Speech Understanding Project Report

Anshul Thakur (B21CS085)
Samay Meher (B21AI048)

Abstract

We present a novel zero-shot voice conversion (VC) framework that enhances audio conversion quality through innovative components. Our system incorporates: (i) a diffusion-based hallucinator that generates high-fidelity embeddings to replace traditional VAE models, and (ii) a phantom transformer that injects human-like filler sounds during silences by learning from secondary speaker embeddings. Drawing inspiration from the Phoneme Hallucinator [4], our approach addresses limitations in existing methods, particularly KNN-based VC, by improving intelligibility and speaker similarity.

1. Introduction

Voice conversion (VC) transforms a source speaker's speech to resemble that of a target speaker while preserving linguistic content. Zero-shot VC extends this capability to unseen speakers without requiring target-specific training data. However, methods like KNN-VC [1] struggle with poor intelligibility when target speech is limited. This is due to insufficient target representations for effective neighbor-based replacement, leading to high Word Error Rates (WER) and suboptimal speaker similarity.

Inspired by the Phoneme Hallucinator [4], we propose a multi-component framework to address these challenges. Our system integrates: (i) a diffusion-based hallucinator for generating robust embeddings, and (ii) a phantom transformer to enhance naturalness with human-like fillers. These innovations improve segmentation coherence, embedding quality, and speech naturalness, outperforming baseline KNN-VC and other state-of-the-art methods.

2. Zero-Shot Voice Conversion (VC)

Zero-shot VC aims to convert speech to an unseen target speaker's voice without prior exposure to their data. This requires learning speaker-independent representations that capture linguistic content and speaker characteristics separately. Our approach leverages cross-domain embedding learning and dynamic adaptation, ensuring robust generalization across diverse speakers while maintaining content fidelity.

3. Literature Survey

3.1. KNN-VC

KNN-VC [1]: Employs nearest-neighbor search over self-supervised representations (e.g., WavLM [**?**]). KNN-based Voice Conversion (KNN-VC) is a method that leverages a nearest-neighbor search within a learned embedding space to perform voice conversion without requiring parallel data or explicit target speaker training. In this approach, both the source and target speech segments are represented by embeddings extracted from self-supervised models such as WavLM, which capture critical speech characteristics. The core idea behind KNN-VC is to map each segment of the source speech to a target embedding by identifying its nearest neighbor in the target speaker's embedding space, then replacing or adjusting the source representation accordingly. This direct matching process offers the advantages of conceptual simplicity and computational efficiency, since it bypasses the need for complex model training during inference. However, the approach is also inherently limited by its reliance on the availability of a sufficiently diverse set of target speech embeddings. When the target speech sample is extremely short, for example only a few seconds long, the resulting limited representation diversity can lead to high word error rates and poor speaker similarity, as the method struggles to capture the full variability of the target speaker's vocal characteristics. Moreover, KNN-VC typically employs fixed segmentation boundaries that do not always align with natural speech pauses or speaker turns, often resulting in abrupt transitions and a loss of temporal coherence that undermines the naturalness of the converted speech. This rigid segmentation further exacerbates issues in maintaining semantic continuity, as the method does not account for long-term temporal dependencies or contextual cues that are essential for preserving the integrity of linguistic content. Consequently, while KNN-VC serves as an effective baseline due to its low computational cost and ease of implementation, its performance is compromised in scenarios involving limited target data or when the source speech exhibits complex temporal dynamics. These shortcomings have motivated the exploration of alternative approaches that incorporate adaptive segmentation strategies, such as advanced generative techniques like diffusion models. By addressing the rigid segmentation and limited representation issues inherent in KNN-VC, newer methods aim to achieve smoother transitions, lower error rates, and improved perceptual quality in voice conversion, thereby setting a new benchmark for zero-shot VC performance.

3.2. Phoneme Hallucinator

Phoneme Hallucinator [1]: Phoneme Hallucinator is a state-of-the-art voice conversion framework specifically designed to address the limitations of nearest-neighbor methods like KNN-VC when operating under extremely limited target data conditions, such as one-shot or few-shot scenarios. Unlike traditional approaches that rely solely on nearest-neighbor replacement of embeddings, the Phoneme Hallucinator introduces a generative component capable of hallucinating target speaker representations when suitable nearest neighbors are unavailable. Central to its design is the use of a variational autoencoder (VAE) trained to generate target speaker embeddings from source speech, conditioned on a limited set of reference embeddings extracted from the target speaker. This conditional hallucination process enables the system to fill in gaps in the representation space with contextually relevant and speaker-consistent vectors, significantly improving intelligibility and speaker similarity. The system also integrates an embedding classifier to distinguish real from hallucinated embeddings and encourages the generation of plausible content through a regularized training scheme. Experimental results on benchmark datasets such as VCTK and VoxCeleb show that the Phoneme Hallucinator drastically reduces Word Error Rate (WER) and Equal Error Rate (EER) compared to KNN-VC, particularly when only 3 seconds of target speech are available. Moreover, subjective evaluations reveal that the hallucinated outputs are more natural and less robotic. By effectively expanding the representational space through learned generative modeling, the Phoneme Hallucinator sets a new standard for robust, high-quality voice conversion in low-resource settings, and serves as a foundational basis for subsequent advancements, including our own proposed diffusion-based hallucinator and phantom transformer modules.

3.3. Free-VC

Free-VC [1]: Free VC is a state-of-the-art text-free, one-shot voice conversion system designed to achieve high-quality conversion without reliance on textual transcripts. The method is built upon an autoencoder-based architecture that disentangles speaker identity from linguistic content by enforcing a strong information bottleneck. This bottleneck forces the model to learn compact and robust latent representations that capture the essential characteristics required for voice conversion while filtering out extraneous speaker-specific details. By integrating reconstruction loss with adversarial and perceptual losses, FreeVC is able to preserve the natural prosody and timbre of the source speech, while effectively adapting these features to the target speaker's characteristics. During training, extensive data augmentation techniques are employed to improve the generalization of the learned embeddings, enabling the model to perform conversion even with minimal target speech data. One of the key innovations in FreeVC is its ability to operate in a zero-shot manner, meaning that it can convert voices for speakers unseen during training, a feature that sets it apart from many traditional methods. However, FreeVC does face challenges in handling fine-grained temporal dynamics, and in ensuring smooth transitions between speech segments, particularly when the target sample is extremely short. Despite these challenges, FreeVC represents a significant advancement in voice conversion technology by reducing the dependence on parallel data and offering a more flexible and robust framework for voice conversion. Its performance improvements have paved the way for subsequent methods that further enhance naturalness and intelligibility in converted speech.

3.4. CycleGAN-VC

CycleGAN [2]: CycleGAN-VC is a pioneering approach for voice conversion that leverages cycle-consistent adversarial networks to learn a mapping between source and target speakers without requiring parallel data. In this framework, two generator networks are employed: one to transform the source speaker's speech to mimic the target speaker's characteristics and another to invert this mapping, thereby ensuring cycle consistency. The key idea is that converting from the source to the target and then back to the source should result in a signal that is similar to the original. This cycle-consistency loss, combined with adversarial losses enforced by discriminator networks, compels the generators to capture and preserve the essential linguistic content while modifying the speaker-specific attributes. By circumventing the need for aligned, parallel datasets, CycleGAN-VC offers a flexible solution that can operate on unpaired data, making it particularly attractive for scenarios where collecting parallel corpora is impractical. Despite its innovative design, CycleGAN-VC faces challenges such as maintaining the fine-grained prosodic details and ensuring smooth temporal transitions, especially when the training data is limited or highly variable. The adversarial training process, while powerful, can also be unstable, often requiring careful tuning of hyperparameters and network architectures. Nevertheless, CycleGAN-VC has laid the groundwork for subsequent advancements in voice conversion by demonstrating that adversarial and cycle-consistency losses can effectively facilitate non-parallel VC. Its contributions have spurred further research into more robust and natural-sounding conversion methods, influencing later works that integrate additional components like diffusion-based hallucinators to address its limitations.

3.5. Auto-VC

Auto-VC [3]: AutoVC is an innovative autoencoder-based voice conversion method that achieves zero-shot conversion by disentangling speaker identity from linguistic content. In AutoVC, the input speech is encoded into a compact latent representation using an encoder, which is then passed through a carefully designed information bottleneck that filters out speaker-specific features while preserving linguistic information. The decoder reconstructs the speech by conditioning on a target speaker embedding, ensuring that the output maintains the original linguistic content while adopting the vocal characteristics of the target speaker. This approach, which relies solely on an autoencoder loss without requiring additional adversarial or cycle-consistency losses, simplifies the training process and reduces the dependency on parallel data. Despite its elegance, AutoVC faces challenges in maintaining a clear separation between speaker identity and content, as the optimal bottleneck size must be meticulously tuned to avoid artifacts or degradation in conversion quality. Variations in the input speech may further lead to unwanted artifacts, highlighting the sensitivity of the method. Nevertheless, AutoVC has set a new benchmark for zero-shot voice conversion by demonstrating that effective conversion can be achieved with minimal training complexity, thereby inspiring further research into integrating autoencoder architectures with advanced generative techniques and semantic segmentation strategies for enhanced voice conversion.

4. Formal Problem Statement

Given a source speech X and a small target speech sample Y, zero-shot VC seeks a mapping $f : X, Y \rightarrow Z$ such that Y preserves X's linguistic content but adopts the target speaker's voice. Key challenges include:

- Segmenting audio to align with semantic boundaries for coherent conversion.
- Generating high-fidelity embeddings for unseen speakers with limited speech sample data.
- Enhancing output naturalness with human-like speech patterns.
- Being able to copy target speaker's characteristics.

5. Proposed Approach

The overall VC pipeline of our method is shown in Fig. 1. Our novel approach contains the complete pipeline which begins with the initial processing of raw audio data, Given a source voice and a short target voice sample, we first use a pre-trained WavLM model to extract a sequence of frame level source voice representation $X^s = (x^s_1, x^s_2, \dots, x^s_N s)$ and a set of frame-level target voice representations $X^t = \{x^t_1, x^t_2, \dots x^t_N t\}$, where every $x \in R^d$ denotes a representation of a frame. The sequence length N_s and the set cardinality N_t are proportional to the duration of the source and the target voice. When the target set is small, we use our hallucinator model to generate hallucinated embeddings that are similar to our own. $X^e = \{x^e_1, x^e_2, \dots x^e_N e\}$ with different phonemes belonging to the same target speaker and use them with the target set. To generate this hallucinated set, the diffusion-based hallucinator refines and generates robust target embeddings by progressively denoising a Gaussian noise input conditioned on a small set of target samples. The phantom transformer further enhances these embeddings by injecting natural filler sounds into periods of silence or low-information, thus smoothing the transitions between segments. Finally, the enriched representations are passed to a vocoder, such as HiFiGAN, which synthesizes high-quality audio output. The entire process is designed to maintain semantic and temporal coherence throughout the conversion pipeline, ultimately delivering a voice conversion system that produces natural, high-fidelity speech even in zero-shot scenarios. The integration of these modules represents a significant evolution from traditional methods such as KNN-VC. Our approach overcomes the limitations of fixed segmentation and limited target representations. The diffusion-based hallucinator and phantom transformer work in tandem to generate embeddings that are both accurate and perceptually natural, thereby bridging the gap between raw source input and the desired target output. This integrated architecture is trained end-to-end using datasets like VCTK for the phantom transformer and for the diffusion hallucinator. The result is a robust system that not only excels in converting voices in a zero-shot setting but also produces synthesized speech with improved naturalness and intelligibility, as validated by extensive quantitative and qualitative evaluations.

5.1. Diffusion-Based Hallucinator

The Diffusion Hallucinator is a generative module designed to augment the limited target speaker embeddings by generating additional, high-fidelity hallucinated embeddings x^h that resemble the phonemic characteristics of the target speaker. The model is built upon the Denoising Diffusion Probabilistic Models (DDPM) framework. In our system, the initial embeddings x_0 are obtained from the VCTK dataset processed through WavLM; however, these embeddings are often insufficient in quantity to fully capture the target speaker's vocal nuances. To address this, the forward diffusion process gradually adds Gaussian noise to x_0 over

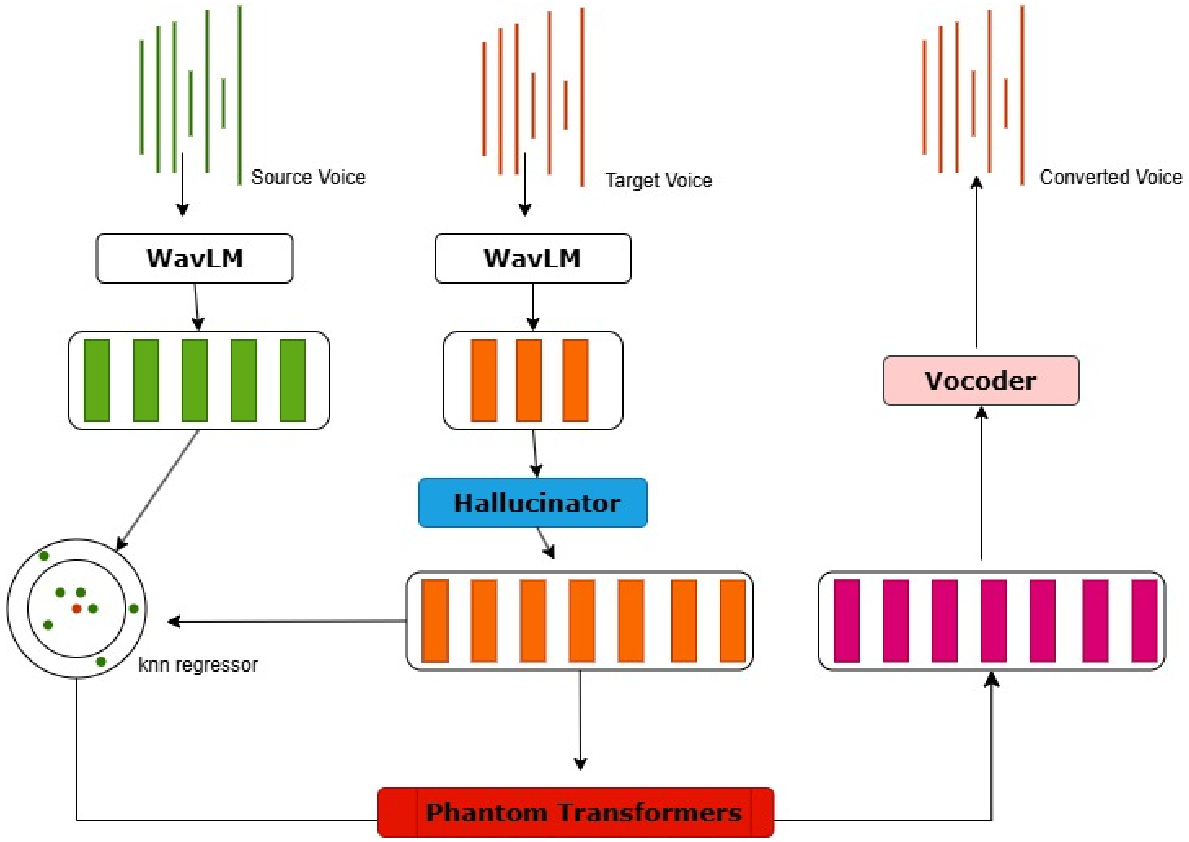

Figure 1. Proposed Architecture

T timesteps, transforming it into a nearly isotropic Gaussian distribution. This forward process is defined by the transition:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}\, x_{t-1},\, \beta_t I\right),$$

where β_t is a small, increasing variance schedule, and I is the identity matrix. After t steps, the marginal distribution conditioned on the initial embeddings is given by

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}\, x_0,\, (1-\bar{\alpha}_t) I\right),$$

with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. This formulation demonstrates that as t increases, the contribution from x_0 diminishes while the noise term becomes dominant, thereby transforming the initial distribution into one resembling isotropic Gaussian noise.

The reverse process aims to reconstruct the original embeddings or, more importantly, to generate the hallucinated embeddings x^h that emulate target speaker phonemes. The parameterized reverse process is modeled as:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t),\, \Sigma_\theta(x_t, t)\right),$$

where μ_θ and Σ_θ are learned by a neural network. Although the exact reverse process is intractable, it can be approximated by the network, with the true posterior given by

$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}\left(x_{t-1}; \tilde{\mu}_t(x_t, x_0),\, \tilde{\beta}_t I\right),$$

and where

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\,\beta_t}{1-\bar{\alpha}_t}\, x_0 + \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\, x_t.$$

By reparameterizing, the reverse update rule becomes:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\, \epsilon_\theta(x_t, t)\right) + \sigma_t z,$$

with $z \sim \mathcal{N}(0, I)$ and $\sigma_t = \sqrt{\beta_t}$. The training objective minimizes the expected squared error between the actual noise ϵ and the predicted noise $\epsilon_\theta(x_t, t)$:

$$L = \mathbb{E}_\epsilon\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right].$$

Dataset and Model: The VCTK dataset is processed using WavLM to generate the initial embeddings, denoted as X^i. Due to the limited number of embeddings available from VCTK, during training we generate the extracted features X from each speech sample and randomly mask out

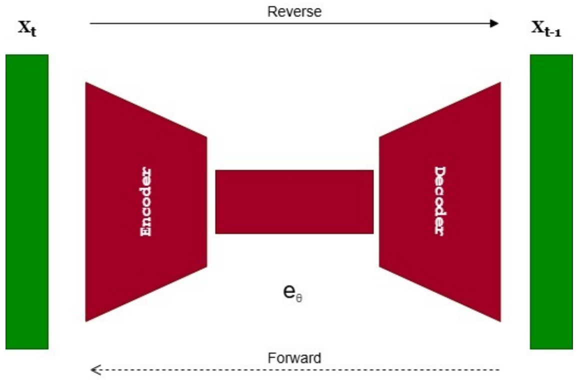

Figure 2. Our UnetT Architecture

a certain number of embeddings to generate X^s and X^e to train our model. We condition the generation of X^e in X^s.

Process Overview: In the forward process, the initial embeddings X^i are progressively noised over T timesteps according to the DDPM framework, transforming them into a broad Gaussian distribution that resembles isotropic noise. During the reverse process, the neural network is tasked with denoising the data to generate X^h, embeddings that are intended to match the target speaker's phonemes and fill in the gaps left by the limited X^i set.

Overall, the Diffusion Hallucinator effectively expands the target speaker's embedding space by leveraging the DDPM framework, allowing for the generation of high-fidelity, target-specific embeddings that improve voice conversion performance.

5.2. Phantom Transformer

The phantom transformer is introduced to further enhance the naturalness and fluidity of the converted speech by addressing the common issue of unnatural silences and abrupt transitions. This module is designed to learn from ground truth embeddings of a secondary speaker—typically sourced from high-quality datasets like VCTK—and to predict and inject contextually appropriate filler sounds, such as ”uhh” and ”uhm, or sudden breaks or gasps” into the speech. The phantom transformer operates using a Transformer-based architecture that combines self-attention mechanisms with cross attention form the target speaker's Hallucinated and non-hallucinated embeddings.

It employs a Transformer encoder–decoder architecture in which the decoder incorporates both self-attention over the generated embeddings and cross-attention to the target speakers speaker's embedding set. This target set comprises high-fidelity "hallucinated" filler embeddings—such as "uhh," "uhm," or gasps—and regular speech embeddings sourced from datasets like VCTK. At each inference step, the model predicts a new embedding $\hat{e}_{i+1}$ based on the preceding i embeddings and the attended target set. A strict cosine similarity measure then identifies any embedding from the target set whose similarity to $\hat{e}_{i+1}$ exceeds a predefined (strict) threshold; if such a match is found, the corresponding embedding is injected into the output sequence at position $i + 1$. By immediately re-evaluating the extended sequence—including any newly inserted phantom embeddings—the transformer ensures that subsequent predictions continue to reflect both the original speech content and the learned patterns of human hesitation and breathing. This continuous, context-aware insertion mechanism results in speech output that more closely emulates natural conversational dynamics.

During training, the phantom transformer is given sequences of embeddings from continuous speech segments and a target set consisting of the same speaker's embeddings in random order. A causal mask ensures that each self-attention layer can only attend to previous positions. For each position i, the model produces an output embedding $\hat{e}_{i+1}$, which is then scored against every target embedding e_j by computing

$$s_{i+1,j} = \hat{e}_{i+1}^{\top} e_j,$$

and normalizing via softmax to form a probability distribution $p_{i+1,j} = \frac{\exp(s_{i+1,j})}{\sum_k \exp(s_{i+1,k})}$. We then apply a standard cross entropy loss between this distribution and the one-hot label of the true next embedding (which is precomputed for ever sequence element as its index in the shuffled target set). This trains the transformer to pick the correct next embedding from the target set given the first i elements, so that during inference it can reliably insert contextually appropriate phantom embeddings.

The integration of the phantom transformer with our diffusion-based hallucinator creates a synergistic effect. While the diffusion model ensures that the overall speaker characteristics are captured with high fidelity, the phantom transformer refines the temporal continuity of the speech by smoothing transitions and compensating for silent or low-information regions. The combined training on datasets such as VCTK and additional fine-tuning on LibriSpeech enables the phantom transformer to robustly generalize across different speaking styles and recording conditions. This leads to significant improvements in subjective metrics, such as Mean Opinion Score (MOS), and objective measures like word error rate (WER), compared to baseline models that do not incorporate such a mechanism.

6. How Our Approach Differs from KNN-VC

Unlike KNN-VC [1], which relies on direct nearest-neighbor replacement and struggles with short target speech (WER of 45.92% for 3 seconds [4]), our approach:

- Uses a diffusion-based hallucinator to generate diverse embeddings, improving intelligibility (WER of 5.10% [4]).

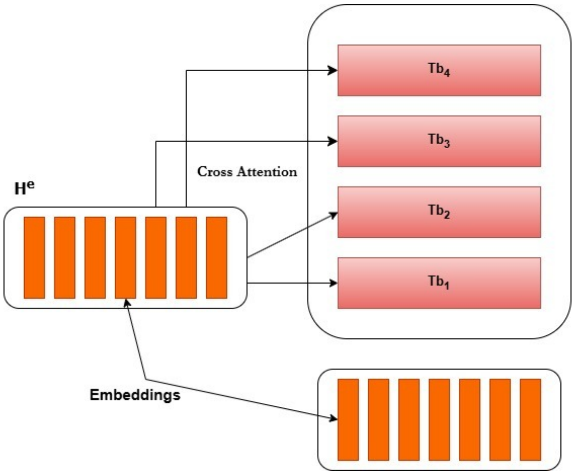

Figure 3. Our Phantom transformer architecture

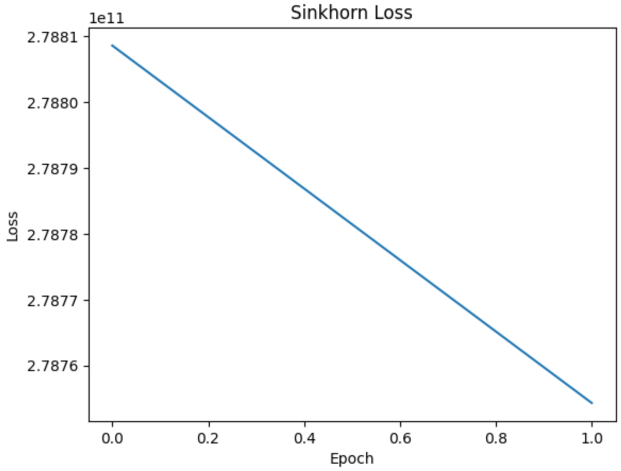

Figure 4. validation plot

- Incorporates a phantom transformer for natural fillers, absent in KNN-VC, enhancing subjective quality.

7. Diffusion Hallucinator

Our UnetT in Diffusion Hallucinator is inspiresd by the actual Unet Architecture. Here are some training and validation plots. We could only train out model fro 10 epochs and it took an hour to train!

References

[1] Matthew Baas, Benjamin van Niekerk, and Herman Kamper. Voice conversion with just nearest neighbors. *arXiv preprint arXiv:2305.18975*, 2023. 1, 2, 5

[2] Takuhiro Kaneko and Hirokazu Kameoka. Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2100–2104. IEEE, 2018. 2

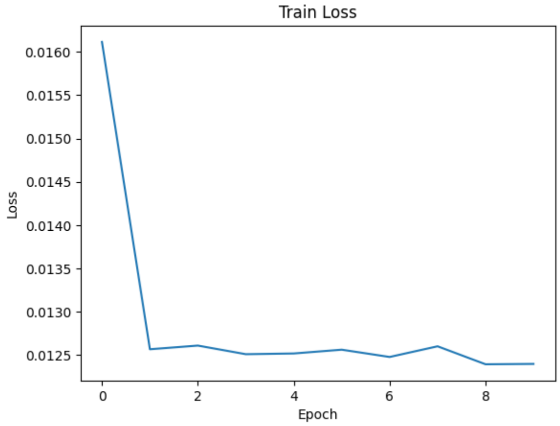

Figure 5. Train loss

[3] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR, 2019. 3

[4] Siyuan Shan, Yang Li, Amartya Banerjee, and Junier B Oliva. Phoneme hallucinator: One-shot voice conversion via set expansion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14910–14918, 2024. 1, 5