Institute of Acoustics
Chinese Academy of Sciences

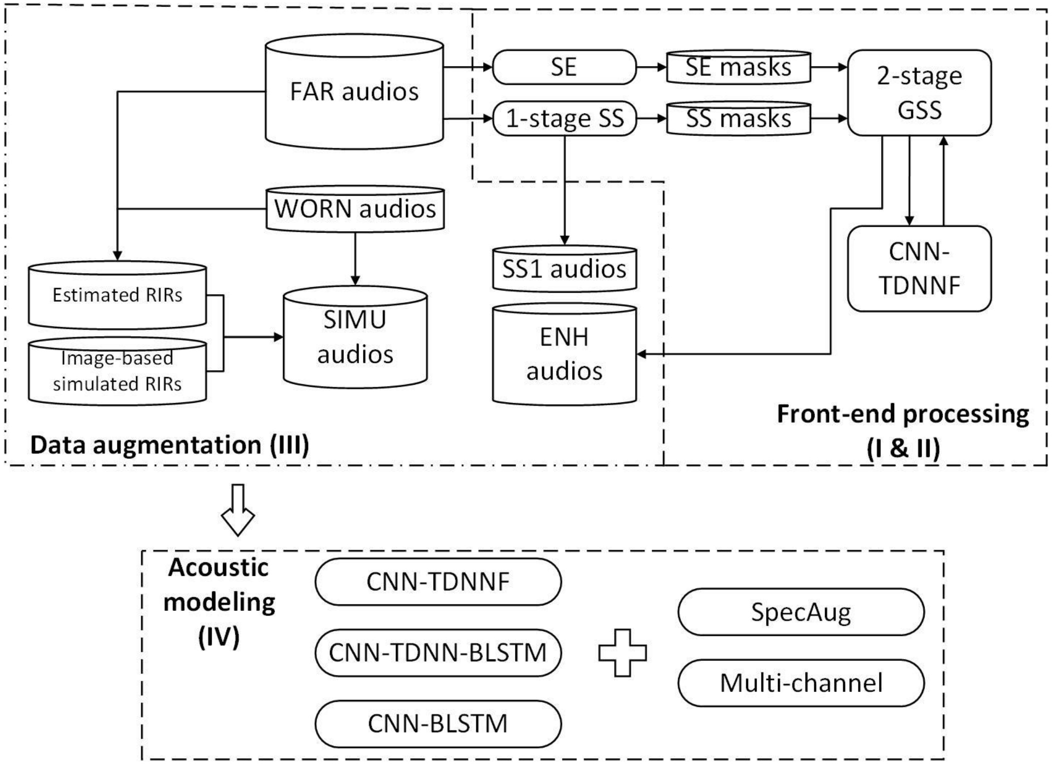
The IOA Systems for CHiME-6 Challenge

Hangting Chen, Pengyuan Zhang, Qian Shi, Zuozhen Liu

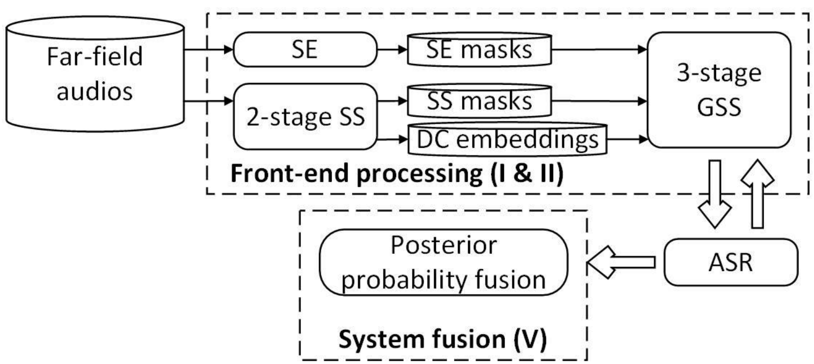
Key Laboratory of Speech Acoustics & Content Understanding, Institute of Acoustics,

Chinese Academy of Sciences, China

System overview



(a) Training phase



(b) Testing phase

Designed for Track 1 A/B

Key components

I. Single-channel speech enhancement
   - SE -> noise mask
   - 1/2-stage SS -> speaker masks and embeddings

II. 2/3-stage GSS in training/testing phase
   - 2-stage GSS with random microphone selection in training
   - 3-stage GSS(*) in testing

III. Data augmentation

IV. Acoustic model
   - 3 types of architectures
   - 2 modules, SpecAug and Multi-channel

V. System fusion

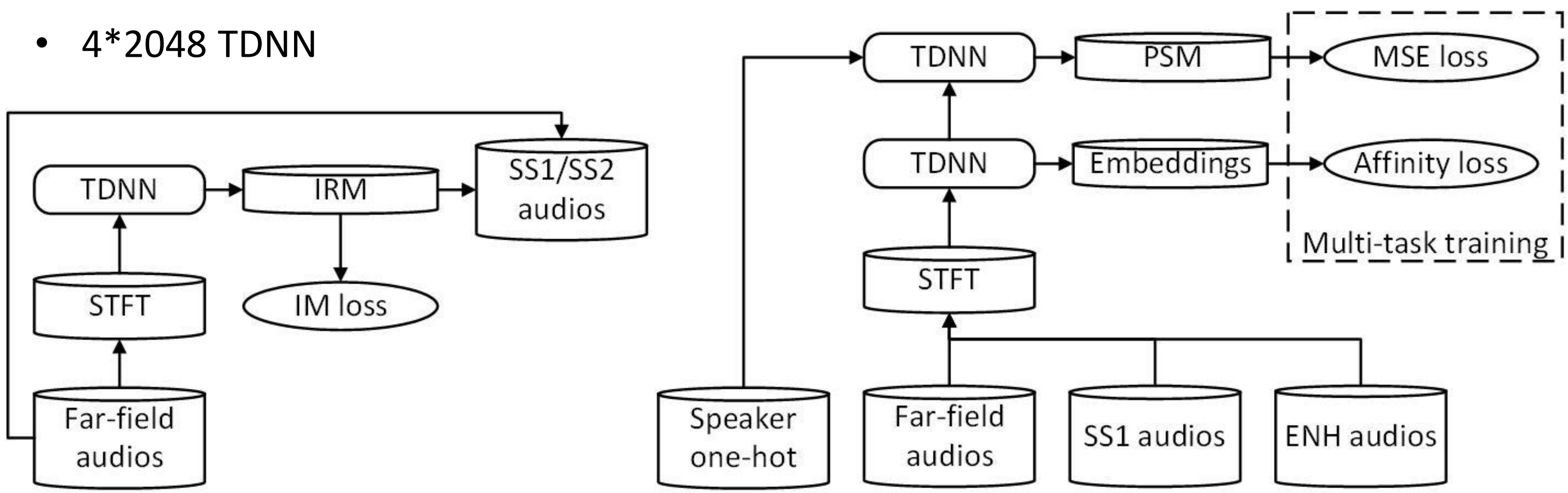
I. Single-channel speech enhancement

- SE model [1]
  - Densely connected progressive learning for TDNN [2]
  - Data
    - Noise data : unlabeled segments filtered by ASR
    - Clean data : Speech segments in original far-field audios, which is not clean actually
    - Loss function : $\left(IRM - \widehat{IRM}\right)^2$
  - Architecture
    - 4*2048 TDNN, 3 progressive output, 1 final output

[1] L. Sun, J. Du, T. Gao, Y. Fang, F. Ma and C. Lee, "A Speaker-Dependent Approach to Separation of Far-Field Multi-Talker Microphone Array Speech for Front-End Processing in the CHiME-5 Challenge," in IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 4, pp. 827-840, Aug. 2019.
[2] T. Gao, J. Du, L. Dai and C. Lee, "Densely Connected Progressive Learning for LSTM-Based Speech Enhancement," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, 2018, pp. 5054-5058.

I. Single-channel speech enhancement

- SS1-spk/SS2-sess model
  - Data
    - Clean data for SS1 : Non-overlap segments
    - Loss for SS1: $(\log(\widehat{IRM}) + \log(|Y| - \log(|X|)))^2$
    - Clean data for SS2/SS2* : Non-overlap segments + SS1 segments + GSS enhanced
    - Loss for SS2*: $(PSM - \widehat{PSM})^2 + (VV^T - BB^T)$
  - Architecture
    - 4*2048 TDNN

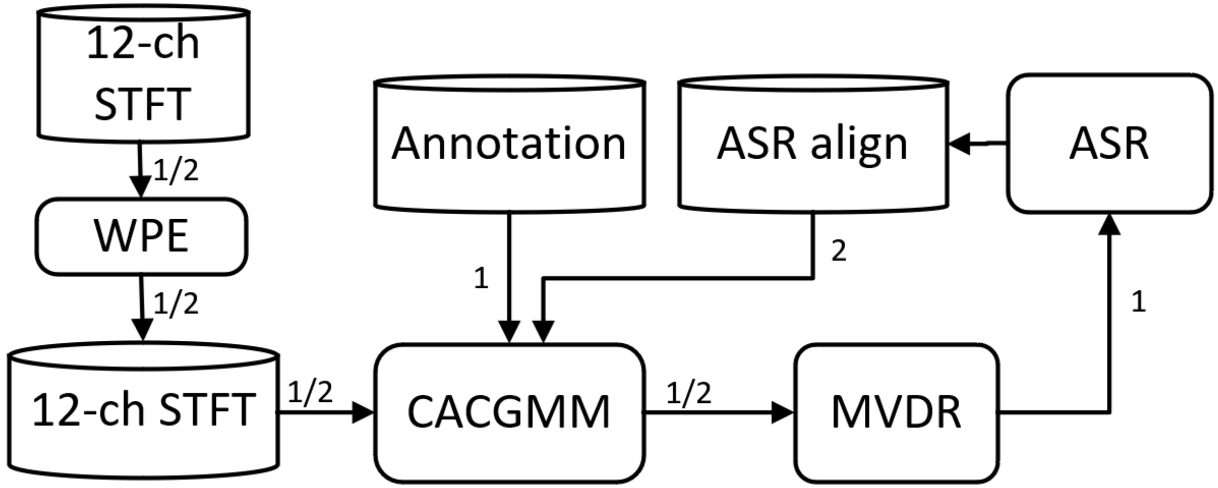


(a) SS1-spk model [1]

(b) SS2-sess model

[1] T. Gao, J. Du, L. Dai and C. Lee, "Densely Connected Progressive Learning for LSTM-Based Speech Enhancement," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, 2018, pp. 5054-5058.

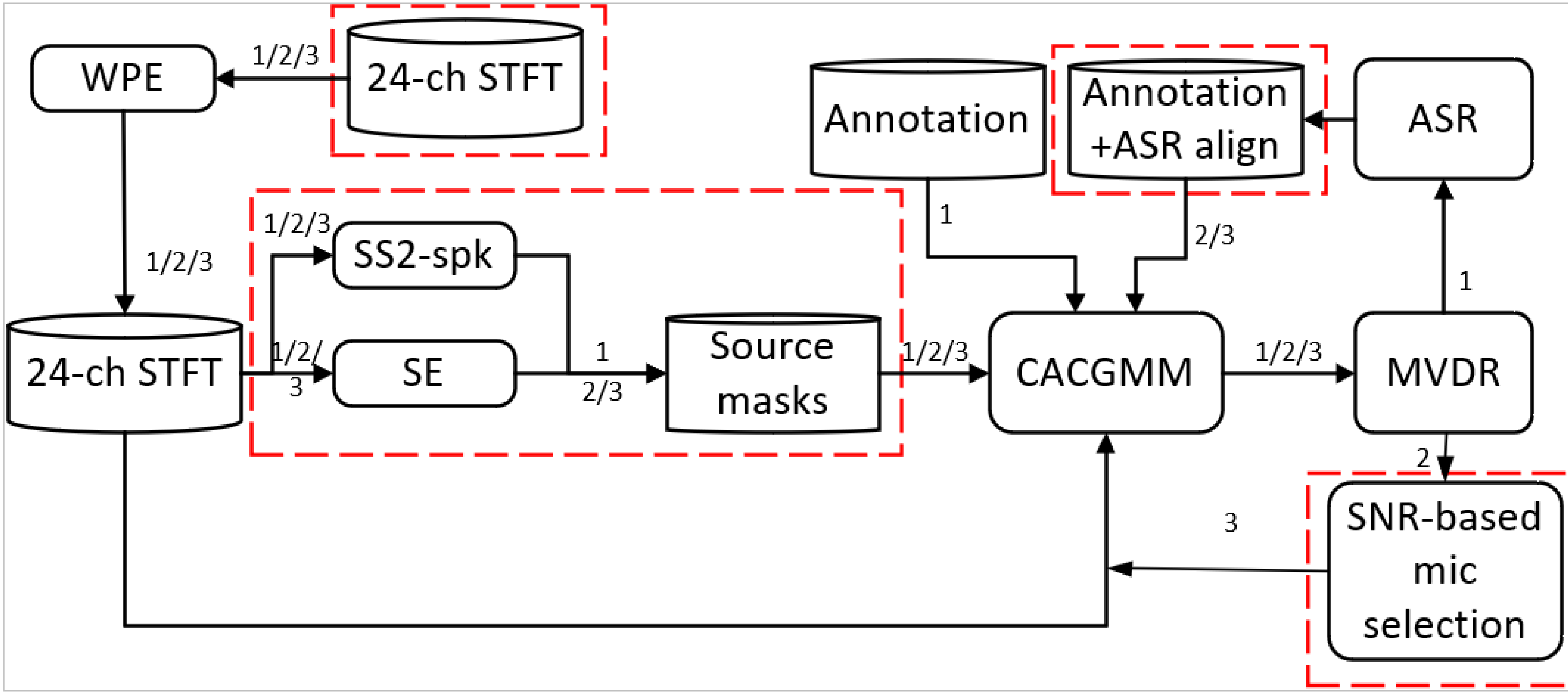
II. 2/3/3*-stage GSS



- Old GSS
- Improvements
    1. Good initialization & 24 mic
    2. Interpolation of annotation and alignment for VAD in each frame t
       $$0.4 \times annot(t) + 0.6 \times align(t)$$
    3. Microphone selection (SNR- or coherency-based [1]) , remove 4/5 from 20/24 mics
    4. Fusion of microphone selection for each microphone i
       $$SNR(i) \ \& \ Coh(i)$$
    5. vMF-CACGMM model [2]
       $$\sum_{t,f} 0.5 \log\left(vMF\left(E_{t,f}\right)\right) + \log\left(CACGMM\left(Y_{t,f}\right)\right)$$

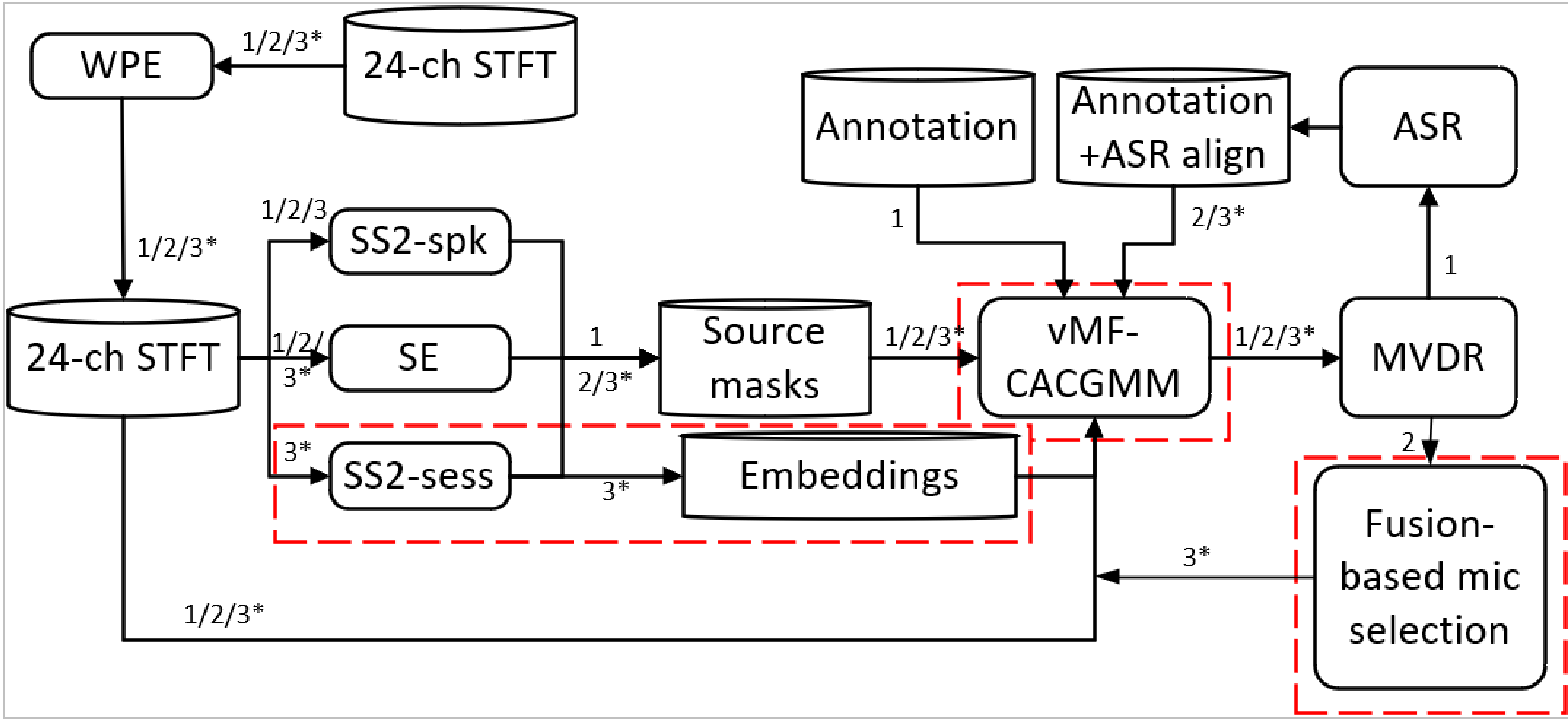
[1] V. M. Tavakoli, J. R. Jensen, M. G. Christensen and J. Benesty, "A Framework for Speech Enhancement With Ad Hoc Microphone Arrays," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 6, pp. 1038-1051, June 2016.
[2] L. Drude and R. Haeb-Umbach, "Integration of Neural Networks and Probabilistic Spatial Models for Acoustic Blind Source Separation," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 815-826, Aug. 2019.

II. 3-stage GSS compared with old GSS



- 3-stage GSS for testing
  - 1st stage → generate ASR alignments
  - 2st stage → generate each mic's SNR
  - 3rd stage → generate 3-stage audios

II. 3*-stage GSS compared with 3-stage GSS



- 3*-stage GSS for testing
  - CACGMM → vMF-CACGMM
  - SNR-based microphone selection → Fusion-based microphone selection

II. Results

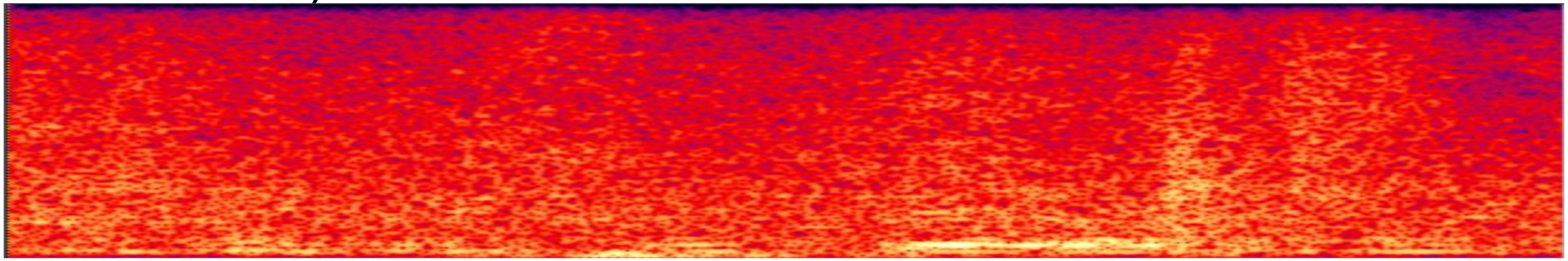
Acoustic model : CNN-TDNNF, Data : worn(2)+oldgss(1)

Init	VAD	Model	microphone selection	Dev. WER(%)	Improvement (%)	Notes
-	Annotation	CACGMM	12 mics	47.67		Baseline
-	Alignment	CACGMM	12 mics	**45.42**	-2.34	Old GSS [1]
SS2	Alignment	CACGMM	12 mics	43.73	-1.69	
SS2	Interpolation	CACGMM	12 mics	43.46	-0.27	
SS2	Interpolation	CACGMM	24 mics	42.59	-1.14	2-stage GSS for testing
SS2	Interpolation	CACGMM	SNR-based	**42.14**	-0.45	3-stage GSS for testing
SS2	Interpolation	CACGMM	Fusion	41.97	-0.17	
SS2	Interpolation	vMF-CACGMM	Fusion	**41.75**	-0.22	3*-stage GSS for testing

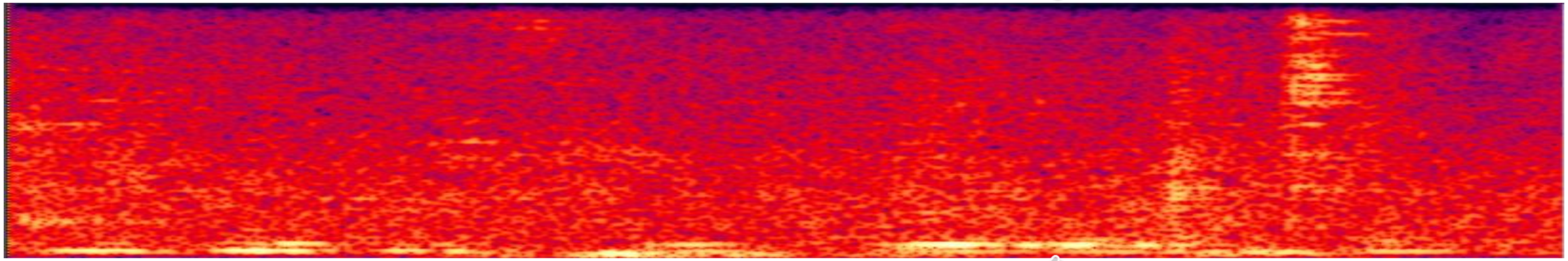
[1] C. Zorilă, C. Boeddeker, R. Doddipatla and R. Haeb-Umbach, "An Investigation into the Effectiveness of Enhancement in ASR Training and Test for Chime-5 Dinner Party Transcription," 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), SG, Singapore, 2019, pp. 47-53.

II. Spectrum of 3-stage and 3*-stage

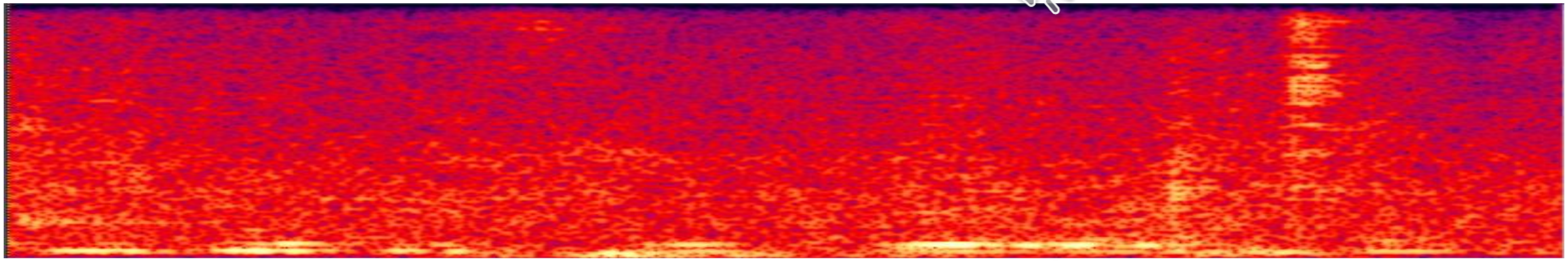
It is the blue, I think



Baseline GSS

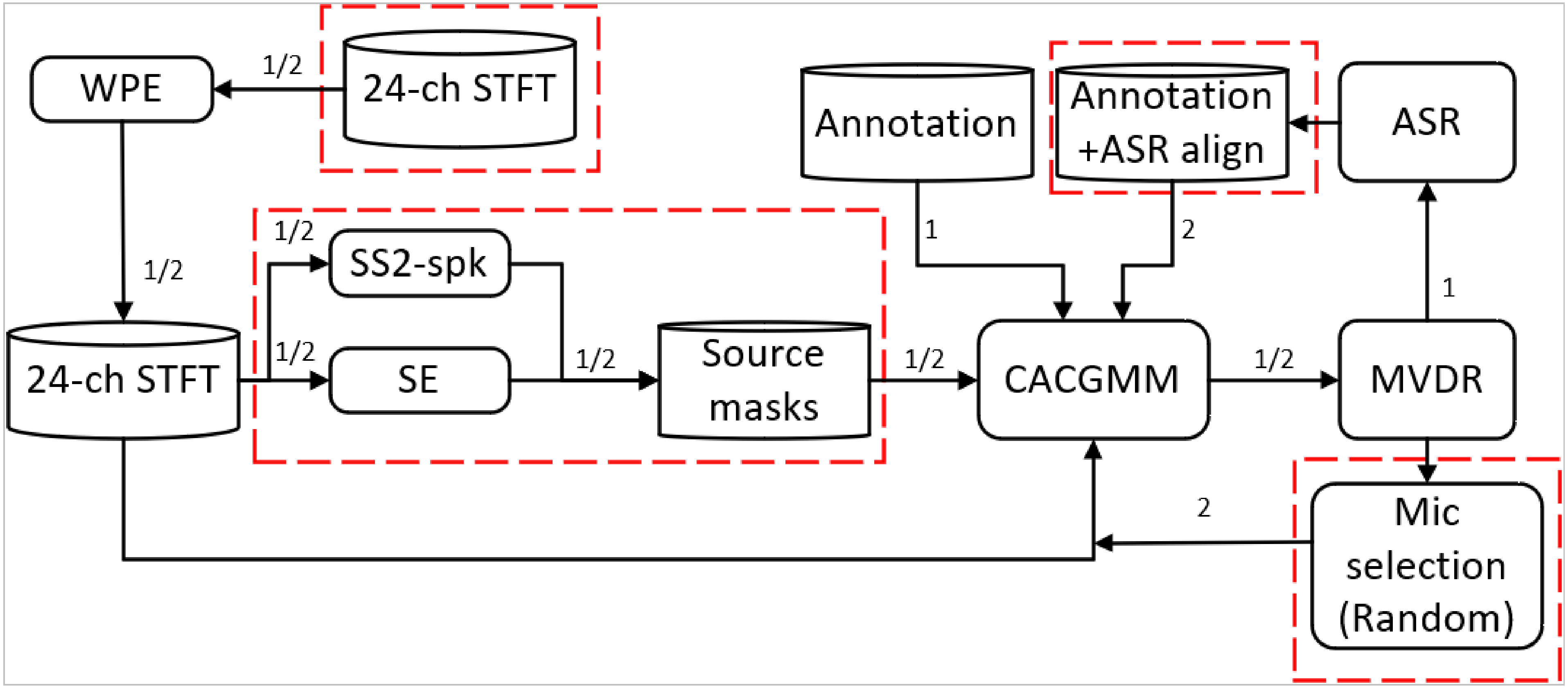


3-stage GSS



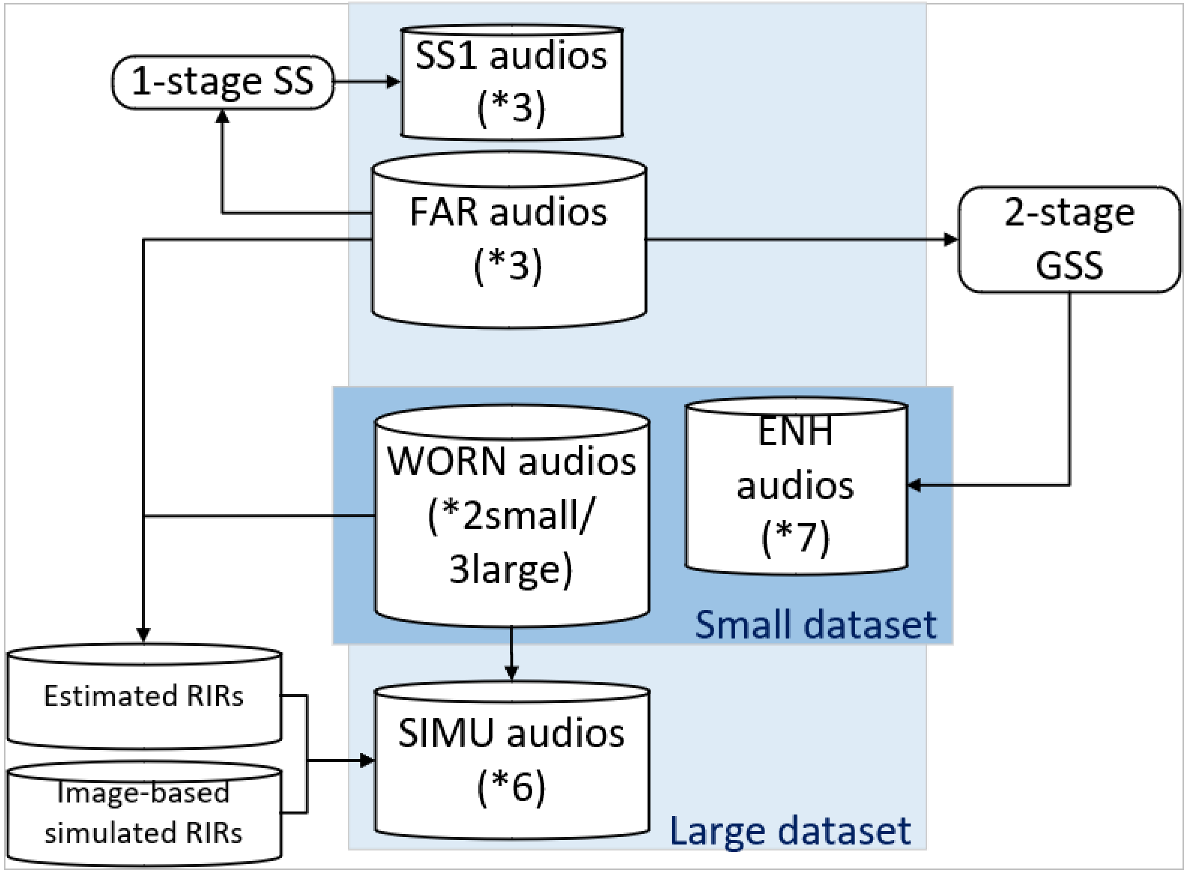
3*-stage GSS

II. 2-stage GSS compared with old GSS



- 2-stage GSS for training
  - Random microphone selection to generate 7-fold data

III. Data augmentation



1. Small dataset for small acoustic models
   - Totally (2+7)*3*40 -> 1000 hours
2. Large dataset for large acoustic models
   - Totally (3+3+3+7+6)*3*40 -> 2600 hours

IV. Acoustic model

2 tricks for training

- Short utterance combination
- Bi-phone tree for chain model, instead of triphone

3 main architectures

CNN-TDNNF
5-layer CNN
9-layer TDNNF

CNN-TDNN-BLSTM
2-layer CNN
8-layer TDNN
3-layer BLSTM
Interleave BLSTM with TDNN

CNN-BLSTM
3-layer CNN
3-layer BLSTM
3-layer DNN

2 modules

SpecAug [1]
Useful for CNN-TDNNF and CNN-BLSTM

4-ch branch
- Inspired from [2]
- Use LPS and magnitude squared coherence (MSC)
- CNN(-BLSTM) instead of TDNN-BLSTM
- Decode use REF array

[1] Park, Daniel S. et al. "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition." Interspeech 2019 (2019): n. pag. Crossref. Web.
[2] N. Kanda, Y. Fujita, S. Horiguchi, R. Ikeshita, K. Nagamatsu and S. Watanabe, "Acoustic Modeling for Distant Multi-talker Speech Recognition with Single- and Multi-channel Branches," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 6630-6634.

IV. Results of data aug. & training tricks

Acoustic model : CNN-TDNNF, Front-end : 3-stage GSS

Data(fold)	Modules	Dev. WER(%)	Improvement
worn(2) +oldGSS(1)	-	42.14	
worn(2) +oldGSS(1)	SpecAug	40.57	-1.57
worn(2) +2-stage GSS(2)	SpecAug	39.91	-0.66
worn(*2) +2-stage GSS(7)	SpecAug	39.79	-0.12
worn(2) +2-stage GSS(7)	SpecAug + short utterance combination	39.20	-0.59
worn(2) +2-stage GSS(7)	SpecAug + short utterance combination + biphone tree	**38.73**	-0.47

IV. Results of acoustic models

Architecture	Dataset	Training settings	3-stage GSS Dev./Eval. WER(%)	3*-stage GSS Dev./Eval. WER(%)
CNN-TDNNF	Small	SpecAug	38.73/40.83	38.45/40.85
+Multichannel-CNN-1	Small	Partial update	38.10/39.16	37.95/38.98
+Multichannel-CNN-1-BLSTM	Small	Partial update	38.93/39.73	38.70/39.77
CNN-TDNNF-BLSTM	Small	SpecAug	38.41/40.04	37.90/39.95
+Multichannel-CNN-1	Small	Partial update	37.98/38.42	**37.76**/**38.27**
CNN-TDNNF-attention	Small	SpecAug	39.72/42.09	39.33/41.83
CNN-TDNN-BLSTM	Large	-	38.15/40.10	37.95/39.81
CNN-TDNN-BRLSTM-1	Large	-	38.19/40.26	37.89/40.16
+Multichannel-CNN-1	Large	Full update	38.27/40.42	38.02/40.16
+Multichannel-CNN-2	Small	Partial update	37.57/**38.60**	**37.29**/38.72
CNN-TDNN-BRLSTM-2	Large	SpecAug	42.15/44.03	41.92/43.70
CNN-BLSTM	Large	SpecAug	36.60/38.63	35.92/38.30
+Multichannel-CNN-1	Small	Partial update	37.47/38.45	37.17/38.30
CNN-BLSTM-deltaLayer	Large	SpecAug	37.69/39.41	37.30/39.27
CNN-BLSTM-resnet	Large	SpecAug	35.86/37.97	**35.54**/**37.95**

V. Fusion

Weighted average of posterior probability

Steps:

1. For each type of acoustic models, conduct average fusion.
2. For different types of models, conduct weighted fusion.
3. For different types of front-end, conduct weighted fusion.

Acoustic model type (#)	**3-stage GSS Dev./Eval. WER(%)**	**3*-stage GSS Dev./Eval. WER(%)**
CNN-TDNNF (3)	36.71/38.79	36.25/38.46
CNN-TDNNF + Multi-channel (3)	36.23/37.13	**36.07/37.05**
CNN-TDNN-BLSTM (3)	36.63/38.86	**36.38**/38.52
CNN-TDNN-BLSTM + Multi-channel (2)	37.02/38.32	36.67/**38.28**
CNN-BLSTM (3) CNN-BLSTM + Multi-channel (1)	34.88/36.37	**34.48/36.36**
Fusion with weight 0.05:0.15:0.1:0.1:0.6	34.18/35.67	**33.76/35.56**
Fusion with weight 0.4:0.6	**33.55/35.11**	
RNN rescore	**32.92/34.53**	

V. Final results & Conclusion

Category	Session	Dining	Kitchen	Living	Ave
A	S02	38.30	38.50	31.59	33.55
	S09	32.25	30.07	29.23	
	S01	29.58	48.49	42.72	35.11
	S21	29.76	39.66	28.60	
B	S02	37.51	38.02	31.06	32.92
	S09	31.54	29.69	28.11	
	S01	28.83	48.61	41.64	34.53
	S21	29.14	39.39	28.03	

- The initialization and microphone selection plays an important role in our front-end. The fusion of different front-end can stably lower the WER.
- The data augmentation is important to increase the capacity of acoustic models.
- The multi-channel branch may help the performance.
- The deep CNNs can bring a better acoustic model.

Thank you