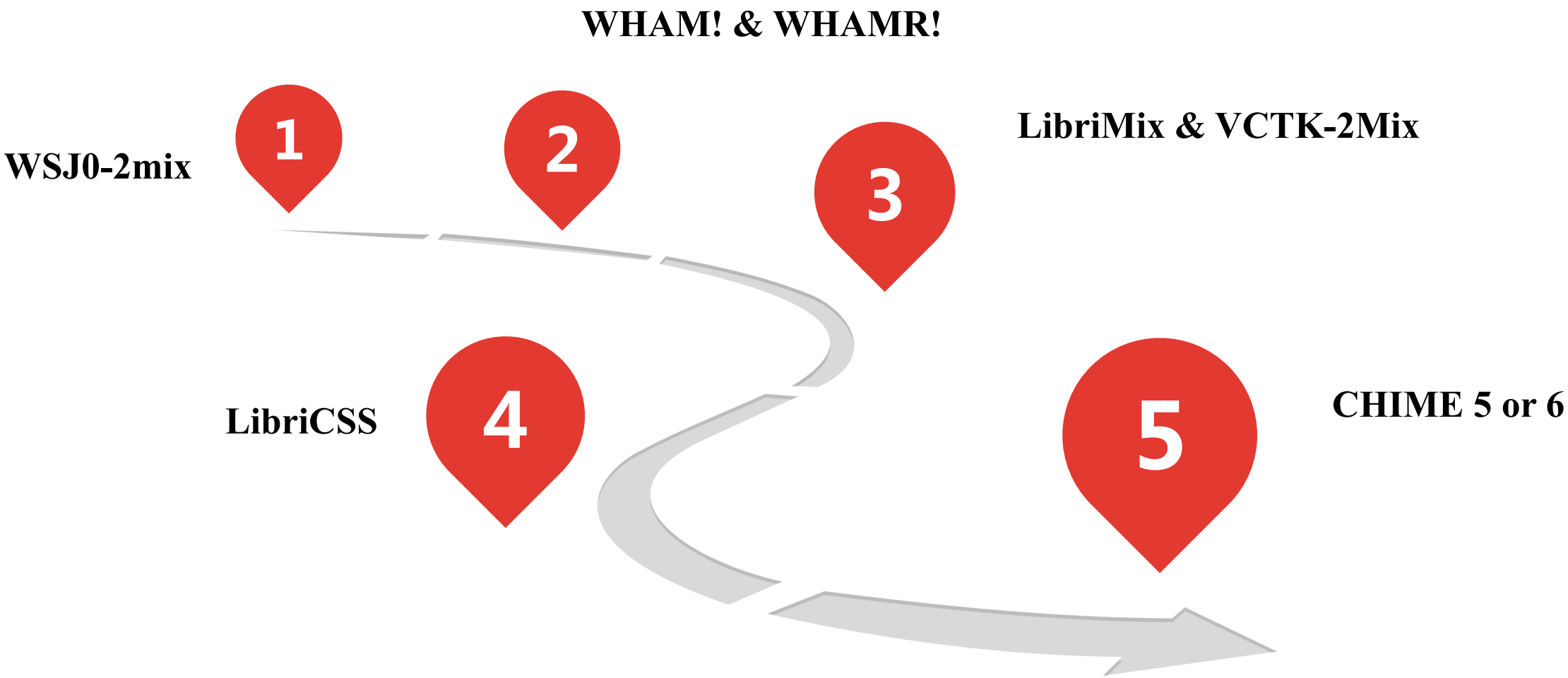
Speech Separation Introduction:

Public Datasets Part

Reporter: Meng Ge

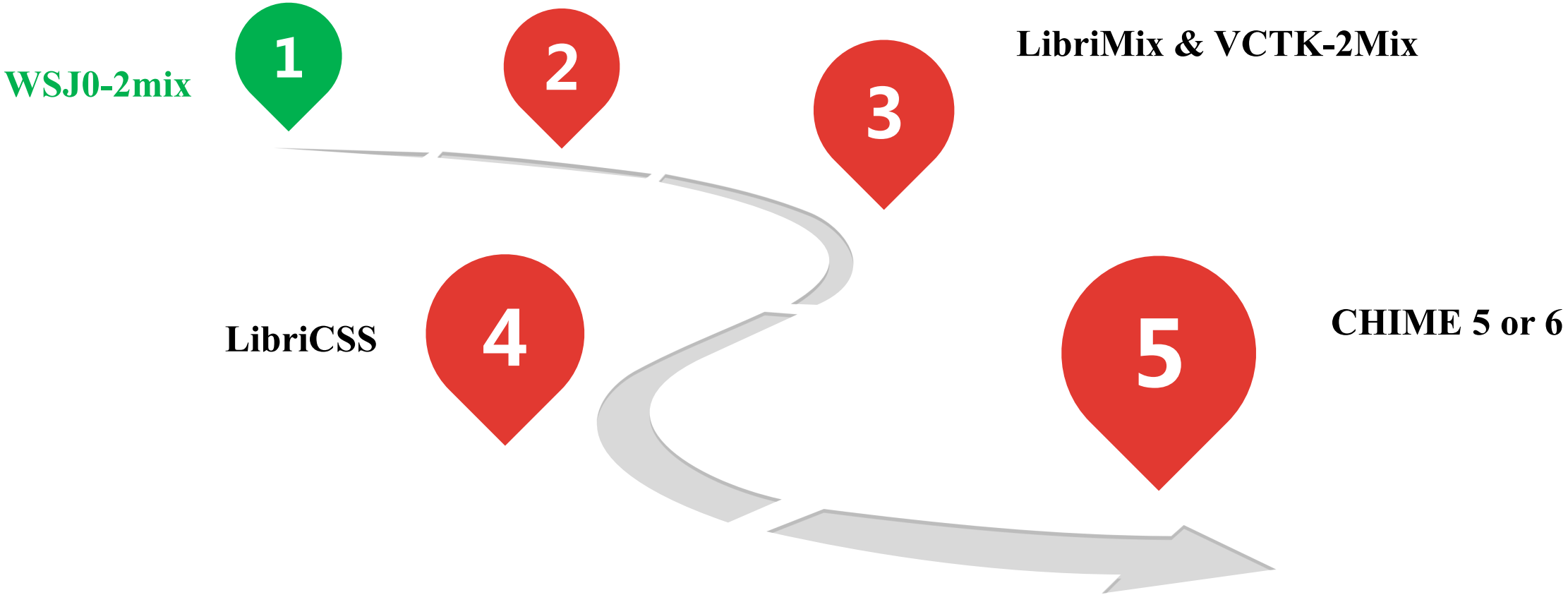
Date: Aug 12, 2020

Outline

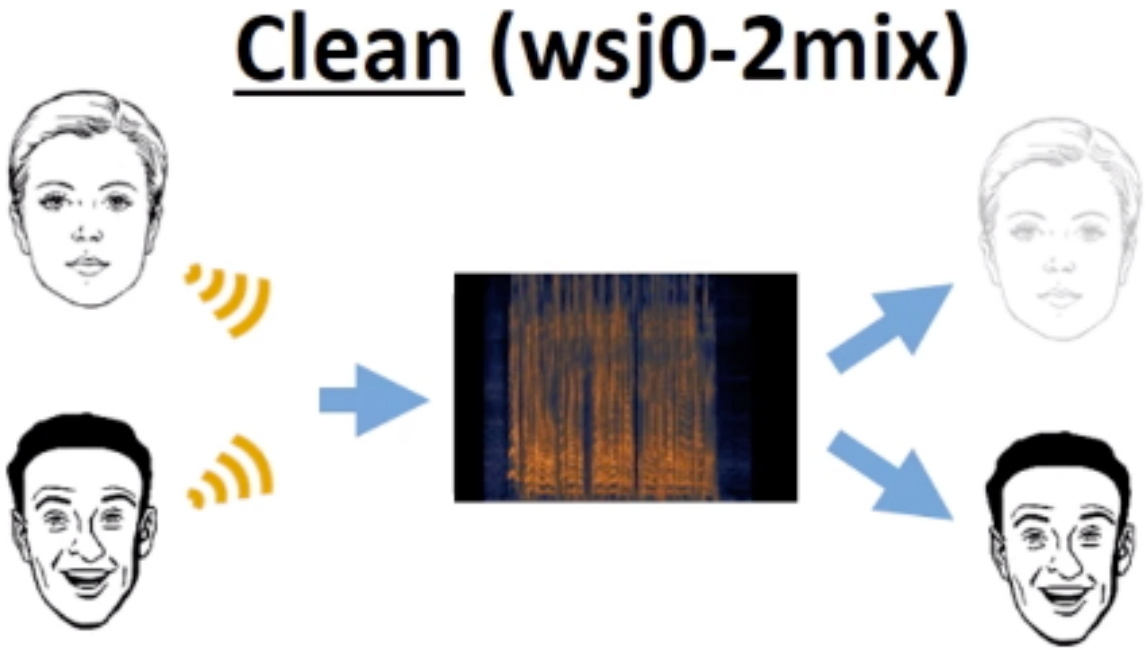

WHAM! & WHAMR!
WSJ0-2mix
1
2
LibriMix & VCTK-2Mix
3
LibriCSS
4
5
CHIME 5 or 6

Outline

WHAM! & WHAMR!



WSJ0-2mix Dataset

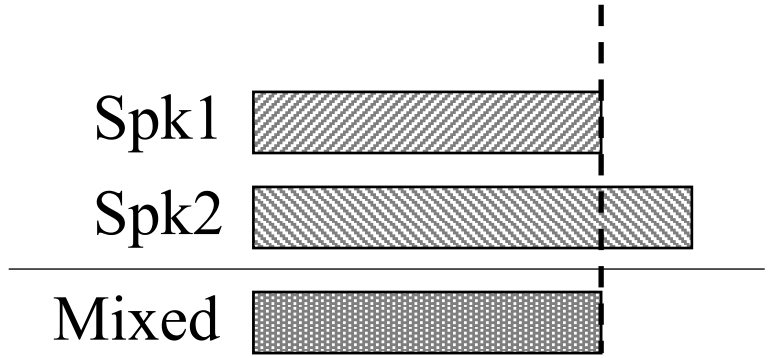


$$y = s_1 + s_2$$

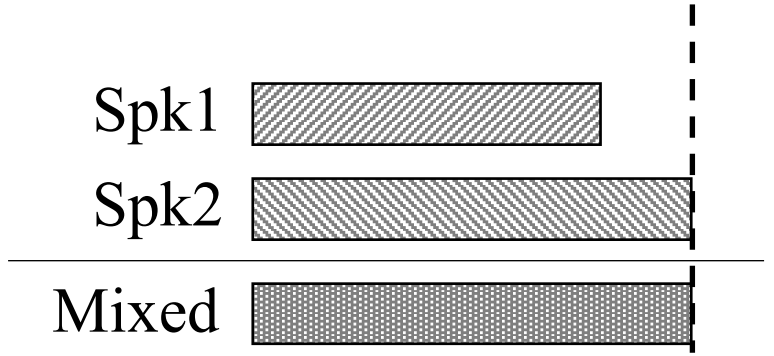
WSJ0-2mix and 3mix

■ WSJ0-2mix and 3mix [1]

- ➢ Derived from WSJ0 corpus
- ➢ 2- and 3-speaker mixtures (artificially generated)
- ➢ Mixed at SNR between 0dB and 5dB
- ➢ 30h train (2w Utt.), 10h dev (5k Utt.), 5h test (3k Utt.)
- ➢ It consists of 101 training speakers
- ➢ Sampling rate: **8k Hz** / 16k Hz
- ➢ Overlap: **Min** / Max
- ➢ Fully Overlap, Utterance-level, Single-channel



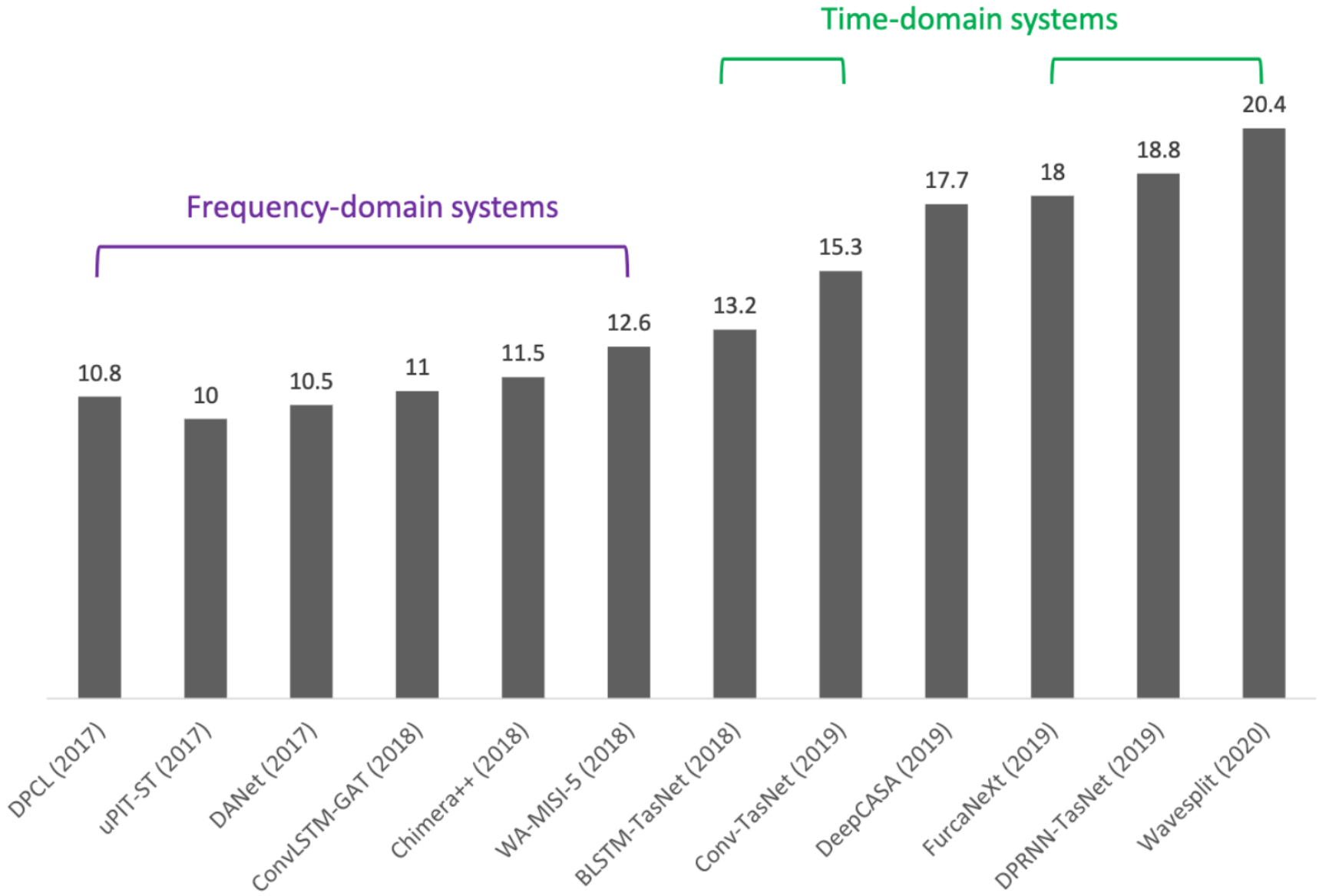
(a) **Min** Case



(b) **Max** Case

[1] Deep Clustering: Discriminative Embeddings for Segmentation and Separation
[2] http://www.merl.com/demos/deep-clustering/create-speaker-mixtures.zip

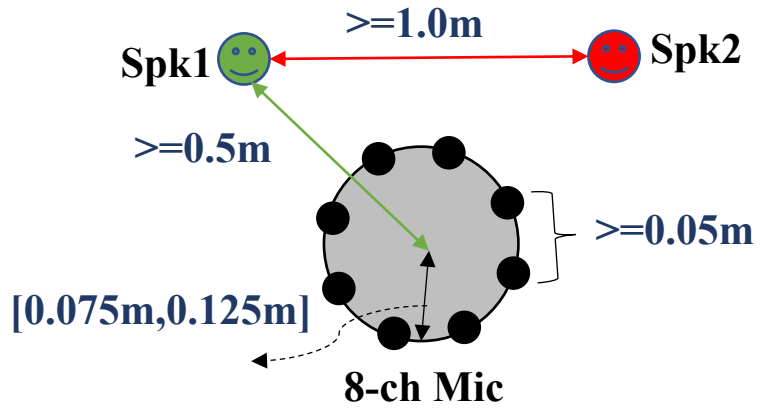
SDRi Result on WSJ0-2mix (8k, Min)

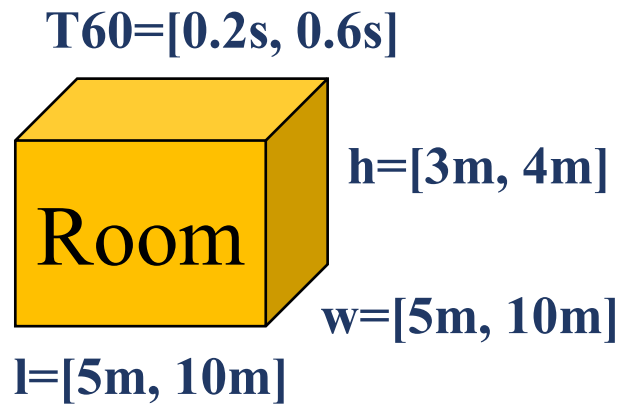


MC-WSJ0-2mix

■ MC-WSJ0-2mix [1]

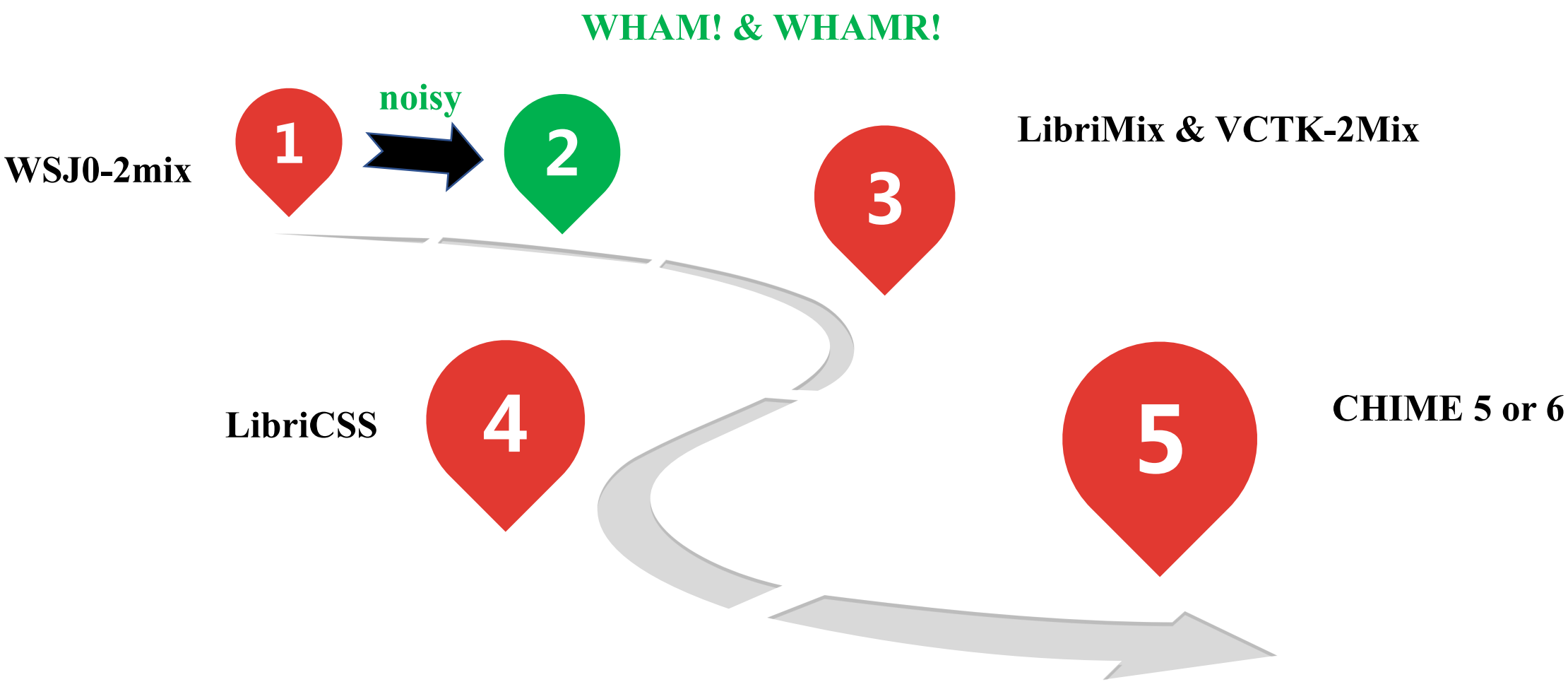
- Multi-channel version of the WSJ0-2mix corpus
- It consists of **8**-channel recordings
- Recordings are generated by convoluting clean speech signal and room impulse response (RIR) [2].
- RIR is simulated with the image method for reverberation time of up to about **600ms**
- Two conditions for T60:
  - ✓ **Anechoic**: the room is assumed anechoic, T60=0.
  - ✓ **Reverb**: the room is assumed reverb, T60 is random.
- Fully Overlap, Utterance-level, Multi-channel





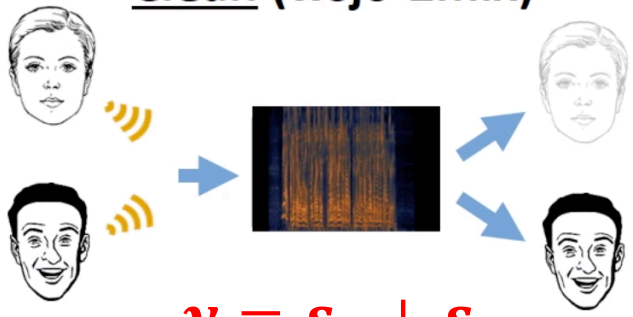
[1] Multi-Channel Deep Clustering: Discriminative Spectral and Spatial Embeddings for Speaker-Independent Speech Separation
[2] https://github.com/ehabets/RIR-Generator [3] http://www.merl.com/demos/deep-clustering/spatialize_wsj0-mix.zip

Outline


WHAM! & WHAMR!
noisy
WSJ0-2mix
1
2
3
LibriMix & VCTK-2Mix
LibriCSS
4
5
CHIME 5 or 6

WHAM! & WHAMR!

Clean (wsj0-2mix)



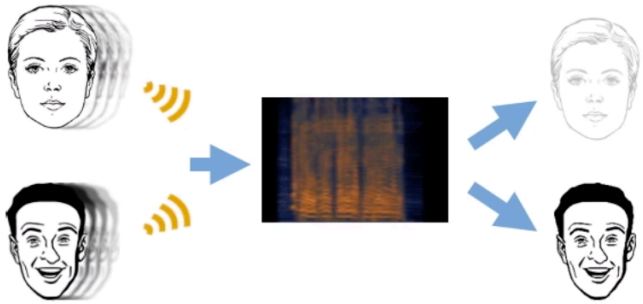
$$y = s_1 + s_2$$

Noisy (WHAM!)

$$y = s_1 + s_2 + n$$

New to WHAMR!

Reverberant



$$y = s_1 * h_1 + s_2 * h_2$$

Noisy and Reverberant

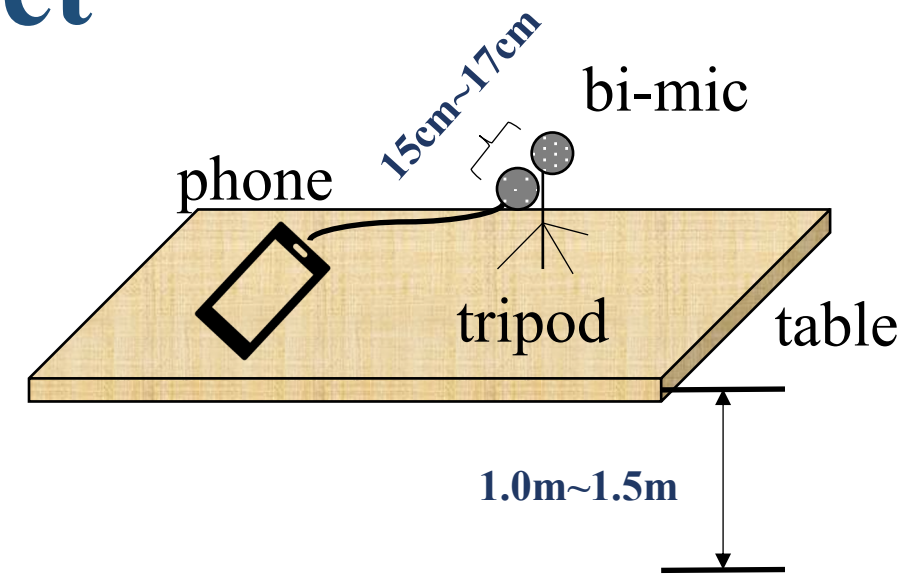
$$y = s_1 * h_1 + s_2 * h_2 + n$$

[1] WHAM!: Extending Speech Separation to Noisy Environments
[2] WHAMR!: Noisy and Reverberant Single-Channel Speech Separation

WHAM! Dataset

■ WHAM!

- WHAM! is based on the WSJ0-2mix
- Pairs each two-speaker utterance in WSJ0-2mix with a unique **noise background scene**, e.g. coffee shop, restaurants, bars.
- Background Audio is recorded using an Apogee Sennheiser **binaural microphone** connected to a **smartphone** (shown in Fig.)
- 48kHz, downsampling to 16kHz or 8kHz.
- Background audio: 80h, 44 locations
- Mixed by a random SNR between -6 and +3 dB
- Fully Overlap, Utterance-level, Single-channel



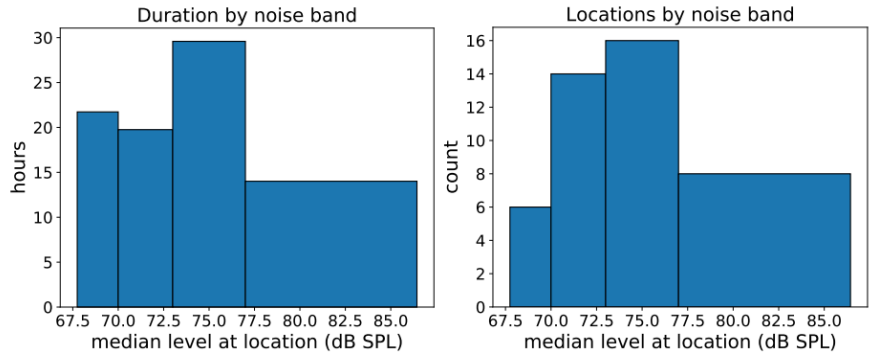


Figure 1: *Histograms of duration and unique locations where background noise was recorded.*

Result on WHAM! Dataset

Table 1: *SI-SDR [dB] oracle performance on WHAM! tasks*

Task	Dataset	Noisy	IRM	IBM	PSF
enhance-single	8 kHz min	-0.9	11.0	11.6	14.7
	16 kHz max	-2.9	11.0	11.6	14.8
enhance-both	8 kHz min	1.2	10.9	11.4	14.6
	16 kHz max	-0.7	10.8	11.4	14.5
separate-clean	8 kHz min	0.0	12.7	13.5	16.4
	16 kHz max	0.0	13.4	14.2	17.1
separate-noisy	8 kHz min	-4.5	8.3	8.9	12.3
	16 kHz max	-5.8	8.5	9.1	12.5

Table 2: *SI-SDR [dB] performance comparison of chimera++ networks on WHAM! tasks, where Δ indicates improvement.*

Task	Dataset	Noisy	Output	Δ
enhance-single	8 kHz min	-0.9	10.2	11.1
	16 kHz max	-2.9	10.0	12.9
enhance-both	8 kHz min	1.2	9.4	8.2
	16 kHz max	-0.7	9.3	10.0
separate-clean	8 kHz min	0.0	11.0	11.0
	16 kHz max	0.0	9.6	9.6
separate-noisy	8 kHz min	-4.5	5.4	9.9
	16 kHz max	-5.8	4.4	10.2

- **enhance-single**: s1+n → s1
- **enhance-both**: s1+s2+n → s1+s2
- **separate-clean**: s1+s2 → s1 & s2
- **separate-noisy**: s1+s2+n → s1 & s2

Results on WHAM! Dataset

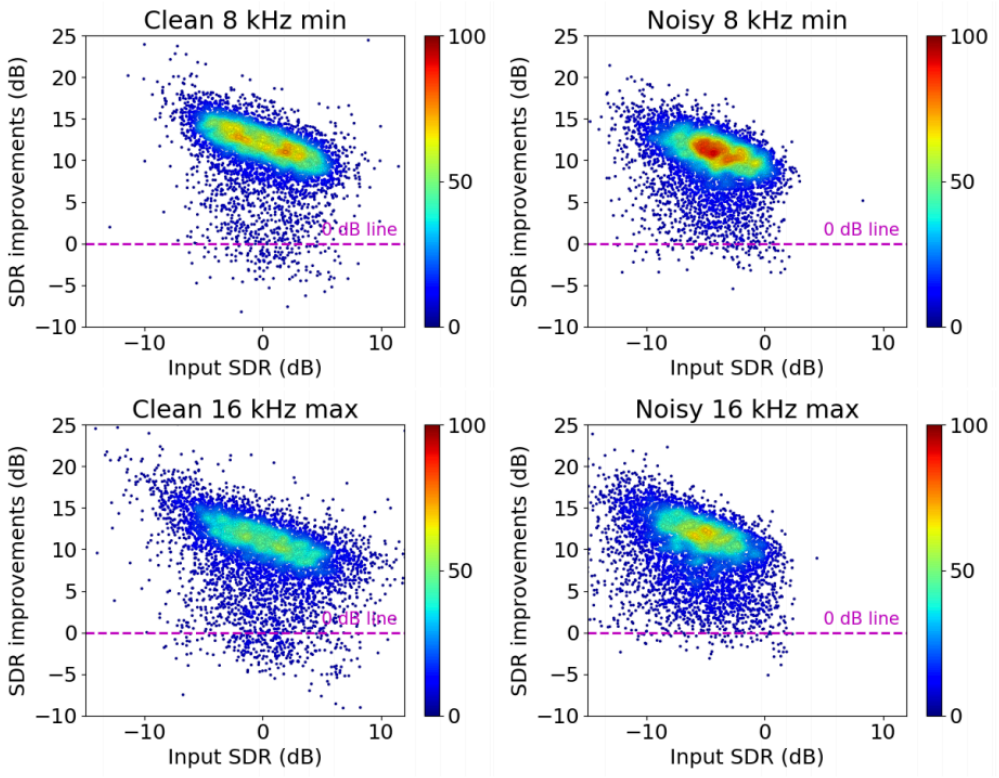


Figure 3: *SI-SDR scatter plots comparing chimera++ performance over different datasets.*

- This suggests that improving the quality of relatively quiet speakers is more difficult in the presence of background noise.

Table 3: *SI-SDR [dB] improvement comparison of different chimera++ objectives for noisy separation on 8 kHz min*

DPCL Objective	DPCL Sources	Δ SI-SDR
n/a (mask inference)	-	8.5
$\mathcal{L}_{DC,C}$	3	9.6
$\mathcal{L}_{DC,N}$	3	9.6
$\mathcal{L}_{DC,W}$	3	**9.9**
$\mathcal{L}_{DC,W}$, 0 weight on noise bins	2	8.4
enh-both + sep-clean	2	9.0
enh-both + sep-clean-finetune	2	**10.3**

Table 4: *SI-SDR [dB] comparison of our implementations of other benchmark networks on the WHAM! separate-clean and separate-noisy tasks*

		separate-clean		separate-noisy	
Model	Dataset	Output	Δ	Output	Δ
chimera++	8 kHz min	11.0	11.0	5.4	9.9
TasNet-BLSTM	8 kHz min	12.5	12.5	5.3	9.8
chimera++	16 kHz max	9.6	9.6	4.4	10.2
1-d conv.	16 kHz max	6.9	6.9	3.0	8.8

WHAMR! Dataset

■ WHAMR!

- An extension of the WHAM! dataset.
- RIR were generated and convolved using pyroomacoustics (a python package) [1].
  - **clean** – anechoic clean mixture to anechoic sources
  - **noisy** – anechoic noisy mixture to anechoic sources
  - **reverberant** – reverberant clean mixture to anechoic sources
  - **noisy and reverberant** – reverberant noisy mixture to anechoic sources
- Fully Overlap, Utterance-level
- Single-Channel

Table 1. Room impulse response parameter sampling distributions. Units for all parameters are meters with the exception of reverberation time (T_{60}) which is in seconds and angles in radians.

Room	L	$\mathcal{U}(5, 10)$
	W	$\mathcal{U}(5, 10)$
	H	$\mathcal{U}(3, 4)$
$\mathbf{T_{60}}$	high	$\mathcal{U}(0.4, 1.0)$
	med.	$\mathcal{U}(0.2, 0.6)$
	low	$\mathcal{U}(0.1, 0.3)$

Mic. Center	L	$\frac{L_{\text{Room}}}{2} + \mathcal{U}(-0.2, 0.2)$
	W	$\frac{W_{\text{Room}}}{2} + \mathcal{U}(-0.2, 0.2)$
	H	$\mathcal{U}(0.9, 1.8)$
Mic. Array	sep.	noise mic. separation
	θ	$\mathcal{U}(0, 2\pi)$
Sources	H	$\mathcal{U}(0.9, 1.8)$
	dist.	$\mathcal{U}(0.66, 2)$
	θ	$\mathcal{U}(0, 2\pi)$

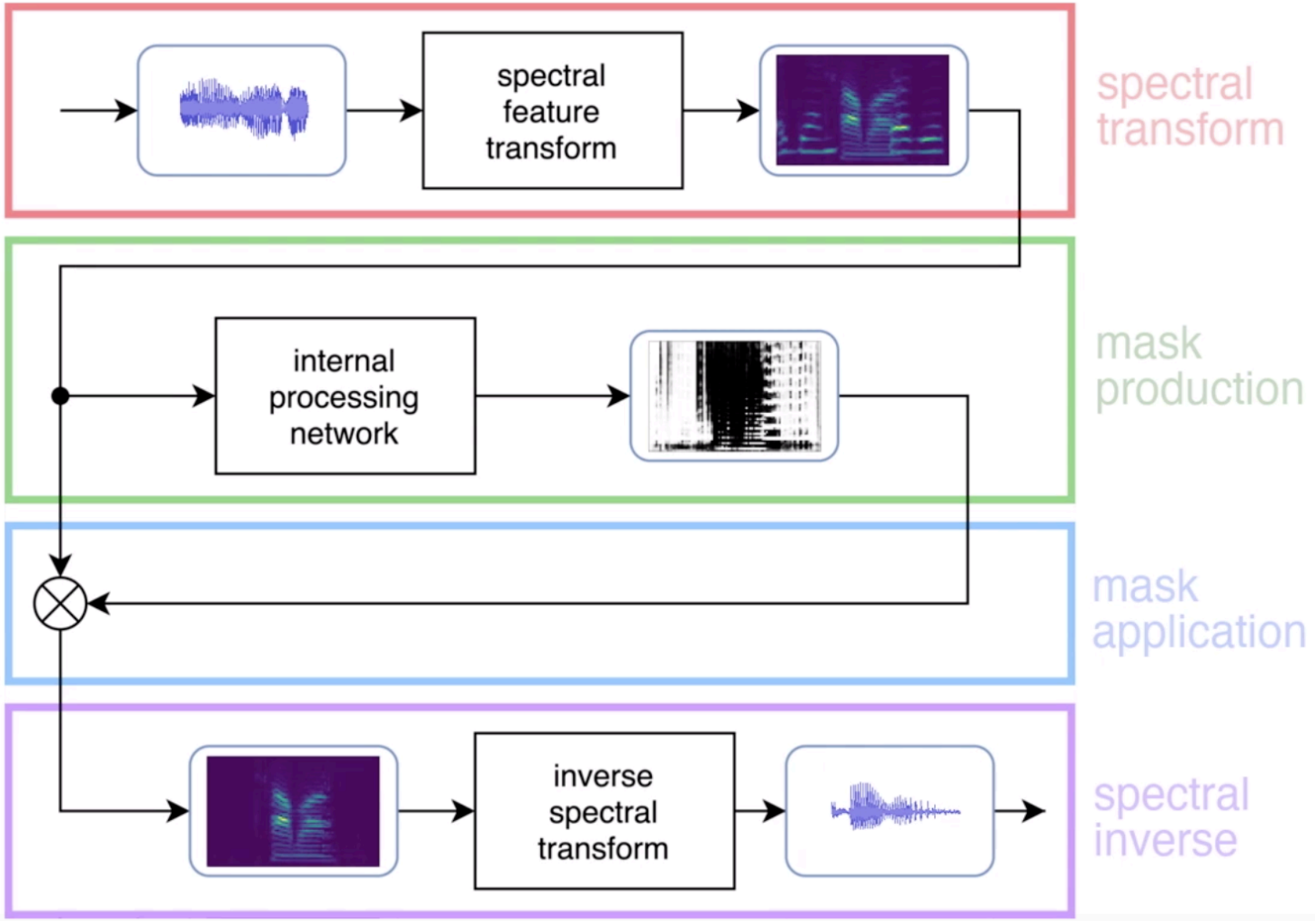
■ Output data organization

1. noise
2. s1_anechoic
3. s2_anechoic
4. s1_reverb
5. s2_reverb
6. mix_single_anechoic
7. mix_clean_anechoic
8. mix_both_anechoic
9. mix_single_reverb
10. mix_clean_reverb
11. mix_both_reverb

[1] Pyroomacoustics: A python package for audio room simulation and array processing algorithms

Single System on WHAMR!

- Paired transforms between waveform and time-frequency spectral domain
- Network produces spectral mask which suppresses interfering sources or noise/reverberation



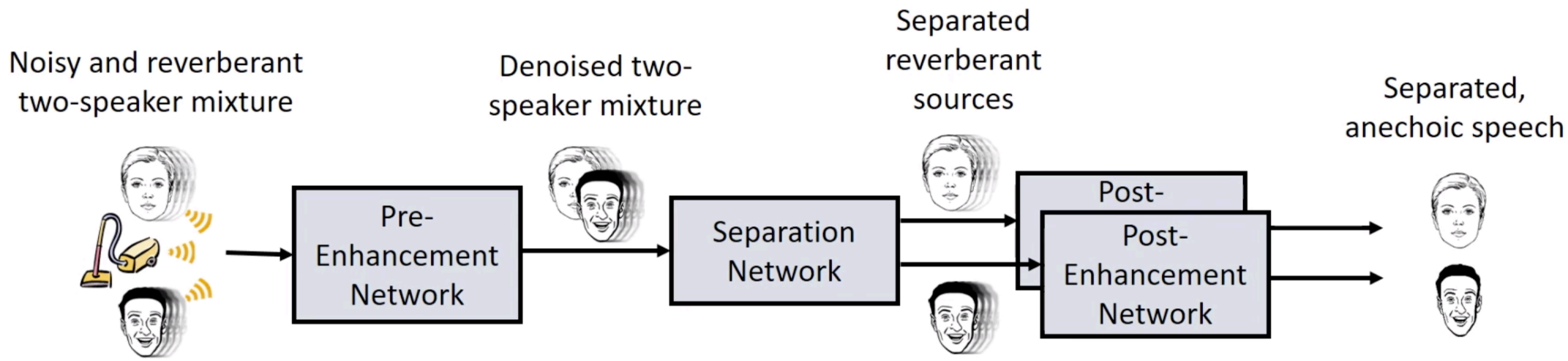
Results on Single System

SI-SDR [dB] of Core Separation Conditions using Single Model

	Input			Conv-TasNet		TasNet-BLSTM	
	Noise	Reverb	Input	Output	Δ	Output	Δ
			0.0	12.9	12.9	**14.2**	**14.2**
	✓		−4.5	7.0	11.5	**7.5**	**12.0**
		✓	−3.3	4.3	7.6	**5.6**	**8.9**
	✓	✓	−6.1	2.2	8.3	**3.0**	**9.2**

Cascaded Systems on WHAMR!

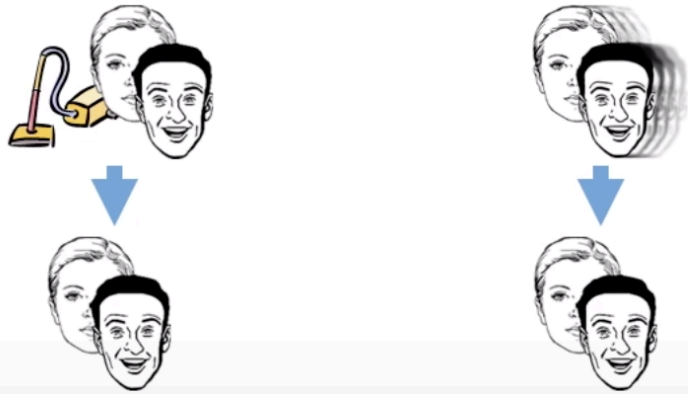
Cascaded Systems (example)



Results on Enhancement Part

SI-SDR [dB] of Enhancement of Overlapping Speech

Net		Denoise		Dereverb	
Feature	Processor	Output	Δ	Output	Δ
Learned	TCN	10.8	9.6	7.2	3.2
Learned	BLSTM	**11.2**	**10.1**	**8.5**	**4.4**
STFT	TCN	8.4	7.2	4.0	0.0
STFT	BLSTM	9.5	8.4	5.9	1.8
Input SI-SDR:		1.2		4.0	



Results on Cascaded Systems

Table 5. Comparison of cascaded models. A dash indicates speech separation without denoising/dereverberation, while × indicates no enhancement sub-model was used. Results are sorted by increasing performance. The highlighted rows indicate the non-cascaded single-model baseline.

System		SI-SDR	
Pre-Enh. Removes	Separate Speech while Removing	Output	Δ
×	noise	7.5	12.0
noise	–	**8.1**	**12.6**
Input SI-SDR:		−4.5	

(a) noisy condition

System			SI-SDR	
Pre-Enh. Removes	Separate Speech while Removing	Post-Enh. Removes	Output	Δ
×	rev.	×	5.6	8.9
rev.	–	×	6.4	9.7
×	–	rev.	**6.6**	**9.9**
Input SI-SDR:			−3.3	

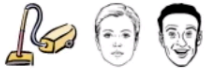
(b) reverberant condition

System			SI-SDR	
Pre-Enh. Removes	Separate speech while removing	Post-Enh. Removes	Output	Δ
×	noise, rev.	×	3.0	9.2
noise	rev.	×	3.5	9.7
noise, rev.	–	×	3.6	9.7
rev.	noise	×	3.7	9.8
×	noise	rev.	3.7	9.8
noise	–	rev.	**4.0**	**10.1**
Input SI-SDR:			−6.1	

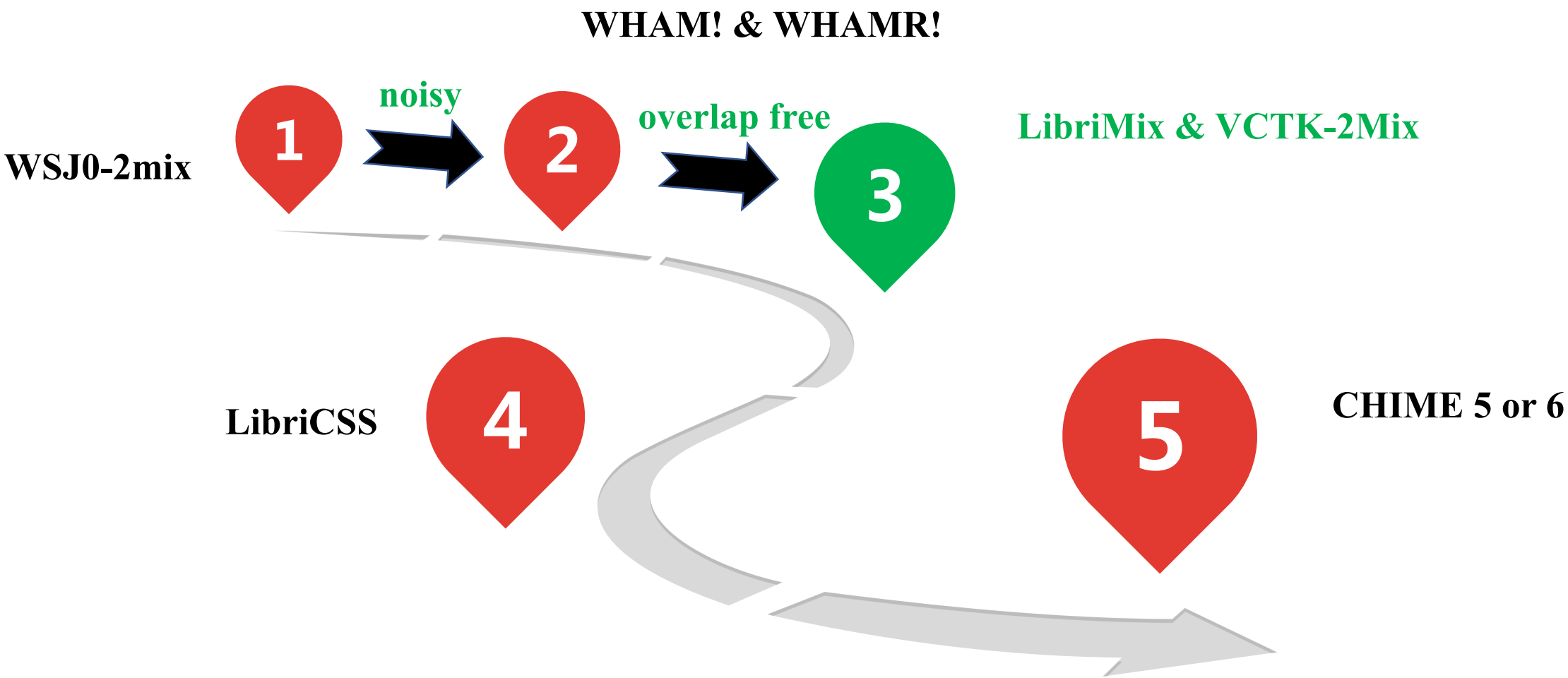
(c) noisy and reverberant condition

Results on Tuned Cascaded Systems

SI-SDR [dB] of Tuned Cascaded Systems

	Input			Best System w/o Tuning		Tuned	
	Noise	Reverb	Input	Output	Δ	Output	Δ
			0.0	14.2	14.2	–	–
	✓		−4.5	8.1	12.6	8.3	12.9
		✓	−3.3	6.6	9.9	7.0	10.3
	✓	✓	−6.1	4.0	10.1	4.7	10.8

Outline


WHAM! & WHAMR!
WSJ0-2mix
1
noisy
2
overlap free
3
LibriMix & VCTK-2Mix
LibriCSS
4
5
CHIME 5 or 6

LibriMix & VCTK-2Mix Dataset

Table 1: *Statistics of original speech datasets.*

Dataset	Split	Hours	per-spk minutes	# Speakers
WSJ0	si_tr_s	25	15	101
	si_dt_05	1.5	11	8
	si_et_05	2.3	14	10
LibriSpeech clean	train-360	364	25	921
	train-100	101	25	251
	dev	5.4	8	40
	test	5.4	8	40
VCTK	test	44	24	109

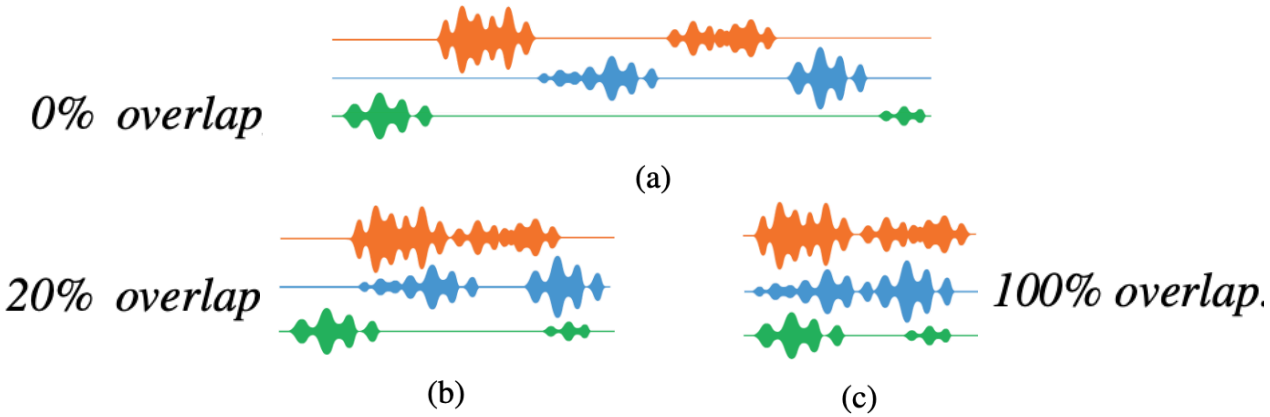


Table 2: *Statistics of derived speech separation datasets.*

Dataset	Split	# Utterances	Hours
wsj0-{2,3}mix	train	20,000	30
	dev	5,000	8
	test	3,000	5
Libri2Mix *[2]*	train-360	50,800	212
	train-100	13,900	58
	dev	3,000	11
	test	3,000	11
Libri3Mix *[2]*	train-360	33,900	146
	train-100	9,300	40
	dev	3,000	11
	test	3,000	11
SparseLibri2Mix *[4]*	test	3,000	6
SparseLibri3Mix *[4]*	test	3,000	6
VCTK-2mix *[3]*	test	3,000	9

[1] LibriMix: An Open-Source Dataset for Generalizable Speech Separation
[2] https://github.com/JorisCos/LibriMix
[3] https://github.com/JorisCos/VCTK-2Mix
[4] https://github.com/popcornell/SparseLibriMix

Results

■ Results on LibriMix

Table 4: *SI-SDR$_i$ (dB) achieved on LibriMix (SI-SDR for the "Input" column).*

	mode	Input	IRM	IBM	Conv-TasNet
2spk-C	8k min	0.0	12.9	13.7	14.7
	16k max	0.0	14.1	14.5	16
2spk-N	8k min	-2.0	12	12.6	12
	16k max	-2.8	13.4	13.7	13.5
3spk-C	8k min	-3.4	13.1	13.9	12.1
	16k max	-3.7	14.5	14.9	13
3spk-N	8k min	-4.4	12.6	13.3	10.4
	16k max	-5.2	14.1	14.4	10.9

■ Results on SparseLibriMix

Table 5: *SI-SDR$_i$ (dB) achieved on SparseLibriMix (8kHz). Conv-TasNet is abreviated TCN.*

	2spk-C		2spk-N		3spk-C		3spk-N	
Overlap	IRM	TCN	IRM	TCN	IRM	TCN	IRM	TCN
0%	43.7	31.9	16.1	14.5	44.2	24.8	18.7	13.0
20%	19.6	20.0	14.7	13.9	18.1	15.8	15.6	12.1
40%	16.2	17.6	13.8	13.2	16.4	14.4	14.9	11.7
60%	14.9	16.3	13.3	12.7	15.5	13.8	14.4	11.5
80%	14.2	15.7	13	12.5	14.6	13.1	13.9	11
100%	13.8	15.3	12.7	12.2	14.3	12.5	13.6	10.7

Results

■ Clean separation task

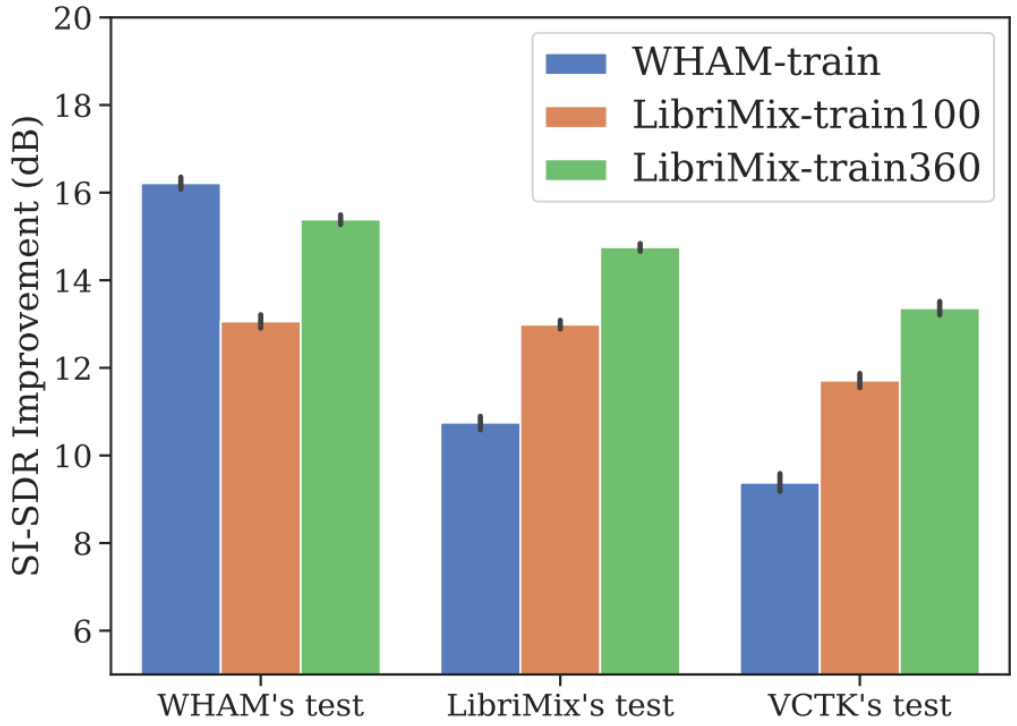


Figure 2: *Cross-dataset evaluation on the clean separation task. Errors bars indicate 95% confidence intervals.*

■ Noisy separation task

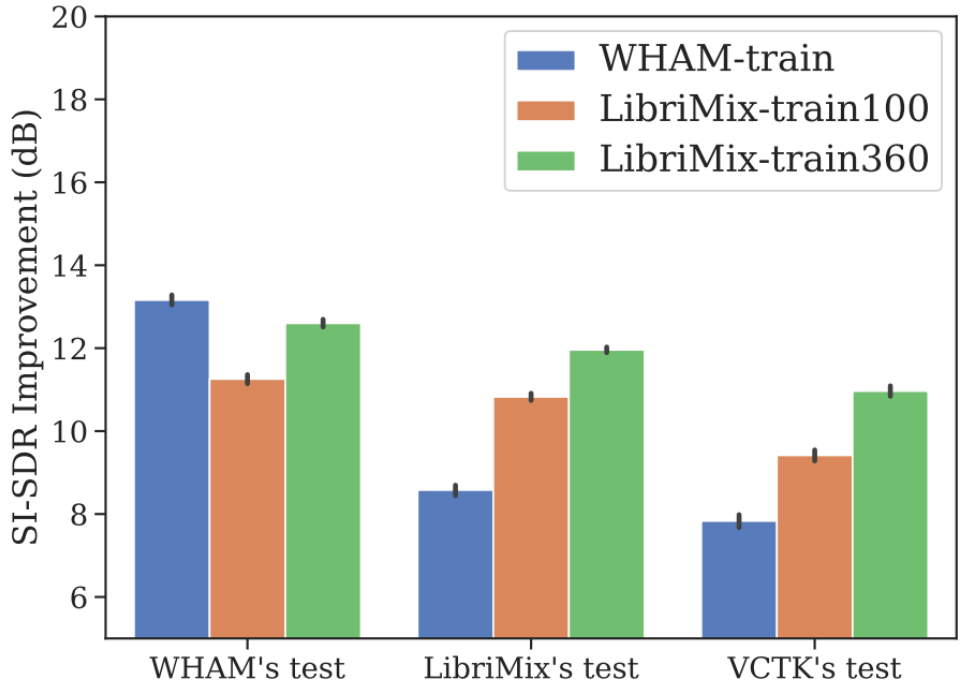
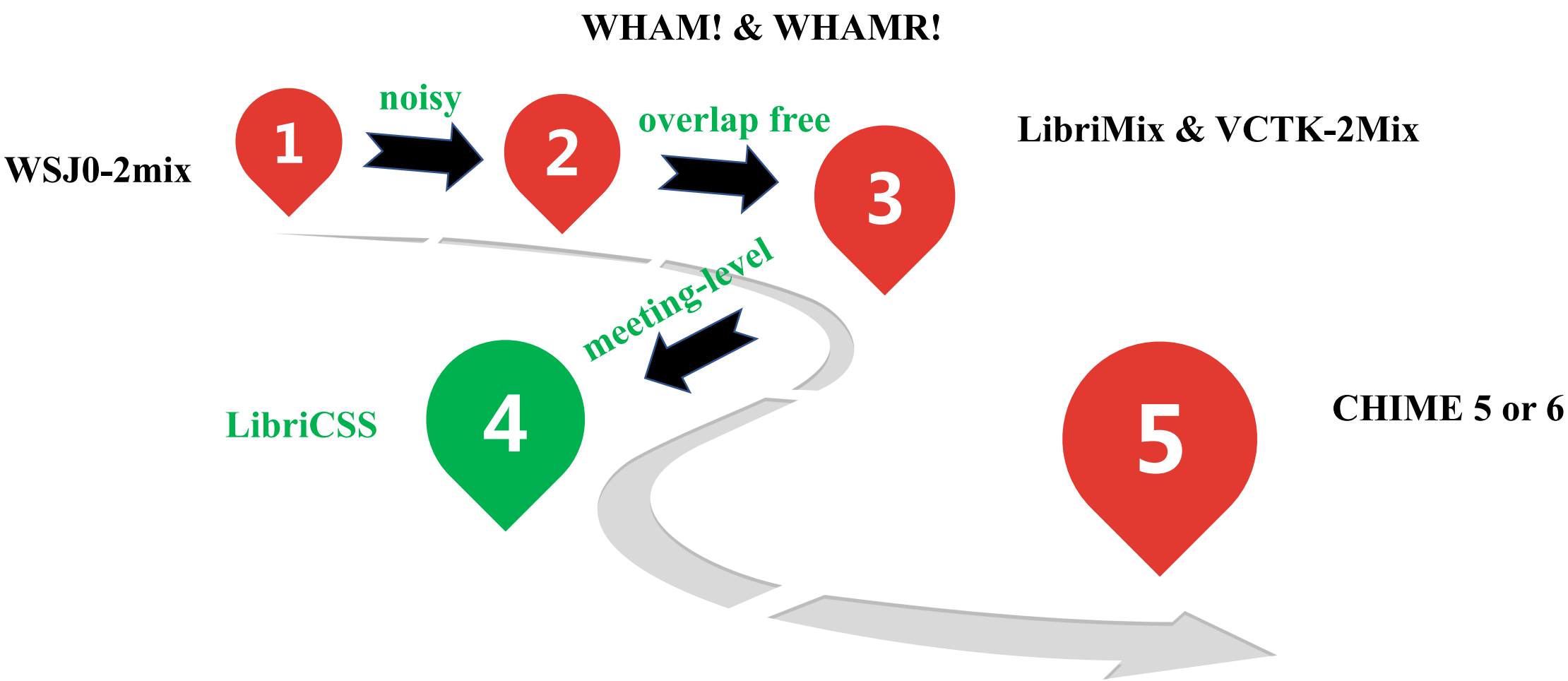


Figure 3: *Cross dataset evaluation on the noisy separation task. Errors bars indicate 95 % confidence interval*

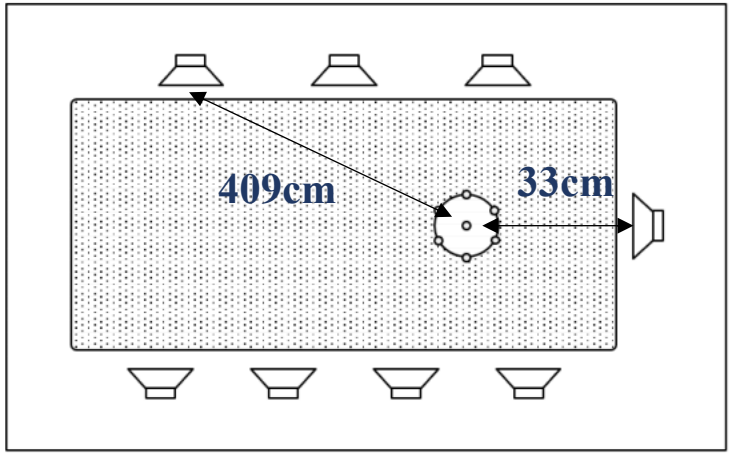
Outline

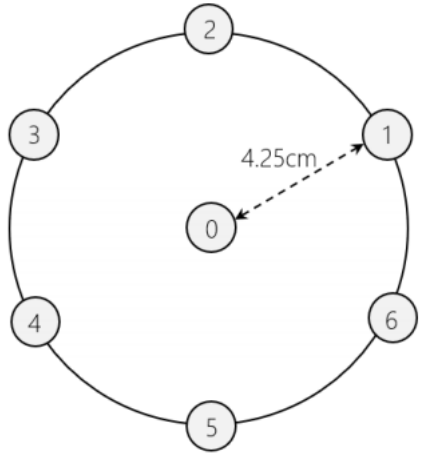

WHAM! & WHAMR!
WSJ0-2mix
1
noisy
2
overlap free
3
LibriMix & VCTK-2Mix
meeting-level
LibriCSS
4
5
CHIME 5 or 6

LibriCSS Dataset

■ LibriCSS

- ➢ Each utterance is taken from LibriSpeech and played back from a loudspeaker placed in a room.
- ➢ Overlap-free, Meeting-wise, Multi-channel
- ➢ Three features:
  - ✓ It is recorded in a room instead of being generated by simulation. The simulated data tend to oversimplify room acoustics especially in multi-channel scenarios.
  - ✓ The dataset encompasses different overlap ratios and silence settings to help analyze how different algorithms work under various overlap conditions
  - ✓ The audio signals are continuously recorded to enable CSS evaluation. Meanwhile, the ground-truth segmentation is also provided, allowing for the conventional utterance-wise evaluation





[1] CONTINUOUS SPEECH SEPARATION: DATASET AND ANALYSIS

LibriCSS Dataset

LibriCSS (more details)

- **10 hours, 10 sessions**. So each session is approximately one hour long.
- **Each session = 6 * 10 min "min sessions"**
- Each min session has **different overlap ratios** (OVRs), ranging from 0 to 40%

$$\mathrm{OVR} = L_{\mathrm{ovl}} / L_{\mathrm{all}}$$

- Each min session has **8 speakers** from 40 speakers in Librispeech "test clean" set
- The total number of **Utts** in each mini session ranges **from 52 to 125**
- **Short silence version**: the inter-utterance silence length is sampled in [0.1s, 0.5s]
- **Long silence version**: the inter-utterance silence length is sampled in [2.9s, 3.0s]

[1] CONTINUOUS SPEECH SEPARATION: DATASET AND ANALYSIS

Results on LibriCSS Dataset

■ Utterance-wise evaluation

- In the utterance-wise evaluation, each utterance is extracted by using ground-truth segmentation information.
- **Alignment (using cross correlation) → cut → separation → ASR**

Table 1. %WERs for utterance-wise evaluation. 0S: 0% overlap with short inter-utterance silence. 0L: 0% overlap with long inter-utterance silence. Our ASR system yields WERs of 4.9% and 5.1% for anechoic versions of 0S and 0L utterances.

System	Overlap ratio in %					
	0S	0L	10	20	30	40
No separation	11.8	11.7	18.8	27.2	35.6	43.3
Mask (1ch)	12.7	12.1	17.6	23.2	30.5	35.6
Mask (7ch)	12.0	11.6	15.6	20.2	25.6	29.4
MVDR (7ch)	8.4	8.3	11.6	15.8	18.7	21.7

■ Continuous-input evaluation

- In the continuous input evaluation mode, separation and recognition are performed on an audio stream without splitting it into individual utterances
- **Perform long-segment-wise decoding instead of truly continuous ASR, about 60s ~ 120s**

Table 2. %WERs for seven-channel continuous input evaluation with different chunking configurations. The dash-separated three numbers of the first column are N_L, N_C, N_R values, respectively. Inherent latency is shown in parentheses.

System	Overlap ratio in %					
	0S	0L	10	20	30	40
No separation	15.4	11.5	21.7	27.0	34.3	40.5
1.2-0.8-0.4 (1.2 s)	11.9	9.7	13.6	15.0	19.9	21.9
1.6-0.8-0.0 (0.8 s)	12.2	9.7	14.7	16.1	20.5	23.1
0.8-0.4-0.4 (0.8 s)	11.5	9.5	13.4	15.8	19.7	21.2

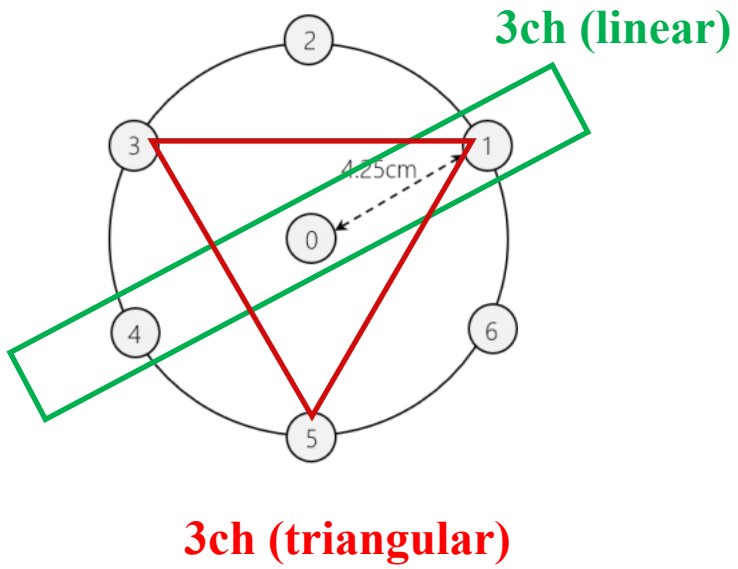
[1] CONTINUOUS SPEECH SEPARATION: DATASET AND ANALYSIS

Results on LibriCSS Dataset

- **The WERs for different microphone setups**

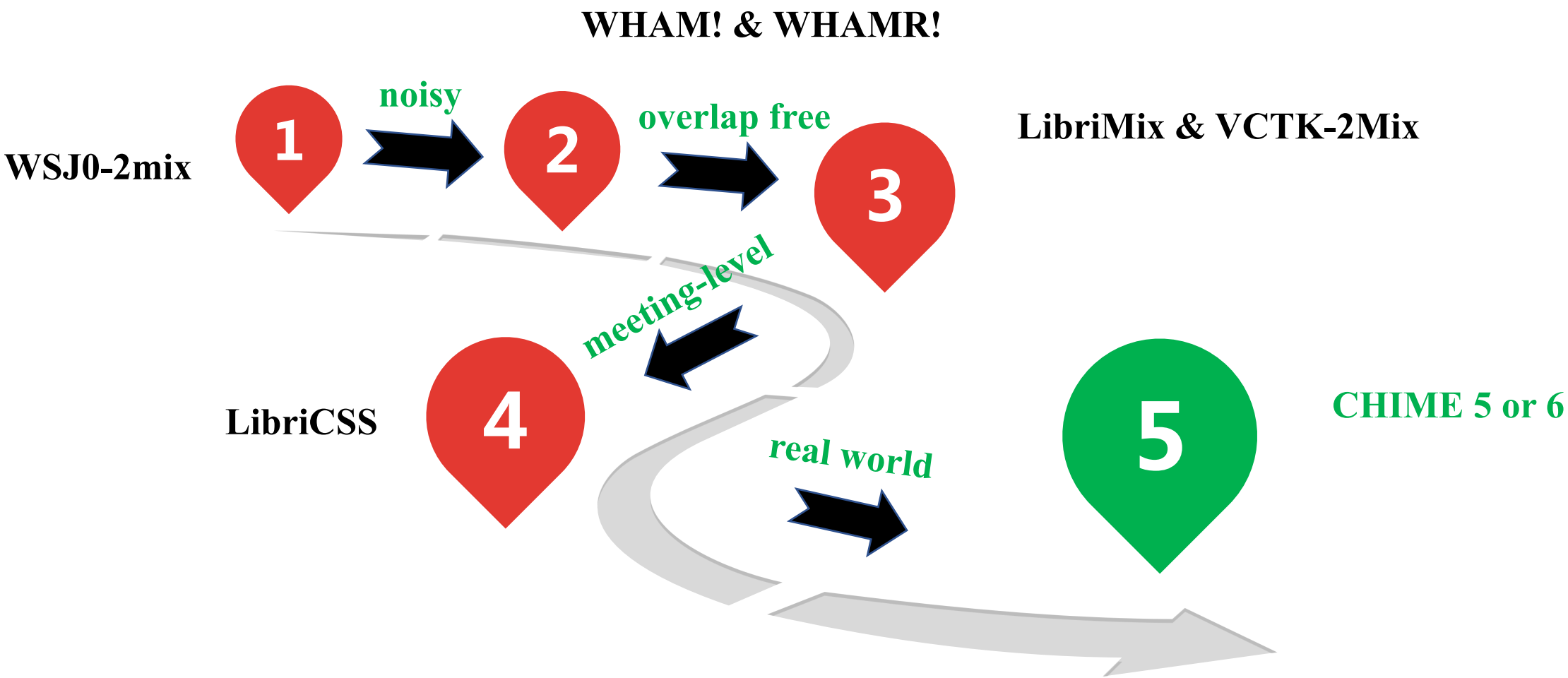
Table 3. %WER impact of number and arrangement of microphones in continuous input case. N_L, N_C, and N_R are set at equivalents of 1.2 s, 0.8 s, and 0.4 s, respectively.

	System	Overlap ratio in %					
		0S	0L	10	20	30	40
	7ch	11.9	9.7	13.6	15.0	19.9	21.9
[0,1,2,4,5]	5ch	12.8	10.5	15.3	17.4	22.8	26.4
[1,3,5]	3ch (triangular)	15.8	10.8	19.4	23.1	28.9	36.0
[1,0,4]	3ch (linear)	15.1	9.8	17.7	20.6	30.1	29.6
	1ch	17.6	16.3	20.9	26.1	32.6	36.1



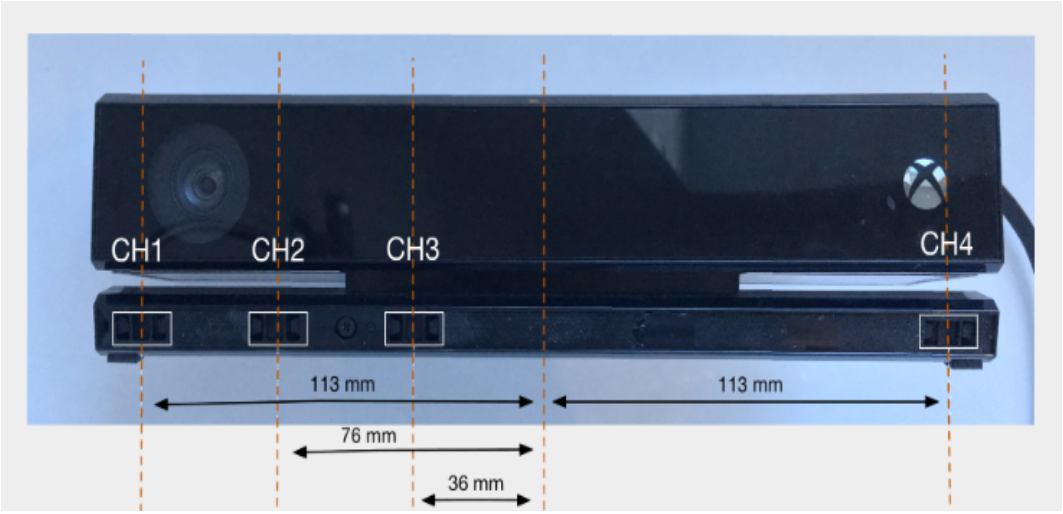
[1] CONTINUOUS SPEECH SEPARATION: DATASET AND ANALYSIS

Outline


WHAM! & WHAMR!
WSJ0-2mix
1
noisy
2
overlap free
3
LibriMix & VCTK-2Mix
meeting-level
LibriCSS
4
real world
5
CHIME 5 or 6

CHiME 5/6 Dataset

- CHiME 6 is built from CHiME 5, which targets the problem of distant microphone conversational speech recognition in everyday home environments. (real cases: no target label)
- Each party has been recorded with a set of six Microsoft Kinect devices. Each Kinect device has a linear array of 4 sample-synchronised microphones and a camera.
- In addition to the Kinects, to facilitate transcription, each participant is wearing a set of Soundman OKM II Classic Studio binaural microphones. The audio from these is recorded via a Soundman A3 adapter onto Tascam DR-05 stereo recorders being worn by the participants.



[1] The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines
[2] CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings

CHiME 5/6 Dataset

- Name format:
  - Binaural microphones: <session ID>_<speaker ID>.wav , e.g., S02_P05.wav
  - Array microphone: <session ID>_<array ID>.CH<channel ID>.wav , e.g., S02_U05.CH1.wav

- Session ID ("session_id")
- Location ("kitchen", "dining", or "living")
- Speaker ID ("speaker")
- Transcription ("words")
- Start time ("start_time")
  - For the binaural microphone recording of that speaker ("original")
  - For all array recordings ("U01", etc.)
  - For all binaural microphone recordings ("P01", etc.)
- End time ("end_time")
- Reference microphone array ID ("ref")

Dataset	Parties	Speakers	Hours	Utterances
Train	16	32	40:33	79,980
Dev	2	8	4:27	7,440
Eval	2	8	5:12	11,028

[1] The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines
[2] CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings

USTC System on CHiME 5

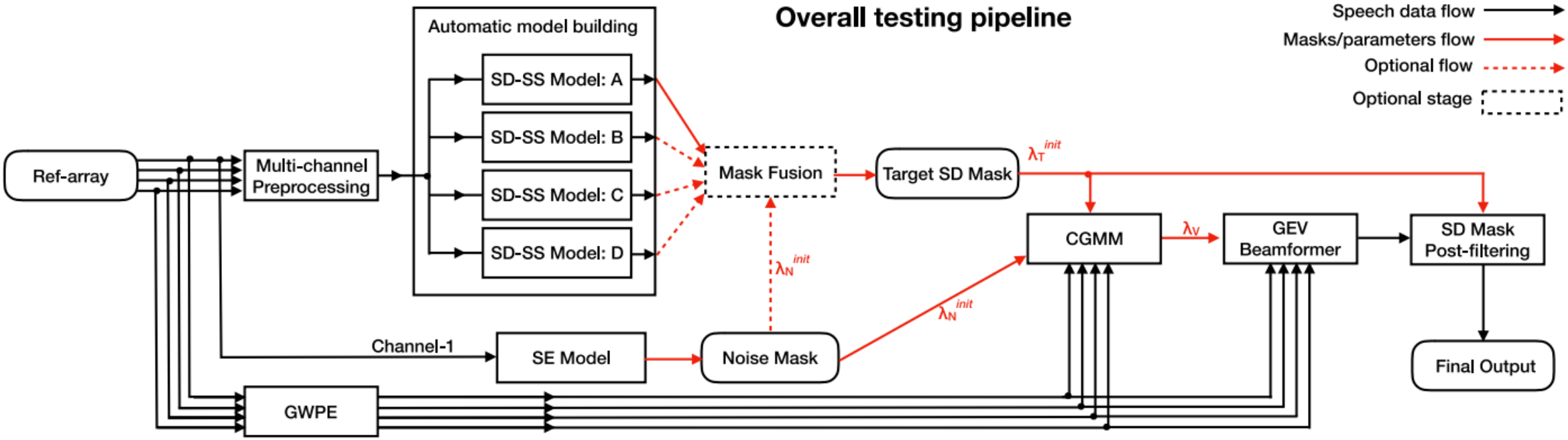
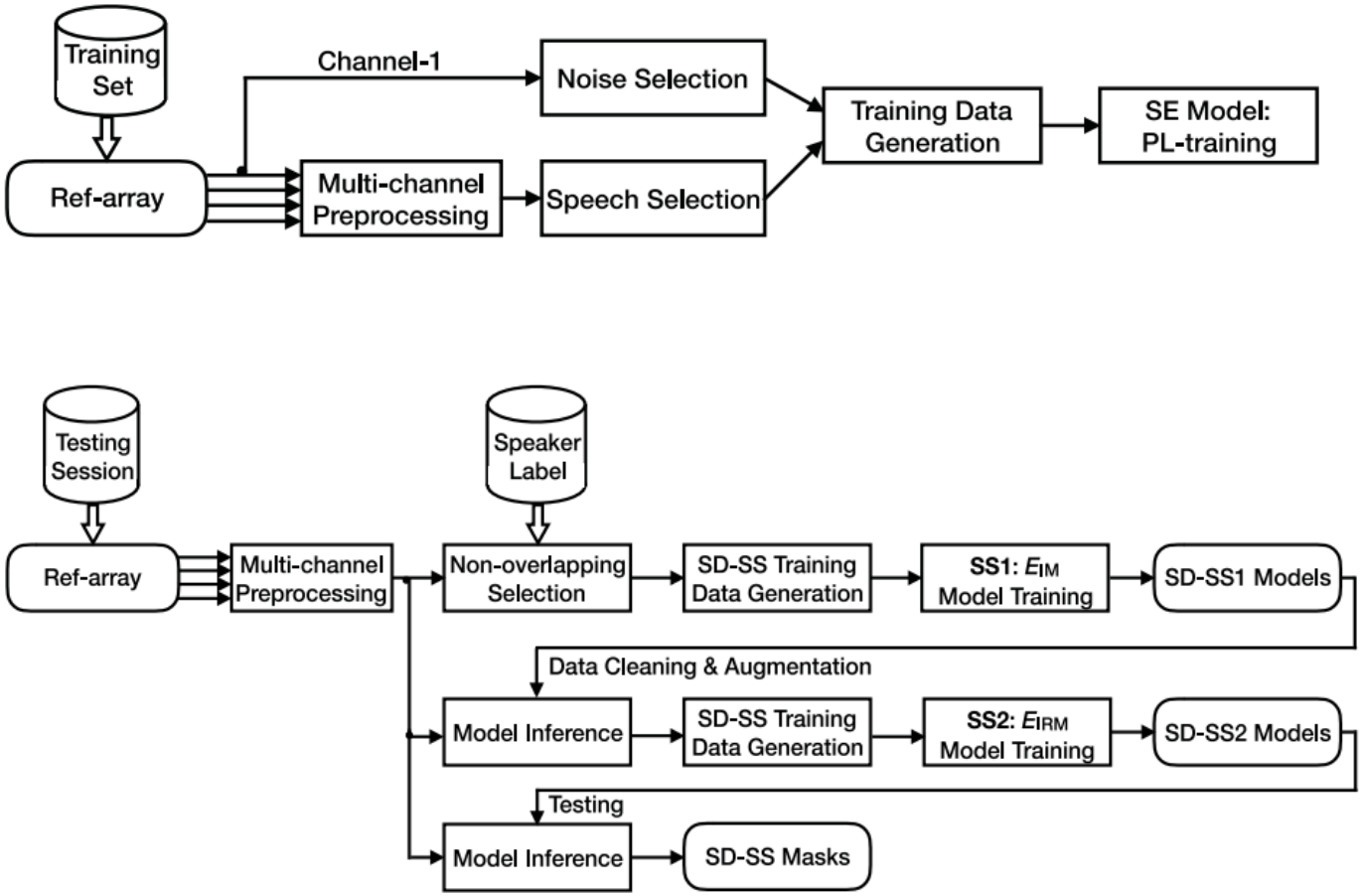


Fig. 2. The overall diagram of our proposed front-end processing system for the CHiME-5 challenge.

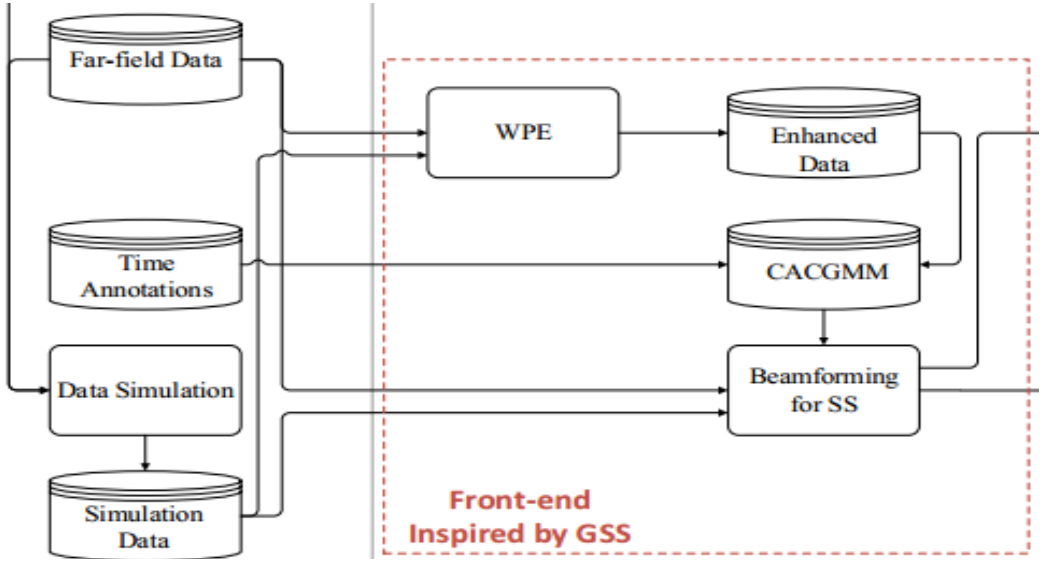
[1] A Speaker-Dependent Approach to Separation of Far-Field Multi-Talker Microphone Array Speech for Front-End Processing in the CHiME-5 Challenge

USTC System on CHiME 5



[1] A Speaker-Dependent Approach to Separation of Far-Field Multi-Talker Microphone Array Speech for Front-End Processing in the CHiME-5 Challenge

USTC System on CHiME 6



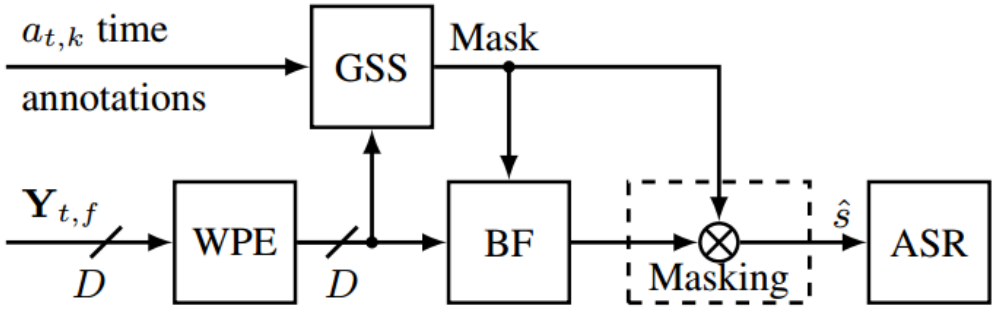


Figure 2: *Overview of speech enhancement system*

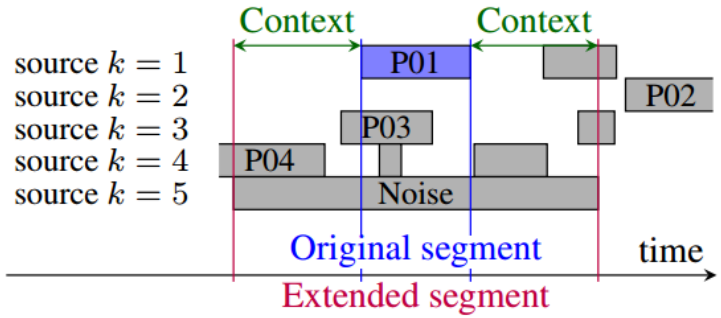


Figure 3: *Time annotation visualization. All utterance segments are aligned on the target array. Relative to a desired utterance is an extended segment selected for the enhancement.*

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	

[1] The USTC-NELSLIP Systems for CHiME-6 Challenge
[2] Guided source separation meets a strong ASR backend: Hitachi/paderborn university joint investigation for dinner party ASR

Summary

- Bridging the advanced research and the real world problems
  - Utterance-level → Meeting-level
  - Fully overlap → Overlap free
  - Single-Channel → Multi-Channel (e.g., Linear, Circle Array)
  - Clean Separation → Noisy Separation (e.g., Noise, Reverberant)
  - Utterance-wise Evaluation (e.g., PESQ, SDR, et. al) → Continuous Input Evaluation (e.g., WER)
  - Simulated Dataset → Real Dataset recorded in real environments


Thank You!