1. Introduction

- Voice conversion (VC): modifies speaker individuality.

- Two VC categories -- parallel & nonparallel:
<table>
<thead>
<tr>
<th>Training data</th>
<th>Alignment</th>
<th>Quality</th>
<th>Practicality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallel</td>
<td>Paired</td>
<td>Yes</td>
<td>Good</td>
</tr>
<tr>
<td>Nonparallel</td>
<td>Unpaired</td>
<td>Yes/No</td>
<td>Bad</td>
</tr>
</tbody>
</table>

- Goal of this research:
  To develop a high-quality nonparallel VC method

2. Current popular parallel VC methods

- DNN-based method:
  
  - Input: Source speaker
  
  - Output: Target speaker

- GAN-based method:
  
  - $D$: minimizes an adversarial loss, while $G$ maximizes it.

  $\mathbb{E} = L_{GAN}(X,Y) + \alpha \cdot MSE(\hat{Y},Y)$

  $L_{GAN}(X,Y) = \mathbb{E}[\log D(Y)] + \mathbb{E}[\log(1 - D(G(X)))]$

  $MSE(\hat{Y},Y) = \mathbb{E}[\|G(X) - Y\|_2]$}

3. Proposed nonparallel VC method

- CycleGAN [Zhu et al., 2017]:
  - Originally developed for unpaired image-to-image translation.
    * $D_1, D_2$: discriminators
    * $G_1, G_2$: generators

  $\mathbb{E}_G = L_{GAN_1}(X,Y) + L_{GAN_2}(\hat{Y},X)$

  $L_{GAN_1}(X,Y) = \mathbb{E}[\log D_1(Y)]$

  $L_{GAN_2}(\hat{Y},X) = \mathbb{E}[\log D_2(\hat{Y})]$

  1. Adversarial loss:

  $\mathbb{E}_G = \mathbb{E}[\log D_1(Y)] + \mathbb{E}[\log(1 - D_2(G_1(X)))]$

  2. Cycle-consistent loss:

  $\mathbb{E}_G = \mathbb{E}[\|G_2(G_1(X)) - X\|_2]$

  $\mathbb{E}_G = \mathbb{E}[\|G_1(G_2(Y)) - Y\|_2]$