List of Abbreviations

AM	Acoustic Model
ASR	Automatic Speech Recognition
ATF	Acoustic Transfer Function
BAN	Blind Analytic Normalization
BLSTM	Bi-directional LSTM
BSS	Blind Source Separation
CACGMM	Complex Angular Central GMM
CD	Cepstral Distortion
CE	Cross Entropy
CNN	Convolutional Neural Network
DAN	Deep Atractor Network
DC	Deep Clustering
DER	Diarization Error Rate
DL	Deep Learning
DNN	Deep Neural Network
DOA	Direction-Of-Arrival
DSP	Digital Signal Processing
EM	Expectation-Maximization
FF	Feed Forward
FWSSNR	Frequency-Weighted Segmental SNR
GEV	Generalized Eigenvalue Decomposition
GMM	Gaussian Mixture Model
ICA	Independent Component Analysis
IVA	Independent Vector Analysis
ILRMA	Independent Low-Rank Matrix Analysis
LP	Linear Prediction
LSTM	Long-Short Term Memory
ML	Maximum Likelihood
MMSE	Minimum Mean Squared Error
MPDR	Minimum Power Distortionless Response

MSE	Mean Squared Error
MVDR	Minimum Variance Distortionless Response
MWF	Multichannel Wiener Filter
NMF	Nonnegative Matrix Factorization
NN	Neural Network
PESQ	Perceptual Evaluation of Speech Quality
PIT	Permutation Invariant Training
PLDA	Probabilistic Linear Discriminant Analysis
PSD	Power Spectral Density
RIR	Room Impulse Response
RNN	Recurrent Neural Network
RSAN	Recursive Selective Attention Network
RTF	Relative Transfer Function
SCER	Speaker Confusion Error Rate
SDW	Speech Distortion Weighted
SDR	Signal to Distortion Ratio
SDW-MWF	Speech Distortion Weighted MWF
SNR	Signal to Noise Ratio
SPP	Speech Presense Probability
STFT	Short-Time Fourier Transformation
STOI	Short-Time Objective Intelligibility
TasNet	Time Domain Audio Separation Network
TF	Time-Frequency
TDOA	Time Difference Of Arrival
TDNN	Time-Delay Neural Network
VAD	Voice Activity Detection
WER	Word Error Rate
WPE	Weighted Prediction Error
WSJ	Wall Street Journal

List of Notations

Mathematical expressions and operations	
$\top$ and H	Non-conjugate and conjugate transpose.
a	A scalar variable.
$\mathbf{a}$	A column vactor.
$\mathbf{A}$	A matrix.
D	A constant.
σ	A scalar parameter, such as a power spectral density (PSD) of a source.
Ψ	A matrix parameter, such as a spatial covariance matrix.
$\mathbb{E}[X]$	Expectation operator.
$\Pr(A = a)$	Probability
$p(x)$	Probability density function
$\mathcal{N}(\mathbf{x}; \mathbf{m}, \mathbf{R})$	Probability distribution of (multi-dimensional) (complex) normal distribution
$\mathrm{tr}\{\mathbf{\Phi}\}$	Trace of a matrix
$\|\cdot\|_2$	Eucredean norm of a vector
$\mathbb{R}$ and $\mathbb{C}$	A set of real scalars, and a set of complex scalars.
$\mathbb{R}^M$ and $\mathbb{R}^{M\times M}$	A set of M dimentional real vectors, and a set of $M \times M$ dimentional real matrices. $\mathbb{C}^M$ and $\mathbb{C}^{M\times M}$ are defined similarly.
$\nabla_{\mathbf{w}} J(\mathbf{w}) \in \mathbb{R}^{N\times 1}$	Gradient in denominator layout: Gradient is a column vector; Note: $\nabla_{\mathbf{w}} J(\mathbf{w}) = \dfrac{\partial}{\partial \mathbf{w}} J(\mathbf{w})$

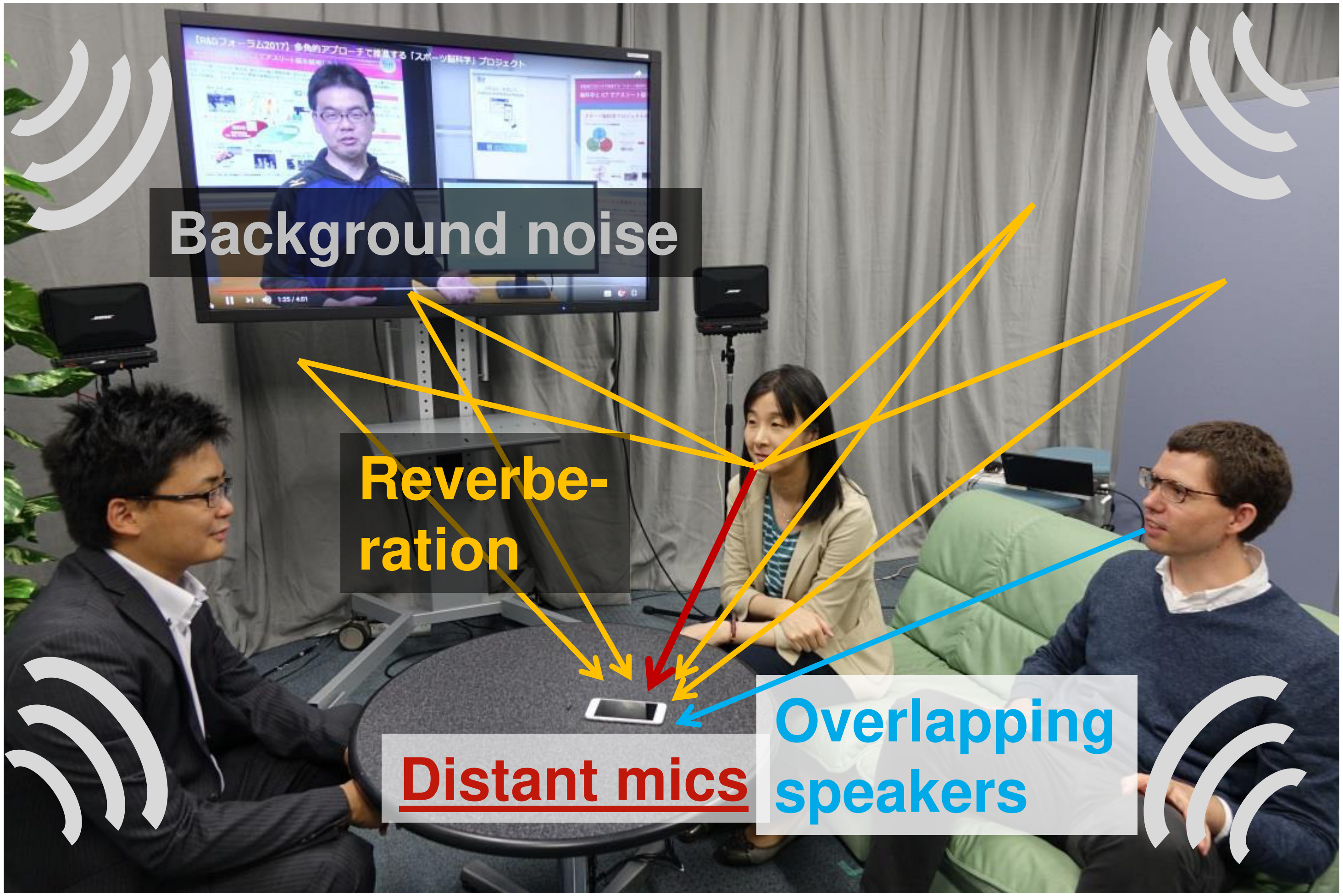
Symbols for Short Time Fourier Transformation (STFT) domain signals	
t, f, m, and i	Indicies of time frames, frequency bins, microphones, and sources.
T, F, M, and I	The numbers of time frames, frequency bins, microphones, and sources.
$s_{t,f}^{(i)} \in \mathbb{C}$	A clean signal for the i-th source.
$x_{m,t,f}^{(i)} \in \mathbb{C}$	A microphone image of the i-th source at the m-th microphone, i.e, noiseless reverberant signal for the source captured at the microphone.
$n_{m,t,f} \in \mathbb{C}$	Diffuse noise.
$y_{m,t,f} \in \mathbb{C}$	A signal captured at the m-th microphone. When I sources and diffuse noise are included, it is typically modeled by $y_{m,t,f} = \sum_{i=1}^{I} x_{m,t,f}^{(i)} + n_{m,t,f}$.
$d_{m,t,f}^{(i)} \in \mathbb{C}$	A part of $x_{m,t,f}^{(i)}$ composed of its direct signal and early reflections.
$r_{m,t,f}^{(i)} \in \mathbb{C}$	A part of $x_{m,t,f}^{(i)}$ composed of its late reverberation.
$\mathbf{y}_{t,f} \in \mathbb{C}^M$	A vector composed of $y_{m,t,f}$ for all m, i.e., $\mathbf{y}_{t,f} = (y_{1,t,f}, \ldots, y_{M,t,f})^\top$. $\mathbf{n}_{t,f}$, $\mathbf{x}_{t,f}^{(i)}$, $\mathbf{d}_{n,f}^{(i)}$, and $\mathbf{r}_{n,f}^{(i)}$ are defined similarly.
$\mathbf{x}_{t,f} \in \mathbb{C}^M$	Sum of $\mathbf{x}_{t,f}^{(i)}$ for all i, namely $\mathbf{x}_{t,f} = \sum_{i=1}^{I} \mathbf{x}_{t,f}^{(i)}$.

Symbols for time domain signals	
$\tilde{t}$ and $\tilde{T}$	A time sample index and the number of time samples in time domain. The same symbols as those for STFT domain signals are used for m, i, M, and I.
$y_m[\tilde{t}]$	A signal captured at the m-th microphone. $x_m^{(i)}[\tilde{t}]$ and $n_m[\tilde{t}]$ are defined similarly.

Part I.
Introduction

Tomohiro Nakatani

Speech recording from a conversation



- Speech enhancement is needed to extract each speaker's voice from various interferences

Applications of speech enhancement

- Hearing assistant
  - Hearing aids
  - Hands-free phones/conferences
- Far-field ASR
  - Home/personal assistants
  - Communication robots
  - Meeting transcription

Deep Learning – One Hammer for all Nails?

Deep Learning is used everywhere

- Speech enhancement, ASR, …

Does this mean we can forget microphone array signal processing?

No!

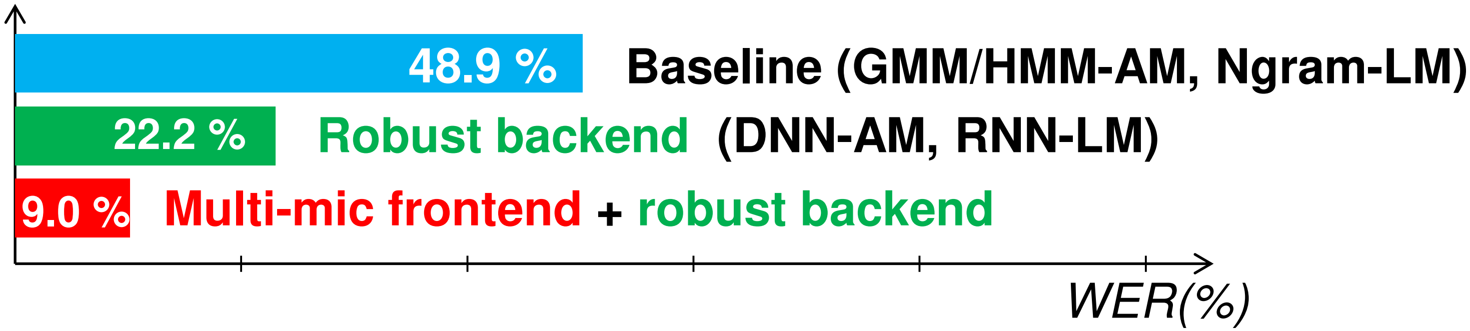
Goal of this talk

- Demonstrate the complementary power of deep neural network (DNN) and microphone array signal processing
- Argue that their integration is very helpful

Quick overview of effectiveness (1/2)

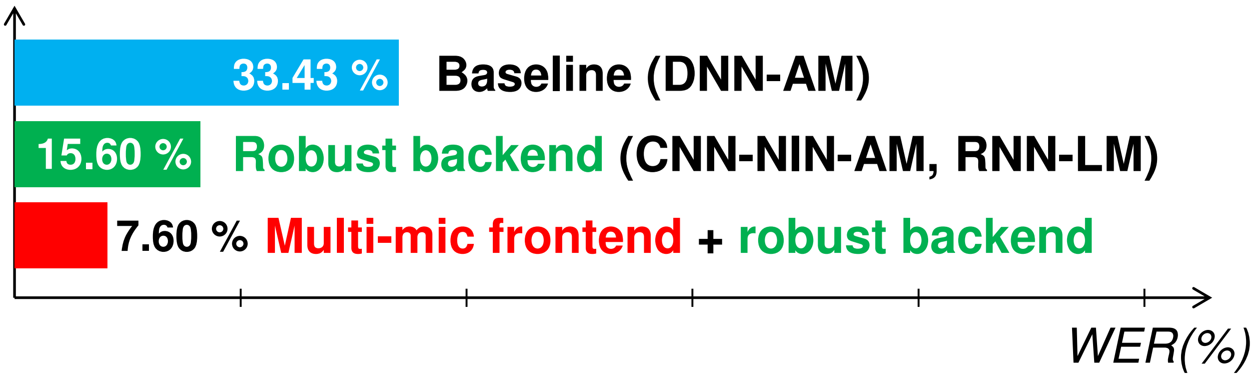
REVERB 2014

[Delcroix et al., 2015]



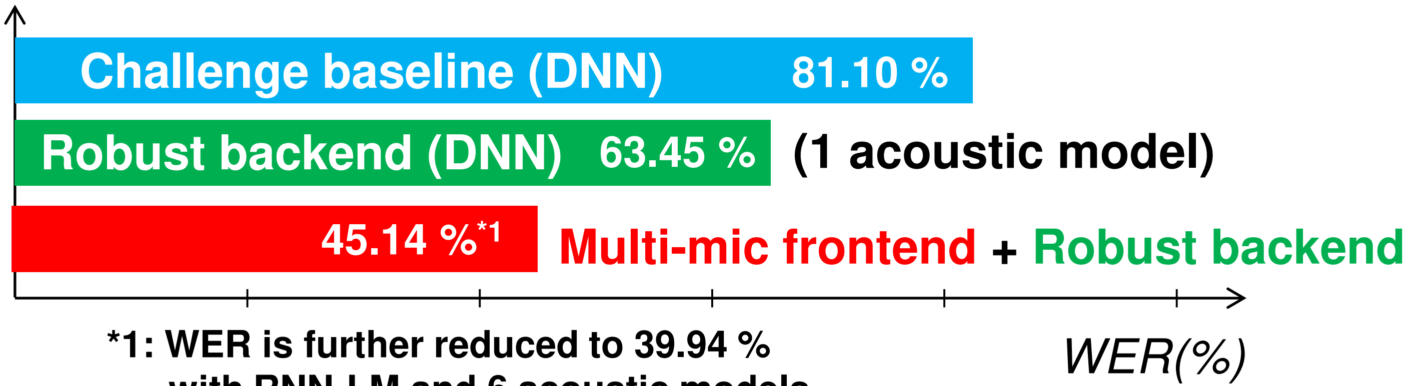
CHiME-3 2015

[Yoshioka et al., 2015]



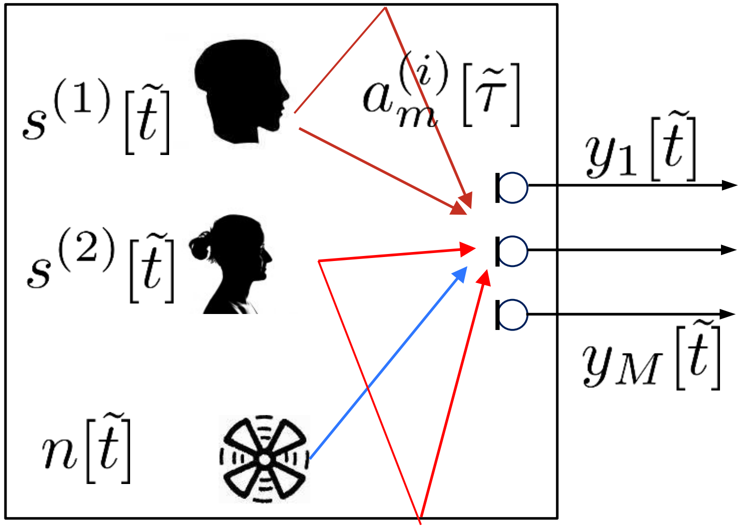
CHiME-5 2018

[Kanda et al., 2019]



***1: WER is further reduced to 39.94 % with RNN-LM and 6 acoustic models.**

Model of recorded speech: time domain



$\tilde{t}$: time index

$s^{(i)}[\tilde{t}]$: i-th source for $1 \le i \le I$

$a_m^{(i)}[\tilde{\tau}]$: room impulse response (RIR) from i-th source to m-th mic

$n[\tilde{t}]$: noise

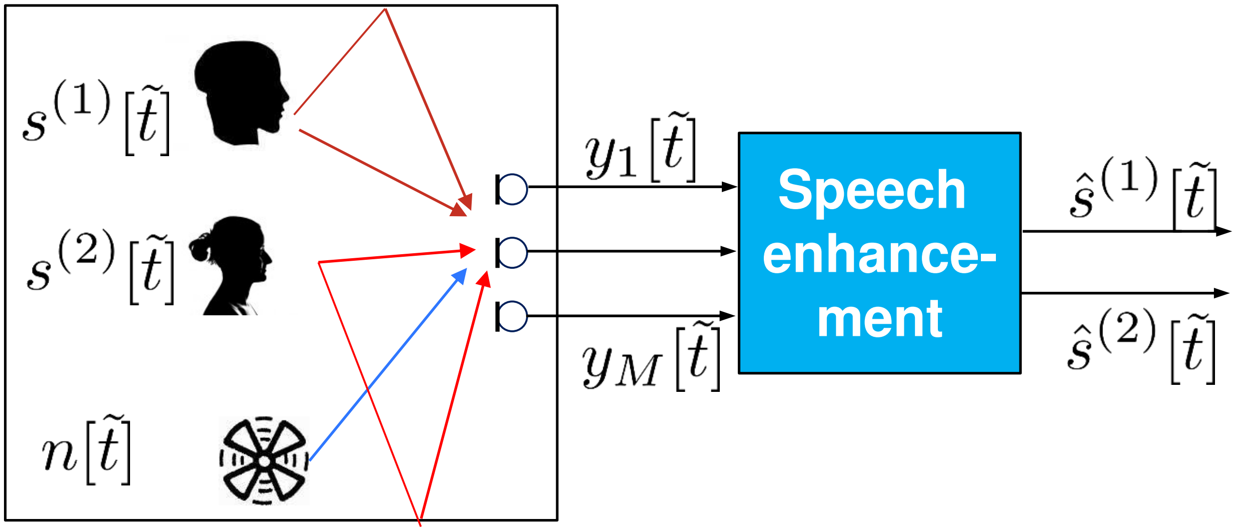
- Observed:

$$y_m[\tilde{t}] = \sum_{i=1}^{I} \left(\sum_{\tilde{\tau}=0}^{L-1} a_m^{(i)}[\tilde{\tau}] s^{(i)}[\tilde{t} - \tilde{\tau}] \right) + n_m[\tilde{t}]; \quad m = 1, \ldots, M$$

$$\mathbf{y}[\tilde{t}] = \sum_{i=1}^{I} \left(\sum_{\tilde{\tau}=0}^{L-1} \mathbf{a}^{(i)}[\tilde{\tau}] s^{(i)}[\tilde{t} - \tilde{\tau}] \right) + \mathbf{n}[\tilde{t}]; \quad \mathbf{y}[\tilde{t}] = \begin{pmatrix} y_1[\tilde{t}] \\ \ldots \\ y_M[\tilde{t}] \end{pmatrix}$$

Goal of speech enhancement

- Denoising – reducing noise
- Dereverberation – reducing reverberation
- Source separation – separating mixtures to individual speeches



- Meeting analysis – diarization (detecting who speaks when) + speech enhancement

Evaluation metrics

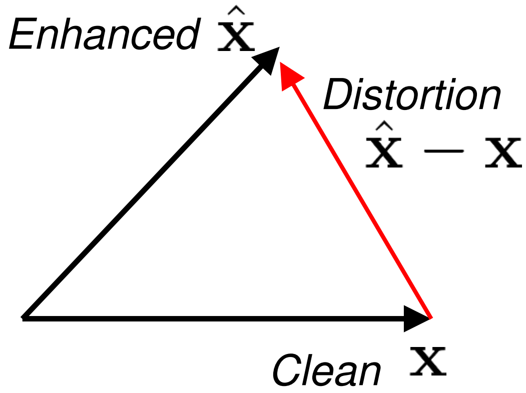
Type	Examples of measures	Pros and cons
Signal level distortion metric	• **Signal to distortion Ratio (SDR)** - Many variations • Frequency-weighted segmental SNR (FWSSNR), cepstral distortion (CD), signal-to-interference ratio (SIR), etc.	• **Most frequently used** • **Not directly reflect perceptual quality/ASR performance** • **Parallel data required** (Incompatible with real recordings)
ASR	• **Word error rate (WER)** and character error rate (CER)	• **Useful for ASR** • **No parallel data required** • **Dependent on ASR systems**
Perceptual quality (listening test)	• Mean opinion score (MOS) • MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA)	• **Reliable** • **Costly** • Dependent on subjects, and test conditions
Perceptual quality (objective measure)	• **PESQ: speech quality** • **STOI: speech intelligibility** • Others : HASPI, EPSM, SIIB, SRMR_norm, GEDI, DNN-based, etc.	• **Perceptually validated** • **Applicability is limited to certain distortion types**

None of them are "perfect"   Do not rely on one !

SDR variations

- BSSEval-SDR [Vincent et al., 2006]

$$\text{BSSEval-SDR}^{(\text{image})} = 10 \log_{10} \frac{\sum_{\tilde{t}} |x[\tilde{t}]|^2}{\sum_{\tilde{t}} |\hat{x}[\tilde{t}] - x[\tilde{t}]|^2}$$



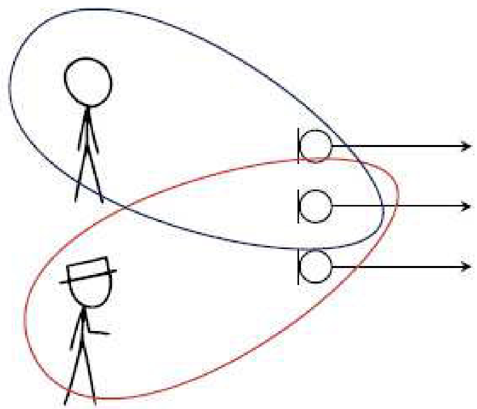
  - Sensitive to scale and phase estimation errors
- Variations
  - Scale-invariant SDR [Le Roux et al., 2019]
    - Invariant to scaling errors
  - Time-invariant filter allowed distortion [Vincent et al., 2006]
    - Invariant to scale and phase estimation errors
- Issues:
  - Smaller but important energy components are almost disregarded, causing mismatch with human perceptual behavior and ASR performance
  - Parallel data composed of clean and noisy signals are required

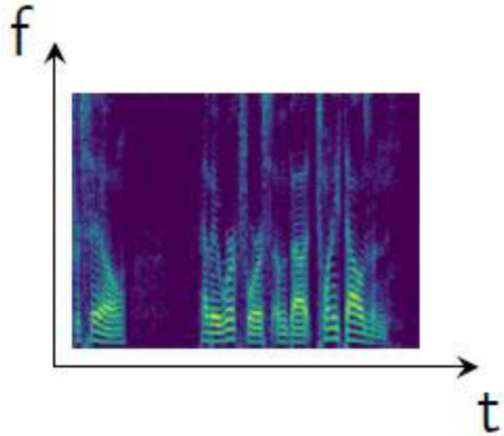
Evaluation metrics

Type	Examples of measures	Pros and cons
Signal level distortion metric	• **Signal to distortion Ratio (SDR)** - Many variations • Frequency-weighted segmental SNR (FWSSNR), cepstral distortion (CD), signal-to-interference ratio (SIR), etc.	• **Most frequently used** • **Not directly reflect perceptual quality/ASR performance** • **Parallel data required** (Incompatible with real recordings)
ASR	• **Word error rate (WER)** and character error rate (CER)	• **Useful for ASR** • **No parallel data required** • **Dependent on ASR systems**
Perceptual quality (listening test)	• Mean opinion score (MOS) • MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA)	• **Reliable** • **Costly** • Dependent on subjects, and test conditions
Perceptual quality (objective measure)	• **PESQ: speech quality** • **STOI: speech intelligibility** • Others : HASPI, EPSM, SIIB, SRMR_norm, GEDI, DNN-based, etc.	• **Perceptually validated** • **Applicability is limited to certain distortion types**

None of them are “perfect” Do not rely on one !

Cues for speech enhancement





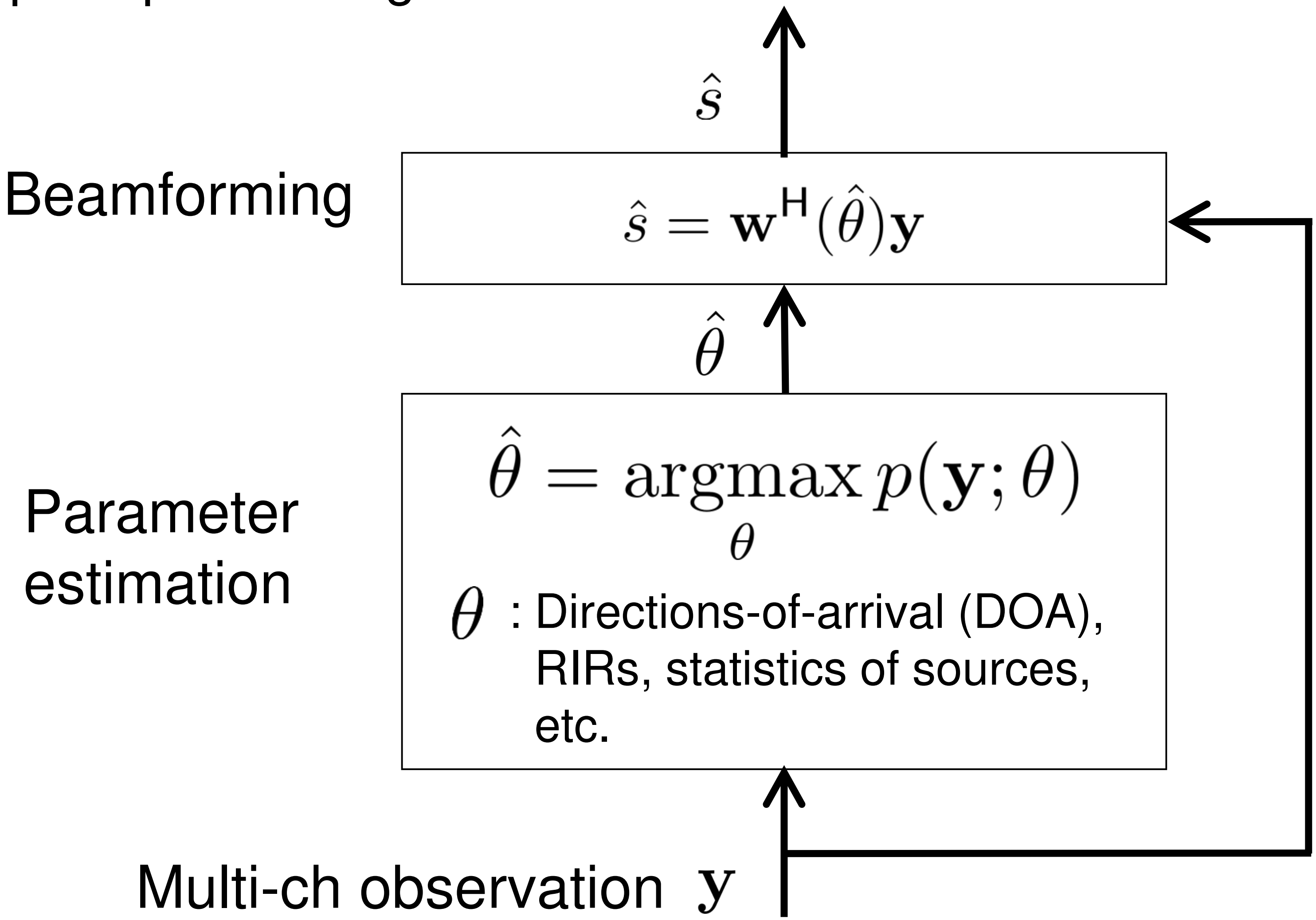
- Spatial
  - Exploits spatial selectivity (multi-channel)
  - Does not exploit speech characteristics (could work for any signal)
- Spectro-temporal
  - Speakers/phonemes have different spectro-temporal characteristics
  - Model speech characteristics

Three approaches to speech enhancement

- Microphone array signal processing
  - Spatial cues
- Neural networks
  - Spectro-temporal cues
- Hybrid of both approaches
  - All cues

Microphone array signal processing (1/2)

- Typical processing flow



Microphone array signal processing (2/2)

- Use generative model to estimate unknown observation system

A generative model: $$p(\mathbf{y};\theta) = \int \underbrace{p(\mathbf{y}|s,\mathbf{n};\theta_r)}_{\text{Room acoustics}}\underbrace{p(s;\theta_s)}_{\text{Speech}}\underbrace{p(\mathbf{n};\theta_n)}_{\text{Noise}}ds\,d\mathbf{n}$$

θ_s : Speech power spectral density, voice activity, etc.

θ_n : Noise power spectral density, etc.

θ_r : Directions-of-arrival (DOAs), room impulse responses (RIRs), etc.

Inverse system: e.g. by maximum likelihood (ML) parameter estimation:

$$\hat{\theta} = \operatorname*{argmax}_{\theta} p(\mathbf{y};\theta)$$

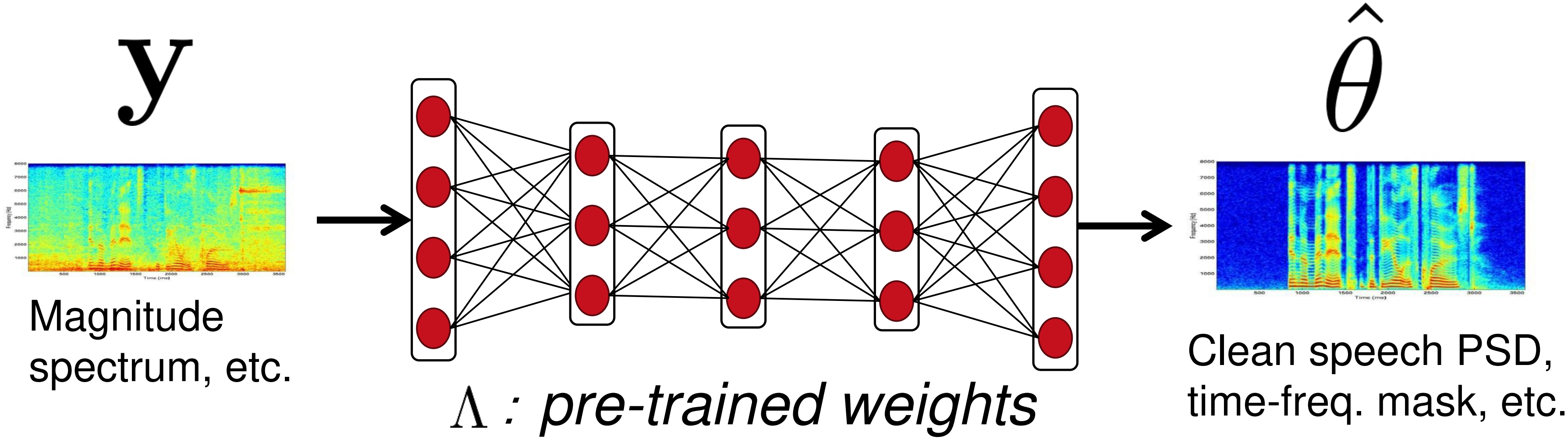
- Beamforming: e.g., by MMSE estimation

$$\hat{s} = \operatorname*{argmin}_{\hat{s}} \int |s - \hat{s}|^2 p(s|\mathbf{y};\hat{\theta})ds = \mathbf{w}^{\mathsf{H}}(\hat{\theta})\mathbf{y}$$

Effective spatial filtering is applicable with no prior info. DOAs or RIRs.

Neural networks

- Train neural networks using huge amount of training data



Robust and accurate spectral estimation is possible

Interpret this as the inverse system of the generative model, that estimates the model parameters from observation.

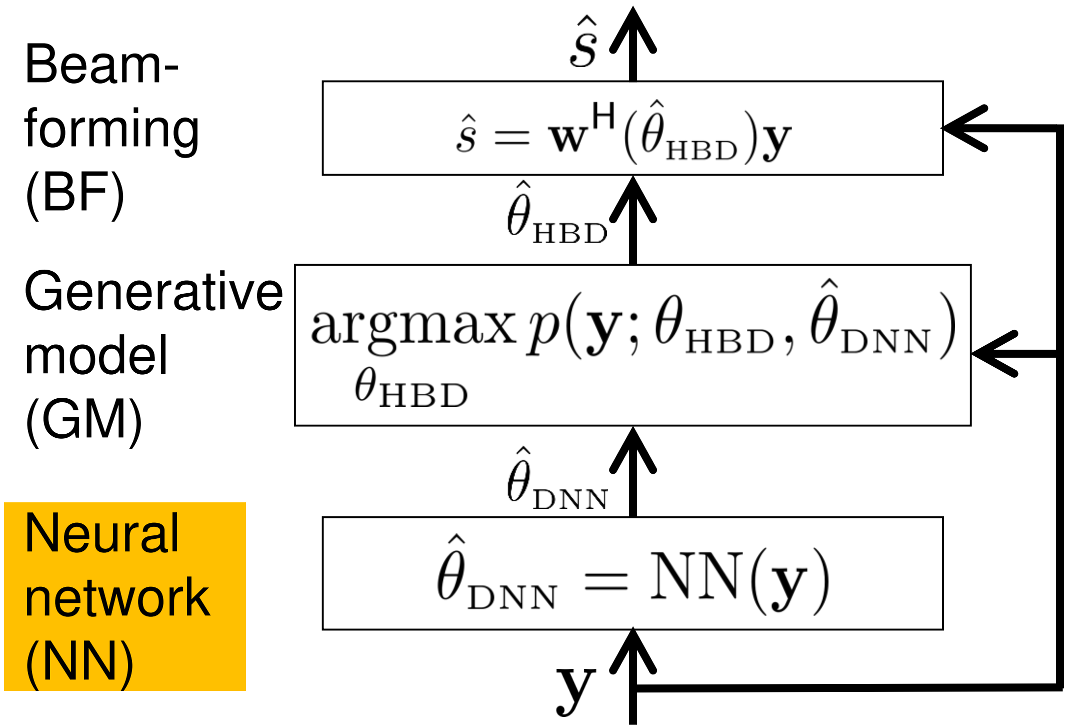
Pros and cons of two approaches

	Microphone array signal processing	Neural networks
Spatial characteristics modeling	• **Strong**	• Moderate (use spatial features as auxiliary input)
Spectro-tempral characteristics modeling (for speech)	• Weak - Permutation problem • No concept of human speech (pros and cons)	• **Very strong** - Strong speech model based on a priori training - Single channel processing applicable
Adaptation to test condition	• **Strong** - Unsupervised learning applicable	• Weak - Poor generalization - Sensitive to mismatch
Interpretability	• **Highly interpretable**	• **Blackbox**

Their pros and cons are highly complementary

Hybrid approaches (1/2)

1) Microphone array boosted by neural networks



- Component-wise optimization
- Joint optimization

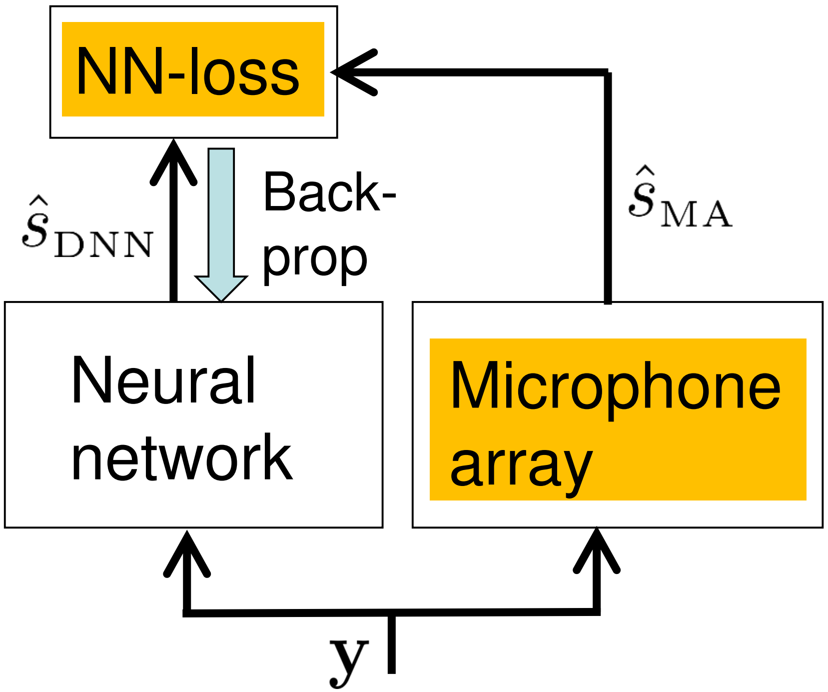
Examples:

- **Mask-based beamforming** (Part II, IV, V, and VI)
  - NN: Mask estimation
  - GM: signal statistics estimation
  - BF: MVDR beamforming
- **DNN-WPE dereverberation** (Part III)
  - NN: PSD estimation
  - GM: Inverse filter estimation
  - BF: Inverse filtering

Achieving state-of-the-art in each example

Hybrid approaches (2/2)

2) Unsupervised learning of neural networks enabled by microphone array



- Approach-1) can be combined after training

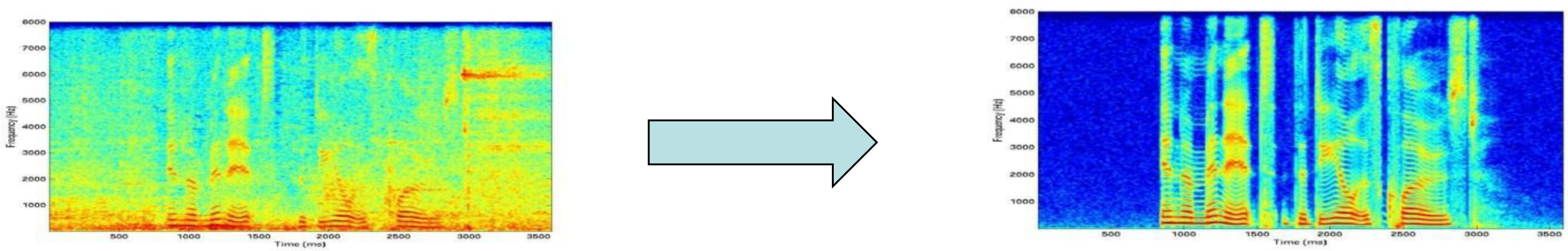
Examples:

- **Unsupervised training of DNN based source separation** (part VI)

Show complementary power of microphone array and DNN

Focus in this tutorial

- This tutorial concentrates on enhancement as a frontend of ASR. This implies different constraints than enhancement for human-to-human communication
  - Less tight latency requirements
    - Utterance-wise processing
    - Quasi-static acoustic scenes assumed
  - Perceptual quality of output less important
    - as long as WER is good
- The solutions here are not readily suitable for enhancing human-to-human speech communication



Benchmarks and Challenges

#targets=1

#targets>1

Real

CHiME CHALLENGE

CHiME-5

AMI CONSORTIUM

DIRHA

REVERB CHALLENGE

CHiME CHALLENGE

CHiME-3/4

MC-WSJ

CSTR

CHiME CHALLENGE

CHiME-1,2

AURORA

wsj0-2mix, WHAM!

MITSUBISHI ELECTRIC
Changes for the Better

Simulation
(Benchmark)

Roles of simulation data vs real recordings

- Simulation data : sounds are mixed on computer
  - Pros:
    - Useful for **data augmentation and training of NN**
    - Parallel data available, **useful for detailed performance analysis**
  - Variations
    - Noise: simulated (e.g., pink/white noise) or recoded
    - Reverb: convolution with simulated/measured RIR
    - Unrealistic data for benchmark: e.g., fixed #speakers keep uttering simultaneously with no noise or reverberation
- Real recordings: all sounds are recorded simultaneously
  - Pros:
    - Includes various varying factors inherently in real recordings
    - **Essential for reliable evaluation**
  - Variations
    - Recordings under controlled conditions for evaluation purposes
    - Recordings of real applications

Popular corpora for speech enhancement

Task	Name of task	Recording condition		
		Environment	#mics (Spk-Mic dist)	Simulated or Real
Denoising	AURORA 4 [Parihar et al., 2002]	Noise in public areas	1 (close mic)	Sim (measured noise, channel distortion)
	CHiME-1/2 [Barker et al., 2013, Vincent et al., 2013]	Home	2 (2m)	Sim (measured noise and RIR)
	CHiME-3/4 [Barker et al., 2017]	Public areas	6 (0.5m)	Sim (measured noise and RIR) + Real
Dereverbe-ration	REVERB [Kinoshita et al., 2016]	Reverberant conference room	1/2/8 (0.5-2m)	Sim (measured noise and RIR) + Real
	Aspire [Harper 2015]	7 different rooms	1/6	Real
	DIRHA [Ravanelli et al. 2015]	Home (distributed mics)	32	Real (distributed mics)
Source separation	wsj0-mix [Hershey et al., 2016]	Mixture of clean signal	1 (close mic)	Sim (no noise, no reverb)
	wsj0-mix [Wang et al., 2018c]	Mixture of anechoic/ reverberated signal	8 (1.3∓0.4m)	Sim (no noise, simulated RIR)
	WHAM! [Wichern et al., 2019]	Noise in public areas	1 (close mic)	Sim (measured noise, no reverb)
	MC-WSJ-AV [Lincoln et al., 2005]	Reverberant conference room	8 (0.5-2m)	Real
Meeting analysis	AMI [Carletta 2006]	Meeting room	8	Real
	CHiME-5 [Barker et al., 2018]	Home (distributed mics)	24	Real
	DIHARD-I,II [Ryant et al., 2019]	Multiple sources, incl. child recs, youtube	1	Real

Software for evaluation

- BSS Eval
  - Matlab: http://bass-db.gforge.inria.fr/bss_eval/
  - Python: https://sigsep.github.io/sigsep-mus-eval/museval.metrics.html
- REVERB challenge (FWSSNR, CD, SRMR, LLR, PESQ)
  - Matlab: https://reverb2014.dereverberation.com/download.html
- Perceptual evaluation of speech quality (PESQ)
  - https://www.itu.int/rec/T-REC-P.862
- Short-Time Objective Intelligibility (STOI)
  - Matlab: http://insy.ewi.tudelft.nl/content/short-time-objective-intelligibility-measure
  - Python: https://github.com/actuallyaswin/stoi

Table of contents

Part II.
Noise Reduction – Beamforming

Reinhold Haeb-Umbach

Speech capture in noisy environments

- Forming a beam of increased sensitivity towards the desired speaker reduces noise and other distortions

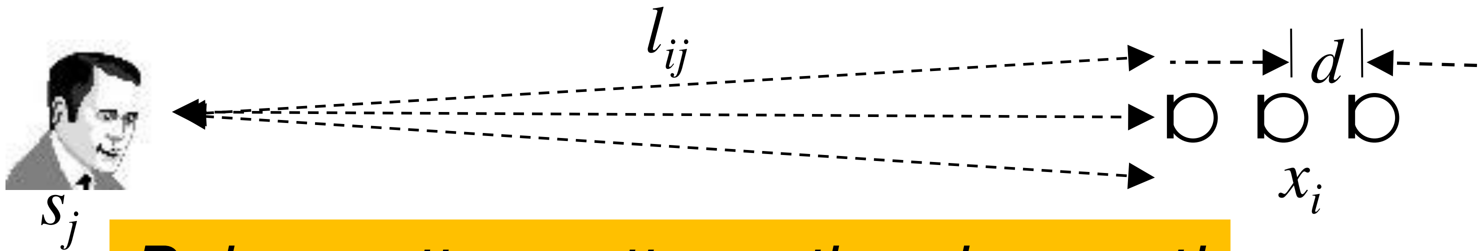
Table of contents in part II

Some physics

- In free space, waveform at point i caused by a waveform emitted at point j

$$x_i[\tilde{t}] = \frac{1}{\sqrt{4\pi l_{ij}}} s_j \left[\tilde{t} - \frac{l_{ij}}{c}\right]$$

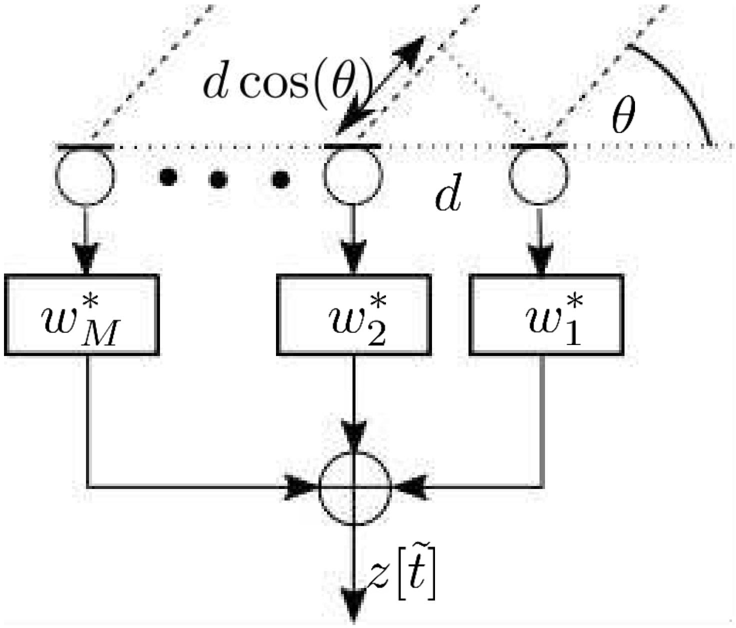
  where l_{ij} is distance from position i to j
- Far-field: l_{ij} much larger than inter-microphone distance d
  - Plane wave
  - Attenuation factor $1/\sqrt{4\pi l_{ij}}$ the same for all mics
  - Signal delay between microphones $\tilde{\tau} = d/c$ where $c \approx 340\,\mathrm{m/s}$
    - Example: for $d = 10\,\mathrm{cm} \Rightarrow \tilde{\tau} = 0.3\,\mathrm{ms} = 4.7$ samples @ 16 kHz



Delay matters, attenuation does not!

Basics of acoustic beamforming

$$s[\tilde{t}] = \mathrm{e}^{j\omega_0 \tilde{t}} = \mathrm{e}^{j\frac{2\pi c}{\lambda_0}\tilde{t}}$$



Signal at *m*th microphone:

$$x_m[\tilde{t}] = s[\tilde{t} - \tilde{\tau}_m] = \mathrm{e}^{j\omega_0(\tilde{t}-\tilde{\tau}_m)}$$

$$\tilde{\tau}_m = \frac{(m-1)d\cos\theta}{c}; \ m = 1, \dots, M$$

Beamformer output:

$$z[\tilde{t}] = \sum_{m=1}^{M} w_m^* x_m[\tilde{t}]$$

$$= \dots$$

$$= \mathrm{e}^{j\omega_0 \tilde{t}} \mathbf{w}^{\mathrm{H}} \mathbf{v}(\theta, \lambda_0)$$

Beamformer coeff.: $\mathbf{w} = [w_1, \dots, w_M]^\top$

Steering vector: $\mathbf{v}(\theta, \lambda_0) = \begin{pmatrix} 1 & \mathrm{e}^{-j2\pi\left(\frac{d}{\lambda_0}\right)\cos(\theta)} & \dots & \mathrm{e}^{-j2\pi\left(\frac{d}{\lambda_0}\right)\cos(\theta)(M-1)} \end{pmatrix}$

Delay-Sum Beamformer (DSB)

- Delay-Sum Beamformer: $\mathbf{w} = \frac{1}{M}\begin{pmatrix}1 & \mathrm{e}^{-j\phi_0} & \cdots & \mathrm{e}^{-j(M-1)\phi_0}\end{pmatrix}^{\mathsf{T}}$

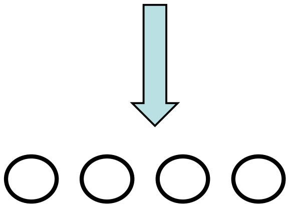
  with phase term $\phi_0 = \omega_0\tau_0 = \omega_0\frac{d\cos\theta_0}{c} = 2\pi\frac{d}{\lambda_0}\cos(\theta_0)$

  – DSB steered towards geometric angle θ_0

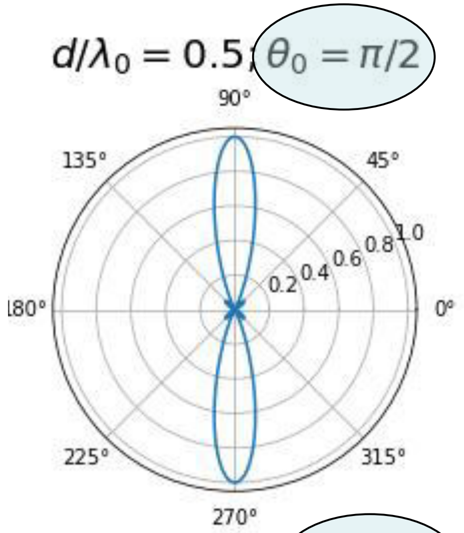
- Beampattern:

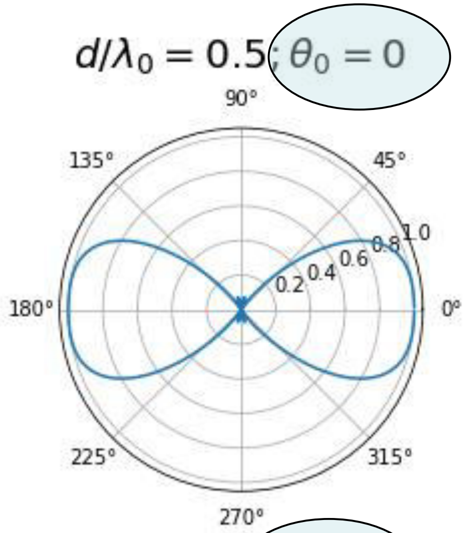
$$\begin{aligned}\left|z[\tilde{t}]\right| &= \left|\mathrm{e}^{j\omega_0\tilde{t}} \cdot \mathbf{w}^{\mathrm{H}}\mathbf{v}\right| \\ &= \cdots \\ &= \frac{1}{M}\left|\frac{\sin\left(\frac{M}{2}2\pi\frac{d}{\lambda_0}(\cos(\theta)-\cos(\theta_0))\right)}{\sin\left(\frac{1}{2}2\pi\frac{d}{\lambda_0}(\cos(\theta)-\cos(\theta_0))\right)}\right|\end{aligned}$$

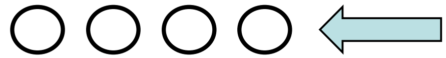
Example beampatterns



Broadside
(here: top/bottom)

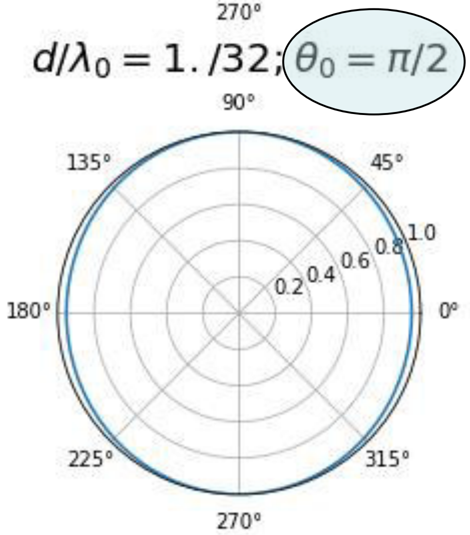


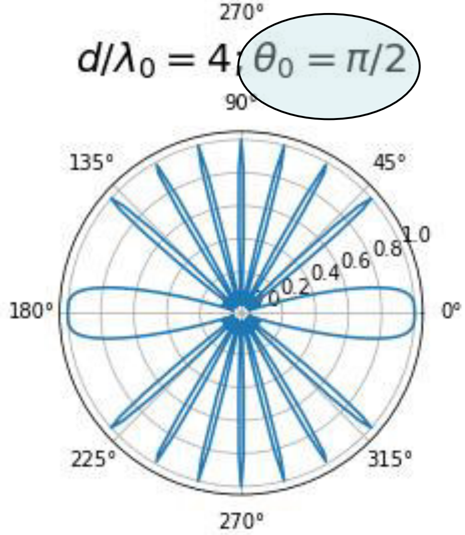




Endfire
(here: left/right)

small
inter-element
distance /
low frequency

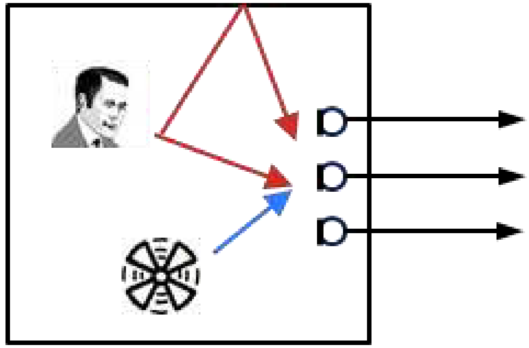




large
inter-element
Distance /
high frequency

From physics to signal processing

Real acoustic environments:



- Reverberation
  - Time differences of arrival (TDOAs) inappropriate
- Wideband beamforming
  - Fourier transform domain processing
- Interferences
  - Need appropriate objective functions
- Unknown and time-varying acoustic environment
  - Estimation of beamformer coefficients

Most common model

- Signal at m-th microphone:

$$x_m[\tilde{t}] = s[\tilde{t} - \tilde{\tau}_m] \;\; \rightarrow \;\; y_m[\tilde{t}] = x_m[\tilde{t}] + n[\tilde{t}] = \sum_{\tilde{\tau}=0}^{\tilde{L}-1} a_m[\tilde{\tau}] s[\tilde{t} - \tilde{\tau}] + n[\tilde{t}]$$

- Short-Time Fourier Transform (STFT): $y_m[\tilde{t}] \;\rightarrow\; y_{m,t,f}$

- Narrowband assumption (multiplicative transfer function approx.): length of acoustic impulse response << STFT analysis window
  - convolution in time domain corresponds to multiplication in STFT domain
- Time-invariant Acoustic Transfer Function (ATF)

$$y_{m,t,f} = a_{m,f} s_{t,f} + n_{t,f}; \quad m = 1, \ldots, M$$

$$\mathbf{y}_{t,f} = \mathbf{a}_f s_{t,f} + \mathbf{n}_{t,f} := \mathbf{x}_{t,f} + \mathbf{n}_{t,f}$$

ATF vs RTF

- Scale ambiguity of ATF

$$\mathbf{x}_{t,f} = \mathbf{a}_f s_{t,f} = (\mathbf{a}_f \cdot C) \cdot s_{t,f}/C; \quad C \in \mathbb{C}$$

- Fix ambiguity: Relative transfer function (RTF)

$$\tilde{\mathbf{a}}_f = \frac{\mathbf{a}_f}{a_{1,f}} = \left(1, \frac{a_{2,f}}{a_{1,f}}, \ldots, \frac{a_{M,f}}{a_{1,f}}\right)^{\mathsf{T}}$$

$$\Rightarrow \mathbf{x}_{t,f} = \mathbf{a}_f s_{t,f} = \tilde{\mathbf{a}}_f a_{1,f} s_{t,f} = \tilde{\mathbf{a}}_f x_{1,t,f}$$

- Thus our goal is to estimate the *image* of the source at a reference microphone (e.g., mic. #1)

$$x_{1,t,f} = a_{1,f} s_{t,f}$$

  – Thus, we do not attempt to dereverberate the signal!

Optimal beamforming design criteria: MMSE

- Beamformer output: $z_{t,f} = \mathbf{w}_f^{\mathsf{H}} \mathbf{y}_{t,f}$
- MMSE:

$$\min_{\mathbf{w}_f} \mathbb{E}\left[\left|\mathbf{w}_f^{\mathsf{H}}\mathbf{y}_{t,f} - x_{1,t,f}\right|^2\right] = \min_{\mathbf{w}_f} \mathbb{E}\left[\left|\mathbf{w}_f^{\mathsf{H}}\mathbf{x}_{t,f} - x_{1,t,f}\right|^2\right] + \mathbb{E}\left[\left|\mathbf{w}_f^{\mathsf{H}}\mathbf{n}_{t,f}\right|^2\right]$$

Add weight μ

Results in:

$$\mathbf{w}_f^{\text{SDW-MWF}} = \left(\mathbf{\Psi}_{\mathbf{xx},f} + \mu \mathbf{\Psi}_{\mathbf{nn},f}\right)^{-1} \mathbf{\Psi}_{\mathbf{xx},f}\mathbf{u}_1$$

where $\mathbf{\Psi}_{\mathbf{xx},f} = \mathbb{E}\left[\mathbf{x}_{t,f}\mathbf{x}_{t,f}^{\mathsf{H}}\right]$ (spatial covar. matrix of speech)

$\mathbf{\Psi}_{\mathbf{nn},f} = \mathbb{E}\left[\mathbf{n}_{t,f}\mathbf{n}_{t,f}^{\mathsf{H}}\right]$ (spatial covar. matrix of noise)

$\mathbf{u}_1 = [1, 0, \ldots, 0]^\top$ (points to reference microphone)

Speech Distortion Weighted Multi-channel Wiener Filter
(SDW-MWF)

Optimal beamforming design criteria: M(P|V)DR

- MPDR: Minimum Power Distortionless Response:

$$\min_{\mathbf{w}_f} \mathbb{E}\left[\left|\mathbf{w}_f^{\mathsf{H}} \boldsymbol{\Psi}_{\mathbf{yy},f} \mathbf{w}_f\right|^2\right] \text{ subject to } \mathbf{w}_f^{\mathsf{H}} \tilde{\mathbf{a}}_f = 1$$

gives $$\mathbf{w}_f{}^{\mathrm{MPDR}} = \frac{\boldsymbol{\Psi}_{\mathbf{yy},f}^{-1} \tilde{\mathbf{a}}_f}{\tilde{\mathbf{a}}_f^{\mathsf{H}} \boldsymbol{\Psi}_{\mathbf{yy},f}^{-1} \tilde{\mathbf{a}}_f}$$

- MVDR: Minimum Variance Distortionless Response:

$$\min_{\mathbf{w}_f} \mathbb{E}\left[\left|\mathbf{w}_f^{\mathsf{H}} \boldsymbol{\Psi}_{\mathbf{nn},f} \mathbf{w}_f\right|^2\right] \text{ subject to } \mathbf{w}_f^{\mathsf{H}} \tilde{\mathbf{a}}_f = 1$$

gives $$\mathbf{w}_f{}^{\mathrm{MVDR}} = \frac{\boldsymbol{\Psi}_{\mathbf{nn},f}^{-1} \tilde{\mathbf{a}}_f}{\tilde{\mathbf{a}}_f^{\mathsf{H}} \boldsymbol{\Psi}_{\mathbf{nn},f}^{-1} \tilde{\mathbf{a}}_f}$$

Optimal beamforming design criteria: maxSNR

- Maximize output SNR:

$$\max_{\mathbf{w}_f} \frac{\mathbf{w}_f^{\mathsf{H}} \boldsymbol{\Psi}_{\mathbf{xx},f} \mathbf{w}_f}{\mathbf{w}_f^{\mathsf{H}} \boldsymbol{\Psi}_{\mathbf{nn},f} \mathbf{w}_f}$$

leads to generalized eigenvalue problem. $\boldsymbol{\Psi}_{\mathbf{xx},f} \mathbf{w}_f = \lambda \boldsymbol{\Psi}_{\mathbf{nn},f} \mathbf{w}_f$
which can be transformed to ordinary eigenvalue problem by Cholesky factorization: $\boldsymbol{\Psi}_{\mathbf{nn},f} = \mathbf{L}_f \mathbf{L}_f^{\mathsf{H}}$

$$\left(\mathbf{L}_f^{-1} \boldsymbol{\Psi}_{\mathbf{xx},f} \mathbf{L}_f^{-H}\right) \left(\mathbf{L}_f^{H} \mathbf{w}_f\right) = \lambda \left(\mathbf{L}_f^{H} \mathbf{w}_f\right)$$

Solution:

$$\mathbf{w}_f^{\mathrm{maxSNR}} = \mathbf{L}_f^{-H} \mathcal{P}\left(\mathbf{L}_f^{-1} \boldsymbol{\Psi}_{\mathbf{xx},f} \mathbf{L}_f^{-H}\right)$$

(Notation: $\mathcal{P}(\mathbf{A})$: Eigenvector corresponding to largest Eigenvalue of $\mathbf{A}$)

Rank-1 Constraint

Narrowband (rank-1) assumption: $\mathbf{x}_{t,f} = \tilde{\mathbf{a}}_f x_{1,t,f} \Rightarrow \mathbf{\Psi}_{\mathbf{xx},f} = \tilde{\mathbf{a}}_f \tilde{\mathbf{a}}_f^{\mathsf{H}} \sigma^2_{x_1,f}$

Use in SDW-MWF: gives[1]: $\mathbf{w}_f^{\text{r1-SDW-MWF}} = \frac{\mathbf{\Psi}^{-1}_{\mathbf{nn},f} \tilde{\mathbf{a}}_f \tilde{\mathbf{a}}_f^{\mathsf{H}} \sigma^2_{x_1,f}}{\mu + \operatorname{tr}\left\{\mathbf{\Psi}^{-1}_{\mathbf{nn},f} \tilde{\mathbf{a}}_f \tilde{\mathbf{a}}_f^{\mathsf{H}} \sigma^2_{x_1,f}\right\}} \mathbf{u}_1$

With μ=0 we obtain $\mathbf{w}_f^{\text{r1-SDW-MWF-0}} = \frac{\mathbf{\Psi}^{-1}_{\mathbf{nn},f} \tilde{\mathbf{a}}_f}{\tilde{\mathbf{a}}_f^{\mathsf{H}} \mathbf{\Psi}^{-1}_{\mathbf{nn},f} \tilde{\mathbf{a}}_f} = \mathbf{w}^{\text{MVDR}}$

Enforcing rank-1 constraint on maxSNR beamformer gives

$$\begin{aligned} \mathbf{w}_f^{\text{maxSNR}} &= \mathbf{L}_f^{-H} \mathcal{P}\left(\mathbf{L}_f^{-1} \tilde{\mathbf{a}}_f \tilde{\mathbf{a}}_f^{\mathsf{H}} \sigma^2_{x_1,f} \mathbf{L}_f^{-H}\right) = \mathbf{L}_f^{-H} \mathbf{L}_f^{-1} \tilde{\mathbf{a}}_f \\ &= \mathbf{\Psi}^{-1}_{\mathbf{nn},f} \tilde{\mathbf{a}}_f \end{aligned}$$

All beamformers point in same direction
and differ only in complex (freq.dep.) constant

[1] employ matrix inversion lemma

Beamforming Criteria: Discussion

- maxSNR beamformer introduces speech distortions, while MVDR does not
  - Can be compensated by postfilter [Warsitz and Haeb-Umbach, 2007]
- There is no unanimous opinion which of the beamformers performs best for enhancement for ASR
  - Advice: try out all of them
- A good estimate of the spatial covariance matrices is more important

How do we estimate the spatial covariance matrix?

- Spatial covariance estimation:

$$\hat{\mathbf{\Psi}}_{\boldsymbol{\nu}\boldsymbol{\nu},f} = \sum_{t=1}^{T} \gamma_{t,f}^{(\nu)} \mathbf{y}_{t,f} \mathbf{y}_{t,f}^{\mathsf{H}} \Big/ \sum_{t} \gamma_{tf}^{(\boldsymbol{\nu})}; \quad \boldsymbol{\nu} \in \{\mathbf{x}, \mathbf{n}\}$$

where: $\gamma_{t,f}^{(x)} = \hat{\Pr}(M_{t,f}^{(x)} = 1|\mathcal{Y})$ speech presence prob. (SPP), speech mask

$\gamma_{t,f}^{(n)} = \hat{\Pr}(M_{t,f}^{(n)} = 1|\mathcal{Y})$ noise presence prob., noise mask

How do we estimate the RTF?

- Estimation of RTF $\tilde{\mathbf{a}}_f$:
  - Solve above (generalized) eigenvalue problem: $\tilde{\mathbf{a}}_f = \mathbf{\Psi}_{\mathbf{nn},f}\mathbf{w}_f^{\text{maxSNR}}$
  - Exploit nonstationarity of speech [Gannot et al., 2001] – not described here

- Advice: use beamformer formulation, which avoids explicit computation of RTF, e.g.,

$$\mathbf{w}_f^{\text{r1-SDW-MWF}} = \frac{\mathbf{\Psi}_{\mathbf{nn},f}^{-1}\mathbf{\Psi}_{\mathbf{xx},f}}{\mu + \text{tr}\left\{\mathbf{\Psi}_{\mathbf{nn},f}^{-1}\mathbf{\Psi}_{\mathbf{xx},f}\right\}}\mathbf{u}_1$$

[Souden et al., 2010]

Summary: processing steps

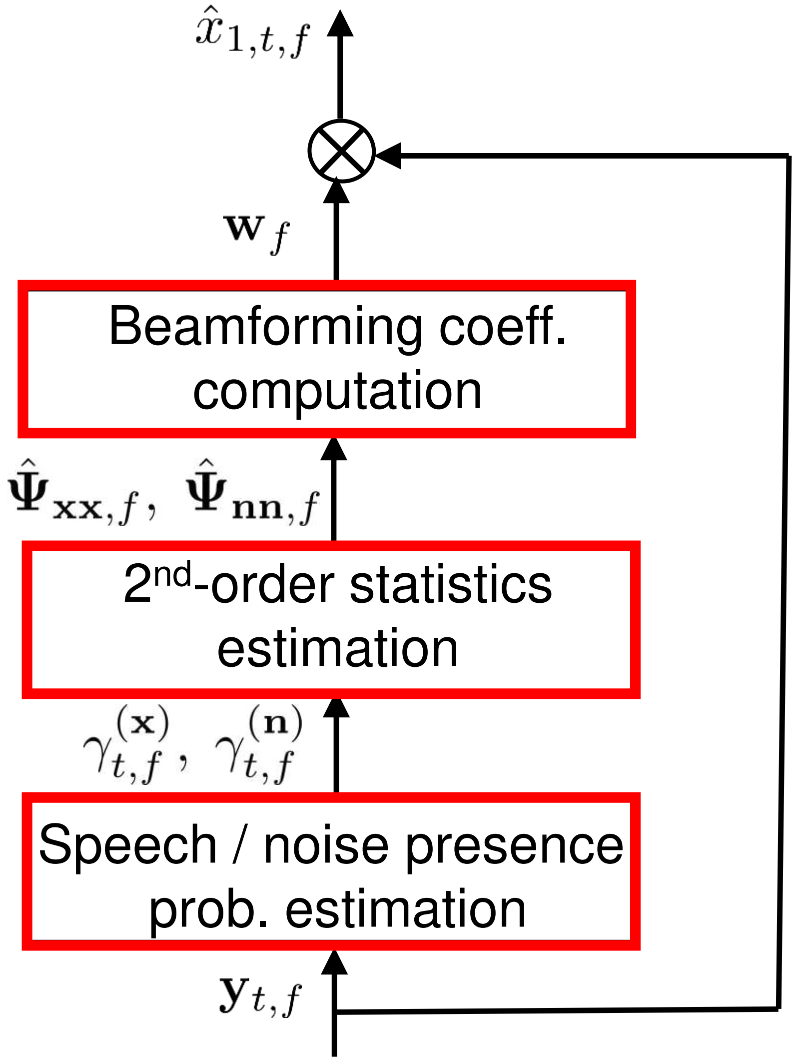
$$\hat{x}_{1,t,f} = \mathbf{w}_f^{\mathsf{H}} \mathbf{y}_{t,f}$$

e.g.: $$\mathbf{w}_f^{\text{r1-SDW-MWF}} = \frac{\hat{\mathbf{\Psi}}_{\mathbf{nn},f}^{-1} \hat{\mathbf{\Psi}}_{\mathbf{xx},f}}{\mu + \operatorname{tr}\left\{\hat{\mathbf{\Psi}}_{\mathbf{nn},f}^{-1} \hat{\mathbf{\Psi}}_{\mathbf{xx},f}\right\}} \mathbf{u}_1$$

$$\hat{\mathbf{\Psi}}_{\mathbf{xx},f} = \sum_t \gamma_{t,f}^{(\mathbf{x})} \mathbf{y}_{tf} \mathbf{y}_{tf}^{\mathrm{H}} \Big/ \sum_t \gamma_{tf}^{(\mathbf{x})}$$

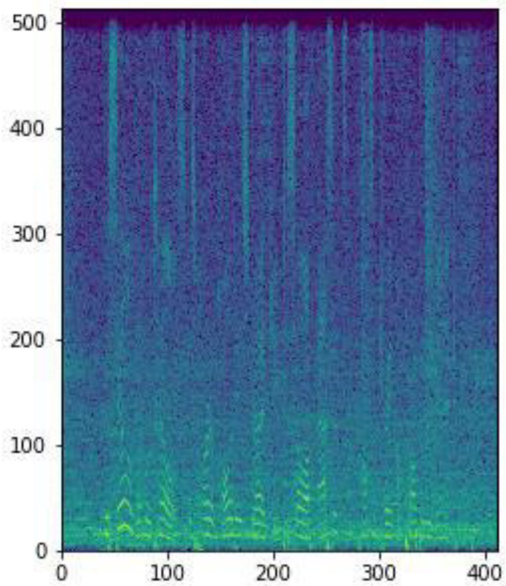
$$\hat{\mathbf{\Psi}}_{\mathbf{nn},f} = \sum_t \gamma_{t,f}^{(\mathbf{n})} \mathbf{y}_{tf} \mathbf{y}_{tf}^{\mathrm{H}} \Big/ \sum_t \gamma_{tf}^{(\mathbf{n})}$$

to be discussed next!



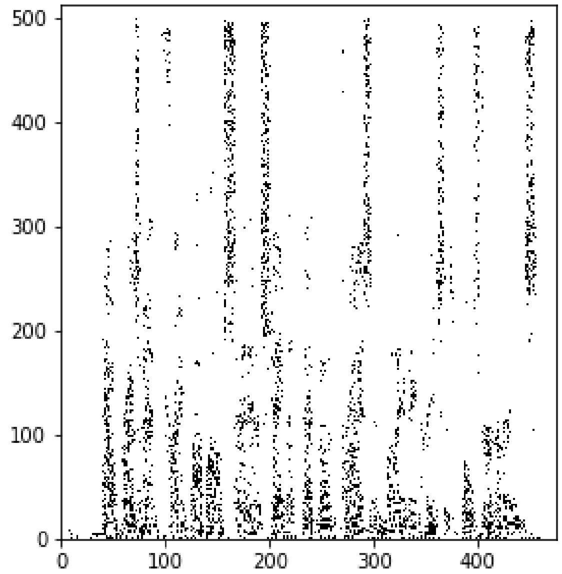
Speech Presence Probability (SPP) / mask estimation

Given:

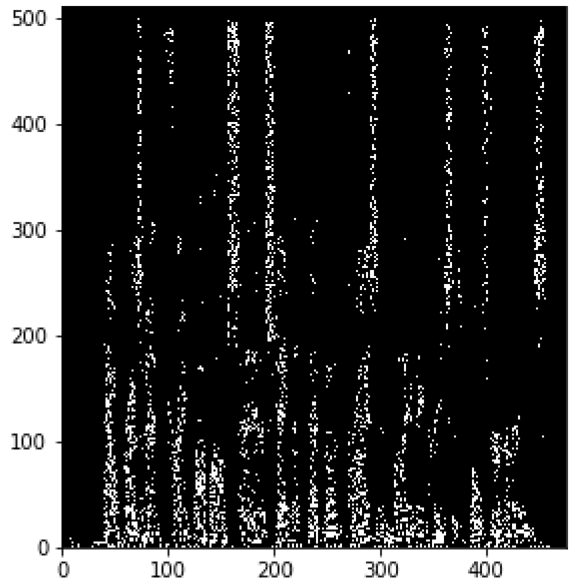


$\mathbf{y}_{t,f}$

Wanted:



$\gamma_{t,f}^{(\mathbf{x})}$



$\gamma_{t,f}^{(\mathbf{n})}$

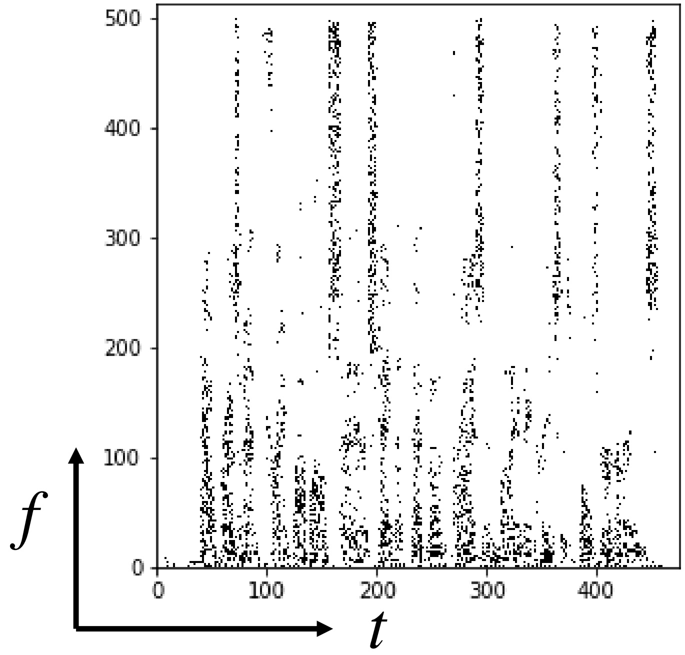
- Estimate for each tf-bin, the probability that it contains speech and the probability that it contains noise, using
  - spatial information
  - or spectral information
  - or both

Options for SPP estimation

- ~~Hand-crafted spectro-temporal smoothing~~
- Spatial mixture models
- Neural networks

Spatial mixture model

- Sparsity assumption [Yilmaz and Rickard, 2004]
  - 90% of the speech power is concentrated in 10% of the tf-bins
  - sparsity most pronounced for STFT window lengths of approx 64 ms



$$M_{t,f} := M_{t,f}^{(x)} = 1 - M_{t,f}^{(n)} \in \{0,1\}$$

$$\gamma_{t,f}^{(i)} := \hat{\Pr}(M_{t,f} = i | \mathbf{y}_{t,f}); i \in \{0,1\}$$

- Mixture model for vector of microphone signals $\mathbf{y}_{t,f}$ or for representation derived from it

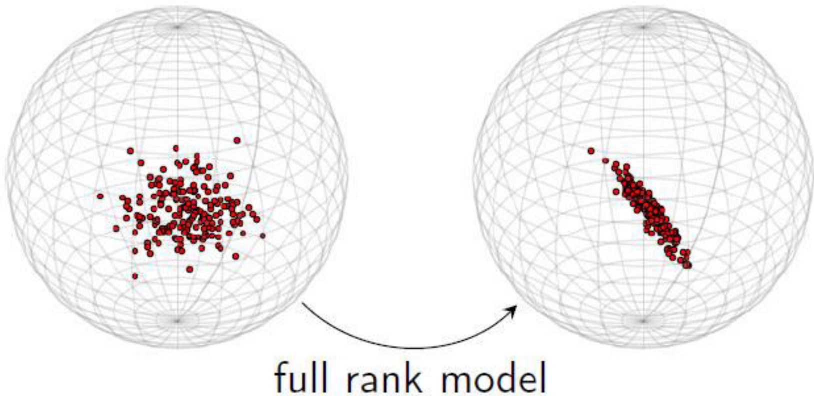
$$p(\mathbf{y}_{t,f}) = \sum_{i=0}^{1} \Pr(M_{t,f} = i) p(\mathbf{y}_{t,f} | M_{t,f} = i)$$

Example spatial mixture model

- Complex angular central Gaussian (cACG) Mixture Model for normalized observation vector $\tilde{\mathbf{y}}_{t,f} = \mathbf{y}_{t,f}/\|\mathbf{y}_{t,f}\|$ [Ito et al., 2016]:

$$p(\tilde{\mathbf{y}}_{t,f}) = \sum_{i=0}^{1} \Pr(M_{t,f} = i) p(\tilde{\mathbf{y}}_{t,f} | M_{t,f} = i) = \sum_i \pi_f^{(i)} \mathrm{cACG}(\tilde{\mathbf{y}}_{t,f}; \mathbf{B}_f^{(i)})$$

$$\mathrm{cACG}(\tilde{\mathbf{y}}_{t,f}; \mathbf{B}_f^{(i)}) = \frac{(M-1)!}{2\pi^M \det \mathbf{B}_f^{(i)}} \frac{1}{(\tilde{\mathbf{y}}_{t,f}^{\mathsf{H}} (\mathbf{B}_f^{(i)})^{-1} \tilde{\mathbf{y}}_{t,f})^M}$$

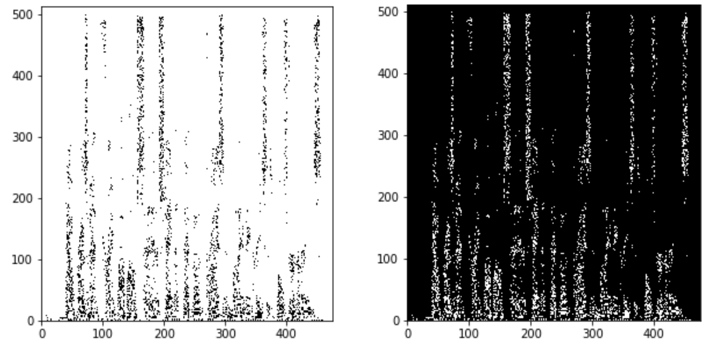


Parameter estimation

- Parameter Estimation via Expectation Maximization (EM) alg.
  - E-step: estimate source activity indicator $\gamma_{t,f}^{(i)}$ for all t, f and $i = 0, 1$
  - M-step: estimate model parameters: $\pi_f^{(i)}, \mathbf{B}_f^{(i)};\ i \in \{0, 1\}$
  - Iterate until convergence
- Actually, we are only interested in $\gamma_{t,f}^{(i)}$

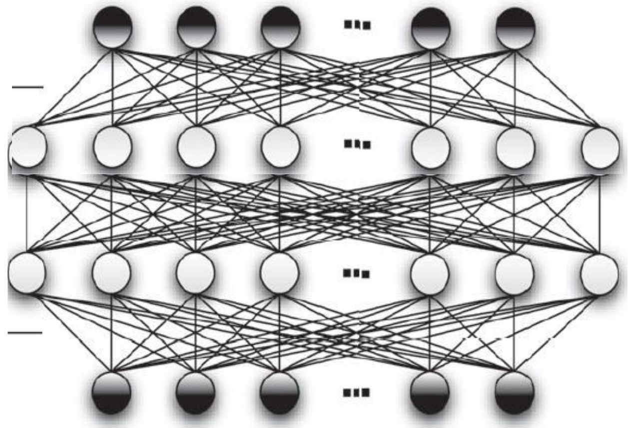
Note: separate EM for each frequency causes frequency permutation problem:
In one frequency i=1 may stand for speech, in another for noise!
Permutation solver required, e.g. [Sawada et al., 2011]
(or use permutation-free model with time-variant mixture weights [Ito et al., 2013])

SPP estimation with neural network

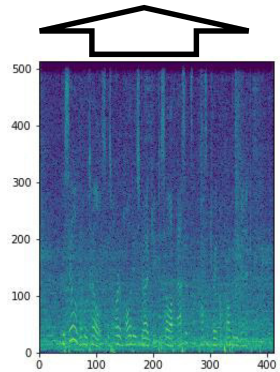


- SPP as supervised learning problem
  - Mask estimation formulated as classification problem
  - Objective function: binary cross entropy:

$$J(\theta) = -\sum_{\nu \in \{x,n\}} \sum_{t,f} \left(M_{t,f}^{(\nu)} \log \gamma_{t,f}^{(\nu)}(\theta) + (1 - M_{t,f}^{(\nu)}) \log(1 - \gamma_{t,f}^{(\nu)}(\theta)) \right)$$



- Note: masks need not sum up to one!



PADERBORN UNIVERSITY NTT

Example configuration

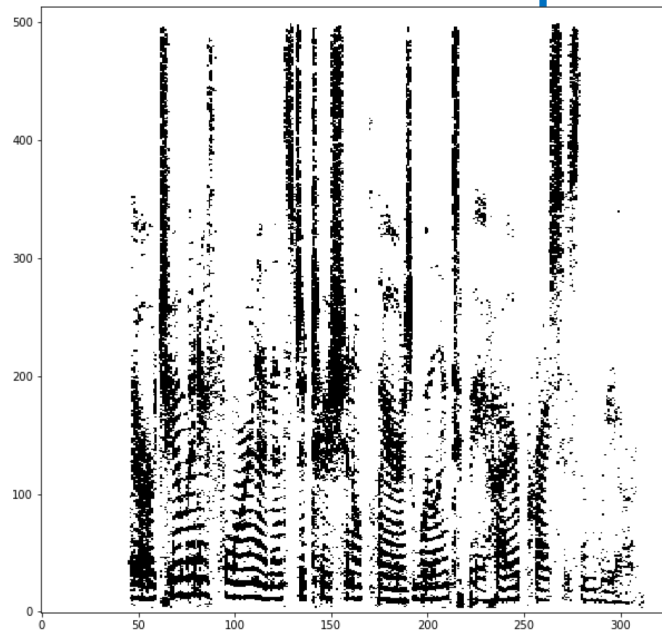
- Input: spectral magnitudes $|\mathbf{y}_{t,f}|$

Layer	Units	Type	Non-linearity	$p_{dropout}$
L1	256	BLSTM	Tanh	0.5
L2	513	FF	ReLU	0.5
L3	513	FF	ReLU	0.5
L4	1026	FF	Sigmoid	0.0

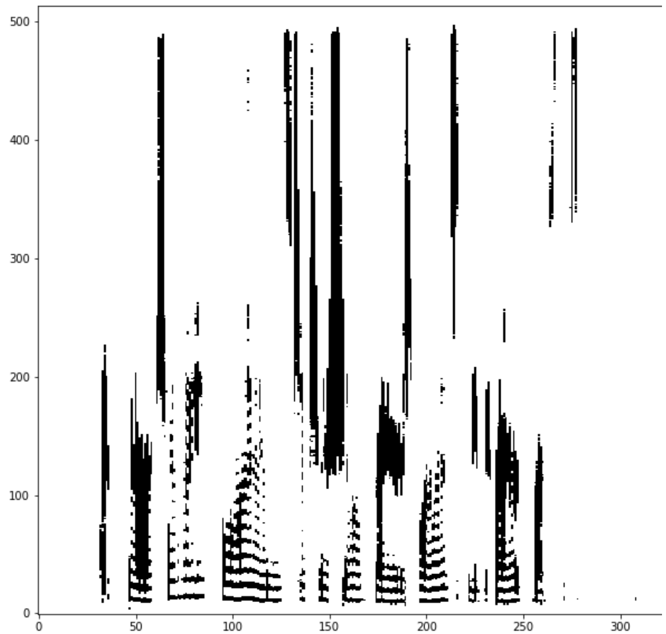
- Output: speech and noise masks $\gamma_{t,f}^{(x)}, \gamma_{t,f}^{(n)}$

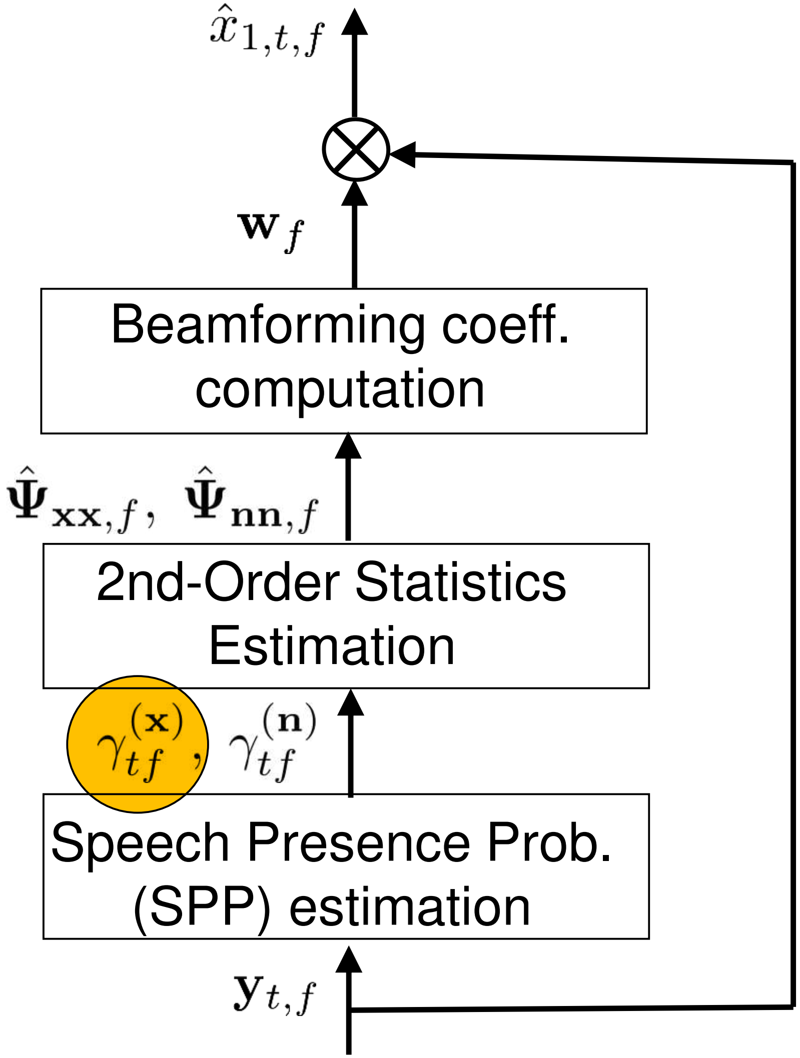
Example masks

Target speech mask $M_{t,f}^{(x)}$



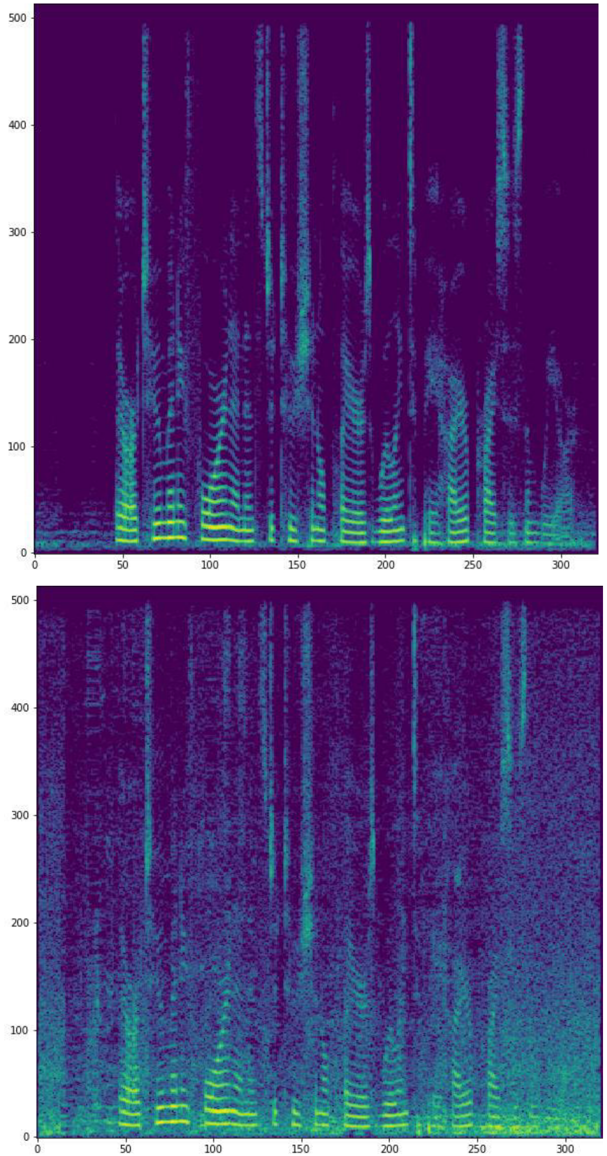
Estimated speech mask $\gamma_{t,f}^{(x)}$

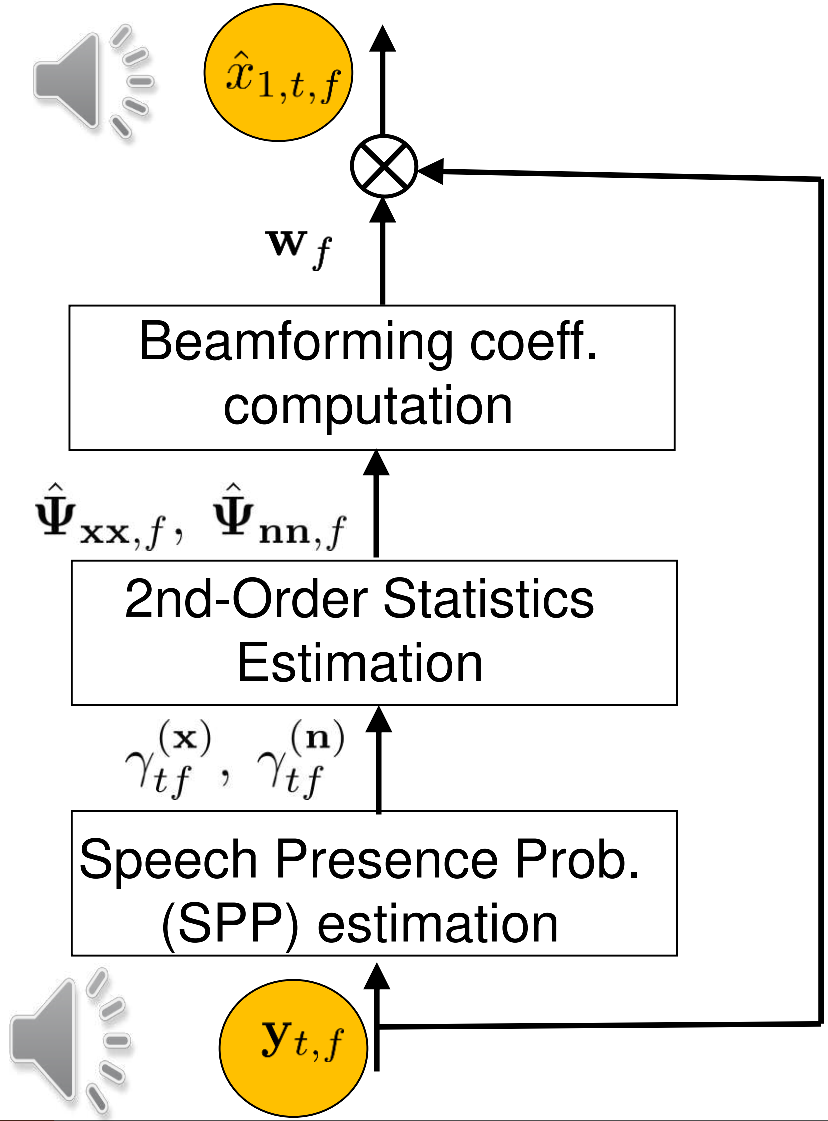




Demonstration NN-based mask estimation

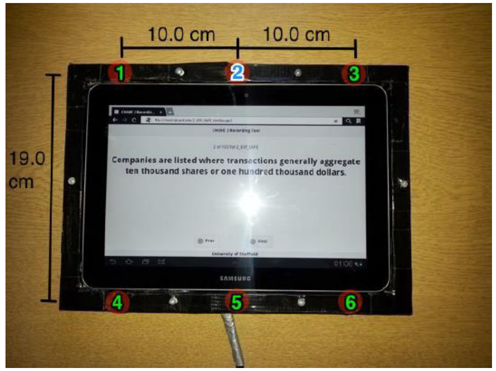
CHiME-3: Utterance ID:
f04_051c0112_str





ASR results: Spatial mixture model mask estimation

- CHiME-3 (2015) [Barker et al., 2017]
  - WSJ utterances
  - „Fixed“ speaker positions
  - Low reverberation
  - Noisy environment: bus, café, street, pedestrian
  - Trng set size: 18 hrs x 6 channels



- The winning system [Yoshioka et al., 2015, Higuchi et al., 2016] used a cACGMM spatial mixture model:

WER [%]	Dev Real	Test Real
No beamforming	9.0	15.6
DSB with DoA estimation	9.4	16.2
Spatial mixture model	4.8	8.9

ASR results: Neural network mask estimation

- CHiME-3 [Heymann et al., 2015]
  - Absolute WER values not comparable with last slide (different acoustic model, language model, data augmentation)

WER [%]	**Dev Real**	**Test Real**
No beamforming	18.7	33.2
NN supported beamforming	10.4	16.5

- CHiME-4 (2016):
  - All top 5 systems used mask-based beamforming (either NN or spatial mixture model)

Extensions

- Spatial mixture models
  - Other mixture models, e.g., Watson MM [Tran Vu and Haeb-Umbach, 2010]
  - On test utterance, with NN-based masks as priors $\Pr(M_{t,f} = i)$ [Nakatani et al., 2017]

- NN-Supported Beamforming
  - Cross-channel features, e.g., [Liu et al., 2018]
  - Block-online processing, e.g., [Boeddeker et al., 2018]
  - Used for dereverberation [Heymann et al., 2017b]

Pros and cons of two mask estimation methods

	Spatial mixture models	Neural networks
Spatial characteristics modeling	• **Strong**	• Moderate (use of cross-channel features at input)
Spectro-temporal characteristics modeling (for speech)	• Weak - Permutation problem • No concept of human speech (pros and cons)	• **Very strong** - Strong speech model based training
#channels required	• Multi-channel	• Single channel
Leverage training data	• No training phase	• **Yes**, but parallel data required
Adaptation to test condition	• **Strong** - Unsupervised learning applicable	• Weak - Poor generalization - Sensitive to mismatch

Table of contents in part II

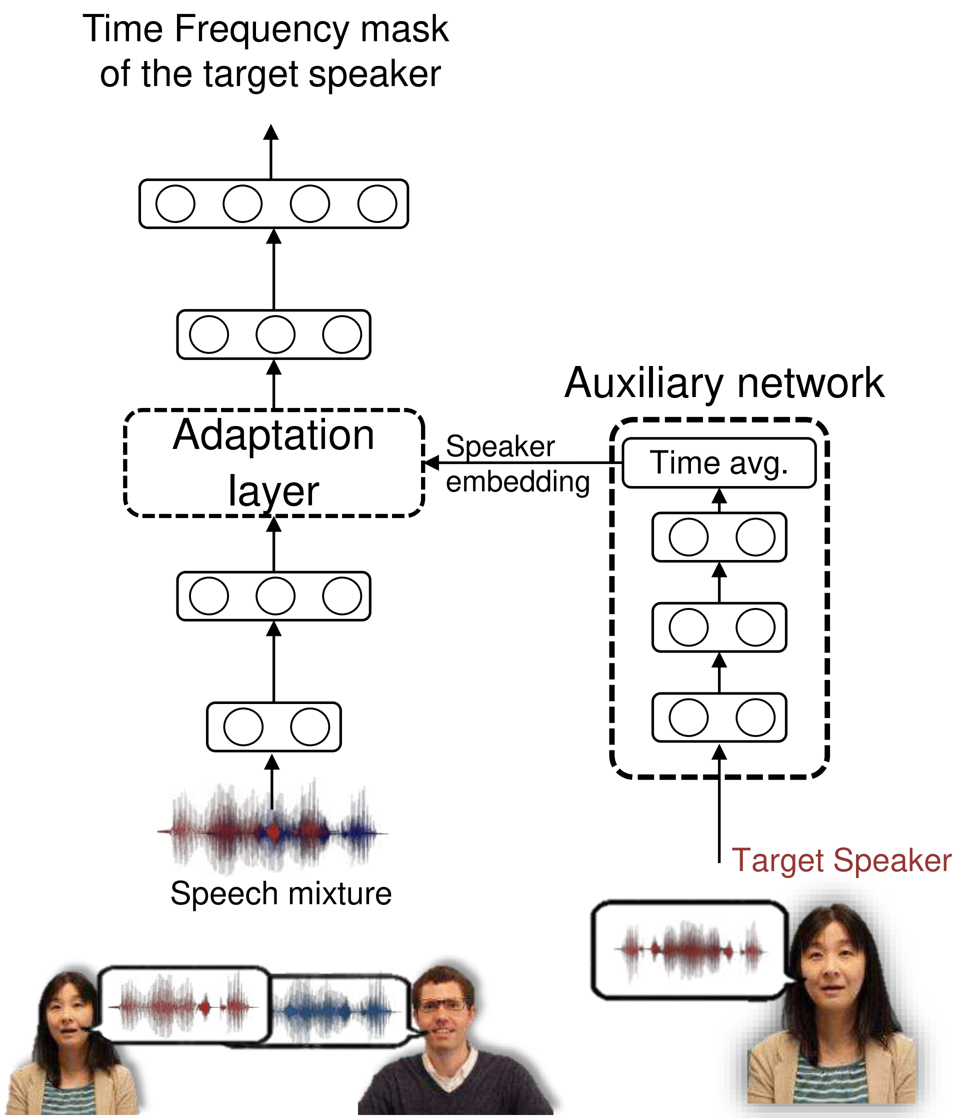
Speaker-Conditioned Spectrogram Masking

- In many application, we may be interested in recognizing speech from a target speaker even if there is noise or other people speaking, e.g., smart speaker

→ Target speaker extraction
  - Known target speaker position → use beamformer to extract speech from that direction
  - Unknown target speaker position → extract speaker based on his/her speech characteristics (SpeakerBeam)

- Idea of SpeakerBeam
  - NN for mask estimation can well discriminate a target speaker from noise, but not when interference is another speaker
  - This can be improved if the mask estimator is informed about the speaker to be extracted
  - We assume that we have about 10 sec. of enrollment/adaptation utterance spoken by the target speaker

SpeakerBeam [Zmolikova et al., 2017]



- Adaptation layer
  - Drive NN to output mask for the target speaker only, given target speaker embedding
  - Different implementations possible, e.g. factorized layer, scaling, etc.
- Auxiliary network
  - Compute speaker embedding given the enrollment/adaptation utterance
  - Implemented using sequence summary network [Vesely et al. 2016]
  - Jointly optimized with mask estimation NN
- SpeakerBeam performs 1ch processing to compute mask, but it can be combined with beamforming for multi-ch processing

Results [Zmolikova et al., 2019]

- WSJ2mix-MC
  - Artificial 2-speaker mixtures from WSJ utterances
  - 1ch no reverberation
  - 8 channels with reverberation T_{60} = 0.2 – 0.6 s

WER [%]	1 ch (no reverb)	8 ch (w/ reverb)
Single speaker	12.2	16.2
Mixtures	73.4	85.2
SpeakerBeam (1ch)	30.6	-
SpeakerBeam + Beamformer	-	22.5
SpeakerBeam + Beamformer (w/ AM joint training)	-	20.7

Software

- Spatial mixture models: https://github.com/fgnt/pb_bss
  - Different spatial mixture models
    - complex angular central Gaussian , complex Watson,von-Mises-Fisher
  - Methods: init, fit, predict
  - Beamformer variants
  - Ref: [Drude and Haeb-Umbach, 2017]

- NN supported acoustic beamforming: https://github.com/fgnt/nn-gev
  - NN-based mask estimator and maxSNR beamformer
  - Ref: [Heymann et al., 2016]
  - Part of Kaldi CHiME-3 baseline

Summary of part II

- Acoustic beamforming as a front-end for ASR
  - Exploits spatial information present in multi-channel input for noise suppression, which typical ASR feature sets (log-mel, cepstral) ignore
  - Leads to significant WER improvements
- SPP / Mask estimation is key component of beamformer
  - Both, spatial mixture models and neural networks are powerful mask estimators with complementary strengths
- Acoustic beamforming followed by DNN-based ASR is a typical representative of a combination of signal processing approaches with deep learning
  - Leads to interpretable, lightweight system compared to a NN with multi-channel input

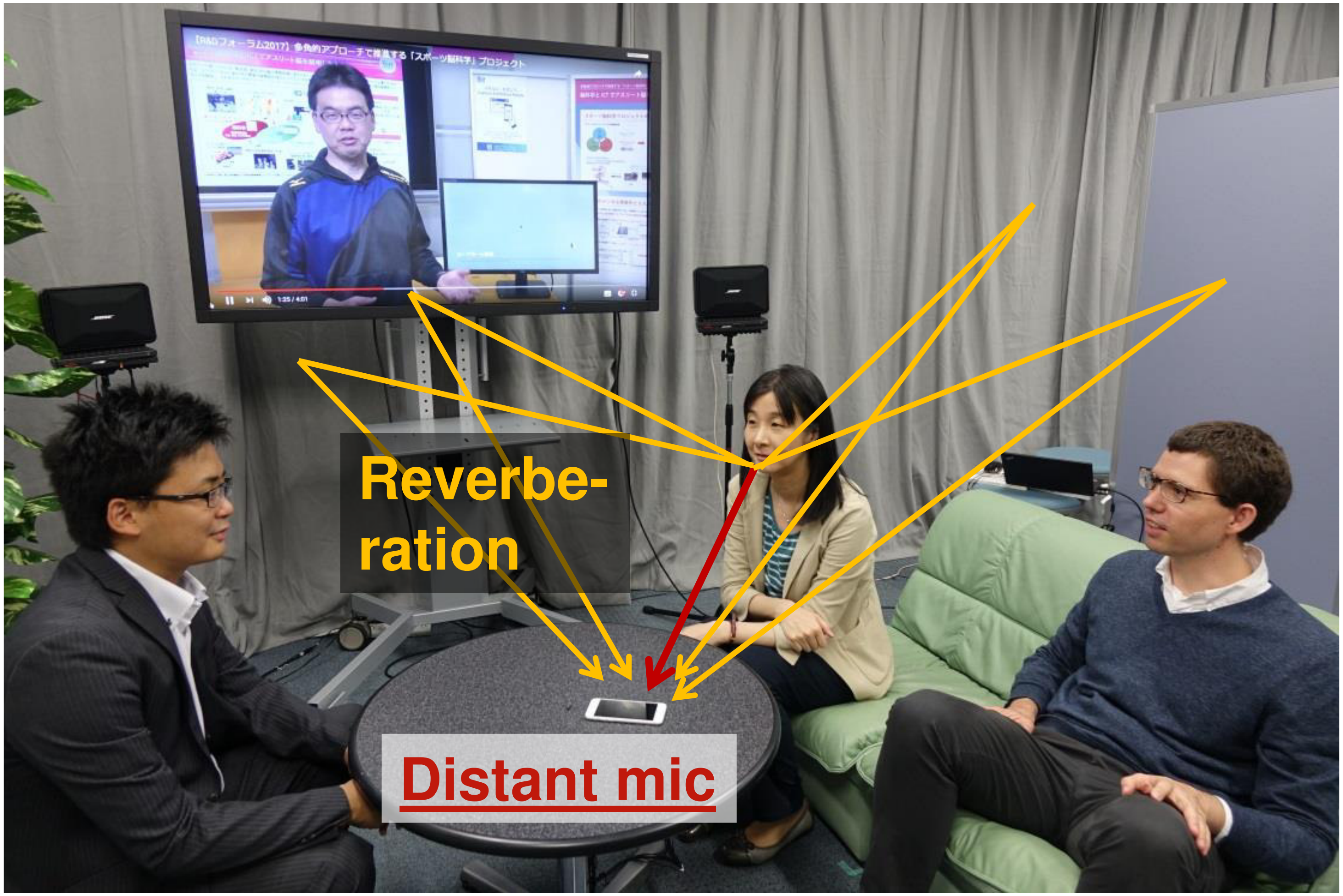
But what about overall optimality? We‘ll come back to that…

Table of contents

Part III.
Dereverberation

Tomohiro Nakatani

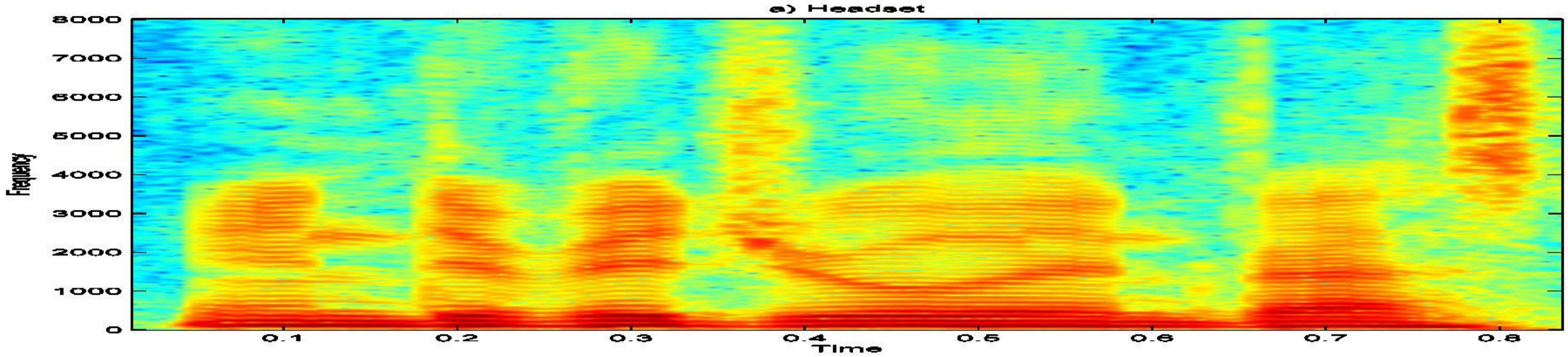
Speech recording in reverberant environments



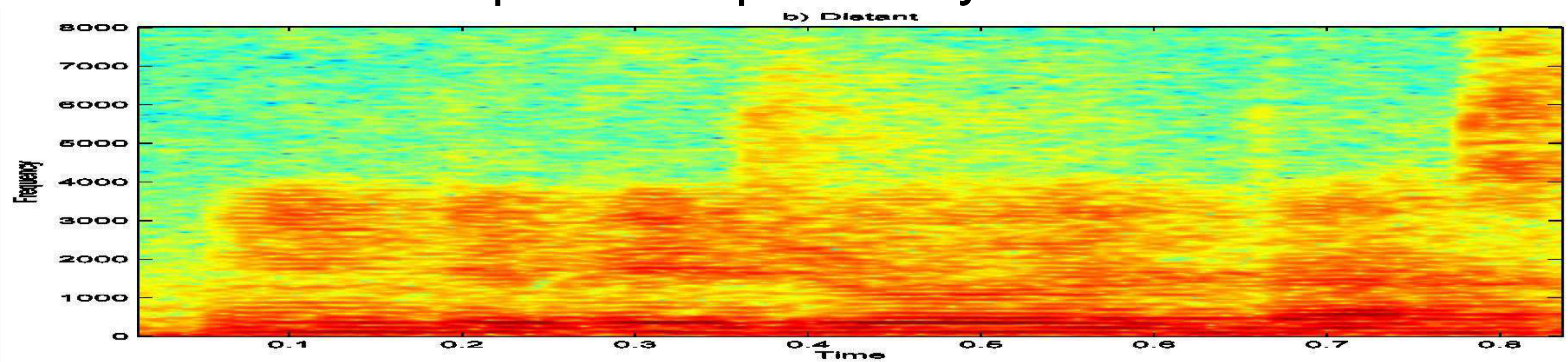
Dereverberation is needed to enhance the quality of recorded speech by reducing reverberation included in it

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

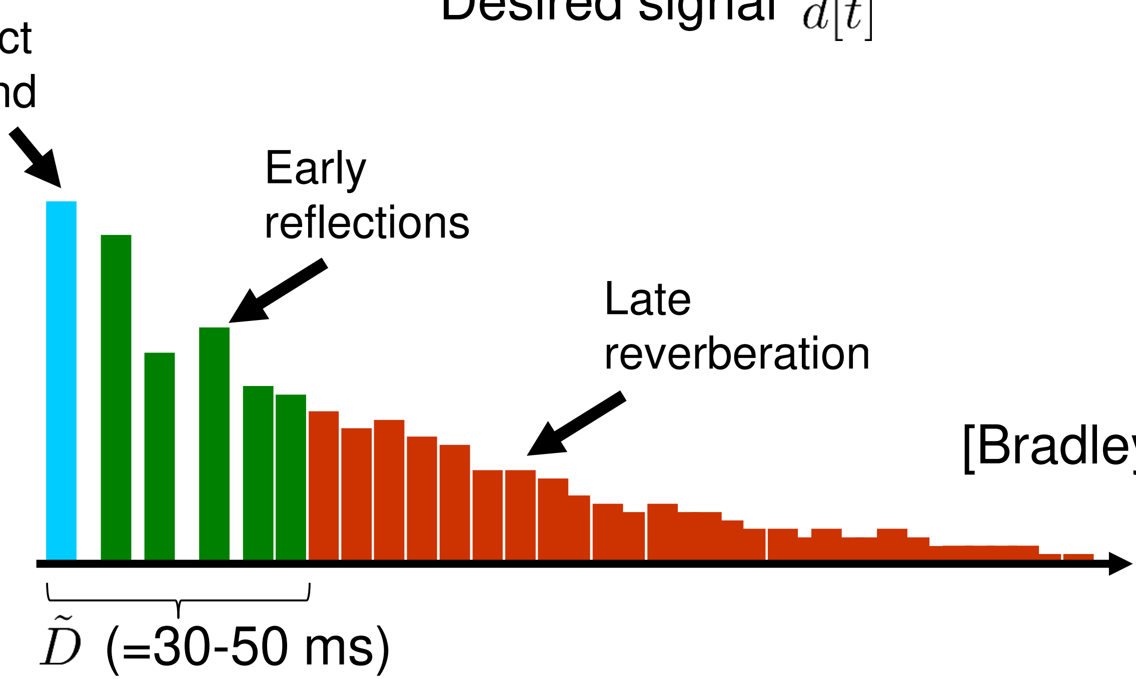
Goal of dereverberation: time domain

Preserve | **Reduce**

Reverberant speech
$$x[\tilde{t}] = \sum_{\tilde{\tau}=0}^{\tilde{L}-1} a[\tilde{\tau}]s[\tilde{t}-\tilde{\tau}] = \underbrace{\boxed{\sum_{\tilde{\tau}=0}^{\tilde{D}-1} a[\tilde{\tau}]s[\tilde{t}-\tilde{\tau}]}}_{\text{Direct sound + Early reflections}} + \underbrace{\boxed{\sum_{\tilde{\tau}=\tilde{D}}^{\tilde{L}-1} a[\tilde{\tau}]s[\tilde{t}-\tilde{\tau}]}}_{\text{Late reverberation}}$$

Desired signal $d[\tilde{t}]$ $\qquad$ $r[\tilde{t}]$

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)

STFT domain (1-ch)

$$x_{t,f} = \sum_{\tau=0}^{L-1} a_{\tau,f} s_{t-\tau,f} = \underbrace{\boxed{\sum_{\tau=0}^{D-1} a_{\tau,f} s_{t-\tau,f}}}_{\text{Desired signal}} + \underbrace{\boxed{\sum_{\tau=D}^{L-1} a_{\tau,f} s_{t-\tau,f}}}_{\text{Late reverberation}}$$

STFT domain (multi-ch)

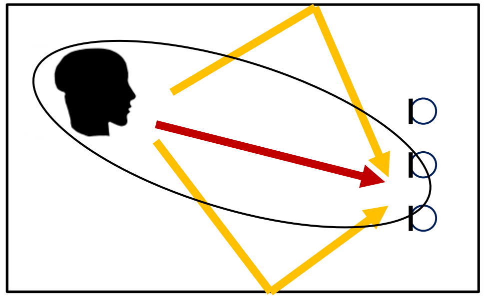
$$\mathbf{x}_{t,f} = \sum_{\tau=0}^{L-1} \mathbf{a}_{\tau,f} s_{t-\tau,f} = \underbrace{\boxed{\sum_{\tau=0}^{D-1} \mathbf{a}_{\tau,f} s_{t-\tau,f}}}_{\mathbf{d}_{t,f}} + \underbrace{\boxed{\sum_{\tau=D}^{L-1} \mathbf{a}_{\tau,f} s_{t-\tau,f}}}_{\mathbf{r}_{t,f}}$$

Convolutional transfer function:

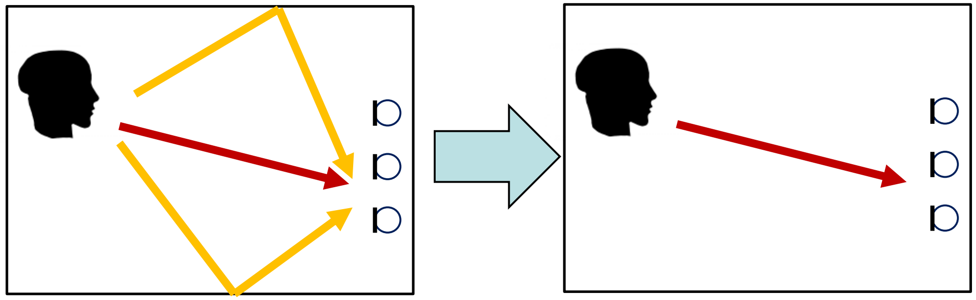
$$\mathbf{a}_{\tau,f} = (a_{1,\tau,f}, a_{2,\tau,f}, \ldots, a_{M,\tau,f})^\top \text{ for } \tau = 0, \ldots, L-1$$

Approaches to dereverberation

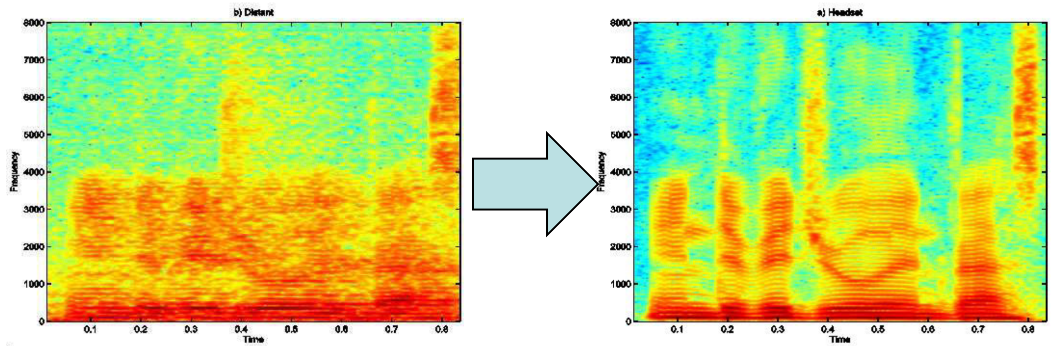
- Beamforming (multi-ch)
  - Enhance desired signal from speaker direction
  - Mostly the same as denoising



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



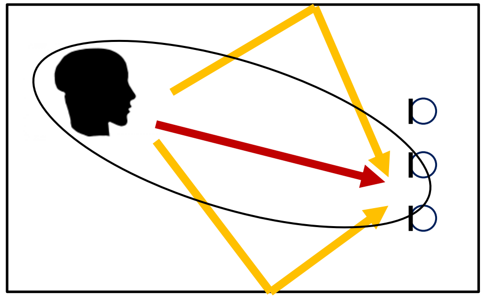
- DNN-based spectral enhancement (1ch)
  - Estimate clean spectrogram
  - Mostly the same as denoising autoencoder



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

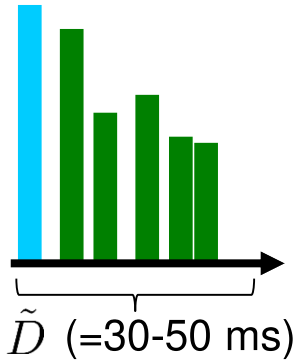
Reverberant Estimated clean

Dereverberation based on beamforming

- Time domain model of desired signal

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

Initial part of the impulse response



STFT

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

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

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

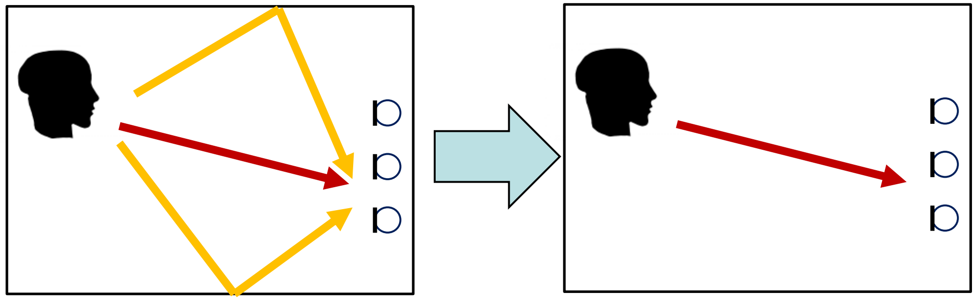
Beamforming is applicable to reduce $\mathbf{r}_{t,f}$

- 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

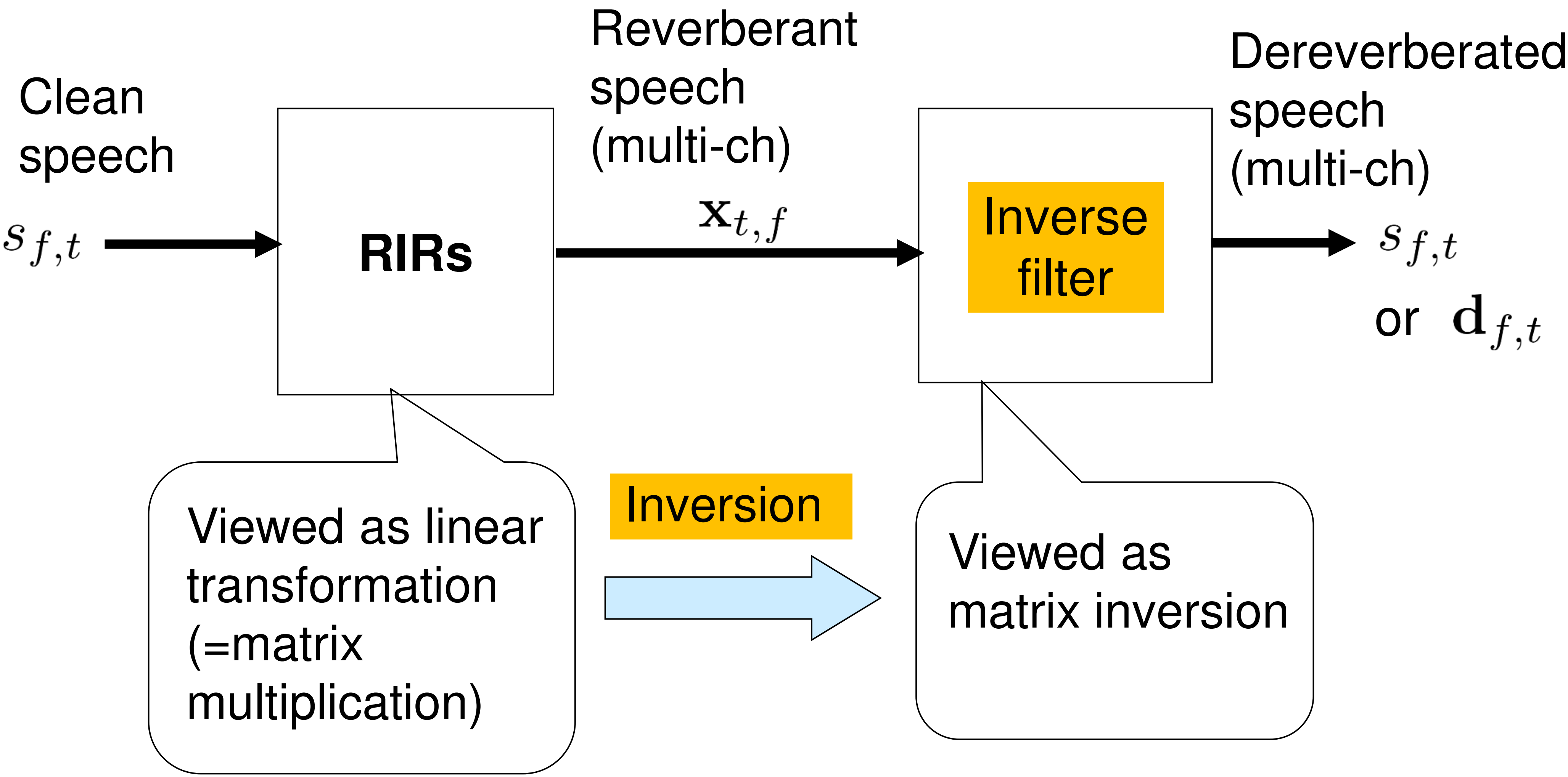
- Beamforming (multi-ch)
  - Enhance desired signal from speaker direction
  - Mostly the same as denoising

- 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

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:

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

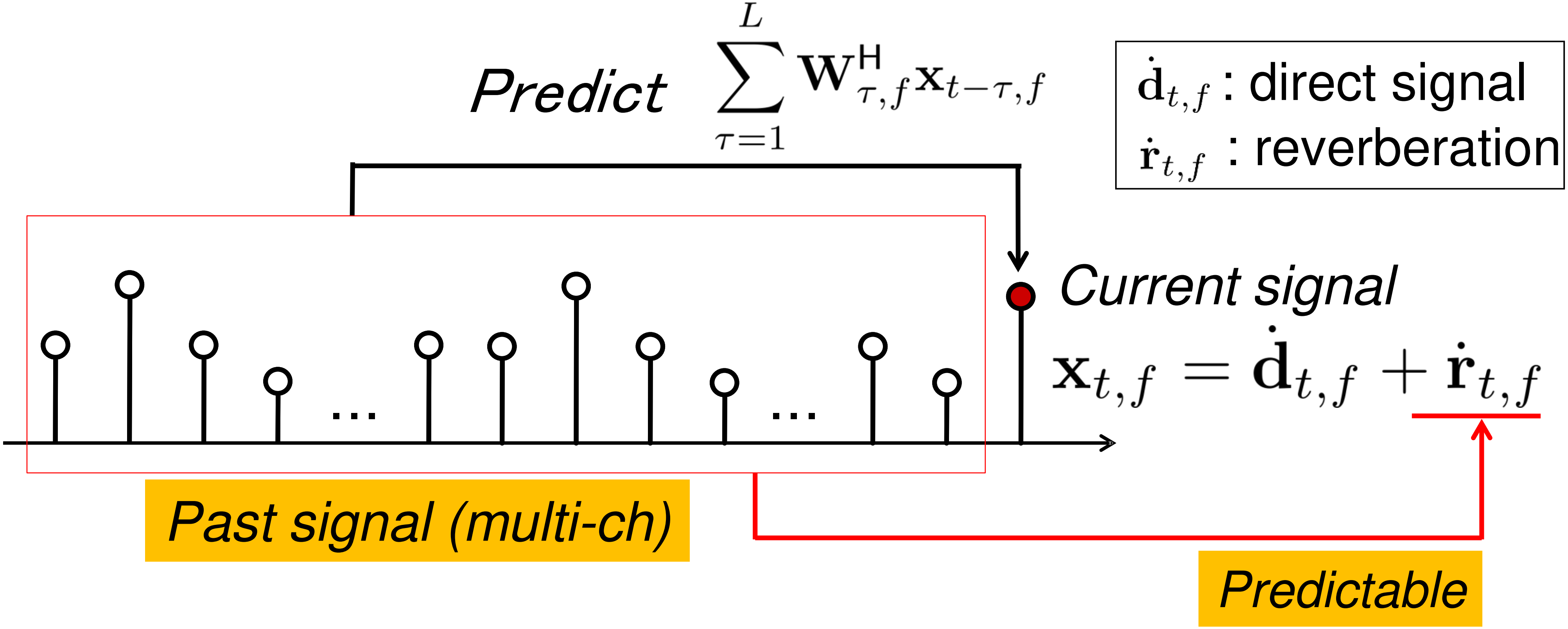
  – When $\mathbf{H}_f$ is full column rank (roughly #mics>1)

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]



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)

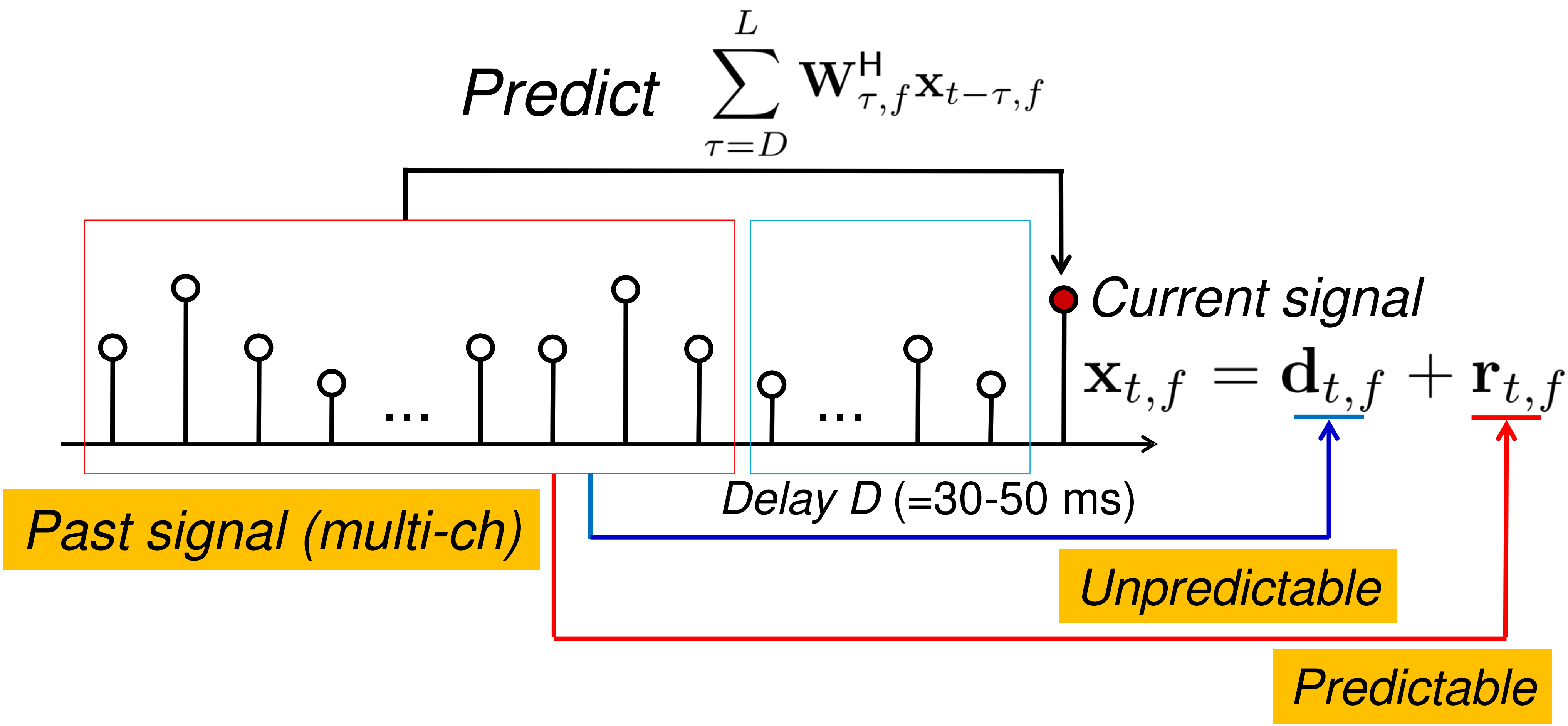
⇒ LP distorts the short-time correlation of speech

  - 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

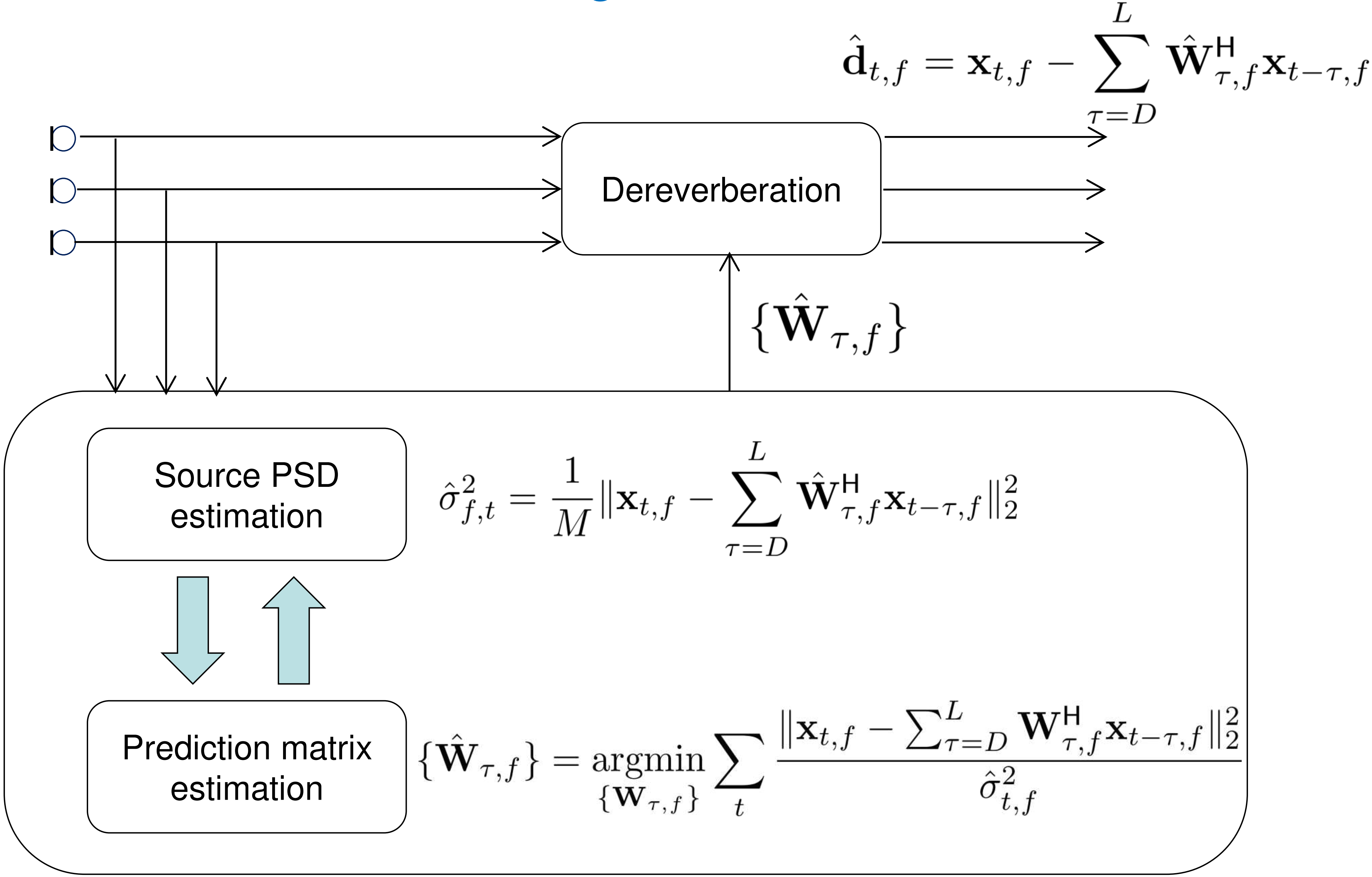
- ML estimation for time-varying Gaussian source

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

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

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

Processing flow of WPE



Why WPE achieves inverse filtering?

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

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

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

Assumption

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

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

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]

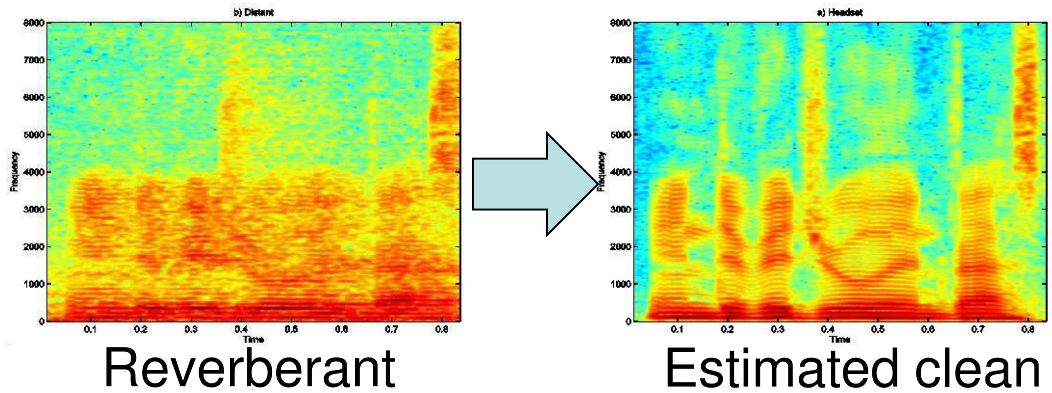
- Frame-by-frame online estimation
  - Recursive least square [Yoshioka et al., 2009], [Caroselli et al., 2017]
  - Kalman filter for joint denoising and dereverberation [Togami and Kawaguchi, 2013], [Braun and Habets, 2018], [Dietzen et al., 2018]

Approaches to dereverberation

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

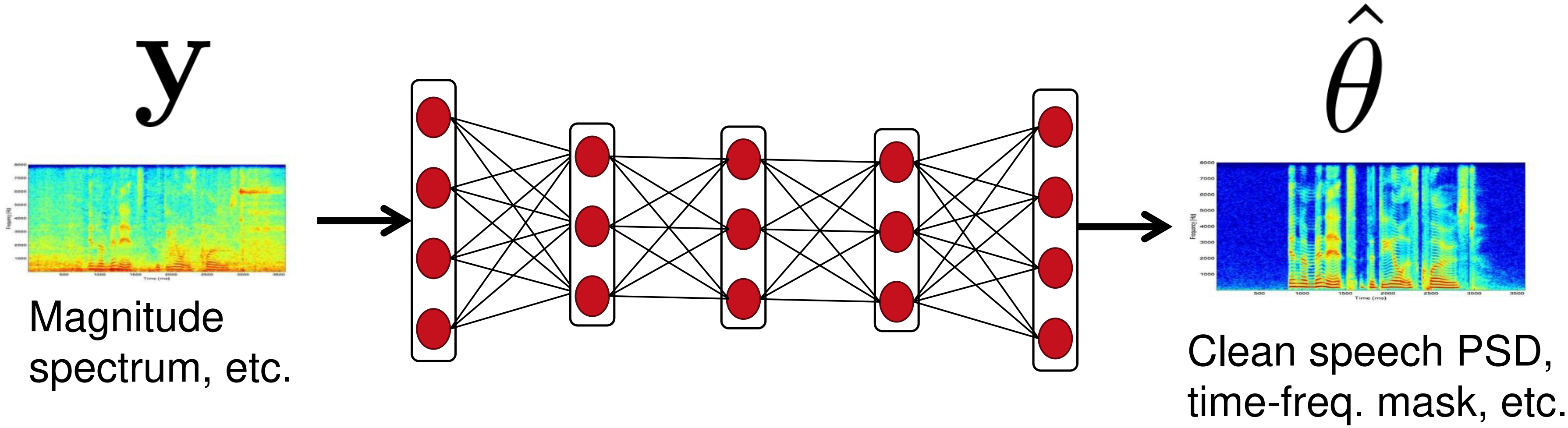
- Blind inverse filtering (multi-ch)
  - Cancel late reverberation
  - (Multi-channel) lnear prediction
    - Weighted prediction error method

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



Neural networks based dereverberation

- Train neural networks based on huge amount of parallel data

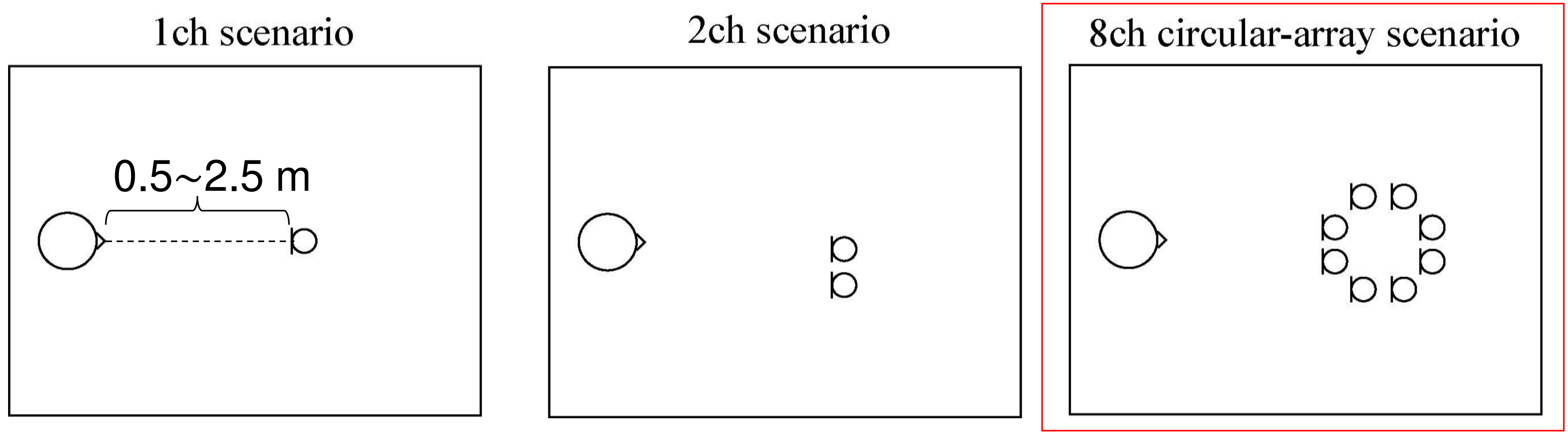


Many variations are proposed depending on tasks (masking/ regression), cost functions, and network structures

[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)

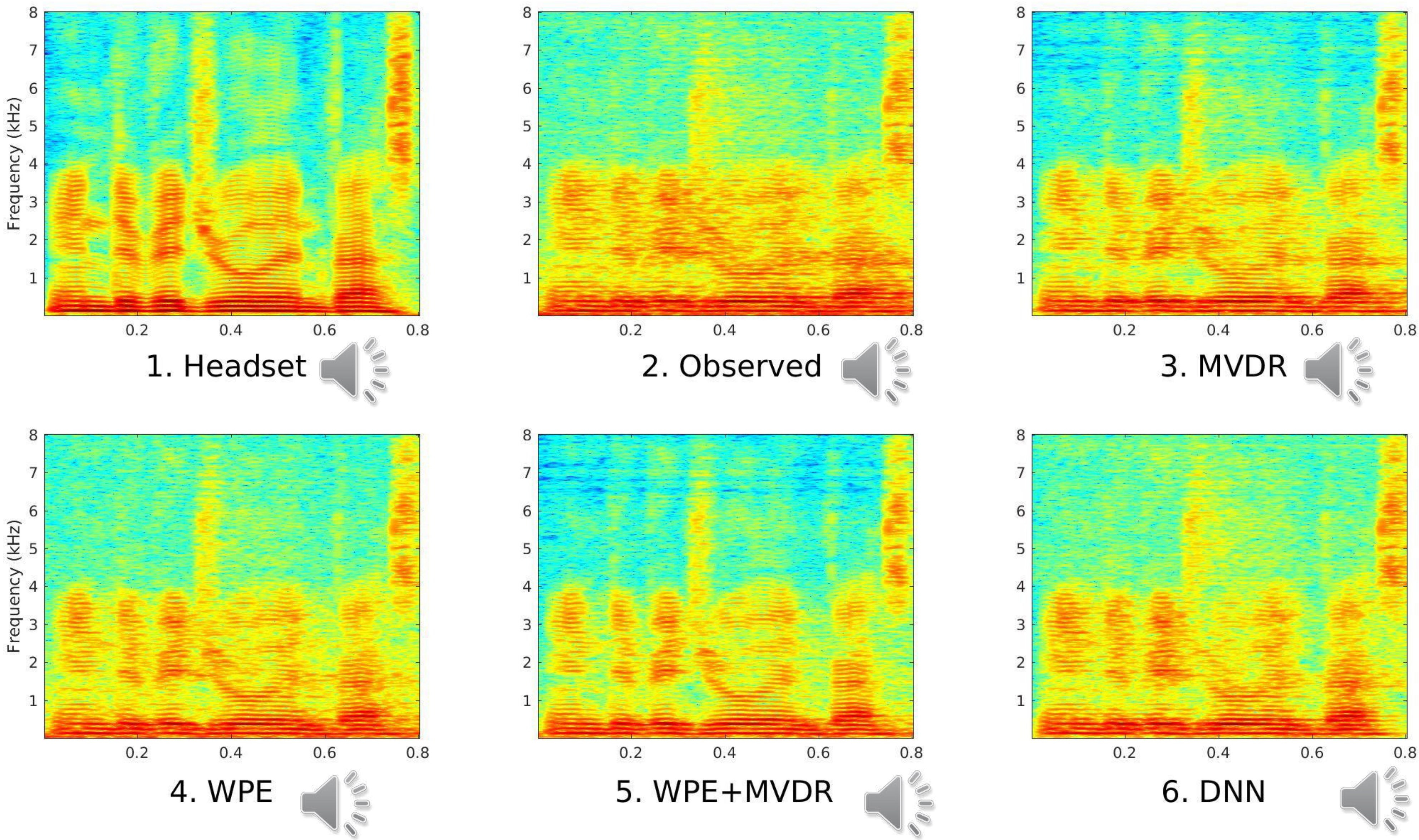


Comparison of three approaches

	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)

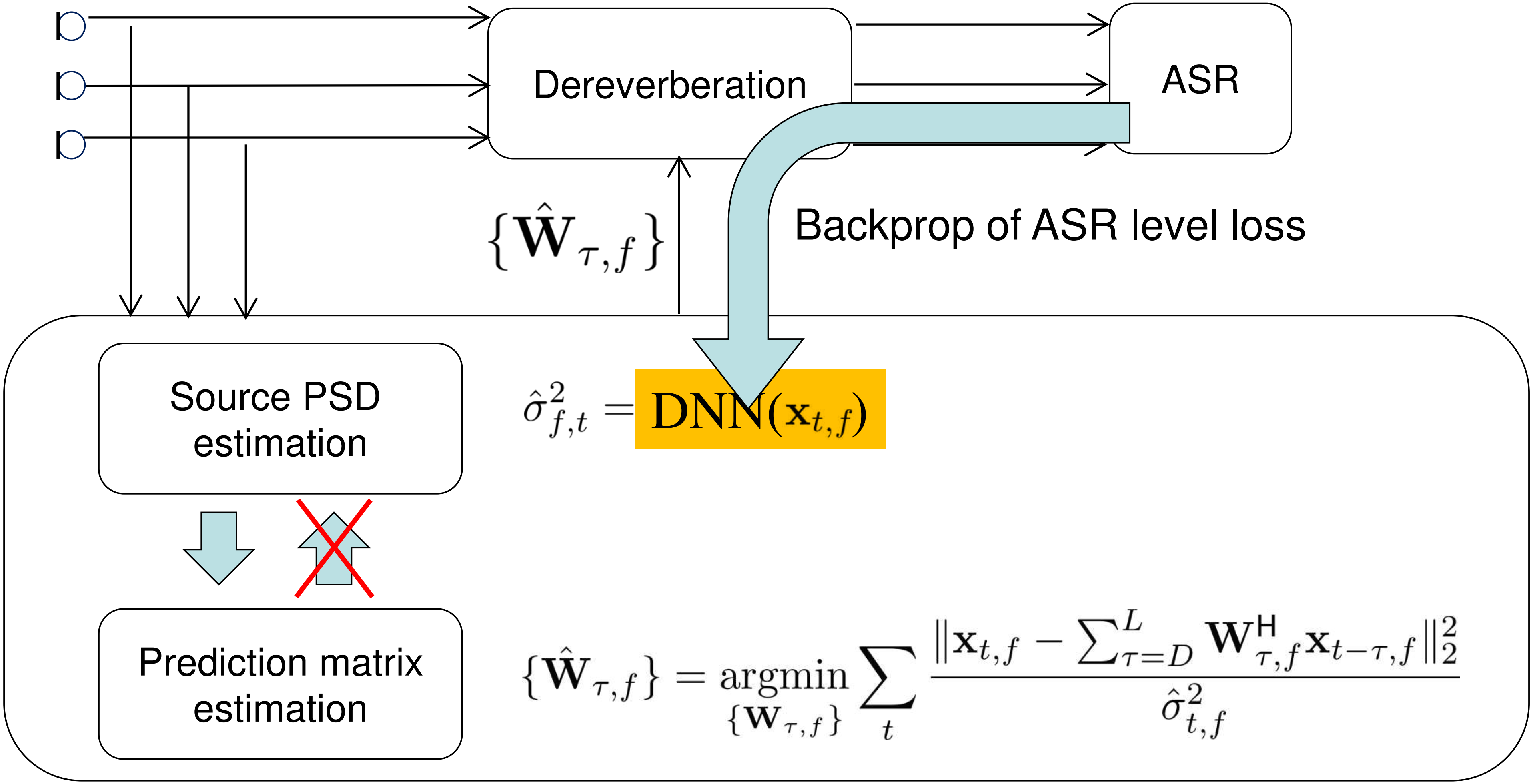
Demonstration



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]

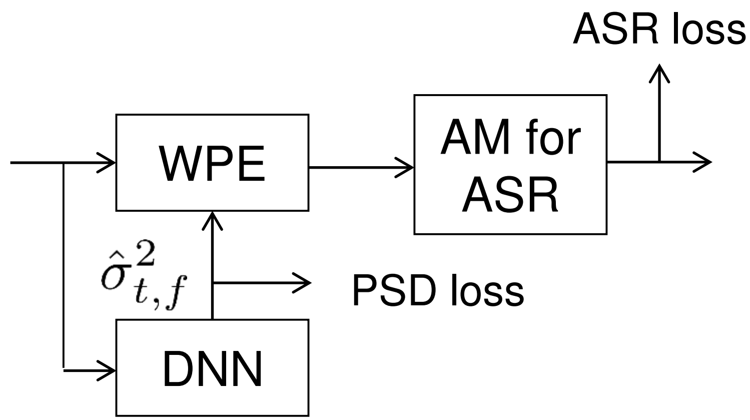


Advantages
1. No iterative estimation → Effective for online processing
2. DNN can be optimized jointly with an ASR system

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

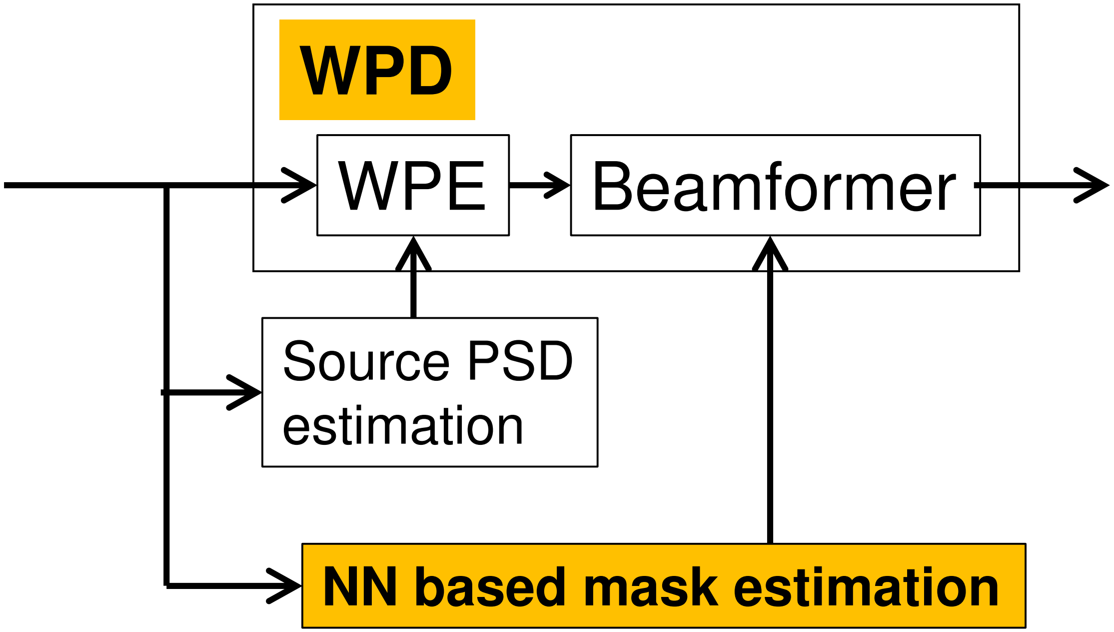


	REVERB (real)		WSJ+VoiceHome	
	Offline	Online	Offline	Online
Unprocessed	17.6		24.3	
WPE	13.0	16.2	18.6	20.0
DNN-WPE (PSD loss)	**10.8**	14.6	18.1	19.3
DNN-WPE (ASR loss)	11.8	**13.4**	**17.7**	**18.4**

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

Part IV.
Source Separation

Reinhold Haeb-Umbach

Problem description

- Known as cocktail party problem [Cherry, 1953]
- Distinguishing speech of different speakers is more difficult than separating speech from noise
- Long history of research

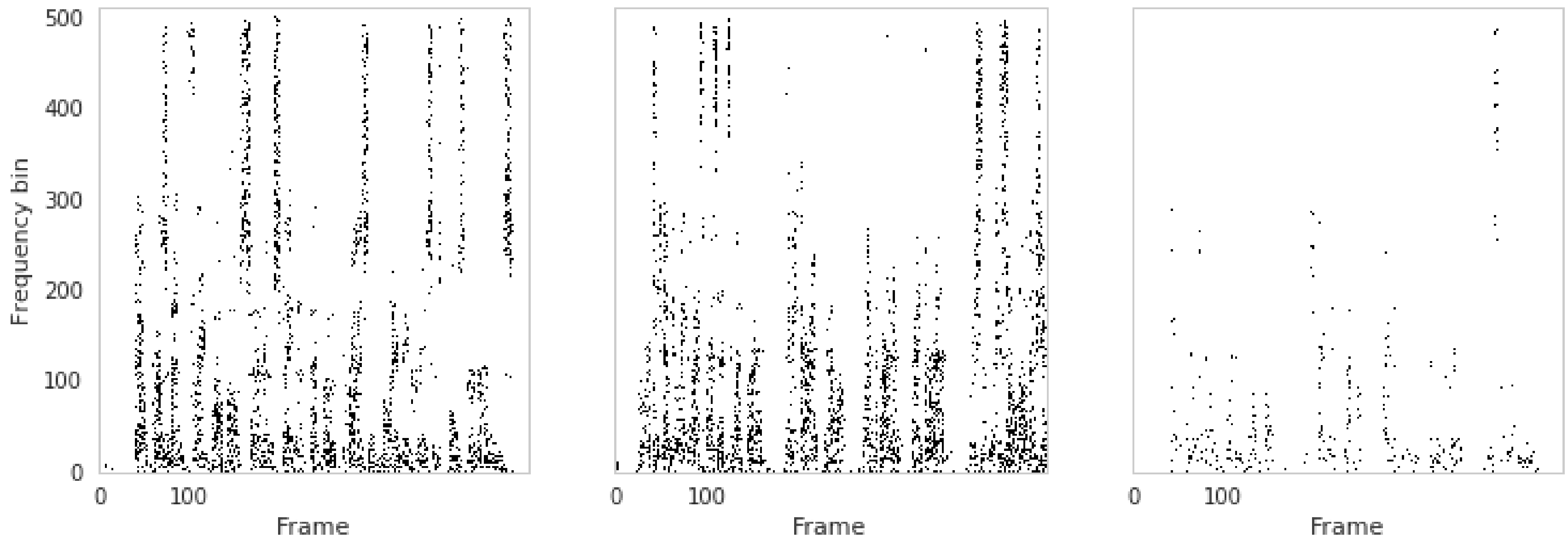
Table of contents in part IV

Blind Source Separation: Taxonomy of Approaches

- ICA (Independent Component Analysis) based
  - Assumption: mutual independence of sources and one or more of the following
    - Non-Gaussianity, non-whiteness, non-stationarity
  - Requires #sensors ≥ #sources
- Sparseness based
  - Assumption: in an appropriate domain, each source does not occupy the whole space, e.g, time-frequency sparseness of speech
  - #sensors can be smaller than #sources
- NMF (Non-negative Matrix Factorization) based
  - Assumption: sources are non-negative and mixing system is additive; sources have low rank
  - Originally single-channel approach, has been extended to multi-channel
- And combinations / variants of them: IVA, ILRMA, IDLMA, …

Here: Blind Speech Separation

- Sparseness based approaches are particularly effective
  - Sparseness of speech in the time-frequency (STFT) domain [Yilmaz and Rickard, 2004]
    - 90% of the speech power is concentrated in 10% of the tf-bins
    - Different speakers populate different tf-bins



Spkr #1 Spkr #2 (Spkr #1) ⊙ (Spkr #2)

BLIND speech separation

Supervised / Guided

- Known mixing system
  - Speaker location
  - Array geometry
  - Acoustic transfer function
- Known diarization
  - On/offset times of speakers
- Known speakers

Blind

- Unknown mixing system
  - Unknown spkr location
  - Unknown array geometry
  - Unknown acoustic transfer function
- Unknown diarization
  - Unknown on/offset times
- Unknown speakers
  - Speaker-independent source separation

Model in STFT domain

- Narrowband assumption
(length of acoustic impulse response << STFT analysis window):

$$\mathbf{y}_{t,f} = \sum_{i=1}^{I} \mathbf{a}_f^{(i)} s_{t,f}^{(i)} + \mathbf{n}_{t,f} =: \sum_{i=1}^{I} \mathbf{x}_{t,f}^{(i)} + \mathbf{n}_{t,f}$$

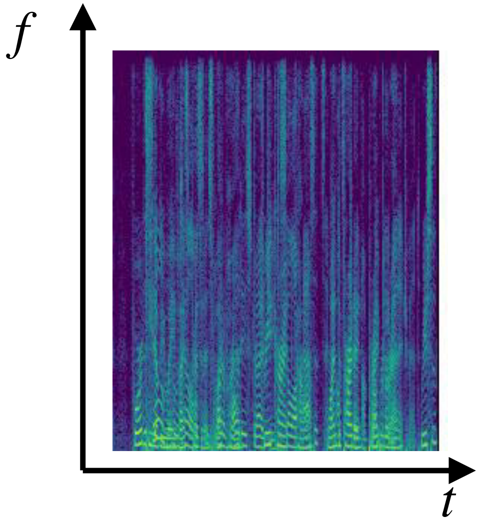
- Often, noise is neglected or treated as an additional source:

$$\mathbf{y}_{t,f} = \sum_{i=1}^{I} \mathbf{x}_{t,f}^{(i)}; \quad \mathbf{y}_{t,f} = \sum_{i=0}^{I} \mathbf{x}_{t,f}^{(i)}$$

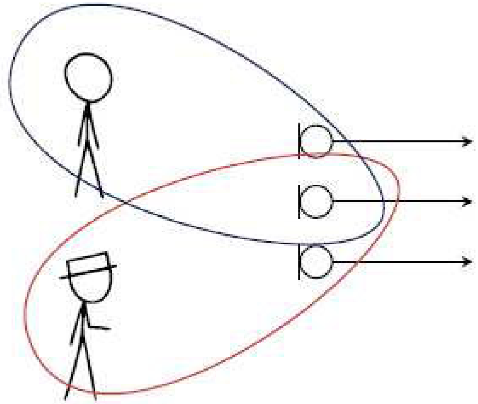
- Our goal is to reconstruct the images of the source signals at a reference microphone (e.g. mic #1):

$$x_{1,t,f}^{(i)}; \ i = 1, \ldots I$$

Separation cues: spectro-temporal vs spatial

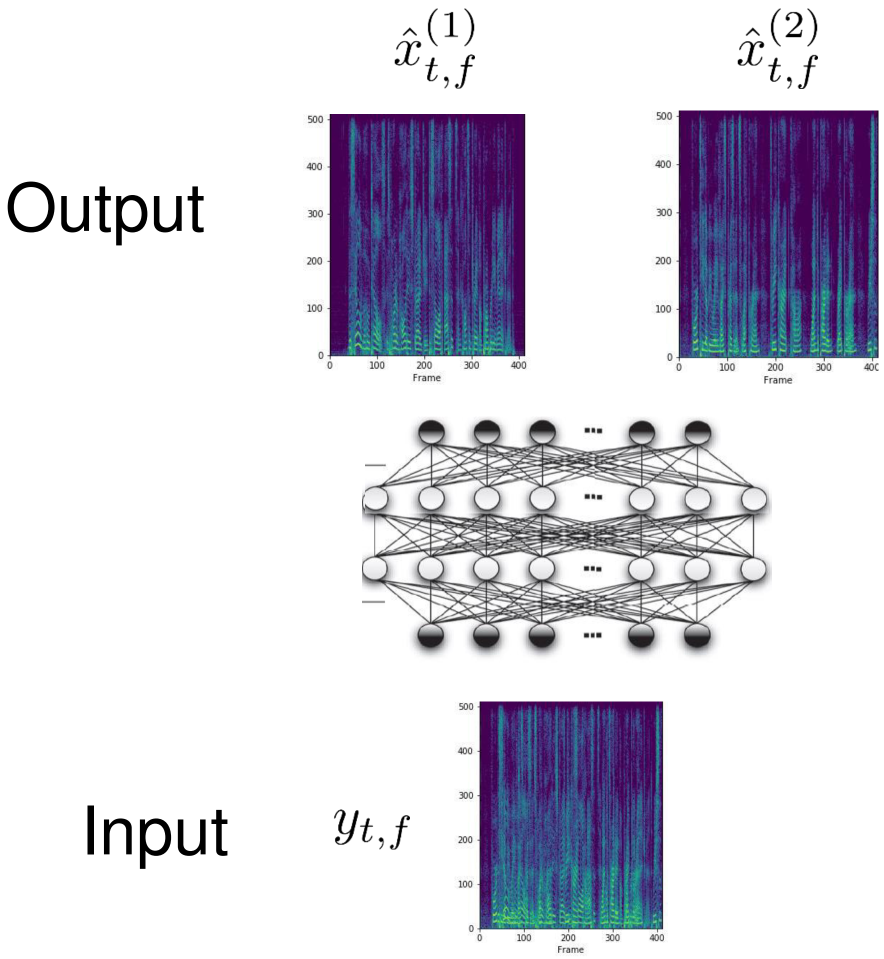


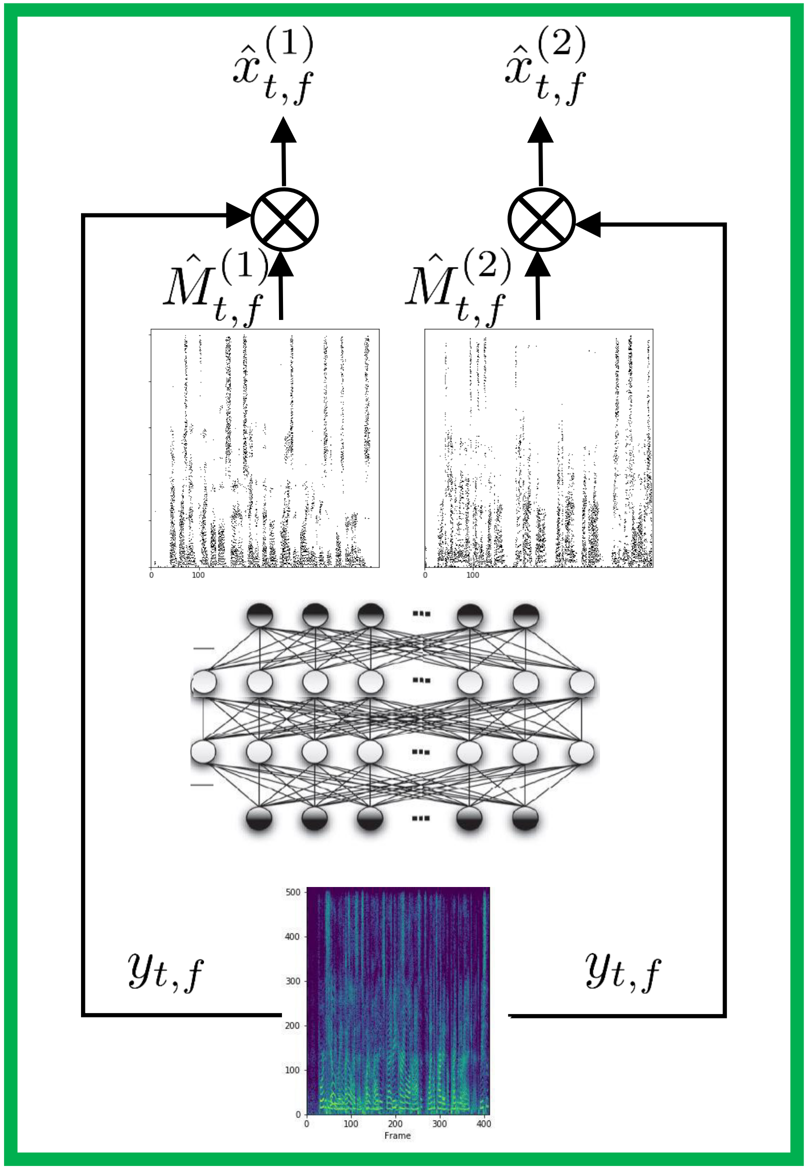
- Spectro-temporal cues
  - Model speech characteristics
  - Can work with **single-channel input**
  - Leverage training data
  - Typically supervised trng
  - DNN based



- Spatial cues
  - Exploits spatial selectivity
  - Requires **multi-channel input**
  - Does not require trng phase
  - Unsupervised learning (EM alg.)
  - Spatial mixture model based

Spectra vs masks as training targets





Mask based extraction performs better than direct signal estimation

Mask estimation

- Predict, for each tf-bin, the presence/absence of a target speaker
- Two types of objective functions
  - Mask approximation, e.g., cross entropy between estimated and ground truth mask
    - Appropriate if we do not need a decision for every tf bin
    - See spatial covariance matrix estimation in beamforming section
    - Does not measure reconstruction error
  - Signal approximation:

$$J(\theta) = \sum_{i,t,f} \left| \hat{x}_{t,f}^{(i)}(\theta) - x_{t,f}^{(i)} \right|^2 = \sum_{i,t,f} \left| \hat{M}_{t,f}^{(i)}(\theta) y_{t,f} - x_{t,f}^{(i)} \right|^2$$

    - Now, the training objective is the reconstruction error

Signal approximation performs better than mask approximation

Masks for signal approximation

$$J(\theta) = \sum_{i,t,f} \left| \hat{x}^{(i)}_{t,f}(\theta) - x^{(i)}_{t,f} \right|^2 = \sum_{i,t,f} \left| \hat{M}^{(i)}_{t,f}(\theta) y_{t,f} - x^{(i)}_{t,f} \right|^2$$

- The optimal mask for the above trng objective is the ideal complex mask $M^{(i)}_{t,f} = \frac{x^{(i)}_{t,f}}{y_{t,f}}$
  - But phase estimation is tricky …
- To avoid phase estimation, use best real-valued approximation to it: *ideal phase-sensitive mask* [Erdogan et al., 2015]

$$M^{(i)}_{t,f} = \Re\left\{ \frac{x^{(i)}_{t,f}}{y_{t,f}} \right\} = \frac{|x^{(i)}_{t,f}|}{|y_{t,f}|} \cos\left[\varphi^{(x^{(i)})}_{t,f} - \varphi^{(y)}_{t,f} \right]$$

  - Thus trng objective fu:

$$\left| \hat{M}^{(i)}_{t,f} y_{t,f} - x^{(i)}_{t,f} \right|^2 \propto \left(\hat{M}^{(i)}_{t,f} |y_{t,f}| - |x^{(i)}_{t,f}| \cos\left[\varphi^{(x^{(i)})}_{t,f} - \varphi^{(y)}_{t,f} \right] \right)^2$$

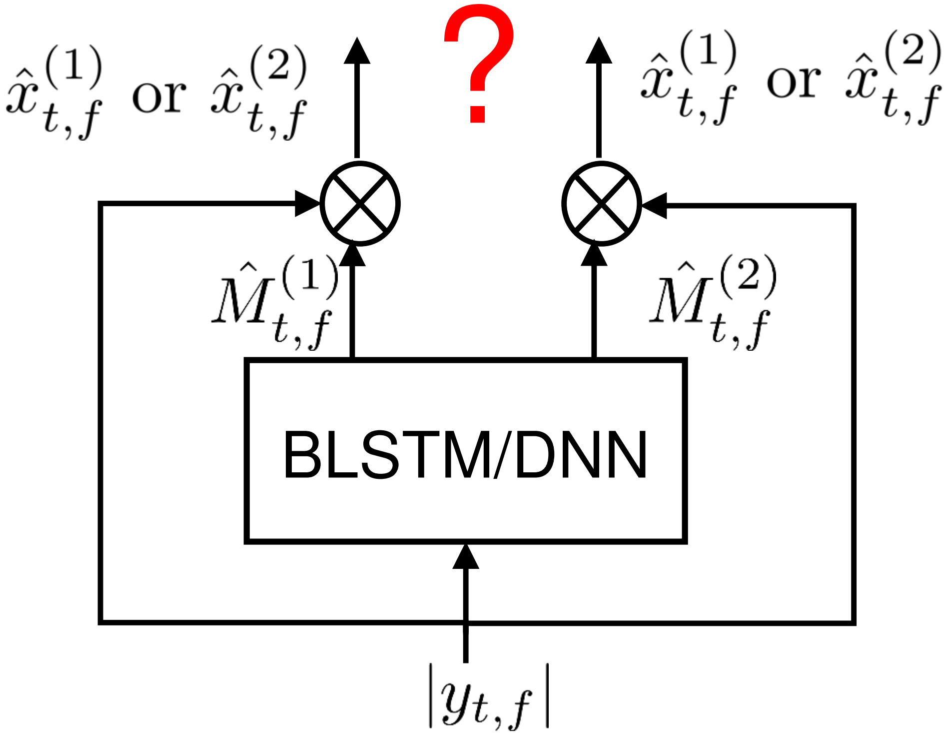
This trng objective has consistently shown better results than Ideal Binary Mask, Ideal Ratio Mask, etc. [Erdogan et al., 2015] [Kolbæk et al., 2017b]

DNN-based single-channel BSS

- Permutation Invariant Training (PIT)
- Deep Clustering (DC)
- Time Domain Audio Separation Network (Tasnet)

Utterance-PIT [Kolbæk et al., 2017b]

- Label ambiguity:



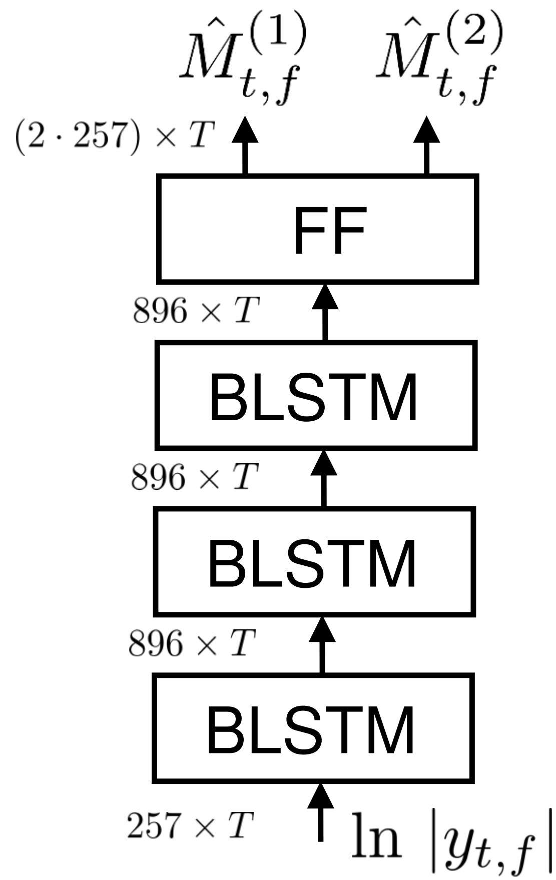
- Compute all permutations between the targets and the estimated sources and find permutation ϕ (over whole utterance) which minimizes MSE

$$J = \min_{\phi \in \mathcal{P}} \sum_{i,t,f} \left| \hat{M}^{(i)}_{t,f} y_{t,f} - x^{(\phi(i))}_{t,f} \right|^2$$

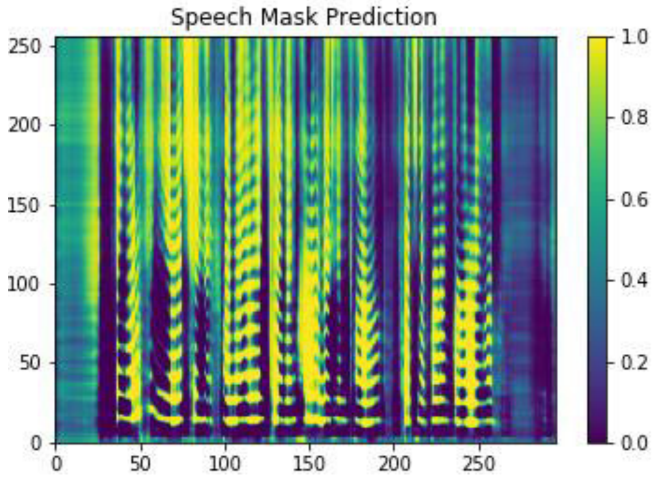
E.g.: $\min \left[\sum_{t,f} \left\{ \left| \hat{M}^{(1)} y - x^{(1)} \right|^2 + \left| \hat{M}^{(2)} y - x^{(2)} \right|^2 \right\}; \sum_{t,f} \left\{ \left| \hat{M}^{(1)} y - x^{(2)} \right|^2 + \left| \hat{M}^{(2)} y - x^{(1)} \right|^2 \right\} \right]$

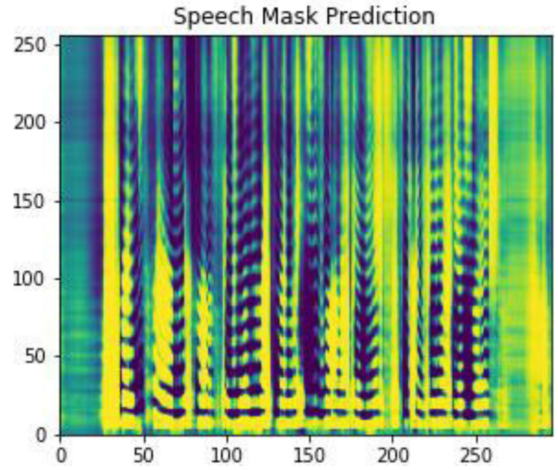
Example configuration

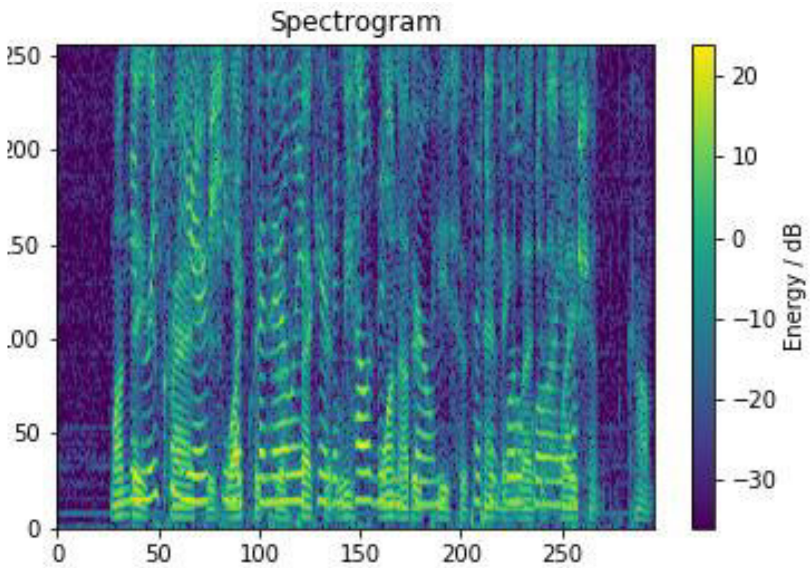
- Example configuration
  - Sampling rate 8 kHz; STFT window size: 64 ms; advance: 16 ms
  - Input: log-spectral magnitude features
  - 3 BLSTM layers with 896 nodes each
  - 1 FF layer with *(I x F)* nodes: I: #spkrs; F: #freq.bins (e.g., *I=2, F=257*); sigmoid output nonlinearity

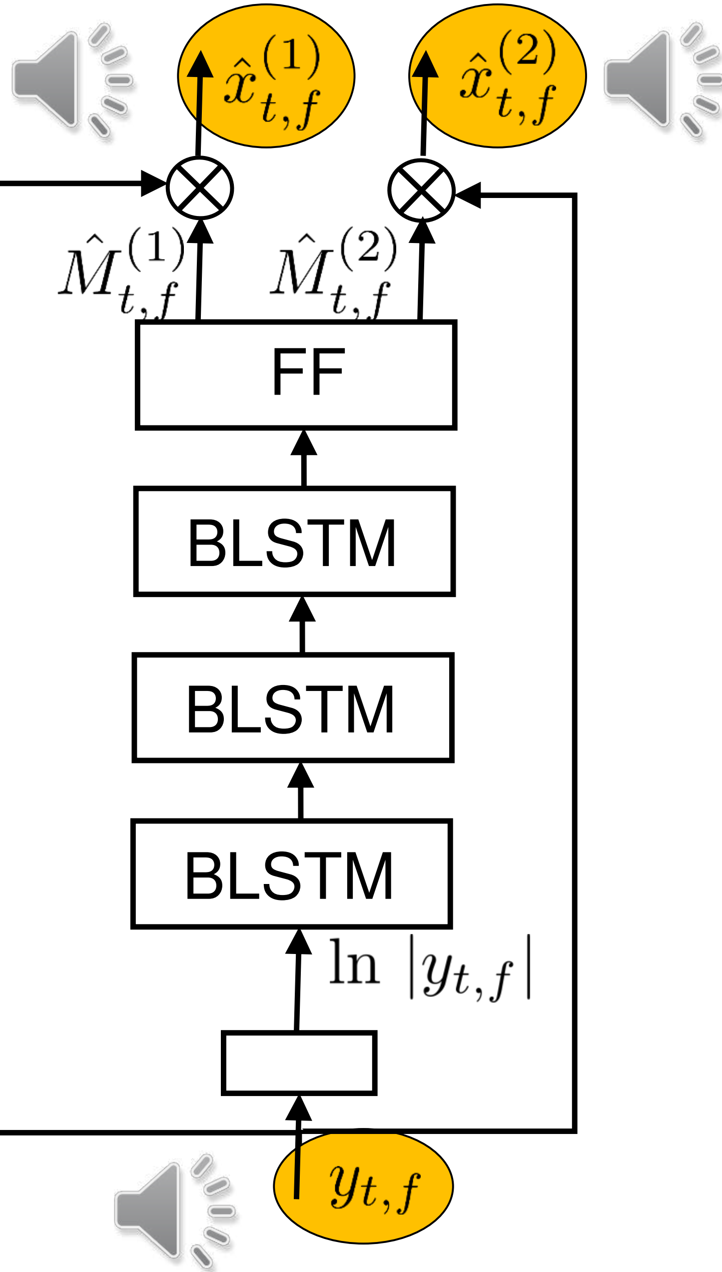


Demonstration



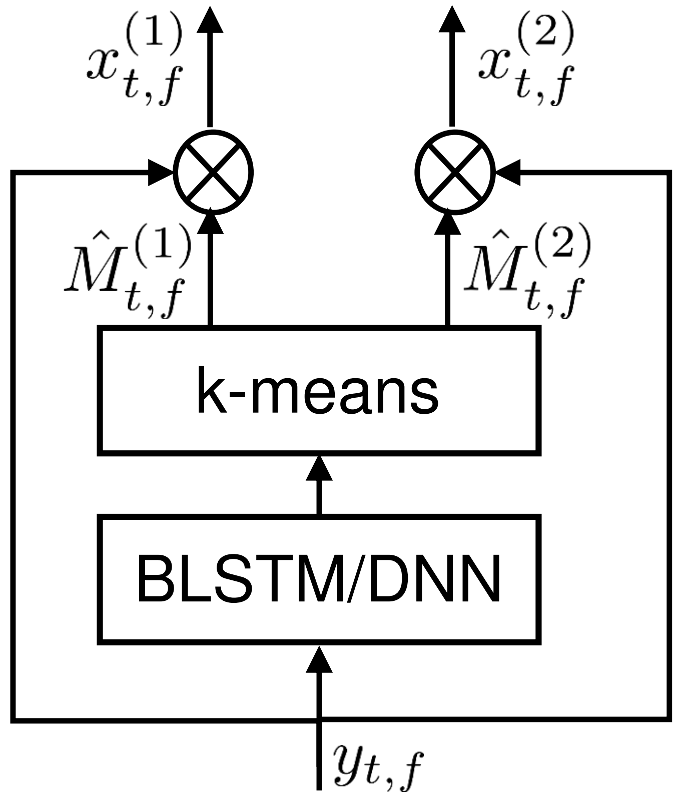






Deep Clustering [Hershey et al., 2016]

- Map each tf-bin to an embedding vector $\mathbf{e}_{t,f}$, where $\|\mathbf{e}_{t,f}\| = 1$
- Goal: tf-bins dominated by the same speaker form a cluster
  - Mapping via BLSTM network
- Mask estimation
  - K-means clustering of embedding vectors: hard assignments
  - Alternatively: estimate mixture model on embedding vectors: soft assignments



Training objective

- Affinity matrix $\mathbf{A}$ of size $(T \cdot F \times T \cdot F)$:
  - $[\mathbf{A}]_{n,n'} = 1$ if n-th and n'-th tf-bin from same speaker
  - n stands for certain time-frequency bin (t,f)
  - E.g, first and third tf-bin occupied by same speaker:

1	0	1	0
0	1	0	0
1	0	1	0
0	0	0	1

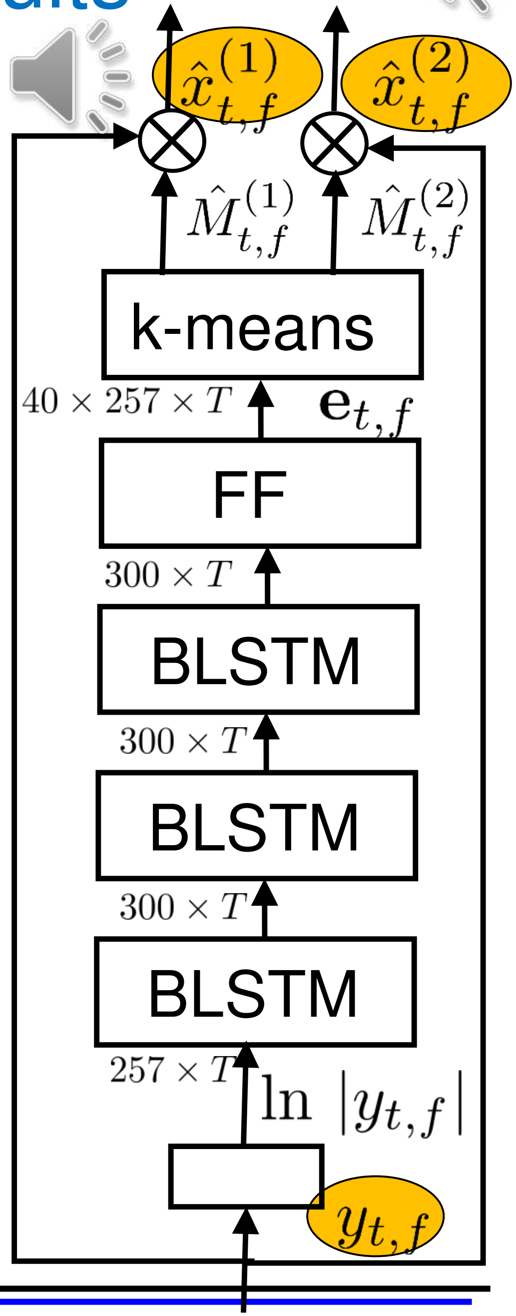
- Training objective: Minimize Frobenius norm of difference between estimated and true *affinity* matrix:

$$J(\theta) = \|\hat{\mathbf{A}}(\theta) - \mathbf{A}\|_{\mathrm{F}}^2$$

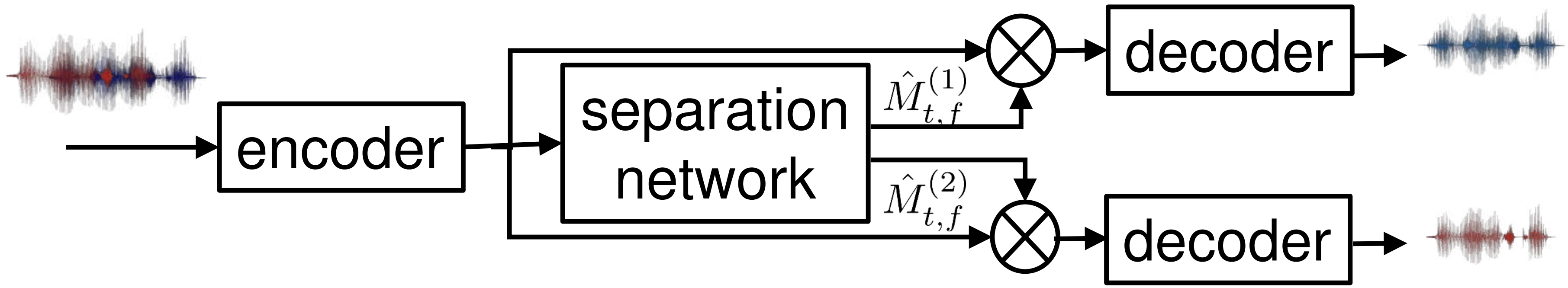
  - Estimated affinity matrix $\hat{\mathbf{A}} = \mathbf{E}\mathbf{E}^\top$, where $\mathbf{E}$ is matrix of embedding vectors $\mathbf{e}_{t,f}$

Example configuration and results

- Example configuration:
  - Embedding network: 3 BLSTM layers with 300 units in each direction
  - Final linear layer with *(K x F) nodes*: *K*: embedding dimension; *F*: #freq.bins (e.g., *K=40, F=257*)



TasNet [Luo and Mesgarani, 2018]



- Time-domain source separation
  - STFT replaced by learnt transformation (encoder):
    - Form segments of speech (e.g. 20 samples, i.e., 2.5 ms)

$$\mathbf{y}[tB] = [y[tB], y[tB-1] \ldots, y[tB-L+1]]^{\mathsf{T}}$$

    - 1-D convolution layers applied to overlapping segments of speech

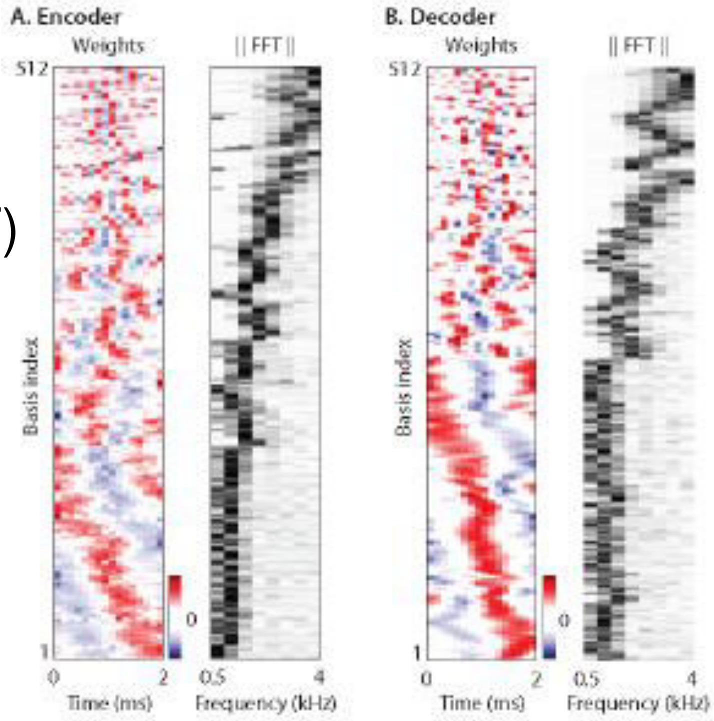
$$\mathbf{w}_t = \mathrm{ReLU}\left(\mathbf{y}[tB] \circledast \mathbf{U}\right); \quad \mathbf{U} \in \mathbb{R}^{N \times L}$$

    - Encoder transforms time-domain signal to nonnegative representation using N encoder basis functions
  - Mask estimation in transform domain
  - Source extraction by masking: $\hat{\mathbf{x}}_t^{(i)} = \mathbf{w}_t \odot \hat{\mathbf{M}}_t^{(i)}$
  - Learned decoder generates waveform: $\hat{\mathbf{x}}^{(i)}[tB] = \hat{\mathbf{x}}_t^{(i)} \circledast \mathbf{V}$

Learned transformations

- Encoder / Decoder
  - No constraint on orthogonality of bases
  - Non-negativity constraint on encoder output
  - Decoder is not inverse of encoder (as in STFT)

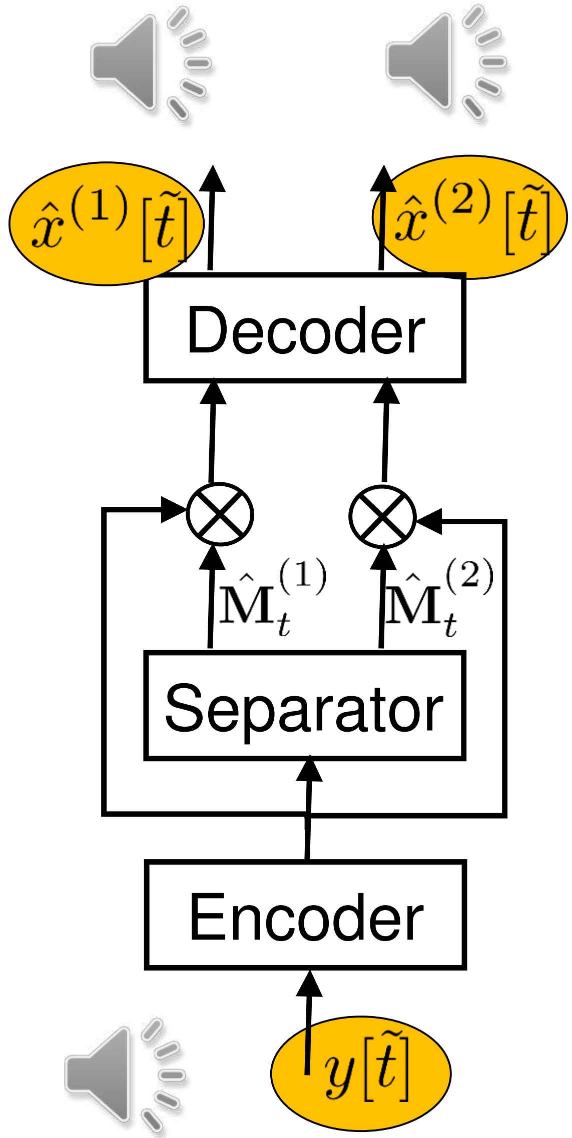
- Can the learned bases be interpreted?
  - Most filters at low frequencies
  - Filters of same frequencies with different phases



Basis functions of encoder/decoder and the magnitudes of their FFT; taken from [Luo and Mesgarani, 2018]

Example configuration and results

- Example configuration
  - Encoder: sampling rate 8 kHz; 1-D convolution operation with
    window of L = 20 (2.5ms); N = 256 basis functions
  - Separator:
    - Stacked 1-D dilated convolutional blocks, see [Luo and Mesgarani, 2018]
  - Decoder: 1-D transposed convolution operations



Discussion

- PIT, DC, TasNet and DAN (Deep Attractor Network) achieve very good speaker independent BSS

Results on wsj0-2mix:
[Le Roux et al., 2018b]

Method	**SDR [dB]**
PIT	(10.0)
DC	10.8
TasNet	14.6

- TasNet naturally incorporates phase restoration, while the others estimate only magnitude spectrum
- TasNet achieves largest SDR improvement
  - Others come close when phase reconstruction component is added
- As a time domain approach TasNet has lowest latency

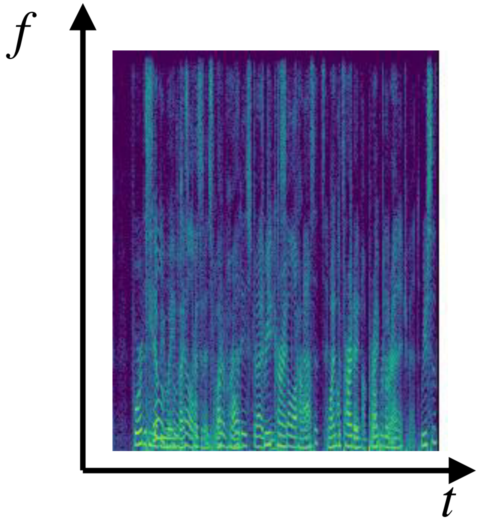
- Number of speakers must be known
  - In PIT, even the network architecture depends on the (max.) no of speakers

Extensions

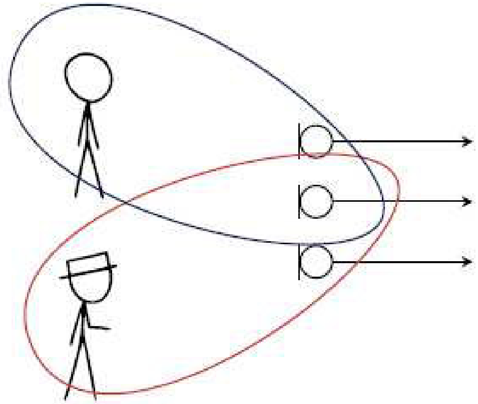
- Combinations of approaches, e.g., PIT network trained with additional DC loss [Wang and Wang, 2019]
- Extension to multi-channel input: use cross-channel features as additional input (e.g. inter-channel phase differences)
- Now that magnitude reconstruction is so good, phase reconstruction has come in the focus of research
  - Time-domain solutions (TasNet)
  - Phase reconstruction at the output of a good magnitude estimation network [Wang et al., 2018b]
  - Estimation of phase masks using discrete representation of phase diff. between noisy and clean phase [Le Roux et al., 2018a]

Table of contents in part IV

Separation cues: spectro-temporal vs spatial



- Spectro-temporal cues
  - ➢ Model speech characteristics
  - ➢ Can work with **single-channel input**
  - ➢ Leverage training data
  - ➢ Typically supervised trng
  - ➢ DNN based



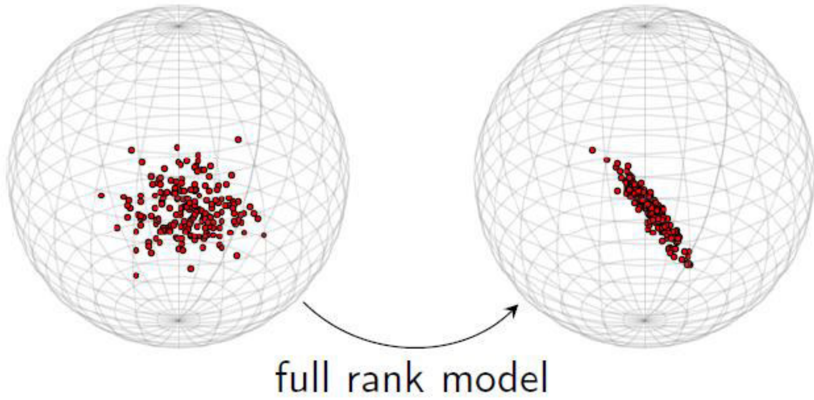
- Spatial cues
  - ➢ Exploits spatial selectivity
  - ➢ Requires **multi-channel input**
  - ➢ Does not require trng phase
  - ➢ Unsupervised learning (EM alg.)
  - ➢ Spatial mixture model based

Spatial mixture model

- Straightforward extension of beamforming case

$$p(\mathbf{y}_{t,f}) = \sum_i \Pr(M_{t,f} = i) p(\mathbf{y}_{t,f} | M_{t,f} = i); \; i \in \{0, 1, \ldots, I\}$$

  - E.g., Complex angular central Gaussian Mixture Model with $I+1$ components



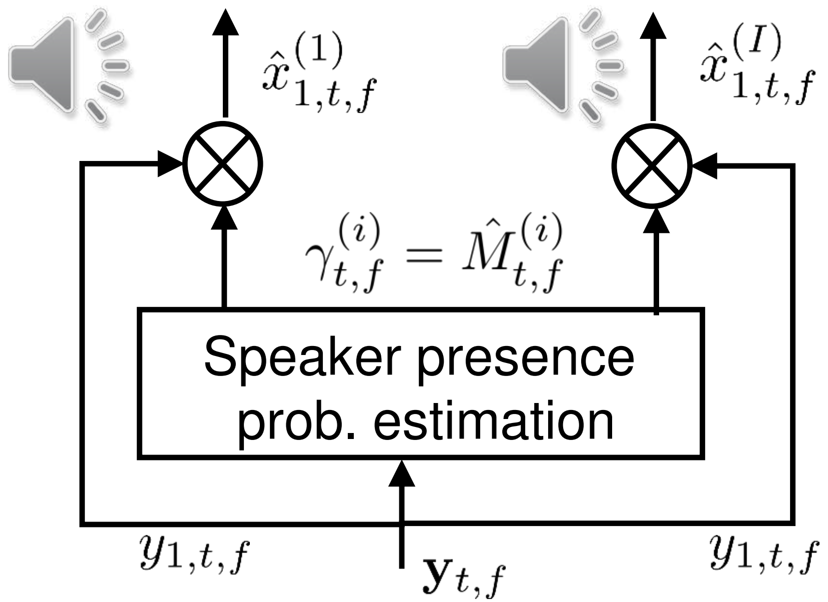
- EM algorithm to estimate speaker presence probabilities

$$\gamma_{t,f}^{(i)} = \hat{\Pr}(M_{t,f} = i | \mathbf{y}_{t,f}) =: \hat{M}_{t,f}^{(i)}$$

Source extraction

by masking

Beamforming achieves better perceptual quality (and WER performance)



by beamforming

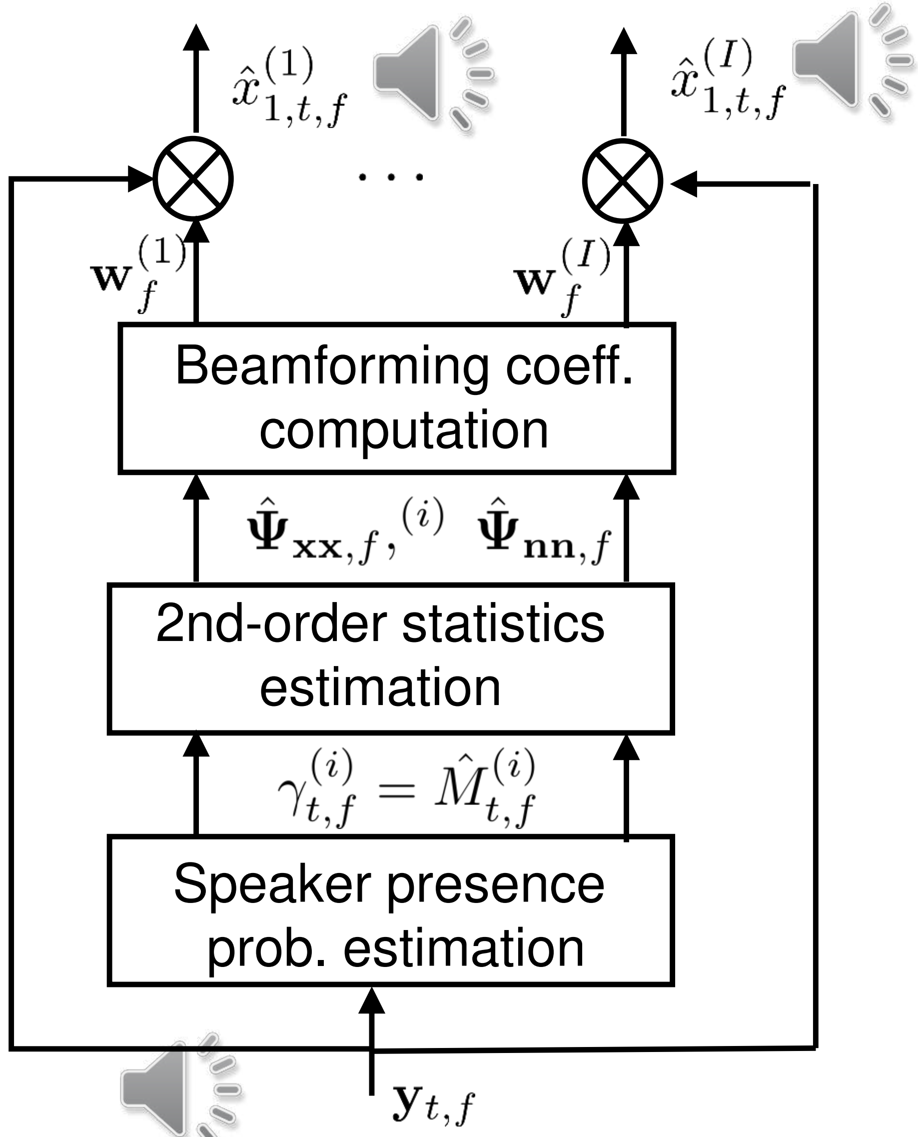
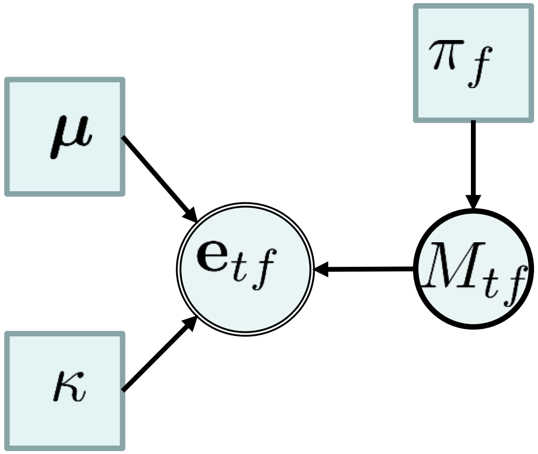


Table of contents in part IV

Integration of Deep Clustering and mixture models

- Goal: combine the strengths of both methods
  - Exploit spectral and spatial cues for separation
  - Leverage trng data and do unsupervised learning on test utterance
- Weak integration
  - Use k-means result of DC as initialization of $\gamma_{t,f}^{(i)}$ (speaker presence prob.) of the spatial mixture model and run EM steps on test utterance
- Strong integration
  - Take embedding vectors $\mathbf{e}_{t,f}$ and microphone signals $\mathbf{y}_{t,f}$ as two observations in a mixture model

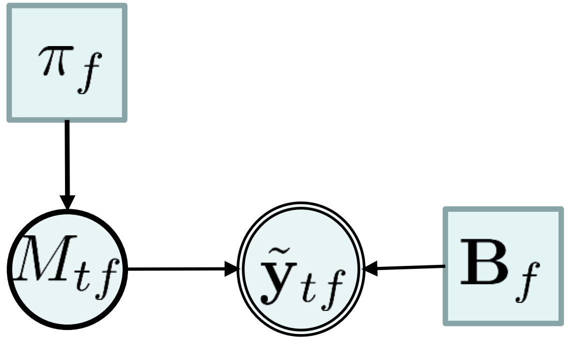
Mixture model for DC embeddings



- Model embedding vectors as r.v.
  - Mixture of von-Mises Fisher distributions
  - K-means replaced by EM

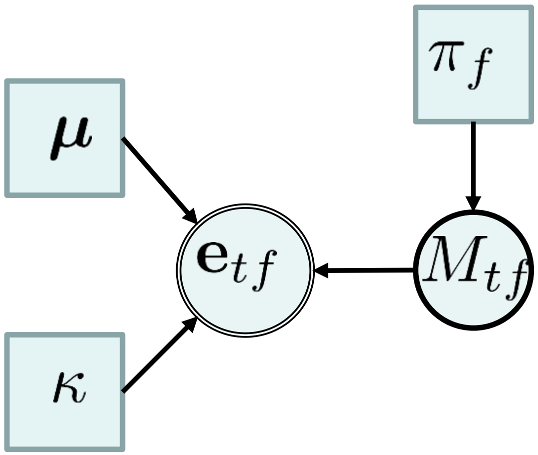
$$p(\mathbf{e}_{tf}) = \sum_i \Pr(M_{t,f} = i) p(\mathbf{e}_{t,f} | M_{tf} = i)$$
$$= \sum_i \pi_f^{(i)} \cdot \mathrm{vMF}(\mathbf{e}_{tf}^{(i)}; \boldsymbol{\mu}^{(i)}, \kappa^{(i)})$$

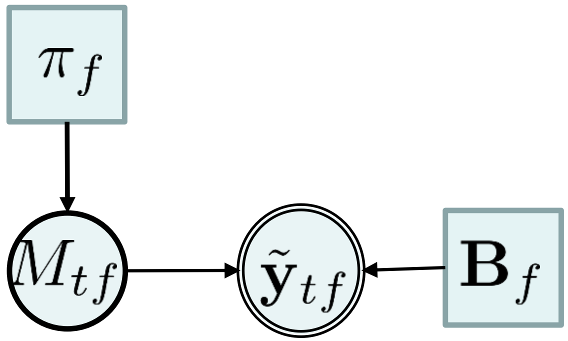
Recall spatial mixture model



$$
\begin{aligned}
p(\tilde{\mathbf{y}}_{t,f}) &= \sum_i \Pr(M_{t,f} = i) p(\tilde{\mathbf{y}}_{t,f} | M_{t,f} = i) \\
&= \sum_i \pi_f^{(i)} \mathrm{cACG}(\tilde{\mathbf{y}}_{t,f}; \mathbf{B}_f^{(i)})
\end{aligned}
$$

Strong integration

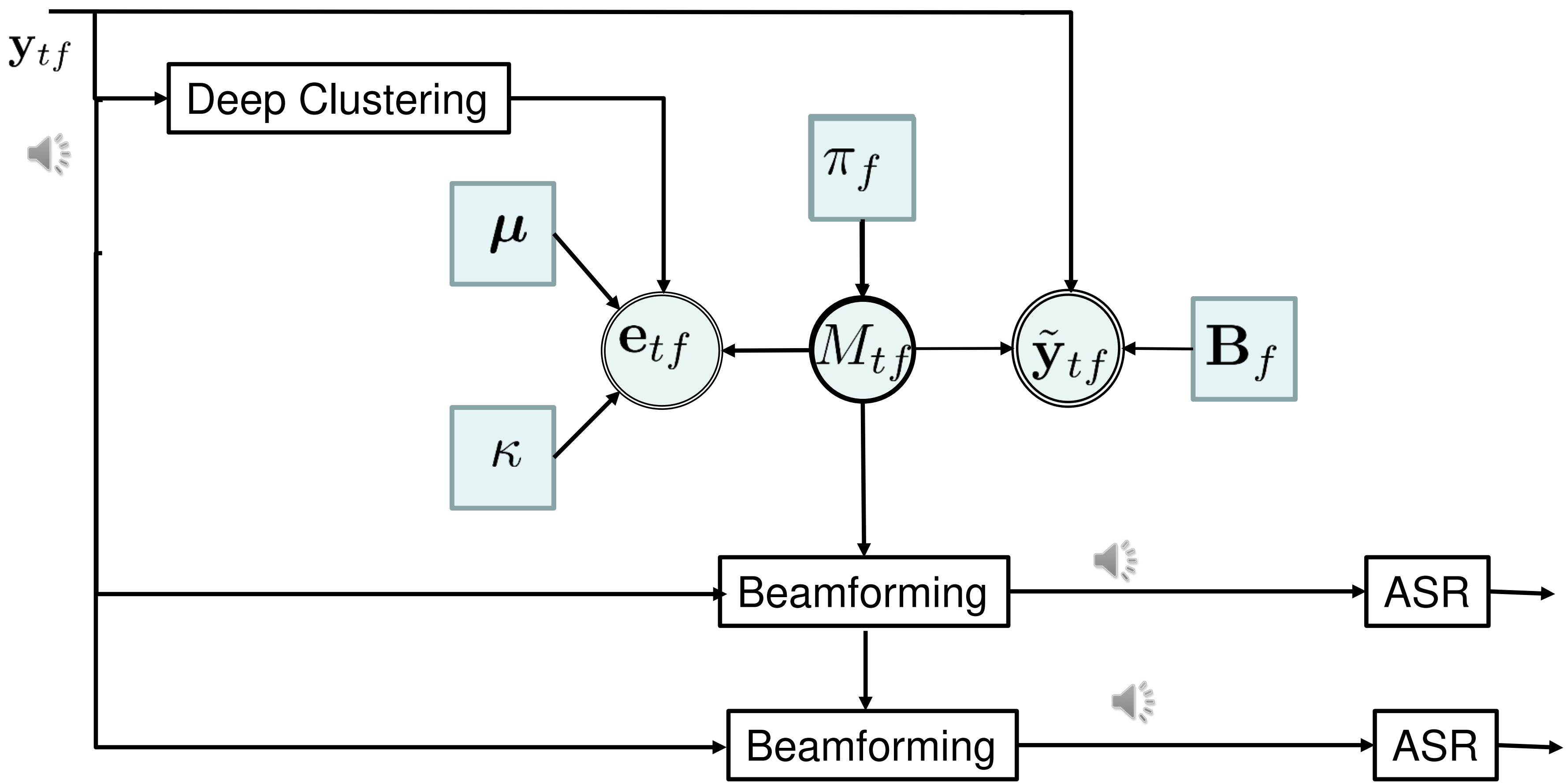




Integrated mixture model

- Coupling via latent class affiliation variable (speaker presence prob.)
- Hypothesis: better estimates when estimated jointly

Overall system



Results [Drude and Haeb-Umbach, 2019]

- Database: spatialized multi-channel wsj-2mix
  - Artificial 2-speaker mixtures from WSJ utterances
  - 8 channels
  - T_{60} = 0.2 – 0.6 s
- Acoustic model trained either on **clean** speech or on **image** of clean speech at reference microphone (includes reverb.)

Model	**WER [%]** Clean	**WER [%]** Image
Spatial mixture model (cACGMM)	40.9	28.2
Deep Clustering (DC)	42.5	26.6
Weak integration	34.4	21.6
Strong integration (DC + cACGMM)	33.4	18.9
oracle	31.1	10.7

Pros and cons of NN and spatial mixture model based BSS

	Spatial mixture models	Neural networks
Spatial characteristics modeling	• **Strong**	• Moderate (use of cross-channel features at input)
Spectro-temporal characteristics modeling (for speech)	• Weak - Permutation problem • No concept of human speech (pros and cons)	• **Very strong** - Strong speech model based on a priori training
#channels required	• Multi-channel	• Single channel
Leverage training data	• No training phase	• **Yes**, but parallel data required
Adaptation to test condition	• **Strong** - Unsupervised learning applicable	• Weak - Poor generalization - Sensitive to mismatch

We have seen the same table before

Software

- Spatial mixture models: https://github.com/fgnt/pb_bss
  - Different spatial mixture models
    - complex angular central Gaussian , complex Watson, von-Mises-Fisher
  - Methods: init, fit, predict
  - Beamformer variants
  - Ref: [Drude and Haeb-Umbach, 2017]

Summary of part IV

- Speaker-independent single-channel DNN-based BSS is a major improvement over earlier approaches
- Source extraction by beamforming produces less artifacts than by masking
- Both DNN-based and spatial mixture model based BSS achieve comparable results when used with beamformer for source extraction
- DNN based and spatial mixture model based BSS have complementary strengths and can be combined
- Often simplifying assumptions:
  - # active speakers known
  - All speakers speak all the time
  - Most investigations on artificially mixed speech and static scenario
  - offline

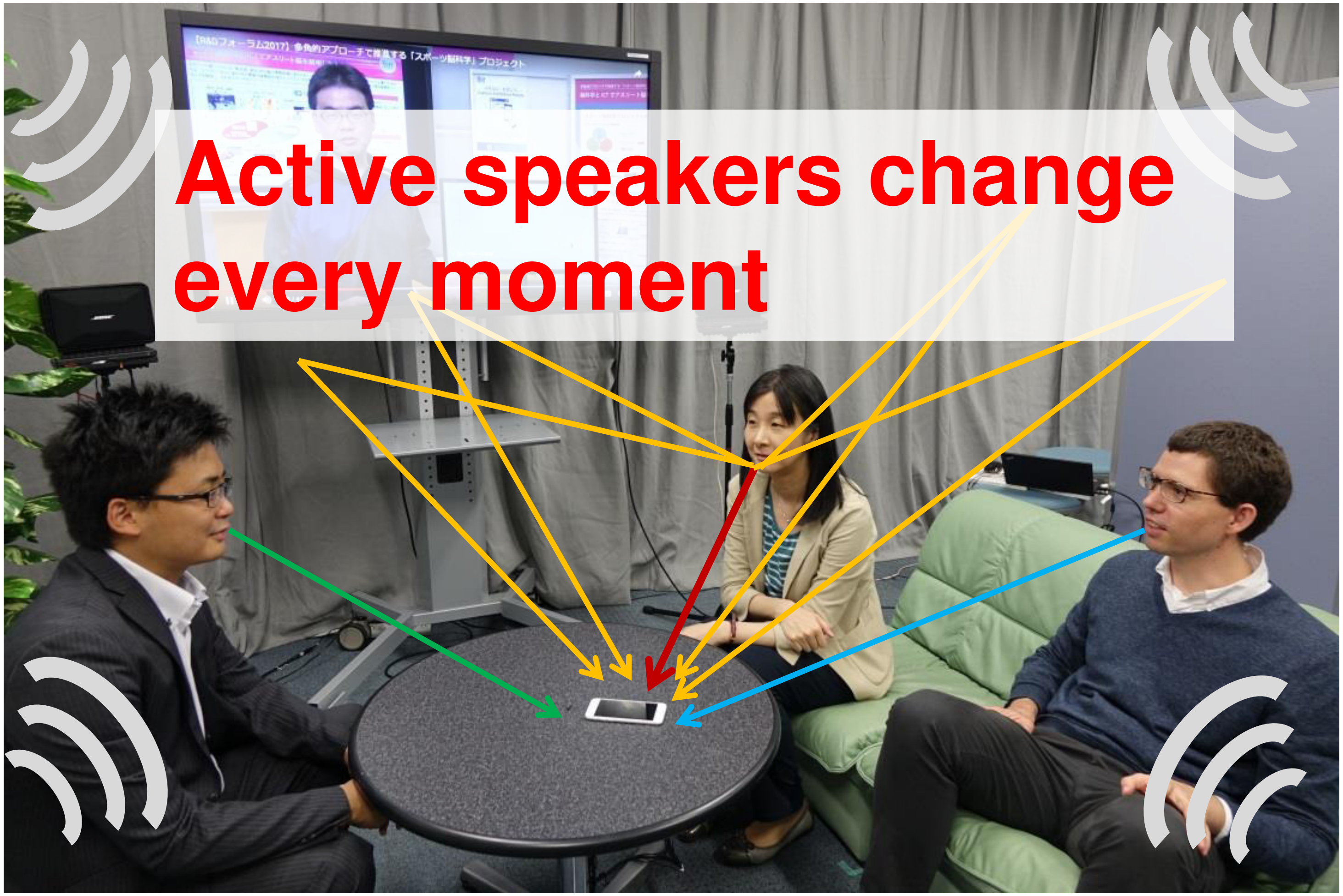
Some of those assumptions will be lifted in the next presentation

Table of contents

Part V.
Meeting Analysis

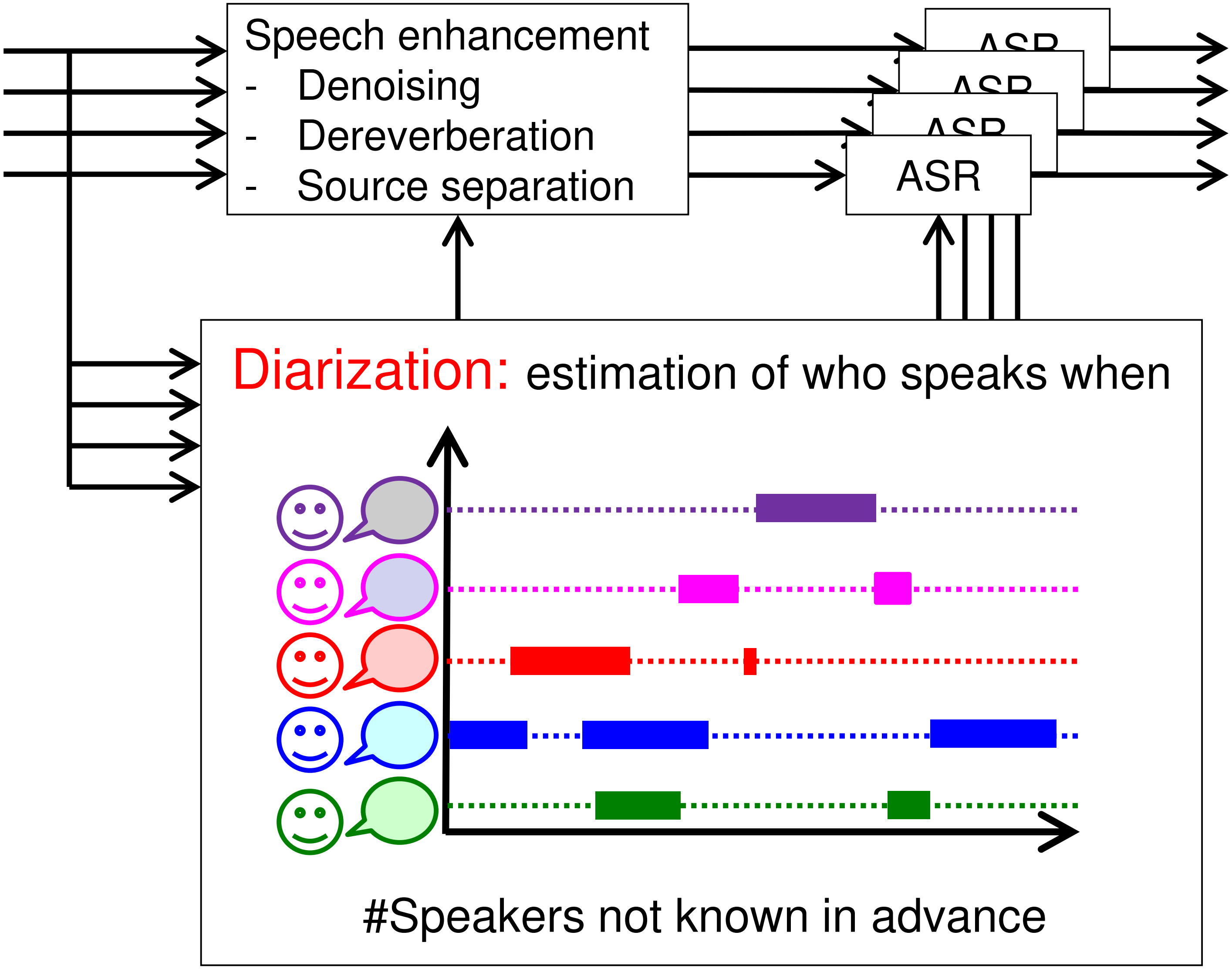
Tomohiro Nakatani

Speech recording in meeting situation



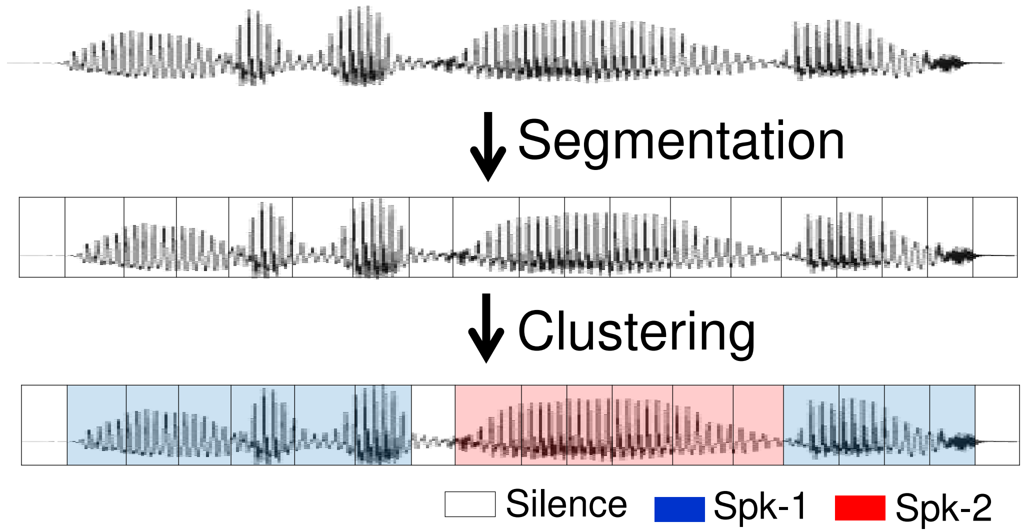
- Estimation of who speaks when (=**diarization**) is crucial for speech enhancement and ASR

Problems in meeting analysis

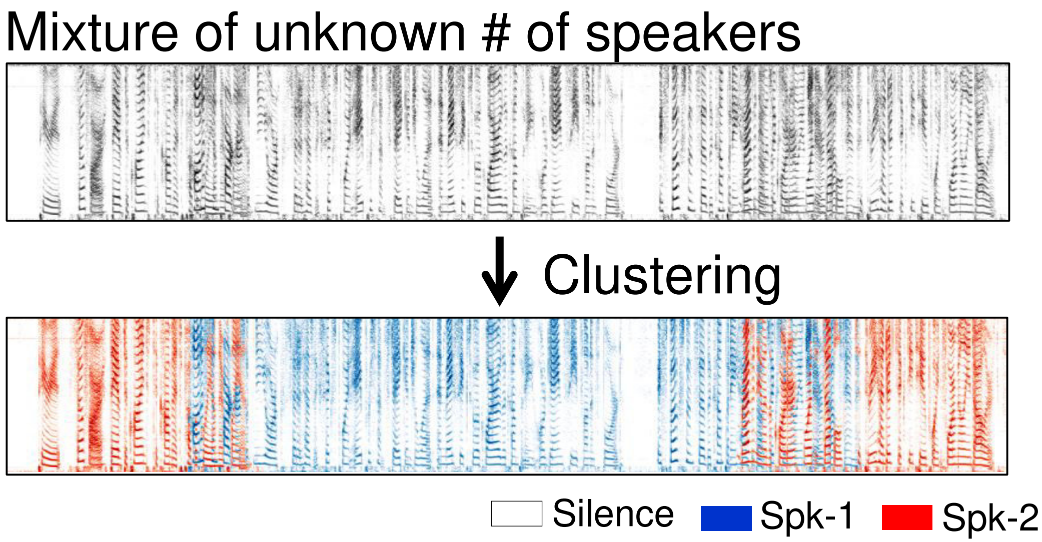


Two approaches to diarization

- Clustering of time segments
  - Based on spectral features
    - MFCC, i-vector, d-vector, x-vector, etc.
  - Speaker overlapping segments are disregarded
  - 1-ch processing

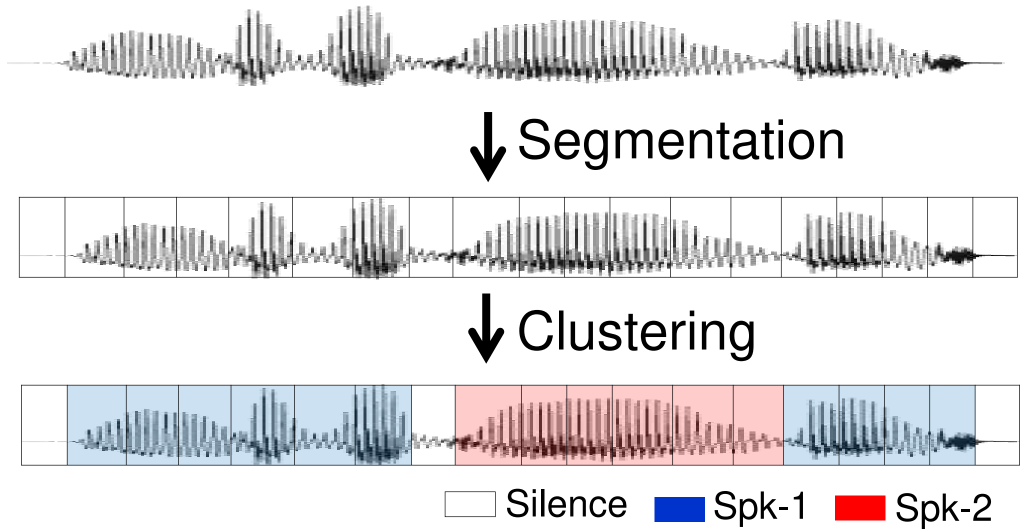


- Clustering of TF points
  - Mask-based source separation for unknown #sources
  - Speaker overlapping segments can be separated
  - 1-ch/multi-ch processings



Approaches to diarization

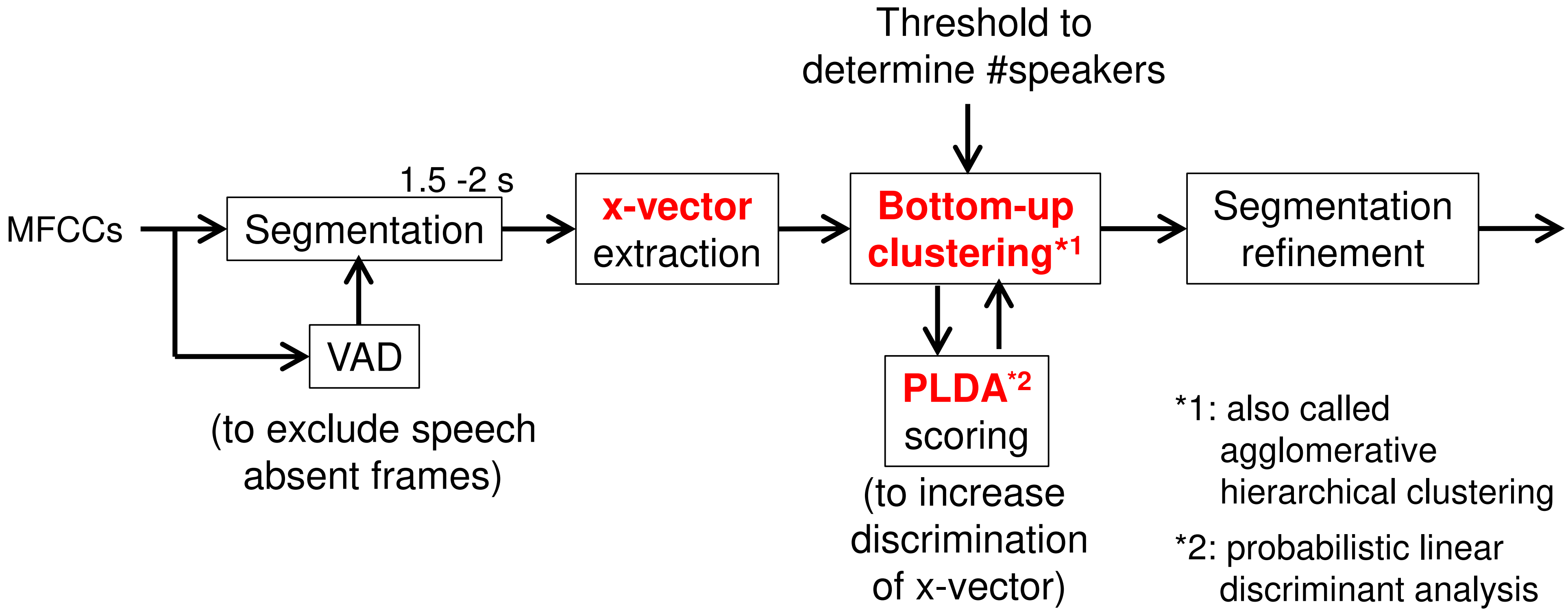
- Clustering of time segments
  - Based on spectral features
    - MFCC, i-vector, d-vector, x-vector, etc.
  - Speaker overlapping segments are disregarded
  - 1-ch processing



- Clustering of TF points
  - Mask-based source separation for unknown #sources
  - Speaker overlapping segments can be separated
  - 1-ch/multi-ch processings

JHU DIHARD challenge system [Sell et al., 2018]

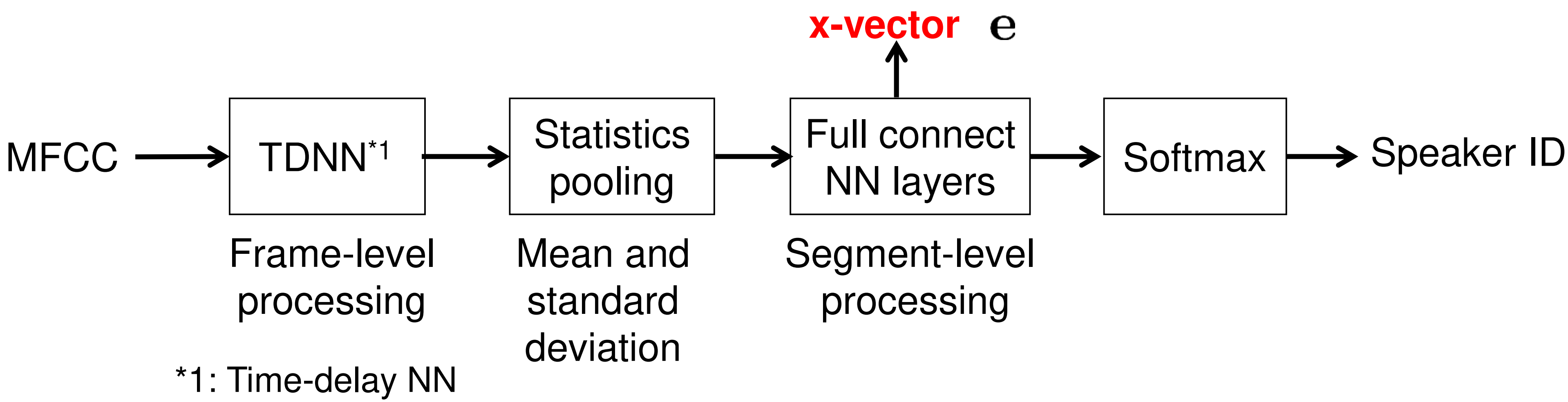
- Best score at Track 1 of DIHARD-I challenge
  - DIHARD-I,II: diarization challenges with HARD corpora [Ryant et al., 2019]



Robust speaker feature extraction and scoring are crucial

x-vector [Snyder et al., 2018]

- A bottleneck feature of speaker verification NN
  - Trained using data augmentation (noise, reverb)



A speaker characteristic essential for speaker verification

PLDA [Silovsky et al., 2011]

- Decompose an x-vector into different factors

$$\mathbf{e} = \underbrace{\mathbf{m}}_{\text{Speaker independent mean}} + \underbrace{\mathbf{F}\mathbf{h}_i}_{\textbf{Speaker inherent feature}} + \underbrace{\mathbf{G}\mathbf{w}_{i,j}}_{\text{Utterance dependent feature}} + \underbrace{\mathbf{n}_{i,j}}_{\text{noise}}$$

i: Speaker index
j: Utterance index
$\mathbf{m}, \mathbf{F}, \mathbf{G}$ and Σ : Model parameters determined in advance using training data

$$p(\mathbf{e} \mid \mathbf{h}_i, \mathbf{w}_{i,j}; \theta) = \mathcal{N}(\mathbf{m} + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{i,j}, \Sigma)$$

Cluster likelihood : $p(\mathbf{e}_1, \ldots, \mathbf{e}_J) = \mathcal{N}(\mathbf{m}', \mathbf{A}\mathbf{A}^\top + \Sigma')$

where $\mathbf{m}' = (\mathbf{m}, \ldots, \mathbf{m})^\top$ $\mathbf{A} = \begin{pmatrix} \mathbf{F} & \mathbf{G} & 0 & \ldots & 0 \\ \mathbf{F} & 0 & \mathbf{G} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \mathbf{F} & 0 & 0 & \ldots & \mathbf{G} \end{pmatrix}$ $\Sigma' = \begin{pmatrix} \Sigma & 0 & \ldots & 0 \\ 0 & \Sigma & \ldots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \ldots & \Sigma \end{pmatrix}$

Diarization can be performed with speaker inherent features

Evaluation metric for diarization

- Diarization error rates (DER) [NIST speech group, 2007]

$$\mathrm{DER} = \frac{\#\text{frames with wrongly estimated speaker}}{\text{total } \#\text{frames}}$$

  – Includes: missed speaker time (MST), false active time (FAT), and speaker error time (SET)

DERs with DIHARD-I challenge [Sell et al., 2018]

Dataset includes: clinical interviews, child language acquisition recordings, YouTube videos, speech in restaurants

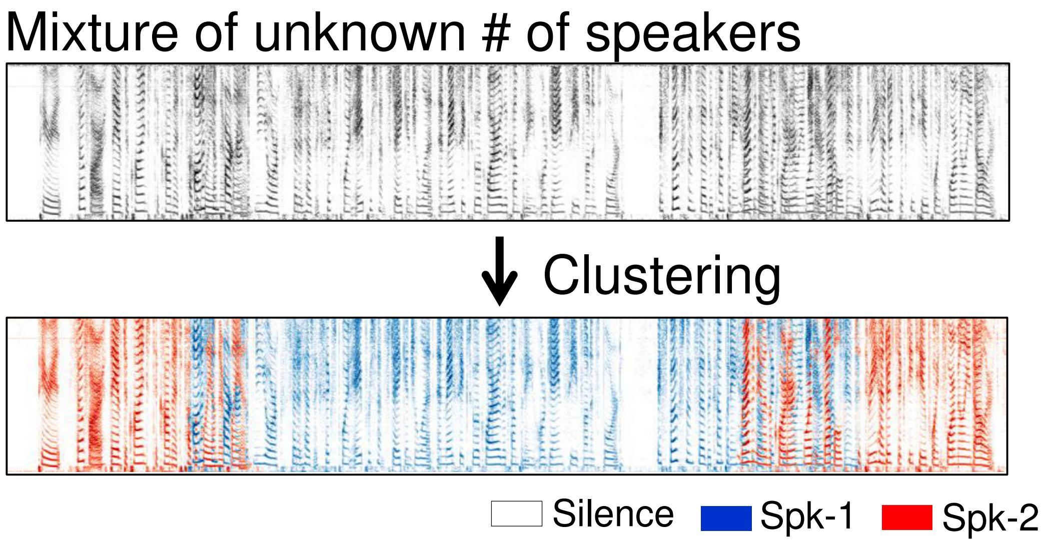
Track1: w/ oracle speech segmentation (Challenge top for Eval: 23.73 %)
Track2: w/o oracle speech segmentation (Challenge top for Eval: 35.51 %)

	Track1	**Track2**
All same speaker	39.01 %	55.93 %
i-vector + PLDA	28.06 %	40.42 %
x-vector + PLDA	25.94 %	39.43 %
x-vector + PLDA, with seg. refinement	**23.73 %**	**37.29 %**

Approaches to diarization

- Clustering of time segments
  - Based on spectral features
    - MFCC, i-vector, d-vector, x-vector, etc.
  - Speaker overlapping segments are disregarded
  - 1-ch processing

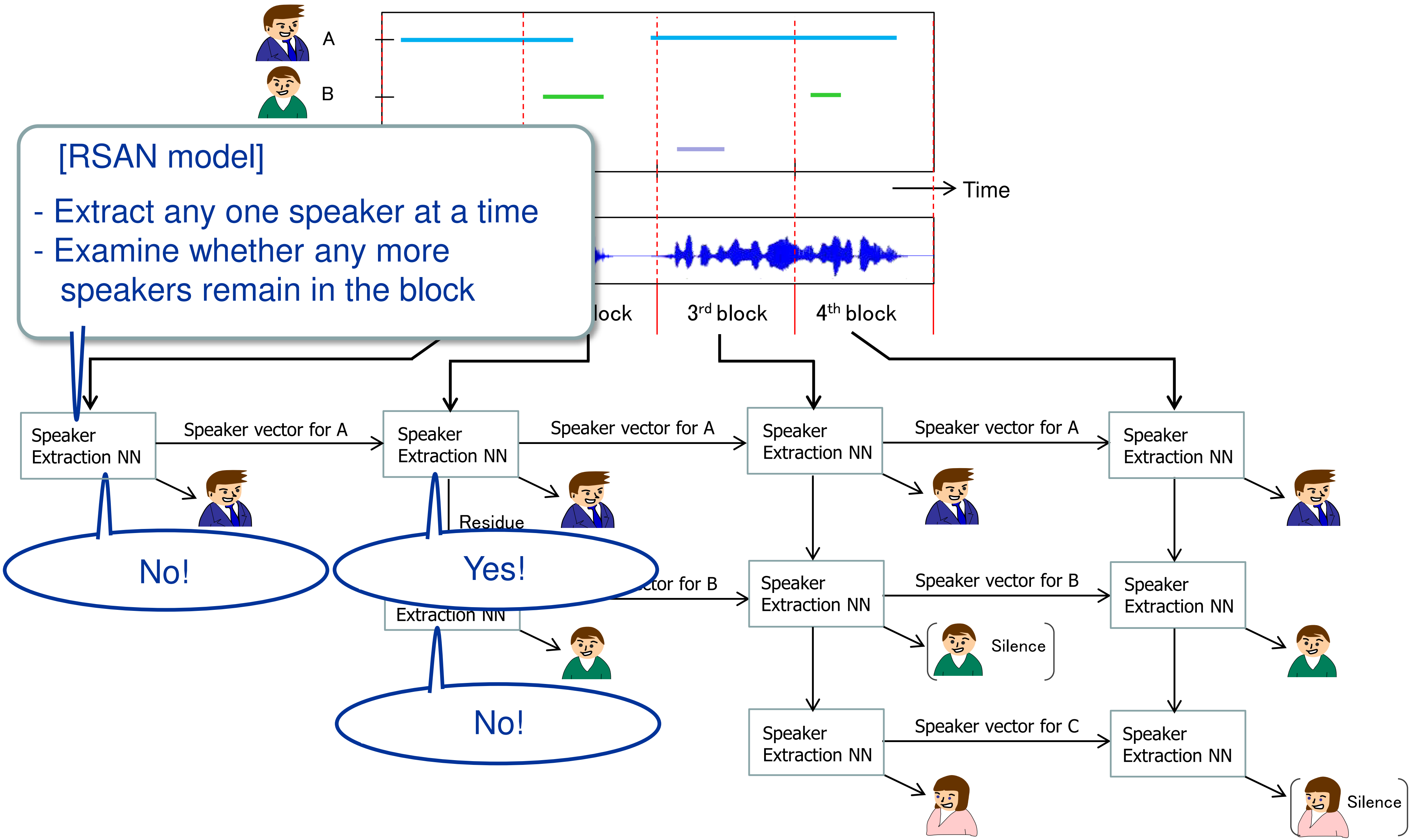
- Clustering of TF points
  - Mask-based source separation for unknown #sources
  - Speaker overlapping segments can be separated
  - 1-ch/multi-ch processings



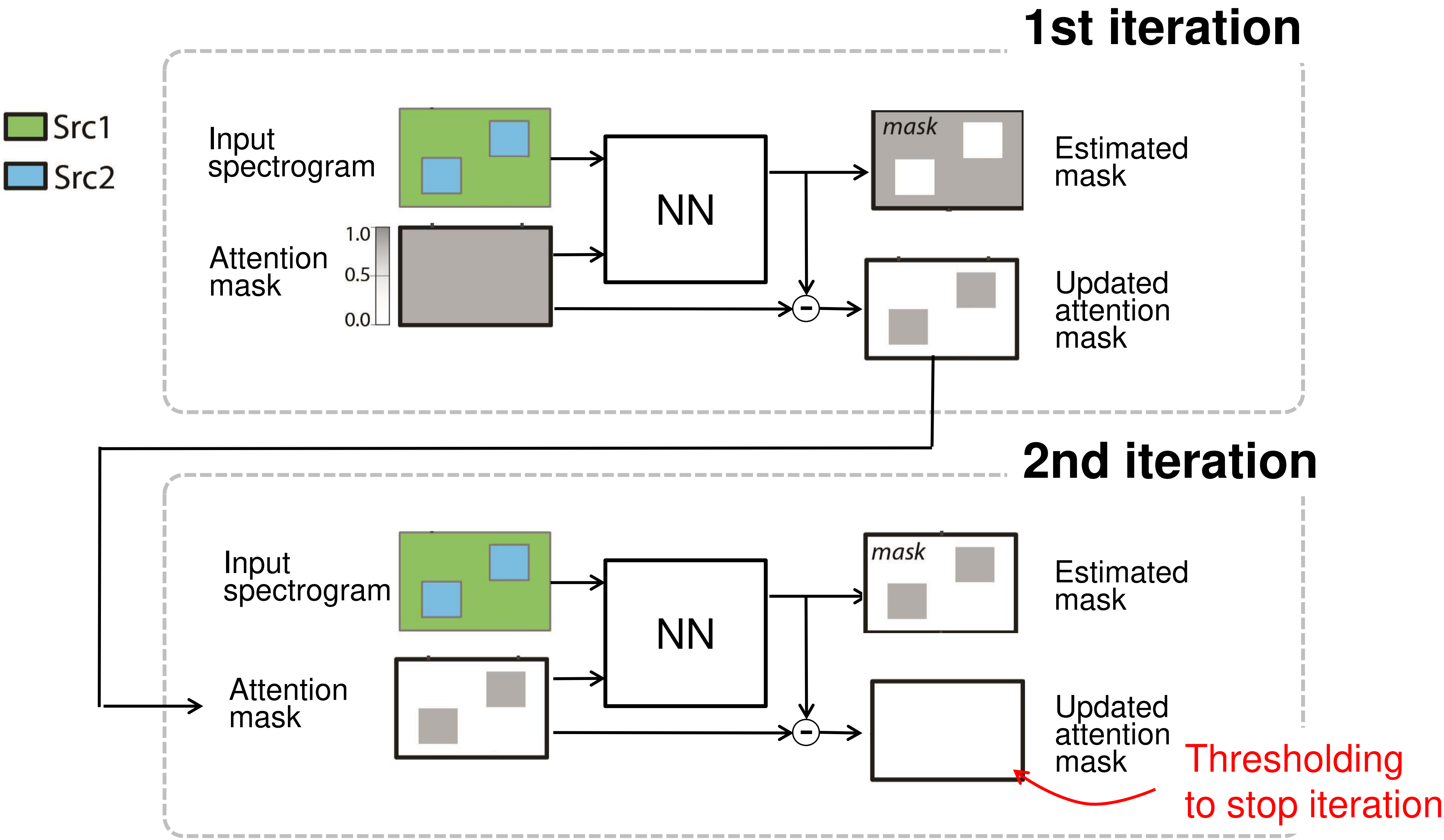
Recurrent Selective Attention Network (RSAN) [Kinoshita et al., 2018, von Neumann et al., 2019]

- Neural network based mask estimator for unknown #speakers
- Perform block online meeting analysis
  - By dynamically assigning a NN to extract a source every time it detects a new source,
- Can be optimized in an end-to-end manner for feature extraction, source counting, diarization, and source separation

Overall online processing flow by RSAN



How to control #iterations at each block



Training of RSAN : loss function

$$\mathcal{L} = \mathcal{L}^{\mathrm{Sep}} + \alpha \mathcal{L}^{\mathrm{Count}}$$

Loss for separation

$$\mathcal{L}^{\mathrm{Sep}} = \sum_i \|\hat{\mathbf{Y}}_i - \mathbf{Y}_i^{ref}\|_2^2$$

$\hat{\mathbf{Y}}_i, \mathbf{Y}_i^{ref}$: Estimated and clean speech spectra

Loss for source counting

$$\mathcal{L}^{\mathrm{Count}} = \max(\mathbf{R}, 0)$$

$$\mathbf{R} = \mathbf{1} - \sum_{\mathbf{i}} \mathbf{M}^{(\mathbf{i})}$$

: Attention mask after masks for all the sources are extracted

Source separation, counting, feature extraction, and diarization are jointly optimized in an end-to-end processing manner

Preliminary results with simulated conversation

Test data:
- Simulated conversation composed of utterances (WSJ)
- Average conversation length: 30 s

	DER	SCER	DER+ SCER
All same speaker	38.8 %	27.4 %	66.2 %
Bottom up clustering of RSAN speaker vectors (batch)	15.8 %	6.2 %	22.0 %
PIT based mask estimation (batch)	9.8 %	**4.4 %**	14.2 %
RSAN (online)	**6.6 %**	4.9 %	**11.5 %**

Speaker confusion error rate (SCER): [von Neumann et al., 2019]

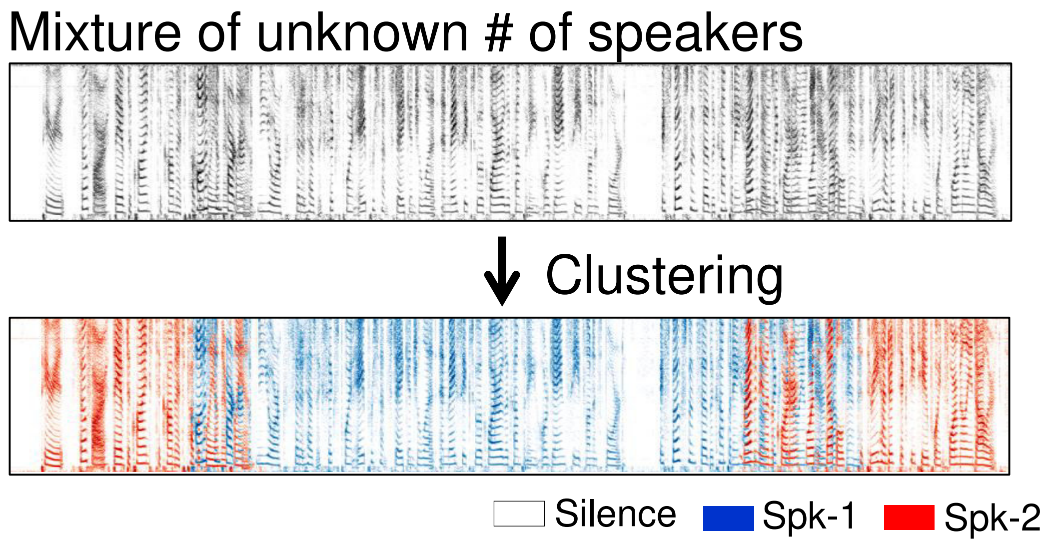
$$\mathrm{SCER} = \frac{\#\text{frames with confused speaker assignments}}{\text{total } \#\text{frames}}$$

– Confused assignments: speakers correctly detected but assigned to wrong clusters
– SCER is not counted by DER, and DER+SCER accounts for total errors

Approaches to diarization

- Clustering of time segments
  - Based on spectral features
    - MFCC, i-vector, d-vector, x-vector, etc.
  - Speaker overlapping segments are disregarded
  - 1-ch processing

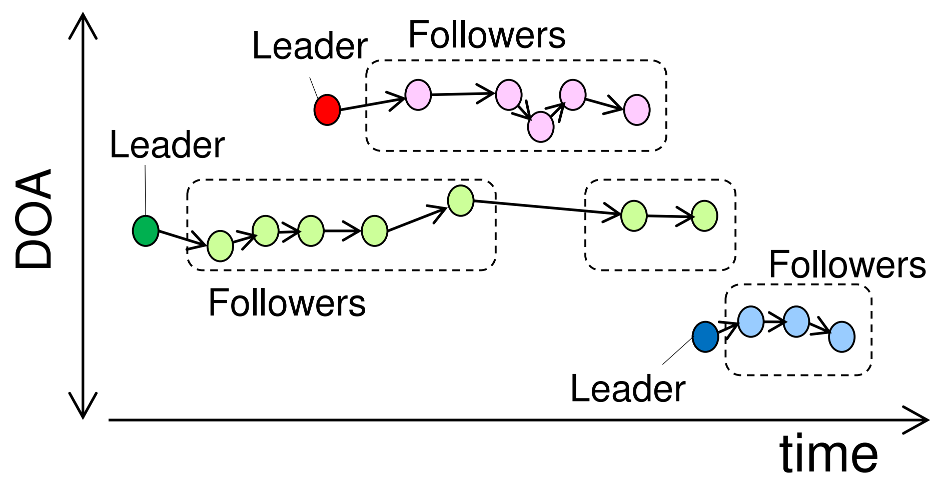
- Clustering of TF points
  - Mask-based source separation for unknown #sources
  - Speaker overlapping segments can be separated
  - 1-ch/multi-ch processings



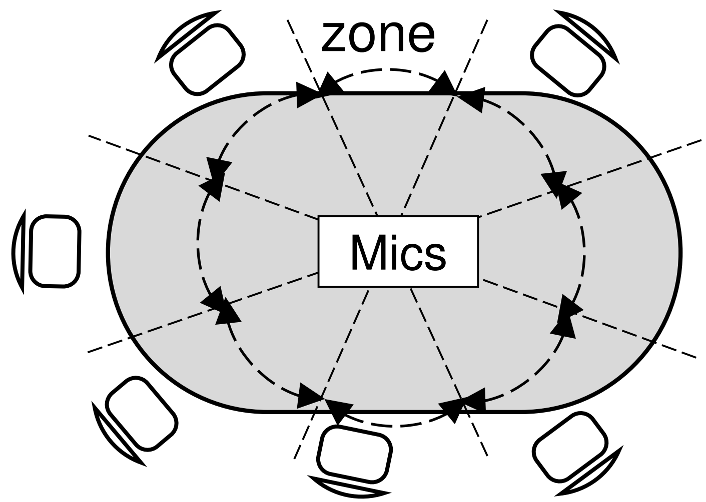
Clustering of TF bins (Multi-ch)

- Features for localization
  - DOAs, and many variants
- Online processing works
  - Multi-target tracking problem
    - Leader-follower clustering [Hori et al., 2012]
    - Probabilistic hypothesis density filter with random finite set [Evers and Naylor, 2018]
  - Zone-based speaker diarization [Fallon and Godsill, 2011, Ito et al., 2017]
    - Divides possible speaker locations into pre-determined zones
    - VAD at each zone results in diarization

Leader-follower clustering



Zones for speaker diarization



Probabilistic spatial dictionary based diarization [Ito et al., 2017]

- Model of signal from each possible speaker location
  - Complex Watson distribution

$$p(\tilde{\mathbf{y}}_{tf}^{(k)}) = \mathcal{W}(\tilde{\mathbf{y}}_{tf}^{(k)}; \kappa_f^{(k)}, \mathbf{a}_f^{(k)})$$

$\mathbf{a}_f^{(k)}$: parameter for RIR (dictionary, pretrained)

$\kappa_f^{(k)}$: parameter for variance (dictionary, pretrained)

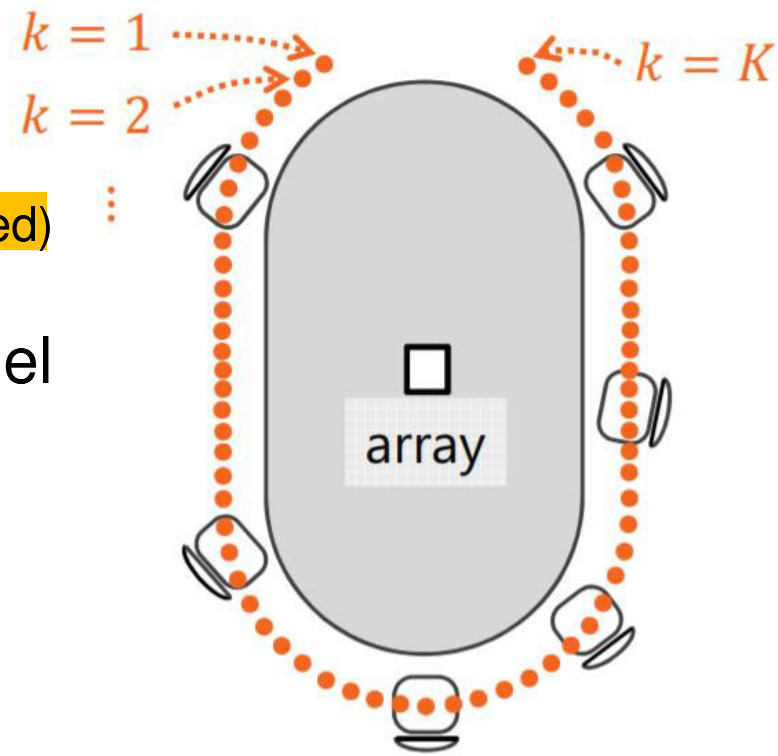
- Model of meeting recording: mixture model

$$p(\tilde{\mathbf{y}}_{tf}) = \sum_{k=1}^{K} \alpha_t^{(k)} \mathcal{W}(\tilde{\mathbf{y}}_{tf}; \kappa_f^{(k)}, \mathbf{a}_f^{(k)})$$

$\alpha_t^{(k)}$: mixture weight (estimated from test data) which indicates active speaker locations

→ Useful for online diarization

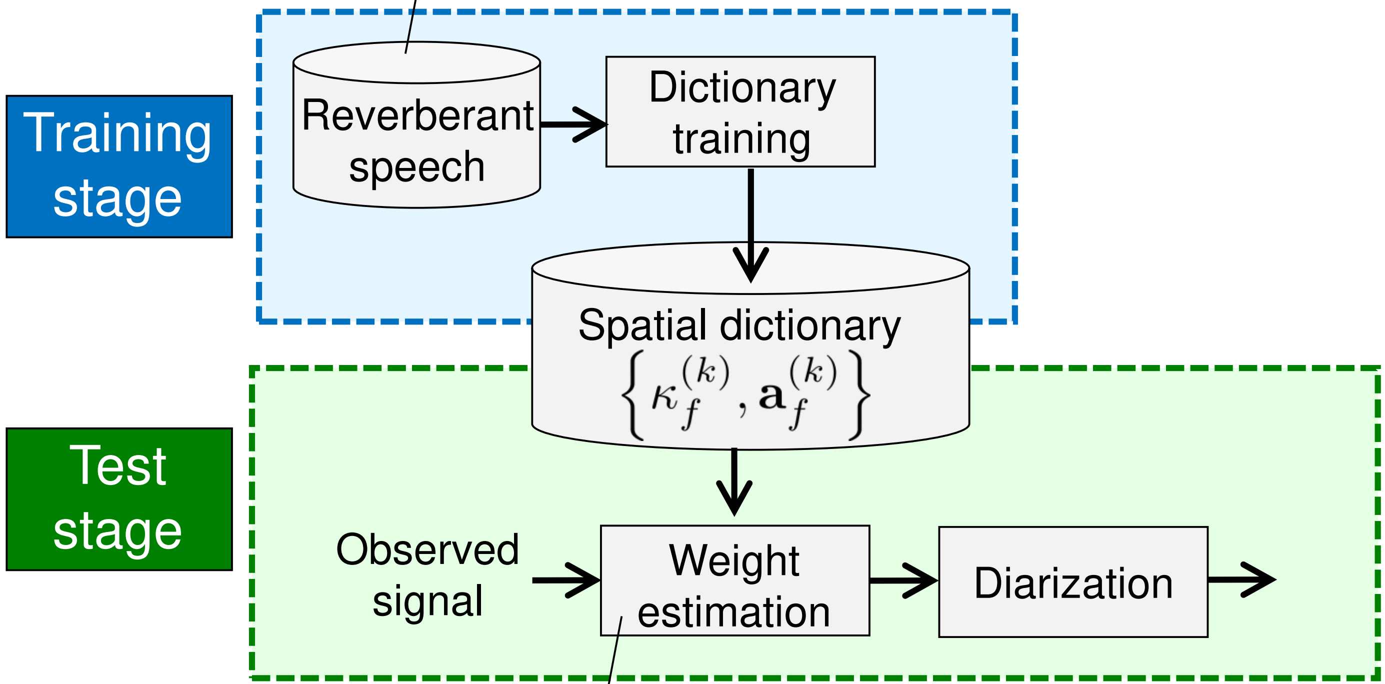
Recording condition



k : possible speaker location

Processing diagram of probabilistic spatial dictionary based diarization

Simulated microphone signals (with a plain wave assumption) can be used for the training



Posterior of source location: $\alpha_t^{(k)} = \sum_f \left\{ \frac{\mathcal{W}\left(\tilde{\mathbf{y}}_{tf}; \kappa_f^{(k)}, \mathbf{a}_f^{(k)}\right)}{\sum_{k'=1}^{K} \mathcal{W}\left(\tilde{\mathbf{y}}_{tf}; \kappa_f^{(k')}, \mathbf{a}_f^{(k')}\right)} \right\}$

DERs under reverberant babble noise condition

Reverberation time: 500 ms
Length of meeting: 15-20 min
SNR: 3-15 dB

#mics: 8
K=65
Information on chair locations is given

Session ID	#Speakers	Noise level (babble noise)	DER	
			Leader-follower clustering [Hori 2012]	Probabilistic spatial dictionary
1	6	No noise	46.8 %	**9.3 %**
2			64.6 %	**12.2 %**
3	5	Low	23.8 %	**17.2 %**
4			47.5 %	**18.9 %**
5	6		62.6 %	**15.6 %**
6	4	High	70.9 %	**27.7 %**
7			73.6 %	**24.8 %**
8	6		67.2 %	**18.9 %**

Discussion

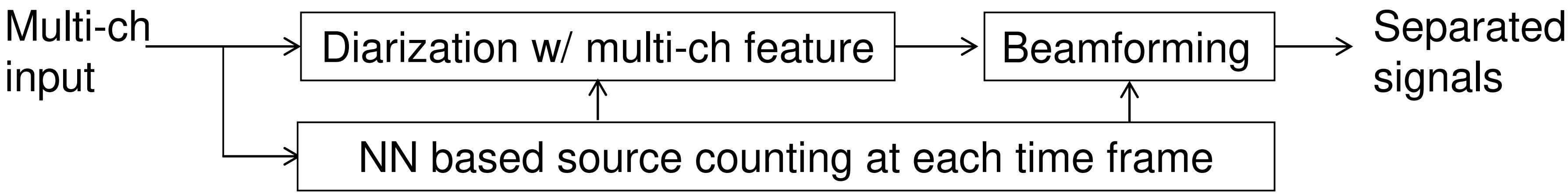
- 1-ch processing
  - Use of neural network is a key to successful diarization
    - End-to-end neural processing is also investigated
  - Treatment of adverse noise conditions is still a challenging problem

- Multi-ch processing
  - Spatial features work effectively even under noisy reverberant envs
  - Hard to track speakers who move with no utterance

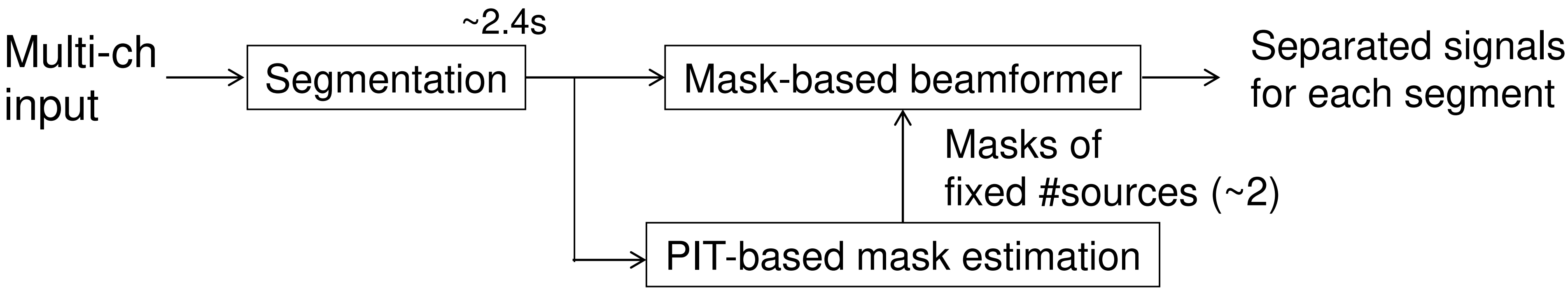
Integration of 1-ch and multi-ch approaches should be explored
- only a few attempts made so far

Meeting analysis based on source separation with integration of NN and microphone array

- NN-based source counting is combined with beamforming [Chazan et al., 2018]



- Segment-wise separation of fixed #sources based on NN and beamforming [Yoshioka et al., 2018]
  - Applicable without performing source counting or diarization



Software

- JHU diarization system (DIHARD-II challenge baseline)
  - https://github.com/iiscleap/DIHARD_2019_baseline_alltracks
  - Based on JHU diarization system developed for the DIHARD-I challenge, and prepared for the DIHARD-II challenge by Ganapathy et al.
  - Segmentation refinement block is omitted

Table of contents

Part VI.
Other Topics

Reinhold Haeb-Umbach

Table of contents in part VI

- NN supported enhancement: Overcome need for parallel clean and distorted training data
  - Motivation
  - Joint training
  - Teacher-student approach
  - Direct optimization of likelihood
- Should we do speech enhancement also on the ASR training data?

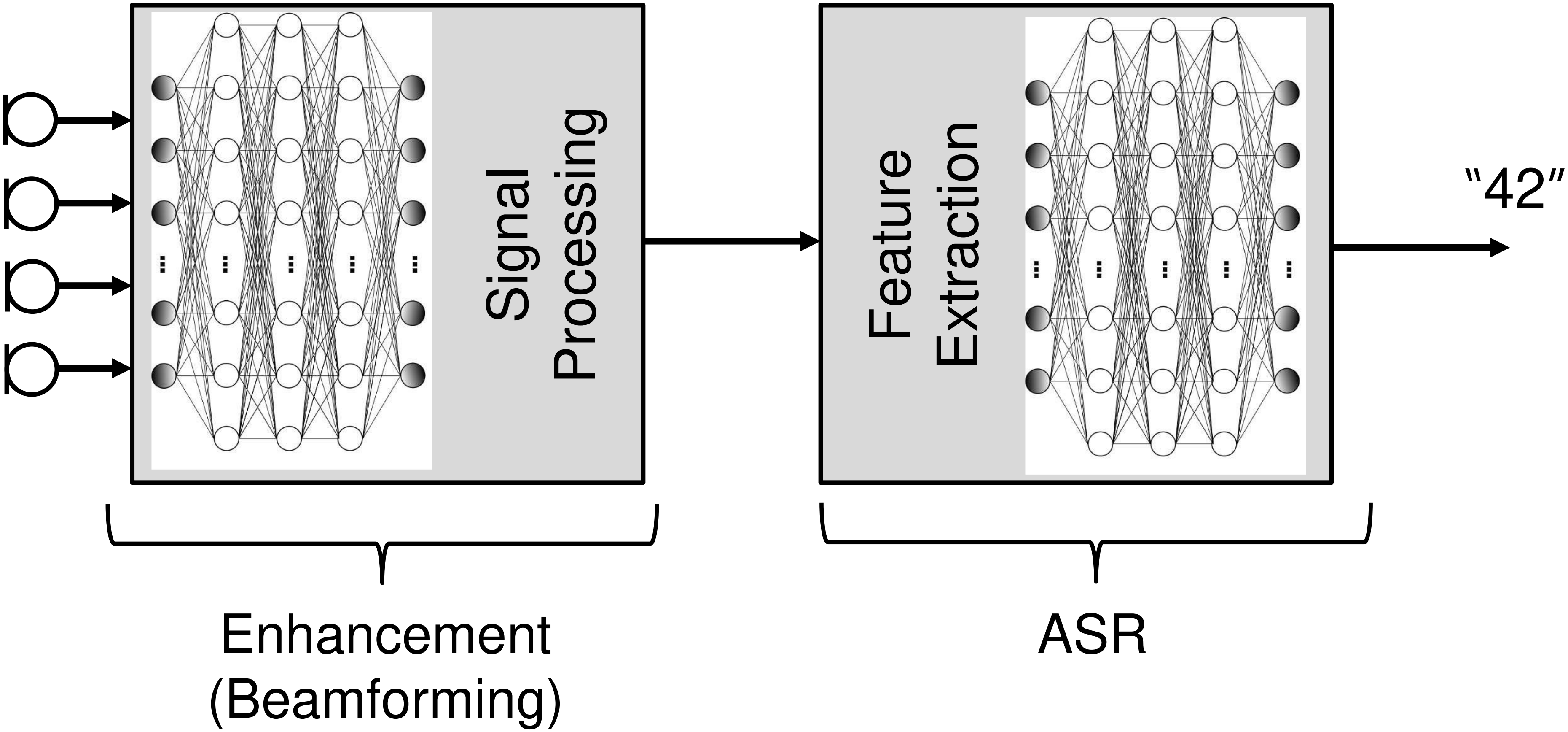
Table of contents in part VI

Motivation

- We have seen different uses of neural networks in enhancement
  - E.g., speech presence probability (mask) estimation
- Those networks were trained by supervised learning
  - Corrupted signal at input
  - Desired/clean signal as target
- This requires parallel (clean and distorted) data
  - Which is unavailable for real recordings of distorted speech
  - Training only on simulated (= artificially distorted) data possible
- Thus
  - No training on real recordings of distorted speech possible
  - Certain effects are hard, if impossible, to realistically simulate
    - e.g., Lombard speech

Goal: Get rid of need for parallel data in NN training!

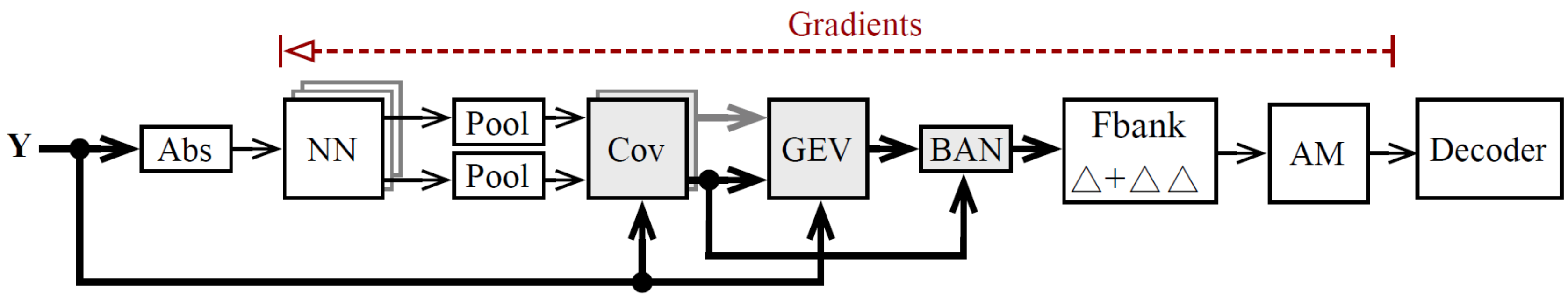
Option 1: Joint training



- Train NNs in front-end and back-end jointly
- Back-propagate gradient of cross entropy loss all the way to enhancement NN

Example NN-supported beamforming

[Heymann et al. 2017a, Ochiai et al., 2017]



- Gradient passed through signal processing tasks
  - ASR feature extraction
  - Beamforming
- Complex-valued gradients
  - See [Boeddeker et al., 2017] for a large collection of complex-valued gradients of various operations

Discussion

- Possible advantages of joint training
  - Parallel clean and noisy data no longer required
  - Training on real recordings of distorted speech
  - Mask estimator trained with criterion closer related to WER

- Possible disadvantages of joint training
  - Weaker acoustic model (AM)
    - Beamforming reduces the number of input channels to one. Thus fewer training data for acoustic model (AM)
    - Beamforming improves SNR, thus AM exposed to less variability
  - Weaker beamformer
    - AM learns to ignore certain distortions, thus beamformer does not need to remove them, meaning that beamforming is less effective in cleaning the data

WER results on CHiME-4

	Beamformer trng	**AM traning**	**Eval Simu**	**Eval Real**
parallel data required				
(a)	i) independent	i) independent on unenh. data	6.8	7.3
(b)	i) independent	ii) indep. on enhanced data	6.6	8.9
no parallel data required				
(c)	i) jointly from scratch	i) jointly from scratch	6.9	9.1
(d)	ii) using gradient from AM	i) separate on unenh. data	7.4	7.6

Training order: first i), then ii)

(a) & (c) Joint training degrades performance, in particular on real data

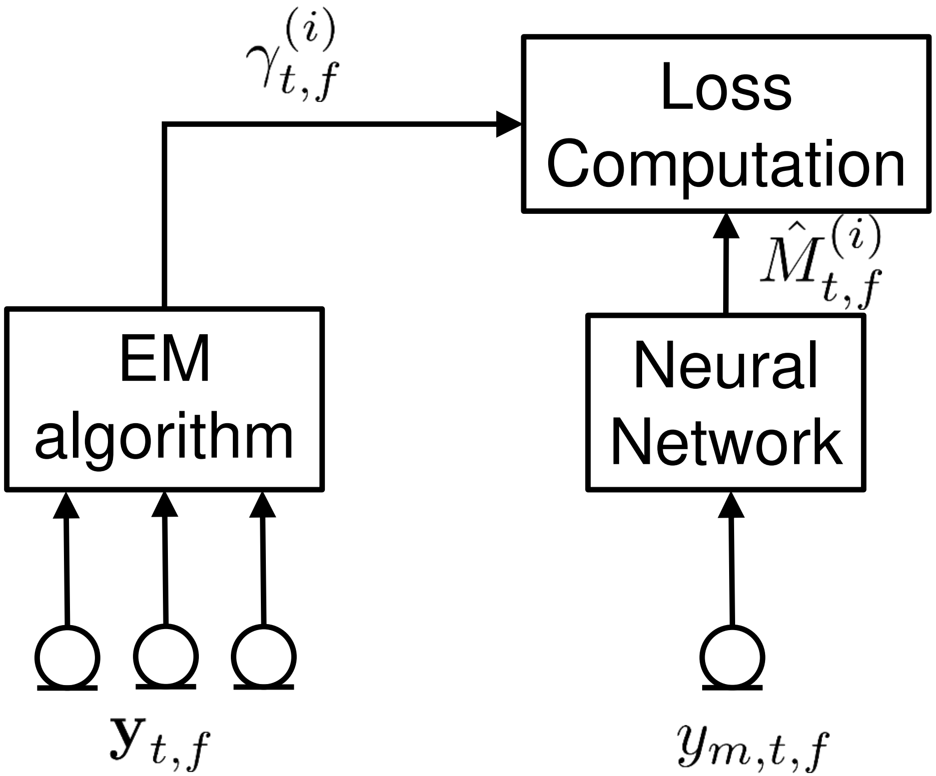
(b) & (d) The cause appears to be the weaker AM;

degradation can be reduced if AM sees enough variability in training

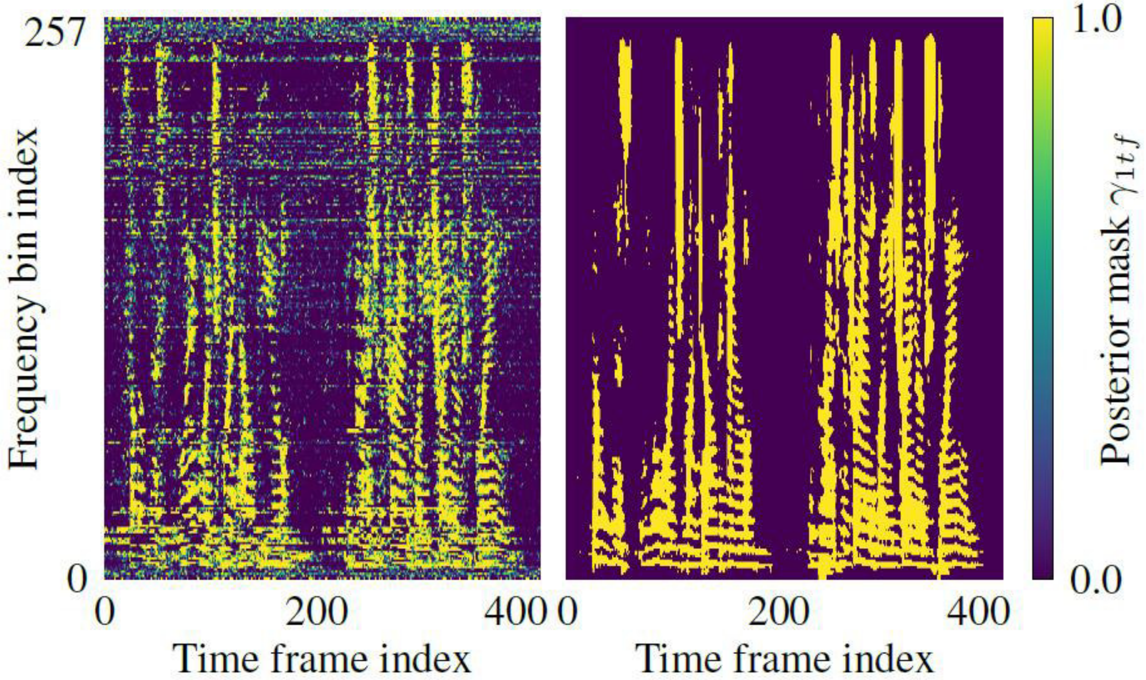
Option 2: Teacher – student approach

[Drude et al., 2019a, Seetharaman et al., 2019, Tzinis et al., 2019]

- Speaker presence probs ($\gamma_{t,f}^{(i)}$) obtained from spatial mixture model used as training targets of NN mask estimator



Example result for BSS [Drude et al., 2019a]



Teacher:
spatial mixture model

Student:
neural network

Results [Drude et al., 2019a]

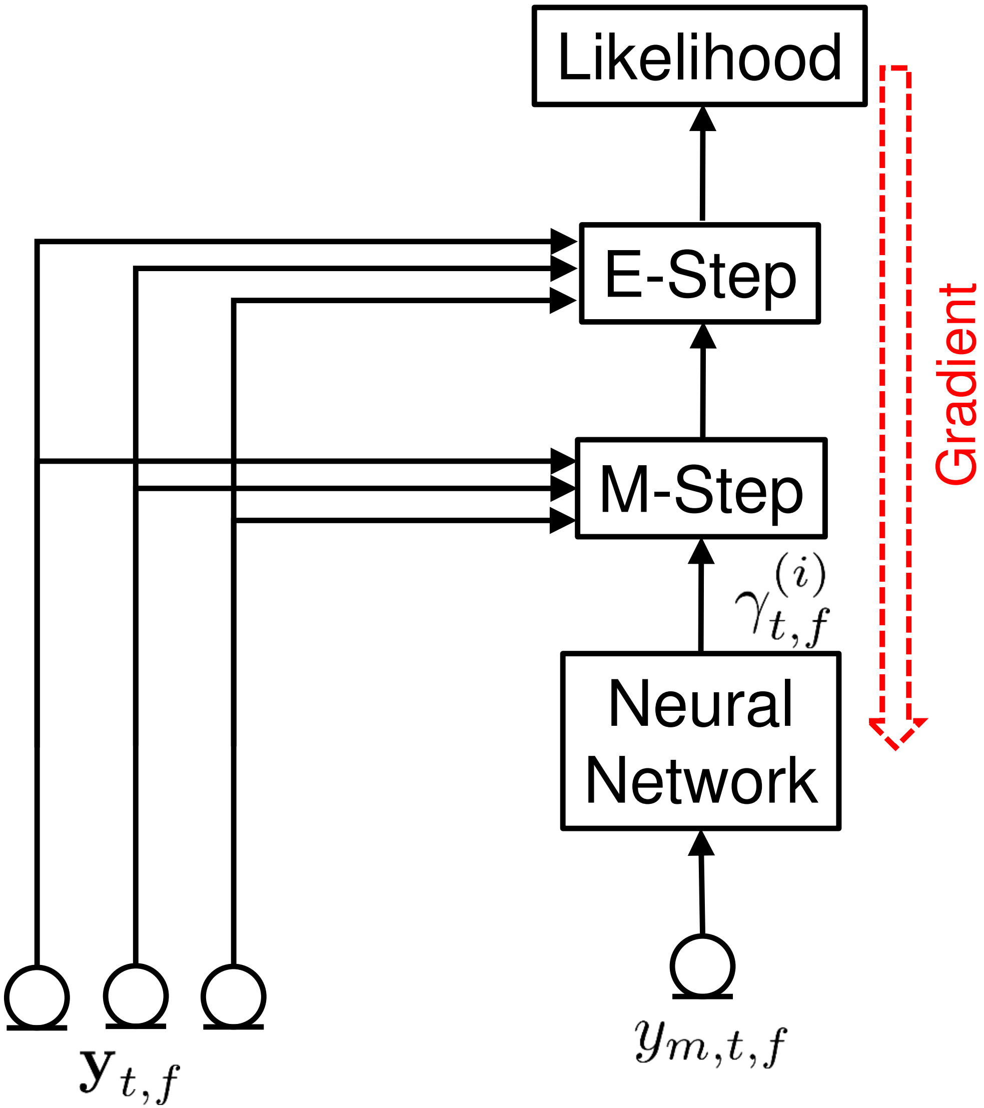
- Database: spatialized multi-channel wsj-2mix
- Source extraction via beamforming

	Model	Training	Initialization on test utt.	WER [%]
(a)	spatial mixt. model	-	random	28.0
(b)	deep clustering	Supervised	-	26.5
(c)	deep clustering	taught by mixt. model	-	29.0
(d)	spatial mixt. model	-	deep clustering from (c)	20.7
(e)	spatial mixture model	-	oracle ideal binary mask	19.9

(d) On test utterance, first apply DC to obtain initial values for $\gamma_{t,f}^{(i)}$.
Then run EM to obtain updated $\gamma_{t,f}^{(i)}$.

Option 3: Direct optimization of likelihood

[Drude et al., 2019b, session Tue-O-3-5]



- Optimize likelilhood of spatial mixture model
- Backpropagate gradient of likelihood through E-step and M-step of spatial mixture model to class affiliation posteriors and then to NN parameters
- Optional: additional EM-step at inference time on test utterance

Results [Drude et al., 2019b]

- Beamforming
- CHiME-4 real test set
- Additional EM step on test utterance

Estimator of $\gamma_{t,f}^{(i)}$	**Training**	**WER [%]**
spatial mixture model	-	13.0
neural network	Oracle masks	7.7
neural network	teacher-student	7.9
neural network	likelihood	7.8

Table of contents in part VI

Enhancement on ASR training data?

Pros:

- Acoustic model can learn artifacts of the enhancement
- Cleaner training data → better alignments → better models

Cons:

- Acoustic model is exposed to less variability
- Can reduce the amount of training data (e.g., if only the beamformed signal is used for training instead of all raw channels)

Example results

- Beamforming on CHiME-4

	Training Data	WER [%] Eval Simu	WER [%] Eval Real
(a)	all six channels	6.8	7.3
(b)	all six channels + beamformed	6.4	7.7
(c)	single channel	6.9	7.6
(d)	beamformed only	6.9	9.6
(e)	clean	11.7	16.3

(a) & (d) enhancement in trng hurts performance, in particular on real data

(c) & (d) The reason is not fewer trng data, but removal of variability

But look at these results

- CHiME-5
  - Extremely degraded: lots of overlapped speech, reverberation, …
  - Weak enhancement: (BeamformIt: variant of Delay-Sum-Beamformer)
  - Strong: guided source separation [Kanda et al., 2019, session Tue-O-3-5]

WER [%] on eval	Enhancement in Test		
Enhancemnt in Trng	**none**	**weak**	**strong**
none	59.9	59.7	51.6
weak (BeamformIt)	59.1	58.5	49.9
strong (GSS)	73.1	69.2	45.7

- Matched is best
- Enhancement in trng beneficial, as long as it is weaker than in test
- If data is extremely poor, enhance for alignment extraction, not for NN training itself

Summary of part VI

- There are several options to avoid the need for parallel clean and noisy training data
  - Direct optimization of likelihood is the (arguably) conceptually most appealing one
    - Sofar only developed for beamforming
  - Joint training of front end NN and acoustic model is tricky

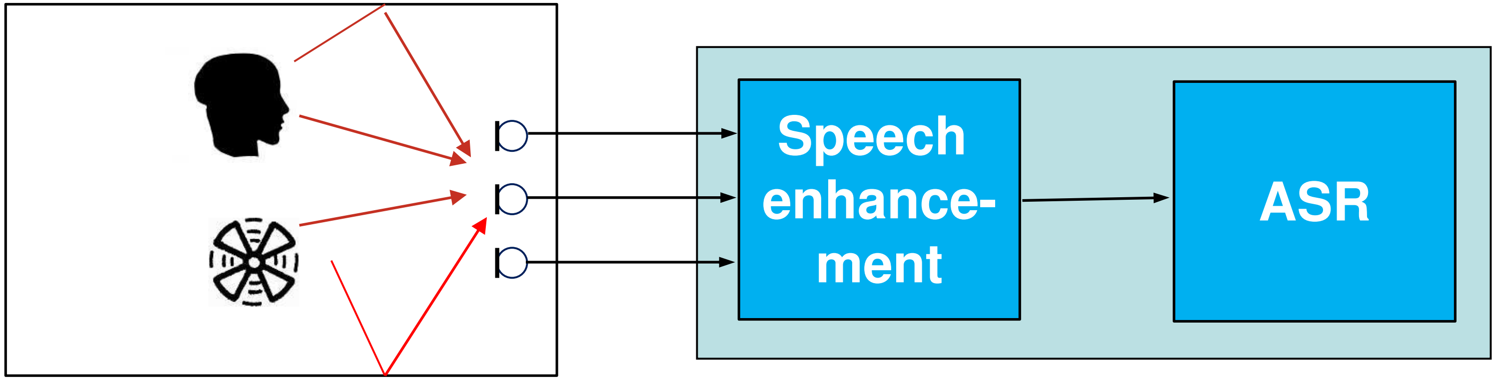
- Enhancement of ASR training data
  - Is only advisable as long as the training data contains still at least as much variability as the test data

Table of contents

Part VII.
Summary

Reinhold Haeb-Umbach &
Tomohiro Nakatani

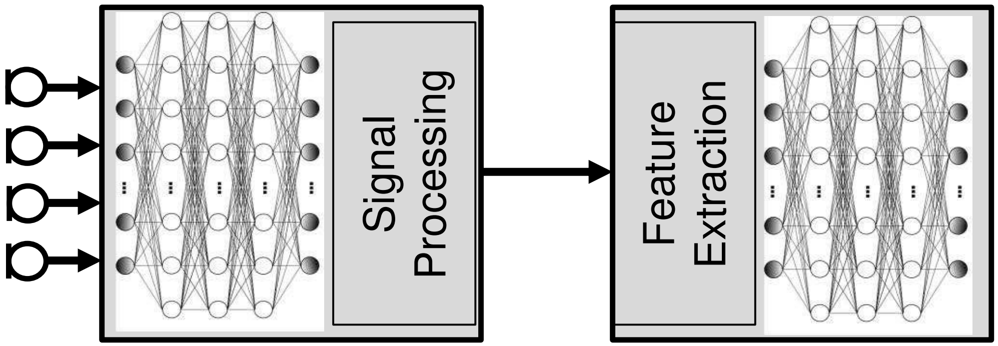
Combination of speech enhancement and ASR



- Speech enhancement for ASR is recommended
  - If phase (spatial) information present in multi-channel input can be exploited, which would be lost in traditional ASR feature repesentations
    - Acoustic beamforming
  - If distortions exist, which introduce huge variability in frame-based ASR processing
    - Reverberation
    - Multiple concurrent speakers
  - Where excellent signal processing solutions exist (which can be further improved by deep learning)
    - MIMO acoustic echo cancellation (not treated in this tutorial)

Speech enhancement by DSP and DNN

- We have seen many examples in this tutorial of combinations of traditional signal processing and deep learning techniques
- Compared to pure DSP they offer several advantages
  - Leverage training data
  - Overcome restrictions of simplifying modeling assumptions otherwise necessary to obtain tractable solutions
- Compared to pure DNN they offer the following advantages
  - Less data hungry
  - Better interpretable
  - Can adapt to test data via unsupervised learning



Trends

- End-to-end trained (enhancement + ASR) systems
- DNNs will gain ever more grounds
  - Future DNNs may include microphone array functionality
  - Compact DNN on device
- Multimodal processing
    - Vision, bio sensors, brain activities, etc.

Future challenges

- Get rid of simplifying assumptions
  - E.g., #speakers constant and known in a mixture
  - Transcribe realistic meeting scenarios
- Leverage huge amounts of unlabeled speech and audio
  - From supervised learning to unsupervised learning enabled by signal processing
- Cope with more challenging environments / applications
  - E.g., CHiME-5 dinner party transcription (WER > 40%)
- Lack of domain/environment specific training data
  - „Speech processing in the wild“

Fortunately, there is still a lot to be done!

Get started[1], and enjoy working in this fascinating field!

Tutorial preparation

[1] Get hands-on experience using the various pointers to software found in this tutorial!

REFERENCES

[Abed-Meraim et al., 1997] Abed-Meraim, K., Moulines, E., and Loubaton, P. (1997). Prediction error method for second-order blind identification. *IEEE Trans. Signal Process.*, (3):694–705.

[Aleksic et al., 2015] Aleksic, P. S., Ghodsi, M., Michaely, A. H., Allauzen, C., Hall, K. B., Roark, B., Rybach, D., and Moreno, P. J. (2015). Bringing contextual information to Google speech recognition. In *Proc. Interspeech*.

[Allen and Berkley, 1979] Allen, J. B. and Berkley, D. (1979). Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Amer.*, 65(4):943–950.

[Andersen et al., 2017] Andersen, A. H., de Haan, J. M., Tan, Z., and Jensen, J. (2017). A non-intrusive short-time objective intelligibility measure. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5085–5089.

[Araki, 2016] Araki, S. (2016). Spatial correlation model based observation vector clustering and MVDR beamforming for meeting recognition. *IEEE ICASSP*, pages 385–389.

[Audio Software Engineering and Siri Speech Team, 2018] Audio Software Engineering and Siri Speech Team (2018). Optimizing Siri on HomePod in far-field settings.

[Avargel and Cohen, 2007] Avargel, Y. and Cohen, I. (2007). On multiplicative transfer function approximation in the short-time fourier transform domain. *IEEE Signal Process. Lett.*, 14:337–340.

[Bahmaninezhad et al., 2019] Bahmaninezhad, F., Wu, J., Gu, R., Zhang, S., Xu, Y., Yu, M., and Yu, D. (2019). A comprehensive study of speech separation: spectrogram vs waveform separation. *CoRR*, abs/1905.07497.

[Barker et al., 2017] Barker, J., Marxer, R., Vincent, E., and Watanabe, S. (2017). The third "CHiME" speech separation and recognition challenge: Analysis and outcomes. *Computer Speech and Language*, 46:605–626.

[Barker et al., 2013] Barker, J., Vincent, E., Ma, N., Christensen, H., and Green, P. (2013). The PASCAL CHiME speech separation and recognition challenge. *Computer Speech & Language*, 27(3):621–633.

[Barker et al., 2018] Barker, J., Watanabe, S., Vincent, E., and Trmal, J. (2018). The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines. In *Proc. Interspeech*, pages 1561–1565.

[Benesty et al., 2001a] Benesty, J., Gänsler, T., Morgan, D., Sondhi, M., and Gay, S. (2001a). *Advances in Network and Acoustic Echo Cancellation*, chapter Multichannel Acoustic Echo Cancellation. Springer.

[Benesty et al., 2001b] Benesty, J., Gänsler, T., Morgan, D., Sondhi, M., and Gay, S. (2001b). *Advances in network and acoustic echo cancellation*. Springer.

[Boeddeker et al., 2018] Boeddeker, C., Erdogan, H., Yoshioka, T., and Haeb-Umbach, R. (2018). Exploring practical aspects of neural mask-based beamforming for far-field speech recognition. In *Proc. ICASSP*.

[Boeddeker et al., 2017] Boeddeker, C., Hanebrink, P., Drude, L., Heymann, J., and Haeb-Umbach, R. (2017). Optimizing neural-network supported acoustic beamforming by algorithmic differentiation. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*.

[Boll, 1979] Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120.

[Bradley et al., 2003] Bradley, J. S., Sato, H., and Picard, M. (2003). On the importance of early reflections for speech in rooms. *The Journal of the Acoustic Sociaty of America*, 113:3233–3244.

[Braun and Habets, 2018] Braun, S. and Habets, E. A. P. (2018). Linear prediction-based online dereverberation and noise reduction using alternating kalman filters. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 26(6):240–251.

[Carletta, 2006] Carletta, J. (2006). Announcing the AMI meeting corpus. *The ELRA Newsletter*, 11(1):3–5.

[Caroselli et al., 2017a] Caroselli, J., Shafran, I., Narayanan, A., and Rose, R. (2017a). Adaptive multichannel dereverberation for automatic speech recognition. In *Proc. Interspeech*.

[Caroselli et al., 2017b] Caroselli, J., Shafran, I., Narayanan, A., and Rose, R. (2017b). Adaptive multichannel dereverberation for automatic speech recognition. In *Proc. Interspeech*.

[Chazan et al., 2018a] Chazan, S. E., Goldberger, J., and Gannot, S. (2018a). DNN-based concurrent speaker detector and its application to speaker extraction with LCMV beamforming. In *Proc. ICASSP*.

[Chazan et al., 2018b] Chazan, S. E., Goldberger, J., and Gannot, S. (2018b). DNN-based concurrent speakers detector and its application to speaker extraction with LCMV beamforming. *IEEE ICASSP*, pages 6712–6716.

[Chen et al., 2014] Chen, G., Parada, C., and Heigold, G. (2014). Small-footprint keyword spotting using deep neural networks. In *Proc. ICASSP*, pages 4087–4091.

[Chen et al., 2017] Chen, Z., Luo, Y., and Mesgarani, N. (2017). Deep attractor network for single-microphone speaker separation. In *Proc. ICASSP*, pages 246–250.

[Cherry, 1953] Cherry, E. (1953). Some experiments on the recognition of speech with one and two ears. *Journal of the Acoustical Society of America*, 25(5):975–979.

[Chetupalli and Sreenivas, 2019] Chetupalli, S. R. and Sreenivas, T. V. (2019). Late reverberation cancellation using bayesian estimation of multi-channel linear predictors and student's t-source prior. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 27(6).

[Chiu et al., 2017] Chiu, C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, K., Jaitly, N., Li, B., Chorowski, J., and Bacchiani, M. (2017). State-of-the-art speech recognition with sequence-to-sequence models. *CoRR*, abs/1712.01769.

[Delcroix et al., 2015] Delcroix, M., Yoshioka, T., Ogawa, A., Kubo, Y., Fujimoto, M., Ito, N., Kinoshita, K., Espi, M., Araki, S., Hori, T., and Nakatani, T. (2015). Strategies for distant speech recognition in reverberant environments. *EURASIP J. Adv. Signal Process*, Article ID 2015:60, doi:10.1186/s13634-015-0245-7.

[Dietzen et al., 2018] Dietzen, T., Doclo, S., Moonen, M., and van Waterschoot, T. (2018). Joint multimicrophone speech dereverberation and noise reduction using integrated sidelobe cancellation and linear prediction. In *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, page 221225.

[Drude et al., 2018] Drude, L., Boeddeker, C., Heymann, J., Kinoshita, K., Delcroix, M., Nakatani, T., and Haeb-Umbach, R. (2018). Integration neural network based beamforming and weighted prediction error dereverberation. In *Proc. Interspeech*.

[Drude and Haeb-Umbach, 2017] Drude, L. and Haeb-Umbach, R. (2017). Tight integration of spatial and spectral features for bss with deep clustering embeddings. In *Proc. Interspeech*.

[Drude and Haeb-Umbach, 2019] Drude, L. and Haeb-Umbach, R. (2019). Integration of neural networks and probabilistic spatial models for acoustic blind source separation. *IEEE J. Sel. Topics Signal Process.*, 13(4):815–826.

[Drude et al., 2019a] Drude, L., Hasenclever, D., and Haeb-Umbach, R. (2019a). Unsupervised training of a deep clustering model for multichannel blind source separation. In *Proc. ICASSP*.

[Drude et al., 2019b] Drude, L., Heymann, J., and Haeb-Umbach, R. (2019b). Unsupervised training of neural mask-based beamforming. In *Proc. Interspeech*.

[Drude et al., 2018] Drude, L., von Neumann, T., and Haeb-Umbach, R. (2018). Deep attractor networks for speaker re-identification and blind source separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11–15.

[Du et al., 2016] Du, J., Tu, Y., Dai, L., and Lee, C. (2016). A regression approach to single-channel speech separation via high-resolution deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(8):1424–1437.

[Duong et al., 2010] Duong, N. Q., Vincent, E., and Gribonval, R. (2010). Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 18(7):1830–1840.

[Elko, 2001] Elko, G. (2001). Microphone arrays. In *Proc. International Workhop on Hands-free Speech Communication*, Kyoto, Japan.

[Ephraim and Malah, 1984] Ephraim, Y. and Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Process.*, 32(6):1109–1121.

[Erdogan et al., 2015] Erdogan, H., Hershey, J. R., Watanabe, S., and Le Roux, J. (2015). Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 708–712.

[Erdogan et al., 2016] Erdogan, H., Hershey, J. R., Watanabe, S., Mandel, M. I., and Le Roux, J. (2016). Improved MVDR beamforming using single-channel mask prediction networks. In *Proc. Interspeech*, pages 1981–1985.

[Evers and Naylor, 2018] Evers, C. and Naylor, P. A. (2018). Acoustic slam. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 26:1484–1498.

[Fallon and Godsill, 2011] Fallon, M. F. and Godsill, S. J. (2011). Acoustic source localization and tracking of a time-varying number of speakers. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 20(4):1409–1415.

[Fernández et al., 2007] Fernández, S., Graves, A., and Schmidhuber, J. (2007). An application of recurrent neural networks to discriminative keyword spotting. In *Proc. of the 17th International Conference on Artificial Neural Networks*, ICANN'07, pages 220–229, Berlin, Heidelberg. Springer-Verlag.

[Gannot, 2010] Gannot, S. (2010). Multi-microphone speech dereverberation using eigen-decomposition. In *Naylor P., Gaubitch N. (eds) Speech Dereverberation. Signals and Commmunication Technology*. Springer, London.

[Gannot et al., 2001] Gannot, S., Burshtein, D., and Weinstein, E. (2001). Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, 49(8):1614–1626.

[Gannot et al., 2017] Gannot, S., Vincent, E., Markovich-Golan, S., and Ozerov, A. (2017). A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 25(4):692–730.

[Gillespie et al., 2001] Gillespie, B. W., Malvar, H. S., and Florêncio, D. A. F. (2001). Speech deconvolution via maximum-kurtosis subband adaptive filtering. In *Proc. ICASSP*, volume 6, pages 3701–3704.

[Guoy et al., 2018] Guoy, J., Kumatani, K., Sun, M., Wu, M., Raju, A., Stroem, N., and Mandal, A. (2018). Time-delayed bottleneck highway networks using a dft feature for keyword spotting. In *Proc. ICASSP*.

[Hadad et al., 2014] Hadad, E., Heese, F., Vary, P., and Gannot, S. (2014). Multichannel audio database in various acoustic environments. *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 313–317.

[Haeb-Umbach, 2018] Haeb-Umbach, R. (2018). Neural network supported acoustic beamforming and source separation for ASR.

[Harper, 2015] Harper, M. (2015). The automatic speech recogition in reverberant environments (ASpIRE) challenge. In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, pages 547–554. IEEE.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

[Hershey et al., 2016] Hershey, J., Chen, Z., Roux, J. L., and Watanabe, S. (2016). Deep clustering: discriminative embeddings for segmentation and separation. In *Proc. ICASSP*. IEEE.

[Heymann et al., 2018] Heymann, J., Bacchiani, M., and Sainath, T. (2018). Performance of mask based statistical beamforming in a smart home scenario. In *Proc. ICASSP*.

[Heymann et al., 2017a] Heymann, J., Drude, L., Boeddeker, C., Hanebrink, P., and Haeb-Umbach, R. (2017a). BEAMNET: End-to-end training of a beamformer-supported multi-channel ASR system. In *Proc. ICASSP*.

[Heymann et al., 2015] Heymann, J., Drude, L., Chinaev, A., and Haeb-Umbach, R. (2015). Blstm supported gev beamformer front-end for the 3rd CHiME challenge. In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*.

[Heymann et al., 2016] Heymann, J., Drude, L., and Haeb-Umbach, R. (2016). Neural network based spectral mask estimation for acoustic beamforming. In *Proc. ICASSP*.

[Heymann et al., 2017b] Heymann, J., Drude, L., and Haeb-Umbach, R. (2017b). A generic neural acoustic beamforming architecture for robust multi-channel speech processing. *Computer Speech & Language*.

[Heymann et al., 2019] Heymann, J., Drude, L., Haeb-Umbach, R., Kinoshita, K., and Nakatani, T. (2019). Joint optimization of neural network-based WPE dereverberation and acoustic model for robust online ASR. *IEEE ICASSP*.

[Higuchi et al., 2016] Higuchi, T., Ito, N., Yoshioka, T., and Nakatani, T. (2016). Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise. In *Proc. ICASSP*, pages 5210–5214.

[Hikichi et al., 2007] Hikichi, T., Delcroix, M., and Miyoshi, M. (2007). Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations. *EURASIP J. Adv. Signal Process.*

[Hinton et al., 2012] Hinton, G., Deng, L., Yu, D., Dahl, G. E., r. Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.*, 29(6):82–97.

[Hori et al., 2012] Hori, T., Araki, S., Yoshioka, T., Fujimoto, M., Watanabe, S., Oba, T., Ogawa, A., Otsuka, K., Mikami, D., Kinoshita, K., Nakatani, T., Nakamura, A., and Yamato, J. (2012). Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 20(2):499–513.

[Isik et al., 2016] Isik, Y., Le Roux, J., Chen, Z., Watanabe, S., and Hershey, J. R. (2016). Single-channel multi-speaker separation using deep clustering. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

[Ito et al., 2017] Ito, N., Araki, S., Delcroix, M., and Nakatani, T. (2017). Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments. In *Proc. ICASSP*.

[Ito et al., 2013] Ito, N., Araki, S., and Nakatani, T. (2013). Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3238–3242.

[Ito et al., 2016] Ito, N., Araki, S., and Nakatani, T. (2016). Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing. In *European Signal Processing Conference (EUSIPCO)*, pages 1153–1157. IEEE.

[J.Barker et al., 2017] J.Barker, Marxer, R., Vincent, E., and Watanabe, S. (2017). Multi-microphone speech recognition in everyday environments. *Computer Speech & Language*, 46:386–387.

[Juang and Nakatani, 2007] Juang, B.-H. and Nakatani, T. (2007). Joint source-channel modeling and estimation for speech dereverberation. In *Prof. International Symposium on Circuits and Systems (ISCAS)*, pages 2990–2993.

[Jukić et al., 2015] Jukić, A., van Waterschoot, T., Gerkmann, T., and Doclo, S. (2015). Multi-channel linear prediction-based speech dereverberation with sparse priors. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 23(9):1509–1520.

[Kanda et al., 2019] Kanda, N., Boeddeker, C., Heitkaemper, J., Fujita, Y., Horiguchi, S., Nagamatsu, K., and Haeb-Umbach, R. (2019). Guided source separation meets a strong ASR backend: Hitachi/Paderborn University joint investigation for dinner party asr. In *Proc. Interspeech*.

[Kim et al., 2017] Kim, C., Misra, A., Chin, K., Hughes, T., Narayanan, A., Sainath, T., and Bacchiani, M. (2017). Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home. In *Proc. Interspeech*, pages 379–383.

[Kinoshita et al., 2016] Kinoshita, K., Delcroix, M., Gannot, S., Habets, E., Haeb-Umbach, R., Kellermann, W., Leutnant, V., Maas, R., Nakatani, T., Raj, B., Sehr, A., and Yoshioka, T. (2016). 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*.

[Kinoshita et al., 2017] Kinoshita, K., Delcroix, M., Kwon, H., Mori, T., and Nakatani, T. (2017). Neural network-based spectrum estimation for online WPE dereverberation. *Proc. Interspeech*.

[Kinoshita et al., 2009] Kinoshita, K., Delcroix, M., Nakatani, T., and Miyoshi, M. (2009). Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 17(4):534–545.

[Kinoshita et al., 2018] Kinoshita, K., Drude, L., Delcroix, M., and Nakatani, T. (2018). Listening to each speaker one by one with recurrent selective hearing networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5064–5068.

[Ko et al., 2015] Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Proc. Interspeech*, pages 3586–3589.

[Kodrasi and Doclo, 2017] Kodrasi, I. and Doclo, S. (2017). EVD-based multi-channel dereverberation of a moving speaker using different RETF estimation methods. In *Proc. Hands-free Speech Communications and Microphone Arrays (HSCMA)*.

[Kodrasi et al., 2013] Kodrasi, I., Goetze, S., and Doclo, S. (2013). Regularization for partial multichannel equalization for speech dereverberation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 21(9):1879–1890.

[Kolbæk et al., 2017a] Kolbæk, M., Tan, Z., and Jensen, J. (2017a). Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1):153–167.

[Kolbæk et al., 2017b] Kolbæk, M., Yu, D., Tan, Z.-H., and Jensen, J. (2017b). Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1901–1913.

[Kristjansson et al., 2006] Kristjansson, T., Hershey, J., Olsen, P., Rennie, S., and Gopinath, R. (2006). Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system. In *International Conference on Spoken Language Processing (SLT)*.

[Kumatani et al., 2012] Kumatani, K., McDonough, J., and Raj, B. (2012). Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. *IEEE Signal Process. Mag.*, 29(6):127–140.

[Kumatani et al., 2017] Kumatani, K., Panchapagesan, S., Wu, M., Kim, M., Strom, N., Tiwari, G., and Mandai, A. (2017). Direct modeling of raw audio with dnns for wake word detection. In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, pages 252–257.

[Le Roux et al., 2018a] Le Roux, J., Wichern, G., Watanabe, S., Sarroff, A. M., and Hershey, J. R. (2018a). Phasebook and friends: Leveraging discrete representations for source separation. *CoRR*, abs/1810.01395.

[Le Roux et al., 2018b] Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. R. (2018b). SDR - half-baked or well done? *CoRR*, abs/1811.02508.

[Le Roux et al., 2019] Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. R. (2019). SDR - half-baked or well done? In *Proc. ICASSP*.

[Lee and Seung, 1999] Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788.

[Li et al., 2015] Li, J., Deng, L., Haeb-Umbach, R., and Gong, Y. (2015). *Robust Automatic Speech Recognition*. Elsevier.

[Lincoln et al., 2005] Lincoln, M., McCowan, I., Vepa, J., and Maganti, H. (2005). The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments. In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*. IEEE.

[Liu et al., 2015] Liu, B., Hoffmeister, B., and Rastrow, A. (2015). Accurate endpointing with expected pause duration. In *Proc. Interspeech*.

[Liu et al., 2018] Liu, Y., Ganguly, A., Kamath, K., and Kristjansson, T. (2018). Neural network based time-frequency masking and steering vector estimation for two-channel MVDR beamforming. In *Proc. ICASSP*, pages 6717–6721.

[Loizou, 2013] Loizou, P. C. (2013). *Speech Enhancement – Theory and Practice*. CRC Press.

[Luo and Mesgarani, 2018] Luo, Y. and Mesgarani, N. (2018). Tasnet: Surpassing ideal time-frequency masking for speech separation. *CoRR*, abs/1809.07454.

[Maas et al., 2016] Maas, R., Parthasarathi, S. H. K., King, B., Huang, R., and Hoffmeister, B. (2016). Anchored speech detection. In *Proc. Interspeech*.

[Maas et al., 2017] Maas, R., Rastrow, A., Goehner, K., Tiwari, G., Joseph, S., and Hoffmeister, B. (2017). Domain-specific utterance end-point detection for speech recognition. In *Proc. Interspeech*.

[Maas et al., 2018] Maas, R., Rastrow, A., Ma, C., Lan, G., Goehner, K., Tiwari, G., Joseph, S., and Hoffmeister, B. (2018). Combining acoustic embeddings and decoding features for end-of-utterance detection in real-time far-field speech recognition systems. In *Proc. ICASSP*.

[Mallidi et al., 2018] Mallidi, S., Maas, R., Goehner, K., A., R., Matsoukas, S., and Hoffmeinster, B. (2018). Device-directed utterance detection. In *Proc. Interspeech*.

[Miyoshi and Kaneda, 1988] Miyoshi, M. and Kaneda, Y. (1988). Inverse filtering of room acoustics. *IEEE Trans. Audio, Speech, Signal, Process.*, 36(2):145152.

[Mogami et al., 2018] Mogami, S., Sumino, H., Kitamura, D., Takamune, N., Takamichi, S., Saruwatari, H., and Ono, N. (2018). Independent deeply learned matrix analysis for multichannel audio source separation. *EUSIPCO*.

[Nakatani, 2015] Nakatani, T. (2015). Boosting distant speech recognition using multiple microphones: Frontend approaches.

[Nakatani et al., 2017] Nakatani, T., Ito, N., Higuchi, T., Araki, S., and Kinoshita, K. (2017). Integrating DNN-based and spatial clustering-based mask estimation for robust mvdr beamforming. In *Proc. ICASSP*, pages 286–290.

[Nakatani and Kinoshita, 2019a] Nakatani, T. and Kinoshita, K. (2019a). A maximum likelihood convolutional beamformer for simultaneous denoising and dereverberation. In *Proc. European Signal Processing Conference (EUSIPCO)*.

[Nakatani and Kinoshita, 2019b] Nakatani, T. and Kinoshita, K. (2019b). Simultaneous denoising and dereverberation for low-latency applications using frame-by-frame online unified convolutional beamformer. In *Proc. Interspeech*.

[Nakatani and Kinoshita, 2019c] Nakatani, T. and Kinoshita, K. (2019c). A unified convolutional beamformer for simultaneous denoising and dereverberation. *IEEE Signal Processing Letters*, 26(6):903–907.

[Nakatani et al., 2012] Nakatani, T., Sehr, A., and Kellermann, W. (2012). Reverberant speech processing for human communication and automatic speech recognition.

[Nakatani et al., 2008] Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., and Juang, B.-H. (2008). Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation. In *Proc. ICASSP*.

[Nakatani et al., 2010] Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., and Juang, B.-H. (2010). Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 18(7):1717–1731.

[Narayanan and Wang, 2013a] Narayanan, A. and Wang, D. (2013a). Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *Proc. ICASSP*, pages 7092–7096.

[Narayanan and Wang, 2013b] Narayanan, A. and Wang, D. (2013b). Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *IEEE ICASSP*, pages 7092–7096.

[Narayanan and Wang, 2014] Narayanan, A. and Wang, D. (2014). Investigation of speech separation as a front-end for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):826–835.

[NIST Speech Group, 2007] NIST Speech Group (2007). Spring 2007 (rt-07) rich transcription meeting recognition evaluation plan.

[Nugraha et al., 2016] Nugraha, A. A., Liutkus, A., and Vincent, E. (2016). Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1652–1664.

[Ochiai et al., 2017] Ochiai, T., Watanabe, S., Hori, T., and Hershey, J. R. (2017). Multichannel end-to-end speech recognition. In *ICML*.

[Ochiai et al., 2017] Ochiai, T., Watanabe, S., Hori, T., Hershey, J. R., and Xiao, X. (2017). Unified architecture for multichannel end-to-end speech recognition with neural beamforming. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1274–1288.

[Parihar and Picone, 2002] Parihar, N. and Picone, J. (2002). Dsr front end lvcsr evaluation - au/384/02. *Aurora Working Group, European Telecommunications Standards Institute*.

[Pedersen et al., 2007] Pedersen, M., Larsen, J., Kjems, U., and Parra, L. (2007). A survey of convolutive blind source separation methods. *Multichannel Speech Processing Handbook*, pages 114–126.

[Petkov et al., 2019] Petkov, P., Tsiaras, V., Doddipatl, R., and Stylianou, Y. (2019). An unsupervised learning approach to neural-net-supported wpe dereverberation. In *Proc. ICASSP*.

[Ravanelli et al., 2015] Ravanelli, M., Cristoforetti, L., Gretter, R., Pellin, M., Sosi, A., and Omologo, M. (2015). The DIRHA-ENGLISH corpus and related tasks for distant-speech recognition in domestic environments. In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, pages 547–554. IEEE.

[Rix et al., 2001] Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (2001). Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, pages 749–752 vol.2.

[Ryant et al., 2019] Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., and Liberman, M. (2019). The second DIHARD diarization challenge: Dataset, task, and baselines. In *Proc. Interspeech*.

[Sainath and Parada, 2015] Sainath, T. N. and Parada, C. (2015). Convolutional neural networks for small-footprint keyword spotting. In *Proc. Interspeech*.

[Sainath et al., 2017] Sainath, T. N., Weiss, R. J., Wilson, K. W., Li, B., Narayanan, A., Variani, E., Bacchiani, M., Shafran, I., Senior, A., Chin, K., Misra, A., and Kim, C. (2017). Multichannel signal processing with deep neural networks for automatic speech recognition. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 25(5):965–979.

[Sainath et al., 2016] Sainath, T. N., Weiss, R. J., Wilson, K. W., Narayanan, A., and Bacchiani, M. (2016). Factored spatial and spectral multichannel raw waveform CLDNNs. In *Proc. ICASSP*, pages 5075–5079.

[Sawada et al., 2011] Sawada, H., Araki, S., and Makino, S. (2011). Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 19(3):516–527.

[Schmid et al., 2012] Schmid, D., Malik, S., and Enzner, G. (2012). An expectation-maximization algorithm for multichannel adaptive speech dereverberation in the frequency-domain. In *Proc. ICASSP*.

[Schwartz et al., 2015] Schwartz, B., Gannot, S., and Habets, E. A. P. (2015). Online speech dereverberation using Kalman filter and EM algorithm. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 23(2):394–406.

[Schwartz et al., 2016] Schwartz, O., Gannot, S., and Habets, E. A. P. (2016). Joint maximum likelihood estimation of late reverberant and speech power spectral density in noisy environments. In *Proc. ICASSP*. IEEE.

[Seetharaman et al., 2019] Seetharaman, P., Wichern, G., Roux, J. L., and Pardo, B. (2019). Bootstrapping single-channel source separation via unsupervised spatial clustering on stereo mixtures.

[Sell et al., 2018] Sell, G., Snyder, D., McCree, A., Garcia-Romero, D., Villalba, J., Maciejewski, M., Manohar, V., Dehak, N., Povey, D., Watanabe, S., and Khudanpur, S. (2018). Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge. In *Proc. Interspeech*.

[Seltzer et al., 2004] Seltzer, M. L., Raj, B., Stern, R. M., et al. (2004). Likelihood-maximizing beamforming for robust hands-free speech recognition. *IEEE Trans. Speech Audio Process.*, 12(5):489–498.

[Shannon et al., 2017] Shannon, M., Simko, G., Chang, S.-y., and Parada, C. (2017). Improved end-of-query detection for streaming speech recognition. In *Proc. Interspeech*.

[Silovsky et al., 2011] Silovsky, J., Prazak, J., Cerva, P., Zdansky, J., and Nouza, J. (2011). Plda-based clustering for speaker diarization of broadcast streams. In *Proc. Interspeech*.

[Snyder et al., 2018] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *Proc. ICASSP*.

[Souden et al., 2010] Souden, M., Benesty, J., and Affes, S. (2010). On optimal frequency-domain multichannel linear filtering for noise reduction. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 18(2):260–276.

[Subramanian et al., 2019] Subramanian, A. S., Wang, X., Watanabe, S., Taniguchi, T., Tran, D., and Fujita, Y. (2019). An investigation of end-to-end multichannel speech recognition for reverberant and mismatch conditions. *arXiv:1904.09049*.

[Taal et al., 2010] Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217.

[Takahashi et al., 2019] Takahashi, N., Parthasaarathy, S., Goswami, N., and Mitsufuji, Y. (2019). Recursive speech separation for unknown number of speakers. *CoRR*, abs/1904.03065.

[Togami and Kawaguchi, 2013] Togami, M. and Kawaguchi, Y. (2013). Noise robust speech dereverberation with kalman smoother. In *Proc. ICASSP*, page 74477451.

[Tran Vu and Haeb-Umbach, 2010] Tran Vu, D. H. and Haeb-Umbach, R. (2010). Blind speech separation employing directional statistics in an expectation maximization framework. In *Proc. ICASSP*, pages 241–244.

[Tzinis et al., 2019] Tzinis, E., Venkataramani, S., and Smaragdis, P. (2019). Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information.

[Variani et al., 2014] Variani, E., Lei, X., McDermott, E., Lopez-Moreno, I., and Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *Proc. ICASSP*.

[Variani et al., 2016] Variani, E., Sainath, T. N., Shafran, I., and Bacchiani, M. (2016). Complex linear projection (clp): A discriminative approach to joint feature extraction and acoustic modeling. In *Proc. Interspeech*.

[Vincent et al., 2013] Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., and Matassoni, M. (2013). The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines. In *Proc. ICASSP*.

[Vincent et al., 2006] Vincent, E., Gribonval, R., and Fvotte, C. (2006). Performance measurement in blind audio source separation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 14(4):1462–1469.

[Vincent et al., 2017] Vincent, E., Watanabe, S., Nugraha, A. A., Barker, J., and Marxer, R. (2017). An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, 46:535–557.

[von Neumann et al., 2019] von Neumann, T., Kinoshita, K., Delcroix, M., Araki, S., Nakatani, T., and Haeb-Umbach, R. (2019). All-neural online source separation, counting, and diarization for meeting analysis. *Proc. ICASSP*.

[Wang and Brown, 2006] Wang, D. and Brown, G. J. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press.

[Wang and Chen, 2017] Wang, D. and Chen, J. (2017). Supervised speech separation based on deep learning: An overview. *CoRR*, abs/1708.07524.

[Wang and Chen, 2018] Wang, D. and Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726.

[Wang et al., 2018a] Wang, J., Chen, J., Su, D., Chen, L., Yu, M., Qian, Y., and Yu, D. (2018a). Deep extractor network for target speaker recovery from single channel speech mixtures. In *Proc. Interspeech*.

[Wang et al., 2017] Wang, Y., Du, J., Dai, L.-R., and Lee, C.-H. (2017). A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25:1535–1546.

[Wang et al., 2014] Wang, Y., Narayanan, A., and Wang, D. (2014). On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1849–1858.

[Wang et al., 2018b] Wang, Z., Roux, J. L., Wang, D., and Hershey, J. R. (2018b). End-to-end speech separation with unfolded iterative phase reconstruction. *CoRR*, abs/1804.10204.

[Wang et al., 2018c] Wang, Z., Vincent, E., Serizel, R., and Yan, Y. (2018c). Rank-1 constrained multichannel wiener filter for speech recognition in noisy environments. *Computer Speech & Language*, 49:37 – 51.

[Wang and Wang, 2019] Wang, Z. and Wang, D. (2019). Combining spectral and spatial features for deep learning based blind speaker separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(2):457–468.

[Wang et al., 2018d] Wang, Z.-Q., Le Roux, J., and Hershey, J. R. (2018d). Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation. In *Proc. ICASSP*. IEEE.

[Wang and Wang, 2018] Wang, Z.-Q. and Wang, D. (2018). All neural multi-channel speech enhancement. In *Proc. Interspeech*, pages 1561–1565.

[Warsitz and Haeb-Umbach, 2007] Warsitz, E. and Haeb-Umbach, R. (2007). Blind acoustic beamforming based on generalized eigenvalue decomposition. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 15(5):1529–1539.

[Warsitz et al., 2008] Warsitz, E., Krueger, A., and Haeb-Umbach, R. (2008). Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller. In *Proc. ICASSP*, pages 73–76.

[Weninger et al., 2014] Weninger, F., Watanabe, S., Le Roux, J., Hershey, J., Tachioka, Y., Geiger, J.T. andSchuller, B., and Rigoll, G. (2014). The MERL/MELCO/TUM system for the REVERB challenge using deep recurrent neural network feature enhancement. In *REVERB challenge workshop*.

[Wichern et al., 2019] Wichern, G., McQuinn, E., Antognini, J., Flynn, M., Zhu, R., Crow, D., Manilow, E., and Roux, J. L. (2019). Wham!: Extending speech separation to noisy environments. In *Proc. Interspeech*.

[Williamson and Wang, 2017] Williamson, D. S. and Wang, D. (2017). Time-frequency masking in the complex domain for speech dereverberation and denoising. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 25(7):1492–1501.

[Woelfel and McDonough, 2009] Woelfel, M. and McDonough, J. (2009). *Distant Speech Recognition*. John Wiley.

[Wu et al., 2017] Wu, B., Li, K., Yang, M., and Lee, C. H. (2017). A reverberation-time-aware approach to speech dereverberation based on deep neural networks. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 25(1):102–111.

[Wu et al., 2018] Wu, M., Panchapagesan, S., Sun, M., Gu, J., Thomas, R., Vitaladevuni, S. N. P., Hoffmeister, B., and Mandal, A. (2018). Monophone-based background modeling for two-stage on-device wake word detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5494–5498.

[Xiao et al., 2016] Xiao, X., Watanabe, S., Erdogan, H., Lu, L., Hershey, J., Seltzer, M. L., Chen, G., Zhang, Y., Mandel, M., and Yu, D. (2016). Deep beamforming networks for multi-channel speech recognition. In *Proc. ICASSP*, pages 5745–5749.

[Xu et al., 2014] Xu, Y., Du, J., Dai, L. R., and Lee, C. H. (2014). An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Process. Lett.*, 21(1):65–68.

[Yilmaz and Rickard, 2004] Yilmaz, O. and Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. Signal Process.*, 52(7):1830–1847.

[Yoshioka et al., 2018] Yoshioka, T., Erdogan, H., Chen, Z., Xiao, X., and Alleva, F. (2018). Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks. *Proc. Interspeech*.

[Yoshioka et al., 2015] Yoshioka, T., Ito, N., Delcroix, M., Ogawa, A., Kinoshita, K., Fujimoto, M., Yu, C., Fabian, W. J., Espi, M., Higuchi, T., Araki, S., and Nakatani, T. (2015). The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices. In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, pages 436–443.

[Yoshioka and Nakatani, 2012] Yoshioka, T. and Nakatani, T. (2012). Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*

[Yoshioka et al., 2012] Yoshioka, T., Sehr, A., Delcroix, M., Kinoshita, K., Maas, R., Nakatani, T., and Kellermann, W. (2012). Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition. *IEEE Signal Process. Mag.*, 29(6):114–126.

[Yoshioka et al., 2009] Yoshioka, T., Tachibana, H., Nakatani, T., and Miyoshi, M. (2009). Adaptive dereverberation of speech signals with speaker-position change detection. In *Proc. ICASSP*, pages 3733–3736. IEEE.

[Yu et al., 2017] Yu, D., Kolbæk, M., Tan, Z.-H., and Jensen, J. (2017). Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *Proc. ICASSP*, pages 241–245. IEEE.

[Zhang and Koishida, 2017] Zhang, C. and Koishida, K. (2017). End-to-end text-independent speaker verification with triplet loss on short utterances. In *Proc. Interspeech*.

[Zhang and Wang, 2018] Zhang, H. and Wang, D. (2018). Deep learning for acoustic echo cancellation in noisy and double-talk scenarios. In *Proc. Interspeech*.

[Zhang et al., 2016] Zhang, S.-X., Chen, Z., Zhao, Y., Li, J., and Gong, Y. (2016). End-to-end attention based text-dependent speaker verification. In *Proc. of IEEE Spoken Language Technology Workshop*.

[Zmolíková et al., 2017] Zmolíková, K., Delcroix, M., Kinoshita, K., Higuchi, T., Ogawa, A., and Nakatani, T. (2017). Speaker-aware neural network based beamformer for speaker extraction in speech mixtures. In *Proc. Interspeech*.

[Zmolikova et al., 2019] Zmolikova, K., Delcroix, M., Kinoshita, K., Ochiai, T., Nakatani, T., Burget, L., and ernock, J. (2019). Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures. *IEEE Journal of Selected Topics in Signal Processing*, pages 1–1.