Guided Source Separation Meets a Strong ASR Backend: Hitachi/Paderborn University Joint Investigation for Dinner Party ASR

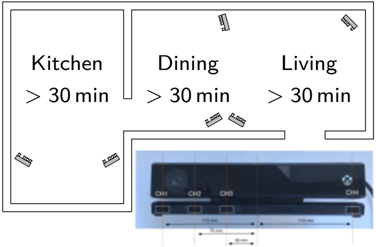
Naoyuki Kanda[1], Christoph Boeddeker[2], Jens Heitkaemper[2], Yusuke Fujita[1], Shota Horiguchi[1], Kenji Nagamatsu[1], Reinhold Haeb-Umbach[2]

[1]Hitachi Ltd., Japan
[2]Paderborn University, Germany

11:00, September 17, 2019

Table of contents

CHiME-5 Dataset: Dinner party automatic speech recognition



- 16+2+2 sessions with ca. 2 h
- 4 participants in each scenario
- 6 Kinect microphone arrays

Difficulties

- Natural conversation
- Overlap
- No simulated data, only in-ear microphone signals
- Realistic recording (e.g. device failure, lost samples)

Motivation

Baseline

81.1 % (DEV) 73.3 % (EVAL)

Motivation

Baseline
81.1 % (DEV) 73.3 % (EVAL)

Challenge

Paderborn (GSS)
61.7 % (DEV) 69.0 % (EVAL)
Pure frontend contribution

Motivation

Baseline
81.1 % (DEV) 73.3 % (EVAL)

Challenge

Challenge

Paderborn (GSS)
61.7 % (DEV) 69.0 % (EVAL)
Pure frontend contribution

Hitachi-JHU (ASR)
54.0 % (DEV) 50.6 % (EVAL)
Main gain from speech recognizer

Motivation

Baseline
81.1 % (DEV) 73.3 % (EVAL)

Challenge

Challenge

Paderborn (GSS)
61.7 % (DEV) 69.0 % (EVAL)
Pure frontend contribution

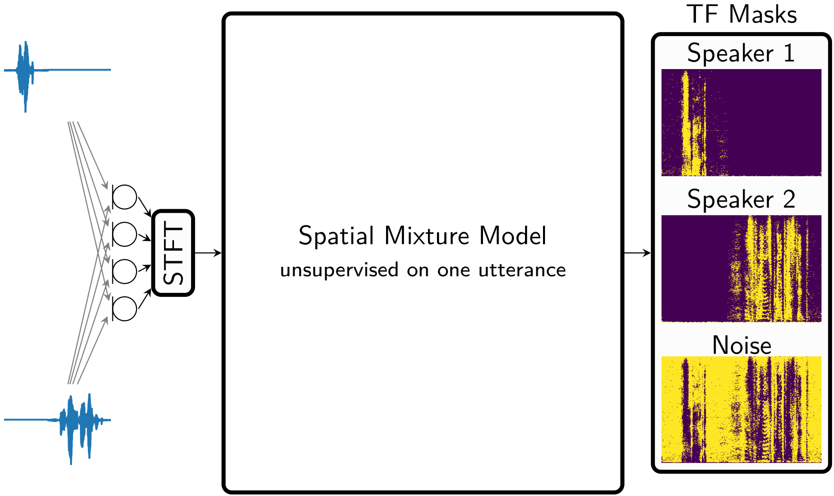
Hitachi-JHU (ASR)
54.0 % (DEV) 50.6 % (EVAL)
Main gain from speech recognizer

Here

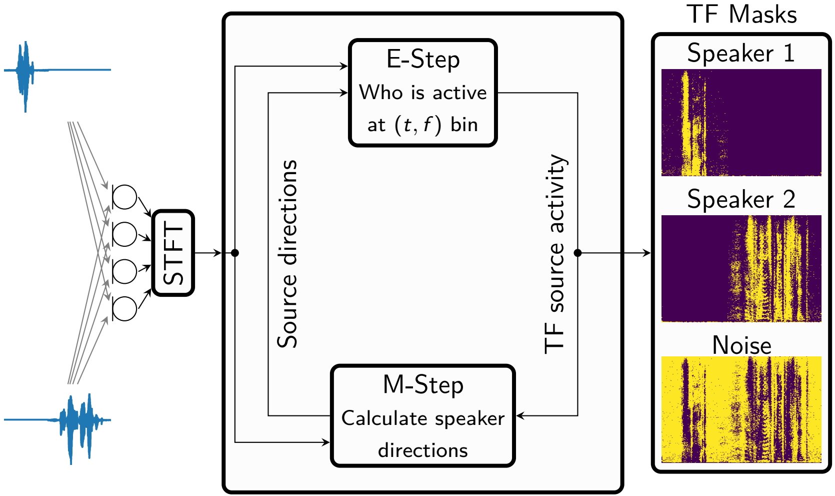
Here

Hitachi-Paderborn
GSS Meets a Strong ASR Backend
Interaction: Do they profit from each other?
Can we use external data to boost the performance?
Can we tune the system?

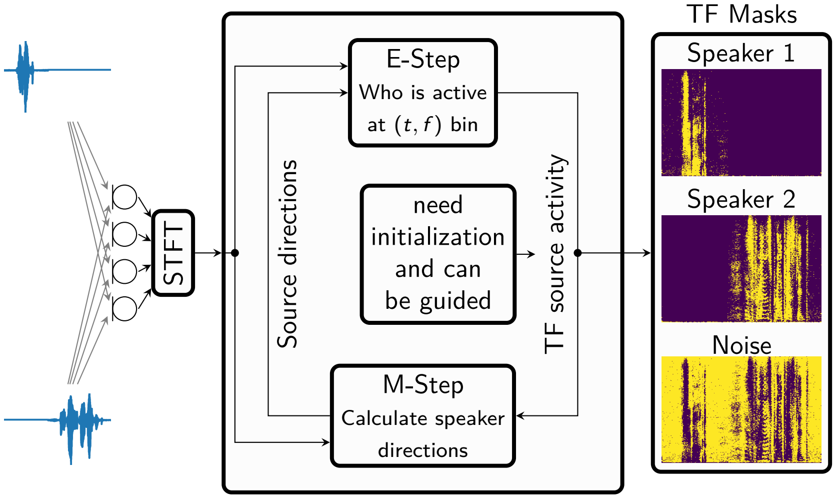
Guided Source Separation (GSS) – Spatial Mixture Model



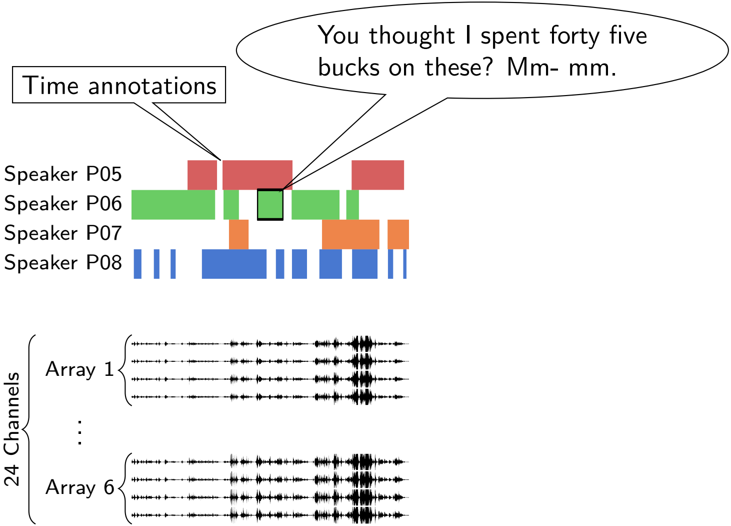
Guided Source Separation (GSS) – Spatial Mixture Model



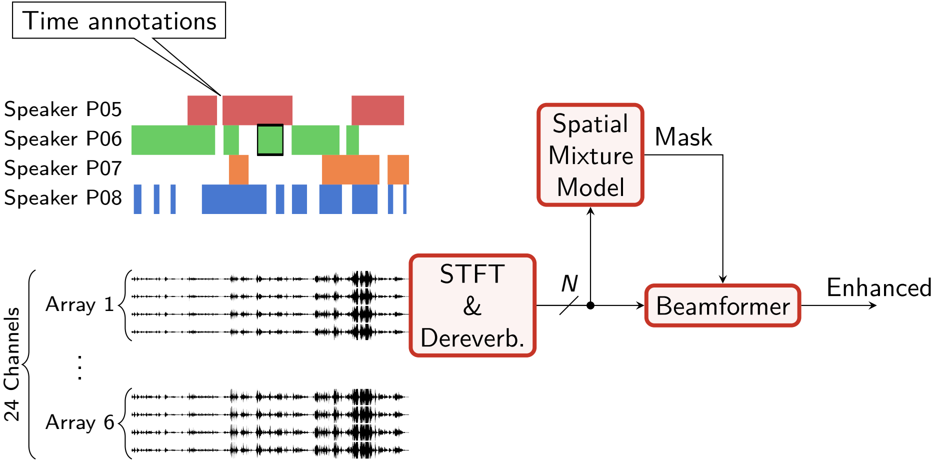
Guided Source Separation (GSS) – Spatial Mixture Model



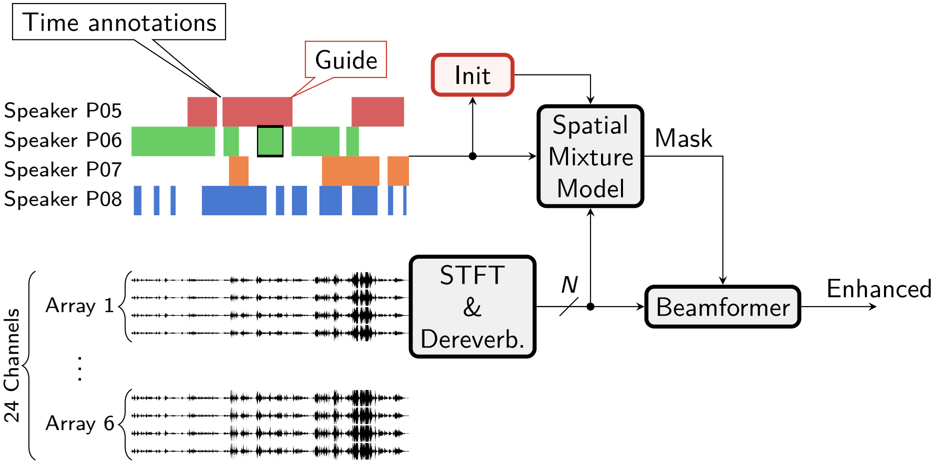
Guided Source Separation (GSS)



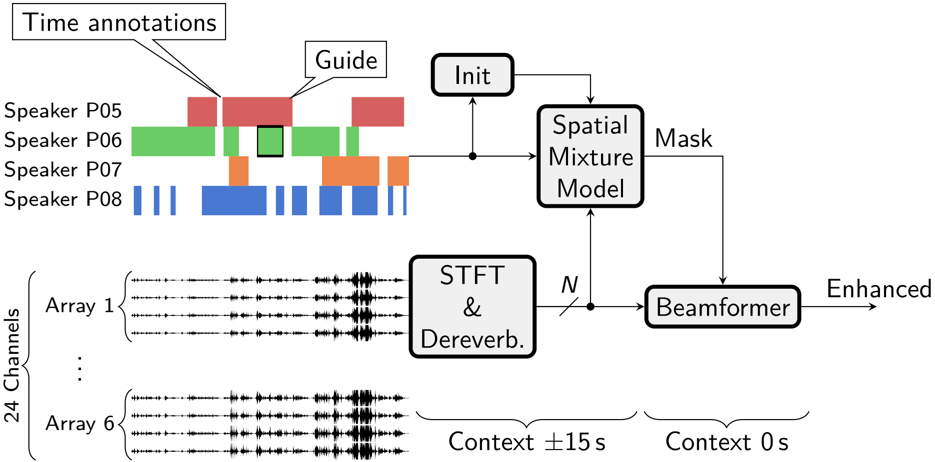
Guided Source Separation (GSS)



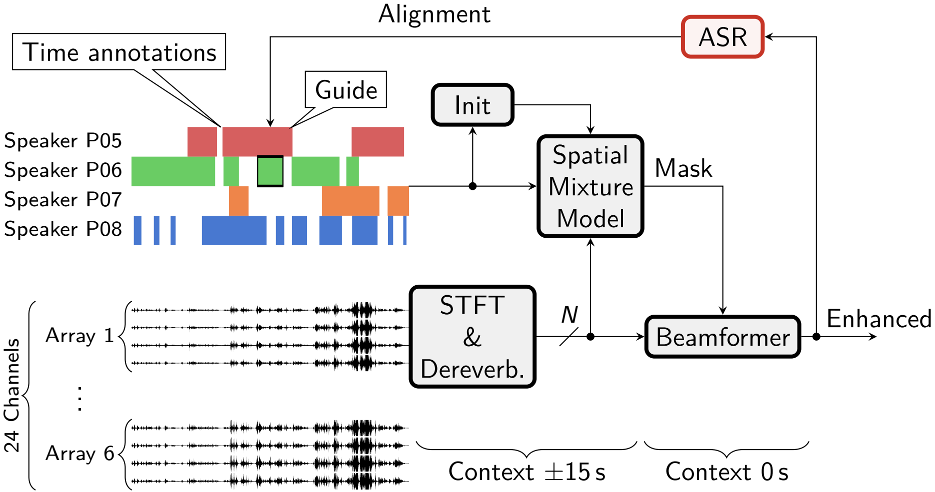
Guided Source Separation (GSS)



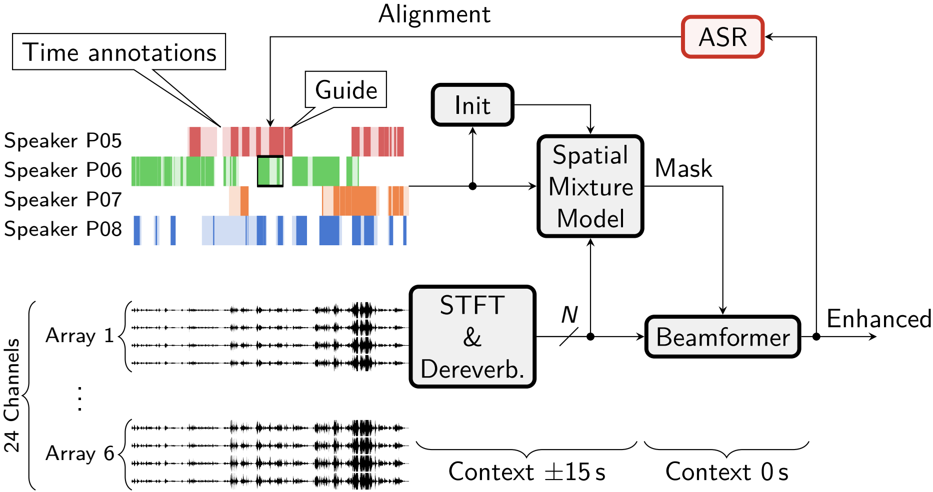
Guided Source Separation (GSS)



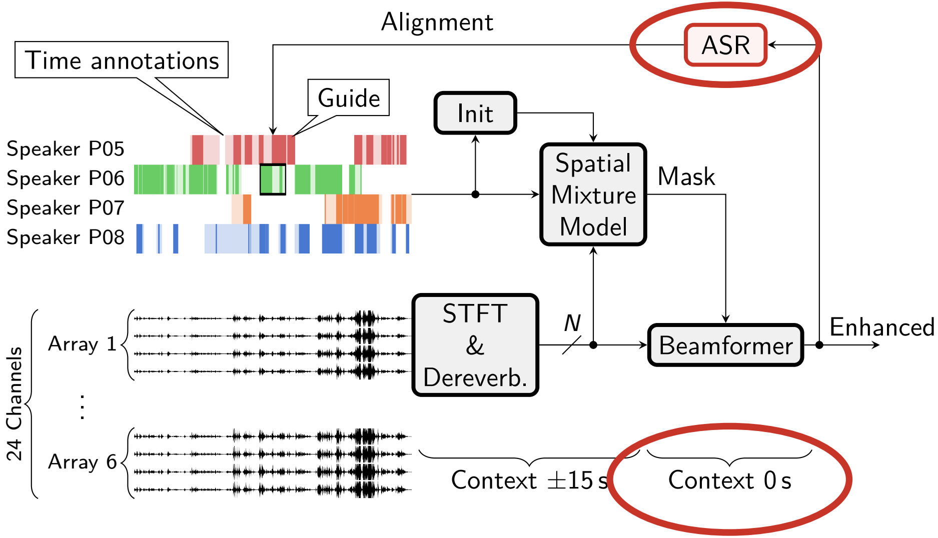
Guided Source Separation (GSS)



Guided Source Separation (GSS)

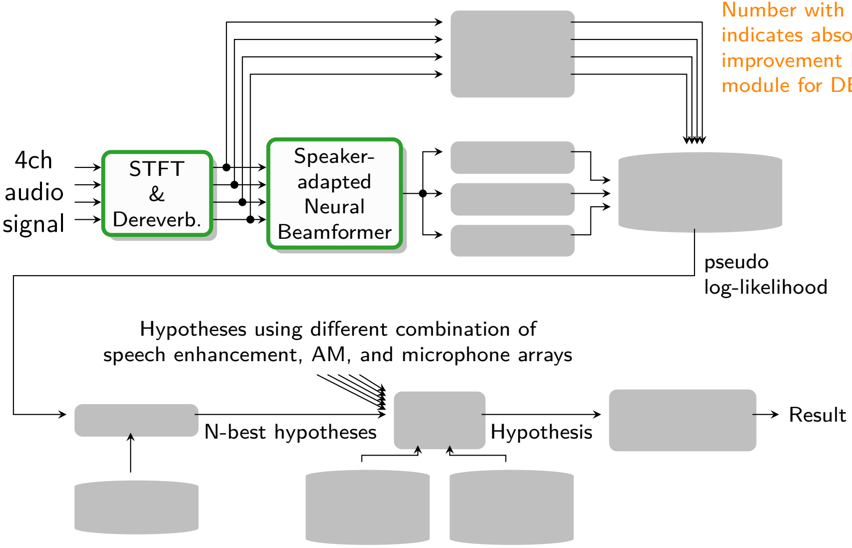


Guided Source Separation (GSS)

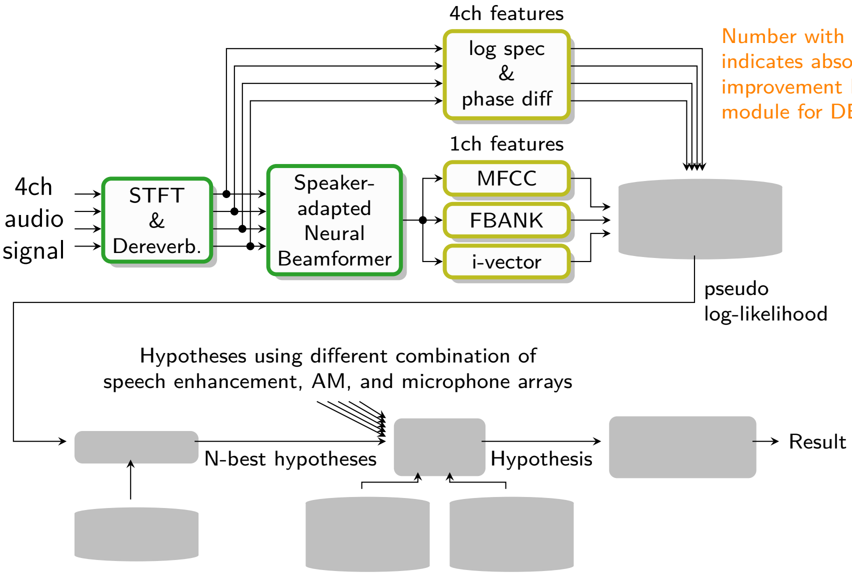


Automatic Speech Recognition (ASR)

Number with % indicates absolute WER improvement by that module for DEV set

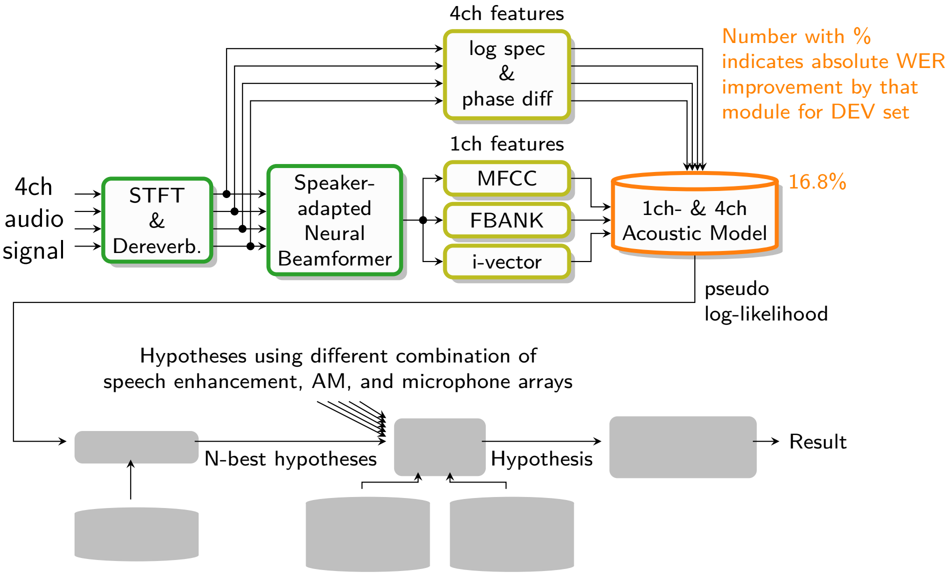


Automatic Speech Recognition (ASR)

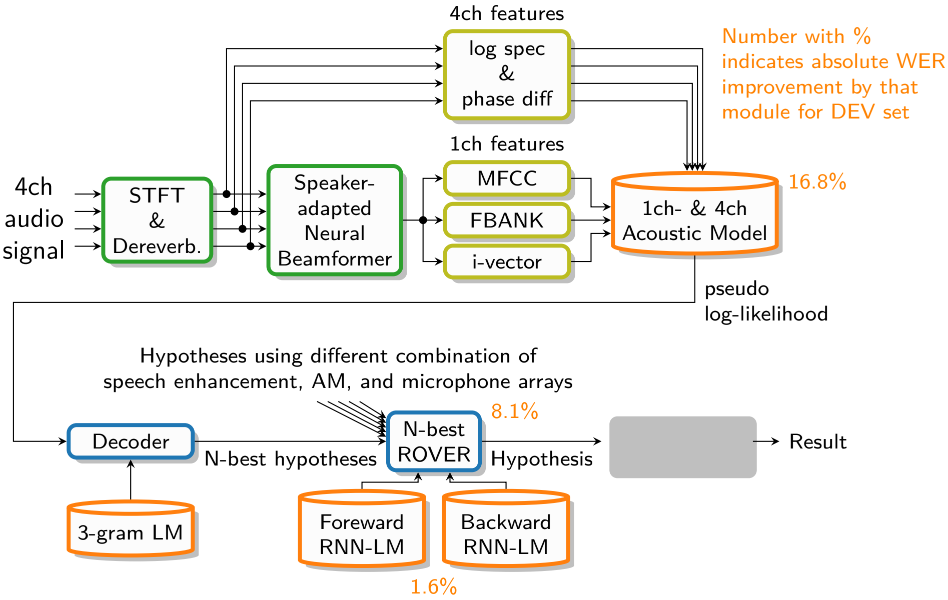


Number with % indicates absolute WER improvement by that module for DEV set

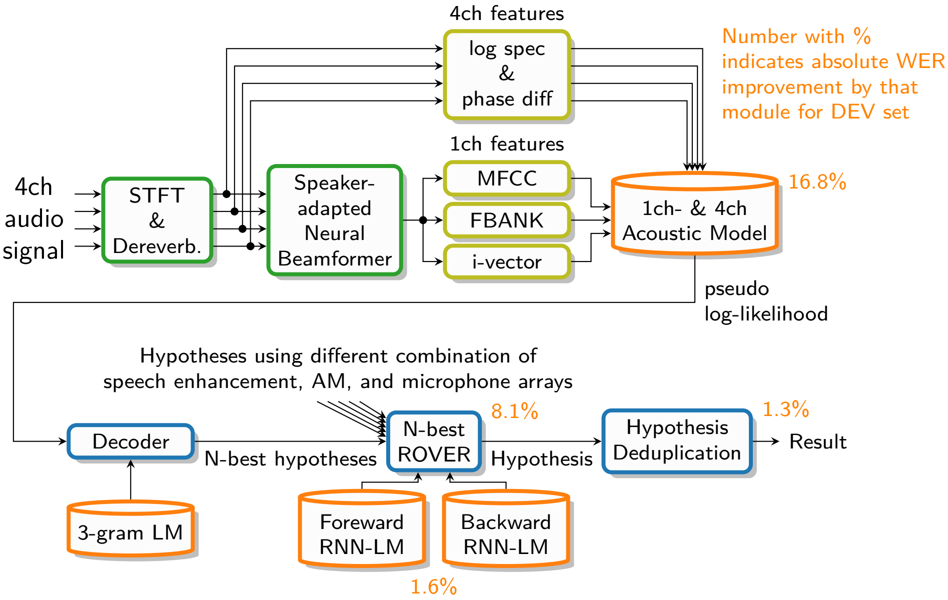
Automatic Speech Recognition (ASR)



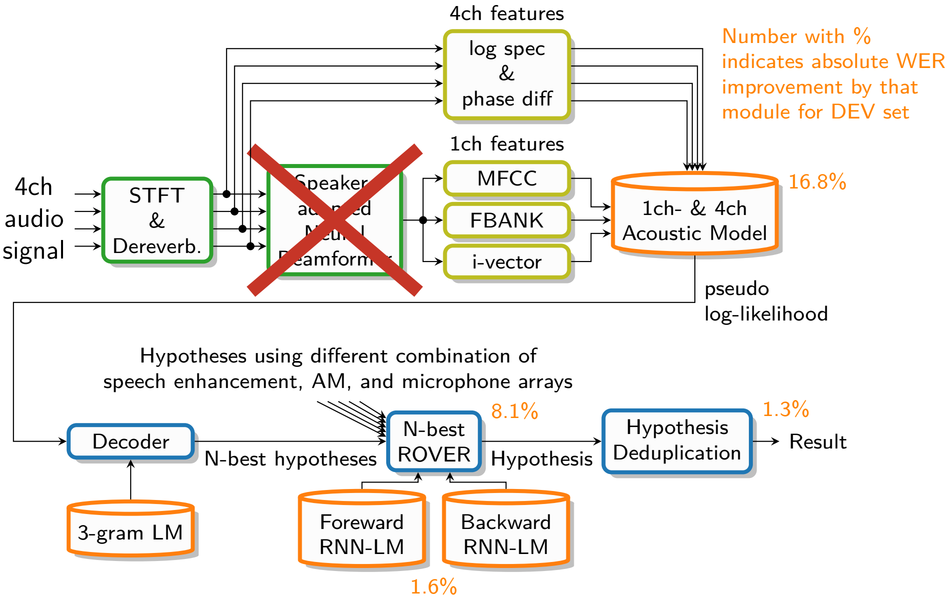
Automatic Speech Recognition (ASR)



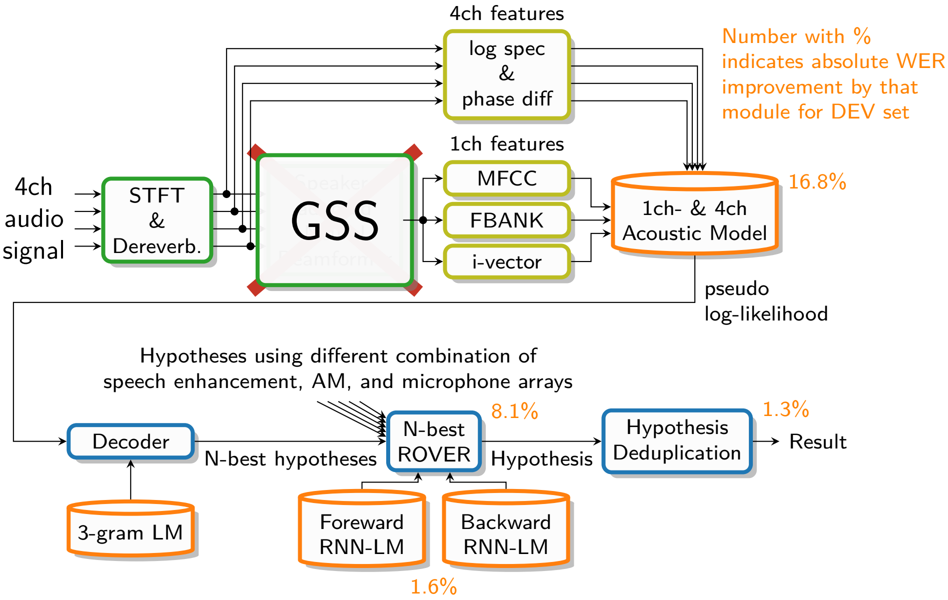
Automatic Speech Recognition (ASR)



Automatic Speech Recognition (ASR)

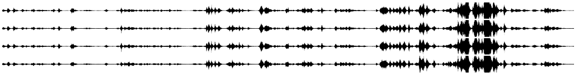


Automatic Speech Recognition (ASR)



ASR – Hypothesis Deduplication (HD)

Same words were sometimes recognized for overlapped utterances



um yeah

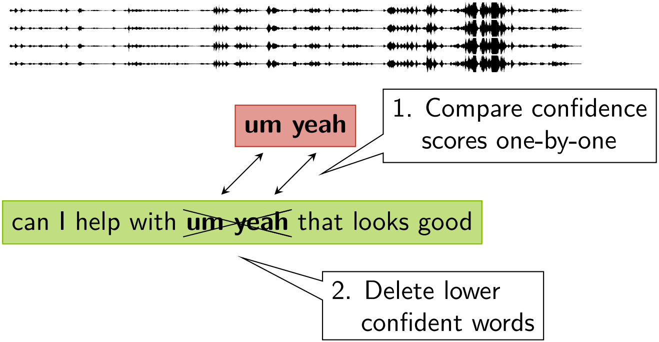
can I help with **um yeah** that looks good

Hypothesis deduplication:

- Compensates weak source separation system
- Indicator which source separation worked better

ASR – Hypothesis Deduplication (HD)

Same words were sometimes recognized for overlapped utterances

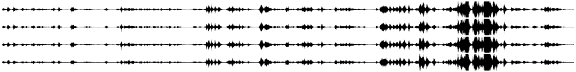


Hypothesis deduplication:

- Compensates weak source separation system
- Indicator which source separation worked better

ASR – Hypothesis Deduplication (HD)

Same words were sometimes recognized for overlapped utterances



um yeah

1. Compare confidence scores one-by-one

can I help with ~~**um yeah**~~ that looks good

2. Delete lower confident words

Hypothesis deduplication:

- Compensates weak source separation system
- Indicator which source separation worked better

Simulation results – WER – Challenge conform

	DEV (%)	EVAL (%)
Baseline (single)	81.1	73.3
USTC/iFlytek	45.0	46.1
Proposed	**39.94**	**41.64**

Speech enhancement (multi array – 24 channels)	DEV (%)	EVAL (%)
Hitachi-JHU	57.50	
Paderborn	49.21	
Paderborn + BF w/o context	46.54	51.99
Paderborn + BF w/o context + ASR feedback	45.14	47.29

Model combination	RNN-LM	HD	DEV (%)	EVAL (%)
			45.14	47.29
✓			41.67	43.70
✓	✓		**39.94**	**41.64**
✓	✓	✓	40.26	42.00

Simulation results – WER – Challenge conform

	DEV (%)	EVAL (%)
Baseline (single)	81.1	73.3
USTC/iFlytek	45.0	46.1
Proposed	**39.94**	**41.64**

Speech enhancement (multi array – 24 channels)	DEV (%)	EVAL (%)
Hitachi-JHU	57.50	
Paderborn	49.21	
Paderborn + BF w/o context	46.54	51.99
Paderborn + BF w/o context + ASR feedback	45.14	47.29

Model combination	RNN-LM	HD	DEV (%)	EVAL (%)
			45.14	47.29
✓			41.67	43.70
✓	✓		**39.94**	**41.64**
✓	✓	✓	40.26	42.00

Simulation results – WER – Challenge conform

	DEV (%)	EVAL (%)
Baseline (single)	81.1	73.3
USTC/iFlytek	45.0	46.1
Proposed	**39.94**	**41.64**

Speech enhancement (multi array – 24 channels)	DEV (%)	EVAL (%)
Hitachi-JHU	57.50	
Paderborn	49.21	
Paderborn + BF w/o context	46.54	51.99
Paderborn + BF w/o context + ASR feedback	45.14	47.29

Model combination	RNN-LM	HD	DEV (%)	EVAL (%)
			45.14	47.29
✓			41.67	43.70
✓	✓		**39.94**	**41.64**
✓	✓	✓	40.26	42.00

Simulation results – WER – External data

ASR difficulty on CHiME-5: training data has only 40 h

AM	Training Data	DEV (%)
CNN-TDNN-LSTM	LibriSpeech (960h)	62.09
Baseline TDNN	CHiME-5 (40h)	58.39
CNN-TDNN-RBiLSTM	CHiME-5 (40h)	45.14

Naive use of 960 hours does not improve performance
Possible causes:

- Conversational/spontaneous speech
- Enhanced speech not close enough to clean speech

Simulation results – WER – External data

3-gram LM Training Data	# of Words	DEV	
		PPL	WER (%)
CHiME-5 (Baseline)	0.4M	155	45.14
CHiME-5 + AMI	1.2M	140	45.10
CHiME-5 + LibriSpeech	9.8M	134	44.49
CHiME-5 + AMI + LibriSpeech	10.6M	131	44.21

Larger gain expected.

Conclusion

+ GSS front-end can boost second best challenge system (Hitachi-JHU)
+ CHiME-5: New best WER (DEV 39.94 % and EVAL 41.64 %)
+ Annotation fine-tuning with ASR
+ Dropping context for beamforming
- ◦ Taking more data for AM and LM needs more investigations

Thank you for listening!

Conclusion

+ GSS front-end can boost second best challenge system (Hitachi-JHU)
+ CHiME-5: New best WER (DEV 39.94 % and EVAL 41.64 %)
+ Annotation fine-tuning with ASR
+ Dropping context for beamforming
∘ Taking more data for AM and LM needs more investigations

Thank you for listening!

CHiME 5 result table

Track	Session		Kitchen	Dining	Living	Overall
Single	Dev	S02	62.33	52.82	44.62	52.07
		S09	51.87	54.02	48.09	
	Eval	S01	60.07	40.88	60.94	47.31
		S21	49.09	38.14	42.67	
Multiple	Dev	S02	46.66	45.07	36.19	39.94
		S09	36.40	39.43	35.33	
	Eval	S01	53.93	35.66	49.78	41.64
		S21	46.43	34.53	36.64	

Single array to multi array

Arrays	Context in BF	
	On	Off
1	58.05	58.13
3	52.30	48.81
6	49.21	46.54