Exploring Generative Models (Variational Autoencoders) for Audio Generation

By: Rabin Nepal (U00901360)
Department of Electrical and Computer Engineering
University of Memphis, TN, USA

1. INTRODUCTION

Speech is a unique trait inherent to humans, allowing us to convey thoughts and ideas through vocal sounds. It serves as the most natural method of communication among humans. And even in the area of human-computer interaction, attempts have been made to integrate speech. However, the main issue within this implementation is that all the verbal replies we get from computers, intelligent systems, or artificial intelligence systems still need to be pre-recorded, and the computers themselves do not actually generate speech data. Hence, there is a need to generate speech data that works without the restrictions possessed by the prerecorded audio data.

1.1 Problem Statement

The field of Artificial Intelligence (AI) has advanced to a stage where we can now reconstruct or generate many forms of data like images, videos, text, and even complete entire sentences using AI and more specifically generative AI models. Most of the research work in generative models, however, has been on the synthesis/generation of textual data, which is not as natural a form of communication as speech data. The textual data format has also limited the access of the technology to people who have the ability to read and write.

This project will explore the usage of Generative models on audio data to shed more light on the potential usage of such models to generate speech, which is a more inherent form of communication to humans. This project's scope is to explore the use of generative AI models, more specifically the VAE (Variational AutoEncoders) model to see how well these models work on audio generation tasks. Through this exploration, the project aims to contribute to the broader understanding of variational autoencoders and their role in audio generation.

1.2 Literature Review

The field of artificial intelligence was predominantly used in the prediction or inference task using the vast amount of data available. The idea that AI can actually be used to generate something new was first introduced by Geoffery Hinton and his team in their 2006 paper [1]. After the introduction of GANs (Generative Adversarial Networks) in 2014 by Ian Goodfellow, a major leap in the domain of generative AI happened leading to major contributions to the field [2]. Deep learning frameworks were later distinguished into two main domains: a) Generative Frameworks and b) Discriminate Frameworks.

a) **Discriminative AI:**
Discriminative AI deals with models that separate given data instances into different classes by learning the differences within the dataset. Image classification through Convolution Neural Is an example of discriminative AI.

b) **Generative AI:**
Generative AI, instead of learning the differences in data, tries to learn the actual distribution of data and tries to generate new data samples using the learned distribution of data

There are many generative architectures that have been developed after its introduction but the two most famous are the VAE (Variational AutoEncoders), which this project aims to utilize for the generation of audio samples, and GANs (Generative Adversarial Networks).

Audio data has been primarily used as an input to the machine learning models, mostly made famous by Google's speech recognition engine. There also have been works in the Speech Synthesis task from the past but mostly focused on text-to-speech synthesis to convert written text into spoken words, trying to mimic the human form of speech. Works like WaveNet by DeepMind [3] can generate natural-sounding raw audio waveforms and use generative AI to do so. Microsoft has also designed a robust speech generation AI framework called FastSpeech with MFCC (Mel-Frequency Cepstral Coefficients) and a feed-forward-based Transformer model to generate natural speech [4].

2. METHODOLOGY

2.1 Block Diagram

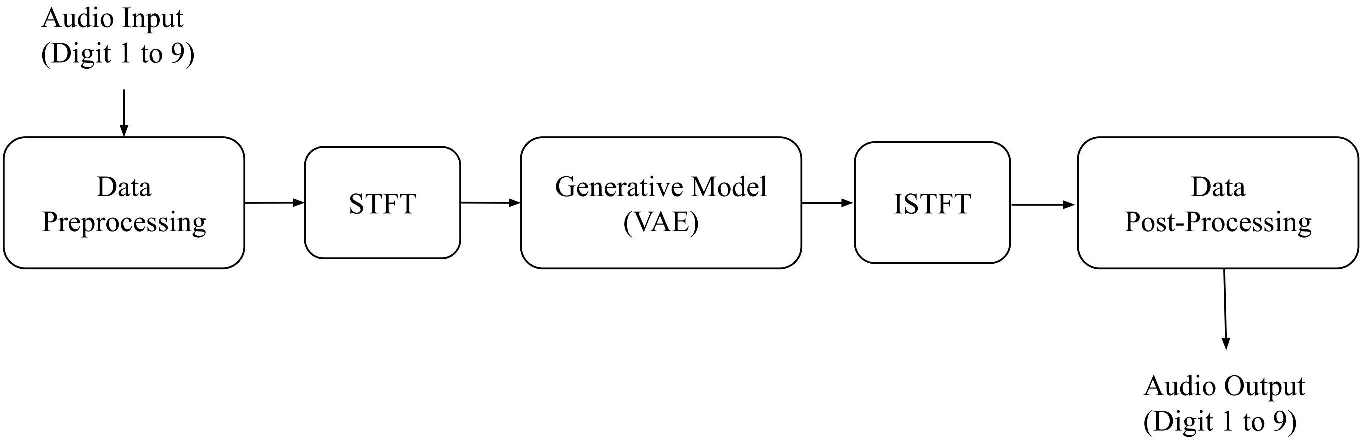

Fig. 1: Block Diagram of the Proposed Audio Generation System

Fig. 1 depicts the proposed system for the audio generation task. The input to the system is an audio sample of spoken digits (explained further in section 2.2). The steps include data preprocessing, feature extraction using STFT (Short Time Fourier Transform), and feeding into a generative model (Variational Autoencoder in this case), which gives a newly generated STFT for generated audio data. The ISTFT (Inverse Short Time Fourier Transform) block reconverts the signal back to an audio signal and this signal is then sent to a data post-processing block which does the necessary post-processing such as increase in amplitude, frequency conversion, etc., and finally, a proper generated audio can be observed from the system. The steps have been discussed further in detail:

2.1.1 Data Preprocessing

Audio data from the dataset can include recordings with different lengths, different amplitudes, and silences or there could be data samples with faulty recordings. The data preprocessing step eliminates this inconsistency in data samples.

2.1.2 STFT

STFT is a way of observing the frequency and phase content of a signal over time. It provides a localized time-frequency representation of the signal that allows us to observe the changes in frequency over time more accurately. This is crucial for understanding the differences among the different audio digits in the dataset.

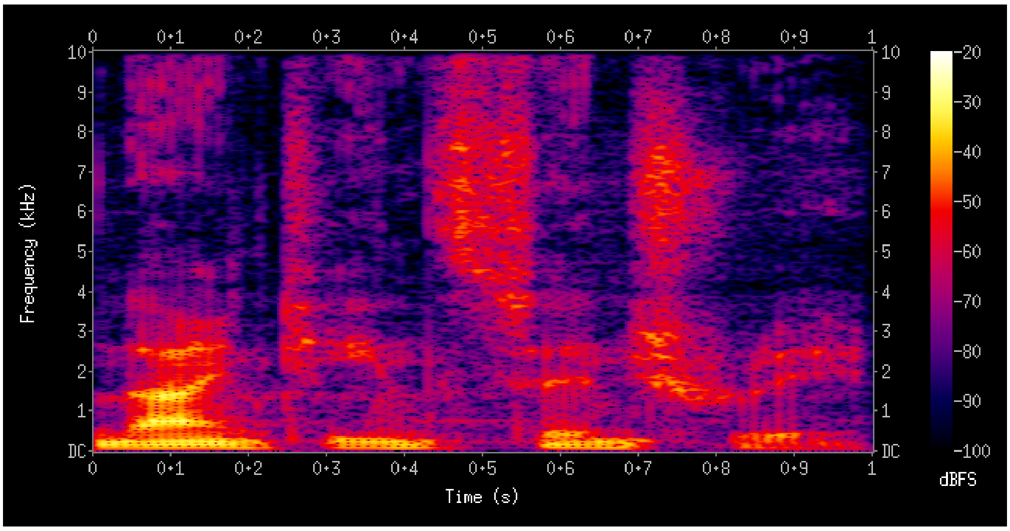

Fig 2: A Sample of STFT Diagram [5]

2.1.3 Generative Model

Variational Autoencoder is the generative model proposed for this study. This model is an unsupervised learning method that can learn the distribution of the data to generate new data. VAE architecture is composed of an encoder and decoder that is trained to minimize the reconstruction error between the encoded-decoded data and initial data. The encoder of VAE is responsible for mapping the dataset into a latent representation, which is a joint probability distribution of the samples. A random sample data is sampled from the latent representation (joint probability distribution) by the decoder. This sampled data is a reconstruction of data from the latent space. The error in reconstruction is calculated using the input data and the reconstructed data, and the model is trained such that the reconstruction error is minimized. [6]

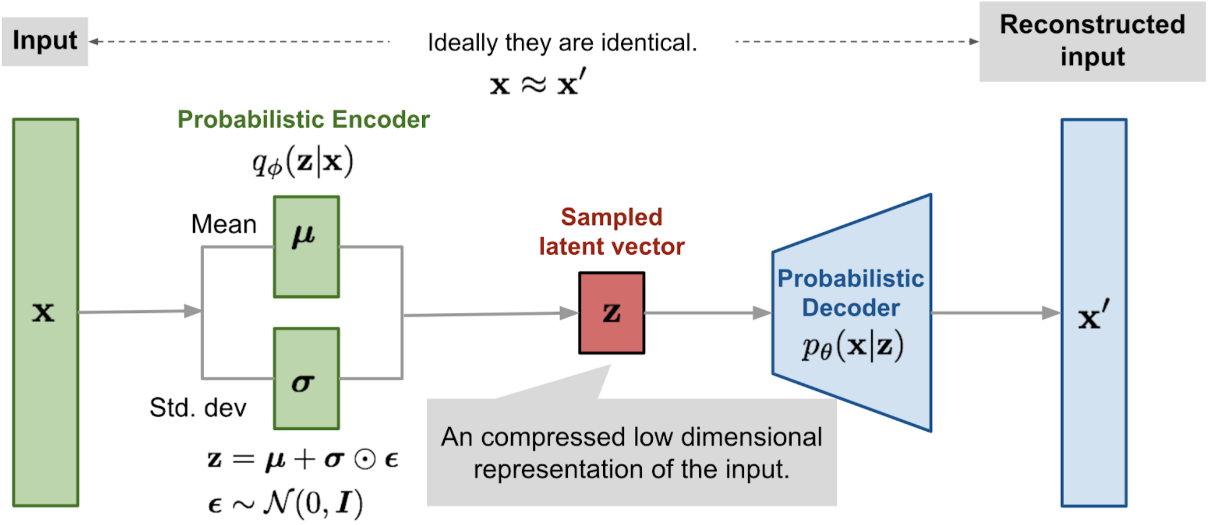
Input
Ideally they are identical.
$\mathbf{x} \approx \mathbf{x}'$
Reconstructed input
Probabilistic Encoder
$q_\phi(\mathbf{z}|\mathbf{x})$
Mean
$\boldsymbol{\mu}$
Std. dev
$\boldsymbol{\sigma}$
$\mathbf{x}$
Sampled latent vector
$\mathbf{z}$
Probabilistic Decoder
$p_\theta(\mathbf{x}|\mathbf{z})$
$\mathbf{x}'$
$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$
$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$
An compressed low dimensional representation of the input.

Fig 3: Architecture of Variational AutoEncoders [7]

2.1.4 ISTFT

The ISTFT block of the system is responsible for recovering the original signal from the transformed signal by STFT. The output of the VAE model is a newly generated STFT of MNIST audio data, which needs to be converted back to an audio signal so that the generated output can be heard.

2.1.5 Data Post-Processing

Data post-processing is the final stage of the system that is used to transform the audible signal from the ISTFT block. The audible signal at this stage may not yet be audible due to the low amplitude of audio data or same alterations in frequency domain during the ISTFT process or any other issue that can make the generated audio not audible. This block aims to eliminate such issues.

2.2 Audio MNIST Dataset

This project will utilize the audio version of the MNIST dataset made available by [8]. AudioMNIST Dataset has an audio recording of all the digits (from 0 to 9), from a total of 60 speakers, totaling 30,000 samples. The dataset of 9.5 hours long and is recorded from speakers with a broad range of accents, including both native and non-native speakers. The recorded sample is 16-bit single-channel audio, sampled at 48KHz.

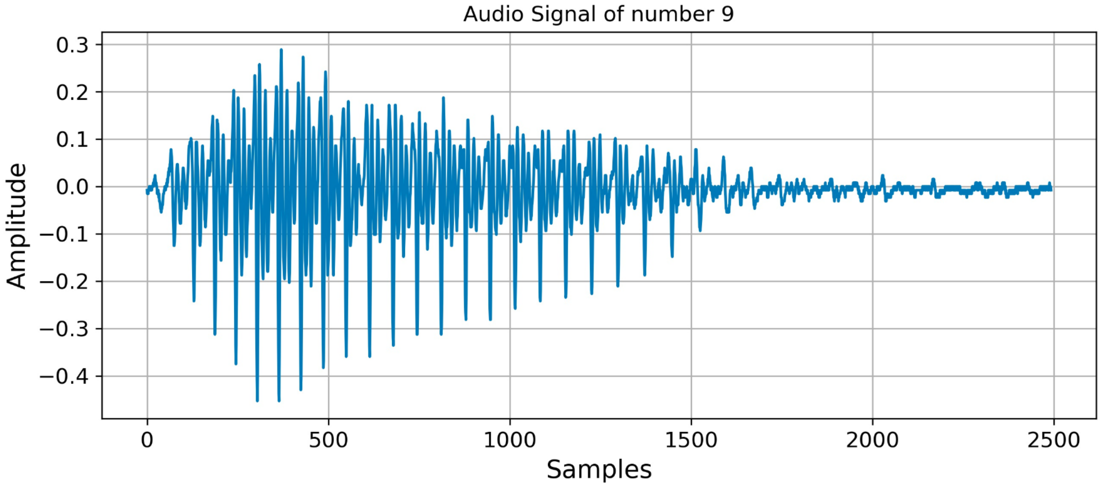

Fig. 4: Plotted Sample Audio Signal of Digit 9 from AudioMNIST Dataset

3. REFERENCES

[1] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A Fast Learning Algorithm for Deep Belief Nets,” *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006, doi: 10.1162/neco.2006.18.7.1527.
[2] I. J. Goodfellow *et al.*, “Generative Adversarial Networks,” *arXiv.org*, Jun. 10, 2014. https://arxiv.org/abs/1406.2661
[3] A. van den Oord *et al.*, “WaveNet: A Generative Model for Raw Audio,” *arXiv.org*, Sep. 12, 2016. https://arxiv.org/abs/1609.03499
[4] Y. Ren *et al.*, “FastSpeech: Fast, Robust and Controllable Text to Speech,” *arXiv.org*, May 22, 2019. https://arxiv.org/abs/1905.09263
[5] Contributors to Wikimedia projects, “Short-time Fourier transform,” *Wikipedia*, Oct. 11, 2022. https://en.wikipedia.org/wiki/Short-time_Fourier_transform (accessed Sep. 30, 2023).
[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.
[7] L. Weng, “From Autoencoder to Beta-VAE,” *Lil’Log*, Aug. 12, 2018. https://lilianweng.github.io/posts/2018-08-12-vae/ (accessed Sep. 30, 2023).
[8] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek, “Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals,” *arXiv.org*, Jul. 09, 2018. https://arxiv.org/abs/1807.03418