DRVC: A FRAMEWORK OF ANY-TO-ANY VOICE CONVERSION WITH SELF-SUPERVISED LEARNING

Authors: Qiqi Wang, Xulong Zhang, Jianzong Wang, Ning Cheng, Jing Xiao

Speaker: Qiqi Wang (virtual)

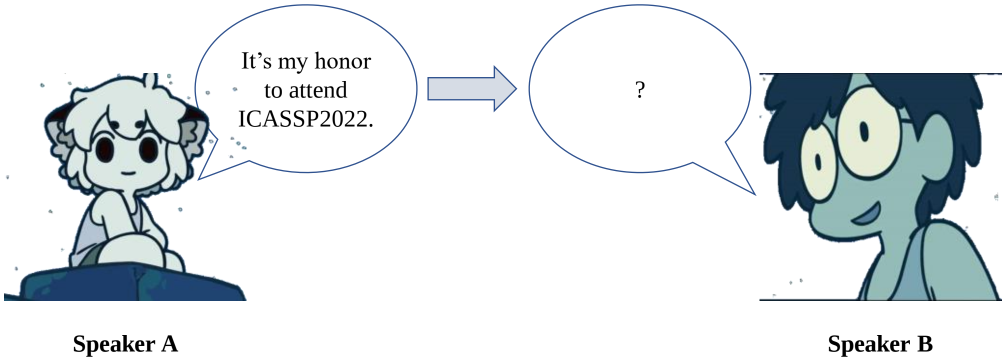
10th April 2022

Outline

- Background
- DRVC
- Experiments
- Conclusion

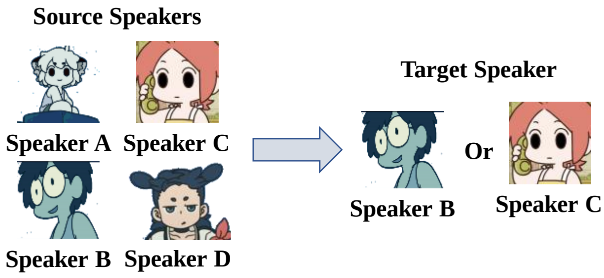
Background

- Voice Conversion

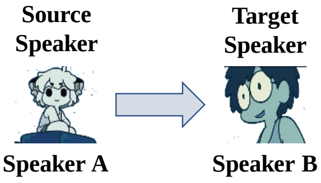


Background

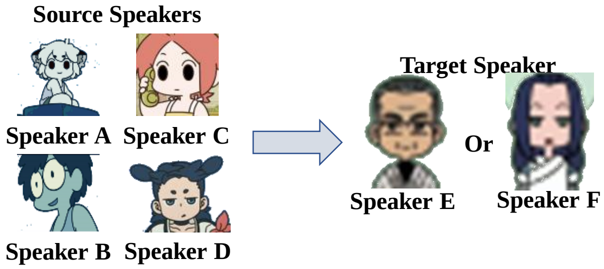
- Voice Conversion



Many-to-Many VC



One-to-one VC

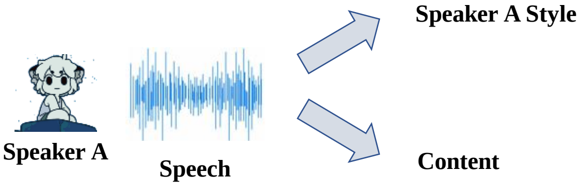


Any-to-Any VC

Background

- Voice Conversion
  - Previous method (Disentanglement-based)

Assumption: *Speech information consists of speaker style and content information*.



Speaker A

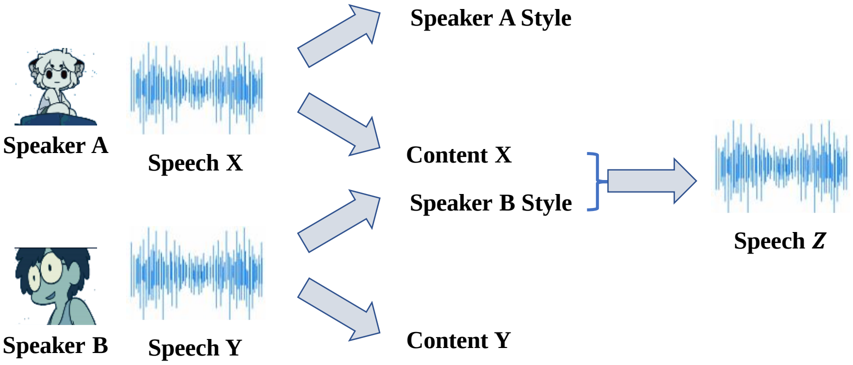
Speech

Speaker A Style

Content

Background

- Voice Conversion
  - Previous method (Disentanglement-based)



Speaker A

Speech X

Speaker A Style

Content X

Speaker B Style

Speech *Z*

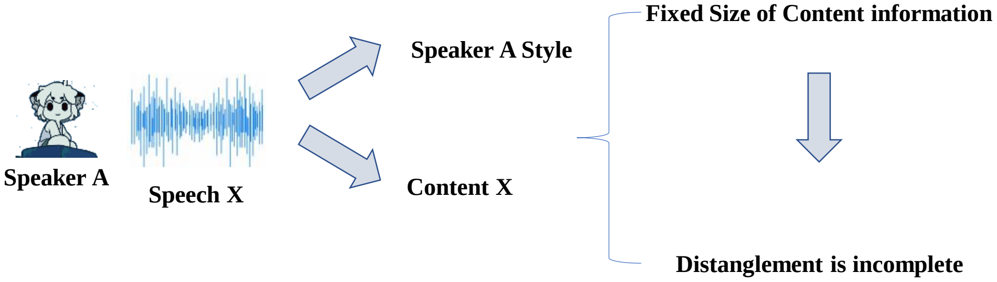
Speaker B

Speech Y

Content Y

Background

- Voice Conversion
  - Shortages



Speaker A

Speech X

Speaker A Style

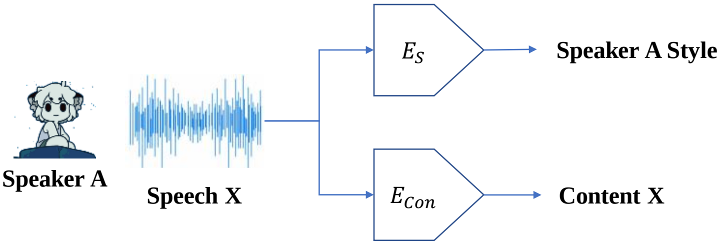
Content X

Fixed Size of Content information

Distanglement is incomplete

DRVC

- Speech Distanglement
  - Two encoders
    - Speaker Style Encoder: E_S
    - Content Encoder: E_{Con}



Speaker A

Speech X

E_S

E_{Con}

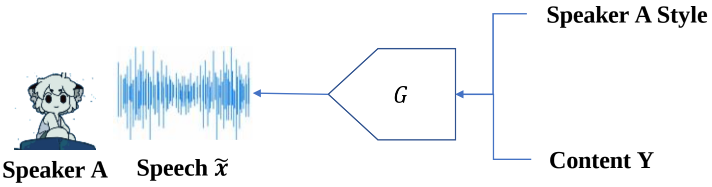
Speaker A Style

Content X

$$\{x_c, x_s\} = \{E_{Con}(X), E_S(X)\}$$

DRVC

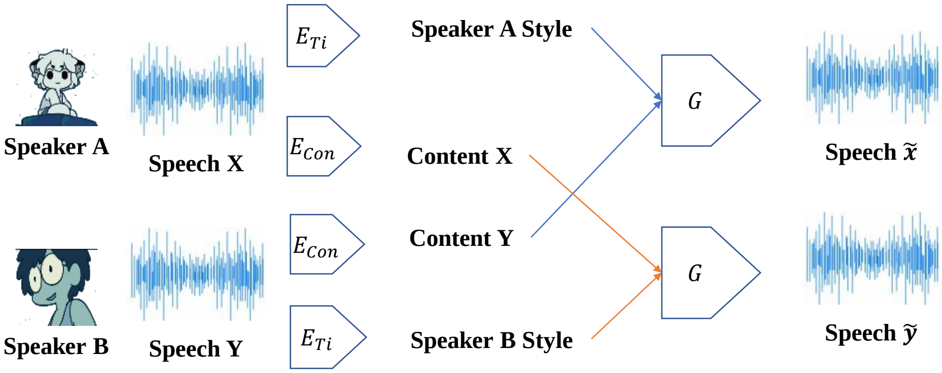
- Speech Distanglement
  - Generator G

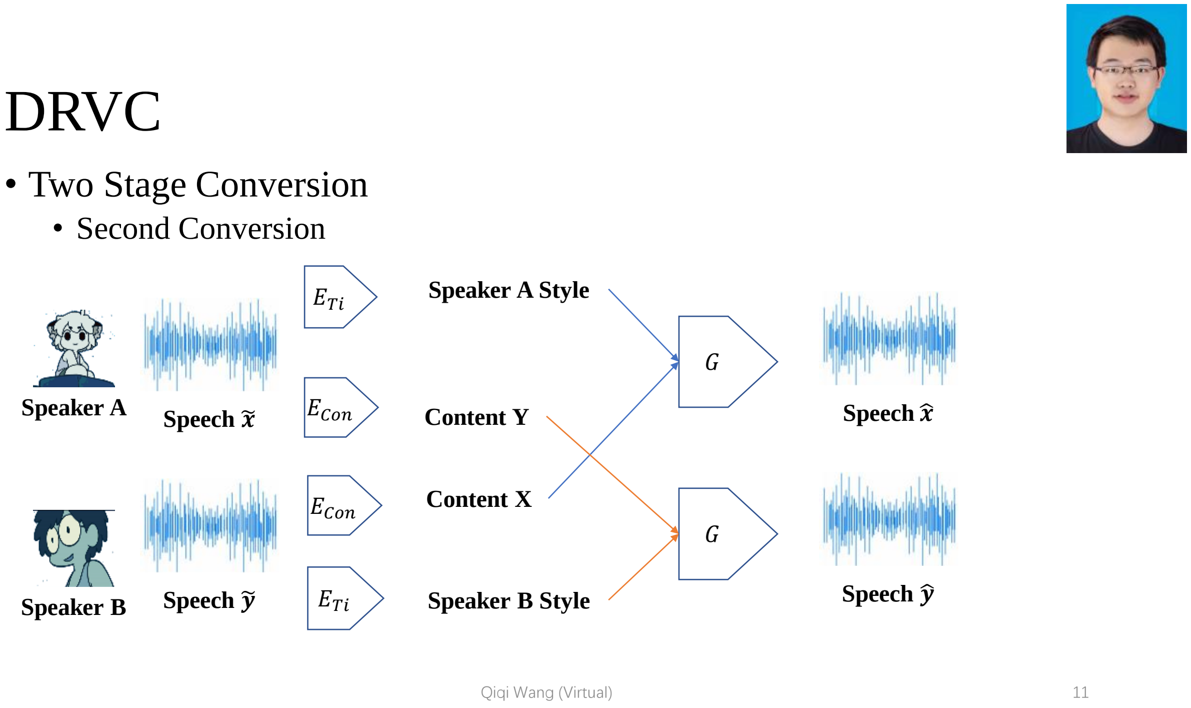


$$\tilde{x} = G(y_c, x_s) = G(\{E_{Con}(Y), E_S(X)\})$$

DRVC

- Two Stage Conversion
  - First Conversion





DRVC

- Two Stage Conversion
  - Second Conversion

Speaker A

Speech $\tilde{x}$

E_{Ti}

Speaker A Style

E_{Con}

Content Y

G

Speech $\hat{x}$

Speaker B

Speech $\tilde{y}$

E_{Con}

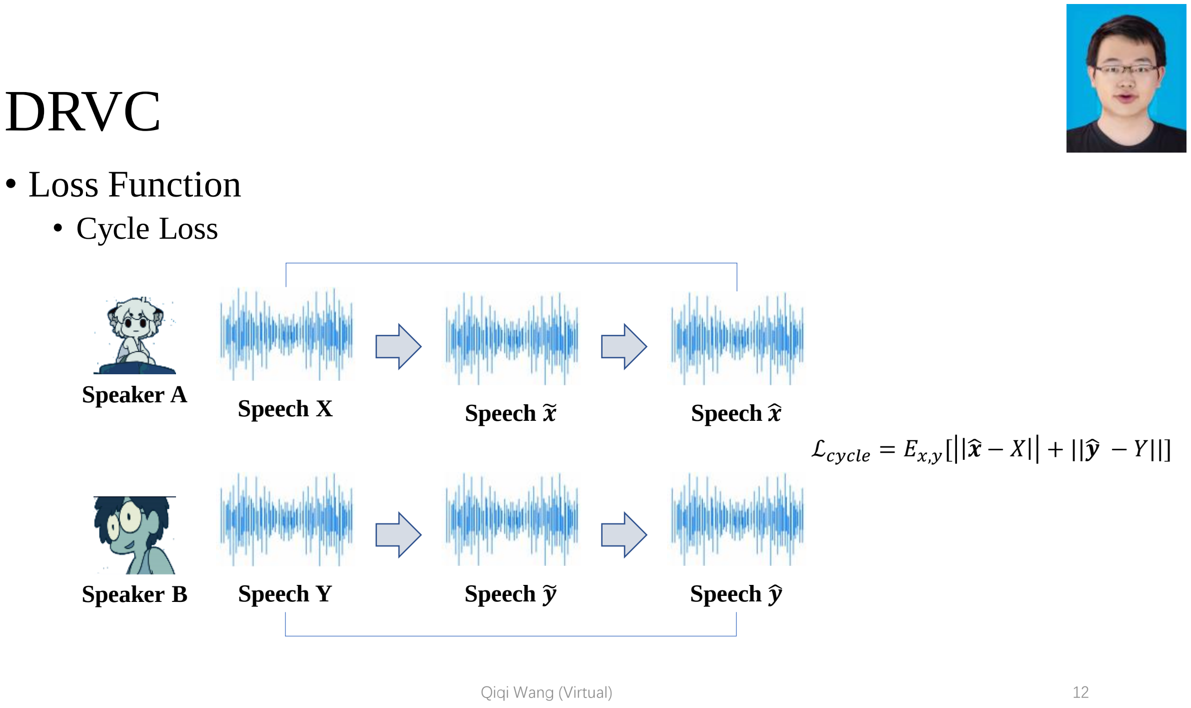
Content X

E_{Ti}

Speaker B Style

G

Speech $\hat{y}$



DRVC

- Loss Function
  - Cycle Loss

Speaker A

Speech X

Speech $\tilde{x}$

Speech $\hat{x}$

Speaker B

Speech Y

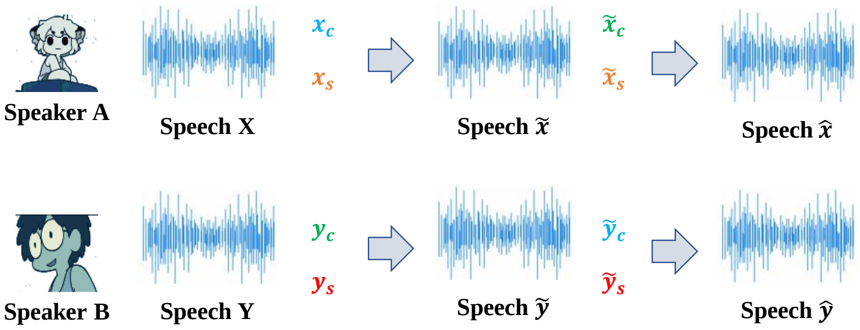
Speech $\tilde{y}$

Speech $\hat{y}$

$$\mathcal{L}_{cycle} = E_{x,y}[||\hat{x} - X|| + ||\hat{y} \ - Y||]$$

DRVC

- Loss Function
  - Same Loss



Speaker A

Speech X

x_c

x_s

Speech $\tilde{x}$

$\tilde{x}_c$

$\tilde{x}_s$

Speech $\hat{x}$

Speaker B

Speech Y

y_c

y_s

Speech $\tilde{y}$

$\tilde{y}_c$

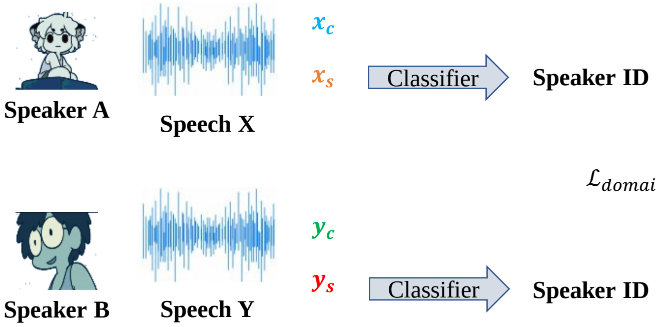
$\tilde{y}_s$

Speech $\hat{y}$

$$\mathcal{L}_{same} = E[|\tilde{y}_c - x_c| + |\tilde{x}_c - y_c|] + E[|\tilde{x}_s - x_s| + |\tilde{y}_s - y_s|]$$

DRVC

- Loss Function
  - Domain Loss



Speaker A

Speech X

x_c

x_s

Classifier

Speaker ID

Speaker B

Speech Y

y_c

y_s

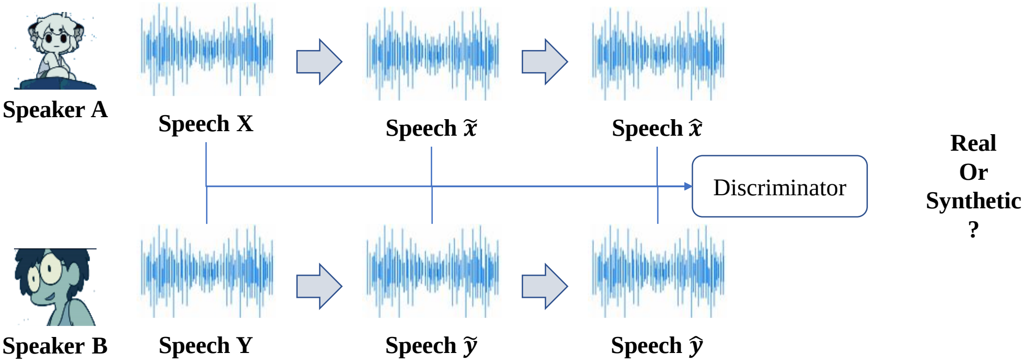
Classifier

Speaker ID

$$\mathcal{L}_{domain} = -\frac{1}{2}\left(\sum_i A(i)C(x_s) + \sum_i B(i)C(y_s)\right)$$

DRVC

- Loss Function
  - Adversarial Loss



Speaker A

Speech X

Speech $\tilde{x}$

Speech $\hat{x}$

Discriminator

Real
Or
Synthetic
?

Speaker B

Speech Y

Speech $\tilde{y}$

Speech $\hat{y}$

Experiments

- Data
  - VCC2018

Target Speakers	
VCC2SF4	VCC2SM4
VCC2TF2	VCC2TM2

Many-to-Many VC

Sources Speakers	
VCC2SF1	VCC2SM1
VCC2SF2	VCC2SM2
VCC2SF4	VCC2SM4
VCC2TF2	VCC2TM2

Target Speakers	
VCC2SF3	VCC2SM3
VCC2TF1	VCC2TM1

Any-to-Any VC

Experiments

- Result
  - MCD & MOS

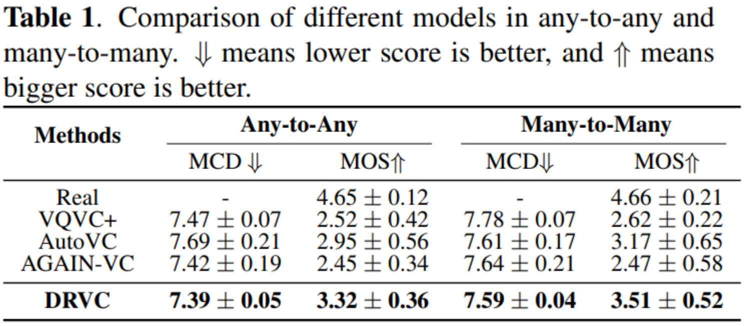
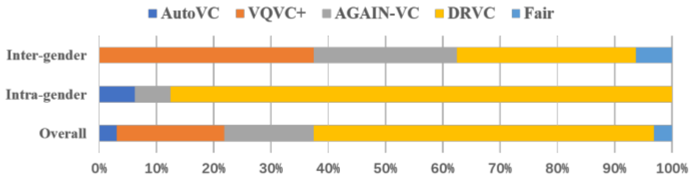


Table 1. Comparison of different models in any-to-any and many-to-many. ⇓ means lower score is better, and ⇑ means bigger score is better.

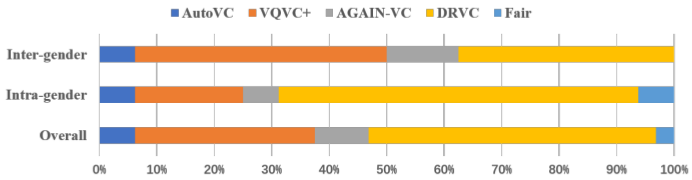
Methods	**Any-to-Any**		**Many-to-Many**	
	MCD ⇓	MOS⇑	MCD⇓	MOS⇑
Real	-	4.65 ± 0.12	-	4.66 ± 0.21
VQVC+	7.47 ± 0.07	2.52 ± 0.42	7.78 ± 0.07	2.62 ± 0.22
AutoVC	7.69 ± 0.21	2.95 ± 0.56	7.61 ± 0.17	3.17 ± 0.65
AGAIN-VC	7.42 ± 0.19	2.45 ± 0.34	7.64 ± 0.21	2.47 ± 0.58
DRVC	**7.39 ± 0.05**	**3.32 ± 0.36**	**7.59 ± 0.04**	**3.51 ± 0.52**

Experiments

- Result
  - Human Evaluation



(a) Any-to-any phase



(b) Many-to-many phase

Experiments

- Result
  - Ablation experiments

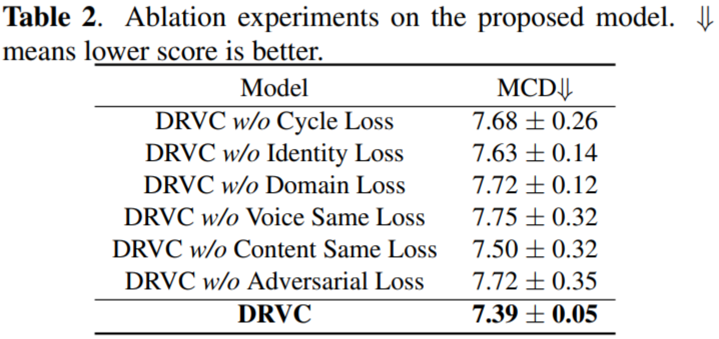


Table 2. Ablation experiments on the proposed model. ⇓ means lower score is better.

Model	MCD⇓
DRVC *w/o* Cycle Loss	7.68 ± 0.26
DRVC *w/o* Identity Loss	7.63 ± 0.14
DRVC *w/o* Domain Loss	7.72 ± 0.12
DRVC *w/o* Voice Same Loss	7.75 ± 0.32
DRVC *w/o* Content Same Loss	7.50 ± 0.32
DRVC *w/o* Adversarial Loss	7.72 ± 0.35
DRVC	**7.39** ± **0.05**

Conclusion

- Contribution
  - We propose a end-to-end framework, DRVC, to address the untangle overlapping problem without circumspection choose the content sizes.
  - Both the subjective and objective results show our model has better performance.

Thanks for you listening

Acknowledge & Notes:

- All anime character images are from the ‘*The Legend of LUOXIAOHEI*’.
- The presentation speech video, including the voice and personal video, is auto synthesis by PingAn Technology Co. Ltd.