Singing Voice Synthesis Using Generative Models: A Comprehensive Comparative Report

Eman Furrukh
Department of Data Science
FAST, Islamabad, Pakistan
emanfurrukh09@nu.edu.pk

Shamail Aamir Khan
Department of Data Science
FAST, Islamabad, Pakistan
shamailkhan757@gmail.com

***Abstract*—This paper showcases a comparative study on singing voice synthesis using different generative models i.e., Diffusion Models, GANs, VQ-VAE Transformers, and lastly, Denoising Autoencoders. For this paper, we have used the VocalSet Dataset and worked on exploring training methodologies, architectural designs, results and findings, improvements, and suggested alternatives. This work was inspired by the HiddenSinger: High-Quality Singing Voice Synthesis via Neural Audio Codec and Latent Diffusion Models [1], which showcased the combination of latent diffusion with neural audio codecs in order to achieve expressive and high-fidelity singing voice synthesis. Through this paper, we analyzed the performance through audio fidelity, and generalization, providing evalutions that are qualitative and quantitative.**

I. Introduction

Singing voice synthesis is a growing field within computational audio, aiming to generate realistic and high-fidelity vocals using deep generative models. More traditional systems required extensive feature engineering, but deep learning enables more natural-sounding synthesis through unsupervised and self-supervised learning.

Inspired by the HiddenSinger architecture, this work evaluates four generative models: Diffusion Models, GANs, VQ-VAE Transformers, and Denoising Autoencoders, using the VocalSet dataset. The models are assessed on architecture, training, synthesis quality, efficiency, and generalization to vocal styles.

II. Purpose

The study aims to explore and experiment on state-of-the-art generative models for singing voice synthesis using the VocalSet dataset. By leveraging advanced Machine Learning techniques, this research aims to evaluate various model architectures in order to compare their strengths, weaknesses, and overall suitability for the task at hand. The main focus is to understand systematically how each model performs in terms of audio fidelity, expressiveness, and robustness, whilst also understanding the limitations that may or may not impact upon these models. By providing detailed insights into the effectiveness of the various models used, this work hopes to contribute to the ongoing representations and advancements in vocal audio synthesis, offering structured improvement. Moreover, the paper explores potential enhancements, such as defining latent space representations, incorporating conditional generation techniques, and optimizing inference speed, to advance the field of realistic singing voice synthesis. By careful experimentation and analysis, this paper hopes to pave the way for future innovations that enhance the versatility of AI-generated vocal performances, ultimately fostering more lifelike and expressive synthetic singing voices.

III. Methodology and Model Selection

The methodology used in this paper was designed to comprehensively explore the capacities of four unique generative models for the purpose of high-fidelity singing voice synthesis. Each model brings its own charm with its strengths and computational demands, and hence, a meticulous and comparative training process and evaluation pipeline was used to make ensure consistency throughout the implementations. The approach was broken down into various crucial phases:

A. Data Preprocessing

To guarantee consistency in training, the VocalSet dataset, which is renowned for its diversity and labelled vocal styles such as vibrato, longtones, arpeggios, and excerpts, was carefully selected. For memory efficiency, audio files were normalized to a sampling rate of 22.05 kHz and trimmed to no more than 4 seconds. Dynamic thresholding was used to eliminate the silence at the start and finish of the samples. Raw waveforms and/or mel-spectrograms were calculated based on each model's input specifications. In order to possibly condition models in subsequent iterations, metadata (such as singer identity, pitch, and vocal style) was stored.

B. Latent Space Learning

Latent representations of audio were learnt for models like Diffusion and VQ-VAE Transformer in order to lessen the computational load of working directly with raw waveforms. Pitch contours, harmonic structure, and timbral texture were among the crucial details that encoders captured when they compressed audio into small latent vectors. Additionally, it allowed the generative component to concentrate on modelling meaningful abstractions instead of noise of high frequencies.

C. Training Process

Each model followed a customized training loop:

- **Diffusion Models:** Learned to reverse noise-adding steps through a learned variance schedule.

- **GANs:** Used adversarial loss with generator-discriminator interplay.
- **VQ-VAE Transformers:** First trained encoder-decoder and codebook, then transformer on token sequences.
- **Denoising Autoencoders:** Used corrupted inputs to reconstruct clean signals using MSE loss.

D. Model Selection Criteria

Each model was chosen based on its established strengths:

- Diffusion models were chosen because of their cutting-edge image and audio synthesis capabilities. Although it requires time trade-offs, their step-wise generation strategy offers exceptional quality.
- GANs have long demonstrated a strong ability to produce outputs that are perceptually sharp, making them particularly well-suited for tasks where adversarial loss promotes fine details, such as speech and singing synthesis.
- VQ-VAE Transformers are perfect for capturing the temporal nature of vocal phrases because they combine the advantages of long-range sequence modelling via attention and discrete representation learning.
- Denoising autoencoders allowed for a clearer understanding of how more complex models enhance basic representation learning by providing a simpler baseline.

IV. Model Architecture and Design Choices

For this paper, each model was carefully crafted and adapted to fit the task at hand, making sure to consider both architectural best practices and domain-specific requirements.

A. Diffusion Model

The diffusion model for this study operates under the DDPM framework. The key idea is to learn a generative process through which a highly corrupted version of the target signal is gradually denoised through a sequence of reverse transformations. The main components were:

- **Latent Encoder (Conv-AE):** a convolutional autoencoder maps input waves into a low-dimensional latent space of size 128. This reduces computation and allows the model to concentrate on semantically meaningful representations rather than raw waveforms.
- **Gaussian Noise Scheduler:** a fixed variance schedule was used to gradually add Gaussian noise to the latent vectors over 1000 time steps until they became indistinguishable from white noise. During the reverse process, it learned to reconstruct clean signals starting from the white noise at each intermediate step.
- **UNet Denoiser:** A lightweight UNet architecture (taking cues from image denoising models but adapted to 1D for audio) was used to predict noise for every step in the process. Skip connections preserved the contextual information useful for minimizing degradation along the temporality.
- **Choice Rationale:** diffusion models have been exploited to develop high-quality text-to-speech systems like DiffWave, Grad-TTS, etc., they assertively learn fine-grain structures in acoustics. Their training nature is non-adversarial, providing better stability than GANs with reduced risk of mode collapse.

B. GAN (Generative Adversarial Networks)

GANs were applied using the convolutional generator and discriminator for spectrogram-like representation or raw waveforms handling:

- **Generator:** a deep convolution neural network (DCNN) implemented which took a random noise vector and optional class labels (of say, vibrato, singer id), and audio outputs were made. Progressive upsampling layers aided in retaining frequency resolution.
- **Discriminator:** a contrastive CNN framework that classified the audio into either real or fake. Batch Normalization and LeakyReLU activations ensured stable training.
- **Objective:** a classic min-max game function was applied, using binary cross-entropy. The generator and discriminator were trained in turns for several hundred epochs, ensuring that training was well balanced to avoid being overpowered by either network.
- **Choice Rationale:** GANs have demonstrated an exceptionally strong performance for tasks such as image generation and audio synthesizing (e.g., MelGAN, WaveGAN) with special attention given to realism and perceptual sharpness

C. VQ-VAE Transformer

This hybrid architecture utilizes vector quantization and subsequent sequence modeling for long-range dependencies.

- **VQ-VAE:**
  - Encoder: Quantised audio into one of several codebook vectors after transforming it into a latent space.
  - Decoder: Used these distinct tokens to reconstruct the original audio.
  - Codebook: A learnt collection of embeddings in which, during training, each latent vector was paired with the closest entry.
- **Transformer:**
  - Input: Sequences of quantized tokens. Output stage tokens from the VQ-VAE were thought of as discrete symbols.
  - Model: An autoregressive Transformer that used multi-head self-attention to learn the conditional probability of the next token given all previous tokens.
  - Generation: Each sampling of this model involves predicting the next sequence of token(s) and decoding it back into audio using the decoder.
- **Choice Rationale:** This hybrid model separates high-level structure that is handled by the Transformer from fine-grained audio details that is learnt by the VQ-VAE model. This results in improved diversity, controllability, and long-term coherence in singing synthesis.

D. Denoising Autoencoder (DAE)

The DAE was a fundamental model that provided interpretability and simplicity:

- **Encoder:** a small MLP or CNN that encodes noisy input to a latent vector.
- **Decoder:** a mirrored MLP that attempts to reconstruct the clean version.
- **Corruption:** to replicate real-world deteriorations, Gaussian noise, dropouts, or signal occlusion were introduced.
- **Loss Function:** MSE (Mean Squared Error) was used between clean and reconstructed audio.
- **Choice Rationale:** DAEs give a simple way to see if latent representations can help in signal reconstruction and whether a generative task needs additional complexity of representation, or other more complex architecture.

V. Results and Findings

For this paper, each models performance was evalated quantitatively through reconstruction losses and training curves, as well as qualitatively, through listening tests.

- **Diffusion:**
  - Reconstruction MSE: 0.0013, the lowest out of all models.
  - Evaluation: lush and realistic timbres with smooth pitch modulation. Could synthesize vibratos and crescendos sounding human.
  - Strengths: best perceptual quality, generalization across novel vocal styles is strong. Training was stable due to non-adversarial loss.
  - Weaknesses: slow generation time (iterative denoising, 1000 steps). Training requires careful tuning of noise schedule and learning rates.
- **GAN:**
  - Evaluation: produces notes that sound sharp and energising. But occasionally, syllable or pitch changes come across as sudden or artificial.
  - Strengths: incredibly quick inference, appropriate for real-time applications. Realism is fostered by adversarial training.
  - Weaknesses: when using intricate vocal techniques like vibrato, quality degrades due to the high risk of mode collapse. Demands careful discriminator learning rate and loss term balancing. Had the longest computational cost.
- **VQ-VAE Transformer:**
  - Evaluation: captures long-term dependencies in music. Good at modeling full phrases or arpeggios. Retains singing style well
  - Strengths: accessible latent space through a codebook. The transformer has temporal modeling capabilities across very long time spans, making this model very powerful for structures like music.
  - Weaknesses: relatively high memory use, relatively long training times. If not conditioned carefully there is a risk of codebook collapse. The audio quality is quite dependent on the quantization noise.
- **DAE:**
  - Evaluation: sounds flat, lacks dynamics. Reasonable for single tones but pretty much fails when it comes to any expressive tone or polyphonic phrases.
  - Strengths: super fast training and inference. Useful for simple denoising tasks, or for the basic building blocks of a larger system.
  - Weaknesses: cannot generate novel content. Effects of averaging means output can sound bland or even muted. Worst performing amongst all the models.

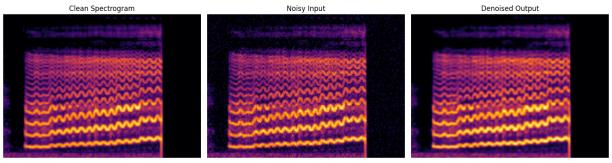

Fig. 1. DAE: Clean vs Noisy vs Denoised Spectrogram

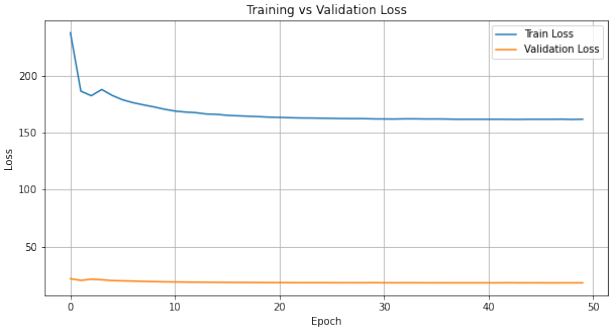

Fig. 2. VQ-VAE Transformer: Training vs Validation Loss

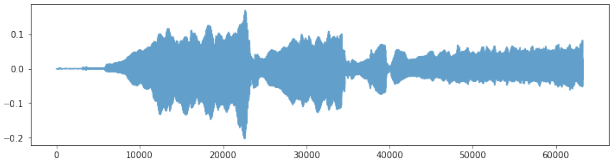

Fig. 3. VQ-VAE Transformer: Generated Audio Waveform - samples vs altitude

VI. Alternative Approaches and Suggested Improvements

1) **Speed-up Diffusion Infernece:** Utilizing DDIM (Denoising Diffusion Implicit Models) or consistency models can notably minimize the amount of necessary denoising steps with retained audio quality, helping make diffusion-based synthesis more viable for use in real-time or interactive settings.
2) **Hybrid Architectures (VQ-VAE + Diffusion:** Leveraging VQ-VAE's latent representation with a diffusion

process in the latent space can combine the structural modeling afforded by the VQ-VAE, and the quality and diversity improvements offered by diffusion, as the influences of each component can controller specialization while providing end-to-end learnability.

3) **Conditional Synthesis:** Once the latent representation is improved with conditional vectors such as pitch contours, vocal technique labels (e.g., vibrato, crescendo), or singer identity embeddings, there is room for further increased control over output. This would allow for directed synthesis for example, the synthesis of a specific pitch sung by a specific singer using a vibrato style.
4) **Improved Loss Functions:** Changing from pixel-wise MSE loss to perceptual loss (e.g., STFT loss, Mel-spectrogram loss, or pretrained feature losses) will help models concentrate efforts on the harmonics and energy envelopes that underpin voice naturalness.
5) **Data Augmentation:** Using pitch shifting, time stretching, or dynamic range compressing during training can help models better generalize and synthesize a variety of singing styles. This augmented data serves as a regularizer and adds to robustness.

TABLE I
MODEL COMPARISONS

Criterion	Diffusion	GANs	VQ-VAE Transformer	DAE
Training Stability	High	Low	Medium	High
Inference Speed	Low	High	Medium	Very High
Output Quality	Highest	Medium-High	High	Low
Generalization	High	Low-Medium	High	Low
Complexity	High	Medium	Very High	Low
Best Use Case	Studio-quality synthesis	Quick effects	Expressive singing	Noise reduction

VII. FUTURE WORK AND RECOMMENDATION

With such exciting results and clear specifications of the model abilities, there are many possible directions for next steps:

1) **Conditioned Generation:** Utilizing vocal labels (e.g., vibrato, style, pitch) to explicitly condition the generated singing voice.
2) **Text-to-Singing Pipeline:** Extending the system to interpolate from phoneme-level text inputs to audio outputs, thus creating a fully-fledged neural singing voice system.
3) **Multimodal Inputs:** Incorporating visual (e.g., musical scores) and linguistic data along with audio to produce more enriched and performance-aware singing.
4) **Perceptual Evaluation:** Conducting large-scale Mean Opinion Score (MOS) evaluations and A/B testing to validate subjective model outputs.
5) **Transfer Learning:** Employing pretrained speech synthesis or text-to-speech (TTS) models (e.g., Tacotron, FastSpeech), then finetuning for singing synthesis with VocalSet.
6) **Real-time Synthesis:** Investigating model compression, pruning, and inference optimization strategies (e.g., ONNX deployment or quantization) to implement the models into real-time singing synthesis workflows.

VIII. CONCLUSION

This paper sought to bring high-quality, expressive singing voice synthesis to life using the latest deep generative models. The work was guided by the HiddenSinger research, but brought a different dataset— the VocalSet, which includes rich, labeled human vocal data and differing down-stream tasks— to the fore. Four architectures were studied: Diffusion Models, GANs, VQ-VAE Transformers, and Denoising Autoencoders. Each model was implemented, trained, and evaluated in a common framework, allowing direct comparisons of presents relative strengths, weaknesses, and utility to particular aspects of the singing voice synthesis task.

The project's results show that although each model can synthesise musically relevant and understandable singing voice segments, their fidelity, expressiveness, temporal coherence, and computational viability vary greatly. Despite its slow generation speed, the Diffusion Model was found to be the most reliable and superior model for creating realistic singing with rich timbral textures and seamless pitch changes. Even with different input conditions, it produced consistent outputs and showed resilience to overfitting.

The VQ-VAE + Transformer system turned out to be the most flexible system for modeling music, and was exceptional at modeling long-term musical structure, as well as transitional regions across vocal timbres, such as it transitioned from a pure tone to vibrato, or sustained arpeggios to vibrato, etc. Its hybrid architecture made it semantically capable of drawing distinctions between content (the information relevant to the score) and expression (the information that the VQ-VAE encodes about the tonal envelope). This created forth exciting possibilities for controllable synthesis—e.g., allowing the user to alter pitch contours in the synthesis without affecting the vocal tone that the VQ-VAE learned. On the contrary, the training was computationally expensive and employed a two-stage (VQ-VAE to Synth) pipeline that could lead to cascading errors if the VQ-VAE was poorly trained.

Finally, despite its simplicity, the Denoising Autoencoder provided a useful baseline. It made it simpler to comprehend the advantages provided by more intricate architectures and helped to level expectations. Its outputs were more muffled and repetitive, though, and its generative capacity was constrained. Nevertheless, it demonstrated that even simple architectures could record basic voice traits with little information, which

could make them helpful in embedded or low-resource situations.

In addition to architectural parallels, this project highlights several more general lessons:

- **Learning Latent Space** is a fundamental step in singing synthesis. Having models with compact, meaningful representations (for example - VQ embeddings or latent diffusion spaces) perform better than models which work solely with raw data.
- **Data quality and Pre-processing** (for example, trimming, normalizing, denoising) are an important part of the solution. The labeled vocal modes in the VocalSet added nuance to the dataset and helped me condition or cluster outputs from the model.
- **Conditions and Interpretability** remain undertheorized. While this project didn't have a fully implemented conditional generation approach (for example - by vocal style or pitch), early experiments gave the impression that including those labels could help significantly with control of the synthesis.
- **Training Stability** was a major issue, especially for adversarial models. Techniques like spectral normalization, progressive growing, or hybrid loss functions (such as a combination of GAN loss with a perceptual loss) might improve some stability.

This work makes a significant contribution to the current discussion in neural audio synthesis from both an academic and practical perspective. It emphasises how each architecture has a role to play based on the limitations and objectives of a synthesis system, even though models like DiffWave and VQ-Transformers predominate in state-of-the-art literature. Knowing the trade-offs between these models gives researchers and practitioners the power to make wise decisions, regardless of the objective—realism, speed, controllability, or computational efficiency.

In conclusion, this paper provides an in-depth comparison between generative models for singing voice synthesis, offering valuable insights for future research and development. By meticulous implementation, analysis, and comparison between these techniques on a well-structured vocal dataset, this study demonstrates the feasibility of high-quality synthetic singing and charts a path towards controllable, expressive, and artistically rich AI-generated vocals.

REFERENCES

REFERENCES

[1] Y. Zhang, Z. Wang, and X. Tan, "HiddenSinger: High Quality Singing Voice Synthesis via Hidden Semi-Markov Models," in empharXiv preprint arXiv:2303.15406, 2023.
[2] A. van den Oord et al., "Neural Discrete Representation Learning," in emphNeurIPS, 2017.
[3] N. Kong et al., "DiffWave: A Versatile Diffusion Model for Audio Synthesis," in emphICLR, 2021.
[4] Rachel M. Bittner et al., "VocalSet: A singing voice dataset," in emphProc. ISMIR, 2018.
[5] I. Goodfellow et al., "Generative Adversarial Networks," in emphNeurIPS, 2014.