Generative Adversarial Network
and its Applications to Signal Processing
and Natural Language Processing

Part II: Speech Signal Processing
Outline of Part II

Speech Signal Generation
- Speech enhancement
- Postfilter, speech synthesis, voice conversion

Speech Signal Recognition
- Speech recognition
- Speaker recognition
- Speech emotion recognition
- Lip reading

Conclusion
Speech Signal Generation (Regression Task)

Paired

Objective function

Conditional GAN

Prior distribution $z$ $\rightarrow$ $G$ $\rightarrow$ Image $x = G(c, z)$

$D$ (better) $\rightarrow$ scalar $x$ is realistic or not + $c$ and $x$ are matched or not

True text-image pairs: (train, Image) 1
     (cat, Image) 0 (train, Image) 0

Cycle-GAN

$G_{X \rightarrow Y}$ $\rightarrow$ $D_X$ $\rightarrow$ $G_Y$ $\rightarrow$ $D_Y$ scalar: belongs to domain $X$ or not

$G_Y$ $\rightarrow$ $D_Y$ $\rightarrow$ $G_X$ scalar: belongs to domain $Y$ or not

as close as possible

as close as possible

[Scott Reed, et al, ICML, 2016]
Speech, Speaker, Emotion Recognition and Lip-reading (Classification Task)

**Domain Adversarial Training**

- Feature extractor (Generator)
- Label predictor
- Discriminator (Domain classifier)

- Not only cheat the domain classifier, but satisfying label predictor at the same time

Successfully applied on image classification

- \( E \) : Clean data
- \( G \) : Accented speech
- \( \tilde{x} \) : Noisy data
- \( y \) : Output label
- \( h(\cdot) \)
- \( \tilde{z} = g(\tilde{x}) \)
- \( g(\cdot) \)

**Acoustic Mismatch**

- Channel distortion
- Accented speech
- Noisy data

Clean data
Outline of Part II

Speech Signal Generation
  • Speech enhancement
  • Postfilter, speech synthesis, voice conversion

Speech Signal Recognition
  • Speech recognition
  • Speaker recognition
  • Speech emotion recognition
  • Lip reading

Conclusion
Speech Enhancement

- Typical objective function
  - Mean square error (MSE) \([\text{Xu et al., TASLP 2015}]\), L1 \([\text{Pascual et al., Interspeech 2017}]\), likelihood \([\text{Chai et al., MLSP 2017}]\), STOI \([\text{Fu et al., TASLP 2018}]\).

- GAN is used as a new objective function to estimate the parameters in \(G\).

- Model structures of \(G\): DNN \([\text{Wang et al. NIPS 2012; Xu et al., SPL 2014}]\), DDAE \([\text{Lu et al., Interspeech 2013}]\), RNN (LSTM) \([\text{Chen et al., Interspeech 2015; Weninger et al., LVA/ICA 2015}]\), CNN \([\text{Fu et al., Interspeech 2016}]\).
Speech Enhancement

- Speech enhancement GAN (SEGAN) [Pascual et al., Interspeech 2017]
Speech Enhancement (SEGAN)

• Experimental results

Table 1: Objective evaluation results.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Noisy</th>
<th>Wiener</th>
<th>SEGAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>PESQ</td>
<td>1.97</td>
<td>2.22</td>
<td>2.16</td>
</tr>
<tr>
<td>CSIG</td>
<td>3.35</td>
<td>3.23</td>
<td>3.48</td>
</tr>
<tr>
<td>CBAK</td>
<td>2.44</td>
<td>2.68</td>
<td>2.94</td>
</tr>
<tr>
<td>COVL</td>
<td>2.63</td>
<td>2.67</td>
<td>2.80</td>
</tr>
<tr>
<td>SSNR</td>
<td>1.68</td>
<td>5.07</td>
<td>7.73</td>
</tr>
</tbody>
</table>

Table 2: Subjective evaluation results.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Noisy</th>
<th>Wiener</th>
<th>SEGAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOS</td>
<td>2.09</td>
<td>2.70</td>
<td>3.18</td>
</tr>
</tbody>
</table>

Fig. 1: Preference test results.

SEGAN yields better speech enhancement results than Noisy and Wiener.
Speech Enhancement

- **Pix2Pix** [Michelsanti et al., Interpsech 2017]
Speech Enhancement (Pix2Pix)

- Spectrogram analysis

Fig. 2: Spectrogram comparison of Pix2Pix with baseline methods.

Pix2Pix outperforms STAT-MMSE and is competitive to DNN SE.
Speech Enhancement (Pix2Pix)

- Objective evaluation and speaker verification test

Table 3: Objective evaluation results.

<table>
<thead>
<tr>
<th>SNR</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>mean</th>
</tr>
</thead>
</table>
| Babble
| (a)  | 1.20| 1.42| 1.79| 2.40| 3.13| 1.99 |
| (b)  | 1.14| 1.31| 1.61| 2.07| 2.65| 1.76 |
| (c)  | 1.25| 1.51| 1.87| 2.31| 2.78| 1.95 |
| (d)  | 1.20| 1.48| 1.98| 2.52| 2.93| 2.02 |
| (e)  | 1.24| 1.52| 1.88| 2.31| 2.78| 1.95 |
| (f)  | 1.20| 1.49| 2.00| 2.53| 2.93| 2.03 |

Table 4: Speaker verification results.

<table>
<thead>
<tr>
<th>SNR</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>clean</th>
<th>mean</th>
</tr>
</thead>
</table>
| Babble
| (a)  | 21.09| 15.99| 13.61| 11.66| 9.18| 6.99 | 13.08|
| (b)  | 17.69| 12.58| 8.17 | 6.53 | 6.27| 5.80 | 9.51 |
| (c)  | 16.99| 10.55| 7.48 | 6.99 | 6.15| 6.12 | 9.05 |
| (d)  | 17.19| 8.84 | 5.44 | 5.05 | 4.63| 3.74 | 7.48 |
| (e)  | 15.99| 8.99 | 6.12 | 6.12 | 5.58| 5.67 | 8.08 |
| (f)  | 15.31| 7.89 | 5.58 | 4.77 | 4.76| 5.44 | 7.29 |
| Airplane
| (a)  | No enhancement | STSA-MMSE | NS-DNN | NS-Pix2Pix | NG-DNN | NG-Pix2Pix |

1. From the objective evaluations, Pix2Pix outperforms Noisy and MMSE and is competitive to DNN SE.
2. From the speaker verification results, Pix2Pix outperforms the baseline models when the clean training data is used.
Speech Enhancement

- Frequency-domain SEGAN (FSEGAN) [Donahue et al., ICASSP 2018]
Speech Enhancement (FSEGAN)

• Spectrogram analysis

Fig. 2: Spectrogram comparison of FSEGAN with L1-trained method.

(a) Noisy speech input $\alpha$
(b) L1-trained output $G'(\alpha)$
(c) Clean speech target $y$
(d) FSEGAN output $G(\alpha)$

FSEGAN reduces both additive noise and reverberant smearing.
Speech Enhancement (FSEGAN)

• ASR results

Table 5: WER (%) of SEGAN and FSEGAN.

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Enhancer</th>
<th>ASR-Clean WER</th>
<th>ASR-MTR WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>None</td>
<td>11.9</td>
<td>14.3</td>
</tr>
<tr>
<td>MTR</td>
<td>None</td>
<td>72.2</td>
<td>20.3</td>
</tr>
<tr>
<td></td>
<td>SEGAN</td>
<td>80.7</td>
<td>52.8</td>
</tr>
<tr>
<td></td>
<td>FSEGAN</td>
<td>33.3</td>
<td>25.4</td>
</tr>
</tbody>
</table>

Table 6: WER (%) of FSEGAN with retrain.

<table>
<thead>
<tr>
<th>Model</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTR Baseline *</td>
<td>20.3</td>
</tr>
<tr>
<td>+ Stereo</td>
<td>19.0</td>
</tr>
<tr>
<td>MTR + FSEGAN Enhancer *</td>
<td>25.4</td>
</tr>
<tr>
<td>+ Retraining</td>
<td>21.0</td>
</tr>
<tr>
<td>+ Hybrid Retraining</td>
<td>17.6</td>
</tr>
<tr>
<td>MTR + L1-trained Enhancer *</td>
<td>21.4</td>
</tr>
<tr>
<td>+ Retraining</td>
<td>18.0</td>
</tr>
<tr>
<td>+ Hybrid Retraining</td>
<td>17.1</td>
</tr>
</tbody>
</table>

1. From Table 5, (1) FSEGAN improves recognition results for ASR-Clean. (2) FSEGAN outperforms SEGAN as front-ends.
2. From Table 6, (1) Hybrid Retraining with FSEGAN outperforms Baseline; (2) FSEGAN retraining slightly underperforms L1–based retraining.
Speech Enhancement

- Adversarial training based mask estimation (ATME)
  [Higuchi et al., ASRU 2017]

\[
V_{\text{Mask}} = E_{s_{\text{fake}}} \left[ \log(1 - D_S(s_{\text{fake}}, \theta)) \right] \\
+ E_{n_{\text{fake}}} \left[ \log(1 - D_N(n_{\text{fake}}, \theta)) \right]
\]
Speech Enhancement (ATME)

• Spectrogram analysis

Fig. 3: Spectrogram comparison of (a) noisy; (b) MMSE with supervision; (c) ATMB without supervision.

The proposed adversarial training mask estimation can capture speech/noise signals without supervised data.
Speech Enhancement (ATME)

• Mask-based beamformer for robust ASR

- The estimated mask parameters are used to compute spatial covariance matrix for MVDR beamformer.

\[ \hat{s}_{f,t} = \mathbf{w}_f^H \mathbf{y}_{f,t} \]  
where \( \hat{s}_{f,t} \) is the enhanced signal, and \( \mathbf{y}_{f,t} \) denotes the observation of \( M \) microphones, \( f \) and \( t \) are frequency and time indices; \( \mathbf{w}_f \) denotes the beamformer coefficient.

- The MVDR solves \( \mathbf{w}_f \) by:

\[ \mathbf{w}_f = \frac{(R_f^{(s+n)})^{-1} \mathbf{h}_f}{\mathbf{h}_f^H (R_f^{(s+n)})^{-1} \mathbf{h}_f} \]

- To estimate \( \mathbf{h}_f \), the spatial covariance matrix of the target signal, \( R_f^{(s)} \), is computed by:

\[ R_f^{(s)} = R_f^{(s+n)} - R_f^{(n)} \]

where \( R_f^{(n)} = \frac{M_f^{(n)} \mathbf{y}_{f,t} \mathbf{y}_{f,t}^H}{\Sigma_{f,t} M_f^{(n)}} \), \( M_f^{(n)} \) was computed by AT.
Speech Enhancement (ATME)

- ASR results

Table 7: WERs (%) for the development and evaluation sets.

<table>
<thead>
<tr>
<th>systems</th>
<th></th>
<th>avg</th>
<th>bus</th>
<th>caf</th>
<th>ped</th>
<th>str</th>
<th></th>
<th>avg</th>
<th>bus</th>
<th>caf</th>
<th>ped</th>
<th>str</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unprocessed</td>
<td></td>
<td>9.01</td>
<td>14.00</td>
<td>7.94</td>
<td>6.03</td>
<td>8.05</td>
<td></td>
<td>15.60</td>
<td>22.55</td>
<td>16.21</td>
<td>12.89</td>
<td>10.74</td>
</tr>
<tr>
<td>Adversarial Training</td>
<td></td>
<td>5.00</td>
<td>7.60</td>
<td>4.09</td>
<td>4.03</td>
<td>4.29</td>
<td></td>
<td>7.58</td>
<td>10.24</td>
<td>7.51</td>
<td>6.20</td>
<td>6.39</td>
</tr>
</tbody>
</table>

1. ATME provides significant improvements over Unprocessed.
2. Unsupervised ATME slightly underperforms supervised MMSE.
Speech Enhancement (AFT)

- Cycle-GAN-based acoustic feature transformation (AFT)
  [Mimura et al., ASRU 2017]

![Diagram of Speech Enhancement (AFT)]

\[ V_{Full} = V_{GAN}(G_{X \rightarrow Y}, D_Y) + V_{GAN}(G_{X \rightarrow Y}, D_Y) + \lambda V_{CyC}(G_{X \rightarrow Y}, G_{Y \rightarrow X}) \]
Speech Enhancement (AFT)

- ASR results on noise robustness and style adaptation

Table 8: Noise robust ASR.

<table>
<thead>
<tr>
<th>acoustic model</th>
<th>feature</th>
<th>cycle loss</th>
<th>$\lambda$ and $\mu$</th>
<th>WER</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>no adapt.</td>
<td>no adapt.</td>
<td>-</td>
<td>-</td>
<td>41.08</td>
<td>(1)</td>
</tr>
<tr>
<td>no adapt.</td>
<td>adapt. with $G_{T \rightarrow S}$</td>
<td>no</td>
<td>1, 1</td>
<td>55.45</td>
<td>(2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>yes</td>
<td>1, 1</td>
<td>37.34</td>
<td>(3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>yes</td>
<td>trained</td>
<td>36.56</td>
<td>(4)</td>
</tr>
<tr>
<td>adapt. with $G_{S \rightarrow T}$</td>
<td>no adapt.</td>
<td>yes</td>
<td>1, 1</td>
<td>35.98</td>
<td>(5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>yes</td>
<td>trained</td>
<td>34.31</td>
<td>(6)</td>
</tr>
</tbody>
</table>

$S$: Clean; $T$: Noisy

Table 9: Speaker style adaptation.

<table>
<thead>
<tr>
<th>source</th>
<th>target</th>
<th>feature</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>JNAS</td>
<td>CSJ-SPS</td>
<td>no adapt.</td>
<td>26.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>adapt. with $G_{T \rightarrow S}$</td>
<td>25.93</td>
</tr>
<tr>
<td>CSJ-APS</td>
<td>CSJ-SPS</td>
<td>no adapt.</td>
<td>17.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>adapt. with $G_{T \rightarrow S}$</td>
<td>16.60</td>
</tr>
</tbody>
</table>

JNAS: Read; CSJ-SPS: Spontaneous (relax); CSJ-APS: Spontaneous (formal);

1. $G_{T \rightarrow S}$ can transform acoustic features and effectively improve ASR results for both noisy and accented speech.
2. $G_{S \rightarrow T}$ can be used for model adaptation and effectively improve ASR results for noisy speech.
Outline of Part II

Speech Signal Generation

• Speech enhancement
• Postfilter, speech synthesis, voice conversion

Speech Signal Recognition

• Speech recognition
• Speaker recognition
• Speech emotion recognition
• Lip reading

Conclusion
Conventional postfilter approaches for $G$ estimation include global variance (GV) [Toda et al., IEICE 2007], variance scaling (VS) [Sil’ en et al., Interpseech 2012], modulation spectrum (MS) [Takamichi et al., ICASSP 2014], DNN with MSE criterion [Chen et al., Interspeech 2014; Chen et al., TASLP 2015].

GAN is used a new objective function to estimate the parameters in $G$. 
Postfilter

• GAN postfilter [Kaneko et al., ICASSP 2017]

Traditional MMSE criterion results in statistical averaging.
GAN is used as a new objective function to estimate the parameters in $G$.
The proposed work intends to further improve the naturalness of synthesized speech or parameters from a synthesizer.
Postfilter (GAN-based Postfilter)

- Spectrogram analysis

Fig. 4: Spectrograms of: (a) NAT (nature); (b) SYN (synthesized); (c) VS (variance scaling); (d) MS (modulation spectrum); (e) MSE; (f) GAN postfilters.

GAN postfilter reconstructs spectral texture similar to the natural one.
Postfilter (GAN-based Postfilter)

- Objective evaluations

Fig. 5: Mel-cepstral trajectories (GANv: GAN was applied in voiced part).

Fig. 6: Averaging difference in modulation spectrum per Mel-cepstral coefficient.

GAN postfilter reconstructs spectral texture similar to the natural one.
### Postfilter (GAN-based Postfilter)

- **Subjective evaluations**

<table>
<thead>
<tr>
<th></th>
<th>Former</th>
<th>Latter</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAN vs. SYN</td>
<td>56.5 ± 4.9</td>
<td>22.0 ± 4.1</td>
<td>21.5 ± 4.0</td>
</tr>
<tr>
<td>GAN vs. GANv</td>
<td>11.3 ± 3.1</td>
<td>37.3 ± 4.8</td>
<td>51.5 ± 4.9</td>
</tr>
<tr>
<td>GAN vs. NAT</td>
<td>16.8 ± 3.7</td>
<td>53.5 ± 4.9</td>
<td>29.8 ± 4.5</td>
</tr>
<tr>
<td>GANv vs. NAT</td>
<td>30.3 ± 4.5</td>
<td>34.5 ± 4.7</td>
<td>35.3 ± 4.7</td>
</tr>
</tbody>
</table>

1. GAN postfilter significantly improves the synthesized speech.
2. GAN postfilter is effective particularly in voiced segments.
3. GANv outperforms GAN and is comparable to NAT.
Postfilter (GAN-postfilter-SFTF)

- GAN post-filter for STFT spectrograms [Kaneko et al., Interspeech 2017]

- GAN postfilter was applied on high-dimensional STFT spectrograms.
- The spectrogram was partitioned into $N$ bands (each band overlaps its neighboring bands).
- The GAN-based postfilter was trained for each band.
- The reconstructed spectrogram from each band was smoothly connected.
Postfilter (GAN-postfilter-SFTF)

- Spectrogram analysis

Fig. 7: Spectrograms of: (1) SYN, (2) GAN, (3) Original (NAT)

GAN postfilter reconstructs spectral texture similar to the natural one.
Speech Synthesis

- Speech synthesis involves linguistic features and speech parameters.

\[ L(c, \hat{c}) = L_G(c, \hat{c}) + \omega_D \frac{E_{LG}}{E_{LD}} L_{D,1}(\hat{c}) \]

Minimum generation error (MGE) with adversarial loss.

\[ L_D(c, \hat{c}) = L_{D,1}(c) + L_{D,0}(\hat{c}) \]

\[ L_{D,1}(c) = -\frac{1}{T} \sum_{t=1}^{T} \log(D(c_t))...\text{NAT} \]

\[ L_{D,0}(\hat{c}) = -\frac{1}{T} \sum_{t=1}^{T} \log(1 - D(\hat{c}_t))...\text{SYN} \]

[Saito et al., ICASSP 2017]
Speech Synthesis (ASV)

- Objective and subjective evaluations

1. The proposed algorithm generates MCCs similar to the natural ones.
2. The proposed algorithm outperforms conventional MGE training.

Fig. 8: Averaged GVs of MCCs.

Fig. 9: Scores of speech quality.
Speech Synthesis

- Speech synthesis with GAN (SS-GAN) [Saito et al., TASLP 2018]

\[ L(c, \hat{c}) = L_G(c, \hat{c}) + \omega_D \frac{E_{LG}}{E_{LD}} L_{D,1}(\hat{c}) \]

Minimum generation error (MGE) with adversarial loss.

\[ L_D(c, \hat{c}) = L_{D,1}(c) + L_{D,0}(\hat{c}) \]
\[ L_{D,1}(c) = -\frac{1}{T} \sum_{t=1}^{T} \log(D(c_t)) \ldots \text{NAT} \]
\[ L_{D,0}(\hat{c}) = -\frac{1}{T} \sum_{t=1}^{T} \log(1 - D(\hat{c}_t)) \ldots \text{SYN} \]
Speech Synthesis (SS-GAN)

- Subjective evaluations

Fig. 10: Scores of speech quality (sp).

Fig. 11: Scores of speech quality (sp and F0).

The proposed algorithm works for both spectral parameters and F0.
Speech Synthesis

- Speech synthesis with GAN glottal waveform model (GlottGAN) [Bollepalli et al., Interspeech 2017]
Speech Synthesis (GlottGAN)

- Objective evaluations

Fig. 12: Glottal pulses generated by GANs.

The proposed GAN-based approach can generate glottal waveforms similar to the natural ones.
Speech Synthesis

- Speech synthesis with GAN & multi-task learning (SS-GAN-MTL) [Yang et al., ASRU 2017]

\[ \text{V}_{GAN}(G, D) = E_{x \sim p_{data}(x)}[\log D(x|y)] + E_{z \sim p_{z}}[\log(1 - D(G(z|y)))|y] \]
\[ \text{V}_{L2}(G) = E_{z \sim p_{z}}[G(z|y) - x]^2 \]
Speech Synthesis (SS-GAN-MTL)

- Speech synthesis with GAN & multi-task learning (SS-GAN-MTL) [Yang et al., ASRU 2017]

\[
V_{GAN}(G, D) = E_{x \sim p_{data}(x)}[\log D_{CE}(x|y, label)] + E_{z \sim p_z}[\log(1 - D_{CE}(G(z|y)))|y, label]
\]

\[
V_{L2}(G) = E_{z \sim p_z}[G(z|y) - x]^2
\]
Speech Synthesis (SS-GAN-MTL)

- Objective and subjective evaluations

Table 11: Objective evaluation results.

<table>
<thead>
<tr>
<th>Methods</th>
<th>MCD (dB)</th>
<th>$F_0$ RMSE (Hz)</th>
<th>V/UV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLSTM</td>
<td>4.624</td>
<td>18.544</td>
<td>6.447</td>
</tr>
<tr>
<td>GAN</td>
<td>4.633</td>
<td>18.678</td>
<td>6.492</td>
</tr>
<tr>
<td>GAN-PC</td>
<td>4.628</td>
<td>18.616</td>
<td>6.464</td>
</tr>
</tbody>
</table>

1. From objective evaluations, no remarkable difference is observed.
2. From subjective evaluations, GAN outperforms BLSTM and ASV, while GAN-PC underperforms GAN.

Fig. 13: The preference score (%).
Voice Conversion

• Convert (transform) speech from source to target

Conventional VC approaches include Gaussian mixture model (GMM) [Toda et al., TASLP 2007], non-negative matrix factorization (NMF) [Wu et al., TASLP 2014; Fu et al., TBME 2017], locally linear embedding (LLE) [Wu et al., Interspeech 2016], restricted Boltzmann machine (RBM) [Chen et al., TASLP 2014], feed forward NN [Desai et al., TASLP 2010], recurrent NN (RNN) [Nakashika et al., Interspeech 2014].
Voice Conversion

- VAW-GAN [Hsu et al., Interspeech 2017]

- Conventional MMSE approaches often encounter the “over-smoothing” issue.
- GAN is used a new objective function to estimate $G$.
- The goal is to increase the naturalness, clarity, similarity of converted speech.

$$V(G, D) = V_{GAN}(G, D) + \lambda V_{VAE}(x|y)$$
Voice Conversion (VAW-GAN)

- Objective and subjective evaluations

VAW-GAN outperforms VAE in terms of objective and subjective evaluations with generating more structured speech.
Voice Conversion

• Sequence-to-sequence VC with learned similarity metric (LSM) [Kaneko et al., Interspeech 2017]

\[
V(C, G, D) = V_{SVC}^{D_l}(C, D) + V_{GAN}(C, G, D)
\]

\[
V_{SVC}^{D_l}(C, D) = \frac{1}{M_l} \| D_l(y) - D_l(C(x)) \|^2
\]
Voice Conversion (LSM)

- Spectrogram analysis

Fig. 16: Comparison of MCCs (upper) and STFT spectrograms (lower).

The spectral textures of LSM are more similar to the target ones.
Voice Conversion (LSM)

- Subjective evaluations

Table 12: Preference scores for naturalness.

<table>
<thead>
<tr>
<th></th>
<th>Former</th>
<th>Latter</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>FVC vs. LSM</td>
<td>17.1 ± 6.3</td>
<td>72.9 ± 7.5</td>
<td>10.0 ± 5.0</td>
</tr>
<tr>
<td>MSE vs. LSM</td>
<td>10.0 ± 5.0</td>
<td>84.3 ± 6.1</td>
<td>5.7 ± 3.9</td>
</tr>
</tbody>
</table>

Table 12: Preference scores for clarity.

<table>
<thead>
<tr>
<th></th>
<th>Former</th>
<th>Latter</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>FVC vs. LSM</td>
<td>32.9 ± 7.9</td>
<td>54.3 ± 8.4</td>
<td>12.9 ± 5.6</td>
</tr>
<tr>
<td>MSE vs. LSM</td>
<td>27.1 ± 7.5</td>
<td>65.0 ± 8.0</td>
<td>7.9 ± 4.5</td>
</tr>
</tbody>
</table>

Fig. 17: Similarity of TGT and SRC with VCs.

LSM outperforms FVC and MSE in terms of subjective evaluations.

Target speaker  
Source speaker
Voice Conversion

- **CycleGAN-VC** [Kaneko et al., arXiv 2017]

\[
V_{Full} = V_{GAN}(G_{X \rightarrow Y}, D_Y) + V_{GAN}(G_{X \rightarrow Y}, D_Y) + \lambda V_{CyC}(G_{X \rightarrow Y}, G_{Y \leftarrow X})
\]
Voice Conversion (CycleGAN-VC)

- Subjective evaluations

Fig. 18: MOS for naturalness.

1. The proposed method uses **non-parallel** data.
2. For naturalness, the proposed method outperforms baseline.
3. For similarity, the proposed method is comparable to the baseline.
Voice Conversion

- Multi-target VC [Chou et al., arxiv 2018]

Stage-1

- $x$ → $\text{Enc} \rightarrow \text{Enc}(x) \rightarrow \text{Dec} \rightarrow y$
- $\text{Dec}(\text{Enc}(x), y)$
- $\text{Dec}(\text{Enc}(x), y')$

Stage-2

- $x$ → $\text{Enc} \rightarrow \text{Enc}(x) \rightarrow \text{Dec} \rightarrow y''$
- $\text{Dec}(\text{Enc}(x), y'')$
- $\text{D+C}$
- $\text{Real data}$
- $\text{F/R}$
- $\text{ID}$
Voice Conversion (Multi-target VC)

- Subjective evaluations

1. The proposed method uses **non-parallel** data.
2. The multi-target VC approach outperforms one-stage only.
3. The multi-target VC approach is comparable to Cycle-GAN-VC in terms of the naturalness and the similarity.
Outline of Part II

Speech Signal Generation
- Speech enhancement
- Postfilter, speech synthesis, voice conversion

Speech Signal Recognition
- Speech recognition
- Speaker recognition
- Speech emotion recognition
- Lip reading

Conclusion
Speech, Speaker, Emotion Recognition and Lip-reading (Classification Task)

Domain Adversarial Training

Output label

\[ y \]

\[ \text{Emb.} \]

\[ \tilde{z} = g(\tilde{x}) \]

\[ E \]

\[ E \]

\[ \tilde{z} \]

\[ h(\cdot) \]

\[ G \]

\[ \tilde{x} \]

\[ g(\cdot) \]

\[ \tilde{x} \]

\[ x \]

\[ \tilde{x} \]

\[ \tilde{x} \]

\[ \text{Clean data} \]

\[ \text{Noisy data} \]

\[ \text{Accented speech} \]

\[ \text{Channel distortion} \]

Acoustic Mismatch

Not only cheat the domain classifier, but satisfying label predictor at the same time

Successfully applied on image classification [Ganin et al, ICML, 2015][Ajakan et al., JMLR, 2016]
Speech Recognition

- Adversarial multi-task learning (AMT)
  [Shinohara Interspeech 2016]

Objective function

\[ V_y = -\sum_i \log P(y_i|x_i; \theta_E, \theta_G) \]
\[ V_z = -\sum_i \log P(z_i|x_i; \theta_E, \theta_D) \]

Model update

\[ \theta_G \leftarrow \theta_G - \epsilon \frac{\partial V_y}{\partial \theta_G} \]
Max classification accuracy

\[ \theta_D \leftarrow \theta_D - \epsilon \frac{\partial V_z}{\partial \theta_D} \]
Max domain accuracy

\[ \theta_E \leftarrow \theta_E - \epsilon \left( \frac{\partial V_y}{\partial \theta_G} \right) + \alpha \frac{\partial V_z}{\partial \theta_G} \]
Max classification accuracy and Min domain accuracy
Speech Recognition (AMT)

- ASR results in known (k) and unknown (unk) noisy conditions

Table 13: WER of DNNs with single-task learning (ST) and AMT.

<table>
<thead>
<tr>
<th>noise</th>
<th>ST</th>
<th>AMT</th>
<th>RERR</th>
</tr>
</thead>
<tbody>
<tr>
<td>car 2000cc</td>
<td>5.83</td>
<td>5.56</td>
<td>4.63</td>
</tr>
<tr>
<td>exhib. booth</td>
<td>6.80</td>
<td>6.66</td>
<td>2.06</td>
</tr>
<tr>
<td>station</td>
<td>7.89</td>
<td>7.76</td>
<td>1.65</td>
</tr>
<tr>
<td>crossing</td>
<td>6.96</td>
<td>6.65</td>
<td>4.45</td>
</tr>
<tr>
<td>car 1500cc</td>
<td>5.58</td>
<td>5.46</td>
<td>2.15</td>
</tr>
<tr>
<td>exhib. aisle</td>
<td>7.71</td>
<td>6.93</td>
<td>10.12</td>
</tr>
<tr>
<td>factory</td>
<td>12.17</td>
<td>12.92</td>
<td>-6.16</td>
</tr>
<tr>
<td>highway</td>
<td>9.73</td>
<td>9.52</td>
<td>2.16</td>
</tr>
<tr>
<td>crowd</td>
<td>6.72</td>
<td>6.40</td>
<td>4.76</td>
</tr>
<tr>
<td>server room</td>
<td>8.54</td>
<td>7.76</td>
<td>9.13</td>
</tr>
<tr>
<td>air cond.</td>
<td>6.96</td>
<td>6.98</td>
<td>-0.29</td>
</tr>
<tr>
<td>elev. hall</td>
<td>9.23</td>
<td>9.60</td>
<td>-4.01</td>
</tr>
<tr>
<td>average</td>
<td>7.84</td>
<td>7.68</td>
<td>2.04</td>
</tr>
</tbody>
</table>

The AMT-DNN outperforms ST-DNN with yielding lower WERs.
Speech Recognition

- Domain adversarial training for accented ASR (DAT)
  [Sun et al., ICASSP2018]

Objective function

\[ V_y = - \sum_i \log P(y_i|x_i; \theta_E, \theta_G) \]
\[ V_z = - \sum_i \log P(z_i|x_i; \theta_E, \theta_D) \]

Model update

\[ \theta_G \leftarrow \theta_G - \epsilon \frac{\partial V_y}{\partial \theta_G} \]
Max classification accuracy

\[ \theta_D \leftarrow \theta_D - \epsilon \frac{\partial V_z}{\partial \theta_D} \]
Max domain accuracy

\[ \theta_E \leftarrow \theta_E - \epsilon \left( \frac{\partial V_y}{\partial \theta_G} \right) + \alpha \frac{\partial V_z}{\partial \theta_G} \]
Max classification accuracy and Min domain accuracy
Speech Recognition (DAT)

- ASR results on accented speech

Table 14: WER of the baseline and adapted model.

<table>
<thead>
<tr>
<th>training data</th>
<th>$\lambda$</th>
<th>STD</th>
<th>FJ</th>
<th>JS</th>
<th>JX</th>
<th>SC</th>
<th>GD</th>
<th>HN</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>STD</td>
<td>-</td>
<td>15.55</td>
<td>23.58</td>
<td>15.75</td>
<td>14.08</td>
<td>15.62</td>
<td>15.32</td>
<td>19.34</td>
<td>17.28</td>
</tr>
<tr>
<td>STD + (600hrs with trans)</td>
<td>-</td>
<td>14.22</td>
<td>14.84</td>
<td>9.41</td>
<td>8.68</td>
<td>9.13</td>
<td>9.62</td>
<td>11.89</td>
<td>10.60</td>
</tr>
<tr>
<td>STD + (600hrs no trans)</td>
<td>0.03</td>
<td>15.37</td>
<td>22.96</td>
<td>14.48</td>
<td>13.79</td>
<td>15.35</td>
<td>14.86</td>
<td>18.24</td>
<td>16.61</td>
</tr>
</tbody>
</table>

STD: standard speech

1. With labeled transcriptions, ASR performance notably improves.
2. DAT is effective in learning features invariant to domain differences with and without labeled transcriptions.
Speech Recognition

• Robust ASR using GAN enhancer (GAN-Enhancer)

[Sriram et al., arXiv 2017]

Cross entropy with L1 Enhancer:
\[ H(h(\tilde{z}), y) + \lambda \frac{||z - \tilde{z}||_1}{||z||_1 + ||\tilde{z}||_1 + \epsilon} \]

Cross entropy with GAN Enhancer:
\[ H(h(\tilde{z}), y) + \lambda V_{adv} (g(x), g(\tilde{x})) \]
Speech Recognition (GAN-Enhancer)

• ASR results on far-field speech:

Fig. 15: WER of GAN enhancer and the baseline methods.

<table>
<thead>
<tr>
<th>Model</th>
<th>Near-Field</th>
<th>Far-Field</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CER</td>
<td>WER</td>
</tr>
<tr>
<td>seq-to-seq</td>
<td>7.43%</td>
<td>21.18%</td>
</tr>
<tr>
<td>seq-to-seq + far-field Augmentation</td>
<td>7.69%</td>
<td>21.32%</td>
</tr>
<tr>
<td>seq-to-seq + $L^1$-Distance Penalty</td>
<td>7.54%</td>
<td>20.45%</td>
</tr>
<tr>
<td>seq-to-seq + GAN Enhancer</td>
<td>7.78%</td>
<td>21.07%</td>
</tr>
</tbody>
</table>

GAN Enhancer outperforms the Augmentation and L1-Enhancer approaches on far-field speech.
Outline of Part II

Speech Signal Generation
- Speech enhancement
- Postfilter, speech synthesis, voice conversion

Speech Signal Recognition
- Speech recognition
- Speaker recognition
- Speech emotion recognition
- Lip reading

Conclusion
Speaker Recognition

- Domain adversarial neural network (DANN)
  [Wang et al., ICASSP 2018]
Speaker Recognition (DANN)

- Recognition results of domain mismatched conditions

Table 16: Performance of DAT and the state-of-the-art methods.

<table>
<thead>
<tr>
<th>Systems#</th>
<th>Adaptation Methods</th>
<th>EER%</th>
<th>DCF10 [21]</th>
<th>DCF08</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>–</td>
<td>9.35</td>
<td>0.724</td>
<td>0.520</td>
</tr>
<tr>
<td>2</td>
<td>–</td>
<td>5.66</td>
<td>0.633</td>
<td>0.427</td>
</tr>
<tr>
<td>3</td>
<td>Interpolated [6] [12]</td>
<td>6.55</td>
<td>0.652</td>
<td>0.454</td>
</tr>
<tr>
<td>4</td>
<td>IDV [9] [12]</td>
<td>6.15</td>
<td>0.676</td>
<td>0.476</td>
</tr>
<tr>
<td>5</td>
<td>DICN [11] [12]</td>
<td>4.99</td>
<td>0.623</td>
<td>0.416</td>
</tr>
<tr>
<td>6</td>
<td>DAE [22] [12]</td>
<td>4.81</td>
<td>0.610</td>
<td>0.398</td>
</tr>
<tr>
<td>7</td>
<td>AED [12]</td>
<td>4.50</td>
<td>0.589</td>
<td>0.362</td>
</tr>
<tr>
<td>8</td>
<td>DAT</td>
<td>3.73</td>
<td>0.541</td>
<td>0.335</td>
</tr>
</tbody>
</table>

The DAT approach outperforms other methods with achieving lowest EER and DCF scores.
Outline of Part II

Speech Signal Generation
- Speech enhancement
- Postfilter, speech synthesis, voice conversion

Speech Signal Recognition
- Speech recognition
- Speaker recognition
- Speech emotion recognition
- Lip reading

Conclusion
Emotion Recognition

- Adversarial AE for emotion recognition (AAE-ER)
  
  \[ Sahu \text{ et al., Interspeech 2017} \]

AE with GAN:

\[ H(h(z), x) + \lambda V_{GAN}(q, g(x)) \]

The distribution of code vectors
Emotion Recognition (AAE-ER)

- Recognition results of domain mismatched conditions:

Table 17: Classification results on different systems.

<table>
<thead>
<tr>
<th></th>
<th>OpenSmile features (1582-D)</th>
<th>Code vectors (2-D)</th>
<th>Auto-encoder (100-D)</th>
<th>LDA (2-D)</th>
<th>PCA (2-D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UAR (%)</td>
<td>57.88</td>
<td>56.38</td>
<td>53.92</td>
<td>48.67</td>
<td>43.12</td>
</tr>
</tbody>
</table>

Table 18: Classification results on real and synthesized features.

<table>
<thead>
<tr>
<th></th>
<th>UAR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syn.</td>
<td>25.00</td>
</tr>
<tr>
<td>S</td>
<td>33.75</td>
</tr>
<tr>
<td>Syn.</td>
<td>only 57.88 data points</td>
</tr>
</tbody>
</table>

1. AAE alone could not yield performance improvements.
2. Using synthetic data from AAE can yield higher UAR.
Outline of Part II

Speech Signal Generation

• Speech enhancement
• Postfilter, speech synthesis, voice conversion

Speech Signal Recognition

• Speech recognition
• Speaker recognition
• Speech emotion recognition
• Lip reading

Conclusion
Lip-reading

• Domain adversarial training for lip-reading (DAT-LR)
  [Wand et al., arXiv 2017]

Output 1
  Words

Output 2
  Speaker

Objective function

\[ V_y = -\sum_i \log P(y_i|x_i; \theta_E, \theta_G) \]
\[ V_z = -\sum_i \log P(z_i|x_i; \theta_E, \theta_D) \]

Model update

\[ \theta_G \leftarrow \theta_G - \epsilon \frac{\partial V_y}{\partial \theta_G} \]
\[ \theta_D \leftarrow \theta_D - \epsilon \frac{\partial V_z}{\partial \theta_D} \]
\[ \theta_E \leftarrow \theta_E - \epsilon \left( \frac{\partial V_y}{\partial \theta_G} \right) + \alpha \frac{\partial V_z}{\partial \theta_G} \]

Max classification accuracy
Max domain accuracy
Max classification accuracy and Min domain accuracy

~80% WAC
Lip-reading (DAT-LR)

- Recognition results of speaker mismatched conditions

Table 19: Performance of DAT and the baseline.

<table>
<thead>
<tr>
<th>Adversarial Training on</th>
<th>Number of training spk</th>
<th>Target Test acc.</th>
<th>Relative Improvement</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>1</td>
<td>18.7%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>39.4%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>46.5%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>All Target Sequences</td>
<td>1</td>
<td>25.4%</td>
<td>35.8%</td>
<td>0.0030*</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>43.6%</td>
<td>10.7%</td>
<td>0.0261*</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>49.3%</td>
<td>6.0%</td>
<td>0.0266*</td>
</tr>
<tr>
<td>50 Target Sequences</td>
<td>1</td>
<td>24.1%</td>
<td>28.9%</td>
<td>0.0045*</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>41.5%</td>
<td>5.3%</td>
<td>0.1367</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>47.0%</td>
<td>1.1%</td>
<td>0.3555</td>
</tr>
</tbody>
</table>

The DAT approach notably enhances the recognition accuracies in different conditions.
Outline of Part II

Speech Signal Generation
- Speech enhancement
- Postfilter, speech synthesis, voice conversion

Speech Signal Recognition
- Speech recognition
- Speaker recognition
- Speech emotion recognition
- Lip reading

Conclusion
Speech Signal Generation (Regression Task)

Paired

Objective function

Conditional GAN

$G$:

$c$: train

Prior distribution $z$

$x = G(c, z)$

$x$ is realistic or not + $c$ and $x$ are matched or not

True text-image pairs:

- (train, ) 1
- (cat, ) 0
- (train, Image) 0

Cycle-GAN

$G_{X \rightarrow Y}$

$G_{Y \rightarrow X}$

Scalar: belongs to domain X or not

$D_X$

$D_Y$

Scalar: belongs to domain Y or not

as close as possible

as close as possible
Speech, Speaker, Emotion Recognition and Lip-reading (Classification Task)

Domain Adversarial Training

Not only cheat the domain classifier, but satisfying label predictor at the same time

Successfully applied on image classification

[ Ganin et al, ICML, 2015][Jakan et al. JMLR, 2016 ]

Output label

\( y \)

\[ \tilde{z} = g(\tilde{x}) \]

\( G \)

\( h(\cdot) \)

\( E \)

\( g(\cdot) \)

\( \tilde{x} \)

Clean data

Acoustic Mismatch

\( \hat{x} \)

Channel distortion

\( \tilde{x} \)

Accented speech

\( \tilde{x} \)

Noisy data
More GANs in Speech

**Diagnosis of autism spectrum**

**Emotion recognition**

**Robust ASR**

**Speaker verification**
References

**Speech enhancement (conventional methods)**

**Speech enhancement (GAN-based methods)**
References

Postfilter (conventional methods)

Postfilter (GAN-based methods)
• Kaneko Takuhiro, Takaki Shinji, Kameoka Hirokazu, and Yamagishi Junichi, Generative Adversarial Network-Based Postfilter for STFT Spectrograms, Interspeech, 2017.
References

VC (conventional methods)

VC (GAN-based methods)
References

ASR
• Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie, Domain Adversarial Training for Accented Speech Recognition, ICASSP, 2018

Speaker recognition
• Qing Wang, Wei Rao, Sining Sun, Lei Xie, Eng Siong Chng, and Haizhou Li, Unsupervised Domain Adaptation via Domain Adversarial Training for Speaker Recognition, ICASSP, 2018.

Emotion recognition

Lipreading
GANs in ICASSP 2018

• Saurabh Sahu, Rahul Gupta, Ganesh Sivaraman, Carol Espy-Wilson, Smoothing Model Predictions using Adversarial Training Procedures for Speech Based Emotion Recognition
• Fuming Fang, Junichi Yamagishi, Isao Echizen, Jaime Lorenzo-Trueba, High-quality Nonparallel Voice Conversion Based on Cycle-consistent Adversarial Network
• Lauri Juvela, Bajibabu Bollepalli, Xin Wang, Hirokazu Kameoka, Manu Airaksinen, Junichi Yamagishi, Paavo Alku, Speech Waveform Synthesis from MFCC Sequences with Generative Adversarial Networks
• Zhong Meng, Jinyu Li, Yifan Gong, Biing-Hwang (Fred) Juang, Adversarial Teacher-Student Learning for Unsupervised Domain Adaptation
• Zhong Meng, Jinyu Li, Zhuo Chen, Yong Zhao, Vadim Mazalov, Yifan Gong, Biing-Hwang (Fred) Juang, Speaker-Invariant Training via Adversarial Learning
• Sen Li, Stephane Villette, Pravin Ramadas, Daniel Sinder, Speech Bandwidth Extension using Generative Adversarial Networks
• Qing Wang, Wei Rao, Sining Sun, Lei Xie, Eng Siong Chng, Haizhou Li, Unsupervised Domain Adaptation via Domain Adversarial Training for Speaker Recognition
• Hu Hu, Tian Tan, Yanmin Qian, Generative Adversarial Networks Based Data Augmentation for Noise Robust Speech Recognition
• Yuki Saito, Shinnosuke Takamichi, Hiroshi Saruwatari, Text-to-speech Synthesis using STFT Spectra Based on Low-/multi-resolution Generative Adversarial Networks
• Jiangyan Yi, Jianhua Tao, Zhengqi Wen, Ye Bai, Adversarial Multilingual Training for Low-resource Speech Recognition
• Taira Tsuchiya, Naohiro Tawara, Tetsuji Ogawa, Tetsunori Kobayashi, Speaker Invariant Feature Extraction for Zero-resource Languages with Adversarial Learning
GANs in ICASSP 2018

- Jing Han, Zixing Zhang, Zhao Ren, Fabien Ringeval, Bjoern Schuller, Towards Conditional Adversarial Training for Predicting Emotions from Speech
- Chenxing Li, Lei Zhu, Shuang Xu, Peng Gao, Bo Xu, CBLDNN-based Speaker-independent Speech Separation via Generative Adversarial Training
- Anuroop Sriram, Heewoo Jun, Yashesh Gaur, Sanjeev Satheesh, Robust Speech Recognition using Generative Adversarial Networks
- Cem Subakan, Paris Smaragdis, Generative Adversarial Source Separation,
- Ashutosh Pandey, Deliang Wang, On Adversarial Training and Loss Functions for Speech Enhancement
- Bin Liu, Shuai Nie, Yaping Zhang, Dengfeng Ke, Shan Liang, Wenju Liu, Boosting Noise Robustness of Acoustic Model via Deep Adversarial Training
- Yang Gao, Rita Singh, Bhiksha Raj, Voice Impersonation using Generative Adversarial Networks
- Aditay Tripathi, Aanchan Mohan, Saket Anand, Maneesh Singh, Adversarial Learning of Raw Speech Features for Domain Invariant Speech Recognition
A promising research direction and still has room for further improvements in the speech signal processing domain

Thank You Very Much

Tsao, Yu  Ph.D.
yu.tsao@citi.sinica.edu.tw
https://www.citi.sinica.edu.tw/pages/yu.tsao/contact_zh.html