HITACHI
Inspire the Next

Unsupervised network adaptation and phonetically-oriented system combination for the CHiME-4 Challenge

09/13/2016

Yusuke Fujita, Takeshi Homma, Masahito Togami

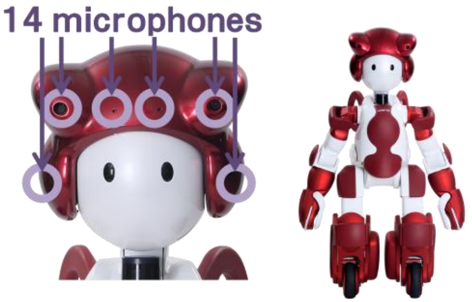
Hitachi Ltd., Research and Development Group, Japan
Hitachi America Ltd.

© Hitachi, Ltd. 2016. All rights reserved.

Background

HITACHI
Inspire the Next

- Hitachi is developing a humanoid robot "EMIEW3" for customer services (ex. airport, station, bank)
  - Distant (1m) ASR
  - Noise robustness in real fields is crucial



- We participated in CHiME-3 challenge
  - Local Gaussian modeling based source separation works well with DNN-based ASR
  - Discriminative system combination outperforms ROVER
  - However, data augmentation, speaker adaptation, and RNNLM examined by top teams have not been applied.
- We followed these state-of-the-art techniques and updated our system for CHiME-4

© Hitachi, Ltd. 2016. All rights reserved.

Contributions

- Local Gaussian modeling based source separation
  - Multi-channel Wiener filter output is utilized for acoustic modeling and frontend speech enhancement
  - Introducing semi-stationarity constraints to non-target sources improves frontend speech enhancement
- Unsupervised deep neural network adaptation
  - Unsupervised re-training of DNN works well for speaker adaptation when using conservative training parameters
- Phonetically-oriented system combination
  - Multiple 1-best sentences are combined considering phonetic similarity improves the system combination performance

© Hitachi, Ltd. 2016. All rights reserved.

Local Gaussian modeling based source separation

HITACHI Inspire the Next

- Multi-channel signal in time-frequency domain

$$x(f,t) = [x_1(f,t), \cdots, x_M(f,t)]^\top \in \mathbb{C}^M$$

f: frequency, t: time (frame), M: # microphones

- Local Gaussian modeling (LGM) [Duong *et al.* 2010]

$$x(f,t) = \sum_{n=1}^{N} c_n(f,t)$$

  – Spatial image of each source

$$c_n(f,t) \sim \mathcal{N}_{\mathbb{C}}(0, \underbrace{v_n(f,t)}_{\text{time-variant activity}} \underbrace{V_n(f)}_{\text{time-invariant spatial correlation matrix}})$$

time-variant **activity**

time-invariant **spatial correlation matrix**

© Hitachi, Ltd. 2016. All rights reserved.

Local Gaussian modeling based source separation

- Multi-channel Wiener filter (MCWF)

$$c_n(f,t) = v_n(f,t)V_n(f)R_x^{-1}(f,t)x(f,t)$$

$R_x(f,t)$: sum of covariance matrix of all sources

$v_n(f,t)$ and $V_n(f)$ are estimated by using EM algorithm

- c.f. Beamforming: $$y(f,t) = \sum_{m=0}^{M} W_m(f)x(f,t) \in \mathbb{C}$$
  - Beamforming outputs single-channel signal : MISO
  - MCWF outputs multi-channel signal :MIMO
- How did we utilize multi-channel signal $c_n(f,t)$?
  - 1. Data augmentation
  - 2. Preprocessor of beamforming

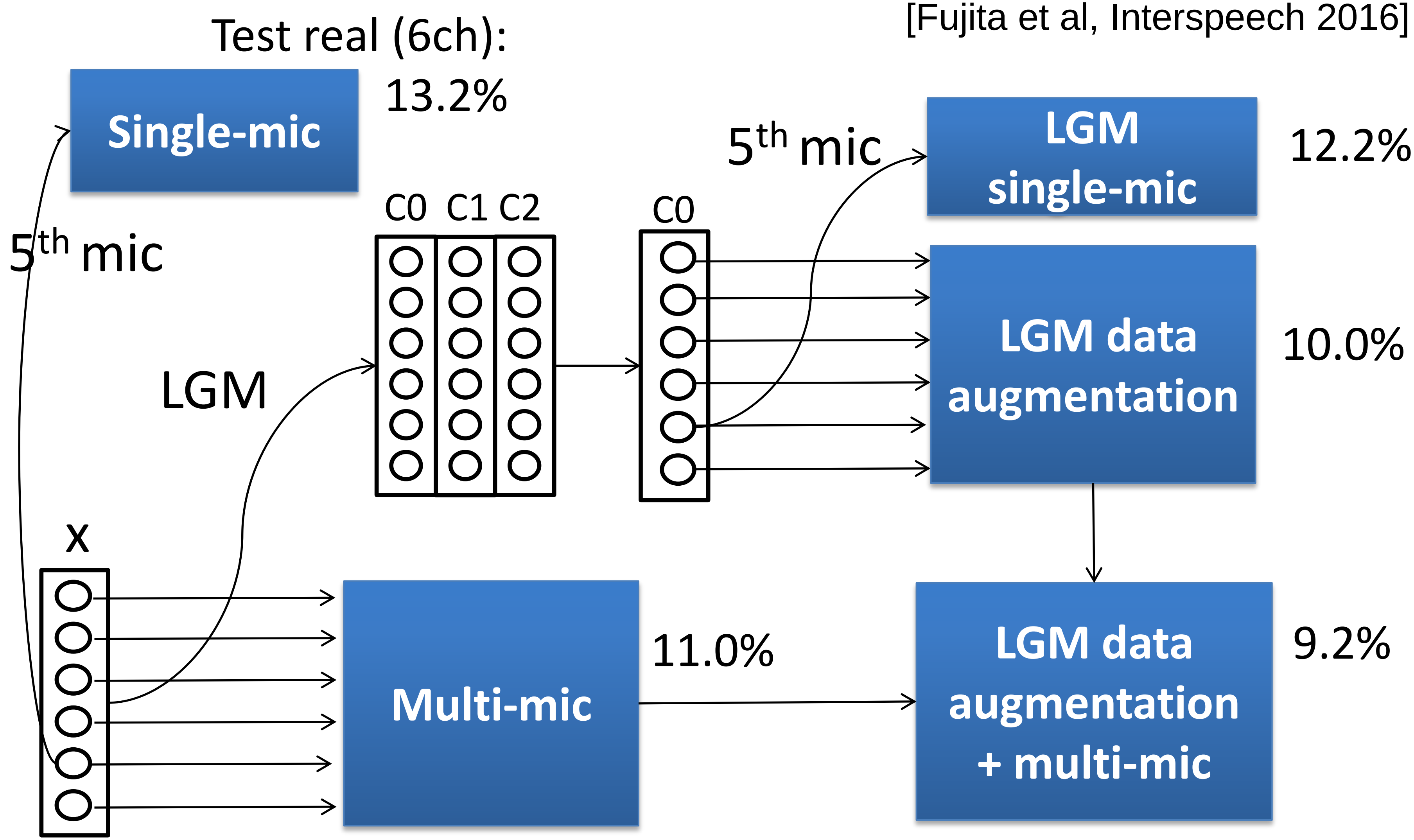
© Hitachi, Ltd. 2016. All rights reserved.

Data augmentation using multi-output separation

HITACHI
Inspire the Next

- All microphone signals from LGM are fed into AM training

[Fujita et al, Interspeech 2016]

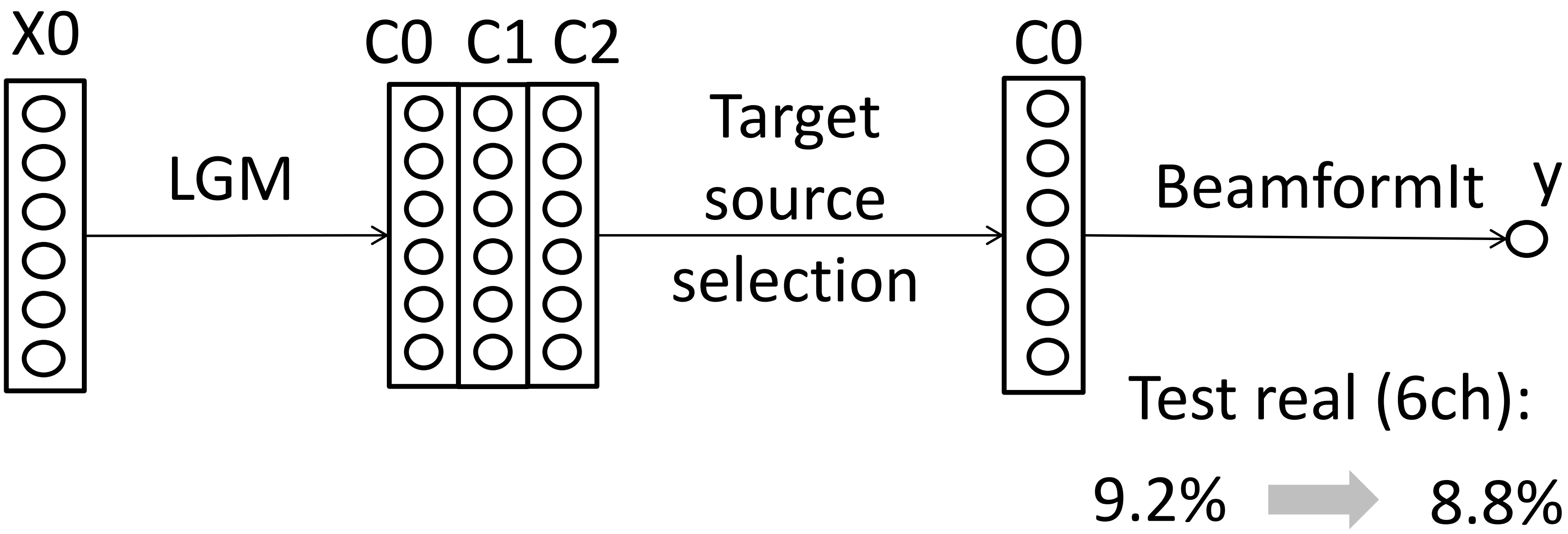


© Hitachi, Ltd. 2016. All rights reserved.

LGM preprocessed beamforming

- $c_n(f,t)$ has 6-ch that holds spatial information
  $\rightarrow$ beamforming technique can be applied
- We used cascading of LGM and BeamformIt

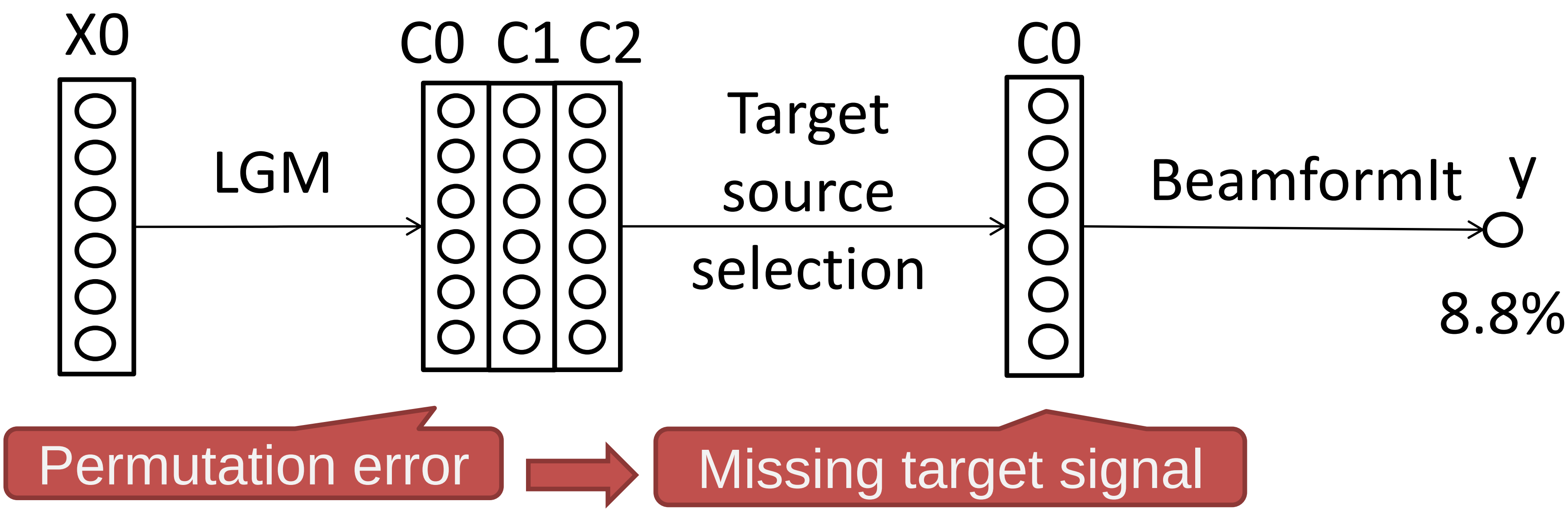
[Fujita et al, Interspeech 2016]



© Hitachi, Ltd. 2016. All rights reserved.

LGM preprocessed beamforming

- Update on target source selection
  - Previous system: target source is selected using SRP-PHAT score on the front direction.
  - Sometimes it failed due to permutation errors and how to hold a tablet device.



  - Introducing permutation-free modification to LGM

© Hitachi, Ltd. 2016. All rights reserved.

Permutation-free Local Gaussian modeling

- Introducing semi-stationary constraints to noise sources
  - Target source $c_0(f,t) \sim \mathcal{N}_{\mathbb{C}}(0, v_0(f,t)V_0(f))$
  - Non-target sources

$$c_n(f,t) \sim \mathcal{N}_{\mathbb{C}}(0, \hat{v}_n(f,t)V_n(f)) \ (n \geq 1)$$

  - Moving average filter is applied to 'activity'

$$\hat{v}_n(f,t) = \sum_{\tau=0}^{T_n} v_n(v, t-\tau)/T_n \qquad T_1 = 3,\ T_2 = 6$$

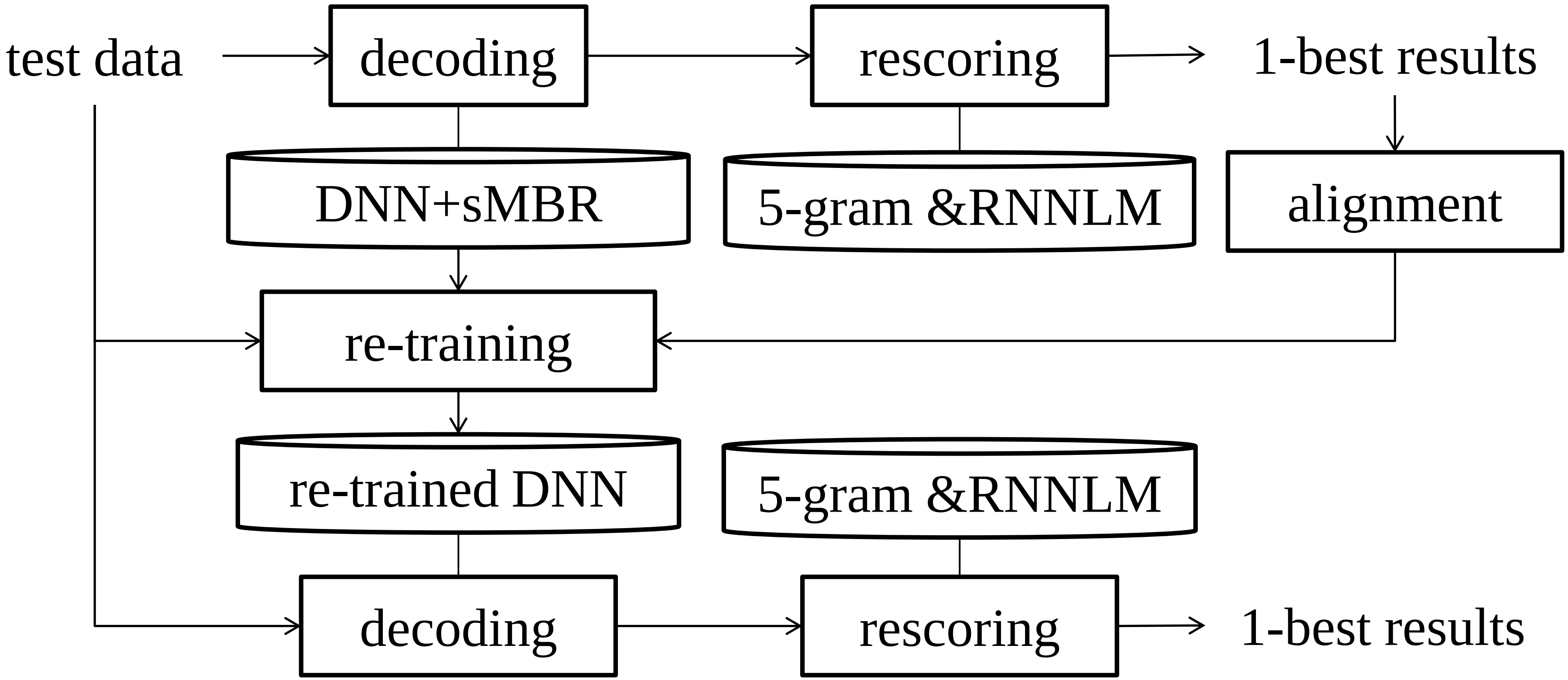
  - Applying the moving average filter in the each EM iteration, target source, i.e. the most active source is extracted onto c0.
  - We no longer select the target source using SRP-PHAT

Test real(6ch): 8.8% → 7.8%

© Hitachi, Ltd. 2016. All rights reserved.

Unsupervised network adaptation

- DNN is self-adapted using 1-best results
- Re-training is performed by using mini-batch SGD with cross entropy criteria [Yoshioka et al, ASRU 2015]



© Hitachi, Ltd. 2016. All rights reserved.

Concerns about re-training of DNN

HITACHI
Inspire the Next

- Unsupervised adaptation fails when the large number of DNN parameters are adapted [Liao, ICASSP 2012]
  - 32M parameters in our case is medium size. We did not try any parameter reduction technique such *as low-rank approximation* nor *partial layer adaptation*
  - Adaptation of entire network works successfully

- Hyper-parameters used in initial training phase is not appropriate for adaptation. We tuned three hyper parameters: learning rate, mini-batch size, and the number of iterations.
  - L2 penalty (weight decay) may be a good option. But we didn’t try it due to time consideration

© Hitachi, Ltd. 2016. All rights reserved.

Unsupervised network adaptation

HITACHI
Inspire the Next

- Small learning rate, early stopping
- Large mini-batch ?

WERs on 6ch track

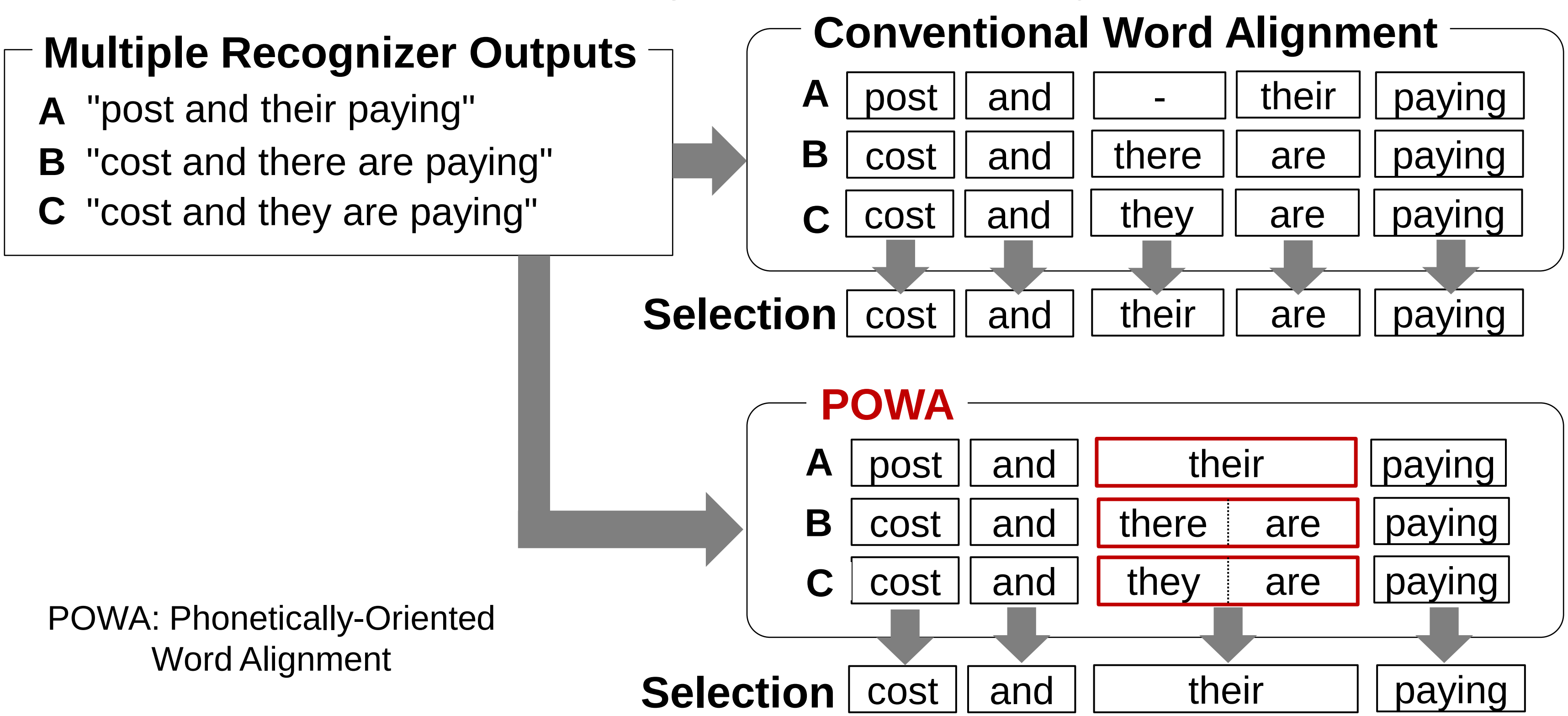
iter	Learn rate	mini-batch	dev avg	dev real	dev simu	test real	test simu
No adaptation			4.85	4.49	5.20	7.78	5.20
1	0.01	256	4.115	3.93	4.3	6.48	5.05
1	0.008	512	4.08	3.95	4.21	6.52	4.97
1	0.001	256	3.7	3.58	3.82	5.56	4.42
1	0.0004	256	3.745	3.6	3.89	5.66	4.47
1	0.0004	512	3.735	3.6	3.87	5.67	4.45
1	0.0004	12000	3.7	3.55	3.85	5.68	4.49
1	0.0001	256	3.865	3.66	4.07	6.23	4.97
2	0.0004	12000	**3.695**	3.58	3.81	**5.56**	4.47
10	0.0004	256	4.305	4.1	4.51	6.9	5.4

© Hitachi, Ltd. 2016. All rights reserved.

Phonetically-oriented System Combination

HITACHI
Inspire the Next

- Combination of 1-best results from various systems
- Word alignment among multiple sentences is important
  - Word based DP matching ⇒ Phonetically-oriented alignment [Ruiz et al, ASRU2015]
- Chunk selection using discriminatively trained model



POWA: Phonetically-Oriented Word Alignment

© Hitachi, Ltd. 2016. All rights reserved.

Phonetically-oriented System Combination

HITACHI
Inspire the Next

Feature vector $x = (x_{cf}^\top, x_{oc}^\top, x_{nl}^\top)^\top$

Geometric mean of confidence score in a chunk

$$x_{cf} = ((\prod_{e=0}^{E} c_e)^{1/E} \; ; 0 < i < H)^\top$$

Co-occerance: whether two chunks are identical

$$x_{oc} = (\delta(w_i, w_j) \; ; 0 < i < j < H)^\top$$

NULL: whether a chunk is NULL

$$x_{nl} = (\delta(w_i, \mathrm{NULL}) \; ; 0 < i < H)^\top$$

Label $y = (\delta(w_h, w_{true}) \; ; 0 < h < H)^\top$

Logistic regression model are trained using development set

© Hitachi, Ltd. 2015. All rights reserved.

Phonetically-oriented System Combination

HITACHI
Inspire the Next

- 12 backend models
  - 4 baselines {GMM, DNN+sMBR, DNN+5-gram, RNNLM}
  - 4 data augmented models
  - 2 adapted DNN models {5-gram, RNNLM}
  - 2 data augmented + adapted DNN models
- 2 frontend speech enhancement
  - Baseline (beamformit)
  - LGM preprocessd beamforming

Test real WERs on 6ch track

Best single recognizer	5.56 %
24-recognizer combination (conventional)	4.75 %
24-recognizer combination with POWA	4.68 %

© Hitachi, Ltd. 2016. All rights reserved.

Summary of experimental evaluation

HITACHI
Inspire the Next

- LGM based system reduce WER especially on 6ch track
- Speaker adaptation is effective when base WER is low
- System combination reduce WER on all tracks.

	Test real WER(%)			Rel. improvement(%)		
system	1ch	2ch	6ch	1ch	2ch	6ch
Baseline	23.59	16.6	11.5	-	-	-
LGM(data augmentad + beamforming)	16.88	12.1	7.78	28.4	27.0	32.1
LGM+ speaker adaptation	13.57	9.09	5.56	19.6	24.8	28.5
system combination	11.42	8.61	4.68	15.8	5.3	15.8

© Hitachi, Ltd. 2016. All rights reserved.

Conclusion

- Local Gaussian modeling based source separation
  - Multi-channel Wiener filter output is useful
  - Introducing semi-stationary constraint to non-target sources improves frontend speech enhancement
  - Achieved up to 32.1% gain from baseline
- Unsupervised deep neural network adaptation
  - Unsupervised re-training of DNN works well for speaker adaptation when using conservative training parameters
  - Achieved up to 28.5% gain on 6ch track
- Phonetically-oriented system combination
  - Word alignment considering phonetic similarity improves system combination
  - Achieved up to 15.8% gain on 6ch track

© Hitachi, Ltd. 2016. All rights reserved.

Future work

HITACHI
Inspire the Next

- Cross-adaptation or Committee-based approach [Kanda et al, Interspeech2016] for speaker adaptation
  – Supervision from other systems gives better performance
- Noise environment adaptation
  – noise adaptation is more desired than speaker adaptation in robot applications; a speaker in front of a robot changes rapidly but noise environment is relatively fixed
- Deep learning based multi-channel Wiener filter and joint training of the filter and acoustic model
  – Many studies on this field are found in Interspeech 2016

© Hitachi, Ltd. 2016. All rights reserved.

HITACHI
Inspire the Next