Voice Conversion (VC) Using Deep Generative Models: Some Advanced Approaches

Sandipan Dhar

PhD Scholar
Department of Computer Science and Engineering
National Institute of Technology Durgapur

Contents:

Overview of VC

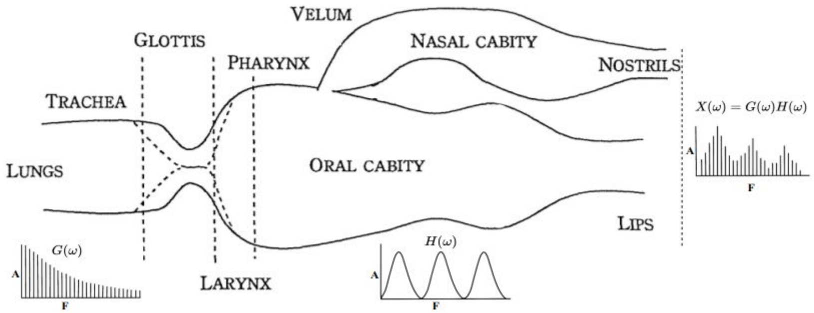


Figure 1: Schematic overview of the human speech production mechanism

$G(\omega)$ is the frequency response of the excitation signal

$H(\omega)$ is the frequency response of the transformation function

$X(\omega)$ is the frequency response of the output speech signal

$$x(t) = \int_0^t g(\tau)h(t-\tau)d\tau \quad (1)$$

$$X(\omega) = G(\omega)H(\omega) \quad (2)$$

[1] Lawrence R. Rabiner, & Ronald W. Schafer (2007). Introduction to Digital Speech Processing. *Found. Trends Signal Process., 1, 1-194.*

Overview of VC (continue)

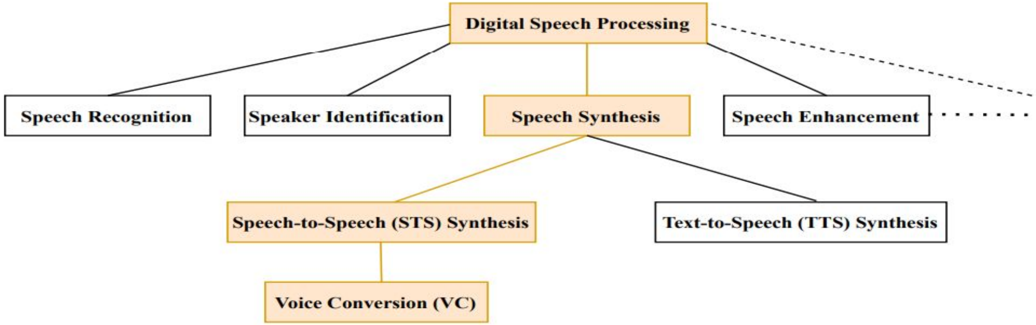


Figure 2: Categorization of the speech synthesis processes

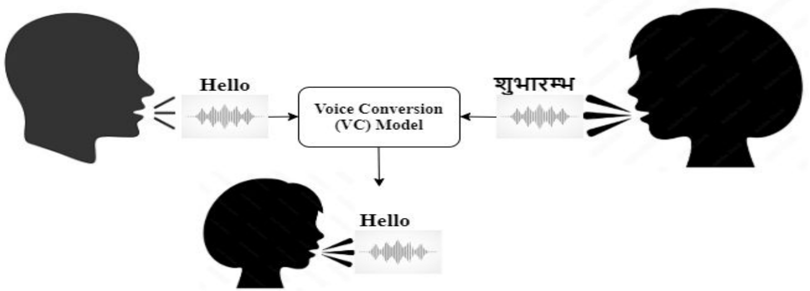
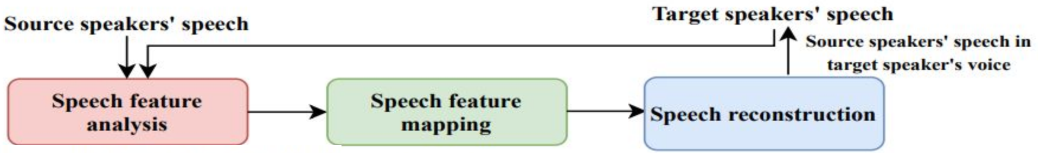


Figure 3: Schematic overview of the VC process

[2] Berrak Sisman, Junichi Yamagishi, Simon King, & Haizhou Li (2020). An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 132-157.*

Overview of VC (continue)



Mel-cepstral coefficients (MCC), Modulation spectra (MS), Mel-frequency cepstral coefficients (MFCC), Mel-spectrogram, etc.

$$F(.) = f_k(...(f_3(f_2(f_1(.)))))$$

$$\mathbf{x} = F(\mathbf{s})$$

$$\hat{\mathbf{y}} = G(\mathbf{x})$$

$$\hat{\mathbf{t}} = F^{-1}(\hat{\mathbf{y}})$$

Statistical and Neural Vocoders:

WORLD, STRAIGHT, Griffin-Lim algorithm (GLA), etc

MelGAN, WaveGAN, etc.

Figure 4 : Three-stage pipeline of the VC process

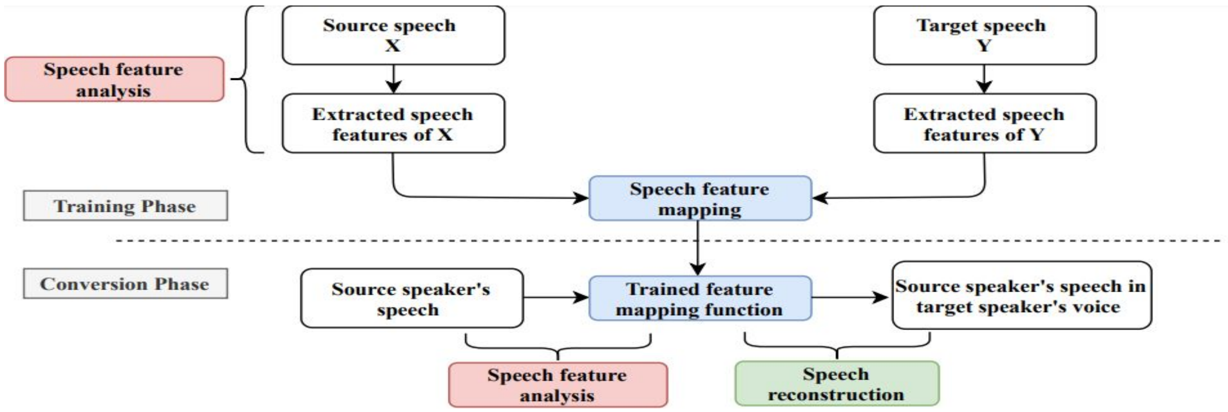


Figure 5: Implementation of the three-stage pipeline in the training and conversion phase of the VC process

[3] M. T. Akhter, P. Banerjee, S. Dhar and N. D. Jana, "An Analysis of Performance Evaluation Metrics for Voice Conversion Models," 2022 IEEE 19th India Council International Conference (INDICON), Kochi, India, 2022, pp. 1-6, doi: 10.1109/INDICON56171.2022.10040000.

Overview of VC (continue)

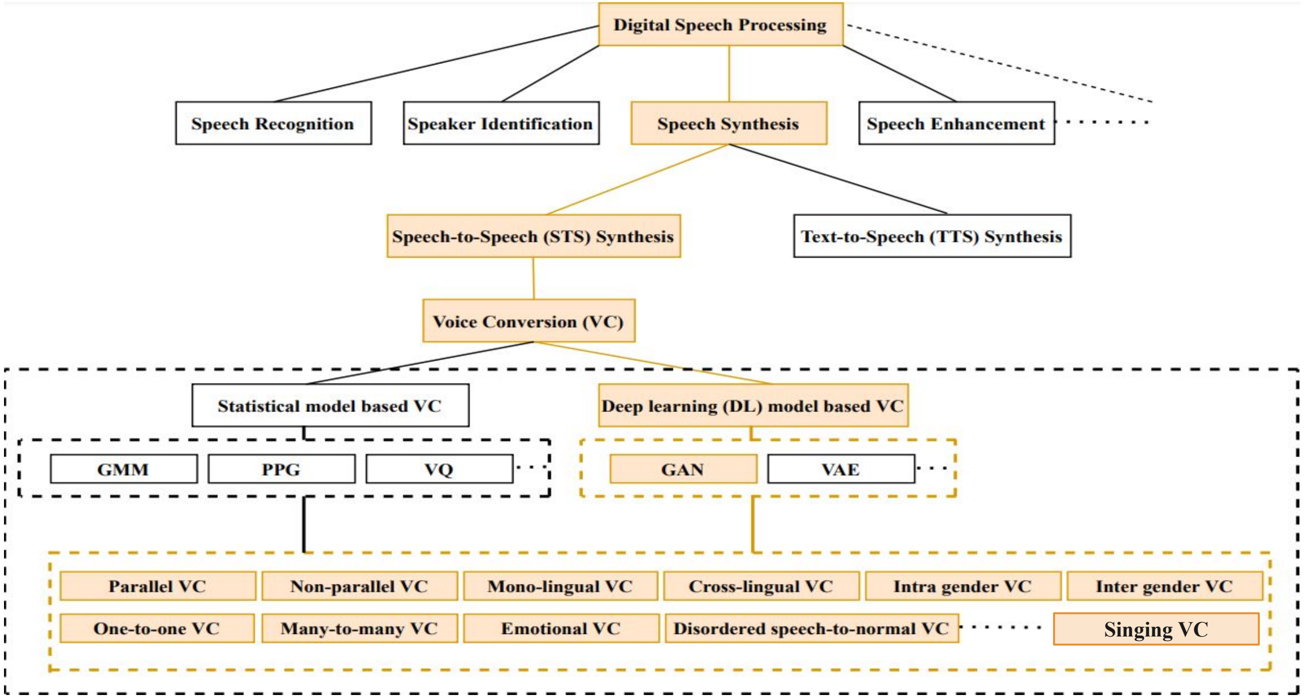


Figure 6: Types of VC processes

Overview of VC (continue)

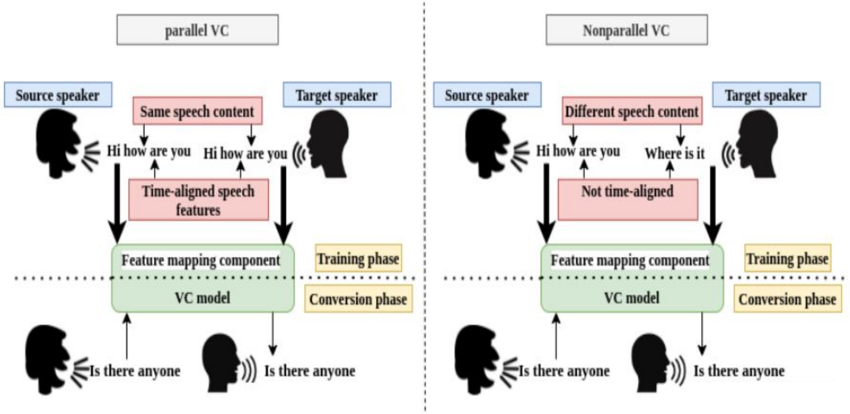


Figure 7: Schematic overview of the parallel VC and non-parallel VC process

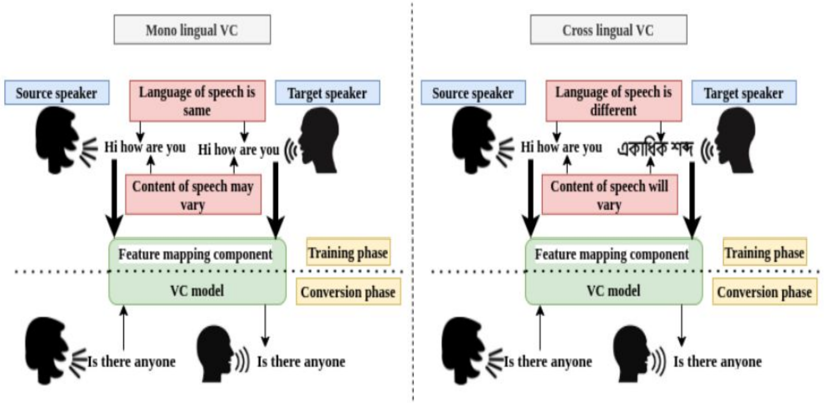


Figure 8: Schematic overview of the mono-lingual and cross-lingual VC

Dataset: Voice Conversion Challenge (VCC) 2016 is a parallel speech dataset consisting of speech samples recorded in US English accents in both male and female voices. Meanwhile, **VCC 2018**, **VCTK**, etc., are non-parallel speech datasets recorded in various English accents (including US English)

Dataset: VCC 2016, **VCC 2018**, **VCTK**, and **CMU ARCTIC** are mono-lingual datasets (recorded in US English accent). On the other hand, **VCC 2020** is a widely used cross-lingual dataset recorded in **English, Finnish, German, and Mandarin**.

[4] S. Dhar, N. D. Jana and S. Das, "An Adaptive-Learning-Based Generative Adversarial Network for One-to-One Voice Conversion," in IEEE Transactions on Artificial Intelligence, vol. 4, no. 1, pp. 92-106, Feb. 2023, doi: 10.1109/TAI.2022.3149858.

Overview of VC (continue)

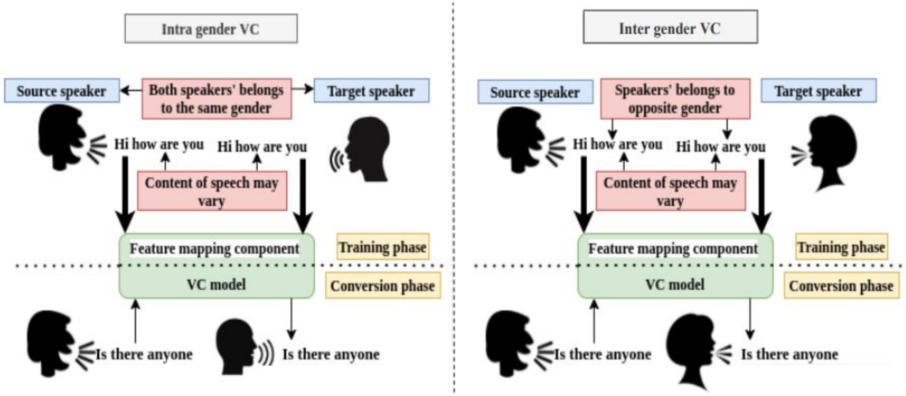


Figure 9: Schematic overview of the intra and inter-gender VC process

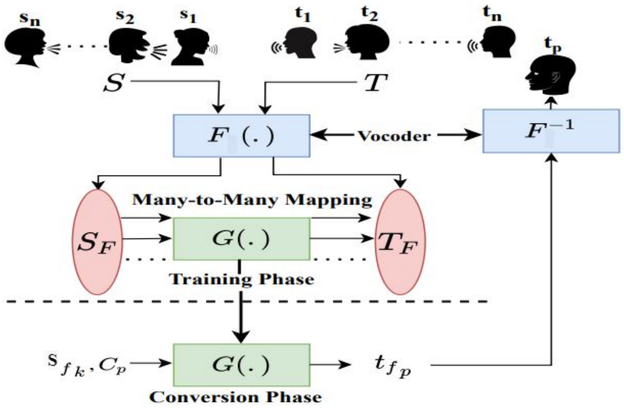


Figure 10: Basic framework of a typical many-to-many VC system

[5] S. Dhar, N. D. Jana and S. Das, "An Adaptive-Learning-Based Generative Adversarial Network for One-to-One Voice Conversion," in IEEE Transactions on Artificial Intelligence, vol. 4, no. 1, pp. 92-106, Feb. 2023, doi: 10.1109/TAI.2022.3149858.

Overview of VC (continue)

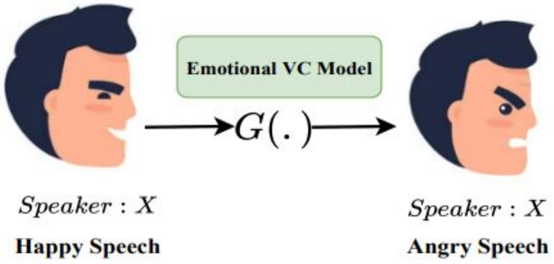


Figure 11: Overview of the emotional VC

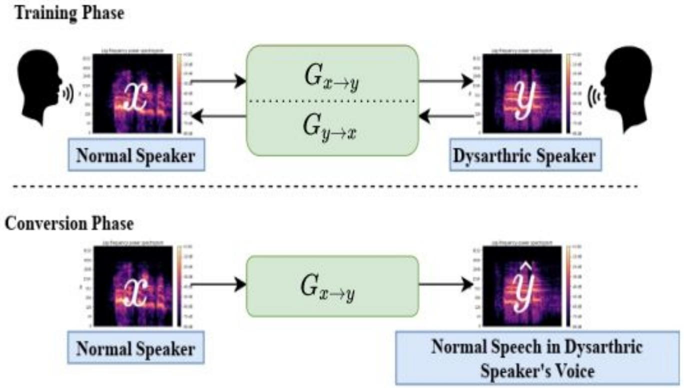


Figure 12: Schematic overview of the dysarthric to normal VC

Dataset: Most of the existing emotional VC datasets, such as the **emotional speech dataset (ESD)**, contain emotions such as **neutral, happy, angry, sad, surprised, etc**.

Dataset: EasyCall corpus dataset is a well known dysarthric speech dataset.

[6] S.Dhar, M. T. Akhter, P. Banerjee, N. D. Jana and S. Das, "FID-RPRGAN-VC: Fréchet Inception Distance Loss based Region-wise Position Normalized Relativistic GAN for Non-Parallel Voice Conversion," 2023 15th Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Taipei, Taiwan, 2023, pp. 350-356, doi: 10.1109/APSIPAASC58517.2023.10317438.

Overview of VC (continue)

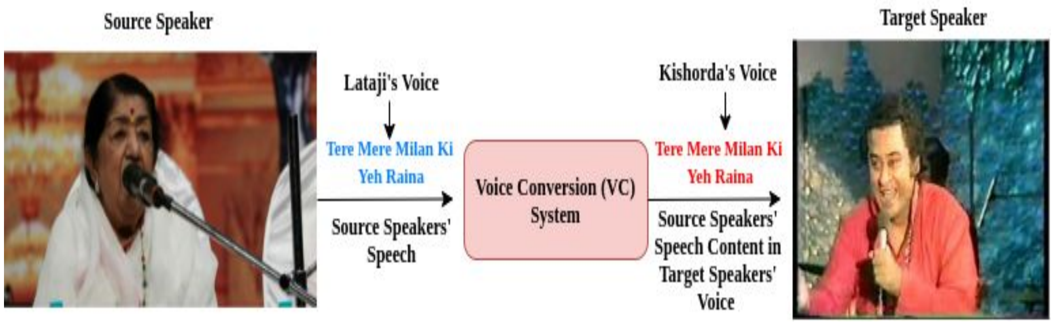


Figure 13: Singing voice conversion

Dataset: **Singing Voice Conversion Challenge (SVCC) 2023** Dataset, each speaker records 10 songs from a selection of 20 songs, making the dataset semi-parallel.

[7] State-of-the-art Singing Voice Conversion methods (Link: https://medium.com/qosmo-lab/state-of-the-art-singing-voice-conversion-methods-12f01b35405b).

Overview of VC (continue)

Structural Similarity Index (SSIM),
Spectral Loss,
Speech to Noise Ratio (SNR), etc.

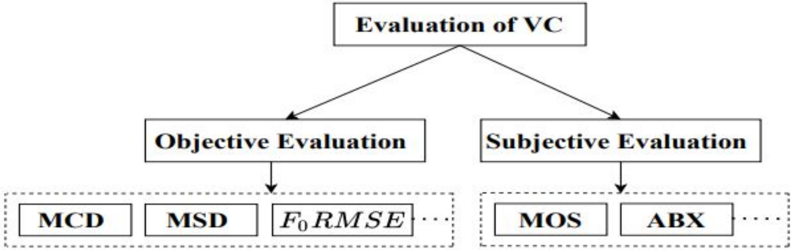


Figure 14: Evaluation metrics for VC model generated speech samples

Mel-Cepstral Distortion (MCD):

$$MCD[dB] = \frac{10}{\log\ 10}\sqrt{2\sum_{d=1}^{k}\left(mcc_d^t - mcc_d^{\hat{t}}\right)^2} \quad (3)$$

{It measures the global structural differences between the spectral features}

Modulation Spectra Distance (MSD):

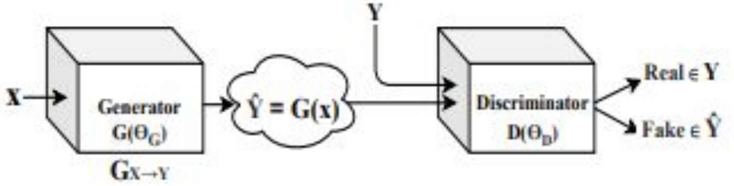
$$MSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(s(\mathbf{y})_i^t - s(\mathbf{y})_i^{\hat{t}}\right)^2} \quad (4)$$

{It measuring the local structural difference between the original and the converted speech samples in spectral domain}

$$\log\ F_0\ RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\log\ F_{0_i}^t - \log\ F_{0_i}^{\hat{t}}\right)^2} \quad (5)$$

[3] M. T. Akhter, P. Banerjee, S. Dhar and N. D. Jana, "An Analysis of Performance Evaluation Metrics for Voice Conversion Models," 2022 IEEE 19th India Council International Conference (INDICON), Kochi, India, 2022, pp. 1-6, doi: 10.1109/INDICON56171.2022.10040000.

Generative Adversarial Network (GAN) based VC



$$\mathcal{L}_{CEL}(\mathbf{y}, \hat{\mathbf{y}}) = \min_{G} \max_{D} \left[\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{y}}}(log D_{\mathbf{y}}) + \mathbb{E}_{\hat{\mathbf{y}}\sim p_{\hat{\mathbf{y}}}}(log(1 - D_{\hat{\mathbf{y}}}))\right] \quad (6)$$

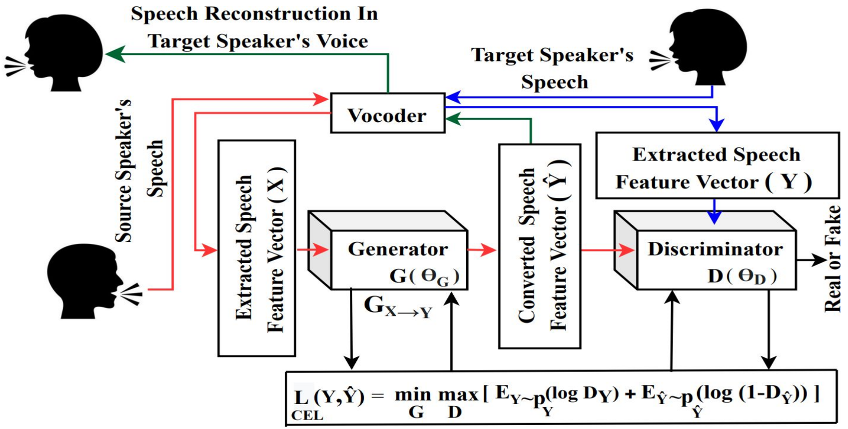


Figure 15: Basic framework of a typical GAN-based VC system

[8] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., & Bengio, Y. (2014). Generative Adversarial Nets. *Neural Information Processing Systems*.

GAN based Baseline VC models

Baseline Models (non-parallel VC):

- CycleGAN-VC (One-to-One) [EUSIPCO 2018]
  - ❑ CycleGAN-VC2 [ICASSP 2019]
  - ❑ CycleGAN-VC3 [Interspeech 2020]
  - ❑ MaskCycleGAN-VC [ICASSP 2021]

- StarGAN-VC (Many-to-Many) [Spoken Language Technology Workshop (SLT) 2018]
  - ❑ StarGAN-VC2 [Interspeech 2020]

[9] CycleGAN-VC: Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks https://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc/

GAN based Baseline VC models (continue)

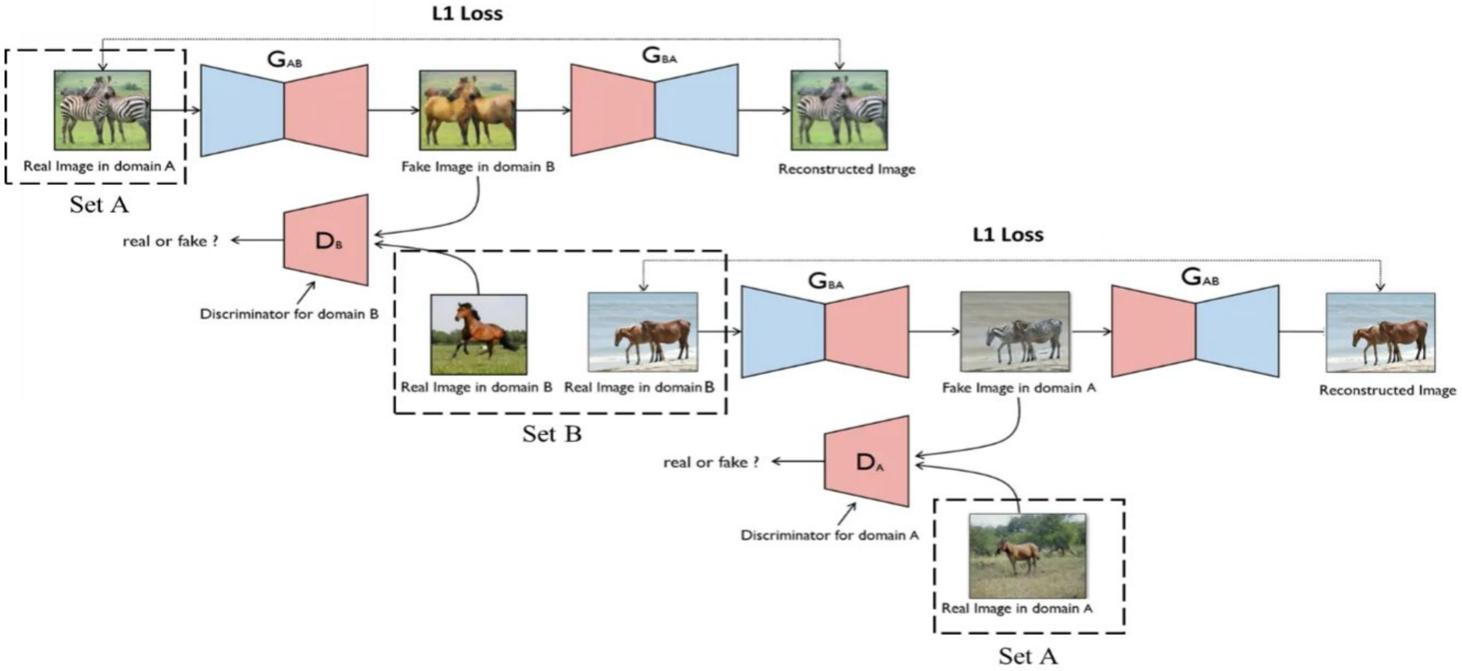


Figure 16: Schematic Overview of CycleGAN

[10] Zhu, Jun-Yan et al. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks." *2017 IEEE International Conference on Computer Vision (ICCV)* (2017): 2242-2251.

GAN based Baseline VC models (continue)

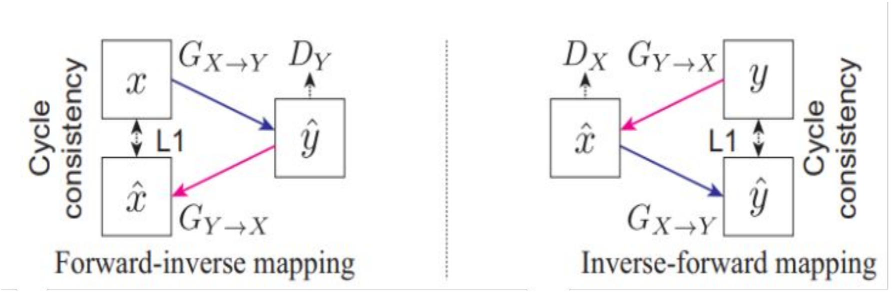


Figure 17: Schematic Overview of CycleGAN-VC Model

Adversarial loss:

$$\mathcal{L}_{adv}(G_{X\to Y}, D_Y) = \mathbb{E}_{y\sim P_{\mathrm{Data}}(y)}[\log D_Y(y)] + \mathbb{E}_{x\sim P_{\mathrm{Data}}(x)}[\log(1 - D_Y(G_{X\to Y}(x)))]. \quad (7)$$

Cycle-consistency loss:

$$\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X}) = \mathbb{E}_{x\sim P_{\mathrm{Data}}(x)}[\|G_{Y\to X}(G_{X\to Y}(x)) - x\|_1] + \mathbb{E}_{y\sim P_{\mathrm{Data}}(y)}[\|G_{X\to Y}(G_{Y\to X}(y)) - y\|_1]. \quad (8)$$

Identity-mapping loss:

$$\mathcal{L}_{id}(G_{X\to Y}, G_{Y\to X}) = \mathbb{E}_{y\sim P_{\mathrm{Data}}(y)}[\|G_{X\to Y}(y) - y\|_1] + \mathbb{E}_{x\sim P_{\mathrm{Data}}(x)}[\|G_{Y\to X}(x) - x\|_1], \quad (9)$$

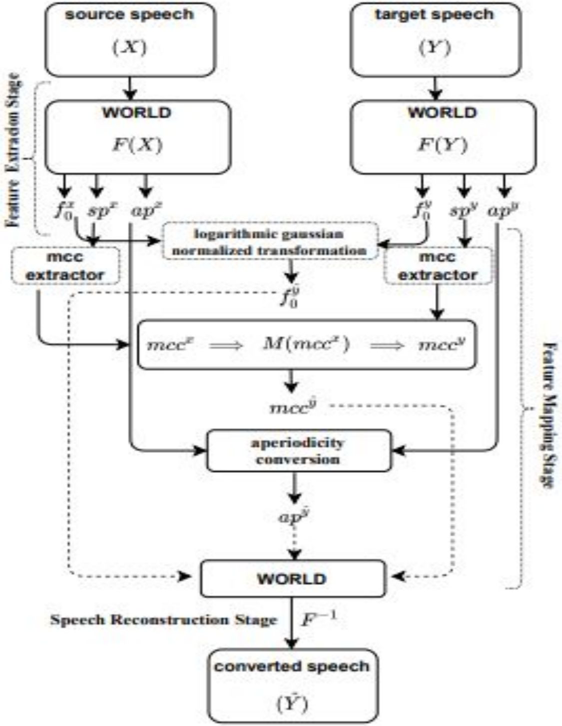


Figure 18: Working Mechanism of CycleGAN-VC Model

[11] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks," 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 2018, pp. 2100-2104, doi: 10.23919/EUSIPCO.2018.8553236.

GAN based Baseline VC models (continue)

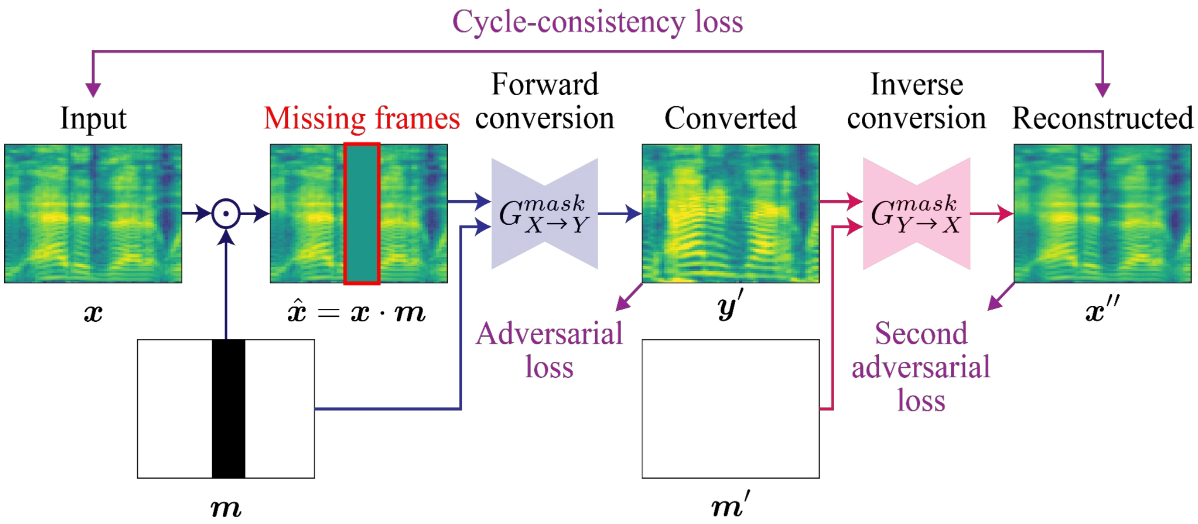


Figure 19: Working Mechanism of MaskCycleGAN-VC Model with MelGAN vocoder

[12] T. Kaneko, H. Kameoka, K. Tanaka and N. Hojo, "Maskcyclegan-VC: Learning Non-Parallel Voice Conversion with Filling in Frames," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 5919-5923, doi: 10.1109/ICASSP39728.2021.9414851.

Motivation

The motivation of the work:

- Prior research employed features such as MCCs, MS[13] etc., to compute feature-specific loss for training GAN-based VC models. Notably, the utilization of the Fréchet inception distance (FID) as a loss function in VC research has been relatively unexplored, primarily confined to the domain of image synthesis.
- In a standard GAN, the discriminator is typically represented as $D(x) = sigmoid(C(x))$. Conversely, the relativistic discriminator considers both real and fake data pairs $\hat{x} = (x_r, x_f)$, and it is defined as $D(\hat{x}) = sigmoid(C(x_r) - C(x_f))$. This approach assesses that real data is more authentic than randomly generated fake data and provides a scope to employ in GAN-based VC to explore its impact.

The contributions of the proposed work:

- Utilisation of a hybrid normalization technique named as region-wise positional normalization (RPN).
- Incorporation of Gaussian error gated linear unit (GEGLU) as an activation function.
- Inclusion of relativistic discriminator to trace the similarity between the latent representation of real and generated mel-spectrogram.
- Incorporation of FID metric as a loss function in GAN training.

[13] S. Dhar, M. T. Akhter, P. Banerjee, N. D. Jana and S. Das, "FID-RPRGAN-VC: Fréchet Inception Distance Loss based Region-wise Position Normalized Relativistic GAN for Non-Parallel Voice Conversion," 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Taipei, Taiwan, 2023, pp. 350-356, doi: 10.1109/APSIPAASC58517.2023.10317438.

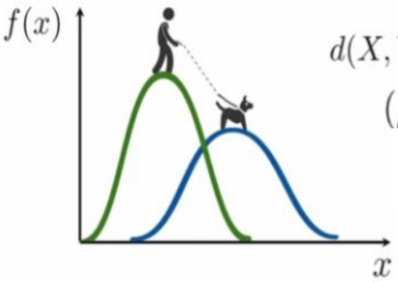
Proposed Model: FID-RPRGAN-VC

Title:- FID-RPRGAN-VC: Fréchet Inception Distance Loss based Region-wise Position Normalized Relativistic GAN for Non-Parallel Voice Conversion
(Accepted in APSIPA-23)

Authors: SANDIPAN DHAR , MD. TOUSIN AKHTER , PADMANABHA BANERJEE, NANDA DULAL JANA , SWAGATAM DAS

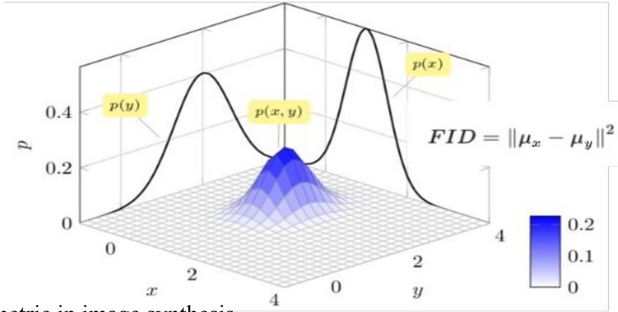
Feature embedding

Frechet Inception Distance(FID) between univariate normal distribution



$$d(X,Y) = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2$$

Frechet Inception Distance(FID) between multivariate normal distribution



$$FID = \|\mu_x - \mu_y\|^2 + T_r(\Sigma_x + \Sigma_y - 2(\Sigma_x\Sigma_y)^{1/2})$$

$\sum_X$ and $\sum_Y$ is the covariance matrix of the feature embeddings

Figure 20: Use of FID metric in image synthesis

[14] J. Lee and M. Lee, "FIDGAN: A Generative Adversarial Network with An Inception Distance," 2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Bali, Indonesia, 2023, pp. 397-400, doi: 10.1109/ICAIIC57133.2023.10066964.

Proposed Model: FID-RPRGAN-VC (continue)

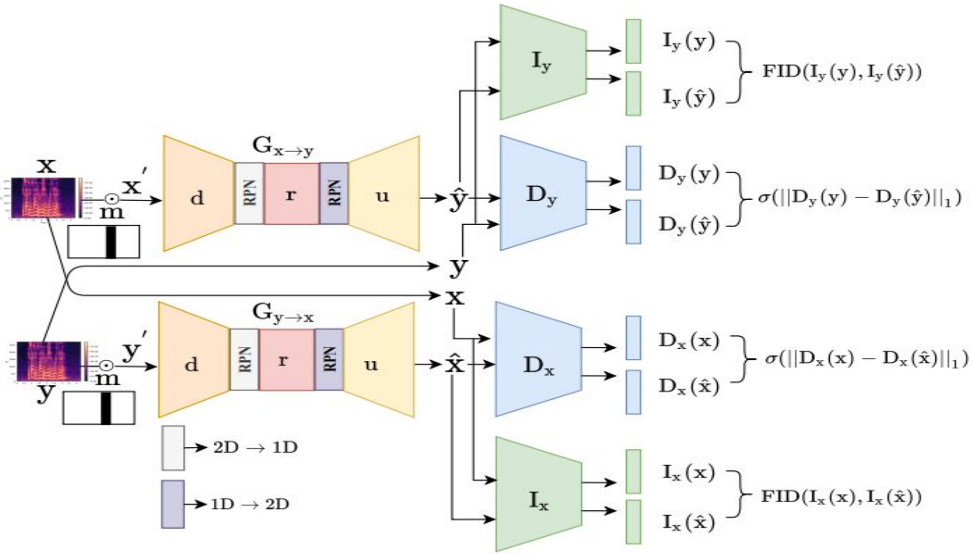


Figure 21: The schematic overview of the proposed FIDRPRGAN-VC model

Proposed Model: FID-RPRGAN-VC (continue)

1. Utilisation of a hybrid normalization technique named as region-wise positional normalization (RPN).
2. Incorporation of Gaussian error gated linear unit (GEGLU) as an activation function.

$$\mathrm{G}(.) \rightarrow \mathrm{u}(\mathrm{u}_{1\rightarrow 2}(\mathrm{r}(\mathrm{d}_{2\rightarrow 1}(\mathrm{d}(.))))). \tag{10}$$

The components of $G(.)$ are represented mathematically below,

$$\mathrm{d}(.) \rightarrow \mathrm{GEGLU}(\mathrm{IN}(\mathrm{Conv2D}(.))), \tag{11}$$

$$\mathrm{d}_{2\rightarrow 1}(.) \rightarrow \mathrm{RPN}(\mathrm{Conv1D}(.)), \tag{12}$$

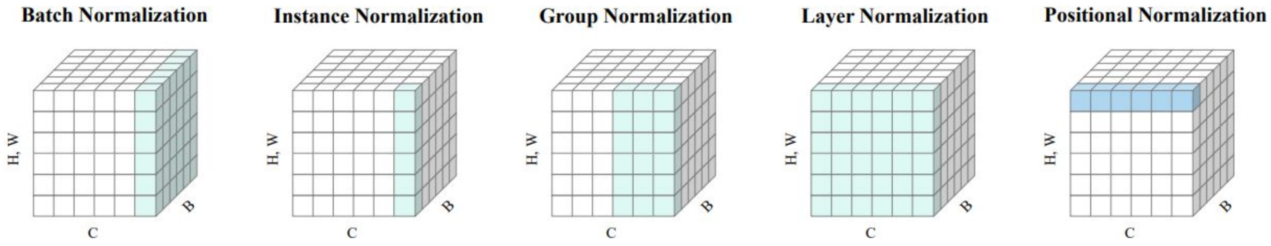
$$\mathrm{r}^{\mathrm{i}} \rightarrow \mathrm{r}^{\mathrm{i}-1} \oplus \mathrm{IN}(\mathrm{Conv1D}(\mathrm{GEGLU}(\mathrm{IN}(\mathrm{Conv1D}(\mathrm{r}^{\mathrm{i}-1}))))), \tag{13}$$

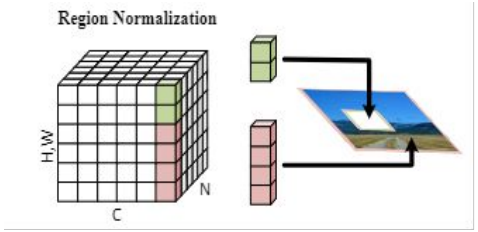
here, r^i indicates the i^{th} residual block.

$$\mathrm{u}_{1\rightarrow 2}(.) \rightarrow \mathrm{RPN}(\mathrm{Conv1D}(.)), \tag{14}$$

$$\mathrm{u}(.) \rightarrow \mathrm{GEGLU}(\mathrm{IN}(\mathrm{PS}(\mathrm{Conv2D}(.)))). \tag{15}$$

Proposed Model: FID-RPRGAN-VC (continue)





Normalization : Address the problem of covariate Shift

$$X'_{b,c,h,w} = \gamma \left(\frac{X_{b,c,h,w} - \mu}{\sigma} \right) + \beta \quad (16)$$

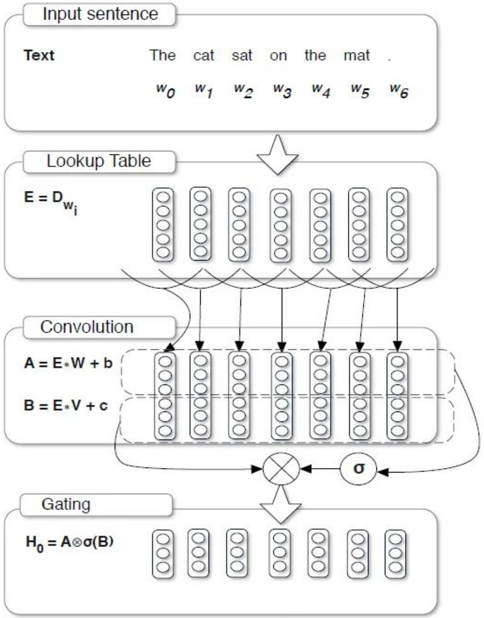
$$\mu_{b,h,w} = \frac{1}{C} \sum_{c=1}^{C} X_{b,c,h,w} \quad (17)$$

$$\sigma_{b,h,w} = \sqrt{\frac{1}{C} \sum_{c=1}^{C} (X_{b,c,h,w} - \mu_{b,h,w})^2 + \epsilon} \quad (18)$$

[15] Yu, Tao et al. "Region Normalization for Image Inpainting." *ArXiv* abs/1911.10375 (2019): n. Pag,
[16] Ulyanov, Dmitry et al. "Instance Normalization: The Missing Ingredient for Fast Stylization." *ArXiv* abs/1607.08022 (2016): n. pag..

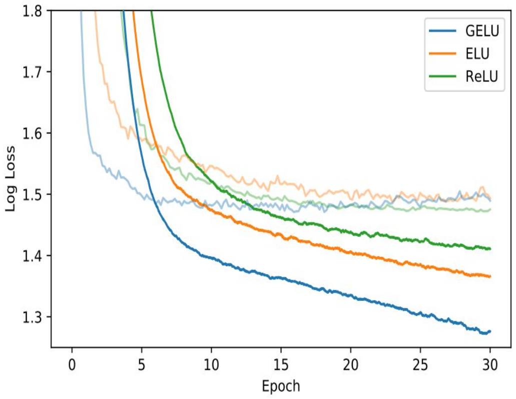
Proposed Model: FID-RPRGAN-VC (continue)

Gated Linear Unit (GLU):



Gaussian Error Gated Linear Unit (GEGLU):

$$\mathrm{GEGLU}(x', W_1, W_2, b_1, b_2) \rightarrow \mathrm{GELU}(x'W_1 + b_1) \otimes (x'W_2 + b_2),$$
$$\mathrm{GELU}(x'W_1 + b_1) \rightarrow (x'W_1 + b_1) \times \sigma(1.702 \times (x'W_1 + b_1)). \quad (19)$$



Φ(x) the standard **Gaussian cumulative distribution function**

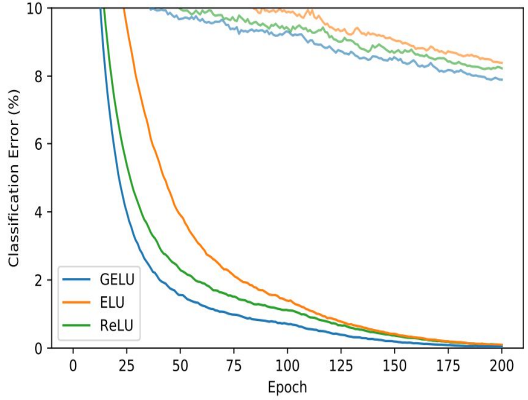
$\mathrm{GELU}(x) = xP(X \leq x) = x\Phi(x)$

We can approximate the GELU with

$$0.5x(1 + \tanh[\sqrt{2/\pi}(x + 0.044715x^3)])$$

or

$$x\sigma(1.702x),$$

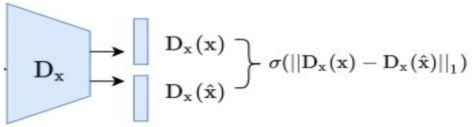


[17] Dauphin, Yann et al. "Language Modeling with Gated Convolutional Networks." *International Conference on Machine Learning* (2016)
[18] Hendrycks, Dan and Kevin Gimpel. "Gaussian Error Linear Units (GELUs)." *arXiv: Learning* (2016): n. pag.

Proposed Model: FID-RPRGAN-VC (continue)

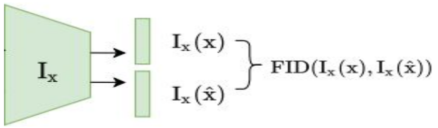
3. Inclusion of Relativistic Discriminator
4. Incorporation of FID metric as a loss function

The architectural framework of the relativistic discriminator D(.) (i.e. both D_X and D_Y) can be written as follows



$$\begin{aligned} &\mathrm{D}(.) \rightarrow \mathrm{Conv2D}(\mathrm{d_l}(\mathrm{GEGLU}(\mathrm{Conv2D}(.)))), \\ &\mathrm{d_l}(.) \rightarrow \mathrm{d_{l-1}}(.). \end{aligned} \tag{20}$$

$$\mathcal{L}_{\mathrm{disc}}^{\mathrm{relx}} = \sigma(\|\mathrm{D_X}(\mathrm{x}) - \mathrm{D_X}(\hat{\mathrm{x}})\|_1), \tag{21}$$



$$\begin{aligned} &\mathcal{L}_{\mathrm{fid}}^{\mathrm{X}} = \mathrm{FID}(\mathrm{I_x}(\mathrm{x}), \mathrm{I_x}(\hat{\mathrm{x}})), \\ &\mathrm{FID}(\mathrm{I_x}(\mathrm{x}), \mathrm{I_x}(\hat{\mathrm{x}})) = ||\mu_{\mathrm{I_x(x)}} - \mu_{\mathrm{I_x}(\hat{\mathrm{x}})}||^2 + \mathrm{T_r}(\Sigma_{\mathrm{I_x(x)}} \\ &+ \Sigma_{\mathrm{I_x}(\hat{\mathrm{x}})} - 2(\Sigma_{\mathrm{I_x(x)}}\Sigma_{\mathrm{I_x}(\hat{\mathrm{x}})})^{1/2}). \end{aligned} \tag{22}$$

[19] Jolicoeur-Martineau, Alexia. "The relativistic discriminator: a key element missing from standard GAN." *ArXiv* abs/1807.00734 (2018): n. pag.

Proposed Model: FID-RPRGAN-VC (continue)

Dataset:

The considered models are trained and tested on **VCC 2018** , and **CMU Arctic** dataset. For the **VCC 2018** dataset, the considered speakers are **VCC2TM1, VCC2SM3, VCC2TF1,** and **VCC2SF3**. Whereas, for the **CMU Arctic dataset**, the regarded speakers are **cmu-us-bld-Arctic, cmu-us-rms-Arctic, cmu-us-clb-Arctic, and cmu-us-slt-Arctic**. For both the datasets, **81** speech samples and **35** speech samples were considered for **training and testing**, respectively. For **CMU-Arctic dataset**, we have considered **116** (i.e. **81 for training and 35 for testing**) speech samples for each speakers such that the **utterances are disjoint**. **The particular setting is considered for making the training process non-parallel**.

Additionally, we evaluated the performance of the proposed model for **Easycall Dysarthric dataset**. The **Easycall dataset** contains parallel speech content for both normal and dysarthric speakers. Here, four speakers belonging to the male and female genders are considered. For each of the speakers, **264 samples were considered for training**, and **66 samples considered for testing**.

Proposed Model: FID-RPRGAN-VC (continue)

Training Details:

For training the **FID-RPRGAN-VC** model, **Adam optimizer** is used with learning rate **0.0001**. The proposed model is trained for **1000** epochs. The mini-batch size is considered as 1. In this work, pretrained **MelGAN vocoder** is used for mel-spectrogram to audible speech synthesis. The size of the **mel-spectrograms** considered here is **2×80×64**. The **mask size** here is taken as **25%** of the input size (along horizontal axis i.e. width).

The complete execution of the proposed **FID-RPRGAN-VC** model is carried out in the **Dell precision 7820 workstation** configured with **ubuntu 18.04 64 bit Operating System**, **Intel Xeon Gold 5215 2.5GHz processor**, **96GB RAM**, and **Nvidia 16GB Quadro RTX5000 graphics**. All the experiments of this work are implemented in **Python 3.6.9** using **Pytorch 1.1.2** and **Numpy 1.19.5**. The audible speech data are preprocessed by using **Librosa 0.9.1**.

Proposed Model: FID-RPRGAN-VC (continue)

Objective Evaluation:

Table 1: MCD, MSD and F_0 RMSE values for **VCC 2018** and **CMU-Arctic** dataset

Dataset	Models	M-M	F-F	M-F	F-M
	MCD				
VCC 2018	**FID-RPRGAN-VC**	**6.40**	**6.45**	**6.53**	**6.73**
	MaskCycleGAN-VC	7.45	6.85	6.76	7.84
CMU-Arctic	**FID-RPRGAN-VC**	**6.97**	**7.48**	**8.09**	**7.97**
	MaskCycleGAN-VC	7.12	7.81	8.20	8.07
	MSD				
VCC 2018	**FID-RPRGAN-VC**	**1.15**	**1.14**	**1.21**	**1.16**
	MaskCycleGAN-VC	1.17	1.18	1.50	1.24
CMU-Arctic	**FID-RPRGAN-VC**	**1.16**	**1.15**	**1.28**	**1.26**
	MaskCycleGAN-VC	1.18	1.19	1.32	1.29
	F_0 RMSE				
VCC 2018	**FID-RPRGAN-VC**	**18.17**	**27.22**	**32.20**	**36.78**
	MaskCycleGAN-VC	18.77	28.37	34.20	38.43
CMU-Arctic	**FID-RPRGAN-VC**	**15.10**	**21.81**	**31.71**	**31.93**
	MaskCycleGAN-VC	15.25	23.62	35.12	34.98

ABS(1) indicates the FID-RPRGAN-VC model **without GEGLU** (replaced by GLU of MaskCycleGAN-VC).

ABS(2) indicates the FID-RPRGAN-VC model **without RPN based generator** (replaced by TFAN of MaskCycleGAN-VC).

ABS(3) denotes the FID-RPRGAN-VC model **without relativistic discriminator** (replaced by MaskCycleGAN-VC discriminator).

ABS(4) denotes the FID-RPRGAN-VC model **without FID loss**.

Table 2: MCD, MSD and F_0 RMSE values for **ablation study**

Models	M-M	F-F	M-F	F-M
MCD				
FID-RPRGAN-VC	**6.40**	**6.45**	**6.53**	**6.73**
ABS(1)	6.42	6.47	6.57	6.76
ABS(2)	6.56	6.72	6.75	7.09
ABS(3)	6.46	6.54	6.60	6.98
ABS(4)	6.50	6.61	6.69	7.11
MSD				
FID-RPRGAN-VC	**1.15**	**1.14**	**1.21**	**1.16**
ABS(1)	1.16	1.16	1.23	1.17
ABS(2)	1.23	1.27	1.30	1.28
ABS(3)	1.20	1.16	1.24	1.23
ABS(4)	1.21	1.23	1.24	1.21
F_0 RMSE				
FID-RPRGAN-VC	**18.17**	**27.22**	**32.20**	**36.78**
ABS(1)	18.23	27.30	32.28	36.87
ABS(2)	18.54	27.65	32.82	38.38
ABS(3)	18.35	28.02	32.61	38.46
ABS(4)	18.41	27.97	32.70	38.47

Proposed Model: FID-RPRGAN-VC (continue)

Subjective Evaluation:

Table 3: **MOS** for naturalness with 95% confidence intervals

Dataset	Models	M-M	F-F	M-F	F-M
VCC 2018	**FID-RPRGAN-VC**	**3.9±0.29**	**3.8±0.18**	**3.1±0.35**	**3.4±0.51**
	MaskCycleGAN-VC	3.1±0.32	2.9±0.42	2.6±0.23	2.9±0.24
CMU-Arctic	**FID-RPRGAN-VC**	**3.7±0.36**	**3.8±0.18**	**3.5±0.23**	**3.8±0.26**
	MaskCycleGAN-VC	2.9±0.15	3.1±0.41	3±0.04	3±0.05

Some Generated Samples are available here:

https://drive.google.com/drive/folders/15FDrPz-w5Ri-0h8fqlEVsPv8pHVpefVQ

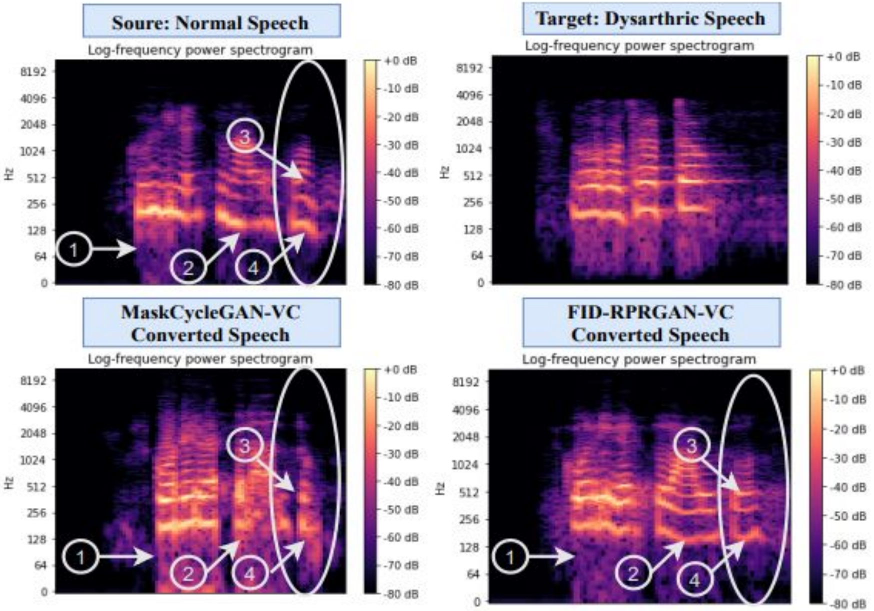


Figure 22: Visual comparison of mel-spectrograms for MaskCycleGAN-VC and FID-RPRGAN-VC converted speech

Conclusion

- In this work, we proposed an improved GAN model for mel-spectrogram based VC and referred to it as the FID-RPRGAN-VC model that consists of a region-wise positional normalized generator, a relativistic discriminator, and a FID loss function.
- These modifications aim to generate mel-spectrograms that capture the target distribution better than the SOTA MaskCycleGAN-VC model.
- The proposed model is tested on VCC 2018, CMU Arctic, and Easycall speech datasets. The objective and subjective evaluation of the FID-RPRGAN-VC generated samples indicates the superiority of the proposed model.
- In the future, the GAN-based VC model can also be investigated for speech enhancement purposes. Moreover, there is also a scope to explore the model for multi-lingual VC.

Some Recent Publications from NIT-Durgapur VC Group

Website: https://sites.google.com/view/nit-dgp-vc-group/home

- **S. Dhar,** N. D. Jana, S. Das**, "An Adaptive-Learning-Based Generative Adversarial Network for One-to-One Voice Conversion," in IEEE Transactions on Artificial Intelligence, vol. 4, no. 1, pp. 92-106, Feb. 2023, doi: 10.1109/TAI.2022.3149858. (Journal link: https://ieeexplore.ieee.org/abstract/document/9709124, Index: SCOPUS, Q1 journal) arXiv version: https://arxiv.org/abs/2104.12159).**

- **S. Dhar**, N. D. Jana and S. Das,**"GLGAN-VC: A Guided Loss based Generative Adversarial Network for Many-To-Many Voice Conversion"**, in **IEEE Transactions on Neural Networks and Learning Systems** (TNNLS), **2023, doi: 10.1109/TNNLS.2023.3335119.** (**Journal link**: https://ieeexplore.ieee.org/document/10339641)

- **S. Dhar**, M. T. Akhter, N. D. Jana and S. Das**."Collective Learning Mechanism based Optimal Transport Generative Adversarial Network for Non-parallel Voice Conversion", Submitted in IEEE Transactions on Neural Networks and Learning Systems (TNNLS), 2023**.

- M. T. Akhter, P. Banerjee, **S.Dhar**, S.Ghosh, N. D. Jana, **"Region Normalized Capsule Network Based Generative Adversarial Network for Non-Parallel Voice Conversion", 25th International Conference on Speech and Computer** Lecture Notes in Computer Science(), vol 14338. Springer, Cham. https://doi.org/10.1007/978-3-031-48309-7_20. **(SPECOM 2023, Dharwad, India), link: https://link.springer.com/chapter/10.1007/978-3-031-48309-7_20.**

- **S. Dhar**, P. Banerjee, N. D. Jana and S. Das**, "Voice Conversion Using Feature Specific Loss Function Based Self-Attentive Generative Adversarial Network," ICASSP 2023 - 2023 IEEE 48th International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10095069, link: https://ieeexplore.ieee.org/abstract/document/10095069.**

Thank You!