Collective Learning-based Optimal Transport GAN with Multi-Level Fine-Grained and Global Discriminators for Voice Conversion

Sandipan Dhar[1], Md. Tousin Akhter[2], Nanda Dulal Jana[3], Swagatam Das[4], Monorama Swain[5], Saurav Chowdhury[6]
[1]*Department of Electrical Engineering, Indian Institute of Technology Bombay, India.*
[2]*Department of Computer Science and Engineering, Indian Institute of Technology Bombay, India.*
[3]*Department of Computer Science and Engineering, National Institute of Technology Durgapur, India.*
[4]*Electronics and Communication Sciences Unit, Indian Statistical Institute, Kolkata, India.*
[5]*Institute of Computational Perception, Johannes Kepler University Linz, Austria.*
[6]*Department of Computer Science and Engineering, Indian Institute of Technology Jodhpur, India.*
Email: sandipandhartsk03@gmail.com, tousin@cse.iitb.ac.in, ndjana.cse@nitdgp.ac.in
swagatam.das@isical.ac.in, monorama.swain@jku.at, sauravchowdhury16.sc@gmail.com

***Abstract*—Generative Adversarial Networks (GANs) have achieved remarkable progress in speech synthesis by effectively modeling the distribution of target data through adversarial learning. However, existing State-of-the-Art (SOTA) GAN-based Voice Conversion (VC) models often exhibit a noticeable gap in the naturalness of generated speech compared to real speech, particularly in cross-gender VC scenarios. Since naturalness reflects an interplay of speaker-specific vocal style, prosody, and linguistic content modeling, capturing these attributes remains a critical challenge in GAN-based VC. To address this, we propose a novel GAN model, named CLOTMUL-VC, which incorporates a dual set of multi-level discriminators. The multi-level discriminators utilizes (i) a set of fine-grained discriminators to capture style, prosody, and content-specific features, (ii) a set of global discriminators to extract high-level feature information from the embedding space of real and generated mel-spectrograms. Unlike the pre-trained fine-grained discriminators, the global discriminators considered in our work are trained jointly with the generator. Therefore, to further improve the learning process of the single generator and multiple discriminator setting, we introduce a collective learning mechanism that utilizes Optimal Transport (OT) loss. The proposed mechanism enables the model to more accurately learn the target distribution by leveraging the principles of OT theory. Experimental evaluations on the VCC 2018, VCTK, and CMU-Arctic datasets demonstrate that the proposed model significantly outperforms existing VC systems in both objective and subjective tests.**

***Index Terms*—Voice Conversion, Collective Learning Mechanism, Optimal Transport Loss, Multi-Level Discriminators, Generative Adversarial Network.**

I. Introduction

Voice Conversion (VC) refers to the process of modifying a source speaker's voice to resemble that of a target speaker, while preserving the original linguistic content. Early approaches in VC research relied on techniques such as Gaussian Mixture Models (GMMs), Phonetic Posteriorgrams (PPGs), and other traditional models [1]. In recent years, the emergence of Generative Adversarial Networks (GANs) [2] due to their strong generative capabilities has made them a compelling alternative for VC tasks. In the context of one-to-one VC, models like ALGAN-VC [3] have shown notable improvements by leveraging adaptive learning mechanisms. More recent GAN-based VC variants such as FLSGAN-VC [4], FID-RPRGAN-VC [5], RNCapsGAN-VC [6], and GLGAN-VC [7] have significantly advanced non-parallel VC performance.

FLSGAN-VC and FID-RPRGAN-VC, in particular, generate target mel-spectrograms by integrating feature-specific evaluation metrics as loss functions into multi-discriminator frameworks, effectively addressing the over-smoothing problem. GLGAN-VC extends similar improvements to the many-to-many VC setting [7]. Additionally, diffusion-based models for voice conversion, such as DiffVC [8] and CycleDiffusion [9], have also demonstrated promising results in recent studies. Despite these advancements, a substantial gap in naturalness remains, especially in cross-gender VC where accurately modeling pitch remains a critical challenge due to inherent pitch differences between male and female voices. In addition to pitch, vocal style is equally important for achieving natural-sounding speech. Most importantly, preserving linguistic content remains a primary goal in VC. However, replicating all these aspects together becomes increasingly difficult when the feature distributions of the source and target speakers differ significantly, which is often the case in cross-gender scenarios. Therefore, there is a scope to incorporate components that effectively capture the target speaker's vocal style and pitch information through feature-specific loss optimization, as well as components that ensure content preservation w.r.t the source speech.

Moreover, it is evident that the majority of GAN-based VC models employ Deep Convolutional Neural Network (DCNN)-based discriminator architectures, which learn hierarchical feature representations from mel-spectrograms, ranging from low to high-level patterns [10]. However, in recent years, Vision

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

Transformer (ViT) [11] has shown significant performance for Speaker Identification (SI) tasks [12] because it obtains the information of local feature distribution more precisely from small patches of input data by utilizing the concept of attention mechanism. In this context, the conformer [13] models have also shown substantial performance improvement for SI and Speech Command Recognition (SCR) tasks. The discriminator in GANs operates much like a speaker verification (SV) model, as it evaluates whether a sample originates from the real target speaker or not, thereby aligning closely with the objectives of SV tasks.

This study proposes a novel GAN framework, termed as CLOTMUL-VC that incorporates a dual set of discriminators, referred to as multi-level discriminators. The multi-level discriminators comprised of (i) a group of Multiple Fine-Grained Discriminators (MFD), each responsible for capturing different aspects of speech, such as style (via a style encoder [14]), prosody (using the pre-trained JDC-Net [15]), and content (using the pre-trained whisper-large-v3 [16]), and (ii) a set of Multiple Global Discriminators (MGD), designed to extract high-level feature information from the embedding space of real and synthesized mel-spectrograms using architectures like DCNN, ViT, and conformer. This setup enables the extraction of fine-grained feature-specific information and global speaker-dependent/-independent abstract information from mel-spectrograms through their respective feature embeddings. We optimize the loss between generated and target embeddings for style and prosody to make the output more similar to the target speaker, while also optimizing the loss between the original source content and generated content embeddings, to preserve the original linguistic content of the source domain.

To improve the training process of MGD, we introduce a collective learning mechanism guided by an optimal transport (OT) loss [17], which leverages the proposed single-generator and multi-discriminator setup. Within MGD, the discriminators collaborate through a weighted learning process, utilizing the Sinkhorn algorithm in OT, which effectively measures distributional differences between source and target features. Importantly, this collective learning is applied only within the MGD. Since MFD captures distinct, non-overlapping features (style, prosody, content), using a weighted loss across them could obscure the relative importance of each feature. The overall design is inspired by multi-agent systems [18] and multi-player game frameworks [19], where multiple specialized agents work in collaboration toward a shared objective. We evaluate the effectiveness of our proposed model on the VCC 2018 [1], CSTR-VCTK [1], and CMU Arctic [20] speech datasets using both objective and subjective metrics. The results show that our proposed model produces speech with improved naturalness and speaker similarity compared to the considered State-of-the-Art (SOTA) VC models.

II. Proposed CLOTMUL-VC Model

This section briefly discusses the proposed CLOTMUL-VC model and its training mechanism. As shown in Fig. 1, $x \in X$ and $y \in Y$ are the mel-spectrograms of the source and target speakers, respectively, for the generator $G_{x \to y}$, and vice-versa. In Fig. 1, d implies the downsample block, r implies the residual block, and u implies the upsample block of the generator. The architectural framework of the generator of the CLOTMUL-VC model is developed keeping the framework of Mask-CycleGAN-VC [21] as the backbone. A multi-head attention (MHA) mechanism is utilized between d and r, and between r and u blocks to effectively model the flow of important features across the layers. This design is inspired by the use of self-attention (SA) mechanism in FLSGAN-VC [4]. The generated mel-spectrograms $\hat{x}$ and $\hat{y}$ are fed to each of the discriminators of the multi-level discriminator setting as shown in Fig. 1 (however, the ASR component uses the input and output of the same generator as its own inputs, i.e., x and $\hat{y}$, and y and $\hat{x}$). The multi-level discriminators in our framework are divided into two sets: MFD and MGD, with a total of six discriminators (3 in MFD and 3 in MGD).

Multiple Fine-Grained Discriminators (MFD): The MFD module comprises a style encoder $S(.)$ adopted from StyleTTS [14], pitch extractor JDCNet $J(.)$ [15], [22], and whisper-large-v3 ASR model $V3(.)$ [16]. These three components act as critics, guiding the generator by evaluating style, pitch, and content loss. We extract the final layer embeddings from $S(.)$ (shape $[1, 48]$) and $J(.)$ (shape $[1, 31, 2]$) for both generated and target mel-spectrograms, and compute Root Mean Square Error (RMSE) loss to derive style and pitch losses respectively. Content loss is calculated utilizing $V3(.)$ as the RMSE between the content embeddings (shape $[1, 1024]$) extracted from the original source and the generated mel-spectrograms produced by the same generator, as shown in Fig. 1. The ensemble of these three losses (style, pitch and content loss), i.e., $\mathcal{L}_{\mathcal{MFD}}$ loss helps the generator better mimic the target characteristics while preserving source content. While $J(.)$ and $V3(.)$ are frozen during training, $S(.)$ is updated via style loss. The complete MFD setup is illustrated in Fig. 1.

Multiple Global Discriminators (MGD): The three discriminators considered in MGD are DCNN (similar to the relativistic discriminator of FID-RPRGAN-VC [5]), ViT [11], and conformer [13] architectures. The ViT and conformer architectures considered in this framework are based on ViT-GAN [11] and CMGAN [13], respectively. The multiple discriminators incorporated in MGD, capture different aspects of feature information from the mel-spectrograms (such as deep convolutional features utilizing DCNN, patch-wise transformer encoded features utilizing ViT, and conformer extracted local-global features). Thus, the different feature embeddings extracted by each of the considered frameworks capture various nuances of the input mel-spectrogram, including both global and local feature distribution related information. Therefore, we introduce a collective learning mechanism to collaboratively learn the feature distribution of the mel-spectrograms by utilizing multiple discriminators (rather than relying on a single model).

Collective Learning Mechanism in MGD: The proposed

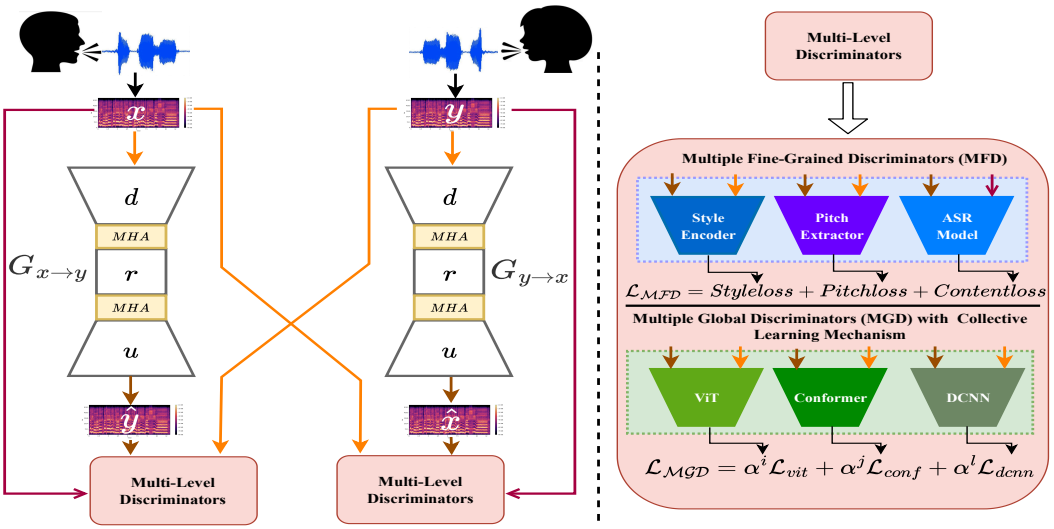

Fig. 1. Overview of the proposed CLOTMUL-VC model. The colored components highlight the modules introduced in the proposed framework.

collective learning mechanism is explained in Algorithm 1. As per Algorithm 1, the initial conditions include the input mel-spectrograms x and y (where x is the source and y is the target for $G_{x\to y}$, and vice versa), along with the corresponding learnable parameters $\boldsymbol{\theta}_d$ of the n discriminators (i.e., $\theta_{d_y}^{k=1\ldots n} \in D_y^{k=1\ldots n}$, and vice versa). Here n is 3, as collective learning mechanism is utilized for MGD training. Thereafter, the discriminator loss for each of the corresponding discriminators ($\mathcal{L}_{d_c^{k=1\ to\ n}} \in D_c^{k=1\ to\ n}$, here $c \in \{x, y\}$, and $D^{k=1\ to\ 3} \in \{ViT, Conformer, DCNN\}$) are obtained in each training epochs, as presented in Algorithm 1 and Fig. 1. The discriminator loss considered in this work is the OT loss function. After that, the total loss $\mathcal{L}_{\mathcal{MGD}_x}$ and $\mathcal{L}_{\mathcal{MGD}_y}$ are obtained by summing up all the corresponding discriminator losses. To find the participation or contribution of each of the discriminators (in terms of loss) in the total loss, the participation weights $\alpha_x^{k=1\ to\ n}$ and $\alpha_y^{k=1\ to\ n}$ are calculated for the respective discriminators (In Fig. 1, the index k is replaced with i, j, l, corresponding to the values 1, 2, and 3, respectively, as also described in Algorithm 1. Here, 1 corresponds to ViT, 2 to Conformer, and 3 to DCNN). As per the Algorithm 1, the highest participation weight is associated with the least discriminator OT loss (i.e., $\alpha \propto \frac{1}{\mathcal{L}_d}$). Afterward, as provided in Algorithm 1, each of the participation weights is multiplied to the corresponding discriminator losses for obtaining the final total discriminator loss $\mathcal{L}_{\mathcal{MGD}_x}$ and $\mathcal{L}_{\mathcal{MGD}_y}$ respectively.

OT Loss in MGD: The OT loss used in MGD, which is trained based on the collective learning mechanism, leverages Optimal Transport (OT) theory [23] to measure the divergence between two probability distributions by minimizing their optimal transport distance, denoted as $\mathcal{W}_c$. This distance is the trace of the product between the optimal transport cost matrix $C \in \mathbb{R}^{N\times N}$ and soft matchings matrix $M \in \mathbb{R}^{N\times N}$ (here, M is obtained using the sinkhorn algorithm, and N is considered as 4) as expressed in Eq. (1) and Eq. (2),

$$\mathcal{W}_c(X, Y) = \inf_{M\in\mathcal{M}} \mathrm{Tr}[MC^T], \tag{1}$$

$$C_{p,q} = c(\mathbf{x}_p, \mathbf{y}_q). \tag{2}$$

Algorithm 1: Collective Learning Mechanism

Input: $x \in X$, $y \in Y$
Output: Optimal parameters $\boldsymbol{\theta}_d^*$
Data: Discriminator OT loss: $\mathcal{L}_{d_c^k}$ ($c \in \{x, y\}$, $k \in \{1, 2, ..., n\}$), Optimizer: Opt, Iterators: $i = 0$ and $j = 0$, Learnable parameters: $\boldsymbol{\theta}_d \in \{\theta_{d_x}^{k=1}, \theta_{d_x}^{k=2}, ..., \theta_{d_x}^{k=n}, \theta_{d_y}^{k=1}, \theta_{d_y}^{k=2}, ..., \theta_{d_y}^{k=n}\}$
while $\boldsymbol{\theta}_d$ *not converged* **do**
  $\mathcal{L}_{d_{y_i}^{k=1}}, \mathcal{L}_{d_{y_i}^{k=2}}, ..., \mathcal{L}_{d_{y_i}^{k=n}} \leftarrow$ *CLOTMUL-VC*$(x, y, \theta_{d_{y_i}}^{k=1}, \theta_{d_{y_i}}^{k=2}, ..., \theta_{d_{y_i}}^{k=n})$;
  $\mathcal{L}_{\mathcal{MGD}_{y_i}} = \sum_{k=1}^{n} \mathcal{L}_{d_{y_i}^k}$;
  $\alpha_{y_i}^k = \frac{\mathcal{L}_{\mathcal{MGD}_{y_i}} - \mathcal{L}_{d_{y_i}^k}}{\mathcal{L}_{\mathcal{MGD}_{y_i}}}$;
  $\mathcal{L}_{y_i} = \sum_{k=1}^{n} \alpha_{y_i}^k \mathcal{L}_{d_{y_i}^k}$;
  $i = i + 1$;
  Update $\theta_{d_{y_i}}^{k=1}, \theta_{d_{y_i}}^{k=2}, ..., \theta_{d_{y_i}}^{k=n}$ based on $\mathcal{L}_{y_i}$ using Opt;
  $\mathcal{L}_{d_{x_j}^{k=1}}, \mathcal{L}_{d_{x_j}^{k=2}}, ..., \mathcal{L}_{d_{x_j}^{k=n}} \leftarrow$ *CLOTMUL-VC*$(y, x, \theta_{d_{x_j}}^{k=1}, \theta_{d_{x_j}}^{k=2}, ..., \theta_{d_{x_j}}^{k=n})$;
  $\mathcal{L}_{\mathcal{MGD}_{x_j}} = \sum_{k=1}^{n} \mathcal{L}_{d_{x_j}^k}$;
  $\alpha_{x_j}^k = \frac{\mathcal{L}_{\mathcal{MGD}_{x_j}} - \mathcal{L}_{d_{x_j}^k}}{\mathcal{L}_{\mathcal{MGD}_{x_j}}}$;
  $\mathcal{L}_{x_j} = \sum_{k=1}^{n} \alpha_{x_j}^k \mathcal{L}_{d_{x_j}^k}$;
  $j = j + 1$;
  Update $\theta_{d_{x_j}}^{k=1}, \theta_{d_{x_j}}^{k=2}, ..., \theta_{d_{x_j}}^{k=n}$ based on $\mathcal{L}_{x_j}$ using Opt;
return optimal parameters $\boldsymbol{\theta}_d^*$;

As depicted in Equation (2), C is the cost associated with transporting the p^{th} data vector $\mathbf{x}_p$ in mini-batch X to the q^{th} data vector $\mathbf{y}_q$ in mini-batch Y. This work defines the cost function c as the cosine distance [23]. The mathematical representation of the OT loss $\mathcal{L}_d$ (i.e., loss associated to each discriminator) is provided in Equation (3),

$$\mathcal{L}_d = \mathcal{W}_c(X, X^{'}) + \mathcal{W}_c(X, Y^{'}) + \mathcal{W}_c(X^{'}, Y) + \mathcal{W}_c(X^{'}, Y^{'}) - 2\mathcal{W}_c(X, X^{'}) - 2\mathcal{W}_c(Y, Y^{'}), \quad (3)$$

where X and $X^{'}$ represent mini-batches independently sampled from the distribution of real class, while Y and $Y^{'}$ denote independent mini-batches sampled from the distribution of generated class. To calculate the discriminator OT loss for each discriminator in MGD, the last layer flatten vector [24] of the respective discriminators are considered. These flattened vectors act as feature representations, preserving key speech characteristics in the latent space as feature embeddings, which vary based on the model architecture.

Generator loss: The generators learn the target distributions by minimizing the least squares loss $\mathcal{L}_g^{x \to y}$ and $\mathcal{L}_g^{y \to x}$ as similar to [5]. Additionally, it incorporates multi-level discriminator feedback through the losses $\mathcal{L}_{\mathcal{MFD}}$ and $\mathcal{L}_{\mathcal{MGD}}$, which improves the generator's ability to efficiently obtain the target distribution. Apart from that, we also use the cycle consistency and identity losses, which are kept similar to [5].

III. Experimental Design

A. Dataset Description and Training Details

In this work, the performance evaluation of each model is conducted on VCC 2018, VCTK, and CMU Arctic speech dataset. The CMU-Arctic dataset is also considered in non-parallel settings [25] considering disjoint utterances. The speakers considered for the VCC 2018 dataset are VCC2-TM1/SM3/TF1/SF3, for the VCTK dataset P-229F2/304M2/306F1/334M1, and for the CMU Arctic dataset cmu-us-bld/-rms/-clb/-slt-Arctic. The training, validation and evaluation (test) sets for each dataset comprised of 81, 35, 25 samples, respectively. The use of a limited set of utterances is intended to assess the models' performance in low-resource scenarios. For training the CLOTMUL-VC model, the standard *adam* optimizer [26] is used with a learning rate 1×10^{-4}. The CLOTMUL-VC model is trained for 500 epochs with mini-batch size 1 and uses input mel-spectrograms of size $2 \times 80 \times 64$. Speech reconstruction is performed using a pre-trained HiFi-GAN vocoder [27].

The proposed model is implemented on two NVIDIA A100 GPUs with 80GB memory. The training process takes approximately 5 GPU days. The framework is implemented in PyTorch 1.1.2 and NumPy 1.19.5. However, while using the pre-trained models switching between different Python environments required.

IV. Results and Discussion

A. Objective Evaluation

The objective evaluation Table presents a comprehensive comparison of six voice conversion (VC) models: proposed CLOTMUL-VC[1], DiffVC, FLSGAN-VC, FID-RPRGAN-VC, RN-CapsGAN-VC, and GLGAN-VC across three considered datasets using five key metrics: MCD [28], MSD [29], F0-RMSE [30], WER [16], and WV-MOS [31]. CLOTMUL-VC demonstrates consistently superior performance across nearly all evaluation criteria, particularly excelling in cross-gender conversion scenarios (M-F and F-M), which are inherently more challenging due to significant differences in prosodic and spectral features between male and female voices. For instance, on the VCC 2018 dataset, CLOTMUL-VC achieves the lowest MCD (7.15 for M-F, 7.10 for F-M) and F0-RMSE (17.6 for M-F, 16.8 for F-M), while also attaining the best WERs (14.6 for M-F and 15.0 for F-M), underscoring its ability to preserve linguistic content and pitch contours more effectively than the baselines. Similar trends are observed in CMU-Arctic and VCTK, where CLOTMUL-VC continues to deliver strong results across all gender pairings, consistently outperforming other models in MCD and WER, and achieving high WV-MOS scores, indicating better naturalness. These gains can be attributed to CLOTMUL-VC's novel architecture that integrates MFD and MGD. Since the MFDs are specifically designed to extract distinct feature embeddings related to style, prosody, and content using dedicated pretrained modules, they enable the model to better capture and preserve these individual aspects of speech, thereby significantly enhancing performance in each corresponding dimension. Our proposed model not only improves intelligibility but also strengthens the model's ability to produce high-quality, speaker-consistent voice conversions for both same-gender and cross-gender scenarios. Among the other models, DiffVC performs competitively in M-M and F-F settings, with relatively good WER and MOS scores, though it lags in F0-RMSE and MCD under cross-gender conditions. FID-RPRGAN-VC shows promise in capturing pitch details but suffers from higher WER and lower WV-MOS scores in M-F and F-M. Overall, Table I shows that the CLOTMUL-VC framework effectively addresses the key challenges of voice conversion.

We also conducted an ablation study on the VCC 2018 dataset to evaluate the contribution of three core components in the proposed CLOTMUL-VC model. In our study ablation 1 removes the MFD, ablation 2 eliminates the MGD, and ablation 3 removes the collective learning mechanism. The results show that ablation 1 leads to the most significant degradation in performance, reflected in higher MCD, MSD, F0-RMSE, and WER, and lower WV-MOS, highlighting the critical role of MFD in fine-grained supervision. Ablation 2 shows degradation drops primarily in terms of MCD and MSD, confirming the importance of global consistency enforced by MGD. Ablation 3 highlights the importance of the collective learning mechanism in maintaining high speech feature similarity, as its removal causes evident deterioration.

[1]The generated speech samples are available here https://drive.google.com/drive/folders/1dtQZO61gVtyW5obn5_RiQwj7Lig3zp-D?usp=sharing

TABLE I

EVALUATION OF VC MODELS ON VCC 2018, CMU-ARCTIC AND CSTR-VCTK DATASETS USING MCD (↓), MSD (↓) , F0-RMSE (↓), WER (↓), AND WV-MOS (↑) METRICS

Dataset	Model	M-M					F-F					M-F					F-M				
		MCD	MSD	F0-RMSE	WER	WV-MOS	MCD	MSD	F0-RMSE	WER	WV-MOS	MCD	MSD	F0-RMSE	WER	MOS	MCD	MSD	F0-RMSE	WER	WV-MOS
VCC 2018	CLOTMUL-VC	**6.45**	1.22	**15.9**	**12.3**	**3.70**	**6.38**	1.14	**15.8**	**11.7**	3.72	**7.15**	**1.39**	**17.6**	**14.6**	**3.60**	**7.10**	1.41	**16.8**	**15.0**	**3.67**
	DiffVC	6.81	**1.20**	19.2	15.4	3.55	7.02	**1.13**	23.5	14.3	**3.75**	7.55	1.48	19.2	16.2	3.45	7.50	**1.40**	19.9	16.5	3.65
	FLSGAN-VC	7.64	1.42	26.8	18.2	3.20	7.69	1.41	27.3	17.8	3.15	8.40	1.52	28.1	19.5	3.10	8.35	1.53	27.6	19.2	3.12
	FID-RPRGAN-VC	6.50	1.25	18.2	16.9	3.45	6.70	1.24	27.2	17.1	3.42	7.30	1.42	32.2	18.4	3.35	**7.10**	**1.40**	36.8	18.6	3.30
	RN-CapsGAN-VC	7.26	1.24	21.4	17.6	3.40	6.63	1.31	22.2	18.1	3.38	8.00	1.45	23.0	19.9	3.36	7.95	1.46	22.6	19.5	3.35
	GLGAN-VC	7.39	1.38	26.2	18.5	3.30	8.10	1.52	25.9	18.3	3.28	8.50	1.59	26.8	20.5	3.22	8.45	1.60	26.5	20.0	3.20
CMU-Arctic	CLOTMUL-VC	7.05	**1.16**	17.4	**12.7**	**3.65**	**7.41**	1.17	**16.2**	**13.2**	**3.80**	**7.75**	**1.30**	18.3	**14.9**	**3.60**	**7.75**	**1.31**	**18.0**	**15.1**	**3.55**
	DiffVC	7.18	1.19	23.6	13.9	3.50	7.69	**1.15**	24.4	14.5	3.60	8.10	1.35	**17.9**	15.8	3.45	8.00	1.36	21.2	16.4	3.42
	FLSGAN-VC	7.81	1.32	28.6	15.2	3.15	7.81	1.43	28.3	15.6	3.12	8.25	1.48	28.9	17.1	3.10	8.20	1.49	28.7	16.9	3.08
	FID-RPRGAN-VC	**6.97**	1.26	**15.1**	14.4	3.42	7.48	1.25	21.8	15.2	3.40	7.85	1.34	31.7	17.0	3.38	7.80	1.35	31.9	17.3	3.35
	RN-CapsGAN-VC	7.23	1.27	22.8	15.3	3.40	7.68	1.33	23.9	15.8	3.38	8.10	1.42	24.7	17.3	3.36	8.05	1.43	23.5	17.0	3.33
	GLGAN-VC	7.12	1.34	26.5	15.9	3.32	8.39	1.49	26.2	16.2	3.30	8.70	1.61	27.1	17.6	3.25	8.65	1.62	26.8	17.4	3.22
CSTR-VCTK	CLOTMUL-VC	**6.60**	**1.38**	**14.7**	**10.3**	**3.75**	**7.13**	1.63	**15.3**	**10.9**	**3.70**	**7.85**	**1.64**	**16.5**	**11.6**	3.60	**7.75**	**1.67**	**16.0**	**11.9**	**3.60**
	DiffVC	6.85	1.40	19.4	11.5	3.60	7.93	**1.62**	20.2	12.3	3.55	8.25	1.70	17.4	13.0	3.62	8.15	1.71	17.0	13.3	**3.60**
	FLSGAN-VC	7.12	1.66	24.3	13.0	3.25	9.42	1.63	24.0	13.4	3.22	9.55	1.82	25.0	14.2	3.18	9.48	1.83	24.7	14.5	3.16
	FID-RPRGAN-VC	6.90	1.52	18.0	11.7	3.48	7.58	1.60	17.6	12.2	3.45	8.10	1.70	19.1	13.6	**3.65**	8.05	1.71	18.9	13.8	3.40
	RN-CapsGAN-VC	7.14	1.45	20.1	11.8	3.50	8.02	1.58	20.6	12.5	3.45	8.45	1.65	21.7	14.1	3.43	8.40	1.70	21.3	14.3	3.40
	GLGAN-VC	6.98	1.50	23.5	12.9	3.40	8.10	1.64	23.2	13.3	3.38	8.55	1.74	24.4	14.4	3.35	8.50	1.75	24.1	14.6	3.32

TABLE II

ABLATION STUDY ON CLOTMUL-VC OVER THE VCC 2018 DATASET USING MCD (↓), MSD (↓), F0-RMSE (↓), WER (↓), AND WV-MOS (↑) METRICS

Model	Ablation	M-M					F-F					M-F					F-M				
		MCD	MSD	F0	WER	WV-MOS	MCD	MSD	F0	WER	WV-MOS	MCD	MSD	F0	WER	WV-MOS	MCD	MSD	F0	WER	WV-MOS
CLOTMUL-VC	Full Model	6.45	1.22	15.9	12.3	3.70	6.38	1.14	15.8	11.7	3.72	7.15	1.39	17.6	14.6	3.60	7.10	1.41	16.8	15.0	3.67
CLOTMUL-VC	Ablation 1	6.70	1.36	17.2	21.5	3.55	6.65	1.33	17.0	20.8	3.58	7.60	1.42	20.5	26.5	3.30	7.55	1.39	19.9	25.8	3.35
CLOTMUL-VC	Ablation 2	6.60	1.31	16.8	15.5	3.65	6.55	1.28	16.3	12.2	3.67	7.40	1.27	17.8	13.7	3.52	7.35	1.29	18.3	19.4	3.55
CLOTMUL-VC	Ablation 3	6.55	1.28	16.0	12.4	3.68	6.50	1.16	15.9	12.1	3.69	7.35	1.42	17.7	13.6	3.54	7.30	1.44	17.6	16.3	3.57

TABLE III

MOS (↑) WITH 95% CONFIDENCE INTERVALS

Dataset	Models	M-M	F-F	M-F	F-M
VCC 2018	Ground Truth	4.52±0.22	4.45±0.25	4.41±0.21	4.38±0.20
	CLOTMUL-VC	**3.88±0.41**	**3.74±0.45**	**3.41±0.28**	**3.52±0.33**
	DiffVC	3.64±0.38	3.49±0.31	3.35±0.26	3.31±0.30
	FLSGAN-VC	3.40±0.34	3.35±0.32	3.19±0.29	3.25±0.31
	FID-RPRGAN-VC	3.52±0.33	3.57±0.37	3.26±0.25	3.29±0.27
	RN-CapsGAN-VC	3.13±0.36	3.10±0.34	3.01±0.28	3.06±0.29
	GLGAN-VC	3.48±0.35	3.39±0.36	3.22±0.30	3.20±0.26
CMU-Arctic	Ground Truth	4.61±0.19	4.59±0.23	4.48±0.22	4.46±0.24
	CLOTMUL-VC	**3.75±0.37**	**3.92±0.41**	**3.53±0.33**	**3.58±0.30**
	DiffVC	3.61±0.35	3.47±0.39	3.45±0.30	3.38±0.27
	FLSGAN-VC	3.33±0.30	3.28±0.31	3.14±0.24	3.18±0.22
	FID-RPRGAN-VC	3.46±0.33	3.51±0.30	3.25±0.28	3.30±0.26
	RN-CapsGAN-VC	3.10±0.28	3.08±0.27	3.03±0.21	3.00±0.20
	GLGAN-VC	3.41±0.32	3.34±0.29	3.18±0.23	3.21±0.25
CSTR-VCTK	Ground Truth	4.58±0.24	4.53±0.27	4.50±0.23	4.42±0.26
	CLOTMUL-VC	**3.90±0.43**	**3.69±0.38**	**3.64±0.36**	**3.55±0.34**
	DiffVC	3.72±0.39	3.55±0.34	3.50±0.33	3.48±0.30
	FLSGAN-VC	3.35±0.29	3.29±0.32	3.24±0.27	3.20±0.28
	FID-RPRGAN-VC	3.50±0.36	3.53±0.35	3.42±0.30	3.40±0.33
	RN-CapsGAN-VC	3.12±0.31	3.06±0.33	3.07±0.25	3.05±0.27
	GLGAN-VC	3.43±0.34	3.38±0.31	3.28±0.26	3.30±0.29

B. Subjective Evaluation

In this work, a total of 17 volunteers participated in the subjective evaluation process. To obtain the MOS values, the volunteers rated randomly chosen speech samples for each VC category (intra-/inter-gender), following the standard MOS calculation method [1]. We provided the MOS values with 95% confidence intervals in Table III for VCC 2018, CMU-Arctic, and CSTR-VCTK datasets. From the Table III, it is evident that the proposed CLOTMUL-VC consistently achieves higher MOS scores across all datasets and gender combinations, outperforming other models. This improvement can be attributed to CLOTMUL-VC's use of optimal transport-based loss in the multi-discriminator setting, which more effectively captures the perceptual alignment between source and target domain.

V. CONCLUSION

This study proposes CLOTMUL-VC, a novel GAN-based VC framework that integrates a multi-discriminator setup and an OT-based collective learning mechanism. The system combines MFD and MGD discriminators to capture style, prosody, and content via specialized pretrained models, while extracting high-level features from mel-spectrograms using diverse architectures. The proposed design enables CLOTMUL-VC to better preserve linguistic content and effectively capture pitch and vocal style of the target speaker, especially in challenging cross-gender scenarios. Extensive evaluations across the datasets demonstrate that CLOTMUL-VC consistently outperforms existing models in objective and subjective tests. However, the inclusion of multiple deep learning components led to an increase in total training time. A promising direction for future work is to extend this approach to multilingual voice conversion while focusing on reducing the training time.

REFERENCES

[1] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM TASLP*, vol. 29, pp. 132–157, 2021. DOI: 10.1109/TASLP.2020.3038524.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," in *NeurIPS 27*, 2014, pp. 2672–2680.

[3] S. Dhar, N. D. Jana, and S. Das, "An adaptive-learning-based generative adversarial network for one-to-one voice conversion," *IEEE TAI*, vol. 4, no. 1, pp. 92–106, 2023. DOI: 10.1109/TAI.2022.3149858.

[4] S. Dhar, P. Banerjee, N. D. Jana, and S. Das, "Voice conversion using feature specific loss function based self-attentive generative adversarial network," in *Proc. 2023 ICASSP*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095069.

[5] 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," in *Proc. 2023 APSIPA ASC*, 2023, pp. 350–356. DOI: 10.1109/APSIPAASC58517.2023.10317438.

[6] M. T. Akhter, P. Banerjee, S. Dhar, S. Ghosh, and N. D. Jana, "Region normalized capsule network based generative adversarial network for non-parallel voice conversion," in *Speech and Computer*, Cham: Springer Nature Switzerland, 2023, pp. 233–244, ISBN: 978-3-031-48309-7.

[7] S. Dhar, N. D. Jana, and S. Das, "Glgan-vc: A guided loss-based generative adversarial network for many-to-many voice conversion," *IEEE TNNLS*, vol. 36, no. 1, pp. 1813–1826, 2025. DOI: 10.1109/TNNLS.2023.3335119.

[8] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," *ArXiv*, vol. abs/2109.13821, 2021.

[9] D. Yook, G. Han, H.-P. Chang, and I.-C. Yoo, "Cyclediffusion: Voice conversion using cycle-consistent diffusion models," *Applied Sciences*, vol. 14, no. 20, 2024, ISSN: 2076-3417. DOI: 10.3390/app14209595.

[10] R. B. Pittala, B. Tejopriya, and E. Pala, "Study of speech recognition using cnn," in *Proc. 2022 Second ICAIS*, 2022, pp. 150–155. DOI: 10.1109/ICAIS53314.2022.9743083.

[11] S. Hirose, N. Wada, J. Katto, and H. Sun, "ViT-GAN: Using vision transformer as discriminator with adaptive data augmentation," in *Proc. 2021 3rd ICCCI*, 2021, pp. 185–189. DOI: 10.1109/ICCCI51764.2021.9486805.

[12] S. Das, S. Dhar, and N. D. Jana, "Convolutional feature based vision transformer model for speech command recognition," in *2023 IEEE 20th INDICON*, 2023, pp. 228–232. DOI: 10.1109/INDICON59947.2023.10440809.

[13] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based metric gan for speech enhancement," in *Proc. Interspeech*, 2022.

[14] Y. A. Li, C. Han, and N. Mesgarani, "Styletts: A style-based generative model for natural and diverse text-to-speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 19, pp. 283–296, 2022.

[15] S. Kum and J. Nam, "Joint detection and classification of singing voice melody using convolutional recurrent neural networks," *Applied Sciences*, vol. 9, no. 7, 2019, ISSN: 2076-3417. DOI: 10.3390/app9071324.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, 2022. DOI: 10.48550/ARXIV.2212.04356.

[17] Q. Li, Z. Wang, G. Li, J. Pang, and G. Xu, "Hilbert sinkhorn divergence for optimal transport," in *Proc. 2021 IEEE/CVF CVPR*, 2021, pp. 3834–3843. DOI: 10.1109/CVPR46437.2021.00383.

[18] A. Ghosh, V. Kulharia, V. P. Namboodiri, P. H. Torr, and P. K. Dokania, "Multi-agent diverse generative adversarial networks," in *Proc. CVPR*, Jun. 2018.

[19] X. Zhang, P. Peng, Y. Zhou, H. Wang, and W. Li, "Evolutionary game-theoretical analysis for general multi-player asymmetric games," *ArXiv*, vol. abs/2206.11114, 2022.

[20] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Proc. Speech Synthesis Workshop*, 2004.

[21] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Maskcyclegan-VC: Learning non-parallel voice conversion with filling in frames," *Proc. 2021 IEEE ICASSP*, pp. 5919–5923, 2021.

[22] Y. A. Li, C. Han, V. S. Raghavan, G. Mischler, and N. Mesgarani, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," *NeurIPS*, vol. 36, pp. 19 594–19 621, 2023.

[23] T. Salimans, H. Zhang, A. Radford, and D. N. Metaxas, "Improving gans using optimal transport," *ArXiv*, vol. abs/1803.05573, 2018.

[24] E. Jeczmionek and P. A. Kowalski, "Flattening layer pruning in convolutional neural networks," *Symmetry*, vol. 13, no. 7, 2021, ISSN: 2073-8994.

[25] T. Kishida and T. Nakashika, "Non-parallel voice conversion based on free-energy minimization of speaker-conditional restricted boltzmann machine," in *Proc. 2022 APSIPA ASC*, 2022, pp. 251–255. DOI: 10.23919/APSIPAASC55919.2022.9980151.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[27] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *ArXiv*, vol. abs/2010.05646, 2020.

[28] M. Cotescu, T. Drugman, G. Huybrechts, J. Lorenzo-Trueba, and A. Moinet, "Voice conversion for whispered speech synthesis," *IEEE Signal Processing Letters*, vol. 27, pp. 186–190, 2020. DOI: 10.1109/LSP.2019.2961213.

[29] T.-h. Huang, J.-h. Lin, and H.-y. Lee, "How far are we from robust voice conversion: A survey," in *Proc. 2021 IEEE SLT Workshop*, 2021, pp. 514–521. DOI: 10.1109/SLT48900.2021.9383498.

[30] S. Dhar, N. D. Jana, and S. Das, "Generative adversarial network based voice conversion: Techniques, challenges, and recent advancements," *ArXiv*, vol. abs/2504.19197, 2025.

[31] P. Andreev, A. Alanov, O. Ivanov, and D. P. Vetrov, "Hifi++: A unified framework for bandwidth extension and speech enhancement," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2022.