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

Speaker A

It’s my honor to attend ICASSP2022.

Speaker B

Qiqi Wang (Virtual)

One-to-one VC

Source Speaker

Speaker A

Target Speaker

Speaker B

Many-to-Many VC

Source Speakers

Speaker A

Speaker C

Speaker B

Speaker D

Target Speaker

Speaker B

Or

Speaker C

Any-to-Any VC

Source Speakers

Speaker A

Speaker C

Speaker B

Speaker D

Target Speaker

Speaker E

Or

Speaker F

Qiqi Wang (Virtual)
Background

- Voice Conversion
  - Previous method (Disentanglement-based)

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

• Voice Conversion
  • Shortages

DRVC

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

\[
\{x_c, x_s\} = \{E_{Con}(X), E_S(X)\}
\]
**DRVC**

- **Speech Distanglement**
  - Generator $G$

\[
\bar{x} = G(y_c, x_s) = G((E_{Con}(Y), E_{S}(X)))
\]

**DRVC**

- **Two Stage Conversion**
  - First Conversion
**DRVC**

- Two Stage Conversion
  - Second Conversion

![Diagram of DRVC](image)

**DRVC**

- Loss Function
  - Cycle Loss

![Diagram of DRVC](image)

\[ L_{cycle} = E_{x,y} [||\tilde{x} - x|| + ||\tilde{y} - y||] \]
DRVC

• Loss Function
  • Same Loss

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

Speaker A
Speech X
\(x_c\)
\(x_s\)

Speaker B
Speech Y
\(y_c\)
\(y_s\)

DRVC

• Loss Function
  • Domain Loss

\[
\mathcal{L}_{\text{domain}} = -\frac{1}{2} \left( \sum_i A(i)\mathcal{L}(x_i) + \sum_i B(i)\mathcal{L}(y_i) \right)
\]

Classifier
Speaker ID

Speaker A
Speech X
\(x_c\)
\(x_s\)

Classifier
Speaker ID

Speaker B
Speech Y
\(y_c\)
\(y_s\)

Qiqi Wang (Virtual)
DRVC

• Loss Function
  • Adversarial Loss

Experiments

• Data
  • VCC2018

| Sources Speakers | Target Speakers |
|------------------|-----------------|
| VCC2SF1          | VCC2SM1         |
| VCC2SF2          | VCC2SM2         |
| VCC2SF4          | VCC2SM4         |
| VCC2TF2          | VCC2TM2         |

Many-to-Many VC

| Target Speakers |
|-----------------|
| VCC2SF4         |
| VCC2TF2         |

Any-to-Any VC

| Target Speakers |
|-----------------|
| VCC2SF3         |
| VCC2TF1         |
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
Experiments

• Result
  • Ablation experiments

| Model                   | MCD \(\downarrow\) |
|-------------------------|---------------------|
| 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 LUOXIAOHUI’.
- The presentation speech video, including the voice and personal video, is auto synthesis by PingAn Technology Co. Ltd.