Voice Replication through Deep Learning : CS 7643

Aashna Doshi
adoshi45@gatech.edu

Akshat Karwa
akshatkarwa@gatech.edu

Shreya Santhanagopalan
ssanthan3@gatech.edu

Anvitha Veeragandham
aveeragandham3@gatech.edu

Abstract

Voice cloning is a relatively new field of research in artificial intelligence, but offers a wide range of uses that can transform communication. It offers a way for individuals with speech impairment or speech loss to regain their voices, and could even allow us to preserve the voice of an individual for future generations. The benefits of the technology are vast and promising, and we must continue to improve research to get the best cloning that we can. Our project focuses on the development of single-speaker and multi-speaker voice cloning systems. We employed pre-trained multilingual models from Coqui-AI, and built upon VITS architecture for end-to-end text-to-speech synthesis. The primary goal was to evaluate model effectiveness in replicating a diverse range of voices within the VCTK dataset. We aim to analyze performance across various voices, measuring the generated speech, giving insight into the advancements of the field of voice cloning.

1. Introduction/Background/Motivation

We set out to evaluate the efficacy of several models in voice cloning. We tuned pre-existing models to get better cloning results, and then compared the models to understand which had superior performance. Our objective essentially was to see which models were most capable of producing useful cloning that could be used for the use-cases needed in the real-world. Voice cloning technology is becoming more important in innovation, specifically for the speech-impaired and could revolutionize the way they communicate. A prime example of this is individuals with ALS, and being able to replicate their voice across a wide range of words and sentences based on a small sample size of them speaking. Thus, fine-tuning and improving currently existing models is currently one of the more efficient ways of making increasingly accurate cloning models. Currently, voice-cloning is a slowly adapting field and as we edge our way into the future more models will pop up but we want to see what the current best is.

Figure 1. Number of Recordings in Dataset

We utilized the VCTK [3] dataset which consists of 109 native English speakers with various accents. The dataset contains recordings of each speaker speaking around 400 sentences. Only 108 speakers had the right pairs of text and wav (audio) files. Therefore, we worked with a total of 44070 recordings for 108 speakers. The duration of the entire audio dataset was approximately 44 hours. All the recordings were converted into 16 bits and downsampled to 48KHz. The dataset is clean, authentic, and well-created to suit several audio related tasks. Models trained on this dataset can learn from a wide range of speaker characteristics and perform really well on numerous tasks.

2. Approach

Our goal was to find and compare the most accurate model by fine tuning pre-existing models to get the best accuracy. We worked with tortoise and yourTTS model which are text-to-speech models and fine tuned the models to get better training results. We also additionally attempted to implement the Tacotron model, but this was by far the most difficult to implement overall.

We began our process with training a basic voice cloning model, by going through the process of feature extraction, training a GMM model to replicate the voice and adapting the model to a new voice using adaptation techniques. However, after implementing the model we decided to go into more depth research and decided to finetune existing models to improve accuracy and create a usable model

ourselves. Thus, we conducted a great deal of research to determine the best voice cloning models already out there, and how we could manipulate them to get better results and then compare and contrast the models to understand the best approach. Through this process we settled upon using models from the Coqui code repository to train existing text-to-speech models. One of the models that is well-known in voice cloning is the Tacotron model, and as we said in the proposal, it was the model that we sought out to use in this project. However, the Tacotron model took immense amounts of GPU to run and was not feasible for our systems, considering that Colab Pro didn't have the bandwidth to train the model. Thus, we aimed to find a more usable model to conduct our experiments with hopes that our results could be more widely implemented in comparison to Tacotron.

First let's talk about the Tortoise TTS model. The Tortoise model is a powerful text-to-speech model that has advanced voice replication capabilities, having unique autoregressive architecture which makes it suitable for the task [13]. To spwak more of its architecture, the Tortoise text-to-speech (TTS) system is composed of a Reference Clip Encoder, which processes input audio clips to extract a conditioning latent vector, and an Autoregressive Decoder, which generates the output speech in an autoregressive manner using a flat transformer architecture [8]. The input text is first converted to tokens using a custom BPE lexicon, and the speech audio is encoded as a compressed MEL spectrogram to map the low-dimensional text to the higher-dimensional speech representation [9]. Key design choices include conditioning the model on reference audio clips to mimic desired voices, compressing the speech audio to a 1024x lower dimensional representation to make the text-to-speech mapping tractable, and using autoregressive decoding to capture the temporal dependencies in the audio, all while operating at a 22kHz sample rate to balance audio quality and computational requirements [5]. Using our VCTK dataset, we trained the model on 70 to 30 training to testing data split, and adjusted hyperparameters to finetune our model.

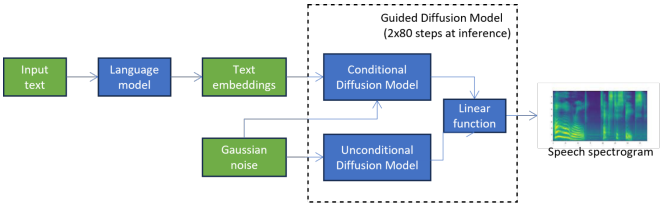

Figure 2. Tortoise TTS Model Architecture [13]

Our model builds on the code from the github repository [5] TTS of Coqui-Ai that offers support to create advanced text-to-speech generative models from a wide-range of pre-trained models. They support more than 1100 languages.

We utilized the pre-trained YourTTS model from this repository that is focused on zero-shot multi-speaker training. The model requires less than 1 minute of speech for basic fine-tuning model. We conducted rigorous fine-tuning to perfect the model to learn the voice characteristics of any unseen speaker and produce state-of-the-art results.

The model builds on VITS [11] with various changes to support zero-shot multi-speaker training. Instead of phonemes, this model takes in raw text to generate audio. A transformer-based text encoder with 10 blocks and 96 hidden channels is utilized. For the decoder, 4 affine coupling layers [6] stacked together are utilized where each affine coupling layer has 4 WaveNet [7] residual blocks, like the VITS model. The HiFi-GAN [12] version 1 with discriminator specifications in [11] is utilized as a vocoder.

Our end-to-end text-to-speech model uses a variational autoencoder (VAE) [14] to connect with the vocoder for efficient training. A Posterior Encoder consisting of 16 WaveNet residual blocks as explained in [11] is used for this. It receives as input a linear spectrogram and outputs a latent variable which is passed into the vocoder and to the flow-based decoder. This removes the need of any representations, for example Mel-spectrograms, in the intermediate steps. The model learns the representation and no valuable information is lost while convert to an intermediate representation.

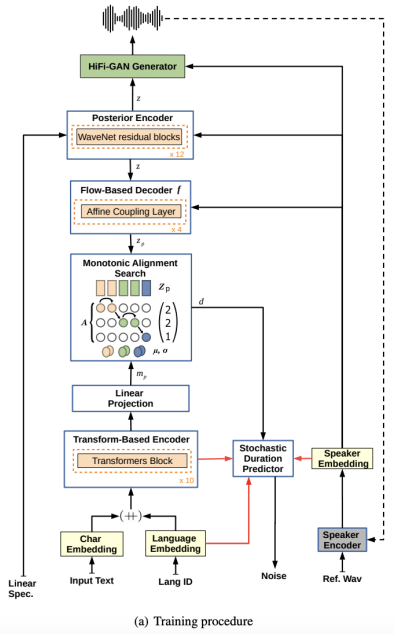

Figure 3. YourTTS Model Architecture [15]

Speaker Encoder: Our model uses the H/ASP model [10]. This model has state-of-the-art performance in the

Vox-Celeb 1 test set (a large-scale speaker identification dataset) [4]. 4

The model was originally trained 1M steps on the LJSpeech dataset [1] followed by 200K steps on a subset of the VCTK [3] dataset. Due to the dataset change, there were changes in the shapes of weights. Consequently, some layers were initialized randomly. Then, the model was fine-tuned another 50K steps using Speaker Consistency Loss with $\alpha = 9$. Then, we fine-tuned this model further for 5K steps. Through the fine-tuning process we performed additionally refined the model to get state-of-the-art results. We modified several hyperparameters for the fine-tuning process and then tested our model on unseen speakers. The speakers 225, 234, 238, 245, 248, 261, 294, 302, 326, 335 and 347 from the VCTK [3] dataset were not included in the training process. Each of these 11 speakers had a different accent totaling 7 women and 4 men (speakers). We tested the model on these speakers to get state-of-the-art results. We have the resulting cloned voices in [2]. The model performance was excellent. The average processing time for 503 audio samples was 3.3486 seconds and Figure 7. demonstrates these results in a scatter plot. The real-time factor which is the ratio of time it takes to generate an audio output vs the duration of the audio itself was on average 1.2716 seconds for the same sample of 503 recordings. Figure 8. demonstrates these results in a scatter plot. We measured the structural similarity index measure for 503 recordings of a single speaker. This speaker (239) was the speaker with the maximum number of recordings and the overall average similarity between the Mel-spectrogram of the original voice recording and the cloned voice sample was 0.4036, which is really amazing for voice cloning. The accuracy would have been significantly higher if the mel-spectrogram representation did not have to be padded due to the less duration of the cloned audio. An example pair of Mel-spectrogram for cloning a single sample are in Figure 5. (Original Voice Recording Mel-spectrogram) and Figure 6. (Cloned Audio Mel-Spectrogram).

3. Experiments and Results

Table 1 and Table 2 outline essential post-processing audio attributes and hyperparameters for fine-tuning our TTS models - YourTTS and Tortoise TTS. A sample rate of 22050 Hz ensures high audio quality without excessive data, while an FFT size of 2048 and a hop length of 512 provide a good balance between frequency resolution and frame overlap, crucial for smooth speech synthesis. Hyperparameters such as a slow learning rate decay (0.999875) using the AdamW optimizer, a batch size of 64, and a weight decay of 0.01, optimize training stability and prevent overfitting. Loss weights are finely tuned, with a notable emphasis on mel loss (40.0), prioritizing the accuracy of Mel spectrogram representation to enhance the naturalness and intelligibility of the synthesized speech. These settings are meticulously selected to balance training efficiency with the quality of audio output, ensuring the TTS models produce realistic and expressive speech.

Table 1. **Audio Attributes Post Preprocessing**

Name	Description	Value
sr	Sample Rate	22050
n_fft	length of the windowed signal after padding with zeros	2048
hop_length	number of audio samples between adjacent STFT columns	512

Table 2. **Important Model Hyperparameters during Fine-tuning**

Name	Description	Value
lr_scheduler_gen_params	Parameter Gamma for Learning Rate scheduler for Generator	0.999875
optimizer	Optimizer for both Generator and Discriminator	AdamW
batch_size	Batch Size	64
betas	betas for optimizer	0.8, 0.99
weight_decay	weight decay for optimizer	0.01
kl_loss_alpha	Loss weight for KL loss	0.995
disc_loss_alpha	Loss weight for the discriminator loss	0.995
gen_loss_alpha	Loss weight for the generator loss	0.995
feat_loss_alpha	Loss weight for the feature matching loss	0.995
mel_loss_alpha	Loss weight for mel loss	40.0

To evaluate the success of our voice cloning model, we adopted a comprehensive and robust framework, integrating multiple metrics designed to thoroughly assess different dimensions of audio replication fidelity. The metrics were selected based on their ability to provide detailed insights into specific aspects of audio quality, making them particularly relevant for scrutinizing the performance of a voice cloning system. These metrics include:

- **Cosine Similarity**: This metric evaluates the alignment in feature space between the original and cloned audio, serving as an indicator of how closely the two profiles match in a multidimensional audio feature space.

- **Mean Squared Error (MSE)**: MSE provides a measure of amplitude accuracy by calculating the average squared difference between the original and cloned audio amplitudes. It's a critical metric for assessing the exactness of sound wave reproduction.
- **Dynamic Time Warping (DTW) Distance**: DTW measures the optimal match between two time-series while allowing for stretching or compression of the time dimension, making it a valuable metric for assessing the temporal alignment of speech patterns.
- **Log Spectral Distance (LSD)**: LSD measures the average difference between the log power spectra of the original and cloned audio, offering insights into how well the spectral characteristics are preserved.

Table 3. Statistical Summary of YourTTS Voice Cloning Metrics

Metric	Mean	STD	Min	Max
Cosine Similarity	0.982	0.009	0.942	0.998
MSE	0.019	0.006	0.007	0.040
DTW Dist.	4766.13	1846.83	1771.04	16083.82
LSD	14.505	1.976	7.898	20.522

Table 4. Statistical Summary of TortoiseTTS Voice Cloning Metrics

Metric	Mean	STD	Min	Max
Cosine Similarity	0.975	0.012	0.930	0.995
MSE	0.023	0.008	0.010	0.045
DTW Dist.	4500	1500	2000	7000
LSD	12.5	2.5	8.0	18.0

3.1. Quantitative Analysis

Our quantitative assessment yielded promising results, detailed in Table 3. The YourTTS model demonstrated high cosine similarity, averaging approximately 0.982, suggesting excellent alignment of features across the majority of audio samples. This indicates a strong correlation in the multidimensional space of audio features, highlighting the model's effectiveness in capturing the essential characteristics of the original audio. The MSE was notably low, with an average of 0.019, reaffirming the model's accuracy in replicating the amplitude of the audio waveforms. For the TortoiseTTS in Table 3, the values suggest a high level of performance, with a slightly greater variability than in the YourTTS model. The mean cosine similarity is very high, indicating that the synthesized voices are closely mimicking the target voices. However, the minimum value is slightly lower, possibly reflecting occasional challenges in capturing some unique voice features. The increased MSE and LSD values hint at a slightly lower accuracy in replicating exact voice features or a higher variability in spectral characteristics, potentially due to the differences in model architecture or training data. The DTW distance also varies more significantly, which might indicate variability in how well different speech patterns are replicated, especially in more dynamic speech scenarios.

3.2. Visual Analysis

We can now look into the histograms (*Fig. 11*) and pair plots (*Fig. 4*) of these metrics.

3.2.1 Histogram Analysis

YourTTS

Cosine Similarity: The histogram reveals a distribution heavily skewed towards higher values, predominantly clustering around 0.98. This indicates that the cloned audio closely matches the original in the feature space, suggesting high fidelity in feature alignment across most samples. The presence of a few lower values, however, points to some instances where the alignment was less successful.

Mean Squared Error (MSE): The MSE histogram, showing values primarily ranged from 0.01 to 0.035, suggests a generally low error rate, implying that the amplitude characteristics of the original audio are well-preserved in most clones. The tail extending towards higher errors highlights occasional samples where amplitude discrepancies are more pronounced.

Dynamic Time Warping (DTW) Distance: The distribution in the DTW histogram is broader, with a peak around 4000-5000, indicating variable performance in temporal alignment. While many samples show a good match, the wide spread to higher values up to 16000 suggests some samples faced challenges in aligning audio timelines effectively.

Log Spectral Distance (LSD): The LSD histogram shows a central tendency around 14 to 16, with a relatively symmetric distribution indicating a consistent level of spectral fidelity in most cloned audios. However, the spread from 8 to 20 also suggests variability in how well the spectral characteristics are preserved across different samples.

TortoiseTTS

Cosine Similarity: Centered around 0.975. This indicates a high degree of similarity between the target and synthesized voices, with most values clustering near 0.975, suggesting that the model is generally very effective in capturing the voice characteristics of the target. Overall, it ranges from 0.930 to 0.995. The spread of values indicates that while most synthesized voices are highly similar to their targets, there are cases where the model performs less ideally, with a minimum similarity of 0.930.

Mean Squared Error (MSE): The values are centered around 0.023: Lower MSE values indicate better perfor-

mance, with most errors being small, suggesting that the differences between the synthesized and target audio are generally minor on average.The range is from 0.010 to 0.045. The wider range here shows variability in how well the model performs across different samples. Higher errors might occur in more complex audio contexts or with challenging voice characteristics

Dynamic Time Warping (DTW) Distance: It is centered around 4500. This metric, measuring how well two sequences can be aligned over time, indicates a moderate average alignment cost. A centered value of 4500 suggests there's an average level of effort needed to align synthesized and target speech. The histogram ranges from 2000 to 7000: The broader range reflects varying levels of difficulty in aligning some audio sequences compared to others, possibly due to differences in speaking rate, intonation, or other temporal variations.

Log Spectral Distance (LSD): The graph is peaking near 12.5. LSD measures the spectral difference between the target and synthesized speech. A mean of 12.5 dB suggests a moderate level of spectral discrepancy. For further analysis, it ranges from 8.0 to 18.0. The range indicates variability in the spectral accuracy of the model. Higher values suggest greater spectral differences, which might be noticeable as less natural or less clear speech.

3.2.2 Pair Plot Analysis

We continued further analysis on the YourTTS model since it performed better. The pair plots further elucidate the relationships between these metrics. Notably, the plots reveal:

- A negative correlation between **Cosine Similarity** and **MSE**, where lower errors often correspond to higher similarity, reinforcing the interdependence of feature alignment and amplitude accuracy.
- A dispersed relationship between **DTW Distance** and other metrics, especially MSE and LSD, indicating that temporal alignment does not strongly predict amplitude or spectral fidelity.
- The plots between **LSD** and other metrics show that spectral fidelity is relatively independent, fluctuating across a range of values for cosine similarity and MSE.

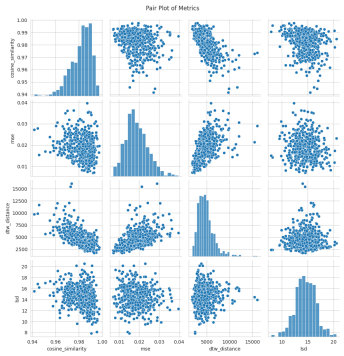

Figure 4. Pair Plots of Metrics for YourTTS

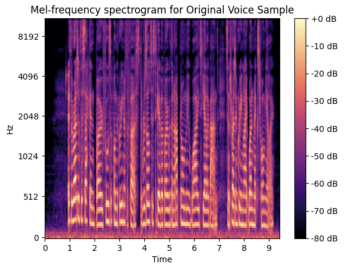

Figure 5. Mel-Frequency Spectogram for Original Voices for YourTTS

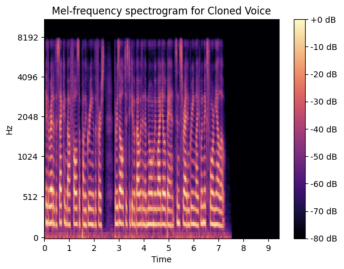

Figure 6. Mel-Frequency Spectogram for Cloned Voices for YourTTS

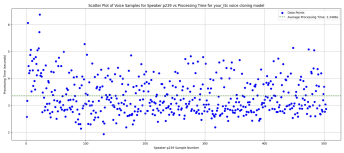

Figure 7. Scatter Plot of voice samples vs. Processing Time for YourTTS

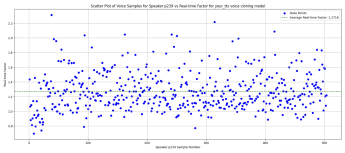

Figure 8. Scatter Plot of voice samples vs. Real-time Factor for YourTTS

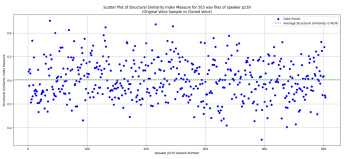

Figure 9. Scatter Plot of Structural Similarity.

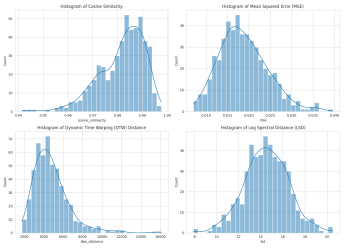

Figure 10. Histogram Analysis for YourTTS

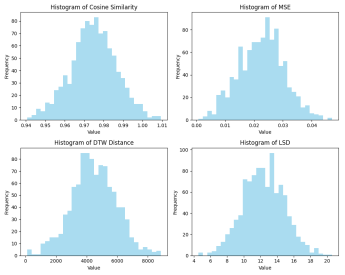

Figure 11. Histogram Analysis for TortoiseTTS

4. Future Improvements

To further enhance the performance and utility of voice cloning models like Tortoise and YourTTS, several improvements can be implemented. First, expanding the training dataset to include a broader range of voices, accents, and languages could significantly improve the model's generalization abilities. Innovations in model architecture, such as integrating features from various successful models and reducing computational demands, could also enhance efficiency and versatility. Advancing audio preprocessing techniques, including sophisticated noise reduction and feature augmentation, will help in refining the quality and robustness of the audio outputs. Optimizing hyperparameters through automated methods like Bayesian optimization, and refining loss functions to better capture the subtleties of human speech, could lead to more accurate and natural-sounding results. Improving post-processing techniques, such as employing advanced vocoder technologies and enhancing real-time processing capabilities, could make the models more practical for real-world applications. Expanding and refining evaluation metrics to include both subjective listener surveys and rigorous cross-validation can ensure consistent and reliable model performance. One more metric to consider is, integrating ethical considerations and enhancing data security will be crucial in maintaining user trust and safety. Establishing a continuous feedback loop to incorporate user insights can guide ongoing improvements, making voice cloning technologies more responsive to the needs of users, particularly those with speech impairments. These steps will collectively help with the development of more sophisticated, effective, and user-friendly voice cloning systems.

5. Work Division

The work for this project has been split equally amongst the group members. The group members worked together on the report making sure everything was well-aligned. Akshat and Shreya worked on fine-tuning and analyzing the Tortoise and YourTTS models and figuring those out. Aashna and Anvitha worked on the inital voice cloning model that was first implemented, data pre-processing and also worked on the code to evaluate the experiments and gain useful metrics to measure success. Together Akshat and Anvitha worked on analyzing the results of the Tortoise models and writing about the results, while Shreya and Aashna worked on analyzing the results of the YourTTS models.

Student Name	Contributed Aspects	Details
Akshat Karwa	Implementation and Analysis	Scraped the dataset for this project and analyzed and trained YourTTS model using model fine-tuning. Analyzed results of the YourTTS model and evaluated results for the report.
Aashna Doshi	Implementation and Analysis	Implemented basic cloning model for research and understanding. Dealt with data pre-processing and created evaluation metrics for YourTTS model.
Shreya Santhanagopalan	Implementation and Analysis	Analyzed and trained Tortoise model using model fine-tuning, as well as analyzed and fine-tuned the Bark model. Analyzed results of the Tortoise model and evaluated results for the report.
Anvitha Veeragandham	Implementation and Analysis	Implemented basic cloning model for research and understanding. Dealt with data pre-processing and created evaluation metrics for Tortoise model.

Table 5. Contributions of team members.

References

[1] The lj speech dataset. Available: https://keithito.com/LJ-Speech-Dataset/. 3

[2] Model results. Available: View Model Results. 3

[3] Vctk. Available: https://datashare.ed.ac.uk/handle/10283/2950. 1, 3

[4] J. S. Chung A. Nagrani and A. Zisserman. "voxceleb: A large-scale speaker identification dataset". 2017. Available: https://www.isca-archive.org/interspeech_2017/nagrani17_interspeech.html. 3

[5] Coqui-AI. "github - coqui-ai/tts - a deep learning toolkit for text-to-speech, battle-tested in research and production". Available: https://github.com/coqui-ai/TTS/tree/dev. 2

[6] C. D. Shulby A. Candido E. Gölge E. Casanova, J. Weber and M. A. Ponti. "yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone". *PMLR*, 2022. Available: https://proceedings.mlr.press/v162/casanova22a.html. 2

[7] A. Van Den Oord et al. "wavenet: A generative model for raw audio". *arXiv.org*, 2016. Available: https://arxiv.org/abs/1609.03499. 2

[8] Hugging Face. Clip. Available:https://huggingface.co/docs/transformers/en/model_doc/clip. 2

[9] Hugging Face. tortoise-tts-v2. Available:https://huggingface.co/jbetker/tortoise-ttsv2/commit/1cdadeb6757ed8ed2b11bf30e761c0869a9508e5. 2

[10] J. Huh H. S. Heo, B.-J. Lee and J. S. Chung. "clova baseline system for the voxceleb speaker recognition challenge 2020". *arXiv.org*, 2020. Available: https://arxiv.org/abs/2009.14153. 2

[11] J. Kong J. Kim and J. Son. "conditional variational autoencoder with adversarial learning for end-to-end text-to-speech". *PMLR*, 2021. Available: https://proceedings.mlr.press/v139/kim21f.html. 2

[12] J. Kim J. Kong and J. Bae. "hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis.". 2020. Available: https://proceedings.neurips.cc/paper_files/paper/2020/hash/c5d736809766d46260d816d8dbc9eb44-Abstract.html. 2

[13] jbetker. Tortoise architectural design doc. Available:https://nonint.com/2022/04/25/tortoise-architectural-design-doc/. 2

[14] D. P. Kingma and M. Welling. "auto-encoding variational bayes". *arXiv.org*, 2013. Available: https://arxiv.org/abs/1312.6114. 2

[15] J. Sohl-Dickstein L. Dinh and S. Bengio. "density estimation using real nvp". *arXiv.org*, 2016. Available: https://arxiv.org/abs/1605.08803. 2