The Hitachi/JHU CHiME-5 system: Advances in speech recognition for everyday home environments using multiple microphone arrays

Naoyuki Kanda[1], Rintaro Ikeshita[1], Shota Horiguchi[1], Yusuke Fujita[1], Kenji Nagamatsu[1], Xiaofei Wang[2], Vimal Manohar[2], Nelson Enrique Yalta Soplin[2], Matthew Maciejewski[2], Szu-Jui Chen[2], Aswin Shanmugam Subramanian[2], Ruizhi Li[2], Zhiqi Wang[2], Jason Naradowsky[2], L. Paola Garcia-Perera[2], Gregory Sell[2]

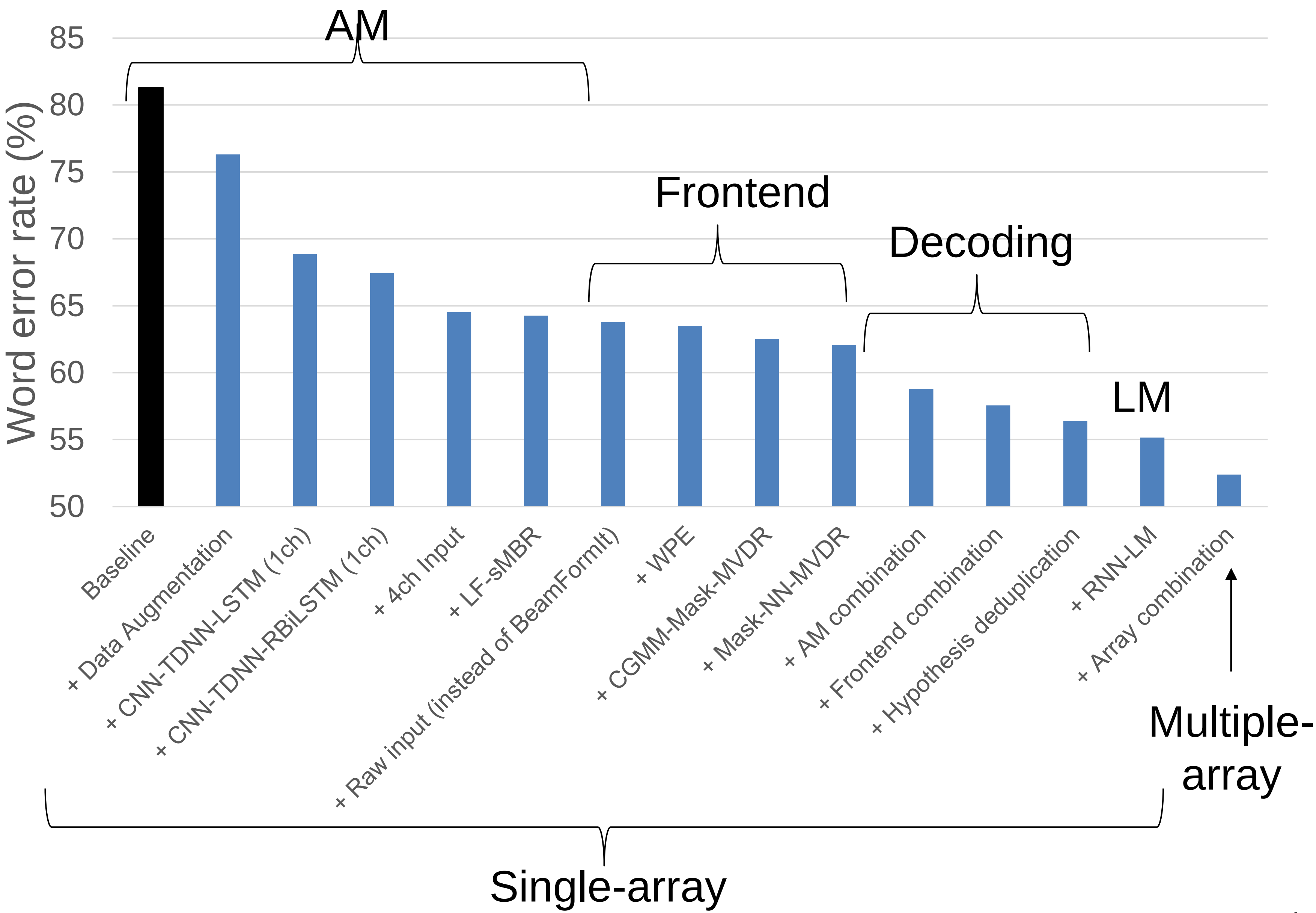
1 HITACHI Inspire the Next

2 JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

Step-by-Step Improvements for Dev

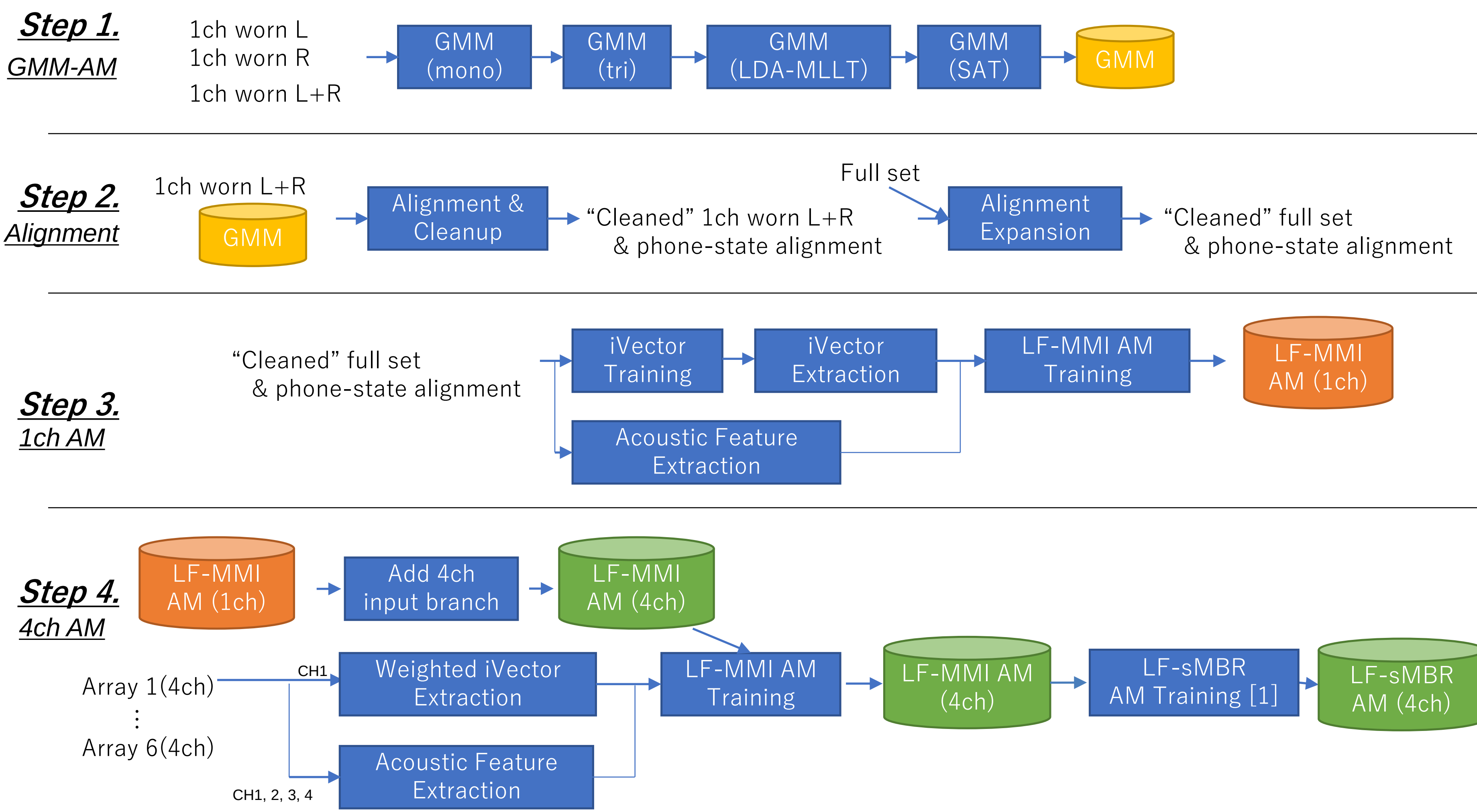
HITACHI Inspire the Next

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING



Acoustic Model Training Pipeline

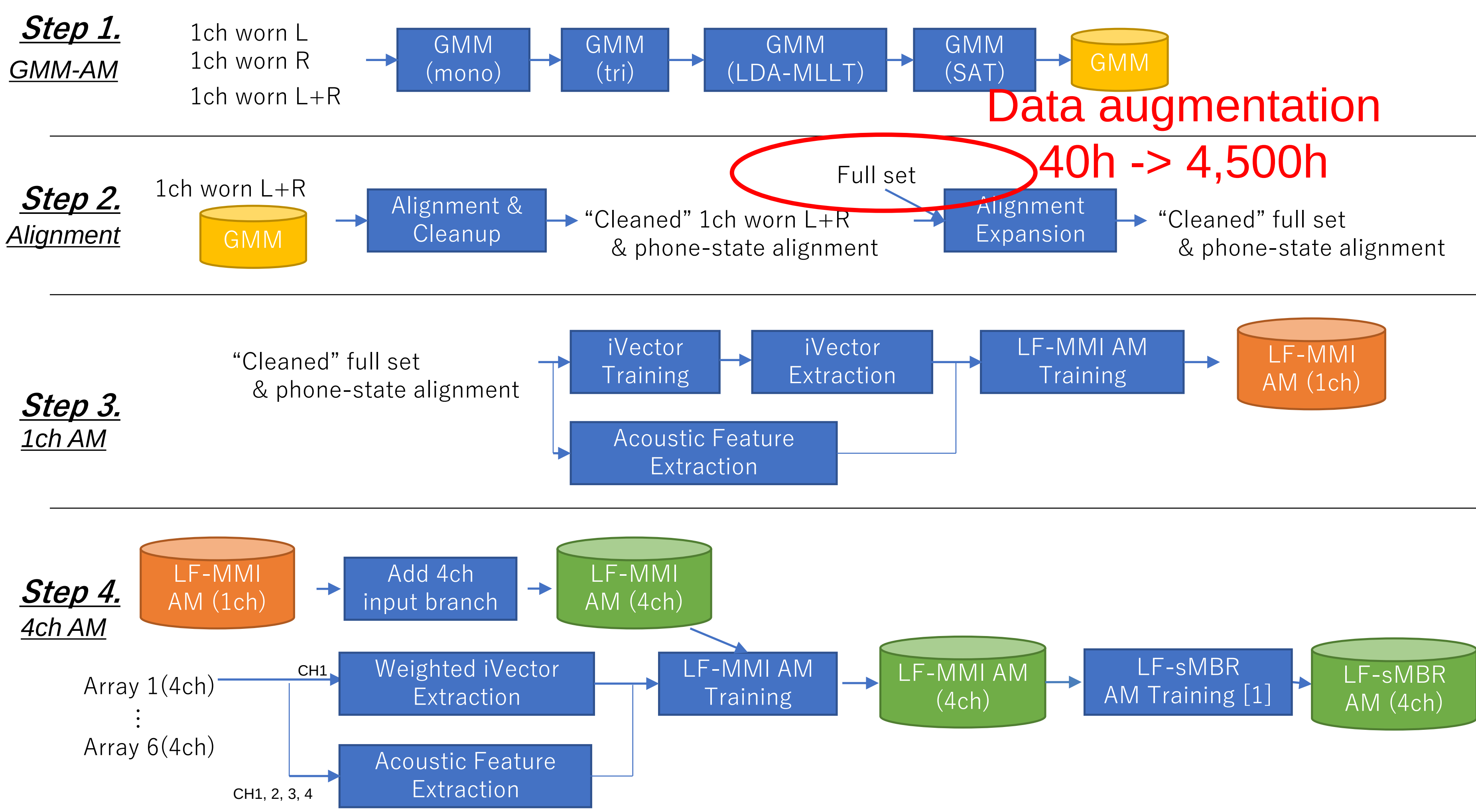
HITACHI Inspire the Next | JOHNS HOPKINS WHITING SCHOOL of ENGINEERING



[1] Naoyuki Kanda, Yusuke Fujita, Kenji Nagamatsu, Lattice-free state-level minimum Bayes risk training of acoustic models, Interspeech 2018.

Acoustic Model Training Pipeline

HITACHI Inspire the Next | JOHNS HOPKINS WHITING SCHOOL of ENGINEERING



[1] Naoyuki Kanda, Yusuke Fujita, Kenji Nagamatsu, Lattice-free state-level minimum Bayes risk training of acoustic models, Interspeech 2018.

Training Data for 1ch AM

Effect of data augmentation with baseline AM

Data			Data Augmentation			Training Epoch	Worn-Dev	Ref-Array-Dev
Worn	Array (Raw, CH1)	Array (BeamFormIt)	Speed & Volume	Reverb. & Noise(*)	Bandpass			
L, R, L+R	1		✓			4	44.05	79.65
L, R, L+R	1	1	✓			4	44.49	78.72
L, R, L+R	1 … 6	1 … 6	✓			4	48.92	78.51
L, R, L+R	1 …. 6	1 … 6	✓	✓		2	45.82	77.26
L, R, L+R	1 … 6	1 … 6	✓	✓	✓	1	45.37	76.31

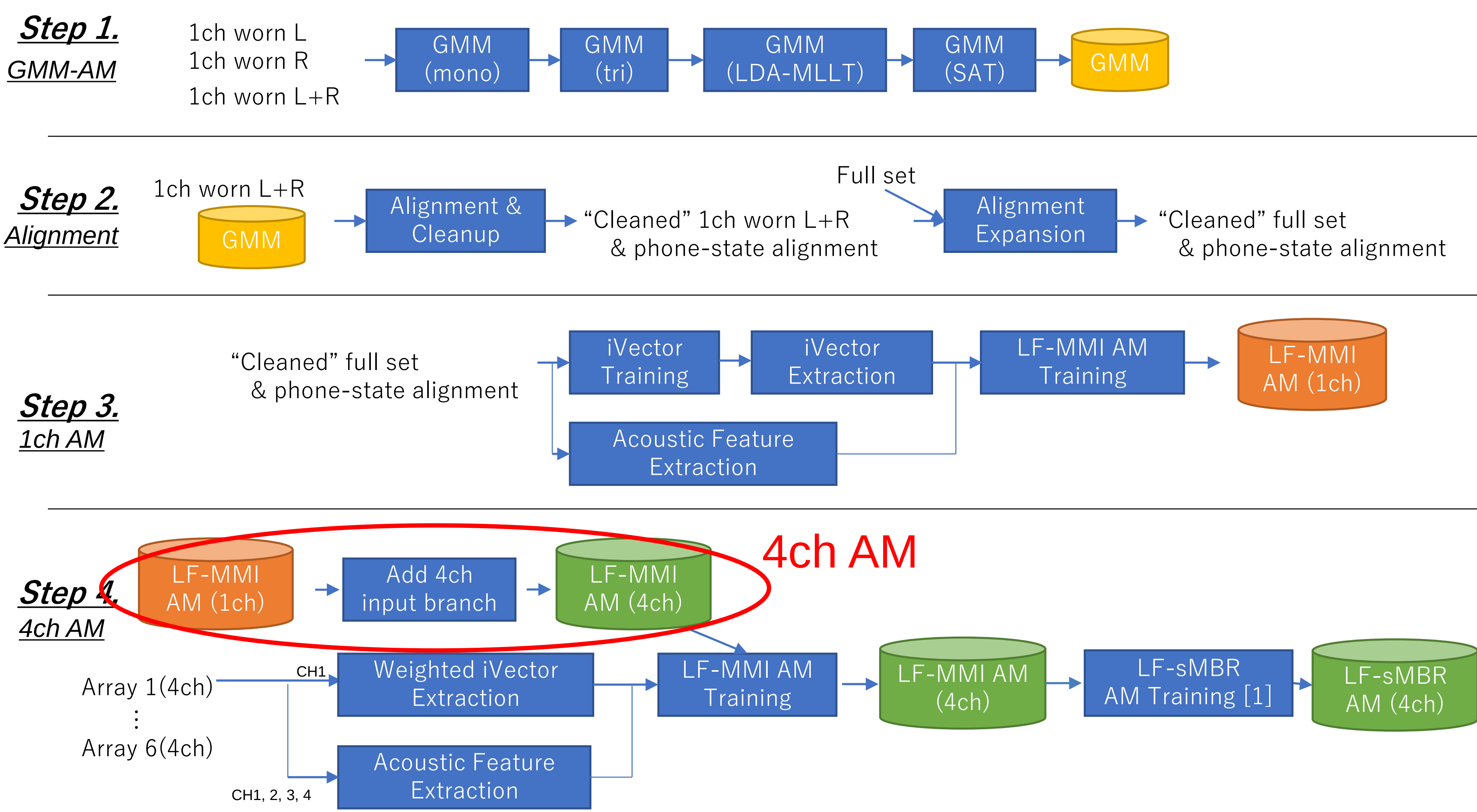
(*) Reverb. & noise perturbation was applied only for worn microphone data.

Speed: 0.9, 1.0, 1.1
Volume: 0.125 – 2.0
Reverberation: Generate impulse responses of simulated rooms by image method.
Follow the settings of {small, medium}-size rooms in [1].
Noise: Add non-speech region of array data with SNR of {20,15,10,5, 0}
Bandpass: Randomly-selected frequency band was cut off.
(leave at least 1,000 Hz band within the range of less than 4,000 Hz)

[1] T. Ko, et al.: A study on data augmentation of reverberant speech for robust speech recognition, Proc. ICASSP, pp. 5220—5224, 2017.

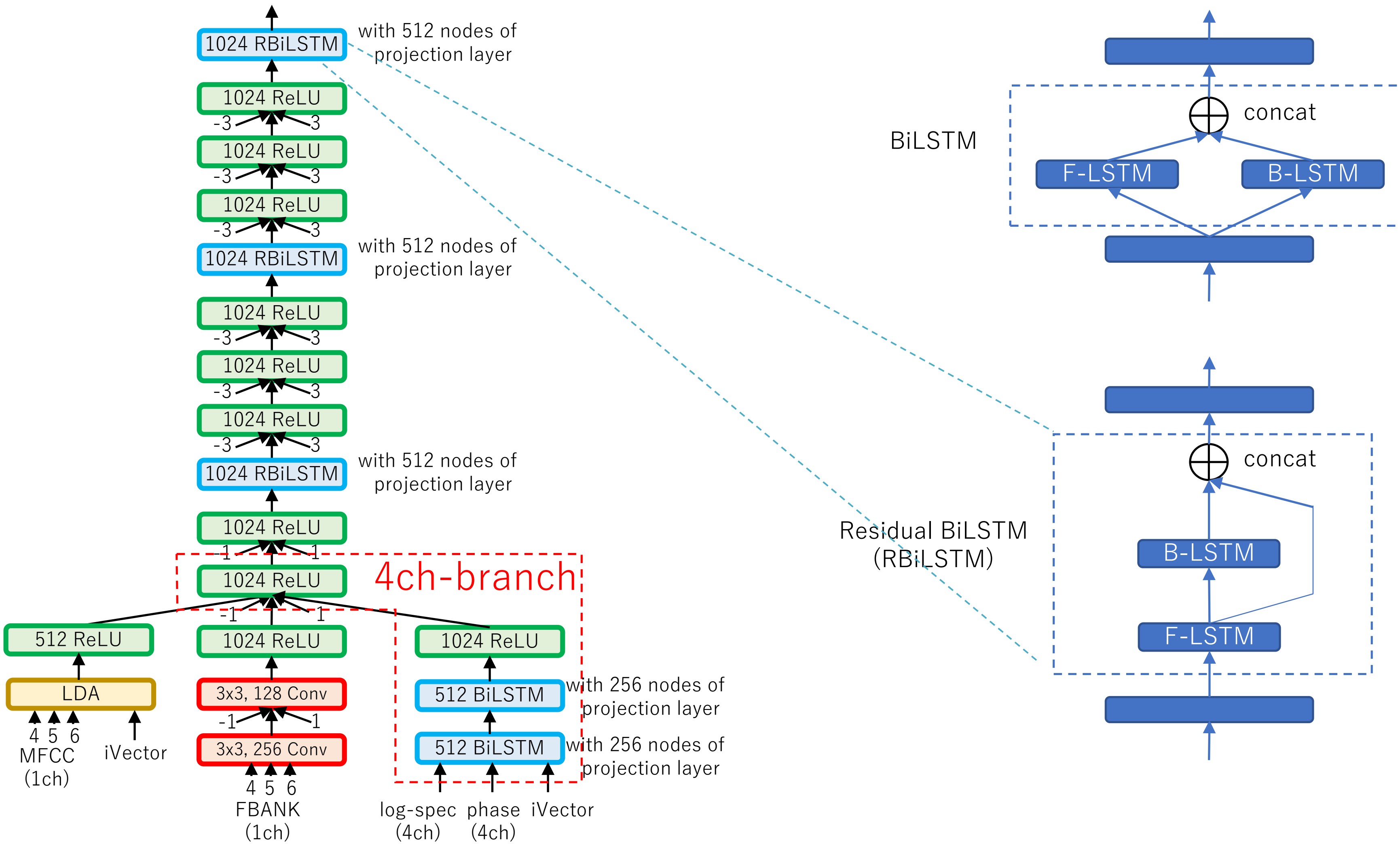
Acoustic Model Training Pipeline

HITACHI Inspire the Next | JOHNS HOPKINS WHITING SCHOOL of ENGINEERING



[1] Naoyuki Kanda, Yusuke Fujita, Kenji Nagamatsu, Lattice-free state-level minimum Bayes risk training of acoustic models, Interspeech 2018.

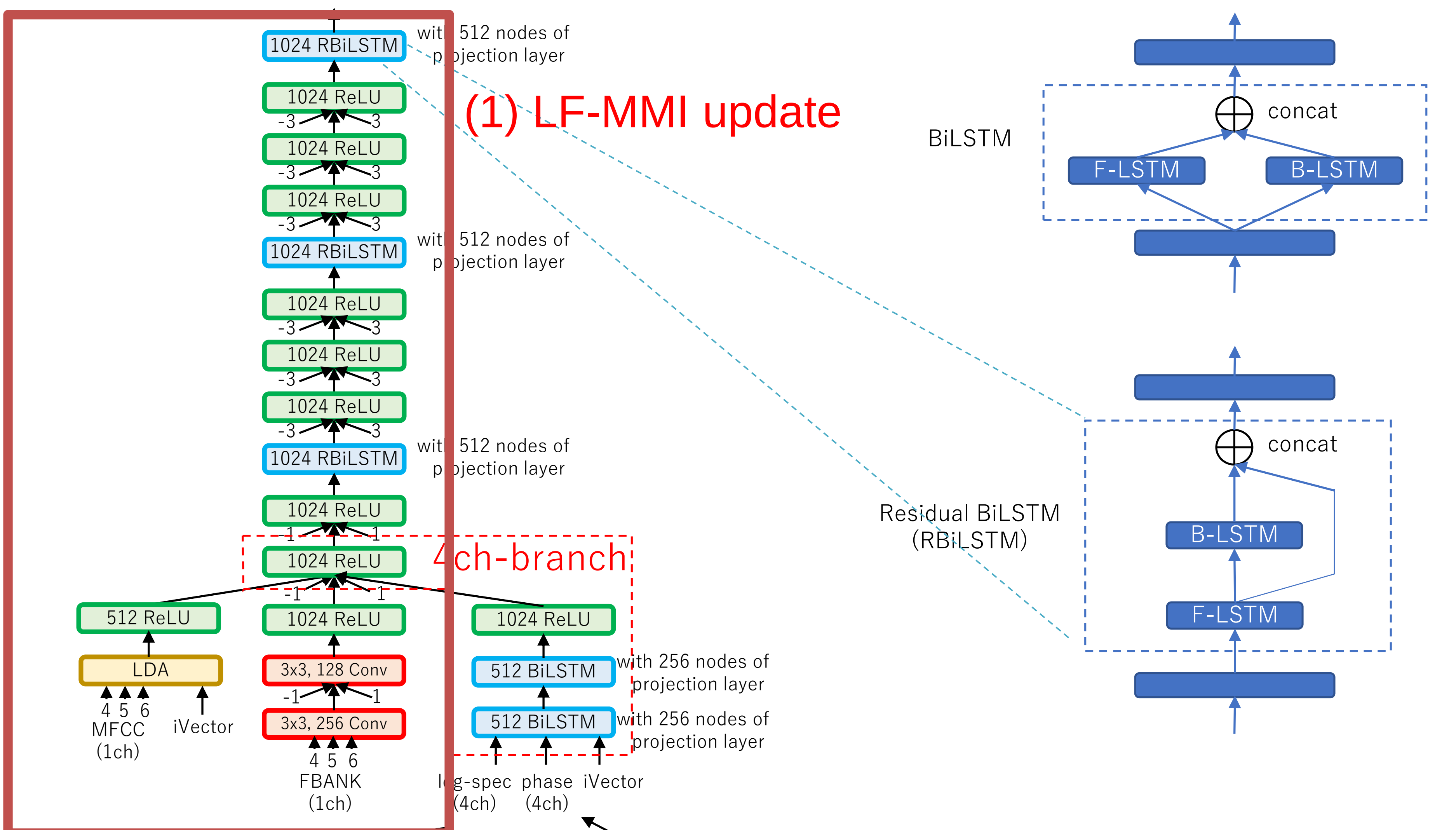
4ch CNN-TDNN-RBiLSTM



$$\log |x_{i,f,t}| \quad i\ (= 1, 2, 3, 4)$$

$$\cos(\angle(x_{i,f,t}) - \angle(x_{1,f,t})) \qquad (i = 2, 3, 4),$$
$$\sin(\angle(x_{i,f,t}) - \angle(x_{1,f,t})) \qquad (i = 2, 3, 4).$$

4ch CNN-TDNN-RBiLSTM



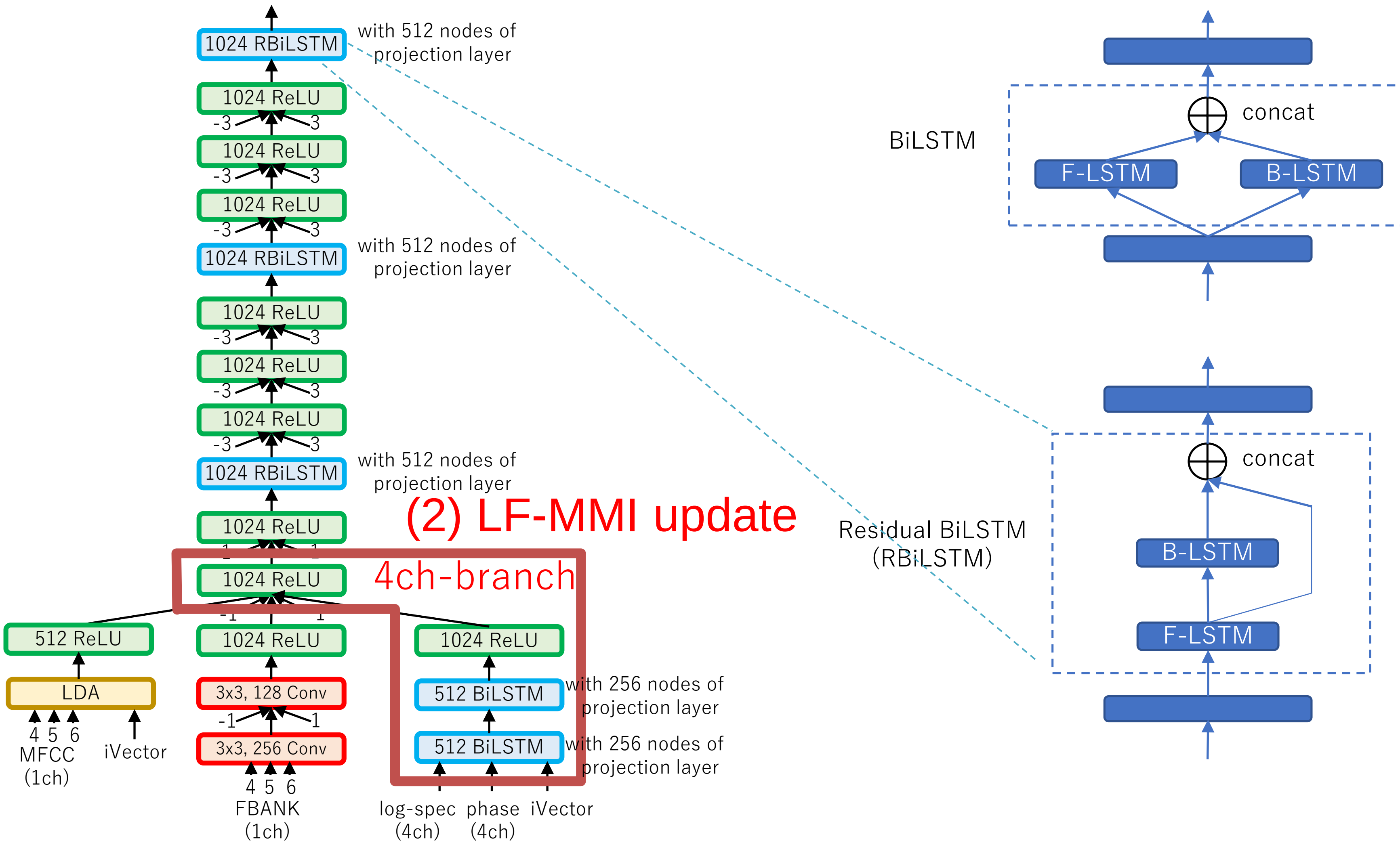
$$\log |x_{i,f,t}| \quad i\ (= 1,2,3,4)$$

$$\cos(\angle(x_{i,f,t}) - \angle(x_{1,f,t})) \qquad (i = 2,3,4),$$
$$\sin(\angle(x_{i,f,t}) - \angle(x_{1,f,t})) \qquad (i = 2,3,4).$$

4ch CNN-TDNN-RBiLSTM

HITACHI Inspire the Next

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING



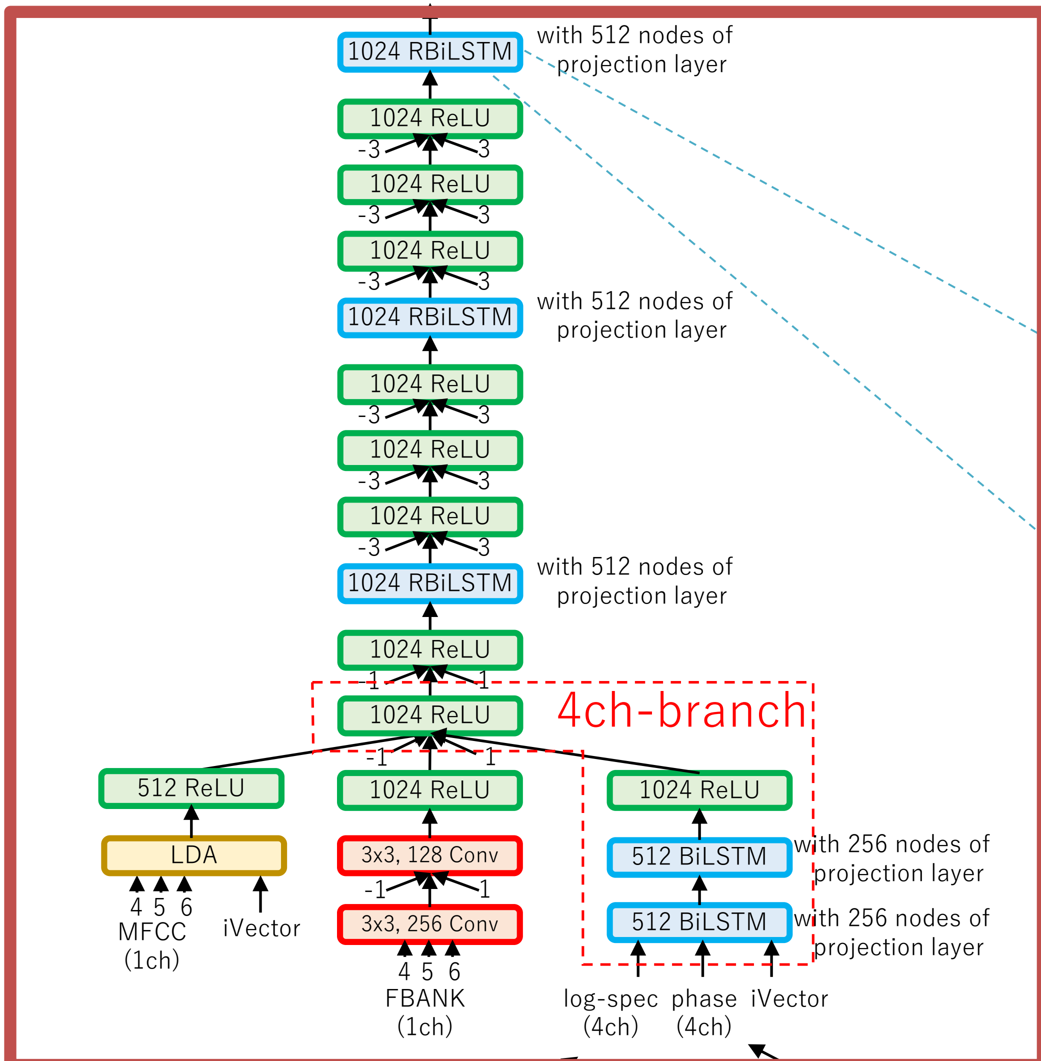
$$\log |x_{i,f,t}| \qquad i \ (= 1, 2, 3, 4)$$

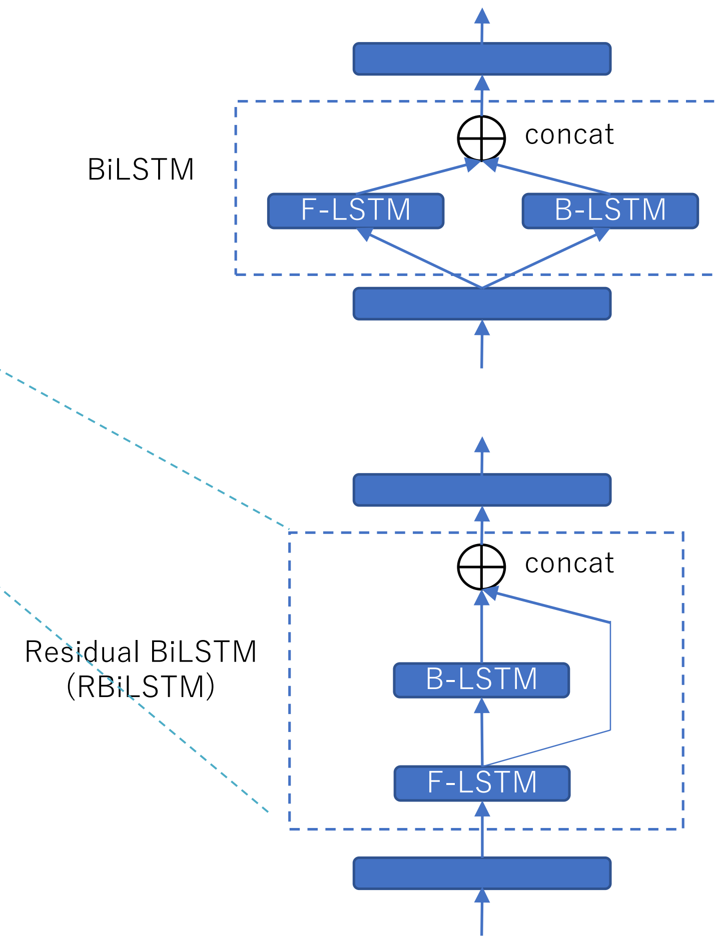
$$\cos(\angle(x_{i,f,t}) - \angle(x_{1,f,t})) \qquad (i = 2, 3, 4),$$
$$\sin(\angle(x_{i,f,t}) - \angle(x_{1,f,t})) \qquad (i = 2, 3, 4).$$

4ch CNN-TDNN-RBiLSTM

HITACHI Inspire the Next

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING





(3) LF-sMBR update

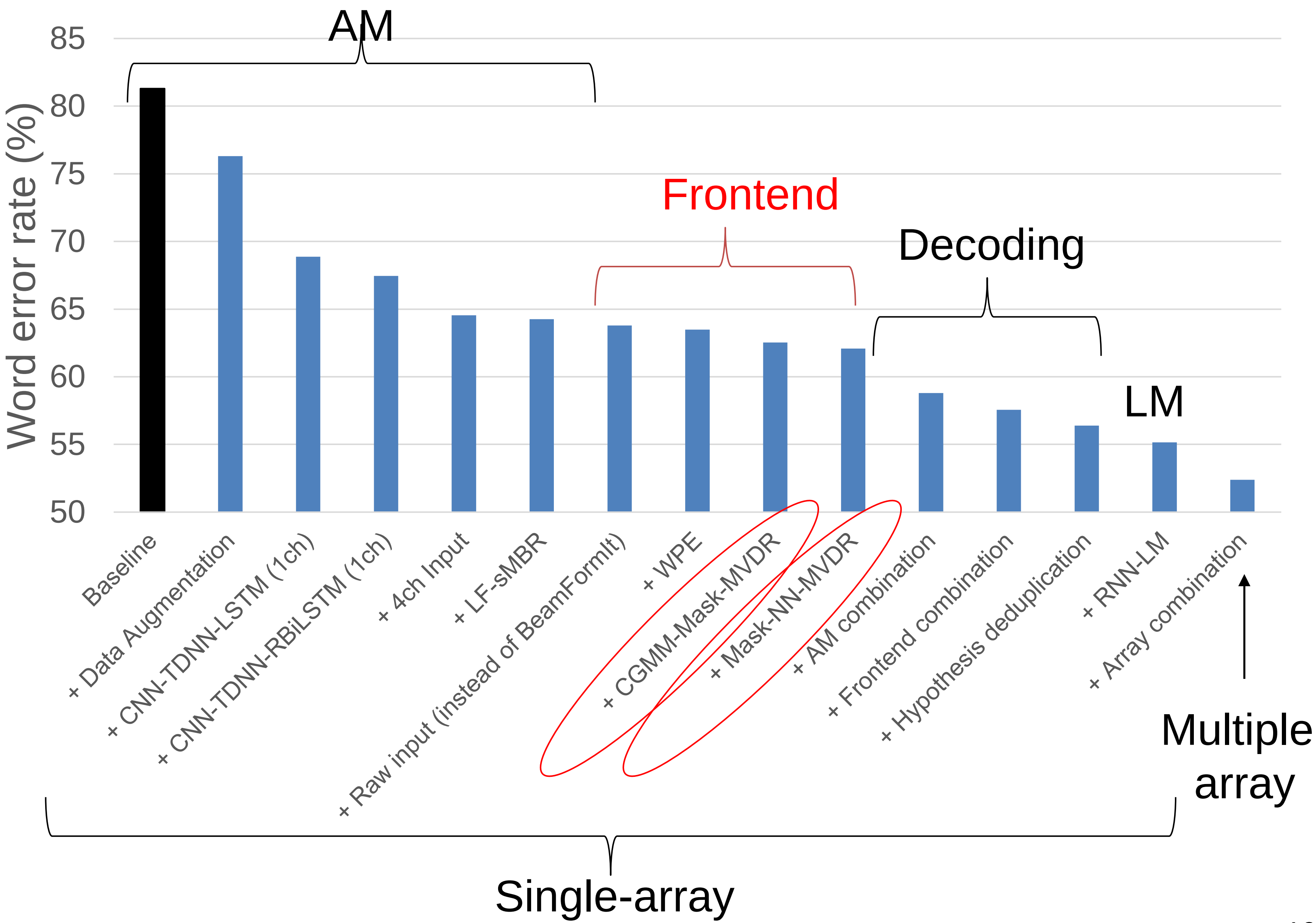
$$\log |x_{i,f,t}| \quad i\ (= 1, 2, 3, 4)$$

$$\cos(\angle(x_{i,f,t}) - \angle(x_{1,f,t})) \qquad (i = 2, 3, 4),$$
$$\sin(\angle(x_{i,f,t}) - \angle(x_{1,f,t})) \qquad (i = 2, 3, 4).$$

Step-by-Step Improvements for Dev

HITACHI Inspire the Next

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

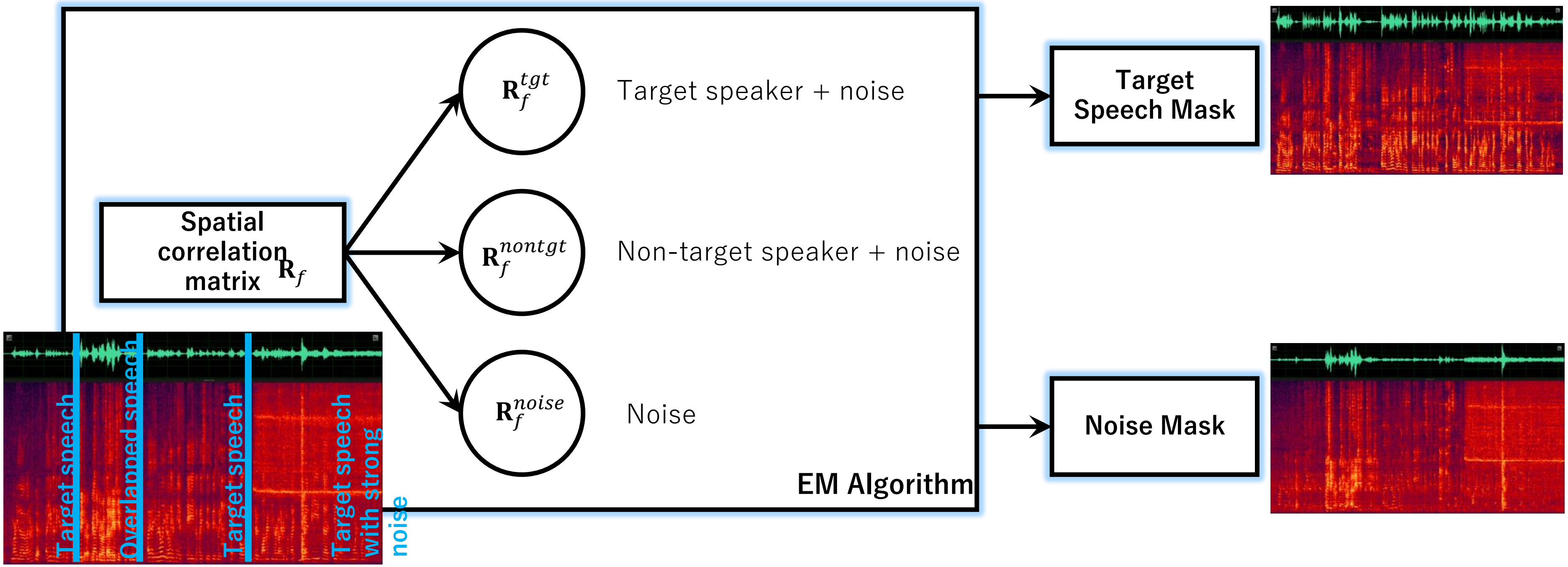


Complex Gaussian Mixture Model

- 3-class mixture: target, non-target, and noise

$$y_f(t) \sim \alpha_{tgt}\mathcal{N}_{\mathbb{C}}\left(0, v_f^{tgt}(t)R_f^{tgt}\right) + \alpha_{nontgt}\mathcal{N}_{\mathbb{C}}\left(0, v_f^{nontgt}(t)R_f^{nontgt}\right) + \alpha_{noise}\mathcal{N}_{\mathbb{C}}\left(0, v_f^{noise}(t)R_f^{noise}\right)$$

- Mask estimation using EM Algorithm ➡ MVDR-based Beamformer



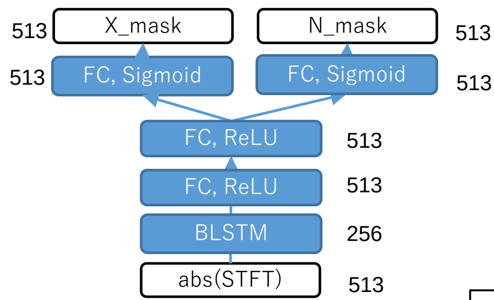
Higuchi, Takuya, et al. "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR. " *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.4 (2017): 780-793.

Mask Estimation Neural Network

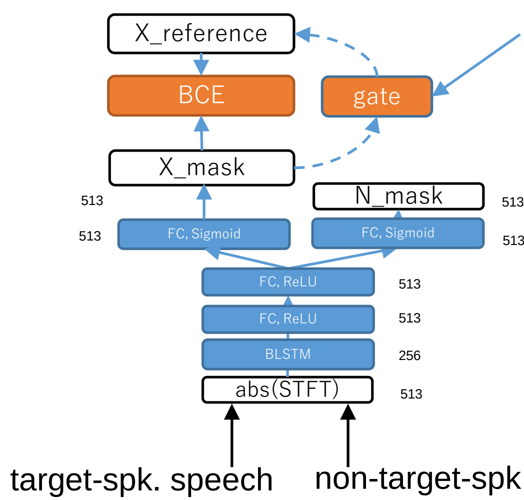
HITACHI Inspire the Next — JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

1. Train mask estimation (ME) network [1][2]

by using mixture of speech (worn non-speaker-overlapped region) and noise (array non-speech region) in the training set



2. Speaker adaptation




```
def gate(x):
    if input.speaker == target_speaker:
        return x
    else:
        return 0
```

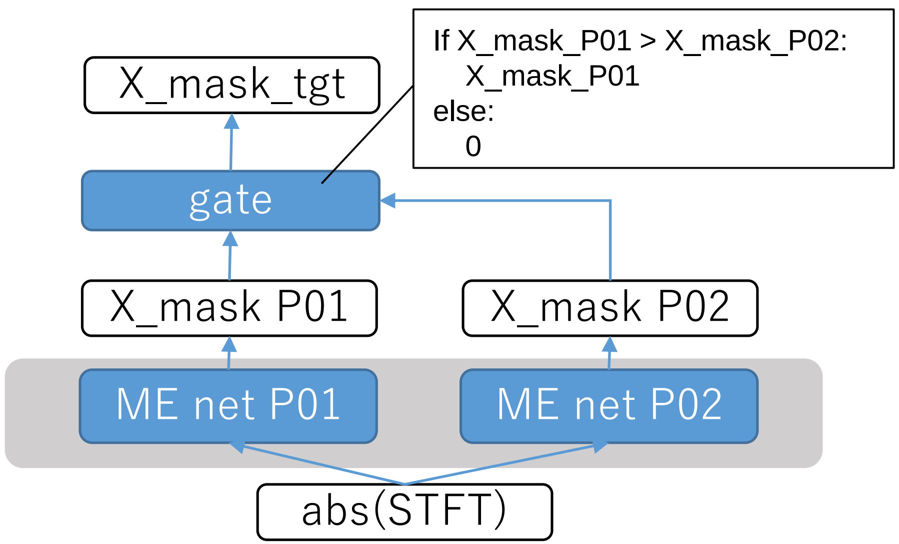
(*) we used only non-overlapped regions for adaptation

[1] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in Proc. ICASSP, 2016, pp. 196-200.
[2] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks." in Proc. Interspeech, 2016, pp. 1981-1985.

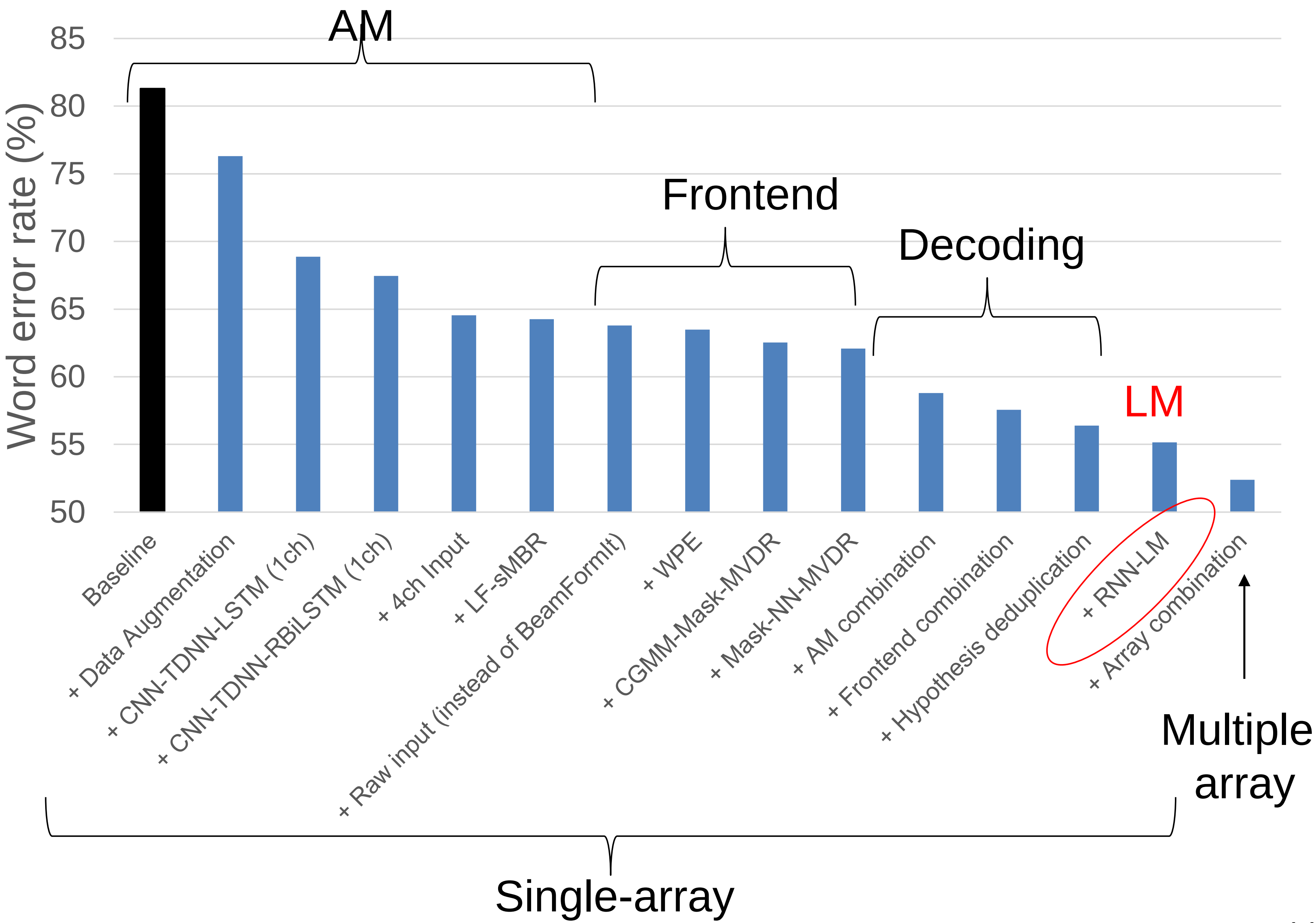
3. Mask inference

Target speaker’s mask is selected only if target speaker’s output value is higher than all other non-targets values.

Example: P01(target) and P02(non-target)



Step-by-Step Improvements for Dev



Language Modeling

- Recurrent neural network based word-LM
  - 2 layer LSTM with 512 nodes, 50% dropout
  - 512 dim embeddings
  - PyTorch implementation

- Official-LM: forward-RNN-LM: backward-RNN-LM
  = 0.5 : 0.25 : 0.25

WER (%) for Dev set

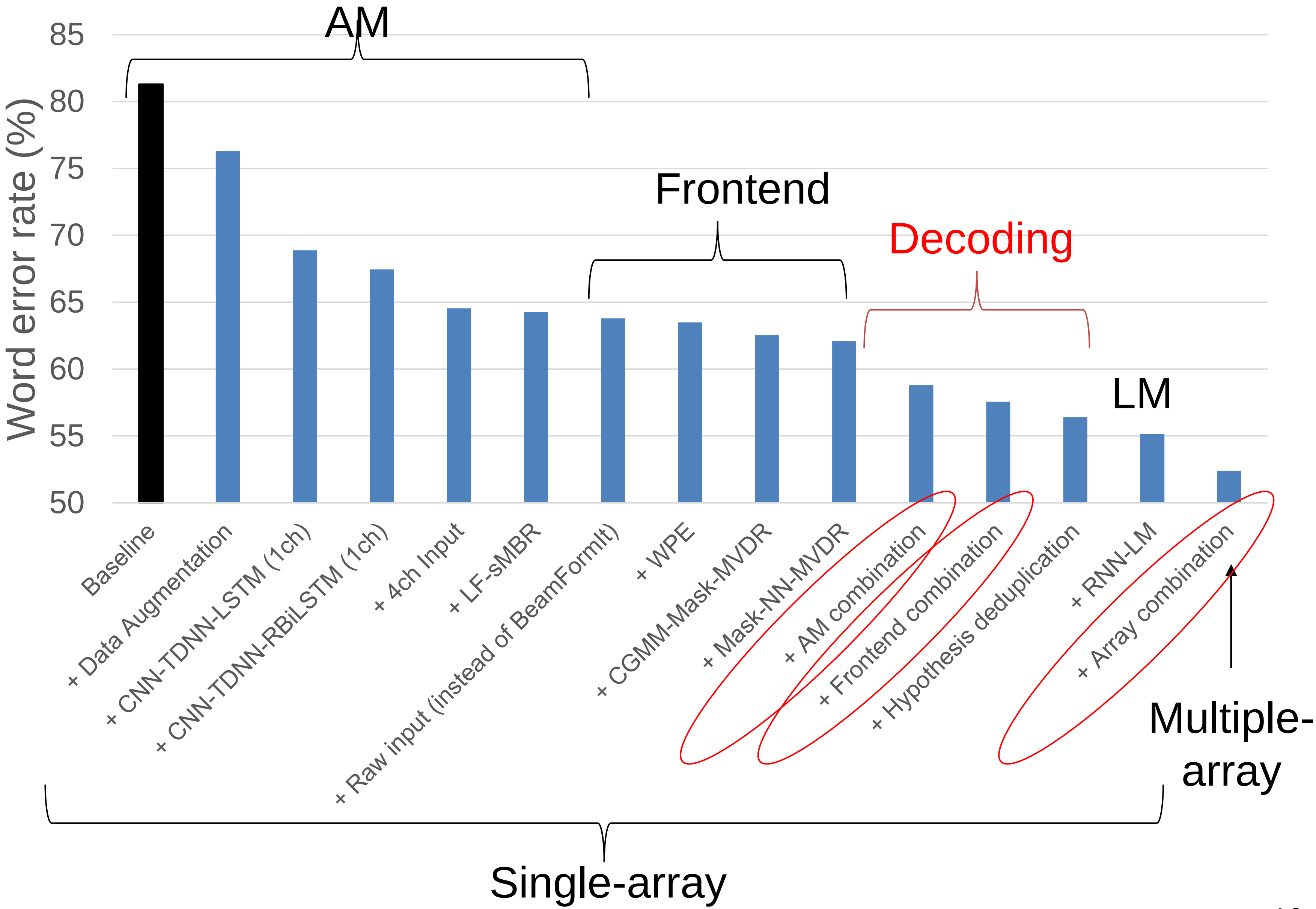
	without RNN-LM		with RNN-LM
Single-array	56.40	1.3% impr.	55.15
Multiple-array	54.00	1.6% impr.	52.38

(*) Results with model combination and hypothesis deduplication

Step-by-Step Improvements for Dev

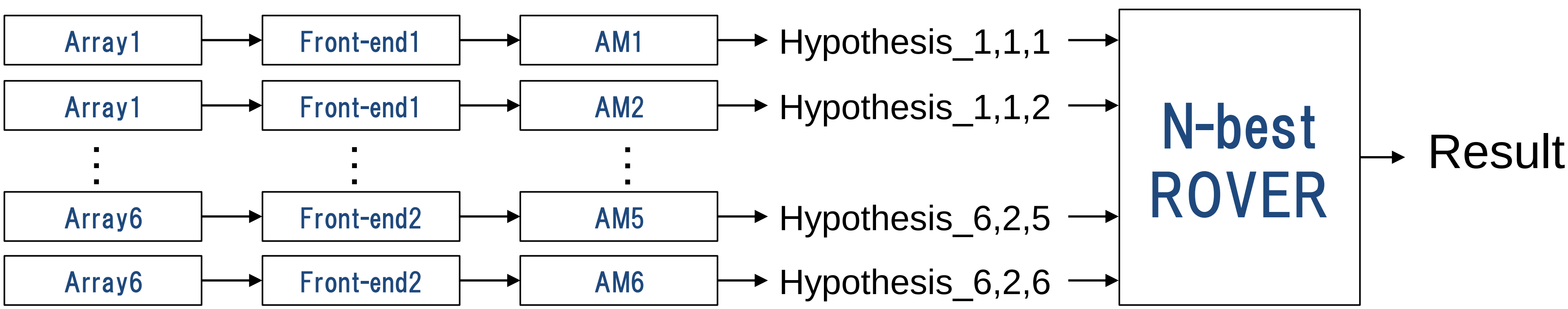
HITACHI Inspire the Next

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING



- Hypotheses combination by N-best ROVER
  - 6 AMs := CNN-TDNN-{LSTM, BiLSTM, RBiLSTM} x {3500, 7000} senones
  - 2 Front-ends := Mask Network, CGMM
  - 6 Arrays

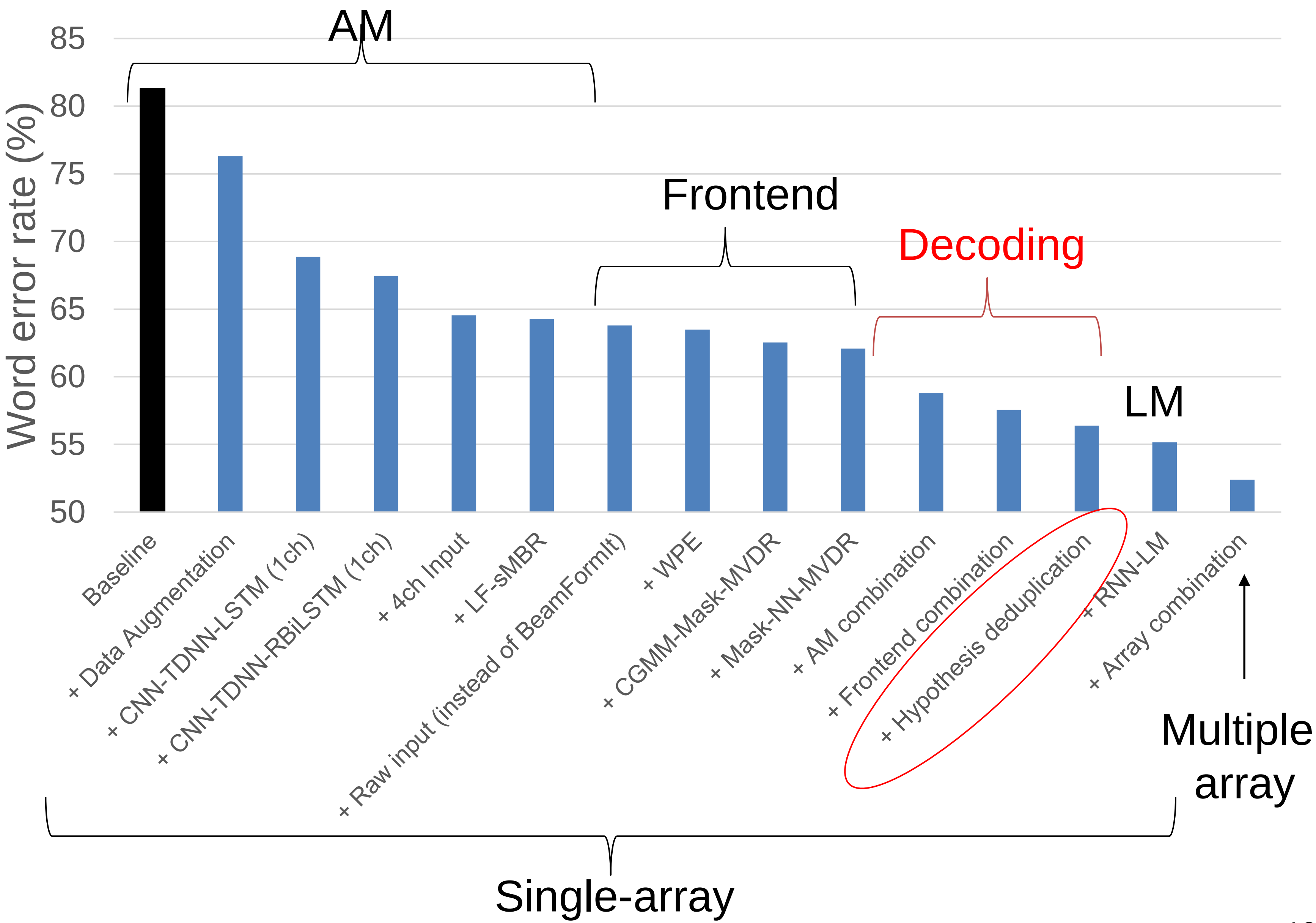
We didn't select array, instead combined hypotheses from each array.



Step-by-Step Improvements for Dev

HITACHI
Inspire the Next

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

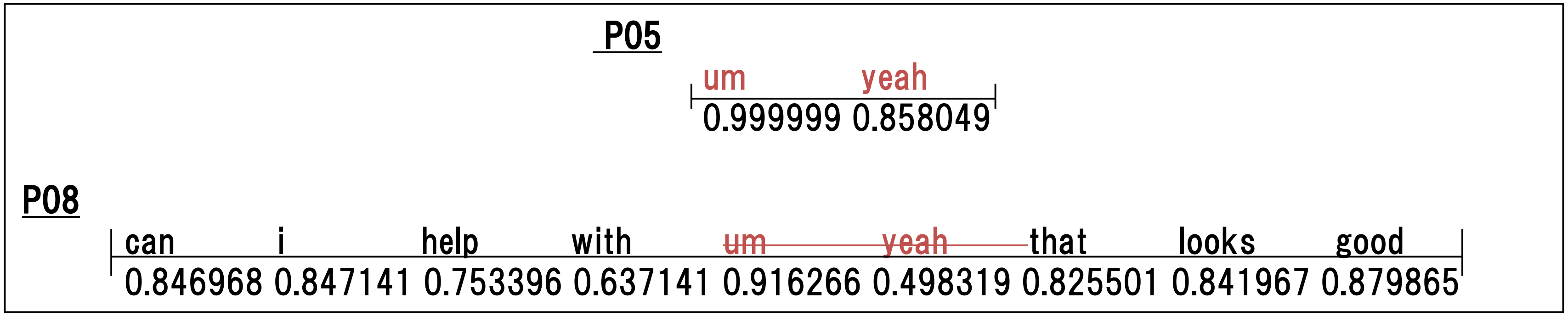


Hypothesis Deduplication (HD)

HITACHI Inspire the Next

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

- Same words were sometimes recognized for overlapped utterances



- Duplicated words with lower confidence were excluded from the hypothesis.
  - HD was applied after ROVER, so precise time boundary could not be used. Minimum edit distance-based word alignment was used to detect word duplication.

WER (%) for Dev set

	without HD	with HD
Single-array	56.44 1.3% impr.	55.15
Multiple-array	53.69 1.3% impr.	52.38

(*) Results with RNN-LM

Final results & Conclusion

Final Results

WER (%)
without RNN-LM / with RNN-LM

Our best eval result

Track	Session		Kitchen	Dining	Living	Overall
Single-array	Dev	S02 S09	66.37 / 65.13 55.89 / 55.24	56.79 / 55.42 55.94 / 54.37	50.89 / 49.54 51.57 / 50.15	**56.40 / 55.15**
	Eval	S01 S21	59.42 / 57.62 52.11 / 49.68	44.18 / 41.81 42.14 / 39.78	63.85 / 62.33 46.71 / 44.59	**50.36 / 48.20**
Multiple-array	Dev	S02 S09	61.05 / 59.31 51.87 / 50.64	54.56 / 52.96 52.46 / 50.69	50.47 / 48.95 52.48 / 50.46	**54.00 / 52.38**
	Eval	S01 S21	59.82 / 57.01 54.70 / 51.59	43.59 / 41.22 44.12 / 42.17	62.28 / 60.67 45.95 / 43.82	**50.59 / 48.24**

Final Results

HITACHI Inspire the Next

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

WER (%)

without RNN-LM / with RNN-LM

Track	Session		Kitchen	Dining	Living	Overall
Single-array	Dev	S02	66.37 / 65.13	56.79 / 55.42	50.89 / 49.54	**56.40 / 55.15**
		S09	55.89 / 55.24	55.94 / 54.37	51.57 / 50.15	
	Eval	S01	59.42 / 57.62	44.18 / 41.81	63.85 / 62.33	**50.36 / 48.20**
		S21	52.11 / 49.68	42.14 / 39.78	46.71 / 44.59	
Multiple-array	Dev	S02	61.05 / 59.31	54.56 / 52.96	50.47 / 48.95	**54.00 / 52.38**
		S09	51.87 / 50.64	52.46 / 50.69	52.48 / 50.46	
	Eval	S01	59.82 / 57.01	43.59 / 41.22	62.28 / 60.67	**50.59 / 48.24**
		S21	54.70 / 51.59	44.12 / 42.17	45.95 / 43.82	

- Array combination by ROVER worked well for dev, but not effective for eval set.
  - Why? Different types of rooms? Speaker-array distance?

- Anyway, better array combination methods should be pursued.

Conclusion

- Our contributions
  - Multiple data augmentation
  - 4-ch AM with Residual BiLSTM
  - Speaker adaptive mask estimation network / CGMM-based beamformer
  - Hypothesis Dedupulication
  - Array combination by ROVER (found not effective for evaluation set)

- Our results
  - 48.2% WER for evaluation set
  - 2nd ranked, with only 2.1 point difference to the best result

Thank you for your attention!

Appendix

Comparison of AM Architectures

Model	Input	Training	Worn-Dev	Ref-Array-Dev
Baseline	1ch	LF-MMI	45.37	76.31
CNN-TDNN-LSTM	1ch	LF-MMI	39.22	68.87
CNN-TDNN-BiLSTM	1ch	LF-MMI	40.04	68.42
CNN-TDNN-RBiLSTM	1ch	LF-MMI	39.21	67.46
CNN-TDNN-RBiLSTM	4ch	LF-MMI	n/a	64.54
CNN-TDNN-RBiLSTM	4ch	LF-sMBR [1]	n/a	64.25

[1] Naoyuki Kanda, Yusuke Fujita, Kenji Nagamatsu, Lattice-free state-level minimum Bayes risk training of acoustic models, Interspeech 2018.

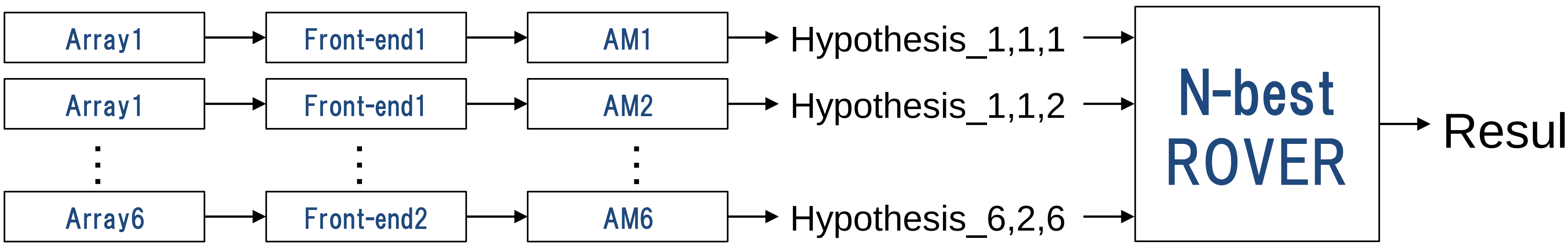
Comparison of Frontend Processing

Front-end for 1ch input	Front-end for 4ch input	Dev
BeamFormIt (= Baseline)	Raw	64.28
Raw	Raw	63.79
WPE	WPE	63.49
CGMM-MVDR	WPE	62.53
Speaker adaptive mask NN-MVDR	WPE	62.09

Hypotheses combination by N-best ROVER

- 6 AMs := CNN-TDNN-{LSTM, BiLSTM, RBiLSTM} x {3500, 7000} senones
- 2 Front-ends := Mask Network, CGMM
- 6 Arrays

We didn't select array. Instead we combined hypotheses from each array.



WER (%) for Dev set

	AM	Array	Frontend	Dev
Single-array	1	1	MaskNet	62.09
Single-array	6	1	MaskNet	58.79
Single-array	6	1	MaskNet, CGMM	57.55
Multiple-array	6	6	MaskNet, CGMM	55.08

(*) Results w/o RNN-LM

Thank you