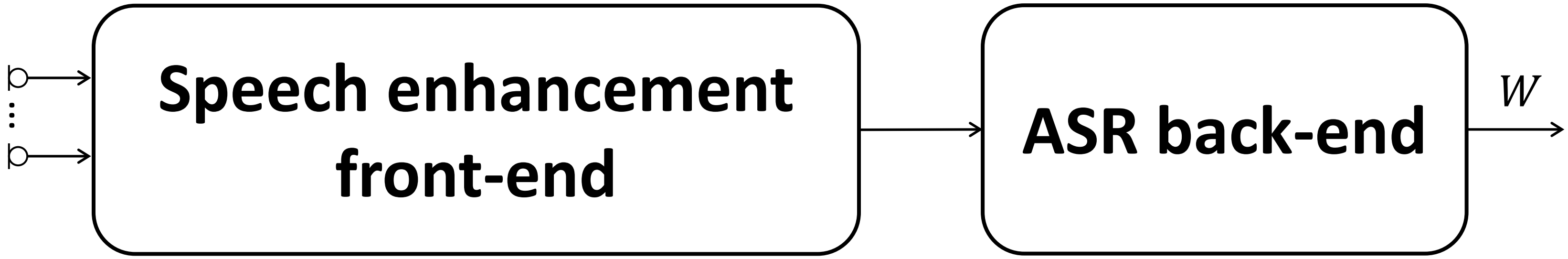
Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge

M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, A. Nakamura

NTT Communication Science Laboratories

Summary

- Combine state-of-the-art SE and ASR techniques



SE task: 2ch & 8ch

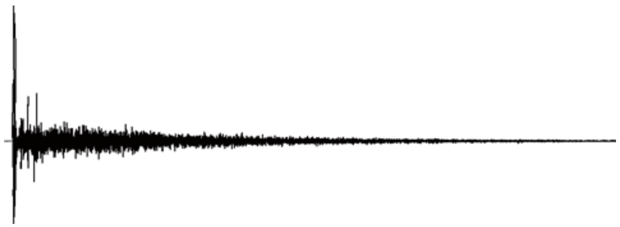
ASR task: 1ch, 2ch & 8ch

- Significant performance improvement:

SE 8ch (FWSegSNR) :	3.62 dB → 10.31 dB
ASR 8ch (WER) :	4.2 % (SimData)
	9.0 % (RealData)

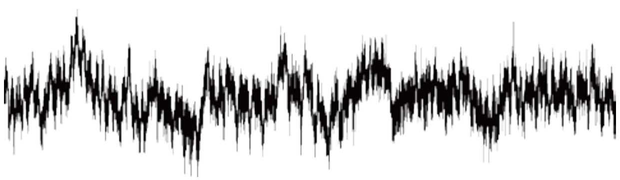
The challenges of the challenge…

1. Severe reverberation



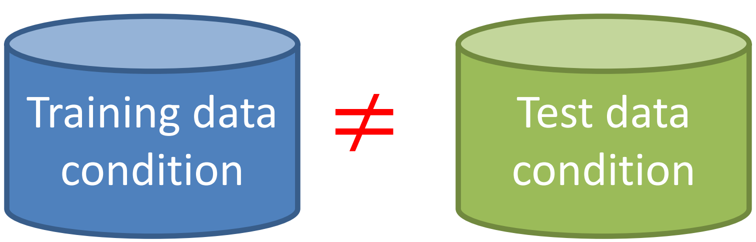
→ Powerful dereverberation

2. Significant amount of noise



→ Noise robust dereverberation
+ additional noise reduction

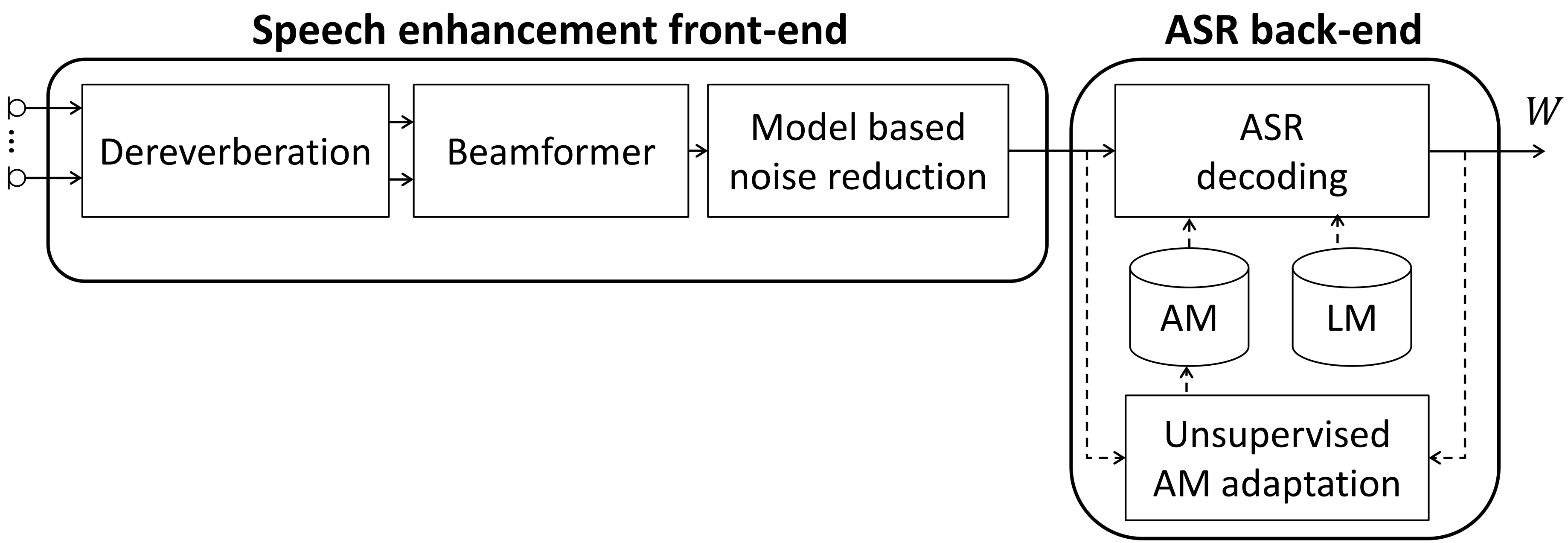
3. Mismatch between training and testing



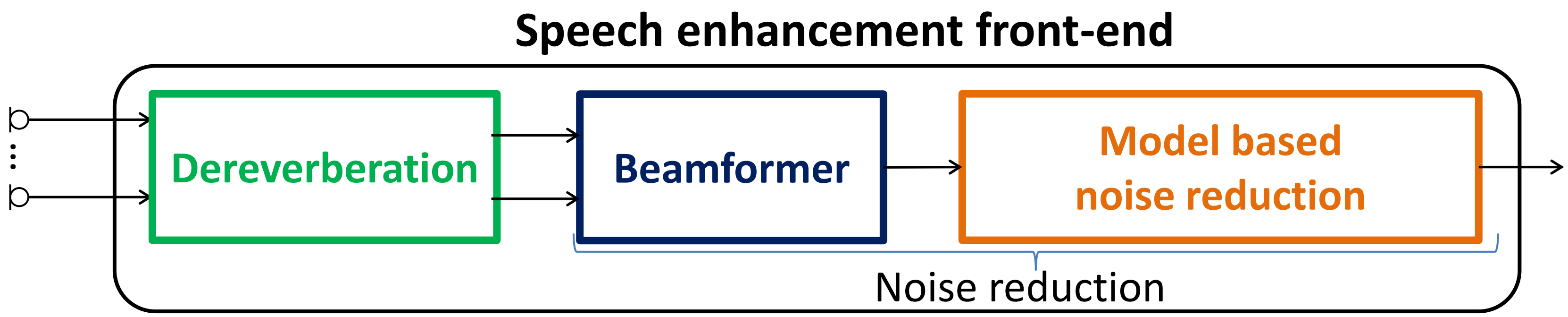
→ High performance w/o overfitting

Extended training data
+ unsupervised environmental adaptation

Proposed system

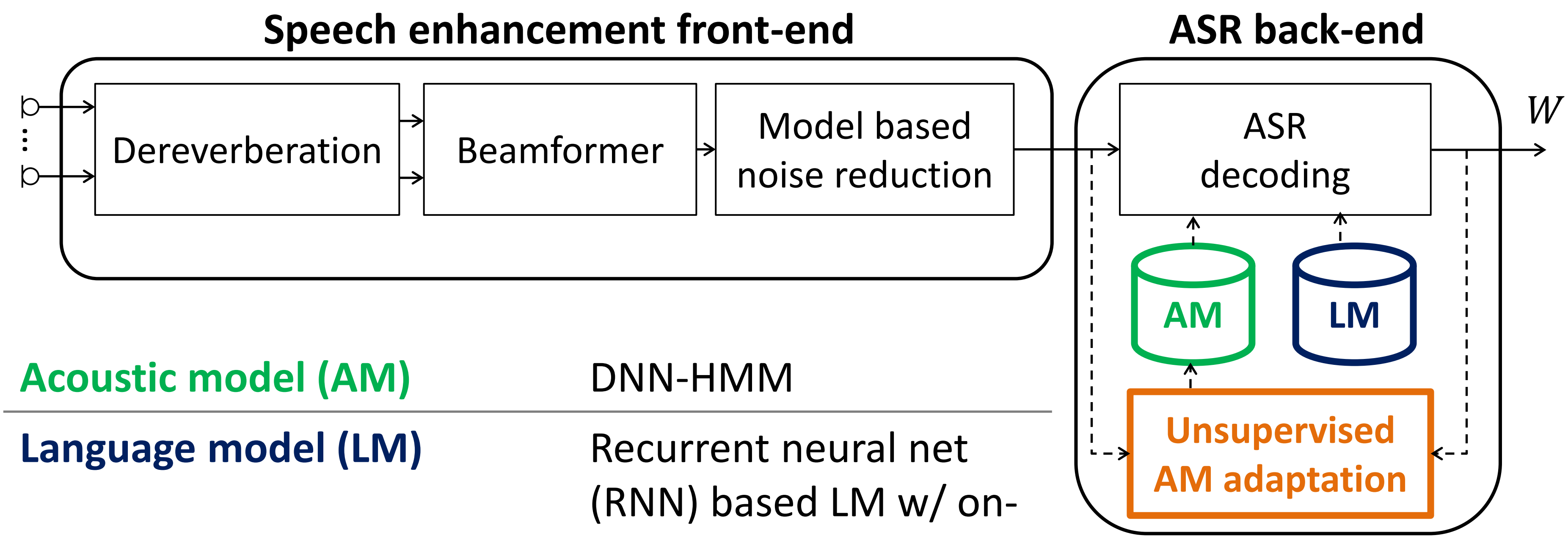


Characteristics of the SE front-end



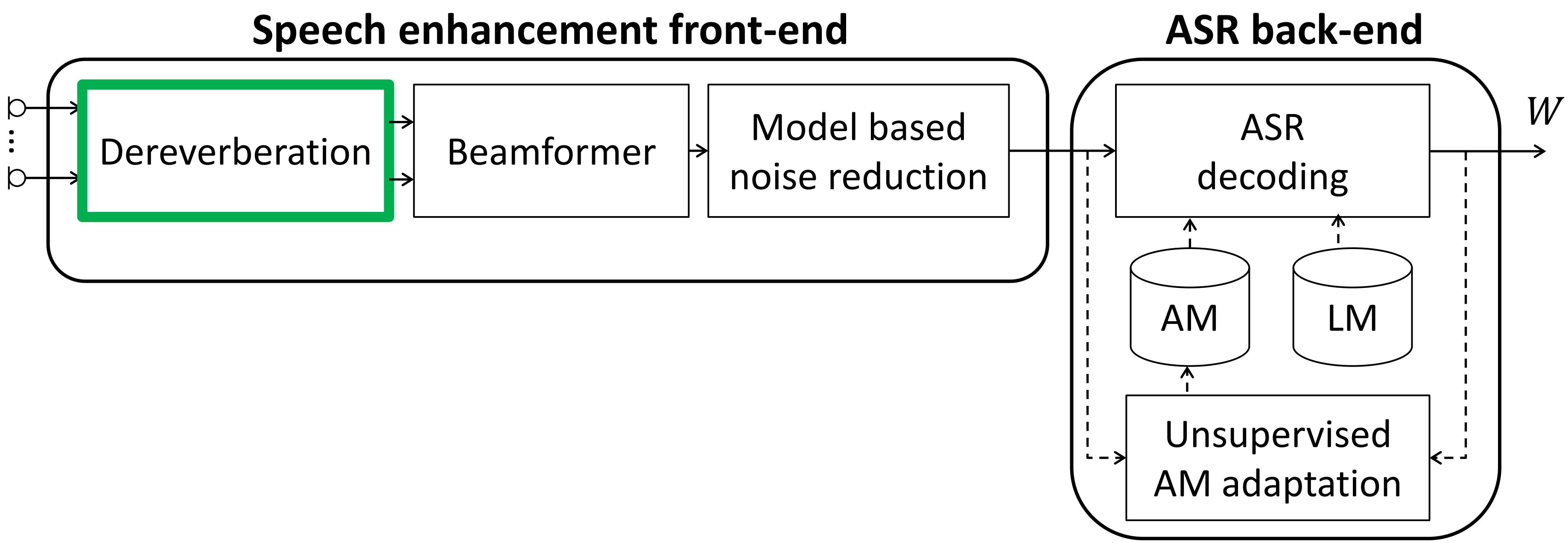
Approach	**Linear-prediction based derev. [Yoshioka, 2012]**	**MVDR beamformer [Souden, 2010]**	**Modified VTS [Fujimoto, 2012]**	**DOLPHIN [Nakatani, 2013]**
Processing	**Linear filtering**	**Linear filtering**	Model-based SE	Model-based SE
Scenario	1ch/2ch/8ch SE/ASR	2ch/8ch SE/ASR	**2ch/8ch** **SE**	**2ch/8ch** **ASR**
Output	**1ch/2ch/8ch**	1ch	1ch	1ch
Proc. mode	Utterance batch	Utterance batch	Utterance batch	Full batch
RTF	0.2-2.8	0.03	0.5	6-10

Characteristics of the ASR back-end



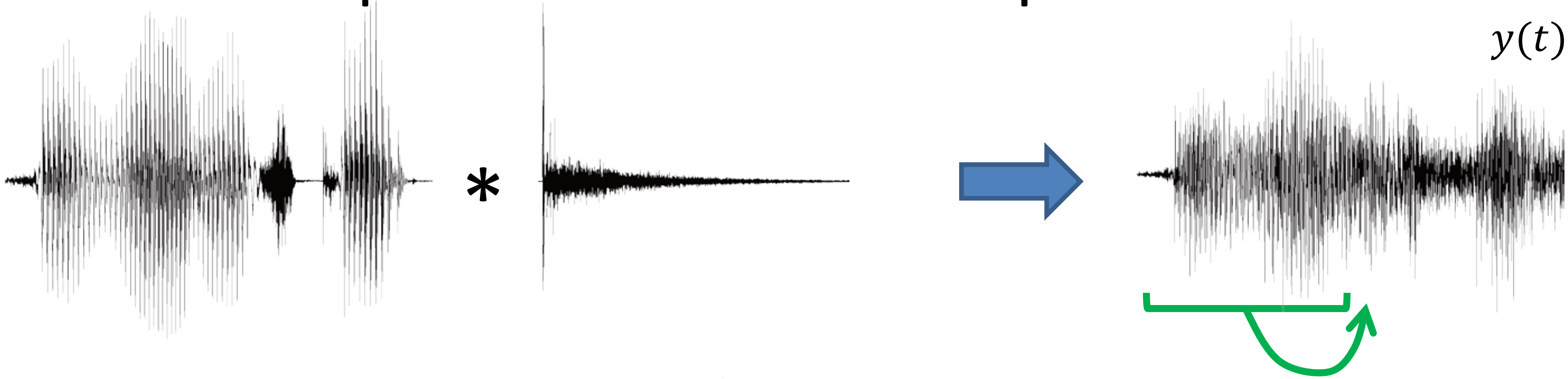
Acoustic model (AM)	DNN-HMM
Language model (LM)	Recurrent neural net (RNN) based LM w/ on-the-fly rescoring [Hori, 2014]
Unsupervised environmental adaptation	Retraining of the input layer of DNN-HMM

Dereverberation



Dereverberation based on linear prediction (1/2)

- Linear prediction for blind equalization



Predict reverberation in current observation from past obs.

- Prediction residual:

$$e(t) = y(t) - \sum_{n=1} g_n \, y(t-n)$$

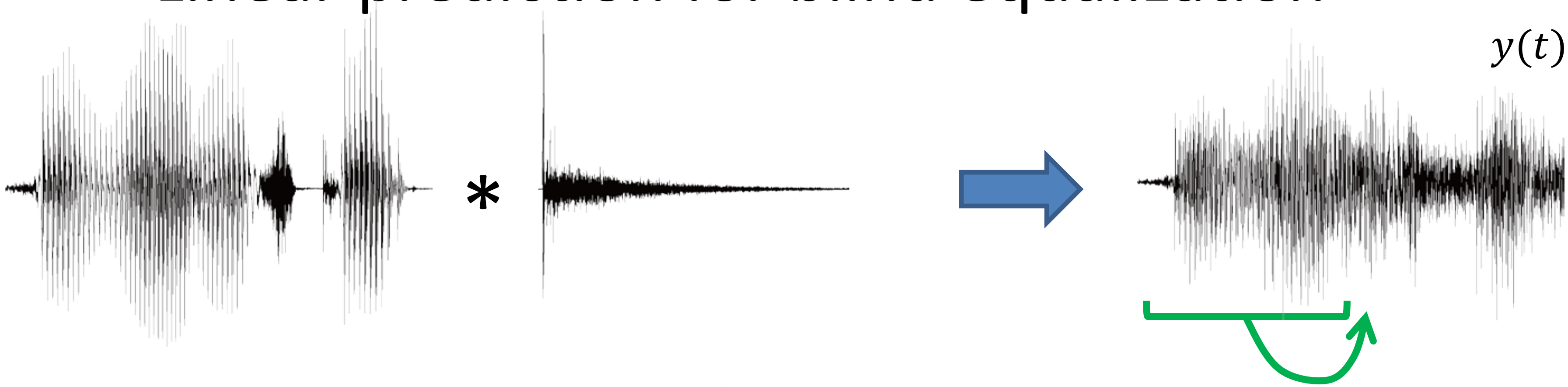
g_n: prediction filter

- Assuming $e(t)$ is Gaussian, g_n can be obtained by minimizing:

$$F = \sum_{t} \left| y(t) - \sum_{n=1}^{N} g_n \, y(t-n) \right|^2$$

Dereverberation based on linear prediction (1/2)

- Linear prediction for blind equalization



Predict reverberation in current observation from past obs.

- Prediction residual:

$$e(t) = y(t) - \sum_{n=1} g_n \, y(t-n)$$

g_n: prediction filter

- Assuming $e(t)$ is Gaussian, g_n can be obtained by minimizing:

$$F = \sum_t \left| y(t) - \sum_{n=1}^{N} g_n \, y(t-n) \right|^2$$

☹ Not well suited for speech dereverberation

Dereverberation based on linear prediction (2/2)

- For speech dereverberation [Yoshioka, 2012]
  - **Introduce time delay τ**

    Do not equalize speech generative process

    → Focus on late reverberation
  - **Better modeling of speech**

    Short time Gaussianity of speech w/ time varying variance $\sigma_t{}^2$

    → Weighted prediction error (WPE)
- g_n , $\sigma_t{}^2$ can be obtained by minimizing:

$$F = \sum_t \frac{|y(t) - \sum_{n=\tau}^{N} g_n \, y(t-n)|^2}{\sigma_t{}^2} + \sum_t \log(\sigma_t{}^2)$$

Dereverberation based on linear prediction (2/2)

- For speech dereverberation [Yoshioka, 2012]
  - **Introduce time delay τ**

    Do not equalize speech generative process

    → Focus on late reverberation
  - **Better modeling of speech**

    Short time Gaussianity of speech w/ time varying variance $\sigma_t{}^2$

    → Weighted prediction error (WPE)
- g_n , $\sigma_t{}^2$ can be obtained by minimizing:

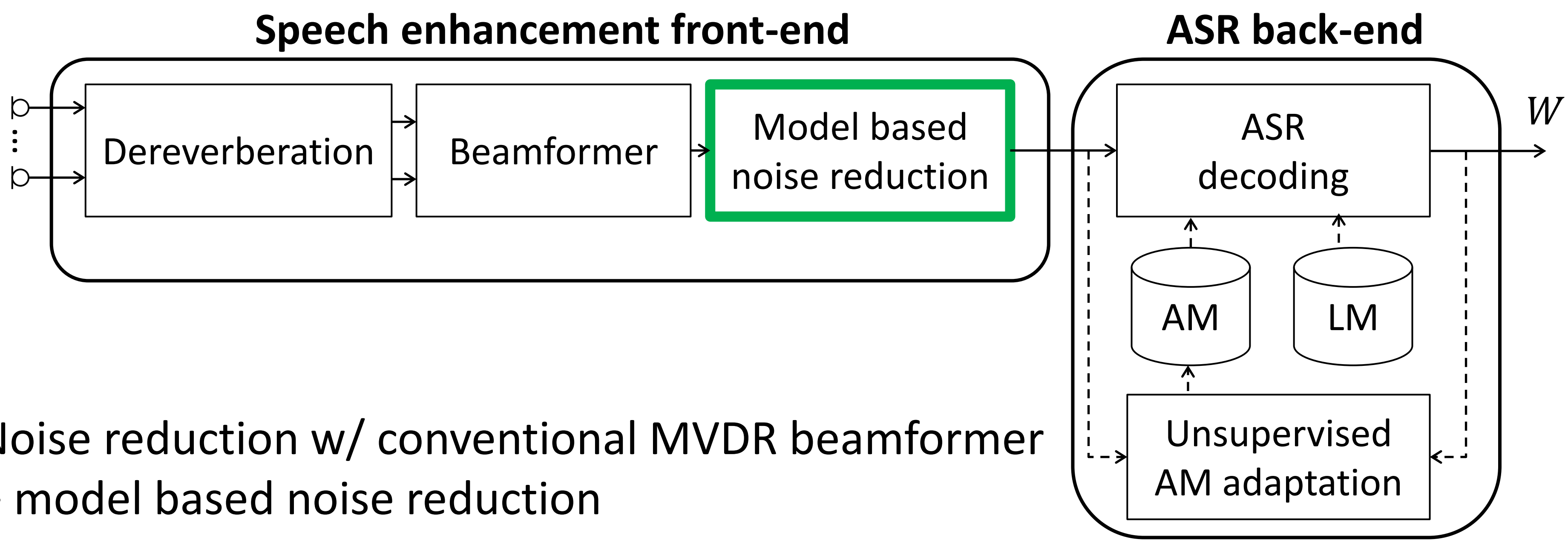
$$F = \sum_t \frac{\left|y(t) - \sum_{n=\tau}^{N} g_n\, y(t-n)\right|^2}{\sigma_t{}^2} + \sum_t \log(\sigma_t{}^2)$$

☺ Precise speech dereverberation

☺ Relatively robust to noise

☺ Implemented in STFT domain → efficient

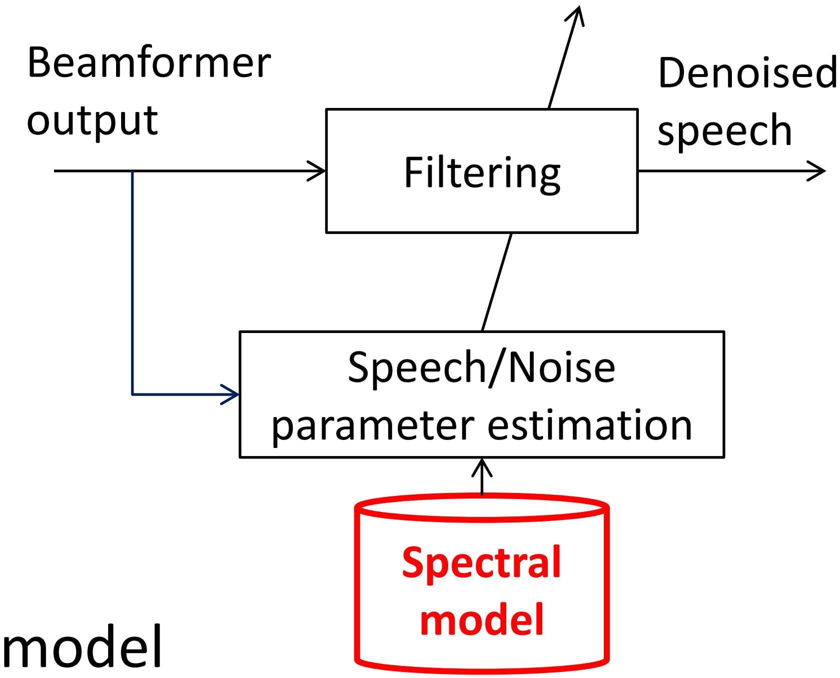
Model based Noise reduction



Noise reduction w/ conventional MVDR beamformer
+ model based noise reduction

Modified VTS (M-VTS)

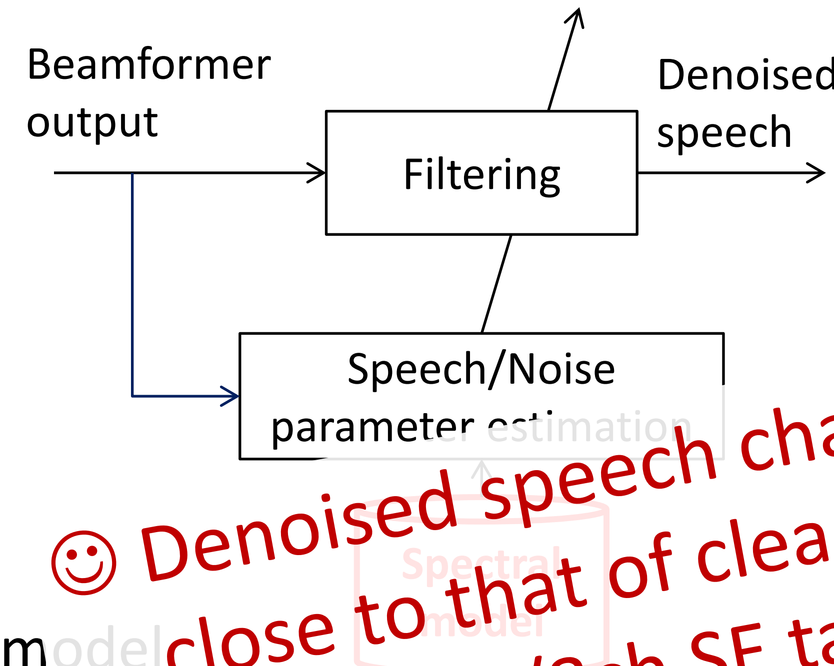
VTS-like model-based noise reduction [Fujimoto, 2012]



- Spectral model
  - Speech model: pre-trained using clean training data + adaptation (per utterance)
  - Noise model: estimated per utterance

Modified VTS (M-VTS)

VTS-like model-based noise reduction [Fujimoto, 2012]



☺ Denoised speech characteristics close to that of clean speech

Used for 2ch/8ch SE tasks

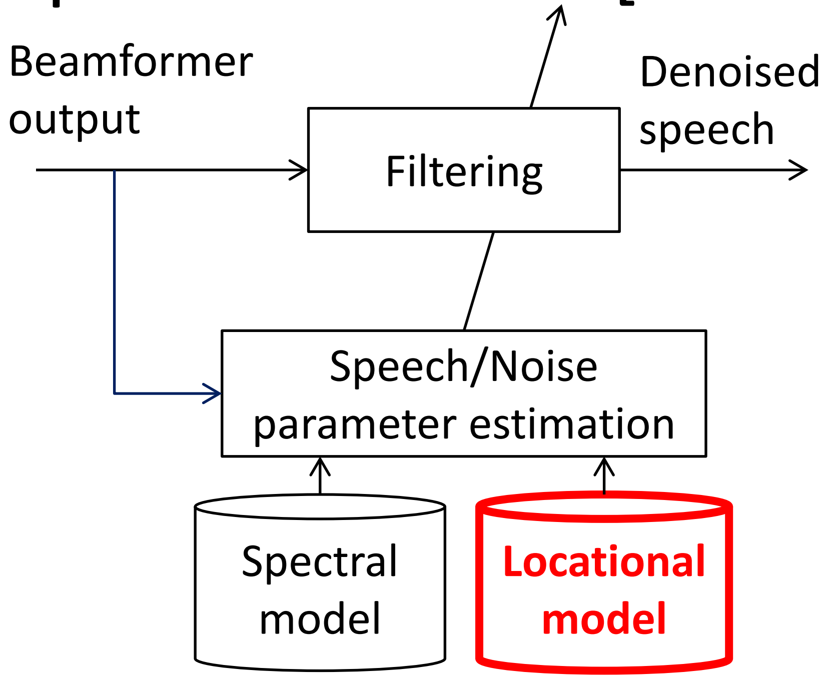
Utterance batch

- Spectral model
  - Speech model: pre-trained using clean training data + adaptation (per utterance)
  - Noise model: estimated per utterance

DOLPHIN*

* Dominance based locational and power-spectral characteristics integration

Combines spectral and locational (DOA feature) models of speech and noise[Nakatani, 2013]

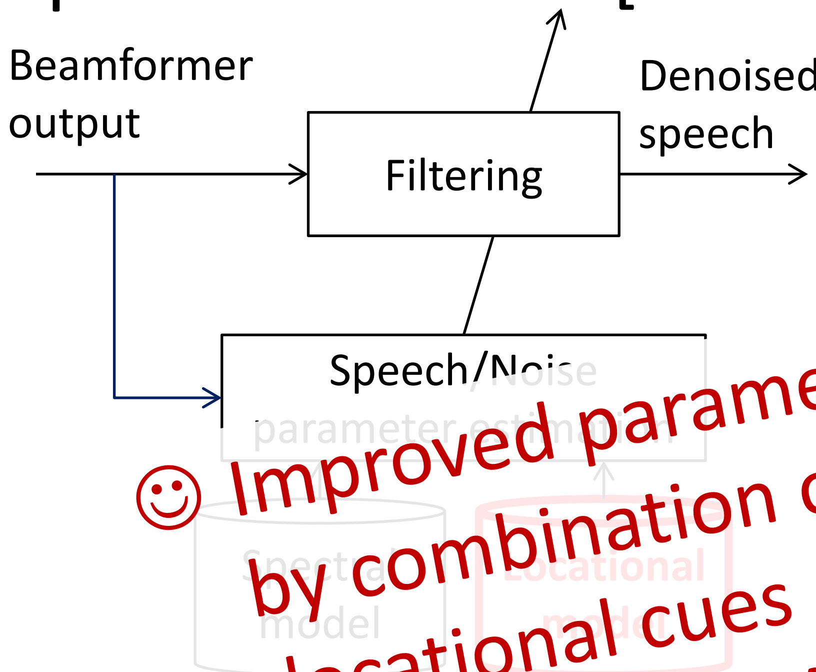


- Locational model estimated from multi-channel dereverberation output

DOLPHIN*

* Dominance based locational and power-spectral characteristics integration

Combines spectral and locational (DOA feature) models of speech and noise[Nakatani, 2013]



☺ Improved parameter estimation by combination of spectral and locational cues

Used for 2ch/8ch ASR tasks

Full batch

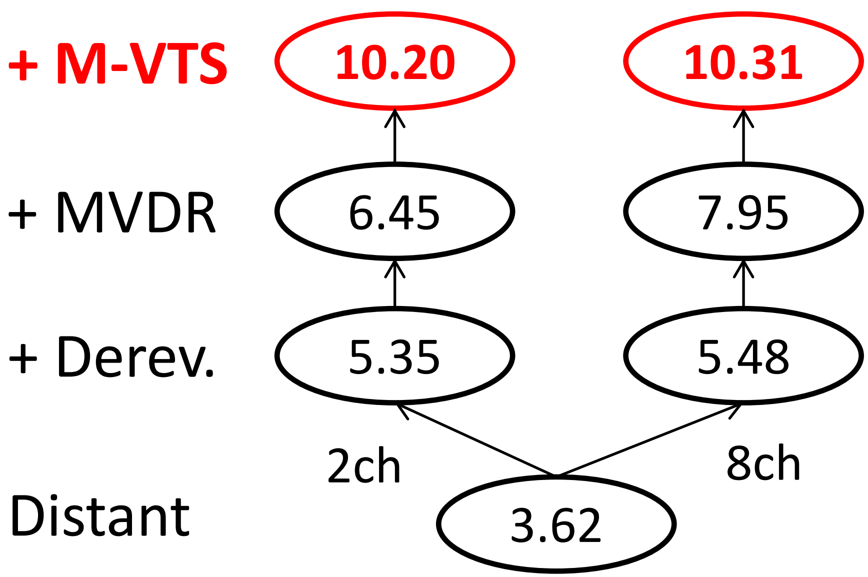
- Locational model estimated from multi-channel dereverberation output

SE results: Objective evaluation (Eval set)

	CD	SRMR	LLR	FWSegSNR	PESQ
w/o front-end	3.97	3.68	0.58	3.62	1.48
Proposed (2ch)	2.34	5.06	0.41	10.20	2.43
Proposed (8ch)	2.25	5.39	0.43	10.31	2.82

Improvement for all measures

Avg. FWSegSNR (dB) *for SimData*



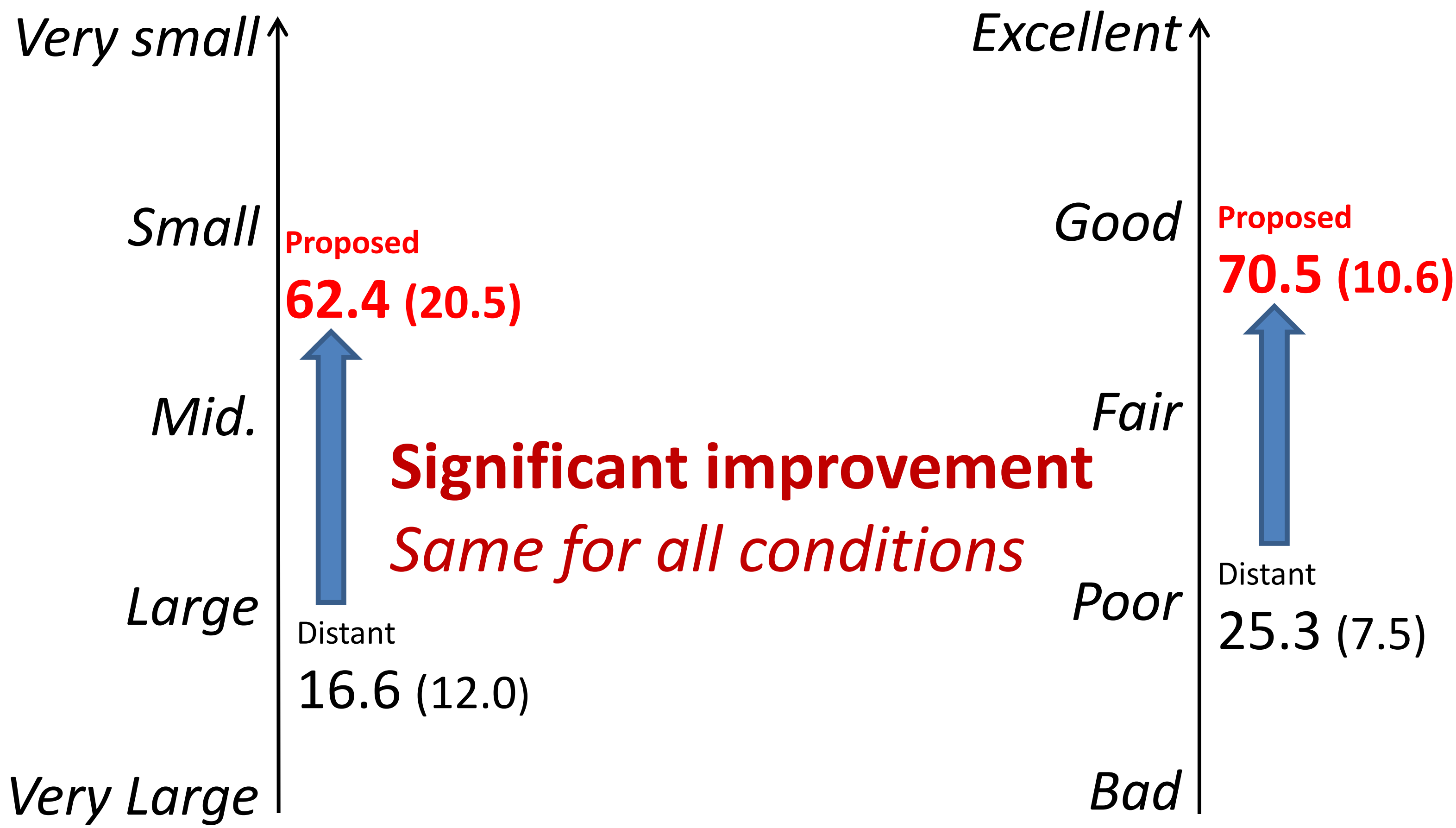
Consistent improvement

SE results: Subjective evaluation (Eval set)

For RealData (far) 8ch

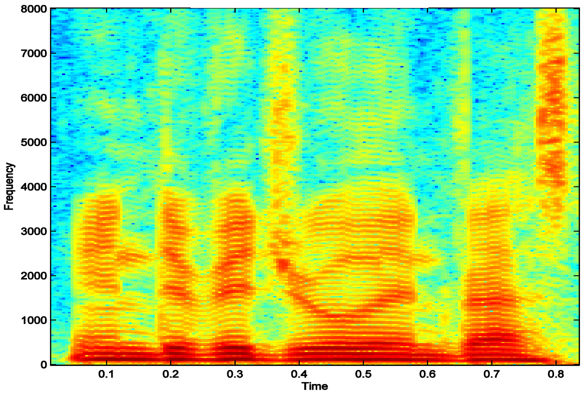
Perceived reverb.

Overall Quality

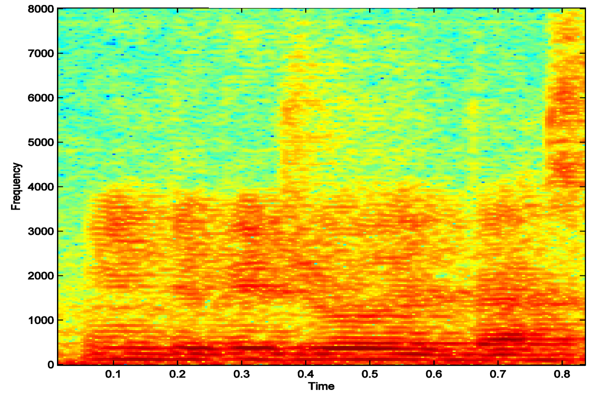


SE results: Spectrograms RealData (Eval)

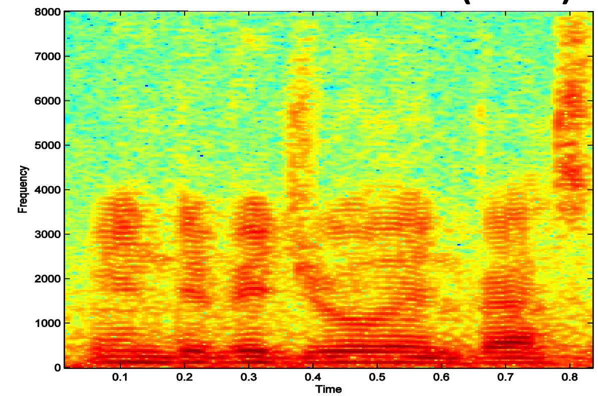
Headset



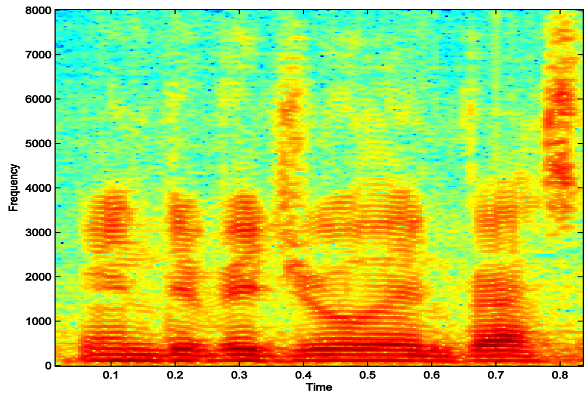
Distant



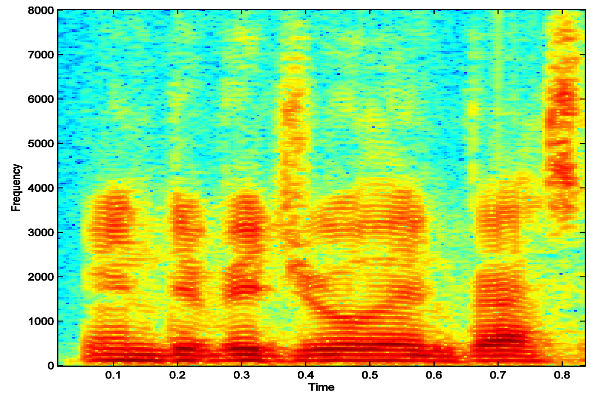
Dereverb. (8ch)



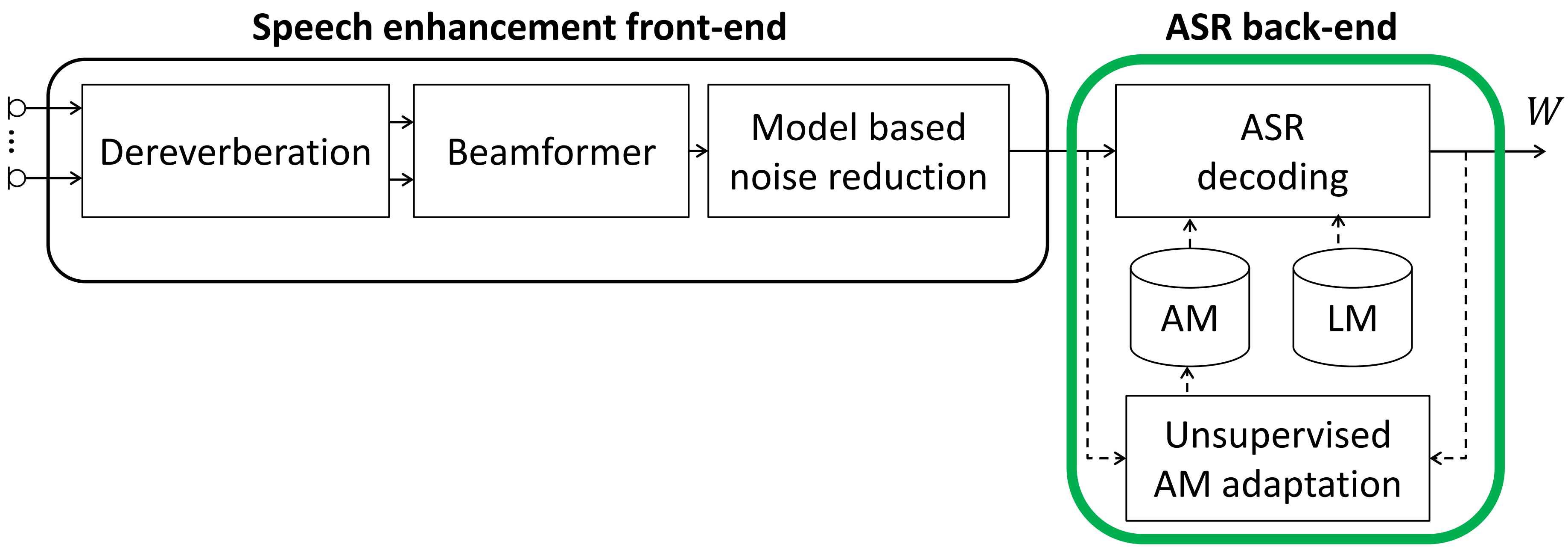
Dereverb.(8ch) + MVDR



Dereverb.(8ch) + MVDR + M-VTS



System overview



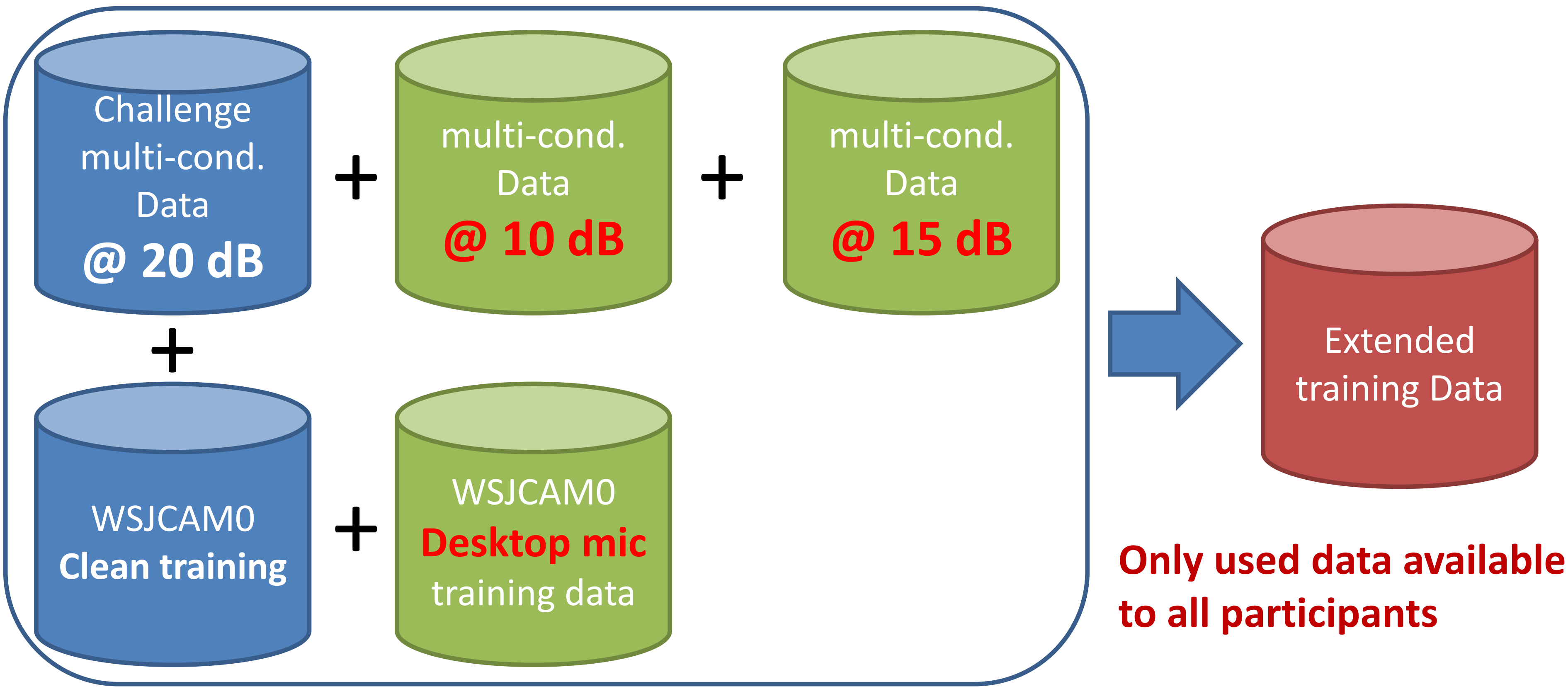
ASR back-end

- DNN-HMM based acoustic model
  - *layer-wise RBM pre-training + fine tuning w/ stochastic gradient descent*
  - *3129 HMM states/ State alignment from clean training data*
  - *7 hidden layers (2048 units)*
- Features
  - *40 Log mel filter-bank coefficients + Δ + $\Delta\Delta$ (120)*
  - *5 left+5 right context (11 frames)*
- Language model LM
  - *Trigram*
  - *RNN with fast decoding using on-the-fly rescoring [Hori, 2014]*
- Unsupervised environmental adaptation
  - *Retrain 1st layer of DNN-HMM w/ small learning rate using labels obtained from a 1st recognition pass*

Extended training data

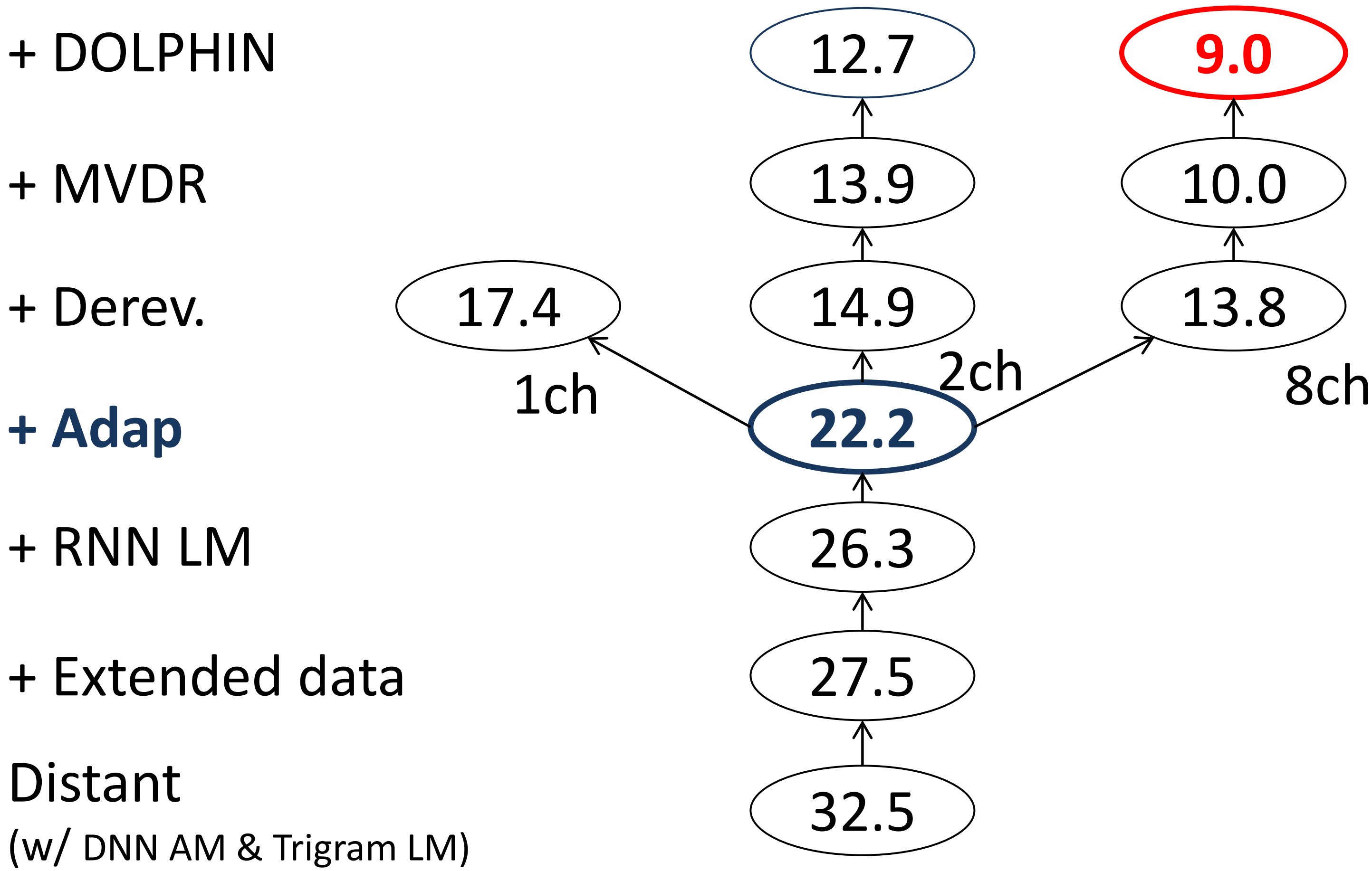
- Combat mismatch between training/SimData and RealData

→ Add acoustic variety during DNN training



Summary of ASR results (Eval)

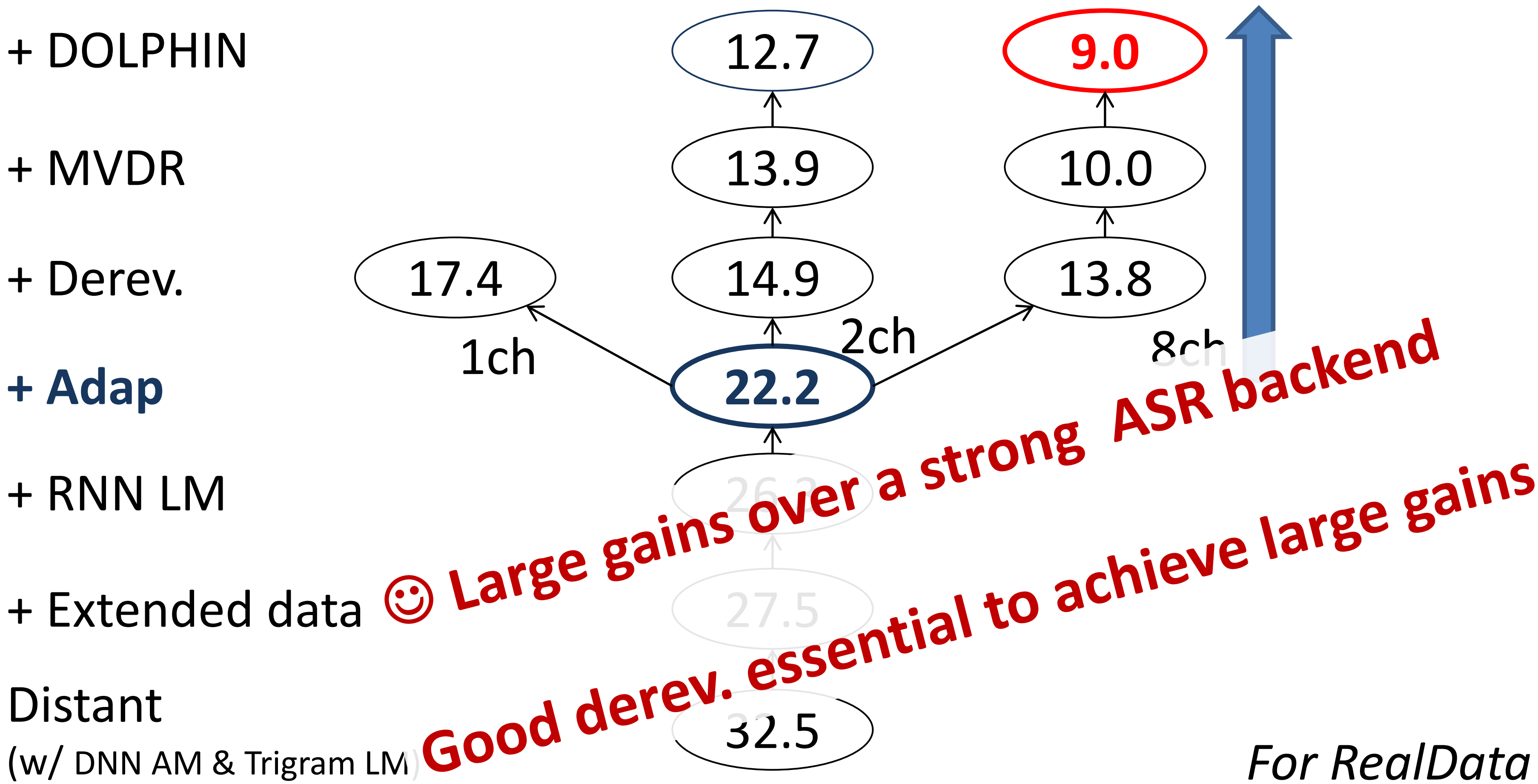
	SimData							RealData		
	Room 1		Room 2		Room 3		**Ave.**	Room1		**Ave.**
	Near	Far	Near	Far	Near	Far		Near	Far	
w/o front-end	3.8	4.4	5.3	8.5	5.8	9.5	**6.2**	21.1	23.3	**22.2**
Best system (8ch)	3.7	4.0	4.0	4.5	4.4	4.8	**4.2**	8.8	9.3	**9.0**



For RealData

Summary of ASR results (Eval)

	SimData							RealData		
	Room 1		Room 2		Room 3		**Ave.**	Room1		**Ave.**
	Near	Far	Near	Far	Near	Far		Near	Far	
w/o front-end	3.8	4.4	5.3	8.5	5.8	9.5	**6.2**	21.1	23.3	**22.2**
Best system (8ch)	3.7	4.0	4.0	4.5	4.4	4.8	**4.2**	8.8	9.3	**9.0**



Conclusion

- Combined advanced SE and ASR techniques for reverberant speech
- Large performance improvement for both SE and ASR tasks

Where are we?

Proposed (1ch)	Proposed (2ch)	Proposed (8ch)	Lapel	Headset (MC-WSJ)	Clean (WSJCAM0)
17.4 %	**12.7 %**	**9.0 %**	8.3 %	**5.9 %**	3.6 %

Room for improvement for robust ASR

Non-environmental mismatch between SimData and RealData

Innovative R&D by NTT

Thank you!