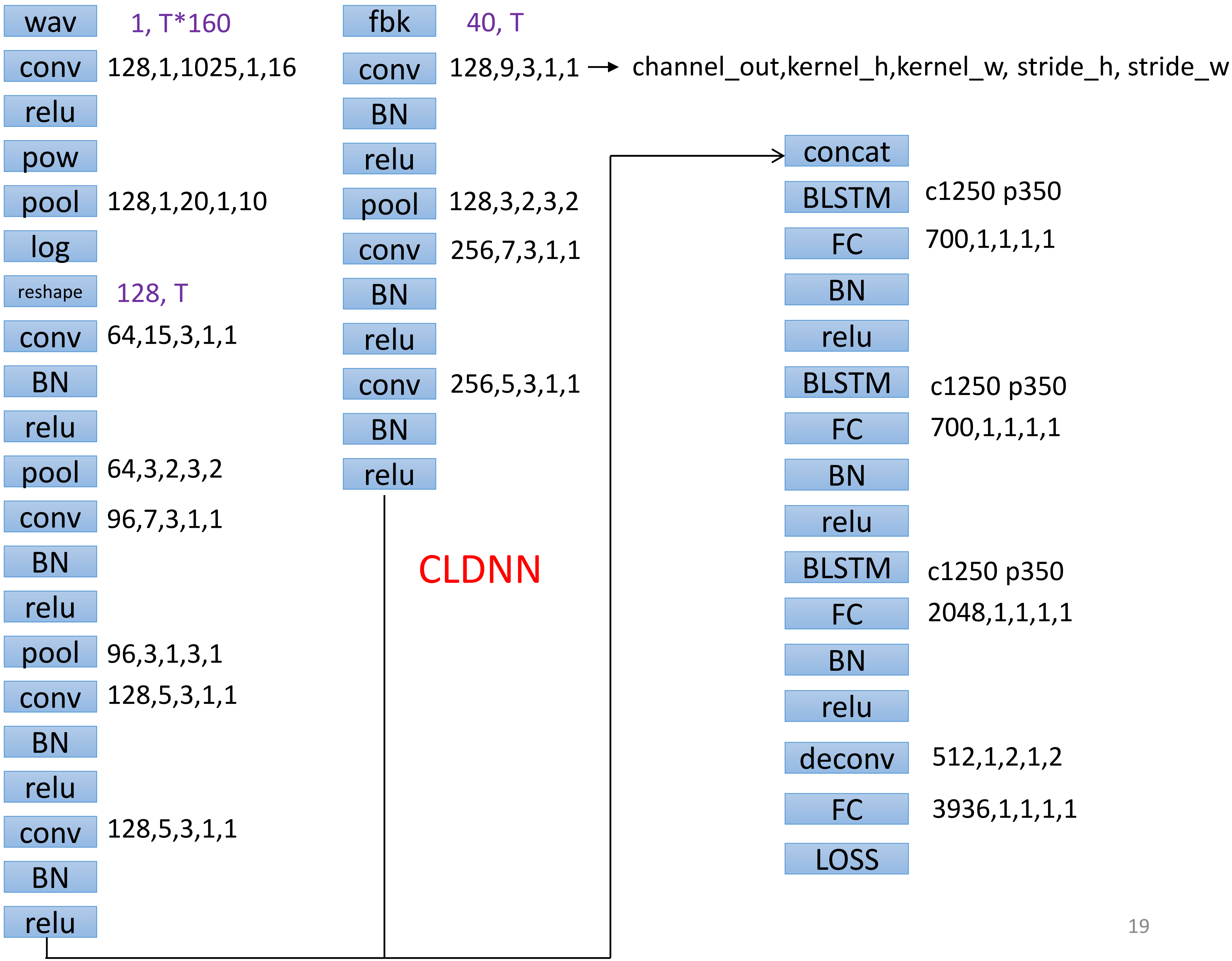
Lattice-Free MMI [1] Based AMs

- LF-BLSTM
  - 5-layer BLSTM network
  - 40-d MFCC
  - 100-d i-vector
- LF-CNN-TDNN-LSTM
  - 2-layer CNN + 9-layer TDNN + 3-layer LSTM network
  - 40-d MFCC
  - 100-d i-vector

[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", in *Proc. Interspeech*, 2016, pp.2751-2755.

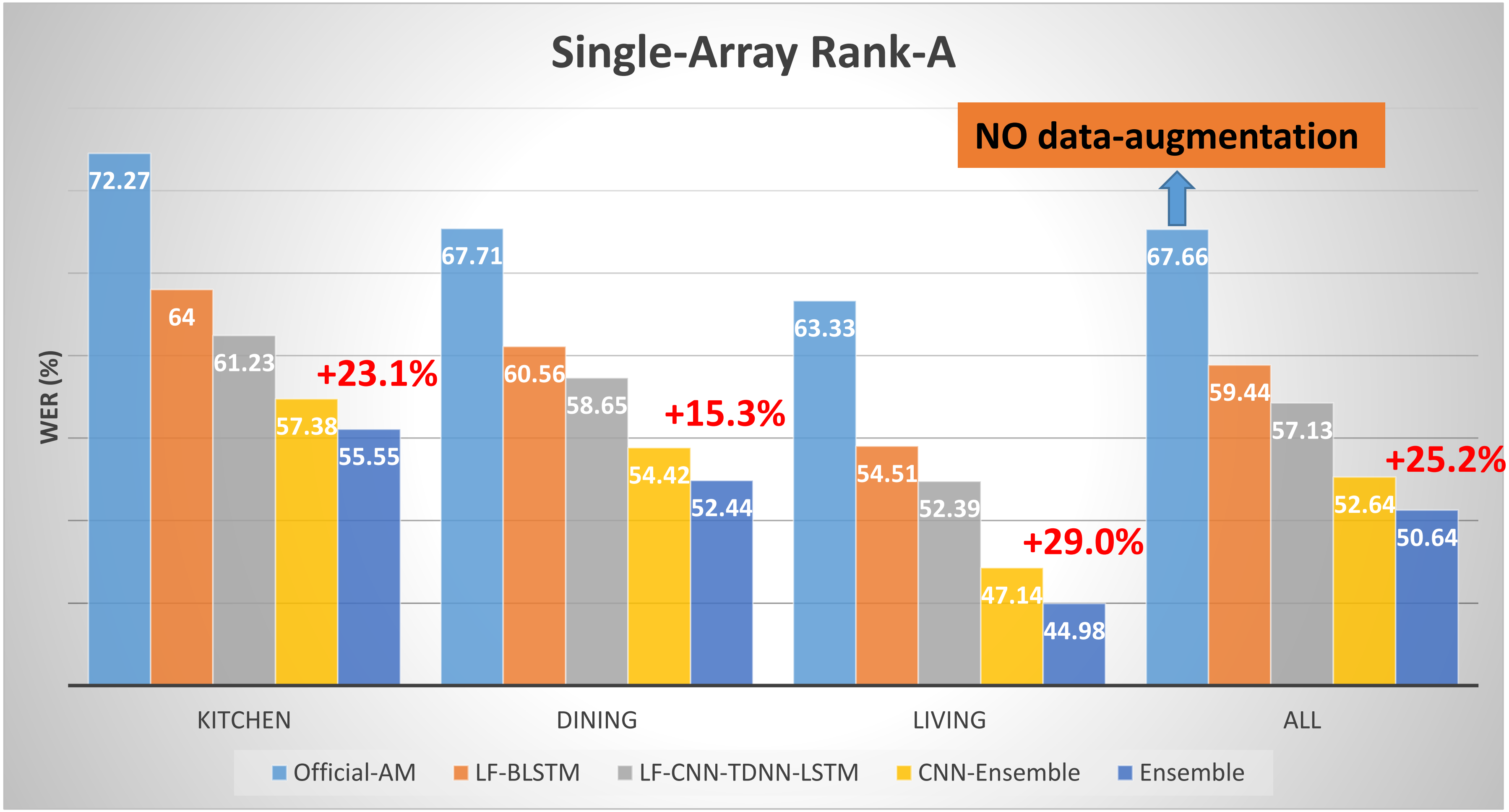
Cross-Entropy Based AMs

- CNN1:
  - CLDNN
  - Input1: 40-d LMFB
  - Input2: Waveform
- CNN2:
  - 50 layers deep fully CNN
  - Input1: 40-d LMFB
  - Input2: Waveform
- CNN3:
  - 50 layers deep fully CNN with gate on feature map
  - Input1: 40-d LMFB
  - Input2: Waveform

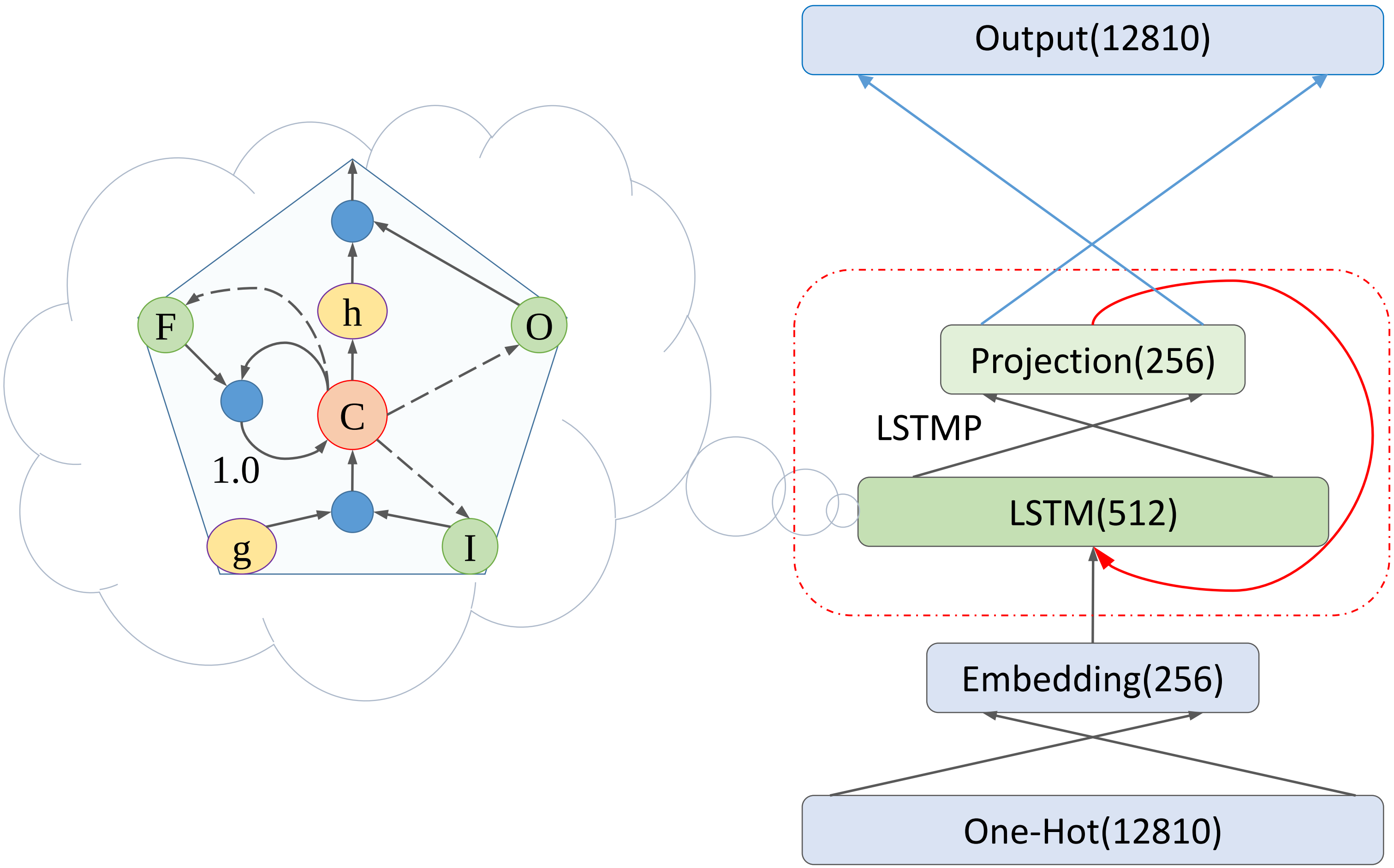

wav
1, T*160
conv
128,1,1025,1,16
relu
pow
pool
128,1,20,1,10
log
reshape
128, T
conv
64,15,3,1,1
BN
relu
pool
64,3,2,3,2
conv
96,7,3,1,1
BN
relu
pool
96,3,1,3,1
conv
128,5,3,1,1
BN
relu
conv
128,5,3,1,1
BN
relu
fbk
40, T
conv
128,9,3,1,1
channel_out,kernel_h,kernel_w, stride_h, stride_w
BN
relu
pool
128,3,2,3,2
conv
256,7,3,1,1
BN
relu
conv
256,5,3,1,1
BN
relu
CLDNN
concat
BLSTM
c1250 p350
FC
700,1,1,1,1
BN
relu
BLSTM
c1250 p350
FC
700,1,1,1,1
BN
relu
BLSTM
c1250 p350
FC
2048,1,1,1,1
BN
relu
deconv
512,1,2,1,2
FC
3936,1,1,1,1
LOSS

AMs with Our Best Front-end

- Ensemble via the state posterior average and lattice combination
- 5-model ensemble (LF-BLSTM, LF-CNN-TDNN-LSTM, CNN-Ensemble)

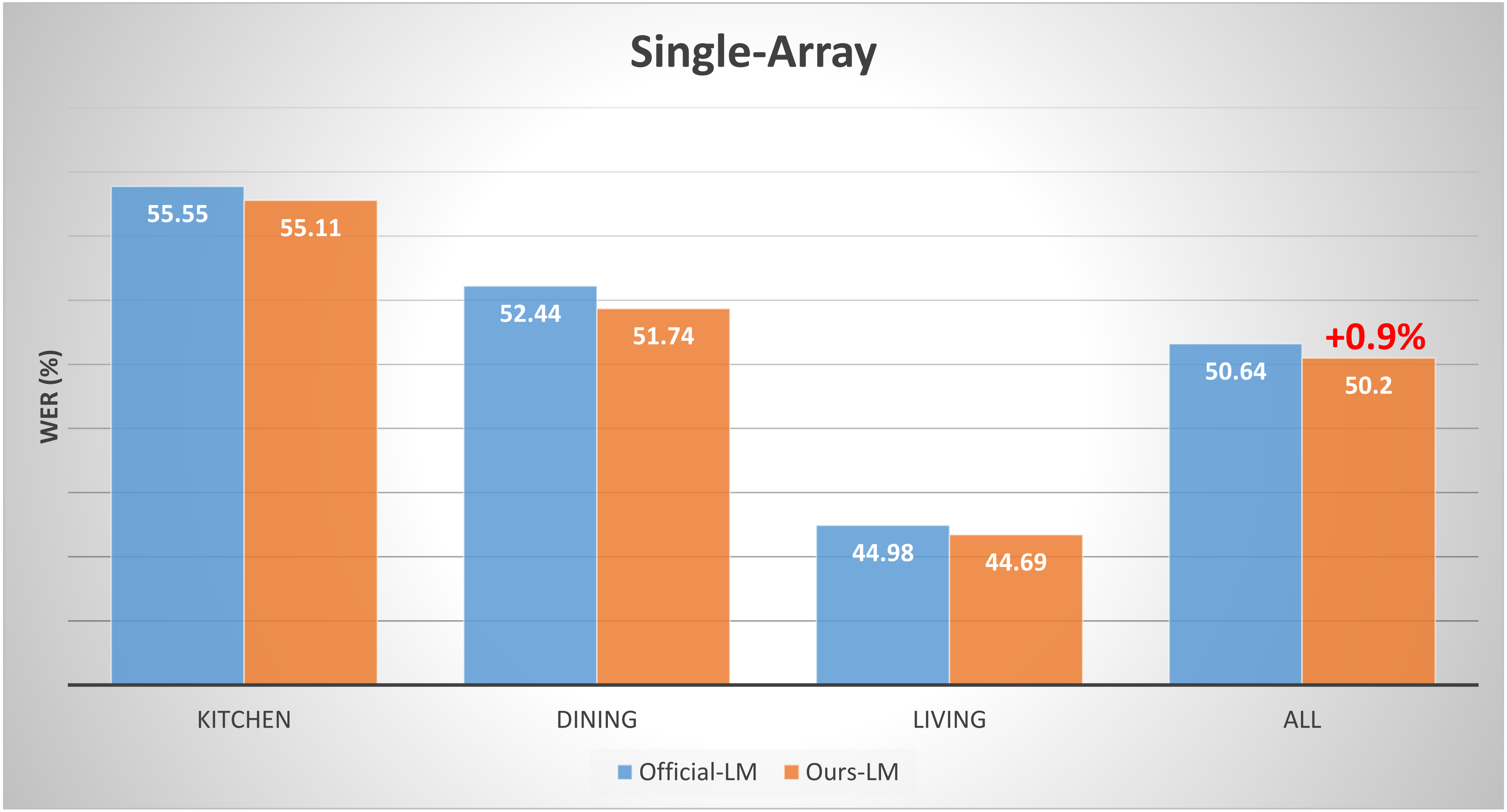


Language Model



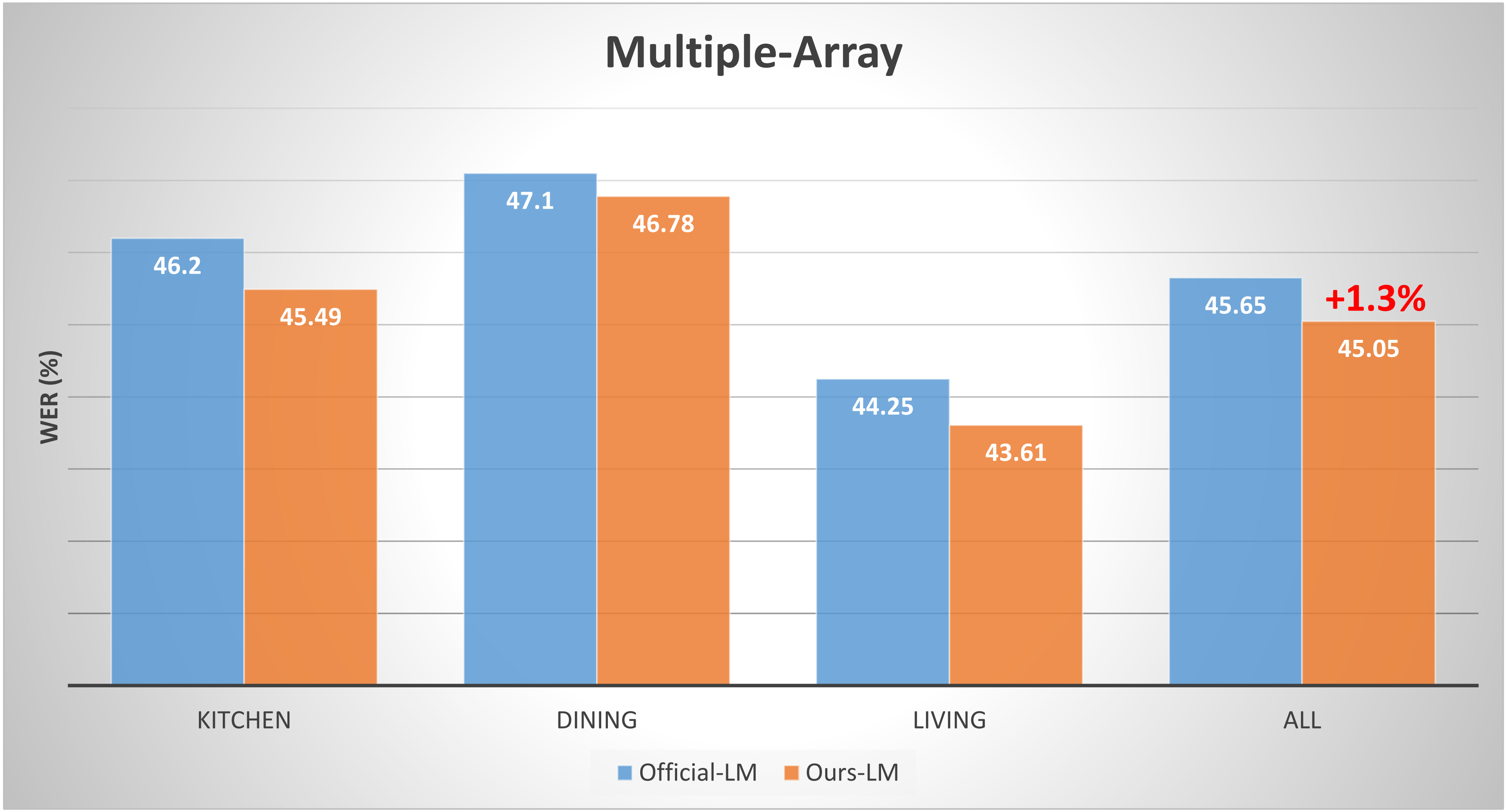
LSTM-LM (Forward) and BLSTM-LM (Forward-Backward) are combined

LMs (Rank A vs. Rank B)

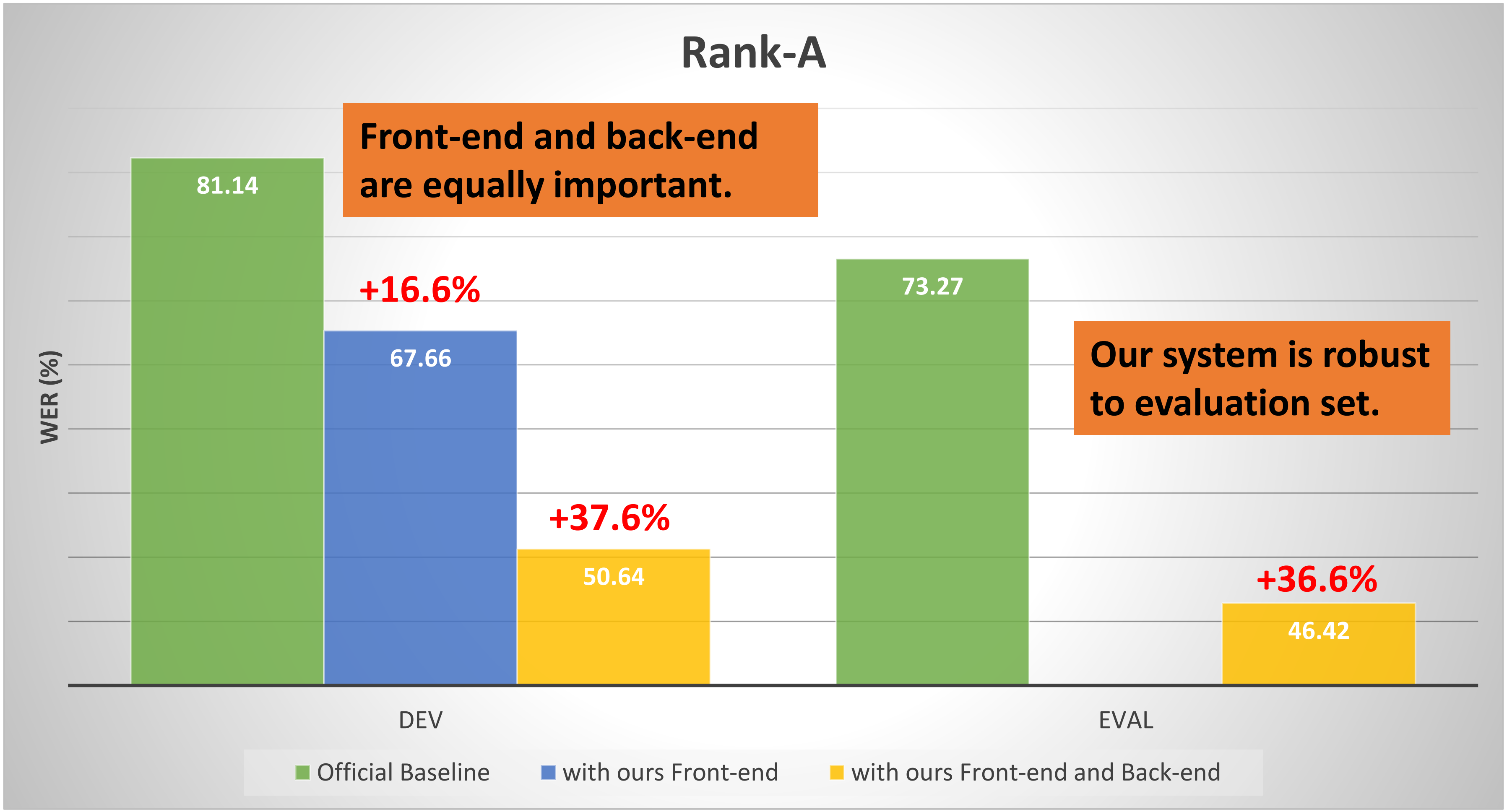


Not significant due to too little training data

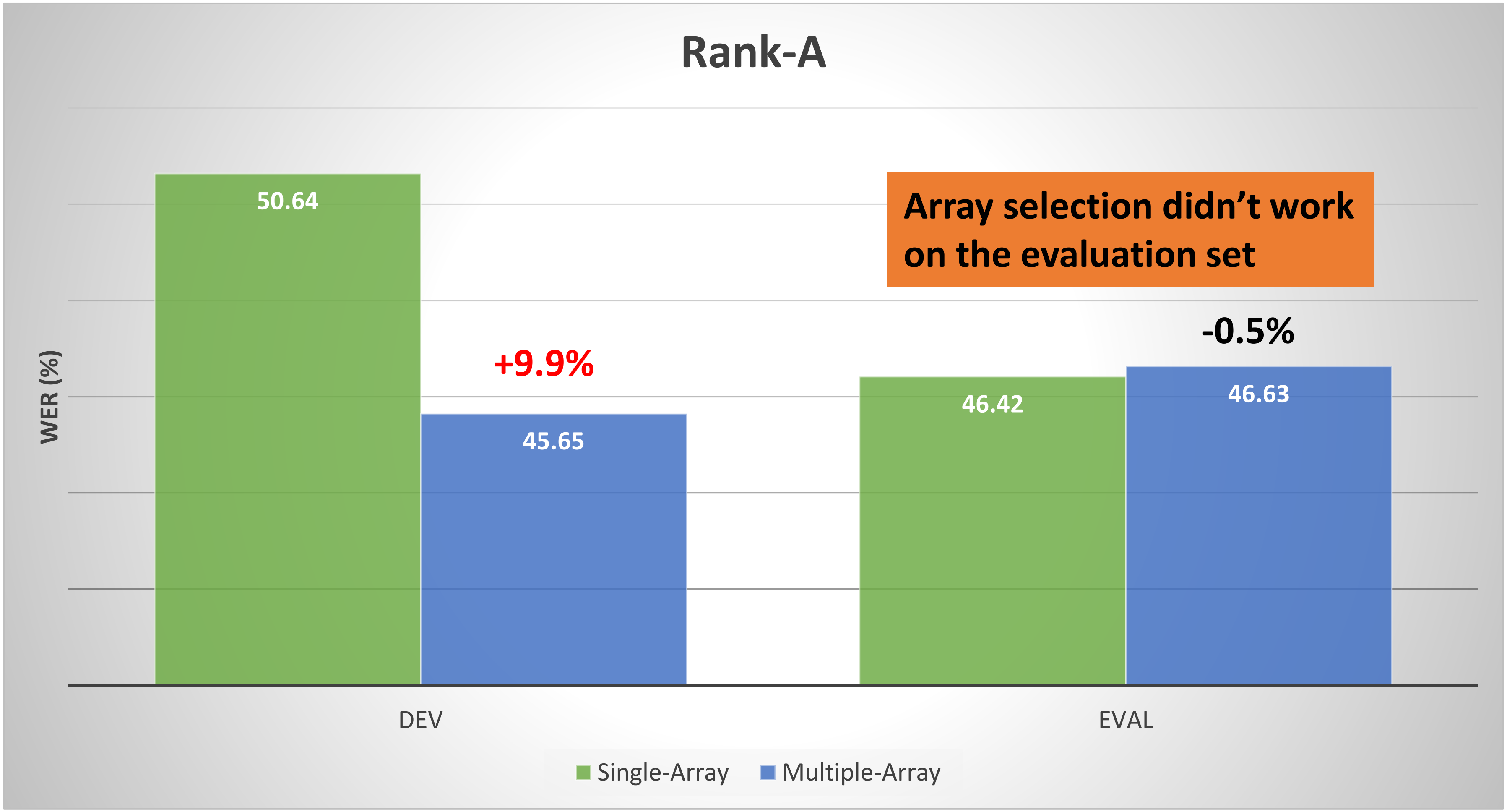
LMs (Rank A vs. Rank B)



Summary of Single-array



Summary of Multiple-array



Summary: Details of Rank-A

Track	Session		Kitchen	Dining	Living	Overall
Single	Dev	S02	57.75	49.43	41.78	50.64
		S09	52.41	56.78	51.36	
	Eval	S01	56.61	38.72	56.70	46.42
		S21	50.41	41.42	42.76	
Multiple	Dev	S02	46.27	46.05	41.15	45.65
		S09	46.11	48.61	50.45	
	Eval	S01	58.73	38.03	55.76	46.63
		S21	52.52	41.57	42.33	

Both systems of single-array and multiple-array are the best

Summary: Details of Rank-B

Track	Session		Kitchen	Dining	Living	Overall
Single	Dev	S02	57.41	48.52	41.49	50.20
		S09	51.83	56.38	51.08	
	Eval	S01	56.26	38.26	56.47	46.11
		S21	50.16	41.44	42.37	
Multiple	Dev	S02	45.41	45.67	40.69	45.05
		S09	45.60	48.37	49.42	
	Eval	S01	58.08	37.11	55.07	46.14
		S21	52.47	41.11	42.20	

Both systems of single-array and multiple-array are the best

Take-Home Message

- Front-End
  - Dinner party scenario is extremely challenging
  - Most previous techniques in CHiME-4 are not working well
  - We design a solution to utilize both traditional and DL techniques
- Acoustic Model
  - Data augmentation is important with contributions from front-end
  - The new design of CLDNN achieves the best performance
  - Different deep architectures are complimentary
- Language Model
  - LSTM-LMs are not effective due to the limited training data
- Multiple-array
  - Our proposed array selection is effective on the development set
  - More analysis should be done on the evaluation set

Acknowledgement

- JSALT 2017
  - Team: Enhancement and Analysis of Conversational Speech
- DIHARD Challenge (Interspeech 2018 Special Session)
- Inspiration for front-end design of CHiME-5 Challenge

Thanks

Q&A