Part I.
Introduction

Tomohiro Nakatani
• 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]
- **Baseline (GMM/HMM-AM, Ngram-LM)**: 48.9%
- **Robust backend (DNN-AM, RNN-LM)**: 22.2%
- **Multi-mic frontend + robust backend**: 9.0%

**CHiME-3 2015**
[Yoshioka et al., 2015]
- **Baseline (DNN-AM)**: 33.43%
- **Robust backend (CNN-NIN-AM, RNN-LM)**: 15.60%
- **Multi-mic frontend + robust backend**: 7.60%

**CHiME-5 2018**
[Kanda et al., 2019]
- **Challenge baseline (DNN)**: 81.10%
- **Robust backend (DNN)**: 63.45% (1 acoustic model)
- **Multi-mic frontend + Robust backend**: 45.14%*1

*1: WER is further reduced to 39.94% with RNN-LM and 6 acoustic models.
Model of recorded speech: time domain

- **Observed:**

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

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

\( \tilde{t} \): 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
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

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

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

<table>
<thead>
<tr>
<th>Type</th>
<th>Examples of measures</th>
<th>Pros and cons</th>
</tr>
</thead>
</table>
| **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

Beamforming

\[ \hat{s} = w^H(\hat{\theta})y \]

Parameter estimation

\[ \hat{\theta} = \arg\max_{\theta} p(y; \theta) \]

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

Multi-ch observation

\( y \)
Microphone array signal processing (2/2)

• Use generative model to estimate unknown observation system

A generative model: 

\[ p(y; \theta) = \int p(y|s, n; \theta_r)p(s; \theta_s)p(n; \theta_n) \, ds \, dn \]

\( \theta_s \) : Speech power spectral density, voice activity, etc.
\( \theta_n \) : Noise power spectral density, etc.
\( \theta_r \) : Directions-of-arrival (DOAs), room impulse responses (RIRs), etc.

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

\[ \hat{\theta} = \arg \max_{\theta} p(y; \theta) \]

• Beamforming: e.g., by MMSE estimation

\[ \hat{s} = \arg \min_{\hat{s}} \int |s - \hat{s}|^2 p(s|y; \hat{\theta}) \, ds = w^H(\hat{\theta})y \]

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

- Train neural networks using huge amount of training data

$\mathbf{y}$

\[ \Lambda : \text{pre-trained weights} \]

$\hat{\theta}$

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

Magnitude spectrum, etc.

Robust and accurate spectral estimation is possible

Interpret this as the inverse system of the generative model, that estimates the model parameters from observation.
<table>
<thead>
<tr>
<th></th>
<th>Microphone array signal processing</th>
<th>Neural networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial characteristics modeling</td>
<td>• Strong</td>
<td>• Moderate (use spatial features as auxiliary input)</td>
</tr>
</tbody>
</table>
| 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
  - 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

Beamforming (BF)
Generative model (GM)
Neural network (NN)

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**

\[
\hat{s} = w^H(\theta_{HBD})y
\]

\[
\arg\max_{\theta_{HBD}, \theta_{DNN}} p(y; \theta_{HBD}, \theta_{DNN})
\]

\[
\hat{\theta}_{DNN} = \text{NN}(y)
\]

- Component-wise optimization
- Joint optimization
Hybrid approaches (2/2)

2) Unsupervised learning of neural networks enabled by microphone array

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

• Approach-1) can be combined after training

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
<table>
<thead>
<tr>
<th>#targets=1</th>
<th>#targets&gt;1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Real</strong></td>
<td><strong>MC-WSJ</strong></td>
</tr>
<tr>
<td><strong>Simulation (Benchmark)</strong></td>
<td><strong>wsj0-2mix, WHAM!</strong></td>
</tr>
</tbody>
</table>

**Benchmarks and Challenges**

- DIRHA
- REVERB CHALLENGE
- CHiME CHALLENGE
- CHiME-3/4
- CHiME-1,2

**CHiME Consortium**

- CHiME-5

**Haeb-Umbach and Nakatani, Speech Enhancement – Introduction**
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

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

• **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

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

Break (30 min)

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

QA