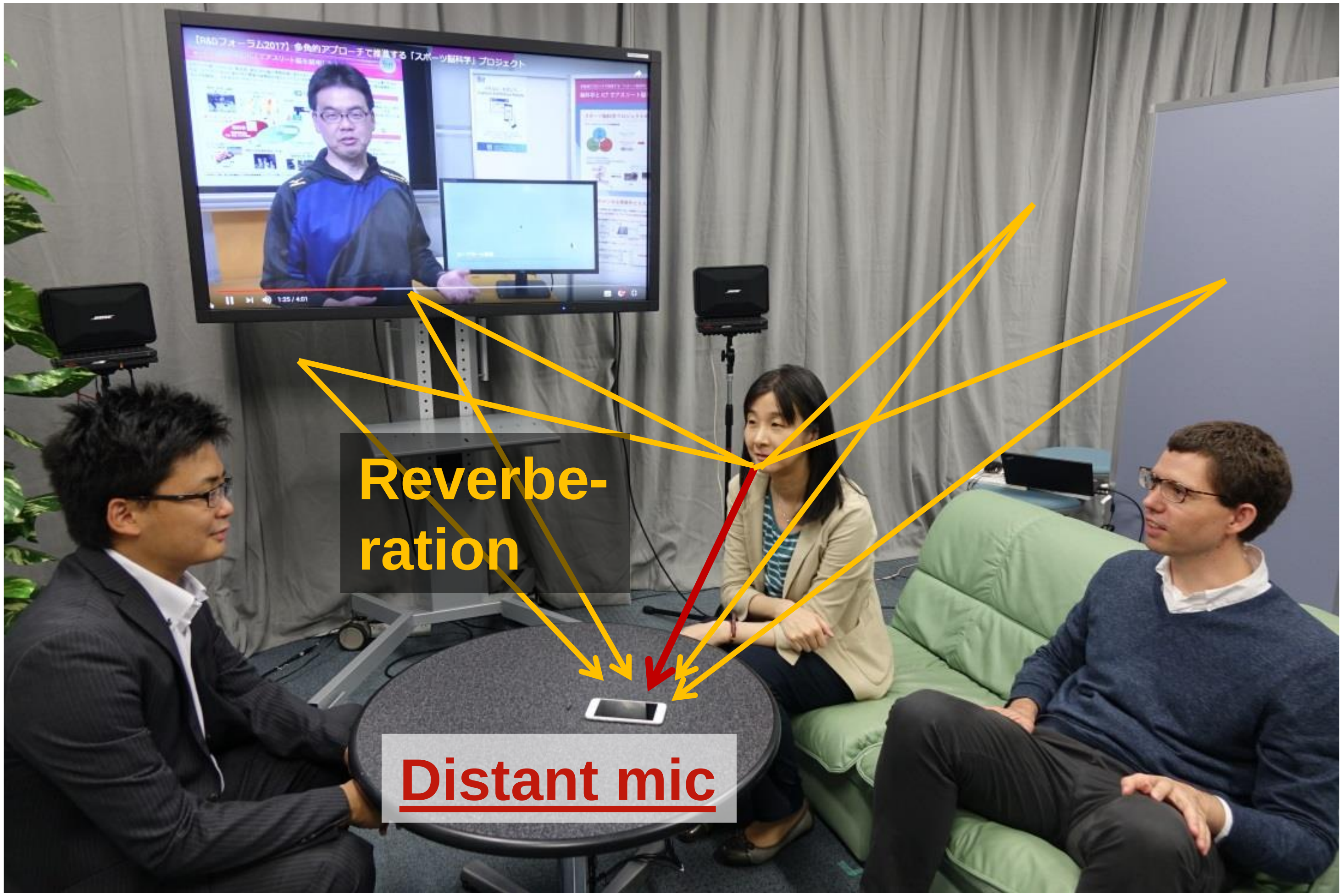
Part III.
Dereverberation

Tomohiro Nakatani

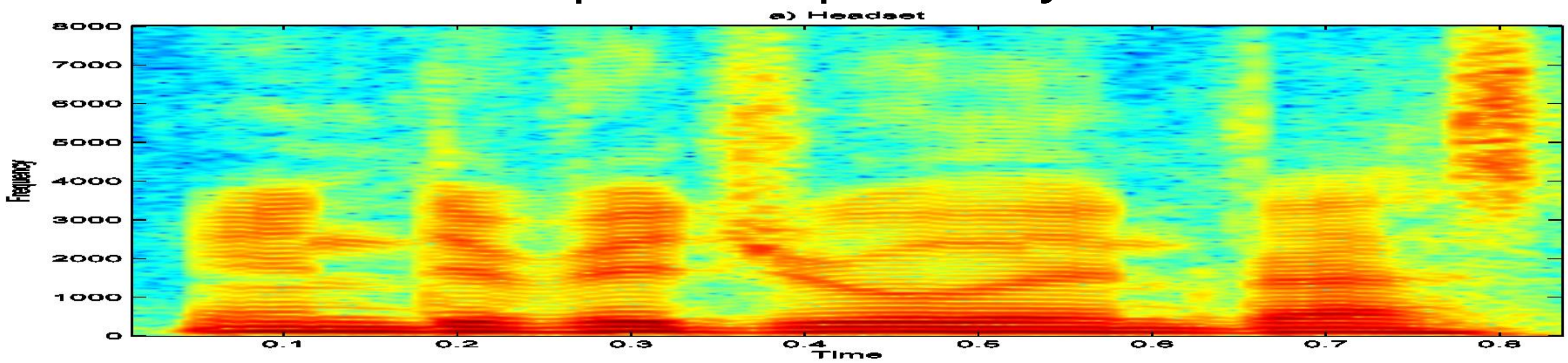
Speech recording in reverberant environments



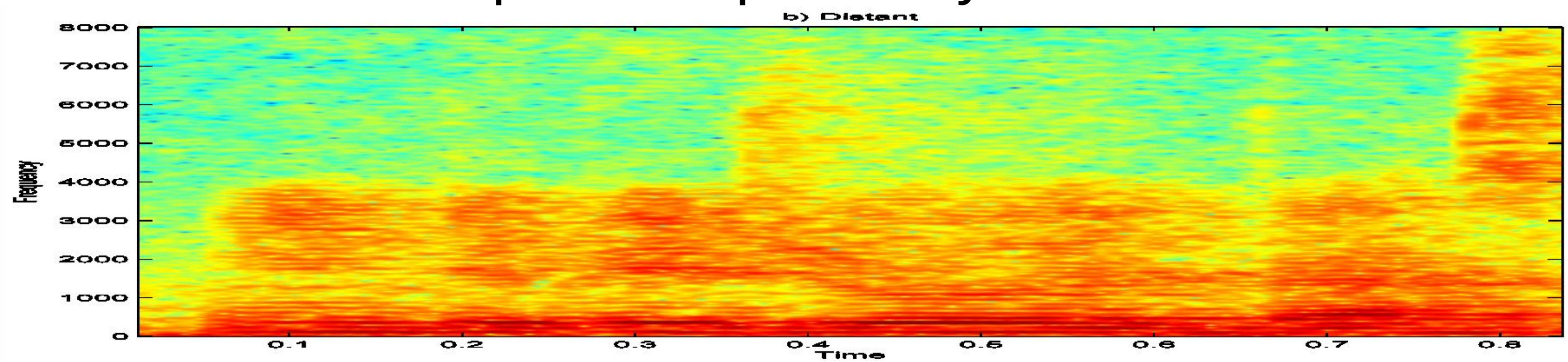
Dereverberation is needed to enhance the quality of recorded speech by reducing reverberation included in it

Effect of reverberation

Non-reverberant speech captured by a headset



Reverberant speech captured by a distant mic



Speech becomes less intelligible and ASR becomes very hard

Table of contents in part III

- Goal of dereverberation
- Approaches to dereverberation
  - Signal processing based approaches
  - A DNN-based approach
- Integration of signal processing and DNN approaches
  - DNN-WPE

Goal of dereverberation: time domain

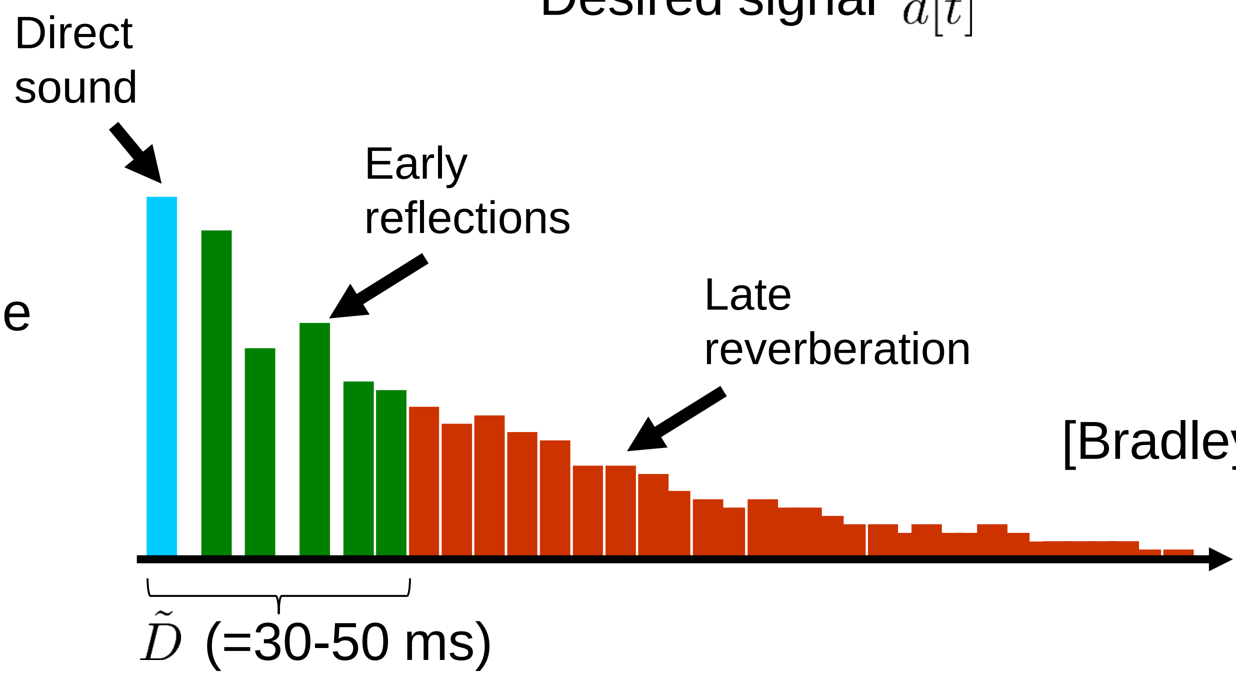
Preserve | Reduce

Reverberant speech

$$x[\tilde{t}] = \sum_{\tilde{\tau}=0}^{\tilde{L}-1} a[\tilde{\tau}]s[\tilde{t}-\tilde{\tau}] = \boxed{\sum_{\tilde{\tau}=0}^{\tilde{D}-1} a[\tilde{\tau}]s[\tilde{t}-\tilde{\tau}]} + \boxed{\sum_{\tilde{\tau}=\tilde{D}}^{\tilde{L}-1} a[\tilde{\tau}]s[\tilde{t}-\tilde{\tau}]}$$

$$\underbrace{\text{Direct sound} + \text{Early reflections}}_{\text{Desired signal } d[\tilde{t}]} \qquad \underbrace{\text{Late reverberation}}_{r[\tilde{t}]}$$

Impulse response $a[\tilde{\tau}]$



[Bradley et al., 2003]

Model of reverberation: STFT domain

- Time domain convolution is approximated by frequency domain convolution at each frequency [Nakatani et al. 2008]
  - If frame shift << analysis window (e.g., frame shift <= analysis window/4)

Desired signal / Late reverberation

STFT domain (1-ch)

$$x_{t,f} = \sum_{\tau=0}^{L-1} a_{\tau,f} s_{t-\tau,f} = \boxed{\sum_{\tau=0}^{D-1} a_{\tau,f} s_{t-\tau,f}} + \boxed{\sum_{\tau=D}^{L-1} a_{\tau,f} s_{t-\tau,f}}$$

STFT domain (multi-ch)

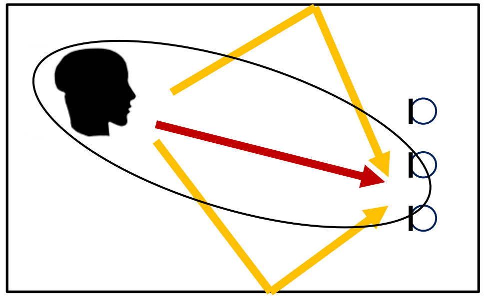
$$\mathbf{x}_{t,f} = \sum_{\tau=0}^{L-1} \mathbf{a}_{\tau,f} s_{t-\tau,f} = \underbrace{\boxed{\sum_{\tau=0}^{D-1} \mathbf{a}_{\tau,f} s_{t-\tau,f}}}_{\mathbf{d}_{t,f}} + \underbrace{\boxed{\sum_{\tau=D}^{L-1} \mathbf{a}_{\tau,f} s_{t-\tau,f}}}_{\mathbf{r}_{t,f}}$$

Convolutional transfer function:

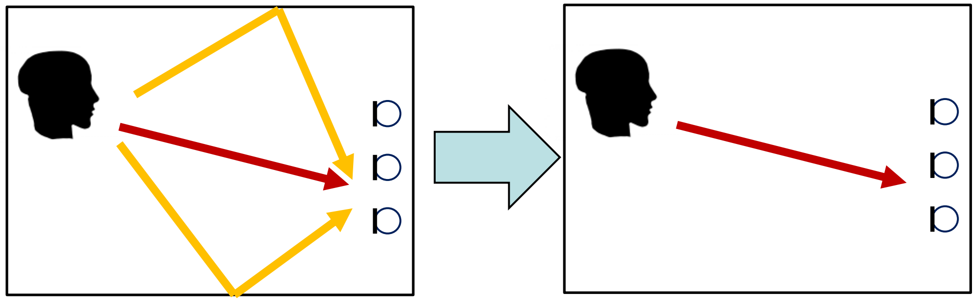
$$\mathbf{a}_{\tau,f} = (a_{1,\tau,f}, a_{2,\tau,f}, \ldots, a_{M,\tau,f})^\top \text{ for } \tau = 0, \ldots, L-1$$

Approaches to dereverberation

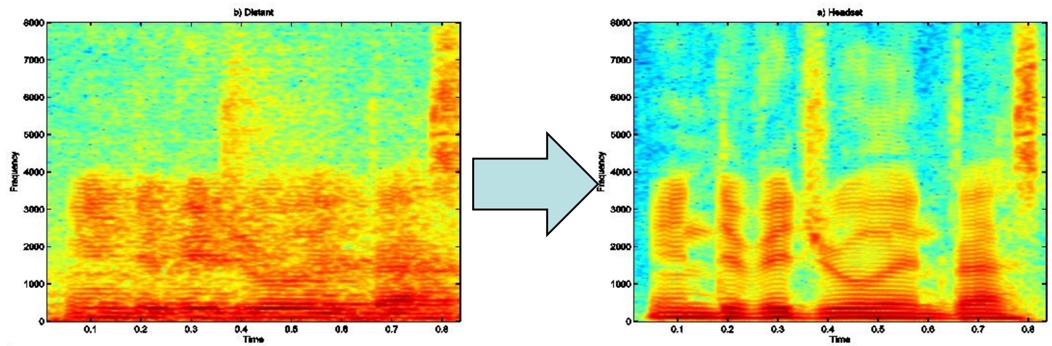
- Beamforming (multi-ch)
  - Enhance desired signal from speaker direction
  - Mostly the same as denoising



- Blind inverse filtering (multi-ch)
  - Cancel late reverberation
  - Multi-channel linear prediction
    - Weighted prediction error (WPE) method



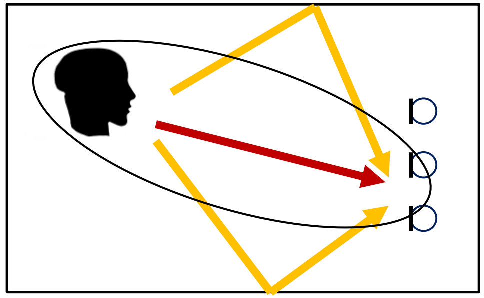
- DNN-based spectral enhancement (1ch)
  - Estimate clean spectrogram
  - Mostly the same as denoising autoencoder



Reverberant   Estimated clean

Approaches to dereverberation

- Beamforming (multi-ch)
  - Enhance desired signal from speaker direction
  - Mostly the same as denoising



- Blind inverse filtering (multi-ch)
  - Cancel late reverberation
  - Multi-channel linear prediction
    - Weighted prediction error (WPE) method

- DNN-based spectral enhancement (1ch)
  - Estimate clean spectrogram
  - Mostly the same as denoising autoencoder

Reverberant

Estimated clean

Dereverberation based on beamforming

- Time domain model of desired signal

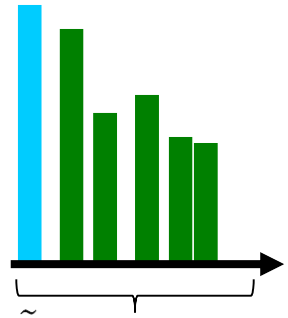
Time domain $$\mathbf{d}[\tilde{t}] = \sum_{\tilde{\tau}=0}^{\tilde{D}} \mathbf{a}[\tilde{\tau}] s[\tilde{t} - \tilde{\tau}]$$

- Assume $\tilde{D}$ << STFT window, then

STFT domain $$\mathbf{d}_{t,f} = \mathbf{a}_f s_{t,f}$$

$$\mathbf{x}_{t,f} = \mathbf{a}_f s_{t,f} + \mathbf{r}_{t,f}$$

Initial part of the impulse response



$\tilde{D}$ (=30-50 ms)

STFT

$\mathbf{a}_f$: acoustic transfer function

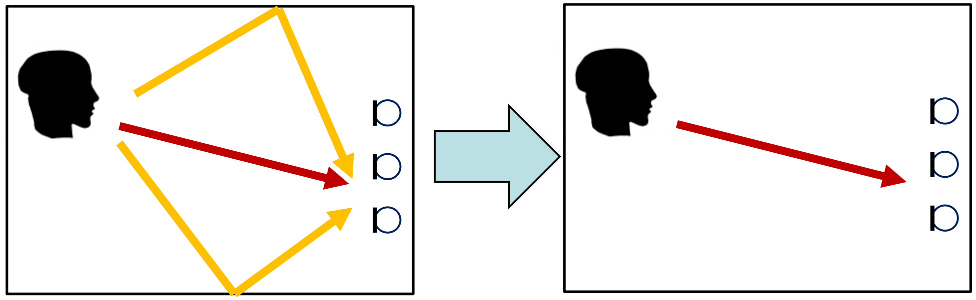
→ Beamforming is applicable to reduce $\mathbf{r}_{t,f}$

- Techniques for estimating spatial covariances, $\Psi_{\mathbf{dd},f}$ and $\Psi_{\mathbf{rr},f}$
  - Maximum-likelihood estimator [Schwartz et al., 2016]
  - Eigen-value decomposition based estimator [Heymann, 2017b, Kodrasi and Doclo, 2017, Nakatani et al., 2019a]

Approaches to dereverberation

- Beamforming (multi-ch)
  - Enhance desired signal from speaker direction
  - Mostly the same as denoising

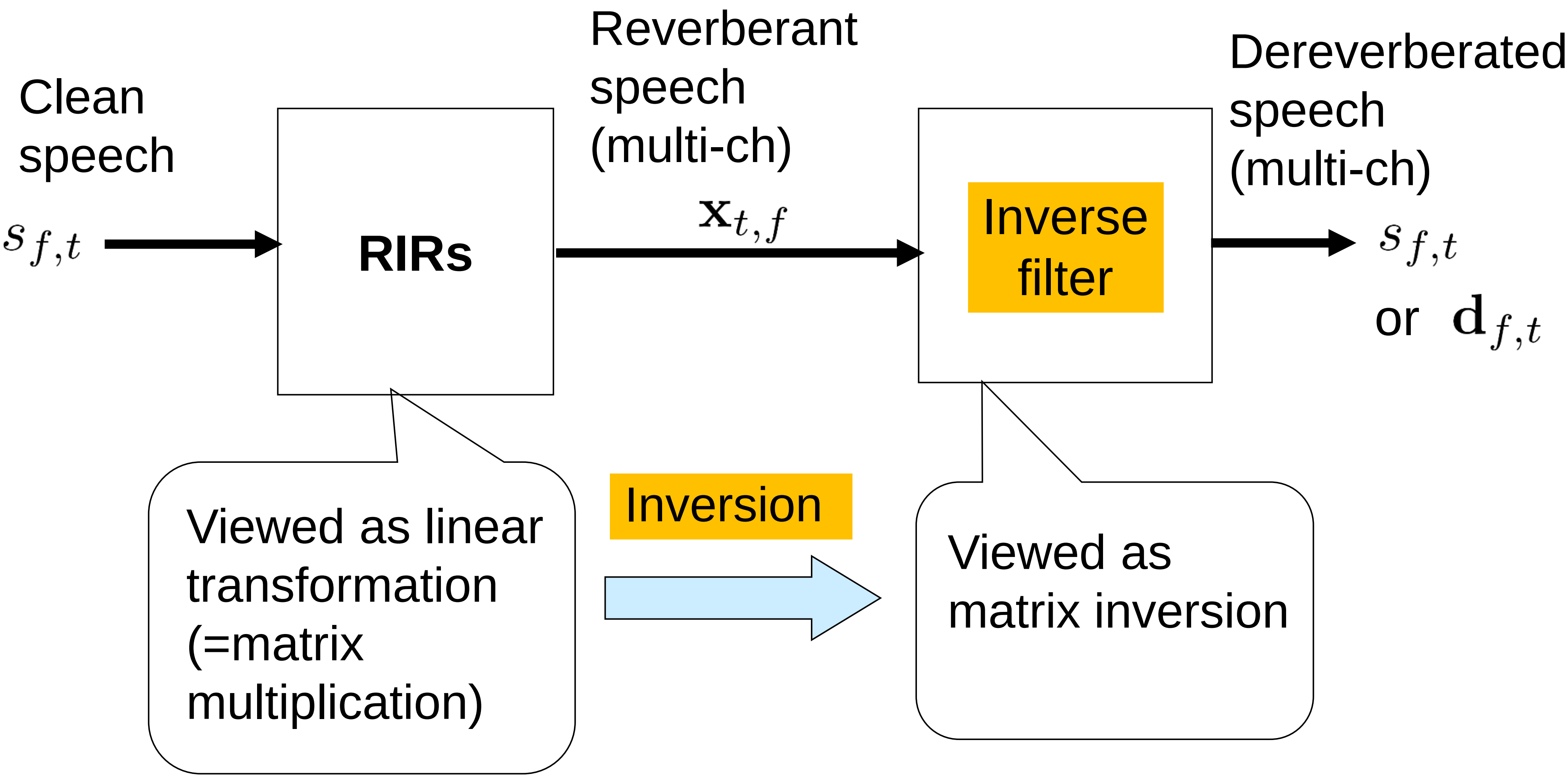
- Blind inverse filtering (multi-ch)
  - Cancel late reverberation
  - Multi-channel linear prediction
    - Weighted prediction error (WPE) method



- DNN-based spectral enhancement (1ch)
  - Estimate clean spectrogram
  - Mostly the same as denoising autoencoder

Reverberant　　Estimated clean

What is inverse filtering



Represent RIR convolution by matrix multiplication

1-ch representation

$$\underbrace{\begin{pmatrix} x_{m,t,f} \\ x_{m,t-1,f} \\ \vdots \\ x_{m,t-K,f} \end{pmatrix}}_{\bar{\mathbf{x}}_{m,t,f} \in \mathbb{C}^{K}} = \underbrace{\begin{pmatrix} a_{m,0,f} & a_{m,1,f} & \cdots & a_{m,L-1,f} & 0 & \cdots & 0 \\ 0 & a_{m,0,f} & a_{m,1,f} & \cdots & a_{m,L-1,f} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \ddots & 0 \\ 0 & \cdots & 0 & a_{m,0,f} & a_{m,1,f} & \cdots & a_{m,L-1,f} \end{pmatrix}}_{\mathbf{H}_{m,f} \in \mathbb{C}^{K \times K_0}} \underbrace{\begin{pmatrix} s_{t,f} \\ s_{t-1,f} \\ \vdots \\ s_{t-K_0,f} \end{pmatrix}}_{\bar{\mathbf{s}}_{t,f} \in \mathbb{C}^{k_0}}$$

$$\boxed{\bar{\mathbf{x}}_{m,t,f} = \mathbf{H}_{m,f} \bar{\mathbf{s}}_{m,t,f}}$$

$$K_0 = L + K - 1$$

Multi-ch representation

$$\underbrace{\begin{pmatrix} \bar{\mathbf{x}}_{1,t,f} \\ \vdots \\ \bar{\mathbf{x}}_{M,t,f} \end{pmatrix}}_{\bar{\mathbf{x}}_{t,f} \in \mathbb{C}^{KM}} = \underbrace{\begin{pmatrix} \mathbf{H}_{1,f} \\ \vdots \\ \mathbf{H}_{M,f} \end{pmatrix}}_{\mathbf{H}_{f} \in \mathbb{C}^{KM \times K_0}} \bar{\mathbf{s}}_{t,f}$$

$$\boxed{\bar{\mathbf{x}}_{t,f} = \mathbf{H}_{f} \bar{\mathbf{s}}_{t,f}}$$

Existence of inverse filter [Miyoshi and Kaneda, 1988]

- Given $\mathbf{H}_f$, the inverse filter $\bar{\mathbf{W}}_f$ should satisfy

$$\bar{\mathbf{W}}_f^{\mathsf{H}} \mathbf{H}_f = \mathbf{I} \qquad \mathbf{I} \text{ : identity matrix}$$

- Solution exists and is obtained as:

$$\bar{\mathbf{W}}_f^{\mathsf{H}} = (\mathbf{H}_f^{\mathsf{H}} \mathbf{H}_f)^{-1} \mathbf{H}_f^{\mathsf{H}}$$

  – When $\mathbf{H}_f$ is full column rank (roughly #mics>1)

How can we estimate $\bar{\mathbf{W}}_f$ without knowing $\mathbf{H}_f$?

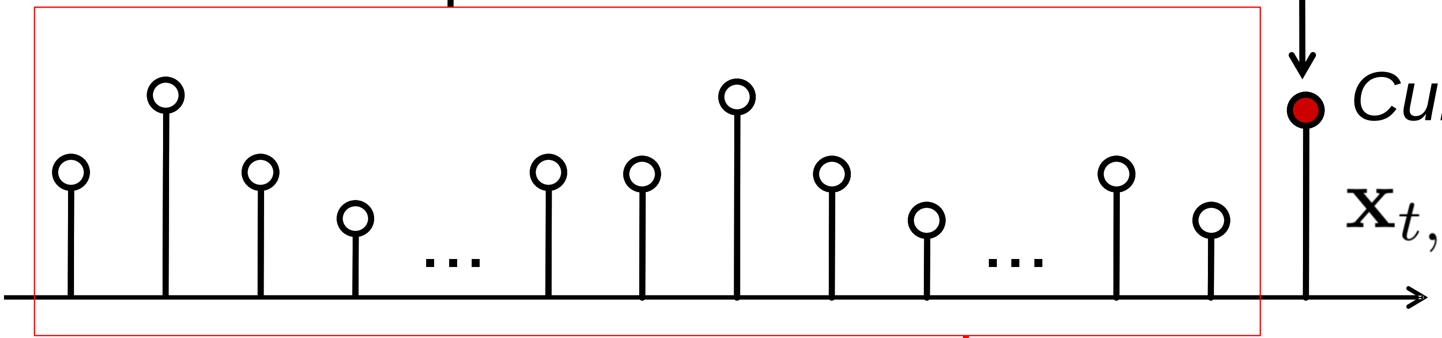
Approaches to blind inverse filtering

- Blind RIR estimation + robust inverse filtering
  - Blind RIR estimation is still an open issue
    - Eigen-decomposition [Gannot, 2010]
    - ML estimation approaches [Juang and Nakatani, 2007, Schmid et al., 2012]
  - Robust inverse filtering
    - Regularization [Hikichi et al., 2007]
    - Partial multichannel equalization [Kodrasi et al., 2013]
- Blind and direct estimation of inverse filter
  - Multichannel linear prediction (LP) based methods
    - Prediction Error (PE) method [Abed-Meraim et al., 1997]
    - Delayed Linear Prediction [Kinoshita et al., 2009]
    - Weighted Prediction Error (WPE) method [Nakatani et al., 2010]
    - Multi-input multi-output (MIMO) WPE method [Yoshioka and Nakatani, 2012]
  - Higher-order decorrelation approaches
    - Kurtosis maximization [Gillespie et al., 2001]

Multichannel LP [Abed-meraim et al, 1997]

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

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



Current signal

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

Past signal (multi-ch)

Predictable

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

Subtract predictable components from observation

Definition of multichannel LP

- Multichannel autoregressive model

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

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

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

$$\{\hat{\mathbf{W}}_{\tau,f}\} = \underset{\{\mathbf{W}_{\tau,f}\}}{\operatorname{argmin}} \sum_{t} \|\mathbf{x}_{t,f} - \sum_{\tau=1}^{L} \mathbf{W}_{\tau,f}^{\mathsf{H}} \mathbf{x}_{t-\tau,f}\|_2^2$$

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

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

Problems in conventional LP

- Speech is not stationary white noise
  - LP assumes the target signal d to be temporally uncorrelated
  - Speech signal exhibits short-term correlation (30-50 ms)

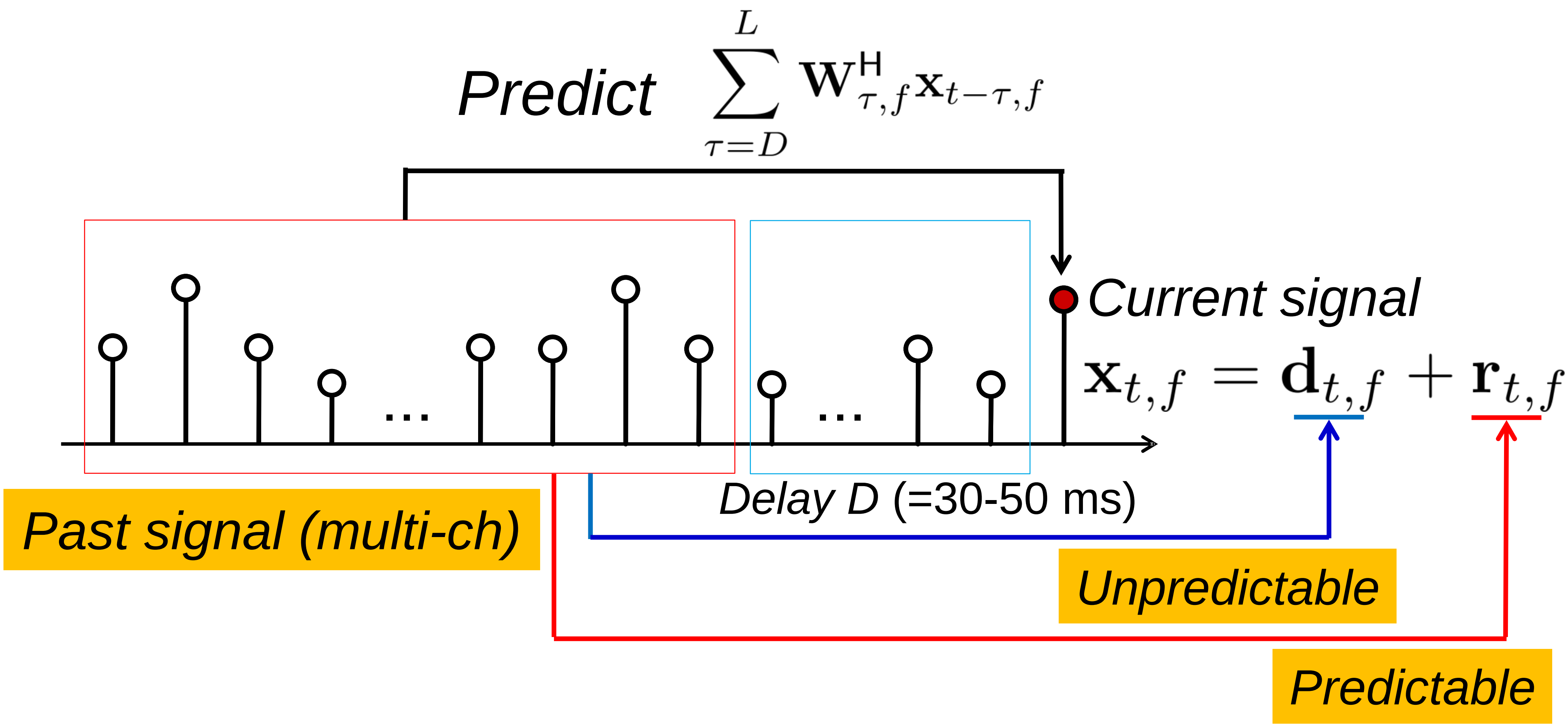
  ⇒ LP distorts the short-time correlation of speech

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

  ⇒ LP destroys the time structure of speech

- Solutions:
  - Use of a prediction delay [Kinoshita et al., 2009]
  - Use of a better speech model [Nakatani et al, 2010]

Delayed LP (DLP) [Kinoshita et al., 2009]



Delayed LP can only predict $\mathbf{r}_{t,f}$ from past signal

➡ Only reduce $\mathbf{r}_{t,f}$

Introduction of better source model
[Nakatani et al., 2010, Yoshioka et al., 2011]

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

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

- ML estimation for time-varying Gaussian source

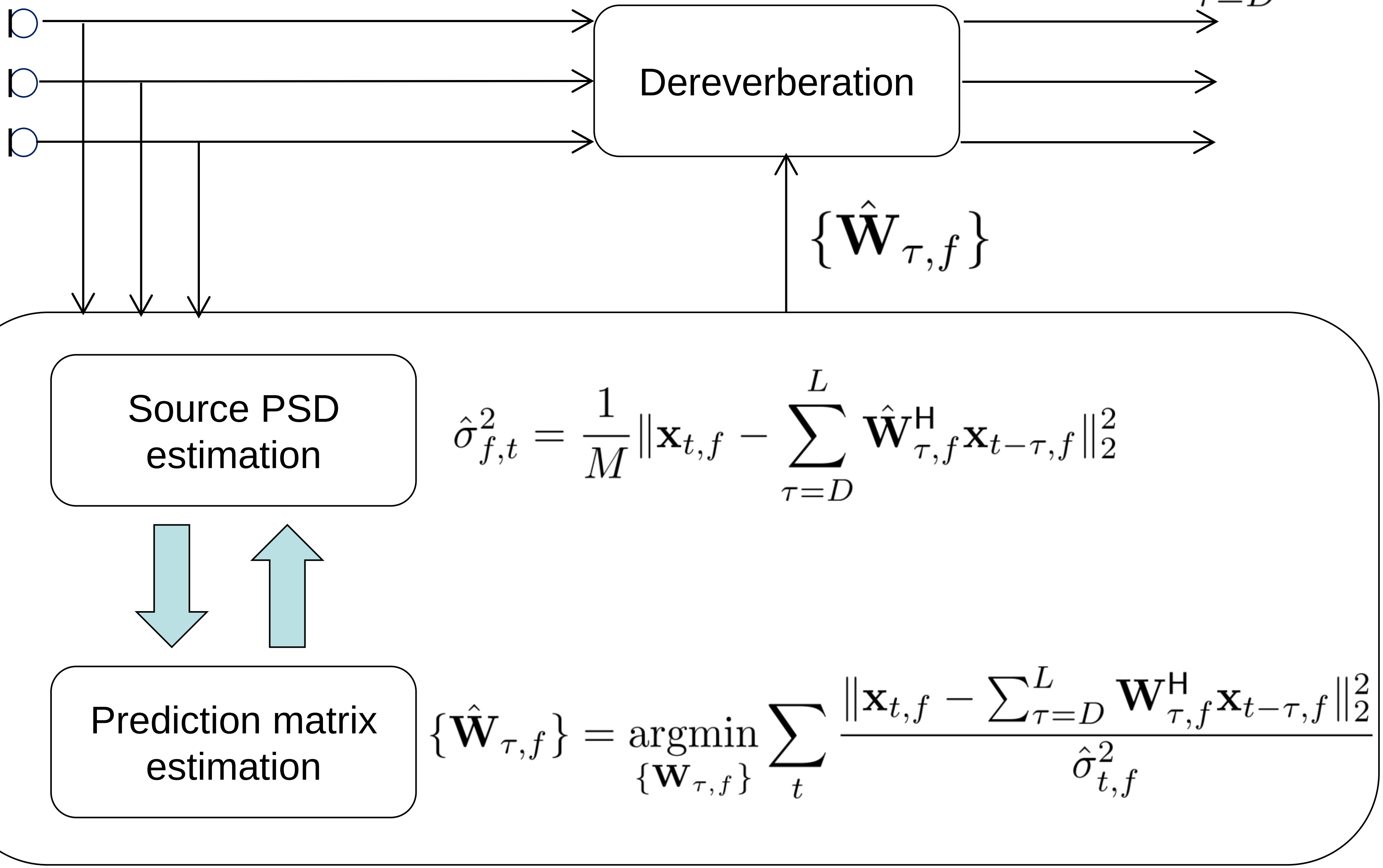
$$\{\hat{\mathbf{W}}_{\tau,f},\hat{\sigma}^2_{t,f}\} = \underset{\{\mathbf{W}_{\tau,f},\sigma^2_{t,f}\}}{\operatorname{argmax}} \prod_t \frac{1}{\pi\sigma^2_{t,f}} \exp\left(\underline{\frac{-\|\mathbf{x}_{t,f} - \sum_{\tau=D}^{L} \mathbf{W}^{\mathsf{H}}_{\tau,f}\mathbf{x}_{t-\tau,f}\|_2^2}{\sigma^2_{t,f}}}\right)$$

Minimization of weighted prediction error **(WPE)**

→ Blind inverse filtering can be achieved based only on a few seconds of observation

Processing flow of WPE

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



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

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

$$\{\hat{\mathbf{W}}_{\tau,f}\} = \underset{\{\mathbf{W}_{\tau,f}\}}{\operatorname{argmin}} \sum_t \frac{\|\mathbf{x}_{t,f} - \sum_{\tau=D}^{L} \mathbf{W}^{\mathsf{H}}_{\tau,f} \mathbf{x}_{t-\tau,f}\|_2^2}{\hat{\sigma}^2_{t,f}}$$

Why WPE achieves inverse filtering?

$$\sum_t \frac{\|\mathbf{x}_{t,f} - \sum_{\tau=D}^{L} \mathbf{W}_{\tau,f}^{\mathsf{H}} \mathbf{x}_{t-\tau,f}\|_2^2}{\sigma_{t,f}^2}$$

$$= \sum_t \frac{\|\mathbf{d}_{t,f} + \mathbf{r}_{t,f} - \sum_{\tau=D}^{L} \mathbf{W}_{\tau,f}^{\mathsf{H}} \mathbf{x}_{t-\tau,f}\|_2^2}{\sigma_t^2}$$

$$= \sum_t \frac{\|\mathbf{d}_{t,f}\|_2^2}{\sigma_{t,f}^2} + \frac{\sum_t \|\mathbf{r}_{t,f} - \sum_{\tau=D}^{L} \mathbf{W}_{\tau,f}^{\mathsf{H}} \mathbf{x}_{t-\tau,f}\|_2^2}{\sigma_{t,f}^2}$$

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

Assumption

$\mathbf{d}_{t,f}$ is not correlated with $\mathbf{r}_{t,f}$ and with $\sum_{\tau=D}^{L} \mathbf{W}_{\tau,f}^{\mathsf{H}} \mathbf{x}_{t-\tau,f}$

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

Existence of $\mathbf{W}_{\tau,f}$ is guaranteed when the inverse filter exists

Extensions

- Elaboration of probabilistic models
  - Sparse prior for speech PSD [Jukic et al., 2015]
  - Bayesian estimation with student-T speech prior [Chetupalli and Sreenivas, 2019]

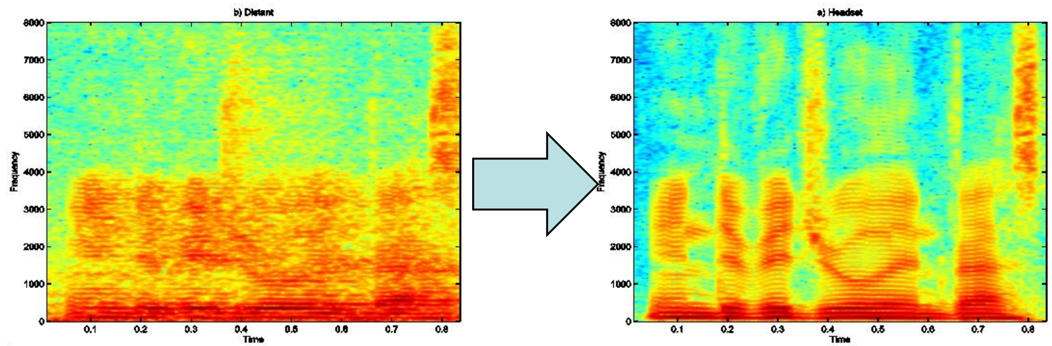
- Frame-by-frame online estimation
  - Recursive least square [Yoshioka et al., 2009], [Caroselli et al., 2017]
  - Kalman filter for joint denoising and dereverberation [Togami and Kawaguchi, 2013], [Braun and Habets, 2018], [Dietzen et al., 2018]

Approaches to dereverberation

- Beamforming (multi-ch)
  - Enhance desired signal while reducing late reverberation
  - Mostly the same as denoising

- Blind inverse filtering (multi-ch)
  - Cancel late reverberation
  - (Multi-channel) lnear prediction
    - Weighted prediction error method

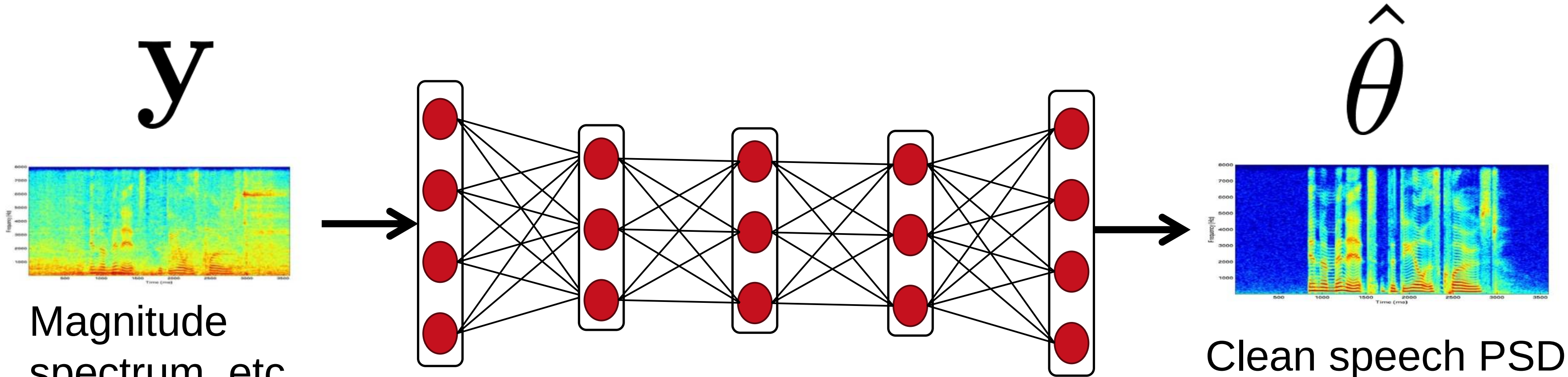
- DNN-based spectral enhancement (1ch)
  - Estimate clean spectrogram
  - Mostly the same as denoising autoencoder



Reverberant   Estimated clean

Neural networks based dereverberation

- Train neural networks based on huge amount of parallel data



Magnitude spectrum, etc.

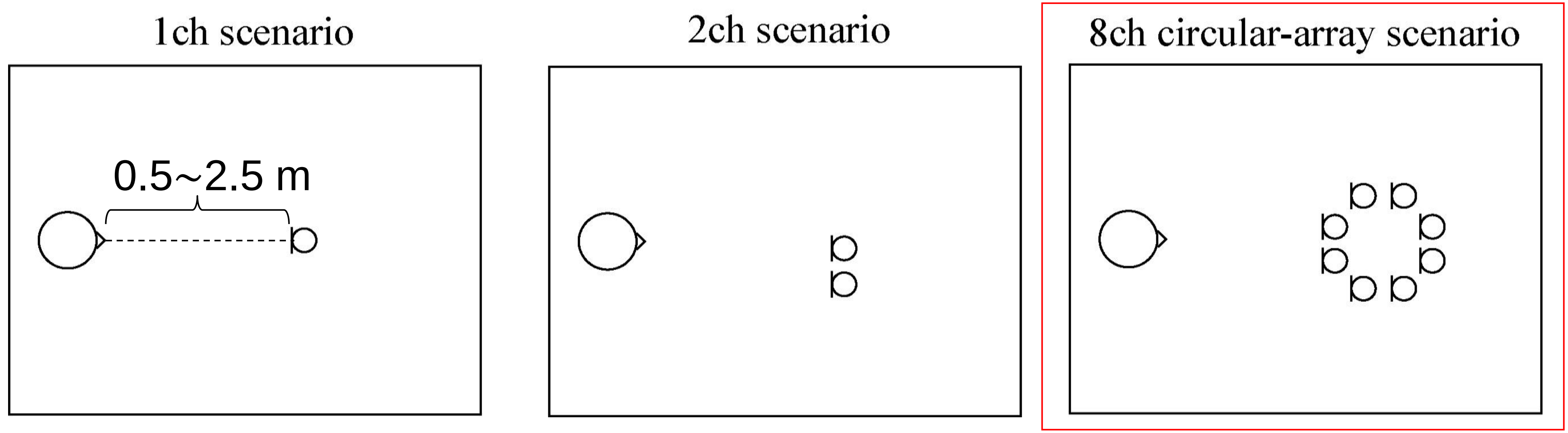
Clean speech PSD, time-freq. mask, etc.

Many variations are proposed depending on tasks (masking/ regression), cost functions, and network structures

[Weninger et al., 2014, Williamson and Wang, 2017]

REVERB Challenge task [Kinoshita et al., 2016]

- Task
  - Speech enhancement
  - ASR
- Acoustic conditions
  - Reverberation (Reverberation time 0.2 to 0.7 s.)
  - Stationary noise （SNR ～20dB)

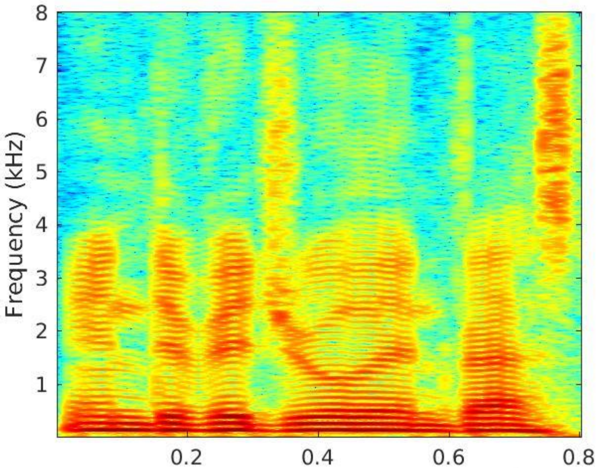


Comparison of three approaches

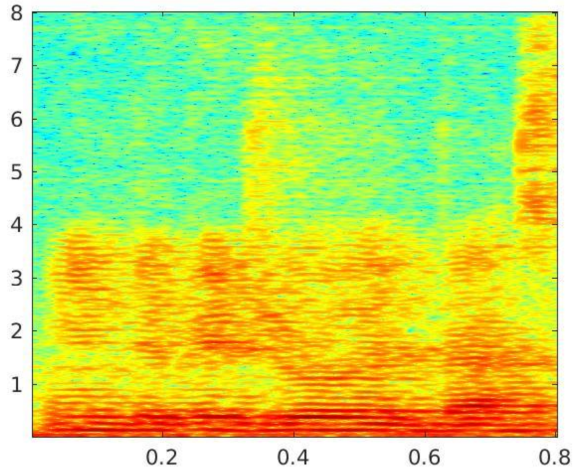
	Simu data				Real data
	FWSSNR	**CD**	**PESQ**	**WER**	**WER**
Observed	3.62 dB	3.97 dB	1.48	5.23 %	18.41 %
MVDR	6.59 dB	3.43 dB	1.75	6.65 %	14.85 %
WPE	4.79 dB	3.74 dB	2.33	4.35 %	13.24 %
WPE+MVDR	7.30 dB	**3.01 dB**	**2.38**	**3.85 %**	**9.90 %**
DNN (soft mask estimation)	**7.52 dB**	3.11 dB	**1.46**	**7.98 %**	**23.38 %**

FWSSNR: Frequency-weighted segmental SNR
CD: Cepstral distortion
PESQ: Perceptual evaluation of speech quality
WER: Word error rate (obtained with Kaldi REVERB baseline)

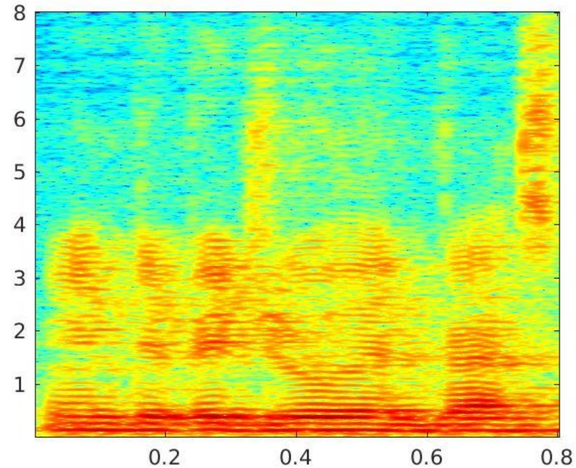
Demonstration



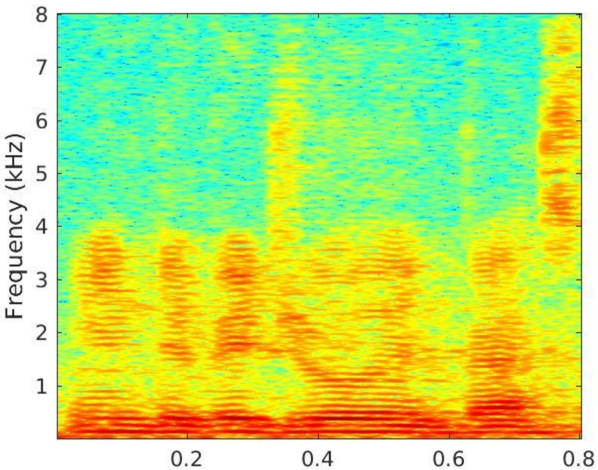
1. Headset



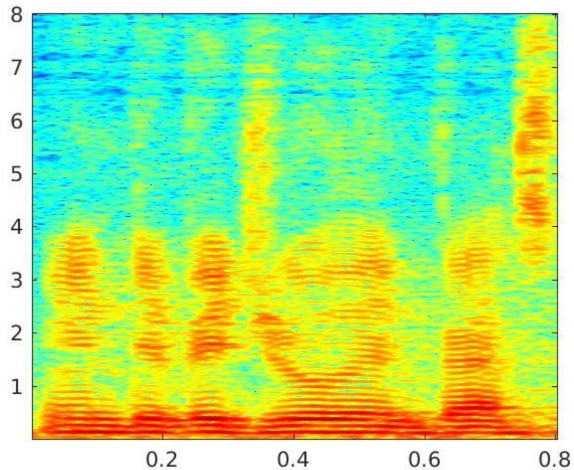
2. Observed



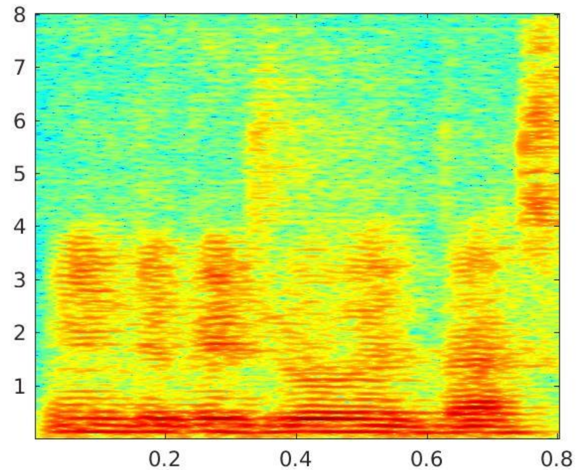
3. MVDR



4. WPE



5. WPE+MVDR

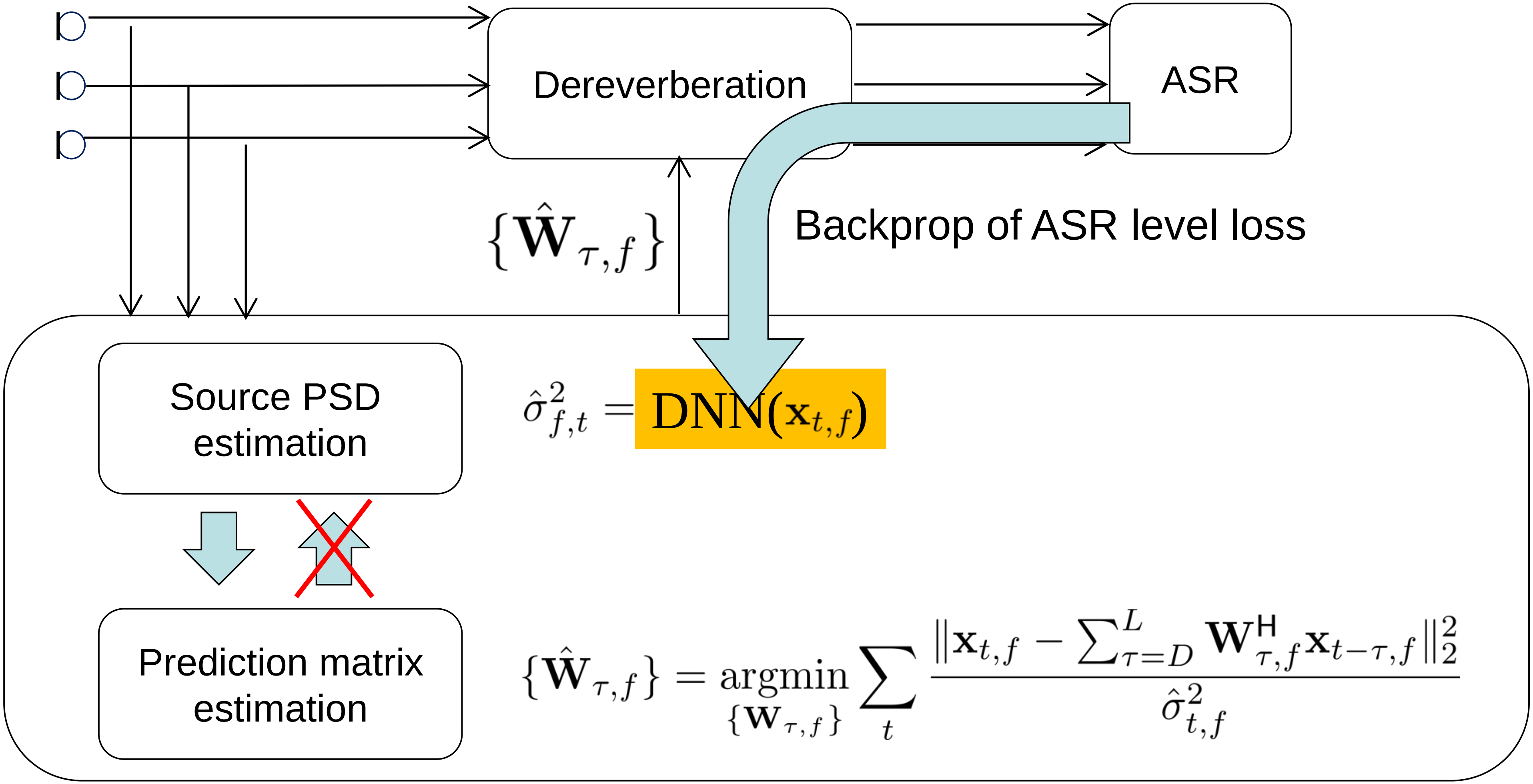


6. DNN

Pros and cons of three approaches

	Pros	Cons
Beamforming	• Low computational complexity • **Capable of simultaneous denoising and dereverberation** • High contribution to ASR	• **Less effective dereverberation**
WPE	• **Effective dereverberation** • **High contribution to ASR**	• No denoising capability • Computationally demanding • **Iteration required for source PSD estimation**
Neural networks	• Effective dereverberation **(source PSD estimation with no iterations)**	• Sensitive to mismatched condition • **Low contribution to ASR**

DNN-WPE [Kinoshita et al., 2017, Heyman et al., 2019]

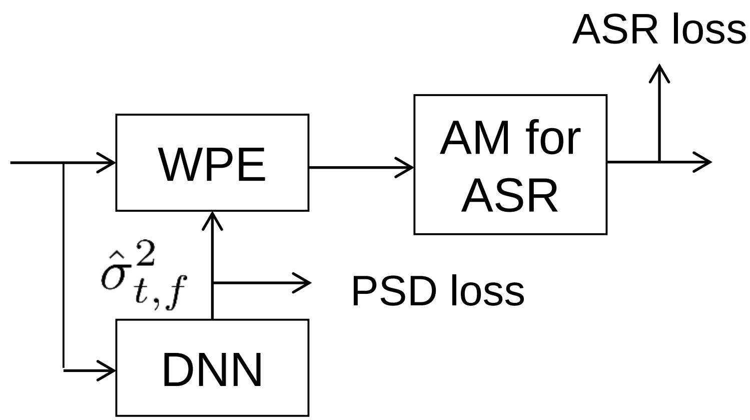


Advantages
1. No iterative estimation → Effective for online processing
2. DNN can be optimized jointly with an ASR system

Effectiveness of DNN-WPE [Heymann et al., 2019]

Training of DNN-WPE

- PSD-loss: MSE of PSD estimates
- ASR-loss: cross entropy of acoustic mode (AM) output

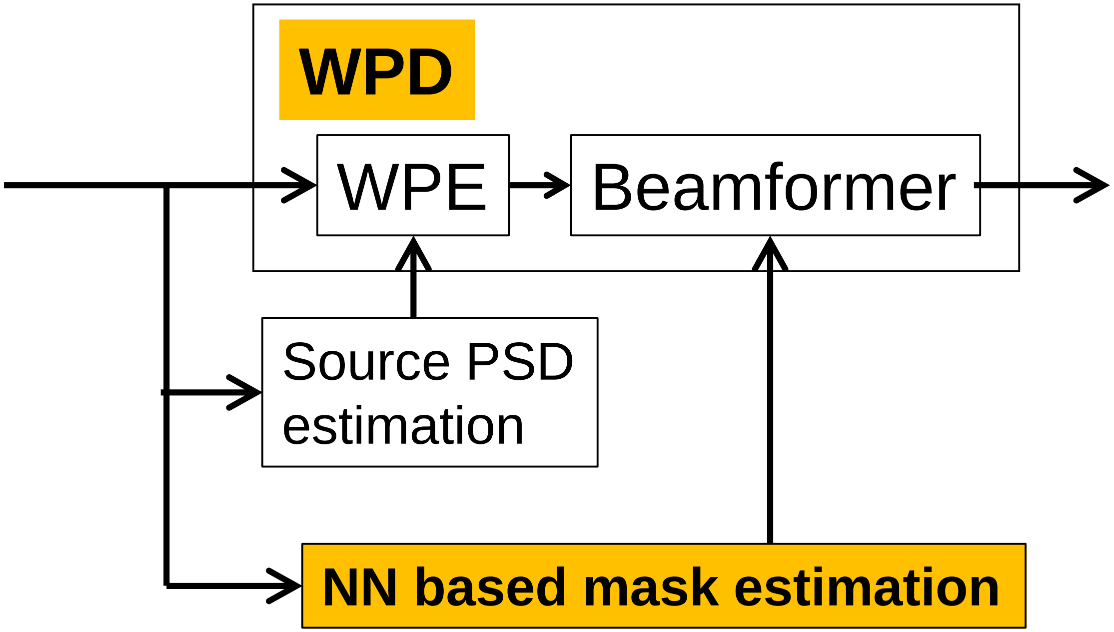


	REVERB (real)		WSJ+VoiceHome	
	Offline	Online	Offline	Online
Unprocessed	17.6		24.3	
WPE	13.0	16.2	18.6	20.0
DNN-WPE (PSD loss)	**10.8**	14.6	18.1	19.3
DNN-WPE (ASR loss)	11.8	**13.4**	**17.7**	**18.4**

Denoising are not performed, and different ASR backend is used.

Frame-online framework for simultaneous denoising and dereverberation

- WPD*[1]: a convolutional beamformer integrates WPE, beamformer, and DNN-based mask estimation



*1: Weighted Power minimization Distortionless response convolutional beamformer

Presentation at Interspeech 2019: 12:40-13:00, Mon, Sep. 16 [Nakatani et al, 2019b]

Software

- WPE
  - Matlab p-code for iterative offline, and block-online processing

    http://www.kecl.ntt.co.jp/icl/signal/wpe/

  - Python code w/ and w/o tensorflow for iterative offline, block-online, and frame-online processing

    https://pypi.org/project/nara-wpe/

- WPE, DNN-WPE
  - Python code with pytorch for offline and frame-online processing

    https://github.com/nttcslab-sp/dnn_wpe

    - Joint optimization of beamforming and dereverberation with end-to-end ASR enabled with espnet (https://github.com/espnet/espnet)

Table of contents

1. Introduction by Tomohiro
2. Noise reduction by Reinhold
3. Dereverberation by Tomohiro

Break (30 min)

4. **Source separation** by Reinhold
5. Meeting analysis by Tomohiro
6. Other topics by Reinhold
7. Summary by Reinhold & Tomohiro

QA