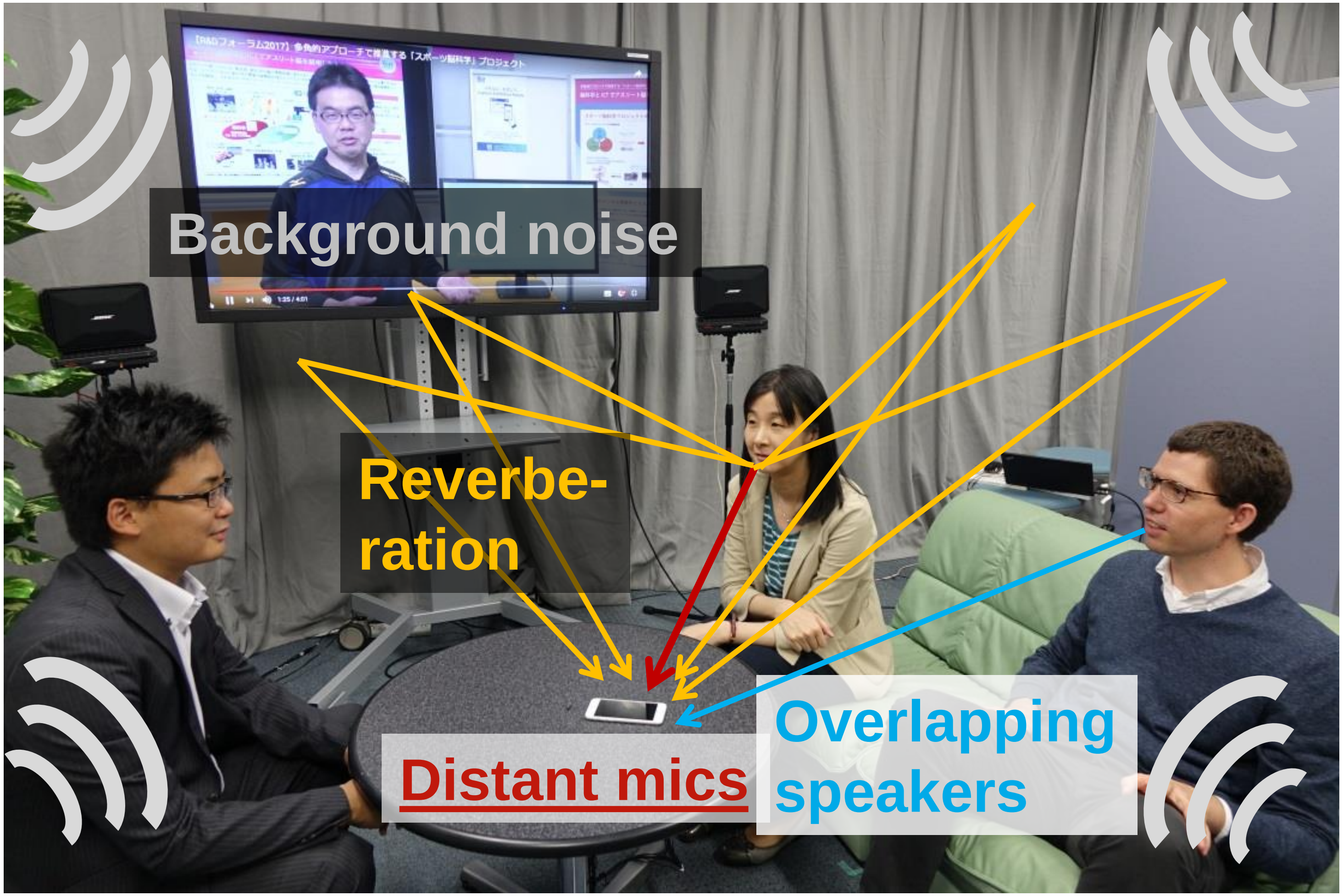
Part I.
Introduction

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!

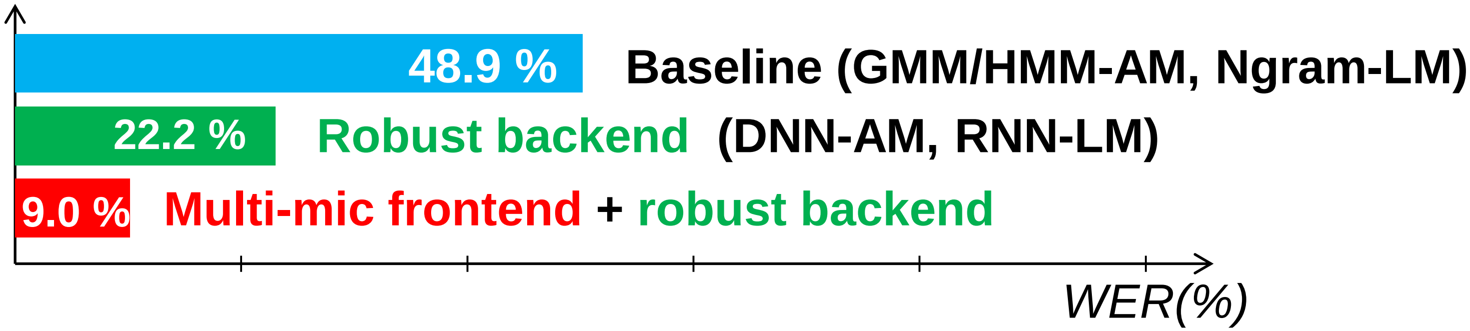
Goal of this talk

- 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)

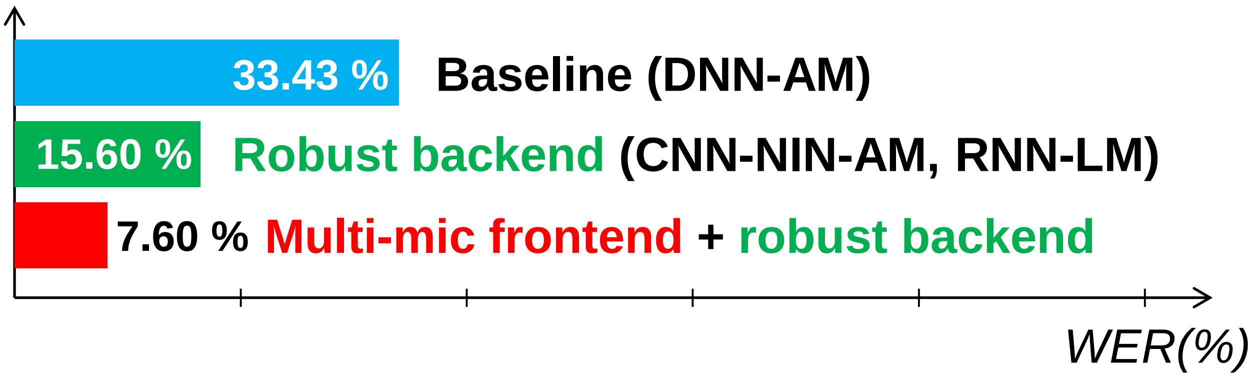
REVERB 2014

[Delcroix et al., 2015]



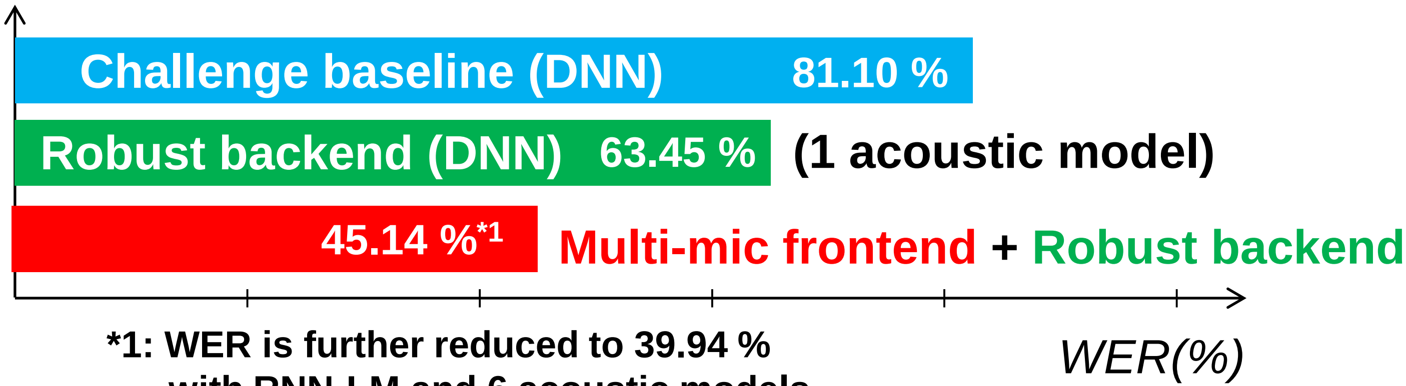
CHiME-3 2015

[Yoshioka et al., 2015]



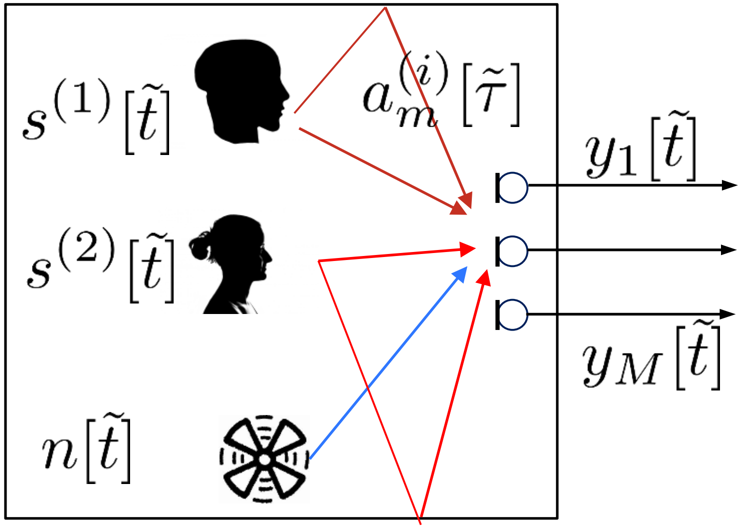
CHiME-5 2018

[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 $1 \leq i \leq I$

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

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

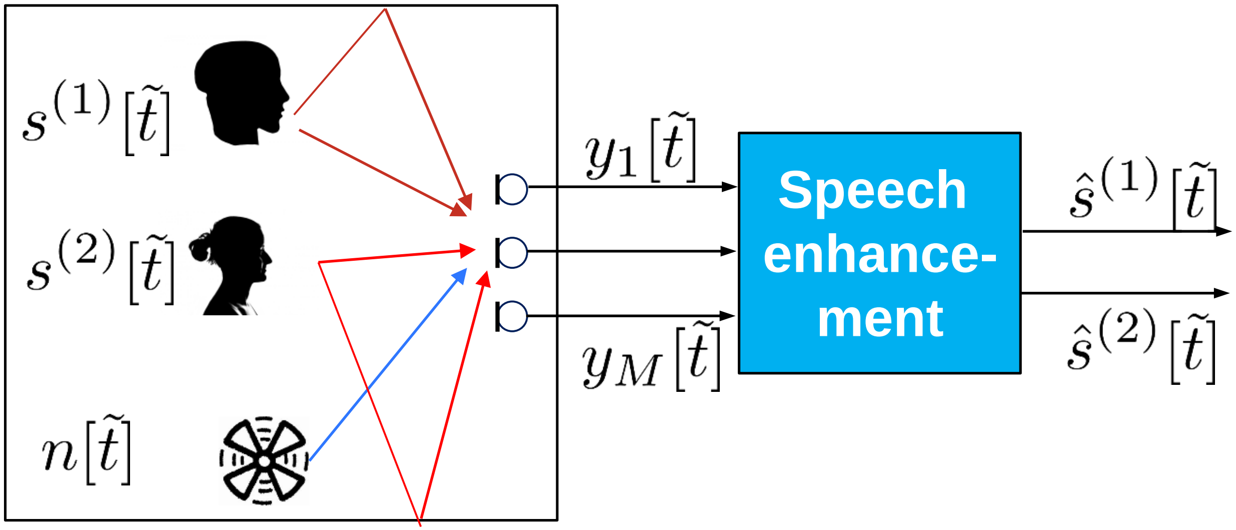
- Observed:

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

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

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

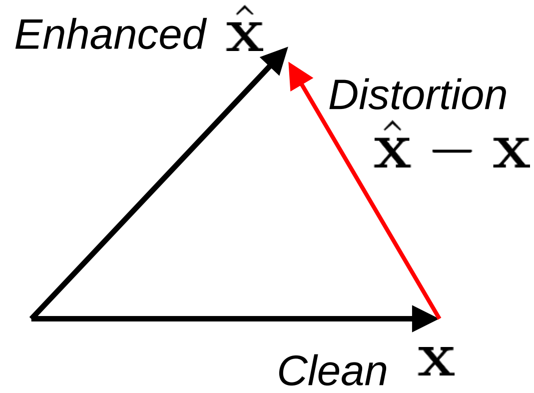
Type	Examples of measures	Pros and cons
Signal level distortion metric	• **Signal to distortion Ratio (SDR)** - Many variations • Frequency-weighted segmental SNR (FWSSNR), cepstral distortion (CD), signal-to-interference ratio (SIR), etc.	• **Most frequently used** • **Not directly reflect perceptual 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]

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



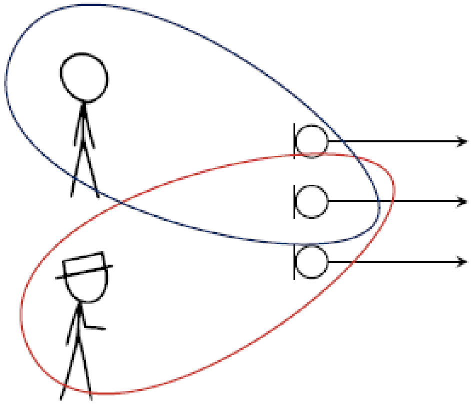
  - Sensitive to scale and phase estimation errors
- Variations
  - Scale-invariant SDR [Le Roux et al., 2019]
    - Invariant to scaling errors
  - Time-invariant filter allowed distortion [Vincent et al., 2006]
    - Invariant to scale and phase estimation errors
- Issues:
  - Smaller but important energy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

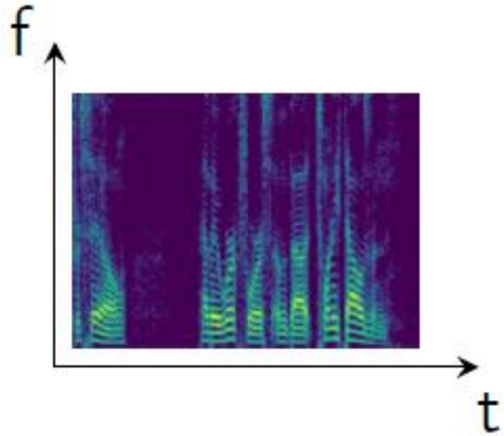
Evaluation metrics

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

$\hat{s}$

Beamforming

$$\hat{s} = \mathbf{w}^{\mathsf{H}}(\hat{\theta})\mathbf{y}$$

$\hat{\theta}$

Parameter estimation

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(\mathbf{y};\theta)$$

θ : Directions-of-arrival (DOA), RIRs, statistics of sources, etc.

Multi-ch observation $\mathbf{y}$

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:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(\mathbf{y};\theta)$$

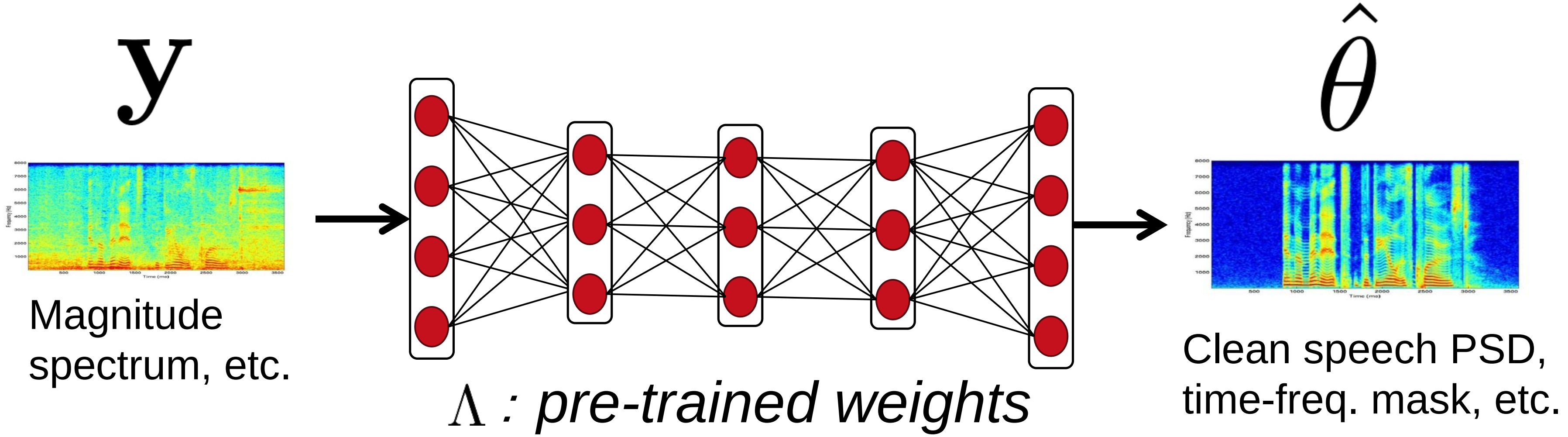
- Beamforming: e.g., by MMSE estimation

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

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.

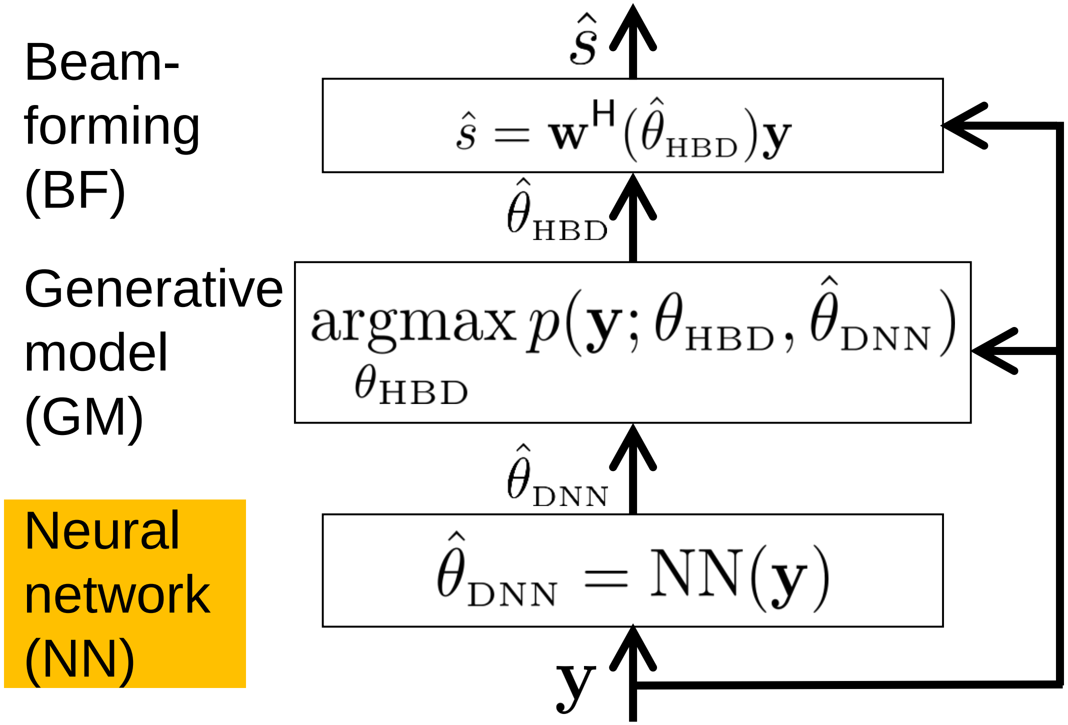
Pros and cons of two approaches

	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

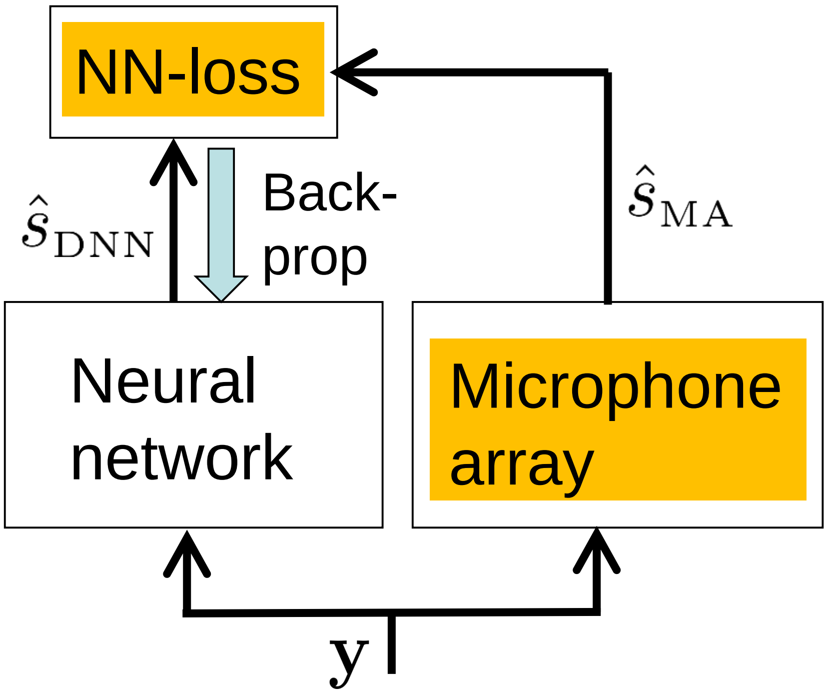
Examples:

- **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

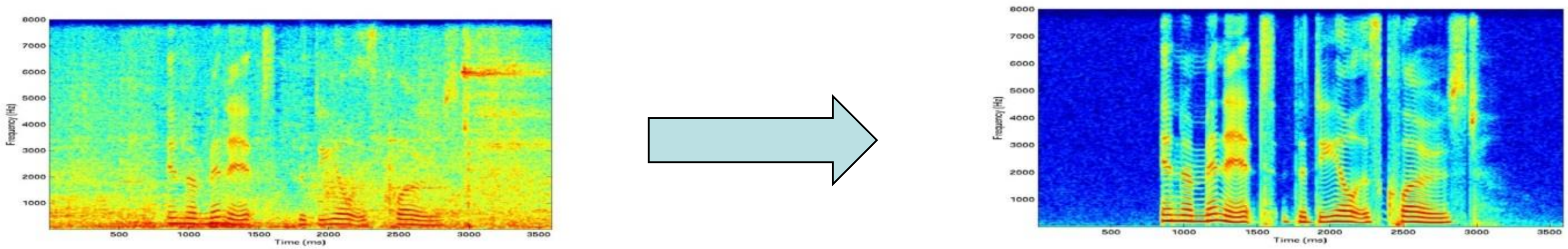
Examples:

- **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	DIRHA REVERB CHALLENGE CHiME CHALLENGE CHiME-3/4	CHiME CHALLENGE CHiME-5 AMI CONSORTIUM MC-WSJ CSTR Centre for Speech Technology Research
Simulation (Benchmark)	CHiME CHALLENGE CHiME-1,2 AURORA	wsj0-2mix, WHAM! MITSUBISHI ELECTRIC Changes for the Better

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