IEEE SPS Webinar, 07:30-09:00 EST, Nov. 14th, 2024

NTT

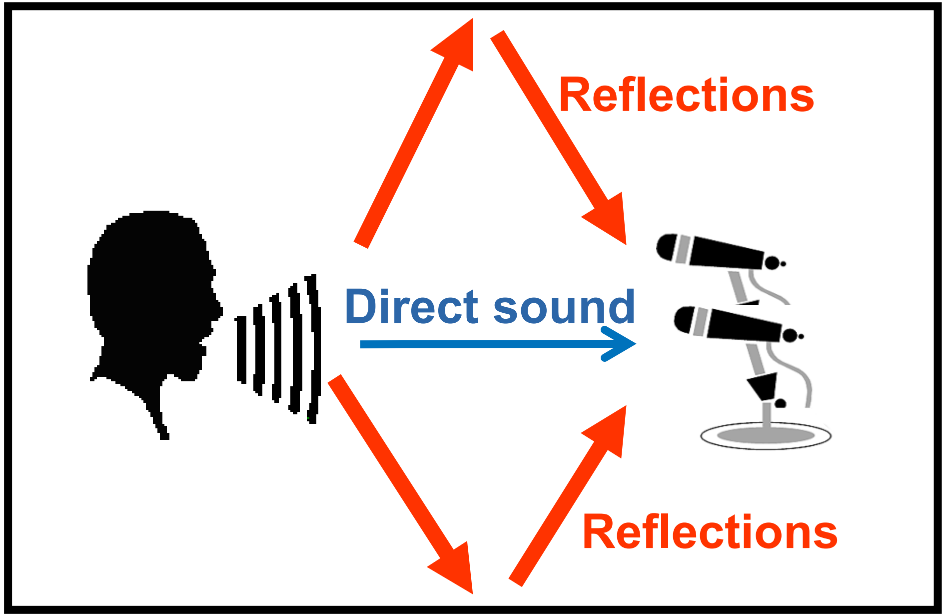
Enhancing Speech Quality: Modern Techniques in Dereverberation

Tomohiro Nakatani

Communication Science Laboratories, NTT Corporation, Japan

What is reverberation?

Reflections from walls, floors, and ceilings

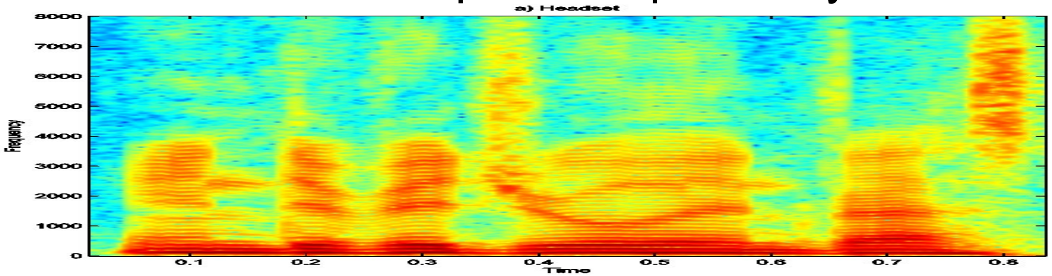


Omnipresent when using a distant mic in an enclosure

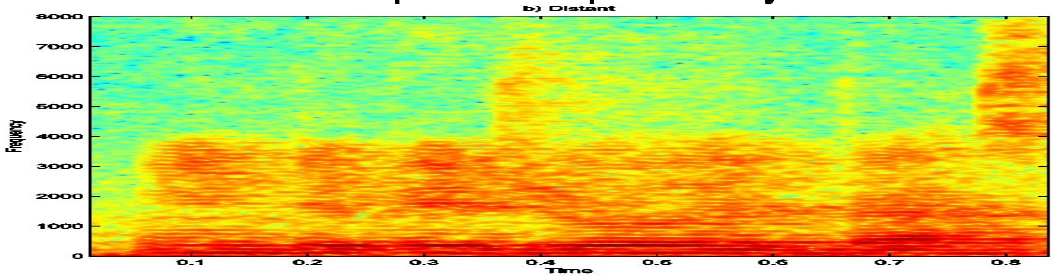
Effect of reverberation (1/2)

Largely modify spectral pattern

Non-reverberant speech captured by a headset



Reverberant speech captured by a distant mic

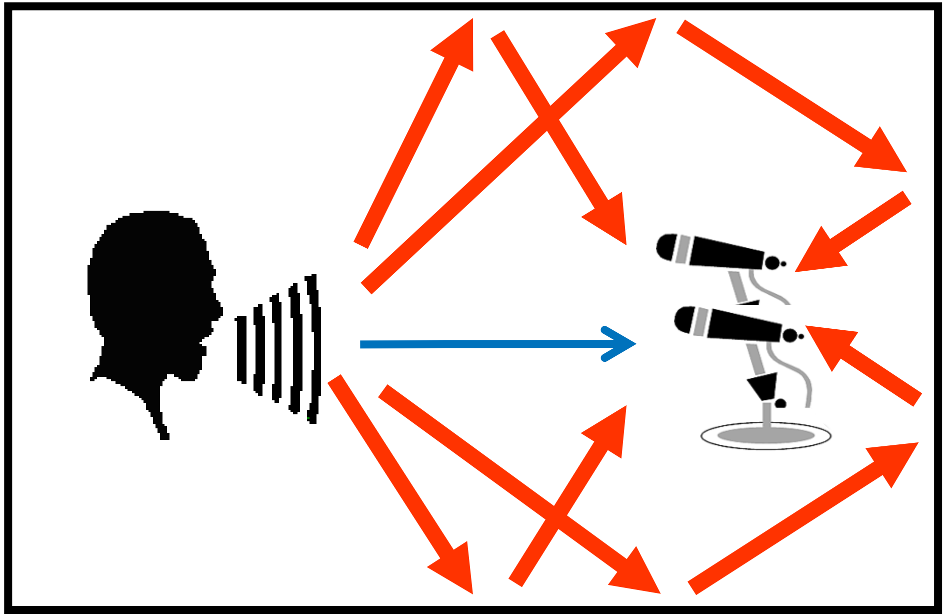


RT60≈0.6 s

- Speech becomes less intelligible for humans
- Automatic Speech Recognition (ASR) becomes very hard

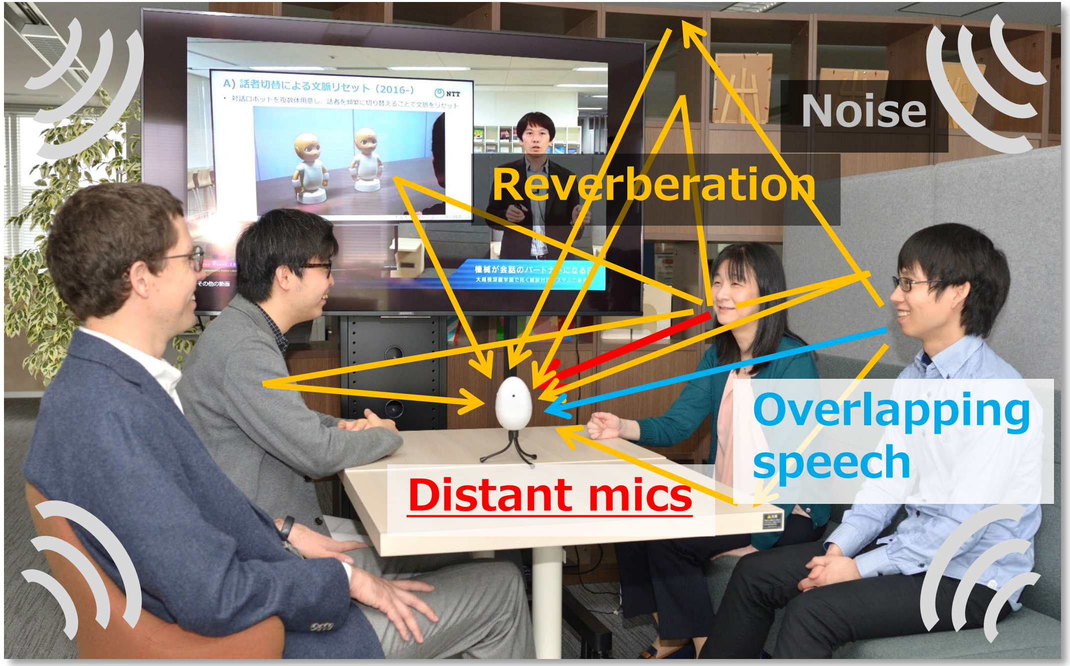
Effect of reverberation (2/2)

Speech arrives at mics from all directions



- Sound localization becomes unclear for humans
- Direction-of-arrival (DOA) estimation becomes challenging

More realistic scenario



Noise + Overlapping speech + Reverberation

Problemss caused by reverberation

- Degrades speech intelligibility and localization for humans
- Degrades performance of speech applications
- Hinders effectiveness of speech preprocessing

Noisy reverberant speech mixture → Speech preprocessing → Speech applications

Speech preprocessing

- Denoising
- Source separation
- Etc.

Degraded by reverberation

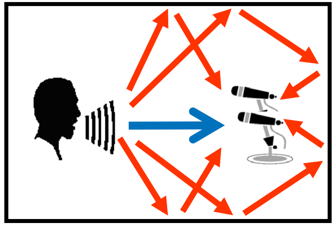
Speech applications

- ASR
- Remote conference
- Etc.

Degraded by reverberation

Role of dereverberation

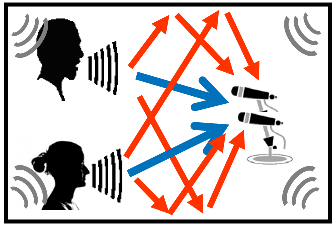
- Reduce reverberation in captured signals to mitigate its negative effects
  - To improve speech intelligibility and localization



Reverberant speech → Dereverb → ***Dereverberated*** speech

Improved intelligibility and localization

  - To improve speech preprocessing and applications



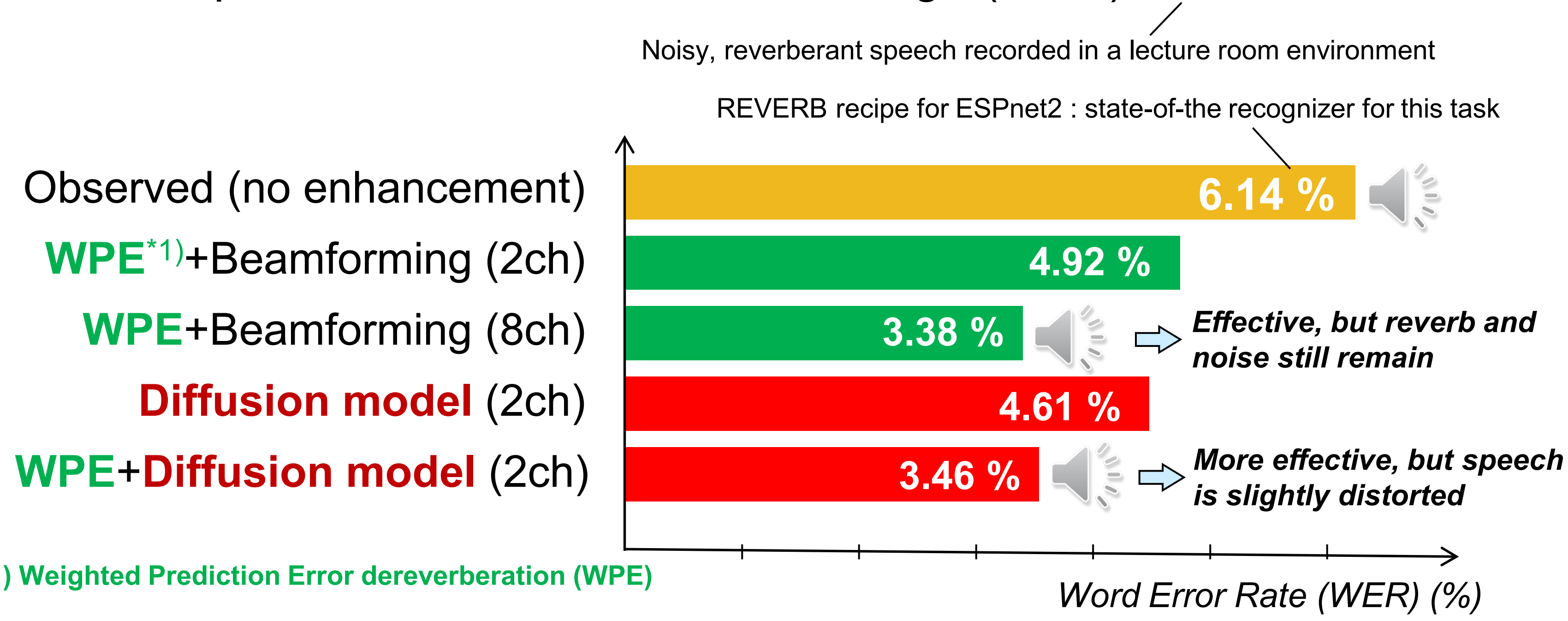
Noisy ***reverberant*** speech mixture → Dereverb → Noisy ***dereverberated*** speech mixture → Speech preprocessing (• Denoising • Source separation) → Speech applications (• ASR • Remote conference)

Improved accuracy

Improved performance

Quick overview of effectiveness

ASR improvement for REVERB Challenge (2014) Real dataset



*1) Weighted Prediction Error dereverberation (WPE)

This webinar puts particular focus on these techniques

Applications of speech dereverberation

A versatile technique to improve quality of speech applications

To enhance human listening

- Hearing aids
- Hands-free remote conference

For computers to understand human conversations

- Smart speaker
- Communication robot
- Meeting recognition

Hearing aids

Remote conference
Minutes generation

Smart speaker

Communication robot

Outline of this talk

1. Approaches to dereverberation
2. Blind inverse filtering-based dereverberation
   - Theoretical background
   - Weighted Prediction Error (WPE) method
   - Extension to joint denoising, dereverberation, and source separation
3. Neural network (NN)-based dereverberation
   - Diffusion model-based joint denoising and dereverberation
   - Integration with WPE and other SE techniques
4. Future challenges and concluding remarks

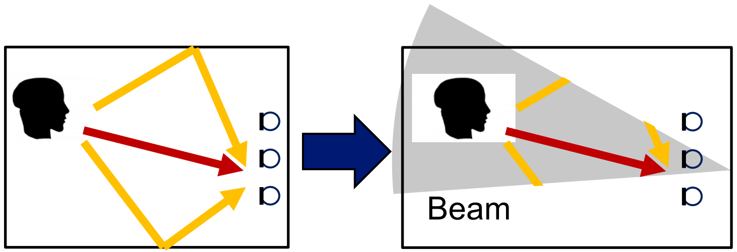
Outline of this talk

1. Approaches to dereverberation
2. Blind inverse filtering-based dereverberation
   - Theoretical background
   - Weighted Prediction Error (WPE) method
   - Extension to joint denoising, dereverberation, and source separation
3. Neural network (NN)-based dereverberation
   - Diffusion model-based joint denoising and dereverberation
   - Integration with WPE and other SE techniques
4. Future challenges and concluding remarks

Signal model-based dereverberation

Beamforming (multi-ch) [Flanagan, 1985]

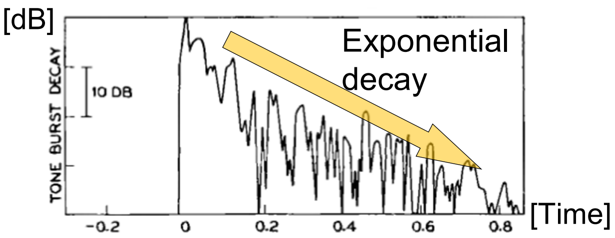
- Model: direct signal comes from source direction
- Solution: enhance signal coming from the source direction
- Requires many mics for large reverb reduction



Power spectral density (PSD) estimation (1-ch)

[Lebart+,2001], [Habets+,2004,2007,2009], [Löllman 2010]

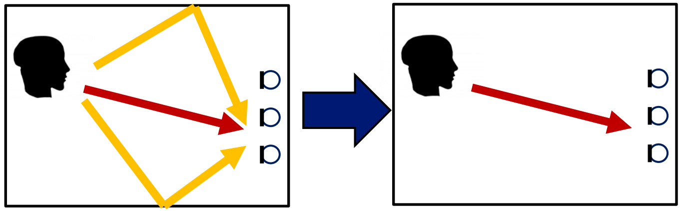
- Model: Energy of reverberation exponentially decays
- Solution: Suppress reverberation PSD in power domain
- Simple and efficient model with marginal effectiveness



Energy decay property of reverberation

Blind inverse filtering (multi-ch)

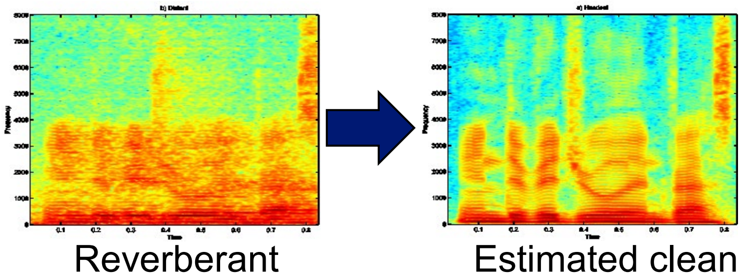
- Model: Convolution with room impulse response (RIR)
- Solution: Apply inverse filter to cancel RIR
  - **Weighted prediction error (WPE) method**
- **One of most effective techniques**



Neural Network (NN)-based dereverberation

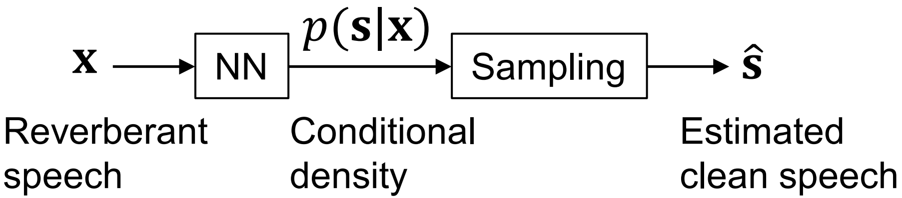
Deterministic prediction (1-ch/multi-ch)

- Train an NN to predict clean speech from reverberant obs. [Weninger+, 2014], [Xu, 2015]
- Use of U-Net [Ronneberger+, 2015] greatly improved the estimation accuracy [Wang, 2021]



Probabilistic prediction (1-ch/multi-ch)

- Train an NN to predict conditional density of clean speech (implicitly or explicitly) from reverberant observation.
- **Diffusion model-based denoising and dereverberation** [Serra+,2022],[Richter+, 2023]
  - An emerging speech enhancement (SE) technique
  - **Can be integrated with signal model-based dereverberation**



Key differences between approaches

NTT

	Blind inverse filtering (Section 2)	NN-based approach (Section 3)	Hybrid (Section 3, and future work)
Prior training	**Not necessary**	**Necessary**	**Necessary**
Adaptability to test condition	**High**	**Limited** (by training data)	**Medium**
Dereverb performance	**Limited** (by signal model)	**High** (Under matched conditions)	**Very high** (Yet depending on conditions)

Outline of this talk

1. Approaches to dereverberation
2. Blind inverse filtering-based dereverberation
   - Theoretical background
   - Weighted Prediction Error (WPE) method
   - Extension to joint denoising, dereverberation, and source separation
3. Neural network (NN)-based dereverberation
   - Diffusion model-based joint denoising and dereverberation
   - Integration with WPE and other SE techniques
4. Future challenges and concluding remark

Time-domain model of reverberation

NTT

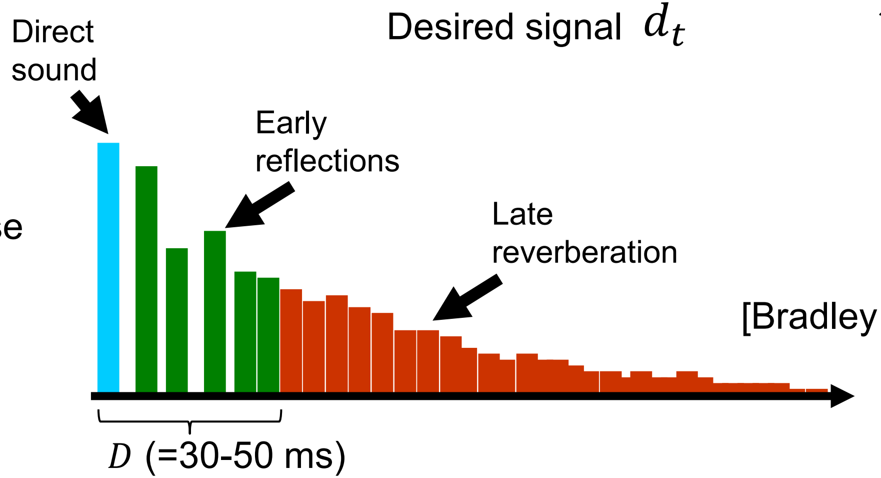
Preserve | Reduce

Reverberant speech $x_t = \sum_{\tau=0}^{L-1} a_\tau s_{t-\tau} = \boxed{\sum_{\tau=0}^{D-1} a_\tau s_{t-\tau}} + \boxed{\sum_{\tau=D}^{L-1} a_\tau s_{t-\tau}}$

Direct sound + Early reflections → Desired signal d_t

Late reverberation → r_t

Impulse response a_τ



[Bradley et al., 2003]

Matrix representation of RIR convolution

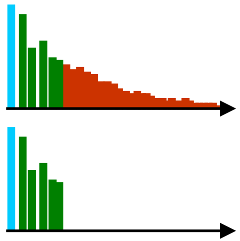
1-ch convolution at mth mic: $\mathbf{x}_{m,t} = \mathbf{H}_m \mathbf{s}_t = \underbrace{\mathbf{H}_m^d \mathbf{s}_t}_{\mathbf{d}_{m,t}} + \underbrace{\mathbf{H}_m^r \mathbf{s}_t}_{\mathbf{r}_{m,t}}$

$$\mathbf{x}_{m,t} = \begin{bmatrix} x_{m,t} \\ x_{m,t-1} \\ \vdots \\ x_{m,t-K} \end{bmatrix} \quad \mathbf{H}_m = \begin{bmatrix} a_{m,0} & a_{m,1} & \cdots & a_{m,L-1} & 0 & \cdots & 0 \\ 0 & a_{m,0} & a_{m,1} & \cdots & a_{m,L-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \ddots & 0 \\ 0 & \cdots & 0 & a_{m,0} & a_{m,1} & \cdots & a_{m,L-1} \end{bmatrix} \in \mathbb{R}^{K \times K_0} \quad \mathbf{s}_t = \begin{bmatrix} s_t \\ s_{t-1} \\ \vdots \\ s_{t-K_0} \end{bmatrix}$$

$K_0 = L + K - 1$

Multi-ch convolution: $\mathbf{x}_t = \mathbf{H}\mathbf{s}_t = \underbrace{\mathbf{H}^d \mathbf{s}_t}_{\mathbf{d}_t} + \underbrace{\mathbf{H}^r \mathbf{s}_t}_{\mathbf{r}_t}$

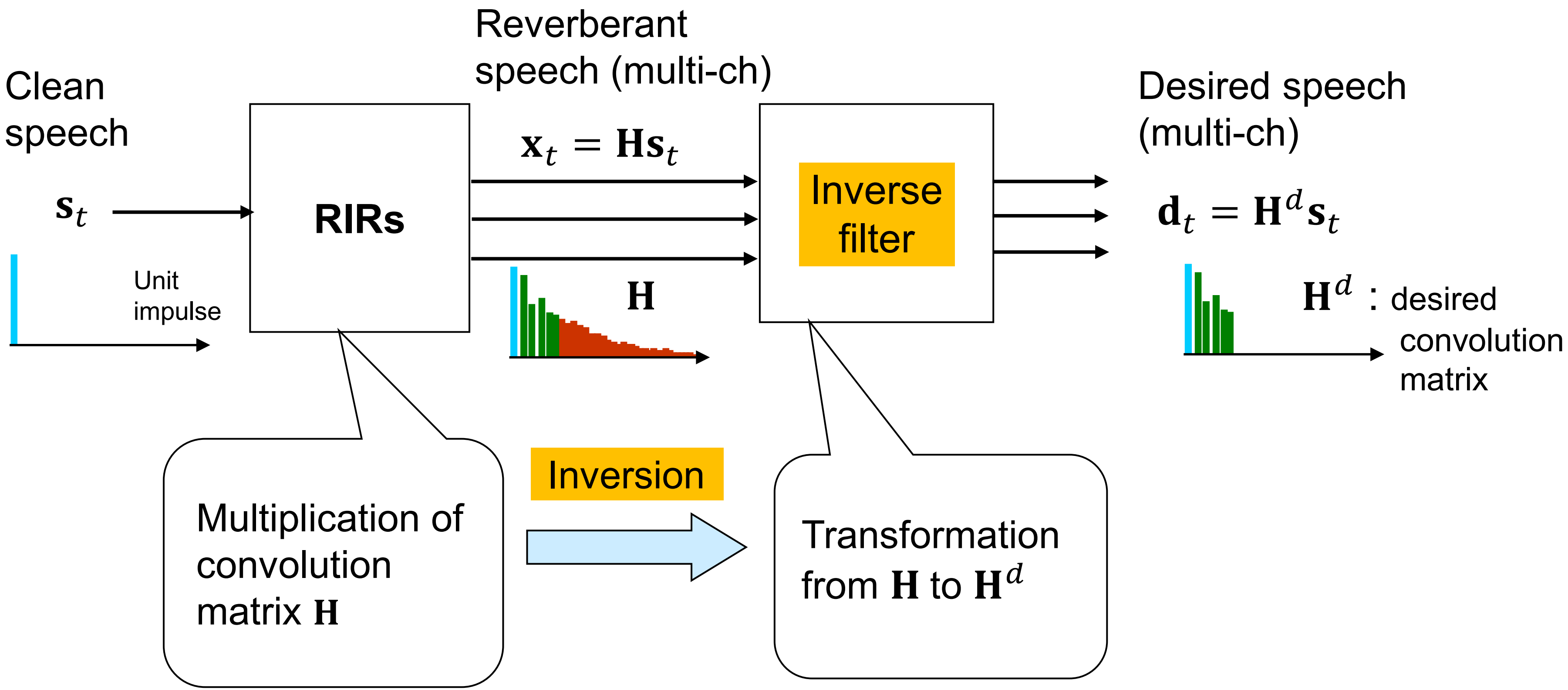
$$\mathbf{x}_t = \begin{bmatrix} \mathbf{x}_{1,t} \\ \vdots \\ \mathbf{x}_{M,t} \end{bmatrix} \quad \mathbf{H} = \begin{bmatrix} \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_M \end{bmatrix} \in \mathbb{R}^{MK \times K_0}$$



$\mathbf{H}$: convolution matrix

$\mathbf{H}^d$: ***desired*** convolution matrix

What is inverse filtering?



Exact inverse filter for given RIR

[Miyoshi and Kaneda, 1988]

- Given $\mathbf{H}$, the inverse filter $\mathbf{W}$ should transform $\mathbf{H}$ to $\mathbf{H}^d$:

$$\mathbf{W}^{\mathsf{H}}\mathbf{H} = \mathbf{H}^d$$

- Solution is obtained using the pseudo-inverse of $\mathbf{H}$ denoted by $\mathbf{H}^+$:

$$\mathbf{W}^{\mathsf{H}} = \mathbf{H}^d\mathbf{H}^+ \quad \text{where} \quad \mathbf{H}^+ = \left(\mathbf{H}^{\mathsf{H}}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathsf{H}}$$

  – When $\mathbf{H}$ is **full column rank** (requiring **#mics**≥ 2)

$\mathbf{H}$ is not given in a blind inverse filtering scenario

The challenge is to estimate $\mathbf{W}$ without knowing $\mathbf{H}$

STFT-domain convolution model

NTT

- For computational efficiency, we decompose time-domain convolution by STFT-domain convolution at each frequency

Time-domain convolution

$$\mathbf{x}_t = \mathbf{H}\mathbf{s}_t$$

Decompose

STFT-domain convolution

$$\left.\begin{aligned} \mathbf{x}_{t,f_1} &= \mathbf{H}_{f_1}\mathbf{s}_{t,f_1} \\ \mathbf{x}_{t,f_2} &= \mathbf{H}_{f_2}\mathbf{s}_{t,f_F} \\ &\vdots \\ \mathbf{x}_{t,f_F} &= \mathbf{H}_{f_F}\mathbf{s}_{t,f_F} \end{aligned}\right\} \text{for each frequency } f$$

  - Valid when frame shift << analysis window [Nakatani+, 2008]

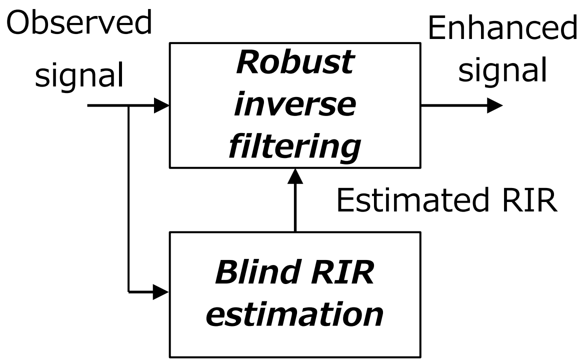
- Exact inverse filter can be defined in the same way as time-domain model

Inverse filtering can be performed separately in each frequency

Approaches to blind inverse filtering

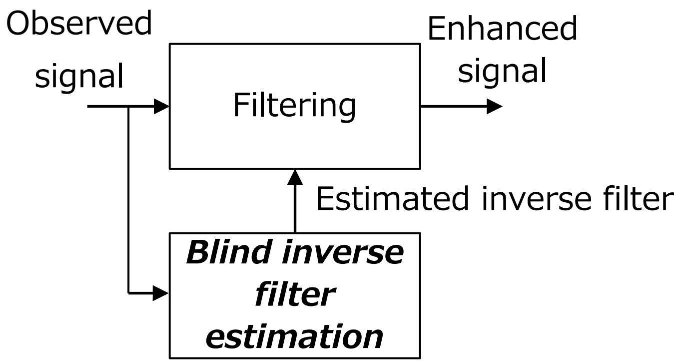
Blind RIR estimation + robust inverse filtering

- Blind RIR estimation is still a challenging problem
  - Eigenvalue decomposition-based [Gannot, 2010]
  - Rank-1 matrix lifting-based joint source and impulse response estimation [Yohena+, 2024]
- Robust inverse filtering for given RIR
  - Regularization [Hikichi+, 2007]
  - Partial multichannel equalization [Kodrasi+, 2013]



Blind and direct estimation of inverse filter

- Multichannel linear prediction (MCLP) based methods
  - Prediction Error (PE) method [Abed-Meraim+, 1997]
  - Delayed Linear Prediction [Kinoshita+, 2009]
  - Weighted Prediction Error (WPE) method [Nakatani+, 2010]
  - Multi-input multi-output (MIMO) WPE method [Yoshioka+, 2012]

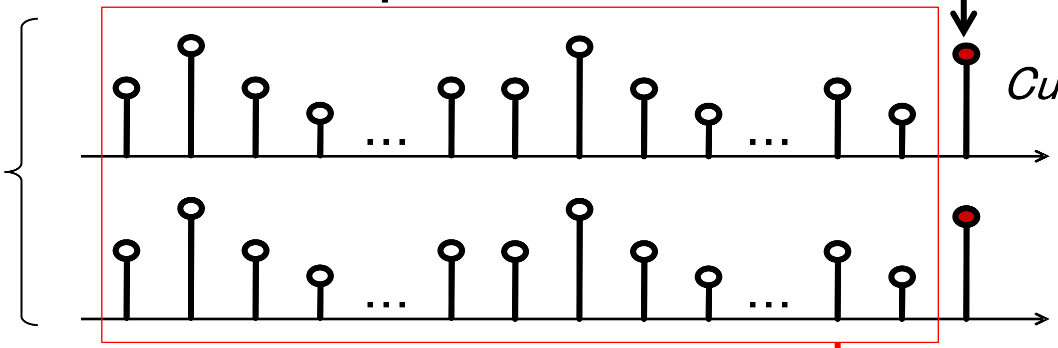


Vanilla MCLP [Abed-meraim+, 1997]

$\dot{\mathbf{d}}_{t,f}$: direct signal
$\dot{\mathbf{r}}_{t,f}$: reverberation

Predict $\sum_{\tau=1} \mathbf{G}_{\tau,f}^{H} \mathbf{x}_{t-\tau,f}$

Multi-ch



Current signal

$$\mathbf{x}_{t,f} = \dot{\mathbf{d}}_{t,f} + \underline{\dot{\mathbf{r}}_{t,f}}$$

Past signal

Predictable

Dereverberation: $\dot{\mathbf{d}}_{t,f} = \mathbf{x}_{t,f} - \underbrace{\sum_{\tau=1} \mathbf{G}_{\tau,f}^{H} \mathbf{x}_{t-\tau,f}}_{\textit{Predicted signal}}$

Subtract predicted signals from observation

Formal definition of vanilla MCLP

Multichannel autoregressive model

$$\mathbf{x}_{t,f} = \sum_{\tau=1}^{L} \mathbf{G}_{\tau,f}^{H} \mathbf{x}_{t-\tau,f} + \dot{\mathbf{d}}_{t,f}$$

$\mathbf{G}_{\tau,f} \in \mathbb{C}^{M \times M}$: prediction matrices.

- Assuming $\dot{\mathbf{d}}_{t,f}$ stationary white noise, Maximum Likelihood (ML) solution becomes

$$\hat{\mathbf{G}}_{\tau,f} = \underset{\{\mathbf{G}_{\tau,f}\}}{\arg\min} \sum_{t=1}^{T} \left\| \mathbf{x}_{t,f} - \sum_{\tau=1}^{L} \mathbf{G}_{\tau,f}^{H} \mathbf{x}_{t-\tau,f} \right\|_2^2$$

- With estimated $\hat{\mathbf{G}}_{\tau,f}$, $\dot{\mathbf{d}}_{t,f}$ is estimated (= inverse filtering) as

$$\hat{\dot{\mathbf{d}}}_{t,f} = \mathbf{x}_{t,f} - \sum_{\tau=1}^{L} \hat{\mathbf{G}}_{\tau,f}^{H} \mathbf{x}_{t-\tau,f}$$

Problems in vanilla MCLP

Speech is not stationary white noise

- » MCLP assumes the desired signal to be temporally uncorrelated
- » Speech signal exhibits short-term correlation (30-50 ms)

→ MCLP distorts the short-time correlation of speech

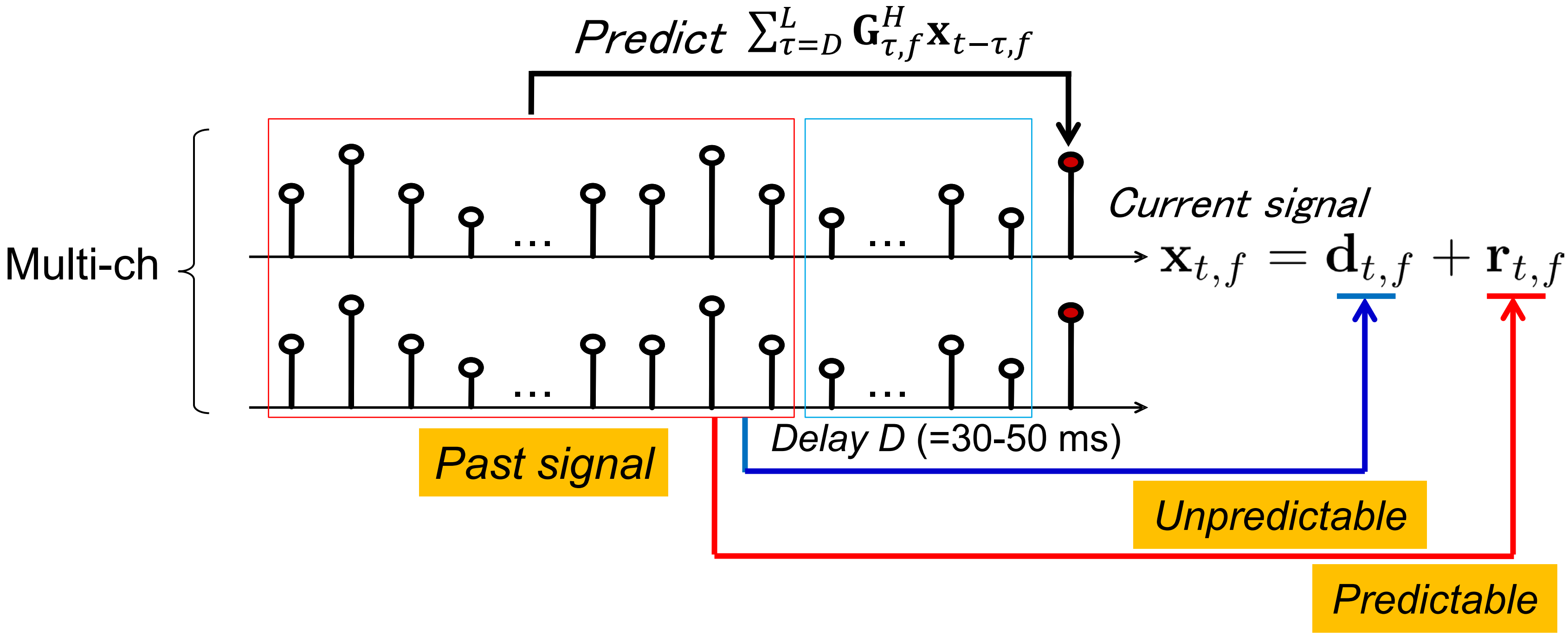
- » MCLP assumes the target signal d to be stationary
- » Speech is not stationary for long-time duration (200-1000 ms)

→ MCLP disrupts the temporal structure of speech

Solutions:

- Use of a prediction delay [Kinoshita+, 2009]
- Use of a non-stationary speech model [Nakatani+, 2010]

Delayed MCLP [Kinoshita+, 2009]



Delayed MCLP can reduce late reverberation $\mathbf{r}_{t,f}$ without distorting temporal correlations of speech

Use of non-stationary source model

[Nakatani+, 2010, Yoshioka+, 2011]

Model of desired signal: time-varying Gaussian (local Gaussian)

$$p(\mathbf{d}_{t,f};\theta) = N_c(\mathbf{d}_{t,f}; 0, \sigma_{t,f}^2 \mathbf{I}) \quad \text{where} \quad \theta = \{\sigma_{t,f}^2\} : \text{source PSD}$$

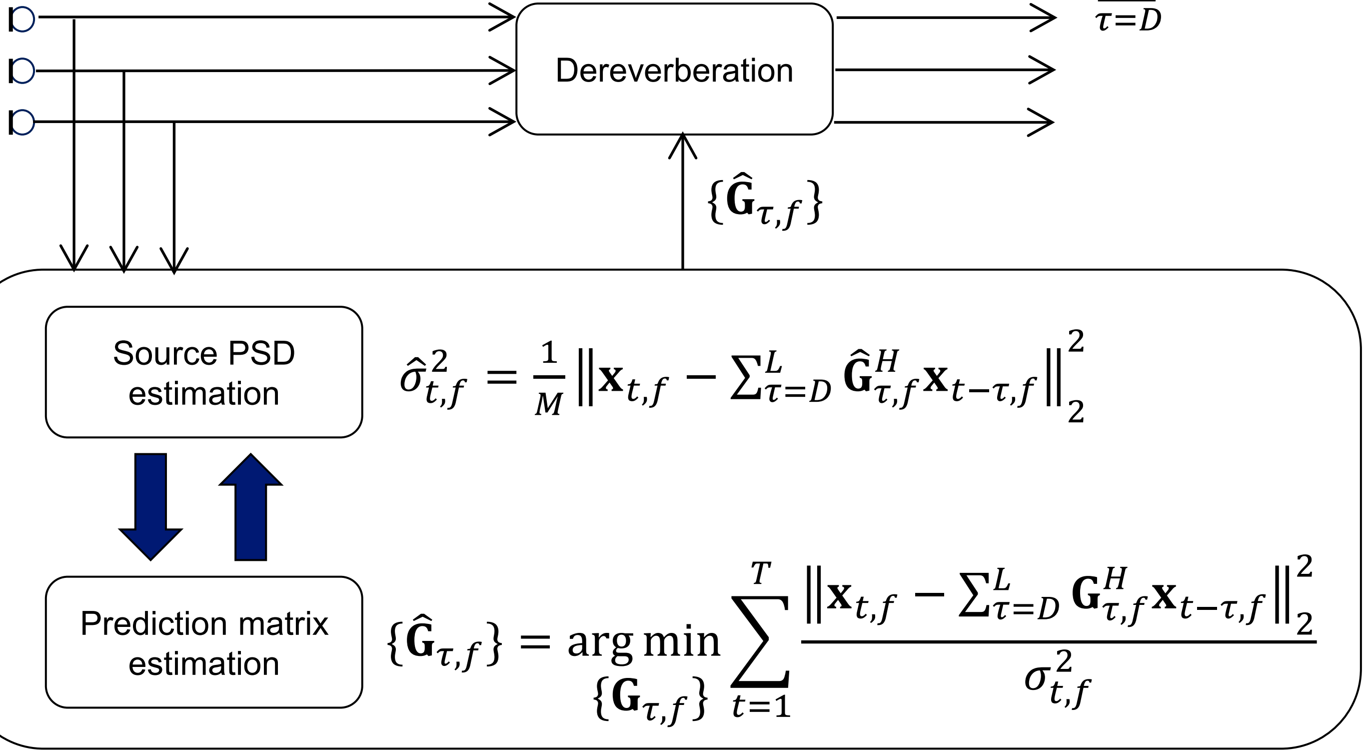
Maximum Likelihood (ML) estimation:

$$\{\hat{G}_{\tau,f}, \hat{\sigma}_{t,f}^2\} = \underset{\{\mathbf{G}_{\tau,f}, \sigma_{t,f}^2\}}{\arg\max} \prod_{t=1}^{T} \frac{1}{\pi \sigma_{t,f}^2} \exp\left(\underbrace{\frac{-\left\| \mathbf{x}_{t,f} - \sum_{\tau=D}^{L} \mathbf{G}_{\tau,f}^H \mathbf{x}_{t-\tau,f} \right\|_2^2}{\sigma_{t,f}^2}}_{\text{Weighted prediction error (WPE)}} \right)$$

→ Can perform dereverberation ***based only on a few seconds of observation***

Processing flow of WPE

$$\hat{\mathbf{d}}_{t,f} = \mathbf{x}_{t,f} - \sum_{\tau=D}^{L} \mathbf{G}_{\tau,f}^{H} \mathbf{x}_{t-\tau,f}$$



$$\{\hat{\mathbf{G}}_{\tau,f}\}$$

$$\hat{\sigma}_{t,f}^{2} = \frac{1}{M} \left\| \mathbf{x}_{t,f} - \sum_{\tau=D}^{L} \hat{\mathbf{G}}_{\tau,f}^{H} \mathbf{x}_{t-\tau,f} \right\|_2^2$$

$$\{\hat{\mathbf{G}}_{\tau,f}\} = \underset{\{\mathbf{G}_{\tau,f}\}}{\arg\min} \sum_{t=1}^{T} \frac{\left\| \mathbf{x}_{t,f} - \sum_{\tau=D}^{L} \mathbf{G}_{\tau,f}^{H} \mathbf{x}_{t-\tau,f} \right\|_2^2}{\sigma_{t,f}^{2}}$$

Does WPE perform inverse filtering?

NTT

$$E\left\{\frac{\left\|\mathbf{x}_{t,f}-\sum_{\tau=D}^{L}\mathbf{G}_{\tau,f}^{H}\mathbf{x}_{t-\tau,f}\right\|_2^2}{\sigma_{t,f}^2}\right\}$$

$$=E\left\{\frac{\left\|\mathbf{d}_{t,f}\right\|_2^2}{\sigma_{t,f}^2}\right\}+E\left\{\frac{\left\|\mathbf{r}_{t,f}-\sum_{\tau=D}^{L}\mathbf{G}_{\tau,f}^{H}\mathbf{x}_{t-\tau,f}\right\|_2^2}{\sigma_{t,f}^2}\right\}$$

Assumption

$\mathbf{d}_{t,f}$ and $\mathbf{r}_{t,f}$ are mutually uncorrelated

$$\geq E\left\{\frac{\left\|\mathbf{d}_{t,f}\right\|_2^2}{\sigma_{t,f}^2}\right\}$$

Minimized when $\underbrace{\mathbf{r}_{t,f}}_{\text{Reverb}}=\sum_{\tau=D}^{L}\underbrace{\mathbf{G}_{\tau,f}^{H}\mathbf{x}_{t-\tau,f}}_{\text{Prediction}}$

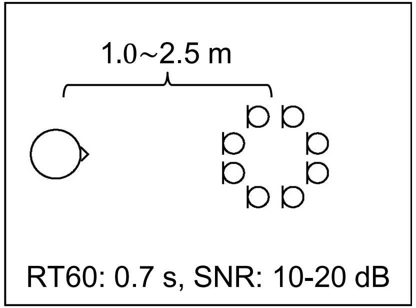
Yes, WPE performs inverse filtering when the inverse filter exists

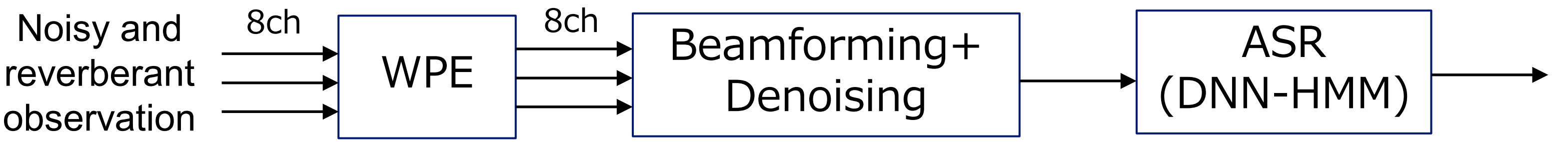
REVERB Challenge : improvement in ASR (2014)

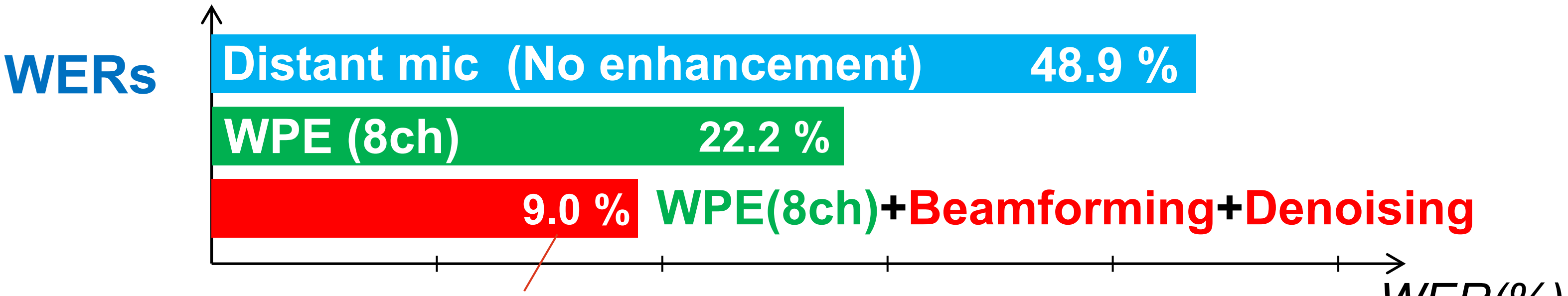
[Kinoshita+, 2016]

- Acoustic conditions
  - Real recordings of read speech
  - Noisy and reverberant lecture rooms
- Processing flow [Delcroix+., 2015]

8ch circular-array scenario





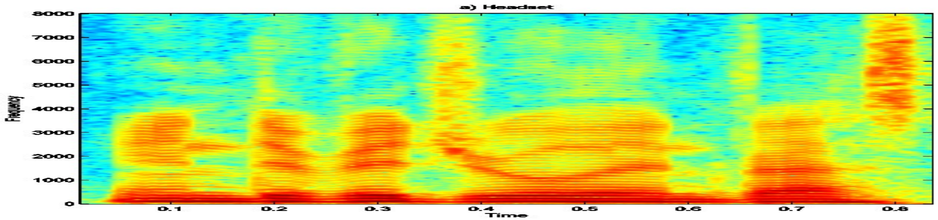


This was the best score at the 2014 challenge

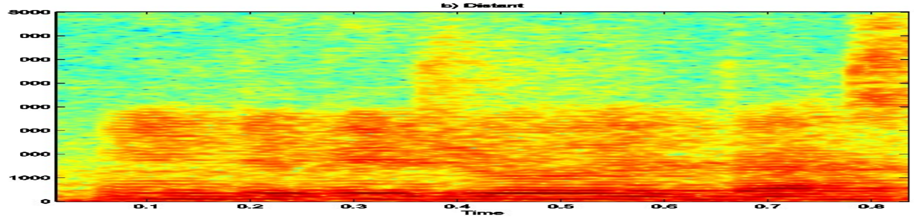
Demonstration（8-mics, Real data）

NTT

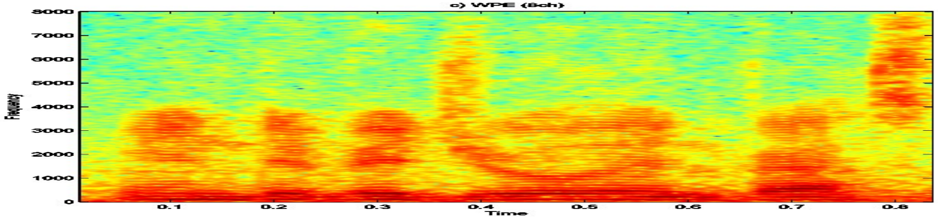
No reverberation
（headset）



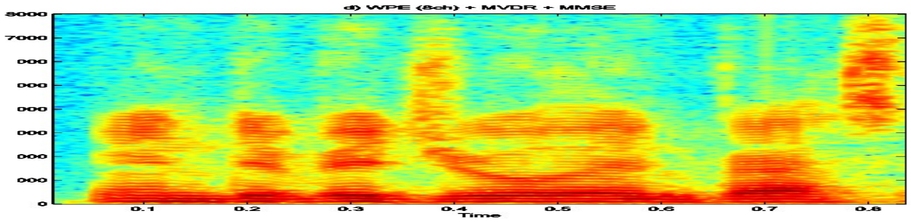
With reverberation
（distant mic）



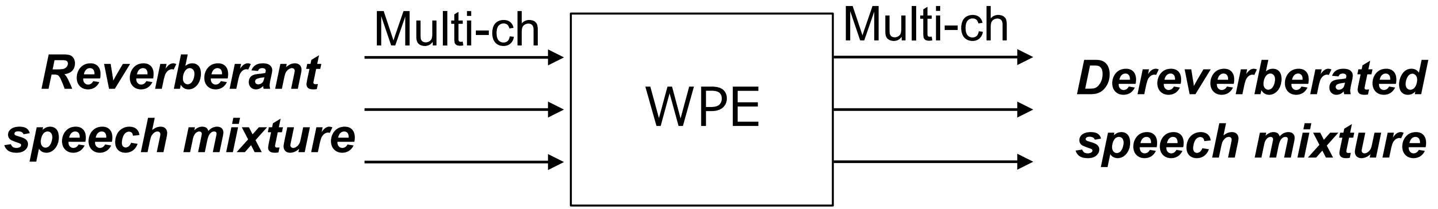
WPE
dereverberation



WPE+beamforming
+denoising



Dereverberation of speech mixture by WPE
[Yoshioka+, 2012]



Existence of such an inverse filter for mixture dereverberation is guaranteed [Miyoshi+, 1988] when

- #mics > #sources
- Convolution matrix for mixture is full column rank

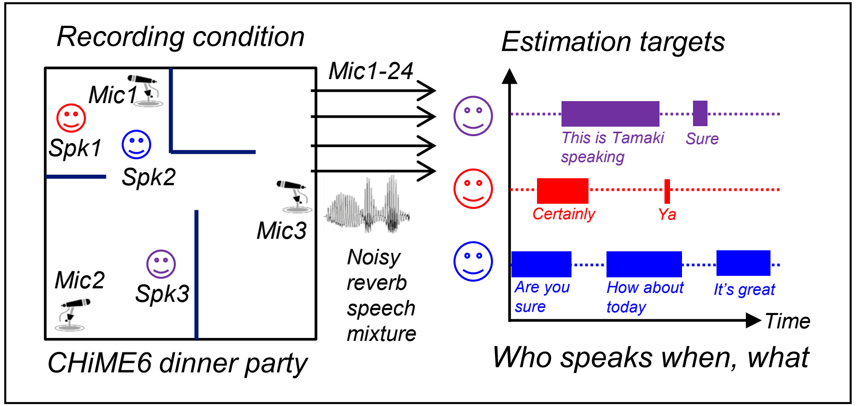
➡ WPE : *versatile dereverberation preprocessor*

Distant ASR challenge CHiME-8 task1 (2024)

[Cornell+, 2024]

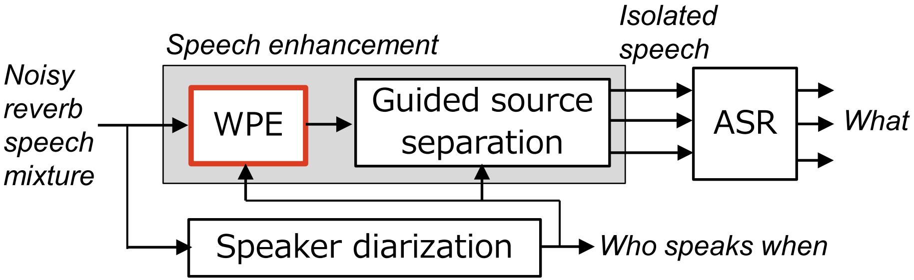
Goal: Estimate who speaks when and what

- In four different conversation scenarios
- Recorded by ***distant distributed mic-arrays***
  - ***Noisy reverberant speech mixture*** with ***unknown number of speakers***



Complex and challenging ASR task

Processing pipeline of baseline system



Results (Dev set): tcpWER*1) of NTT system [Kamo+, 2024]

Scenario (Dataset)	Dinner party1 (CHiME6)	Dinner party2 (DiPCo)	1-to-1 Interview (Mixer 6)	Office meeting (NotSoFar1)	Ave-rage
w/o WPE	21.63	31.22	11.62	9.31	17.52
w/ WPE	**19.80**	**27.33**	**10.13**	**8.93**	**15.85**

*1) Time-constrained minimum permutation WER

Demonstrates effectiveness of WPE for mixture derev.
Further improvement should be included in future work

Extensions of WPE (1/2)

NTT

Elaboration of probabilistic source models

- Sparse prior for speech PSD [Jukic et al., 2015]
- Bayesian estimation with student-T speech prior [Chetupalli+, 2019]

Frame-by-frame online estimation

- Recursive least square [Yoshioka+, 2009], [Caroselli+, 2017]
- Kalman filter for joint denoising and dereverberation [Togami+, 2013], [Braun+, 2018], [Dietzen+, 2018]

Extensions of WPE (2/2)

Dereverberation of more sources than microphones (***under-determined situation***)

- Switching WPE [Ikeshita+, 2021-1, 2021-2]

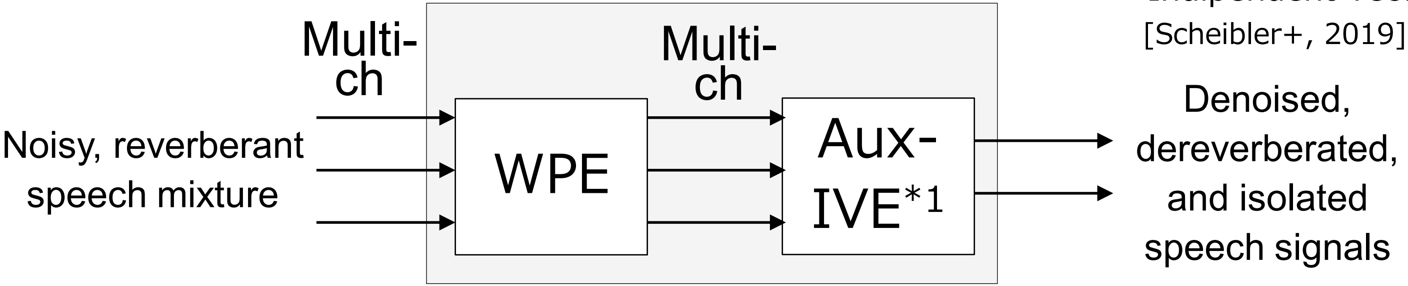
Joint optimization of dereverberation and beamforming

1. Maximum likelihood convolutional beamformer that can jointly perform denoising and dereverberation [Dietzen+, 2018], [Nakatani+, 2019]
2. Integration of WPE and blind source separation (BSS)/extraction (BSE) [Yoshioka+, 2010], [Ikeshita+, 2021], [Nakatani+, 2021]
3. Extension to switching convolutional beamformer [Nakatani+, 2022]

Joint optimization of WPE and BSS/BSE

[Yoshioka+, 2010], [Ikeshita+, 2021], [Nakatani+, 2021]

*1) Auxiliary-function-based Independent Vector Extraction [Scheibler+, 2019], [Ikeshita+, 2020]



Jointly optimize both blocks

Assumptions:

Signals estimated by overall system satisfy:

1. Each speech is time-varying Gaussian
2. Noise is stationary Gaussian
3. Speech signals and noise are mutually independent

Optimiza-tion	fwsSNR↑	STOI↑	WER↓
Separate	5.86 dB	0.83	19.54 %
Joint	**6.16 dB**	**0.84**	**16.31 %**

Results on REVERB-2Mix [Nakatani+, 2021]

Summary of WPE

Advantages:

- Versatile dereverberation preprocessing
  - Applicable to mixed signals
  - Require no prior training or knowledge on recording conditions
    - Highly adaptive to unknown environments

Limitations:

- Performance degrades in high noise conditions
- Relatively a large number of microphones are required for achieving highly accurate processing
- Unable to reduce early reflections

→ Overcome these limitations by using diffusion model-based approach

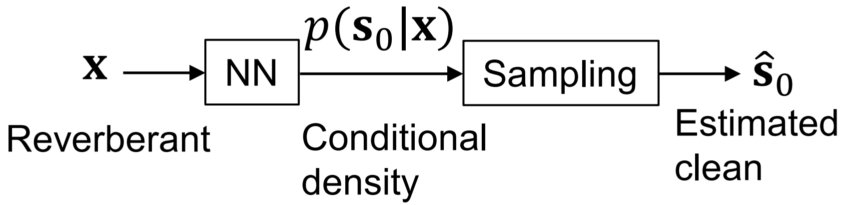
Outline of this talk

1. Approaches to dereverberation
2. Blind inverse filtering-based dereverberation
   - Theoretical background
   - Weighted Prediction Error (WPE) method
   - Extension to joint denoising, dereverberation, and source separation
3. Neural network (NN)-based dereverberation
   - Diffusion model-based joint denoising and dereverberation
   - Integration with WPE and other SE techniques
4. Future challenges and concluding remarks

Diffusion model-based joint denoising and dereverberation

Probabilistic prediction [Serra+,2022],[Richter+, 2023]

- Model $p(\mathbf{s}_0|\mathbf{x})$, i.e., conditional distribution of a clean speech, $\mathbf{s}_0$, given the observed signal, $\mathbf{x}$
- Perform **speech enhancement (SE) by sampling $\mathbf{s}_0$ from $p(\mathbf{s}_0|\mathbf{x})$**
- Score-based Generative Model for Speech Enhancement (SGMSE) [Welker+ 22] [Richter+, 23]



Multi-stream SGMSE (MS-SGMSE) [Nakatani+, 2024]:

- Platform to integrate SE methods using SGMSE
- By conditioning the model with **pre-enhanced signals** by the SE methods

⇒ further improve the SE performance

Conditional density modeled by MS-SGMSE

$$p(\underbrace{\mathbf{s}_0}_{\text{Clean speech}}|\underbrace{\mathbf{x}}_{\text{Observation}},\underbrace{\boldsymbol{X}}_{\text{Pre-enhanced signals}})$$

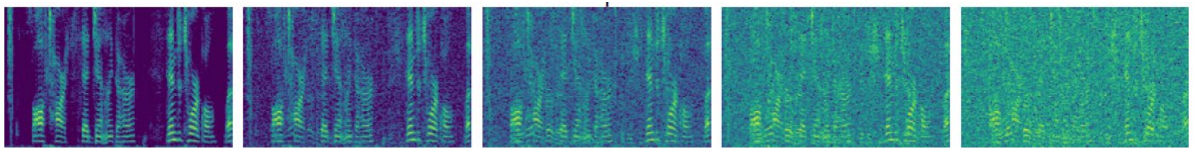
SE using conditional diffusion model

[Song+, 2021],[Richter+,2023]

Forward process

Intermediate state at step *n*

$$\boxed{\mathbf{s}_0 \sim p(\mathbf{s_0}|\mathbf{x},\mathbf{c})} \longrightarrow d\boxed{\mathbf{s}_n} = \mathbf{f}(\mathbf{s}_n,\mathbf{y})dn + g(n)d\mathbf{w} \longrightarrow \mathbf{s}_N = \mathbf{x}+\mathbf{v}$$



$$\mathbf{s}_0 \longleftarrow d\mathbf{s}_n = [-\mathbf{f}(\mathbf{s}_n,\mathbf{y})dt + g(n)^2 \boxed{\nabla_{\mathbf{s}_n} \log p(\mathbf{s}_n|\mathbf{x},\boldsymbol{c})}]dn + g(n)d\bar{\mathbf{w}} \longleftarrow \mathbf{s}_N = \mathbf{x}+\mathbf{v}$$

Score
(modeled by Neural network)

Reverse process

- SE is achieved by the reverse process.

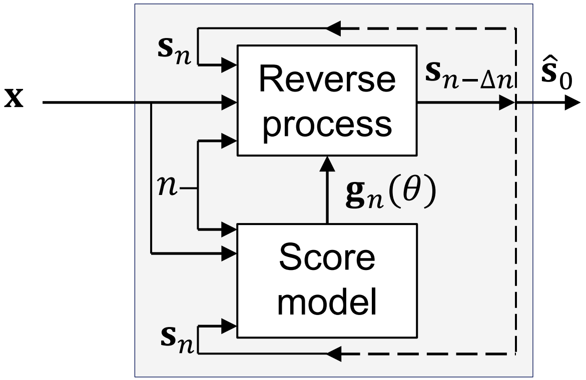
Score model $\mathbf{g}_n(\mathbf{s}_n,\mathbf{x},\boldsymbol{c},n\,;\theta) \simeq \nabla_{\mathbf{s}_t} \log p(\mathbf{s}_n|\mathbf{x},\mathbf{c})$ is all we need.

Loss: $\mathcal{J}^{\text{score}}(\theta) = E\left\|\nabla_{\mathbf{s}_n} \log p(\mathbf{s}_n|\mathbf{x},\boldsymbol{c}) - \mathbf{g}_t\,(\mathbf{s}_n,\mathbf{x},\boldsymbol{c},n;\theta)\right\|_2^2$

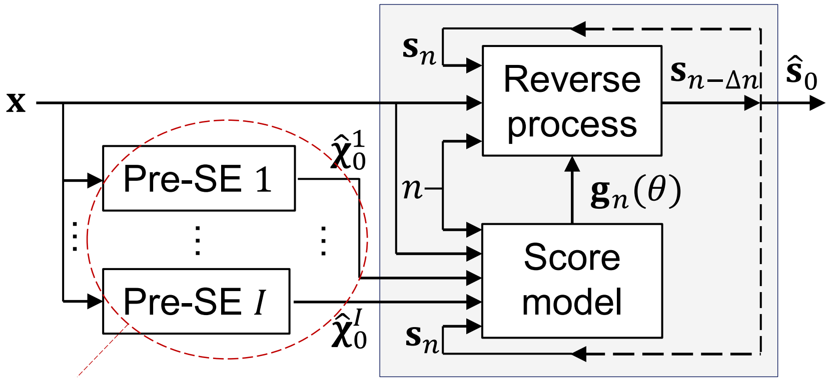
- **MS-SGMSE incorporates X as condition c for integration**

SGMSE and MS-SGMSE

SGMSE



MS-SGMSE



SE methods to be integrated

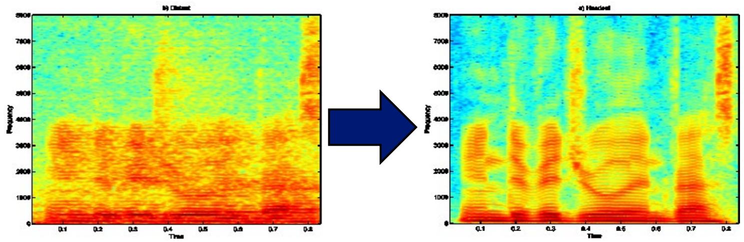
$\boldsymbol{X} = \{\hat{\chi}_0^1, \ldots, \hat{\chi}_0^I\}$: pre-enhanced signals

MS-SGMSE models $p(\mathbf{s}_0|\mathbf{x}, X)$ and improve the accuracy of SE

PRE-SE methods to be integrated

1. **Weighted Prediction Error: WPE**
2. **Complex Spectral Mapping: CSM**

   A deterministic prediction approach



Reverberant　　Predicted clean

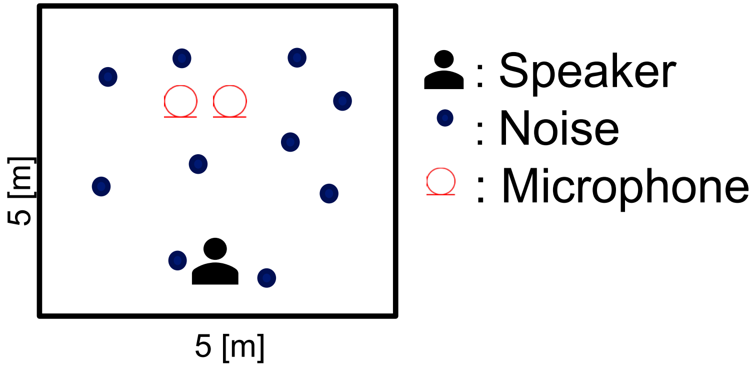
   Training objective:

$$L(\theta) = E[\|\mathrm{Re}(\mathbf{s} - \hat{\mathbf{s}}_0)\|_1 + \|\mathrm{Im}(\mathbf{s}_0 - \hat{\mathbf{s}}_0)\|_1] + \||\mathbf{s}_0| - |\hat{\mathbf{s}}_0|\|_1]$$

3. **Cascade configuration of the above two: WPE-CSM**

Experimental setting

- Experimental conditions
  - Training data: WSJ-CHiME3



# of speakers (WSJ0)	1
# of noises (CHiME3)	10
# of microphones	2
Speaker-mic. Distance [m]	0.5~1.5
Distance between microphones [m]	**0.02~0.14**
Reverberation time [s]	0.2~1.0
SNR [dB]	10~14

- Clean targets: Simulated using room impulse responses truncated at 2 ms.
  - Evaluation data
    - Matched condition: WSJ0-CHiME3 (the same as training data)
    - Mismatched condition: REVERB challenge

Experimental results

SE method	Input stream(s)	Simulated data			Real data
		SI-SDR*2) [dB]	PESQ *3)	ESTOI *4)	WER*5) [%]
Obs	–	-3.5	1.24	0.47	6.14
WPE	Obs	-0.8	1.32	0.55	4.97
CSM	Obs	7.3	2.58	0.86	4.30
WPE-CSM*1)	Obs	8.5	2.75	0.88	4.00
SGMSE	Obs	7.8	2.68	0.86	4.61
Multi-stream SGMSE	Obs, WPE	8.3	2.83	0.88	**3.46**
	Obs, CSM	8.5	2.67	0.87	4.30
	Obs, WPE-CSM	9.3	2.81	0.88	3.92
	Obs, WPE, CSM	9.4	2.84	**0.89**	3.81
	Obs, WPE, CSM, WPE-CSM	**9.8**	**2.85**	**0.89**	3.84

*1) Cascade of WPE and CSM, *2) Scale-Invariant Signal-to-Distortion Ratio,
*3) Perceptual Evaluation of Speech Quality, *4) Extended Short-Time Objective Intelligibility,
*5) Word Error Rate

Summary of diffusion model-based approach

Pros

- Highly accurate joint denoising and dereverberation
  - Direct signal can be recovered
- Further improvement with integration with other SE methods
  - Outperform not only blind inverse filtering approach, but also NN-based deterministic prediction approach

Cons

- Require prior training
  - Still sensitive to mismatch between training and test conditions

Outline of this talk

1. Approaches to dereverberation
2. Blind inverse filtering-based dereverberation
   - Theoretical background
   - Weighted Prediction Error (WPE) method
   - Extension to joint denoising, dereverberation, and source separation
3. Neural network (NN)-based dereverberation
   - Diffusion model-based joint denoising and dereverberation
   - Integration with WPE and other SE techniques
4. Future challenges and concluding remarks

Future challenges

Satisfactory speech quality is not yet achieved for real conversation recordings like CHiME-8 challenge

Challenges	Inverse filtering	NN-based approach
Distributed microphone array scenarios	**Under progress**	-
Mismatches between training and test conditions	-	**Under progress**
Unknown and varying number of speakers and ambient noises	Tighter integration with ***speaker diarization*** and ***audio event detection*** may be the key	
Moving speakers	Not yet well studied	

WPE-SD for spatially distributed microphones

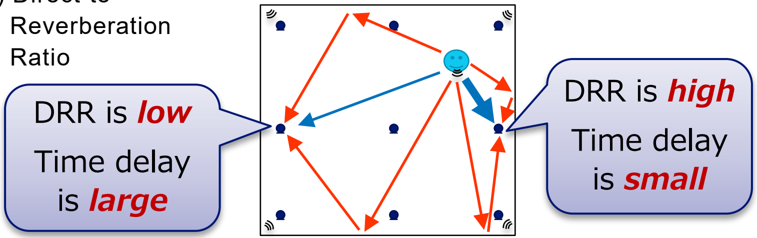
[Lohmann+, 2024]

Problems for distributed microphone scenario:

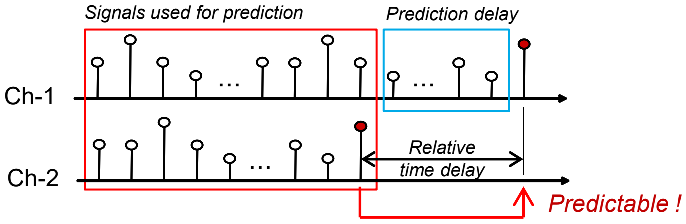
- DRR*1) largely differs depending on mic. locations
  - Performance depends largely on reference microphones

*1) Direct-to-Reverberation Ratio

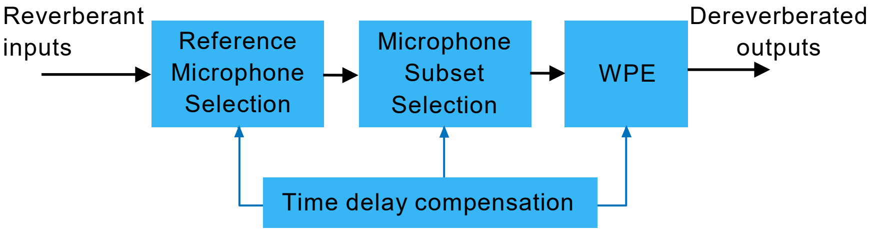
Distributed array

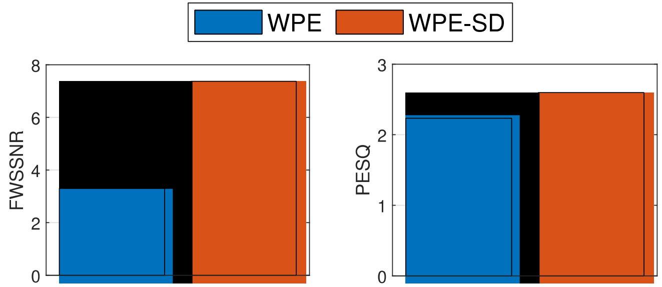


- Time delay between mics may exceed prediction delay
  - Direct signal can be predicted and distorted by WPE



WPE-SD (spatially distributed) [Lohmann+, 2024]





WPE-SD achieves ***large improvement***

⇨ **Future work: extension to more realistic scenarios with noisy reverberant mixtures**

Buddy: unsupervised dereverb with diffusion model (DM)

[Moliner+, 2024]

Jointly estimate clean speech and reverb

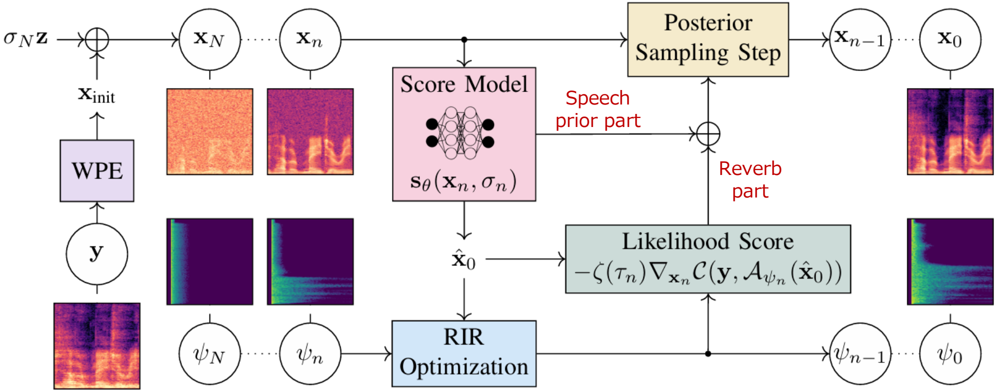
- Modeling clean speech prior $p(\mathbf{x}_0)$ using DM, and
- Reverb by exponential energy-decay model $\mathcal{A}_\psi(\mathbf{x}_0)$

Conditional score of DM

$$\begin{aligned}\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_\tau|\mathbf{y}) &= \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_\tau) + \nabla_{\mathbf{x}_t} \log p(\mathbf{y}|\mathbf{x}_\tau) \\ &\approx \underbrace{\mathbf{s}_\theta(x_t, \sigma_t)}_{\text{Speech prior part}} - \underbrace{\zeta(\tau)\nabla_{\mathbf{x}_t} \mathcal{C}(\mathbf{y}, \mathcal{A}_\psi(\hat{\mathbf{x}}_0))}_{\text{Reverb part}}\end{aligned}$$

Speech prior part (Environment independent)

Reverb part (Environment dependent)



	Matched			Mismatched		
	DNS-MOS	PESQ	ESTOI	DNS-MOS	PESQ	ESTOI
WPE	3.24	1.81	0.57	3.10	1.74	0.54
Buddy (w/ WPE)	**3.76**	**2.30**	**0.66**	**3.74**	**2.24**	**0.65**

Task: Single channel dereverberation (with no noise)

Future work: extension to more realistic scenarios with noisy reverberant mixtures

Concluding remarks

Dereverberation is now a solvable problem:

- Blind inverse filtering is applicable to unknown recording conditions
- NN can perform highly accurate dereverberation when training and test conditions well align

Future work:

- Enhancement of real conversation recordings is still challenging
  - Developing new techniques overcoming current limitations, and integrating various approaches could be the key to the solution

- [Flanagan, 1985] James L. Flanagan, James D Johnston, R Zahn, Gary W. Elko, Computer-steered microphonearrays for sound transduction in large rooms, The Journal of the Acoustical Society of America, 78 (11), 1508-1518, 1985.
- [Lebart+, 2001] K. Lebart, J. M. Boucher, and P. N. Denbigh, A new method based on spectral subtraction for speech dere-verberation, Acta Acustica united with Acustica, 87 (3), 359-366, 2001.
- [Habets+, 2004] Emanuël A.P. Habets, Single-channel speech dereverberation based on spectral subtraction, Proc. ProRISC, 25-26, 2004.
- [Habets+, 2007] Emanuël A.P. Habets, Sharon Gannot and Israel Cohen, Late reverberant spectral variance estimation based on a statistical model, IEEE Signal Processing Letters, 16 (9), 770-773, 2009.
- [Habets+, 2009] Emanuël A.P. Habets, Single-and Multi-Microphone Speech Dereverberation using Spectral Enhancement, Ph.D. Thesis, Technische Universiteit Eindhoven, 2007.
- [Lollman, 2009] Heiner Löllmann, Emre Yilmaz, Marco Jeub, Peter Vary, An improved algorithm for blind reverberation time estimation, Proc. IWAENC, 2010.
- [Weninger+, 2014] FJ Weninger, S Watanabe, J Le Roux, J Hershey, Y Tachioka, JT Geiger, BW Schuller, G Rigoll, The MERL/MELCO/TUM system for the REVERB challenge using deep recurrent neural network feature enhancement, Proc. REVERB Challenge Workshop, 2014.
- [Xu, 2015] Yong Xu, Jun Du, Li-Rong Dai, Chin-Hui Lee, A regression approach to speech enhancement based on deep neural networks, IEEE/ACM Trans. ASLP, 23 (1), 7 – 19, 2015

- [Ronneberger+, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, U-net: Con-volutional networks for biomedical image segmentation, Proc. MICCAI, 234–241, 2015.
- [Wang, 2021] Zhong-Qiu Wang, Peidong Wang, and DeLiang Wang, Multi-microphone complex spectral mapping for utterance-wise and continuousspeech separation, IEEE/ACM Trans. ASLP, 29, 2001-2014, 2021.
- [Serra+, 2022] Joan Serra, Santiago Pascual, Jordi Pons, R. Oguz Araz, and Davide Scaini, Universal speech enhancement with score-based diffusion, arXiv:2206.03065v2, 2022.
- [Richter+, 2023] Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann, “Speech enhancement and dereverberation with diffusion-based generative models, IEEE/ACM Trans. ASLP, 31, 2351–2364, 2023.
- [Misyohi+, 1988] Masato Miyoshi, Yutaka Kaneda, Inverse filtering of room acoustics. IEEE Trans. ASSP, 36(2), 145-152, 1988.
- [Nakatani+, 2008] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, Biing-Hwang Juang, Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation, Proc. IEEE ICASSP, 2008.
- [Gannot, 2010] Sharon Gannot, Multi-microphone Speech Dereverberation Using Eigen-decomposition, In Speech Dereverberation (Naylor P., Gaubitch N. eds), Springer, 2010.
- [Yohena+, 2024] Fumiki Yohena, Kohei Yatabe, SINGLE-CHANNEL BLIND DEREVERBERATION BASED ON RANK-1 MATRIX LIFTING IN TIME-FREQUENCY DOMAIN, Proc. IEEE ICASSP, 2024.

- [Hikichi+, 2007] Takafumi Hikichi, Marc Delcroix, Masato Miyoshi, Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations, EURASIP Journal on Advances in Signal Processing 2007.
- [Kodrasi+, 2013] Ina Kodrasi, Stefan Goetze, and Simon Doclo. A perceptually constrained channel shortening technique for speech dereverberation, Proc. IEEE ICASSP, 2013.
- [Abed-Meraim+, 1997] Karim Abed-Meraim and Eric Moulines and Philippe Loubaton, Prediction error method for second-order blind identification, IEEE Trans. SP, 45 (3), 694-705, 1997.
- [Kinoshita+, 2009] Keisuke Kinoshita, Marc Delcroix, Tomohiro Nakatani, Masato Miyoshi, Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction, IEEE trans. ASLP, 17 (4), 534-545, 2009.
- [Nakatani+, 2010] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, Biing Hwang Juang, Speech dereverberation based on variance-normalized delayed linear prediction, IEEE Trans. ASLP, 18 (7), 1717-1731, 2010.
- [Yoshioka+,2012] Takuya Yoshioka, Tomohiro Nakatani, Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening, IEEE Trans. ASLP, 20 (10), 2707-2720, 2012.
- [Kinoshita+, 2016] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël A P. Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, Armin Sehr, Takuya Yoshioka, A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research, EURASIP Journal on Advances in Signal Processing, 2016.

- [Delcroix+, 2015] Marc Delcroix, Takuya Yoshioka, Atsunori Ogawa, Yotaro Kubo, Masakiyo Fujimoto, Nobutaka Ito, Keisuke Kinoshita, Miquel Espi, Shoko Araki, Takaaki Hori, Tomohiro Nakatani , Strategies for distant speech recognitionin reverberant environments, EURASIP Journal on Advances in Signal Processing, 2015.
- [Cornell+, 2024] Samuele Cornell, Taejin Park, He Huang, Christoph Boeddeker, Xuankai Chang, Matthew Maciejewski, Matthew Wiesner, Paola Garcia, Shinji Watanabe, The CHiME-8 DASRChallenge for Generalizable and Array Agnostic Distant Automatic Speech Recognition and Diarization, Proc. CHiME-8, 2024.
- [Kamo+, 2024] Naoyuki Kamo, Naohiro Tawara, Atsushi Ando, Takatomo Kano, Hiroshi Sato, Rintaro Ikeshita, Takafumi Moriya, Shota Horiguchi, Kohei Matsuura, Atsunori Ogawa, Alexis Plaquet, Takanori Ashihara, Tsubasa Ochiai, Masato Mimura, Marc Delcroix, Tomohiro Nakatani, Taichi Asami, Shoko Araki, NTTMulti-Speaker ASR System for the DASR Task of CHiME-8 Challenge, Proc. CHiME-8, 2024.
- [Jukic+, 2015] Ante Jukić, Toon van Waterschoot, Timo Gerkmann, Simon Doclo, Multi-channel linear prediction-based speech dereverberation with sparse priors, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23 (9), 1509-1520, 2015.
- [Chetupalli+, 2019] Srikanth Raj Chetupalli and Thippur V. Sreenivas, Late Reverberation Cancellation Using Bayesian Estimation of Multi-Channel Linear Predictors and Student's t-Source Prior, IEEE/ACM Trans. ASLP, 27 (6), 2019.
- [Yoshioka+, 2009] Takuya Yoshioka, Hideyuki Tachibana, Tomohiro Nakatani, and Masato Miyoshi, Adaptive dereverberation of speech signals with speaker-position change detection, Proc IEEE ICASSP, 2009.

- [Caroselli+, 2017] Joe Caroselli, Izhak Shafran, Arun Narayanan, Richard Rose, Adaptive Multichannel Dereverberation for Automatic Speech Recognition, Proc. Interspeech, 2017.
- [Togami+, 2013] Masahito Togami and Yohei Kawaguchi, Noise robust speech dereverberation with Kalman smoother, Prog. IEEE ICASSP, 2013.
- [Braun+, 2018] Sebastian Braun and Emanuël AP Habets, Linear prediction-based online dereverberation and noise reduction using alternating Kalman filters, IEEE/ACM Trans. ASLP, 26 (6), 1119-1129, 2018.
- [Dietzen+, 2018] Thomas Dietzen, Simon Doclo, Marc Moonen, Toon Van Waterschoot, Joint multi-microphone speech dereverberation and noise reduction using integrated sidelobe cancellation and linear prediction, Proc. IWAENC, 2018.
- [Ikeshita+, 2021-1] Rintaro Ikeshita, Naoyuki Kamo, and Tomohiro Nakatani, Blind signal dereverberation based on mixture of weighted prediction error models, IEEE Signal Processing Letters, 28, 399-403, 2021.
- [Ikeshita+, 2021-2] Rintaro Ikeshita, Naoyuki Kamo, and Tomohiro Nakatani, Online speech dereverberation using mixture of multichannel linear prediction models, IEEE Signal Processing Letters, 28, 1580-1584, 2021.
- [Nakatani+, 2019] Tomohiro Nakatani and Keisuke Kinoshita, Maximum likelihood convolutional beamformer for simultaneous denoising and dereverberation, Proc. EUSIPCO, 2019.
- [Yoshioka+, 2010] Takuya Yoshioka, Tomohiro Nakatani, Masato Miyoshi, Hiroshi G Okuno, Blind separation and dereverberation of speech mixtures by joint optimization, IEEE Trans. ASLP, 19 (1), 69-84, 2019.

- [Ikeshita+, 2021] Rintaro Ikeshita and Tomohiro Nakatani, Independent vector extraction for fast joint blind source separation and dereverberation, IEEE Signal Processing Letters, 28, 972-976, 2021.
- [Nakatani+, 2021] Tomohiro Nakatani, Rintaro Ikeshita, Keisuke Kinoshita, Hiroshi Sawada, Shoko Araki, Blind and neural network-guided convolutional beamformer for joint denoising, dereverberation, and source separation,Proc. IEEE ICASSP, 6129-6133, 2021.
- [Nakatani+, 2022] Tomohiro Nakatani, Rintaro Ikeshita, Keisuke Kinoshita, Hiroshi Sawada,Naoyuki Kamo, Shoko Araki, Switching independent vector analysis and its extension to blind and spatially guided convolutional beamforming algorithms, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 1032-1047, 2022.
- [Nakatani+, 2024] Tomohiro Nakatani, Naoyuki Kamo, Marc Delcroix, Shoko Araki, Multi-stream diffusion model for probabilistic integration of model-based and data-driven speech enhancement, Proc. IWAENC, 2024.
- [Song+, 2021] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole, Score-based generative modeling through stochastic differential equations, Proc. ICLR, 2021.
- [Lohmann+, 2024] Anselm Lohmann, Toon van Waterschoot, Joerg Bitzer, Simon Doclo, Dereverberation in Acoustic Sensor Networks Using Weighted Prediction Error With Microphone-dependent Prediction Delays, Proc. SOUNDS Workshop (satellite of IWENC-2024), 2024.
- [Moliner+, 2024] Eloi Moliner, Aalto University, Finland; Jean-Marie Lemercier, Simon Welker, Timo Gerkmann, Universität Hamburg, Germany; Vesa Välimäki, BUDDY: SINGLE-CHANNEL BLIND UNSUPERVISED DEREVERBERATION WITH DIFFUSION MODELS, Proc. IWAENC, 2024.