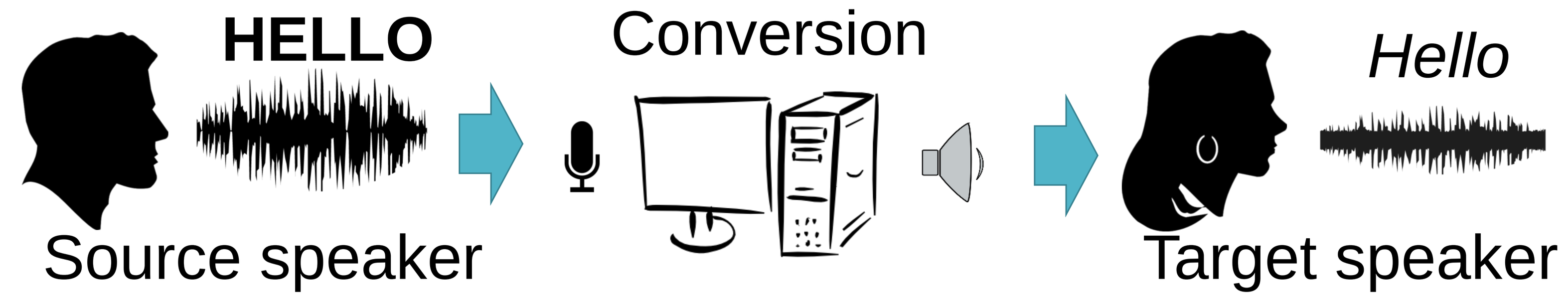
High-quality nonparallel voice conversion based on cycle-consistent adversarial network

Fuming Fang[1], Junichi Yamagishi[1,2], Isao Echizen[1], Jaime Lorenzo-Trueba[1]
[1]National Institute of Informatics, [2]University of Edinburgh

1. Introduction

- Voice conversion (VC): modifies speaker individuality.



- Two VC categories -- parallel & nonparallel:

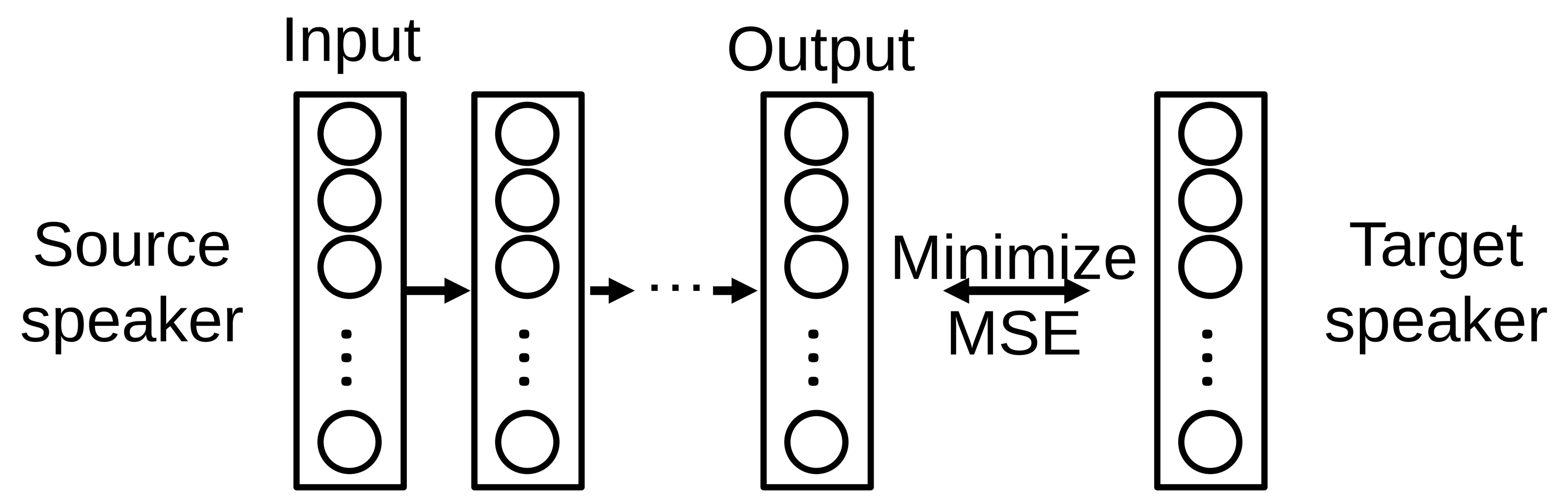
	Training data	Alignment	Quality	Practicality
Parallel	Paired	Yes	Good	Low
Nonparallel	Unpaired	Yes/No	Bad	High

- Goal of this research:

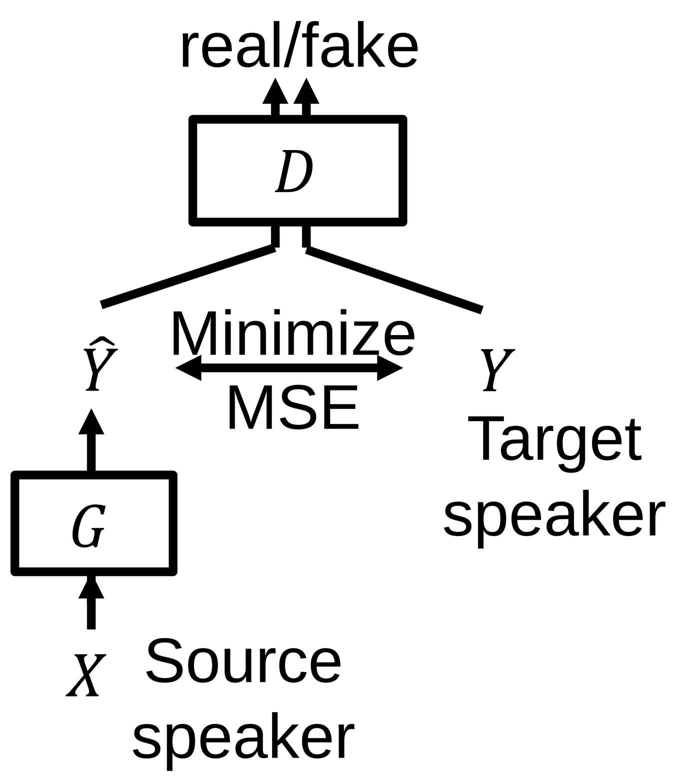
To develop a high-quality nonparallel VC method

2. Current popular parallel VC methods

- DNN-based method:



- GAN-based method:



- G minimizes an adversarial loss, while D maximizes it.

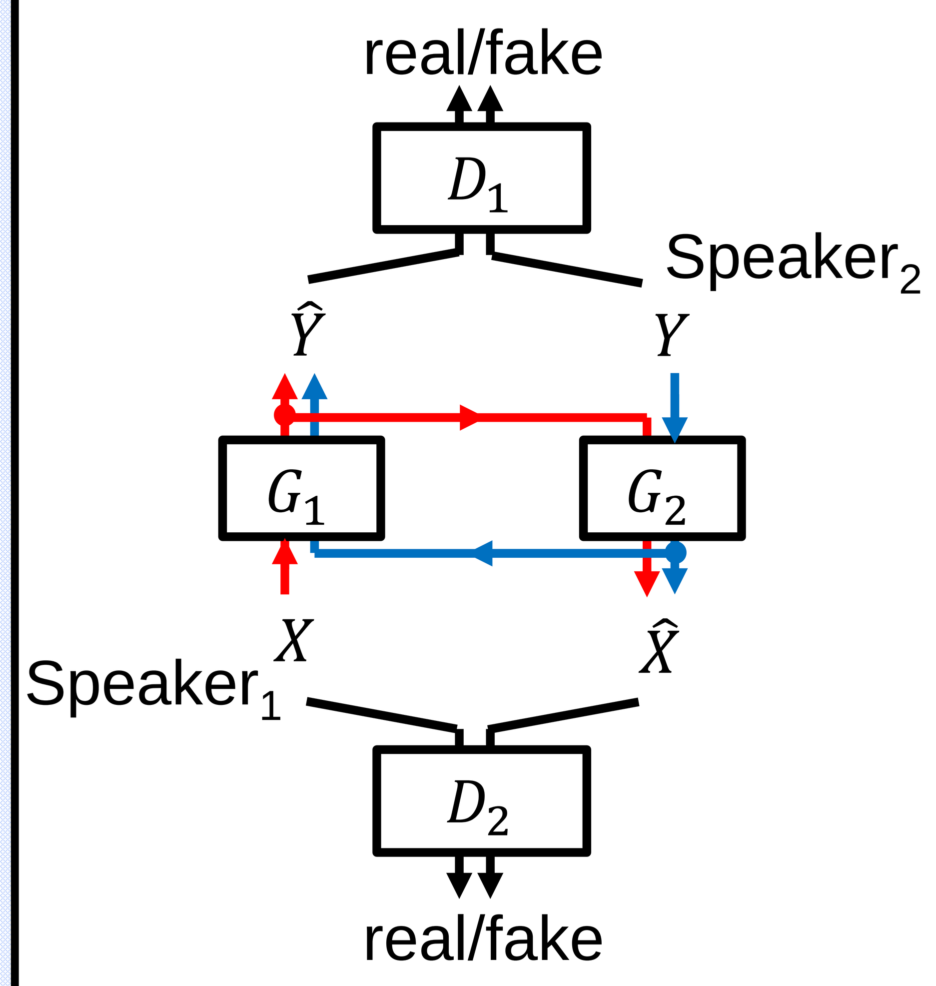
$$loss = L_{GAN}(X,Y) + \alpha \cdot MSE(\hat{Y},Y)$$

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

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

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

$$loss = L_{GAN_1}(X,Y) + L_{GAN_2}(Y,X) + \beta \cdot L_{cyc}(X,Y)$$

① Adversarial loss:

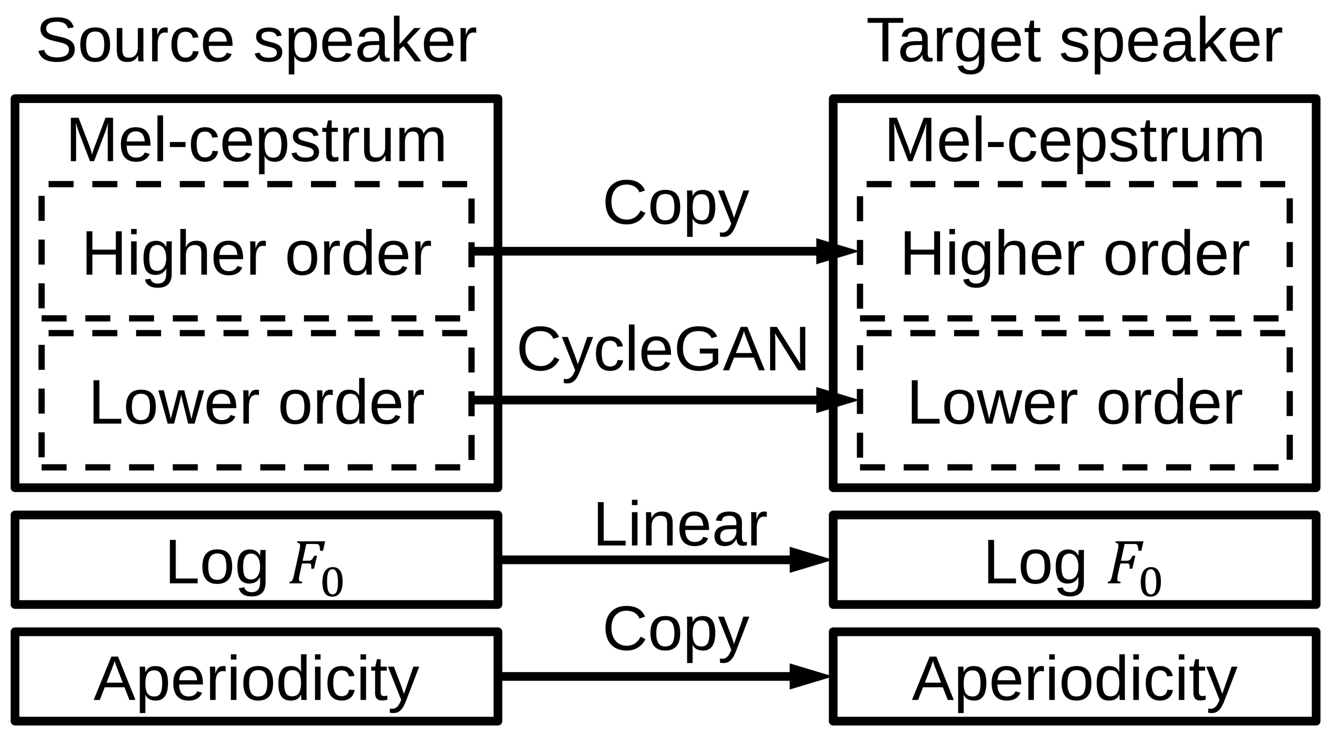
$$L_{GAN_1}(X,Y) = E[\log D_1(Y)] + E[\log(1 - D_1(G_1(X)))]$$

$$L_{GAN_2}(Y,X) = E[\log D_2(X)] + E[\log(1 - D_2(G_2(Y)))]$$

② Cycle-consistent loss:

$$L_{cyc}(X,Y) = E\left[\|G_2(G_1(X)) - X\|_1\right] + E\left[\|G_1(G_2(Y)) - Y\|_1\right]$$

- CycleGAN-based nonparallel VC:
  - Adversarial loss: modifies speaker individuality.
  - Cycle-consistent loss: keeps linguistic contents.

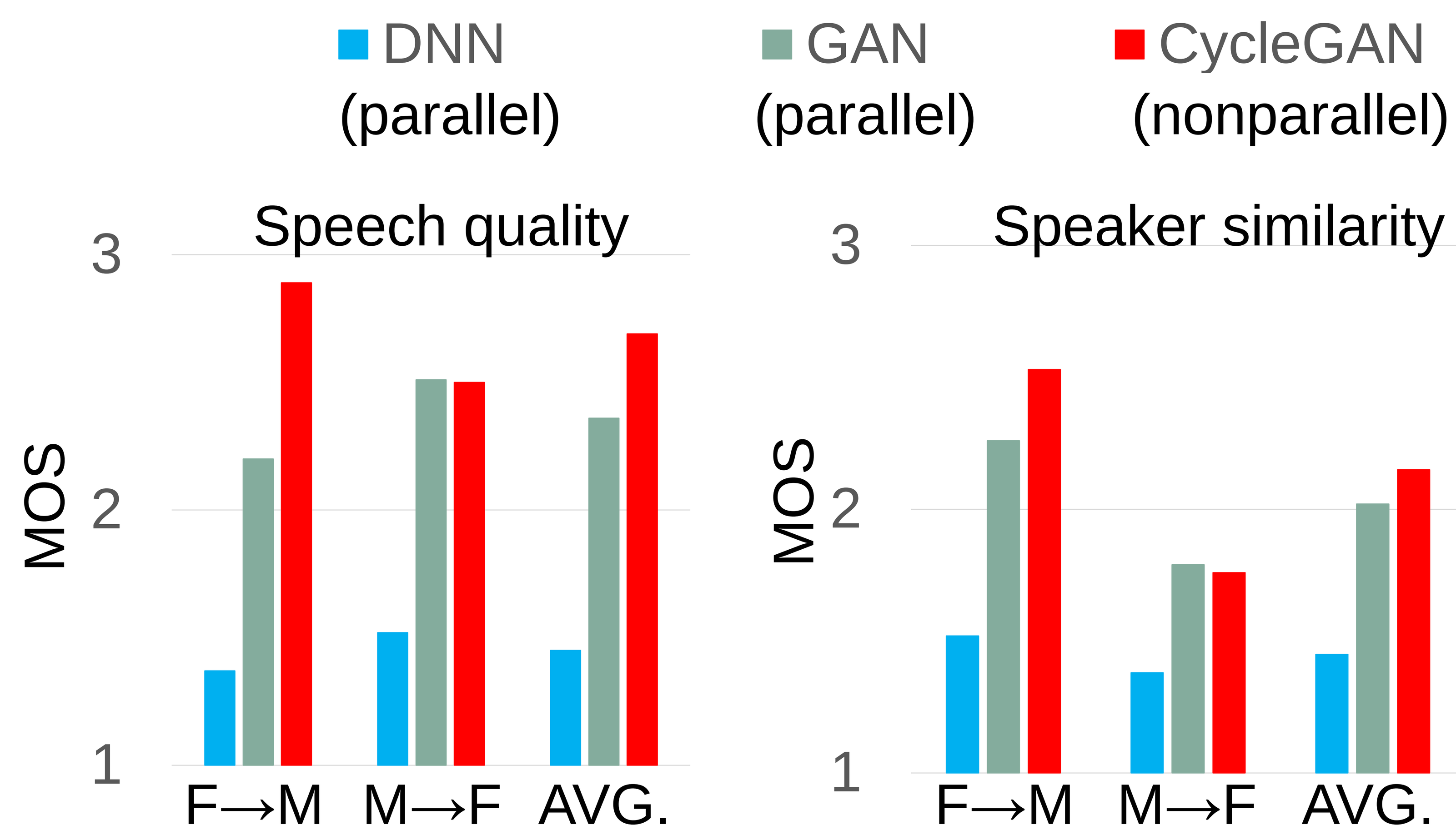


* Higher order: fine structure
* Lower order: spectral envelope
(motivated by highway VC [Saito et al., 2017])

4. Experiment

Baselines	1. DNN-based parallel VC 2. GAN-based parallel VC
Conversion tasks	Female-to-male (F → M) Male-to-female (M →F)
Corpus	ALAGIN Japanese speech database (sampling frequency: 20KHz)
Training data	Paired 200 utterances (parallel case) Unpaired 200 utterances (nonparallel case)
Test data	50 utterances
Feature vectors	75 dimensions (lower order = the first 25 coefficients out of 49, 1st and 2nd derivatives)
# hidden units	128, 256, 256, 128 (sigmoid)
Batch size	128 randomly selected frames

- Result (110 evaluators, 24 data points/utterance):



* CycleGAN (nonparallel) outperformed DNN&GAN (parallel).
* Reason: perhaps no alignment error occurred.

5. Conclusion & Future work

- Proposed a nonparallel VC method based on CycleGAN.
- Proposed method achieved higher performance than parallel VC methods.
- Plan to improve CycleGAN to guarantee linguistic contents invariants.