Deep Learning-Based Breath Sound Synthesis for Medical Simulations

Namrata Harish[1], Natalie Heitkamp[1], Nitya Lagadapati[1], Austin Baird[2], David Hananel[2]
[1]Department of Bioengineering, [2]Department of Surgery, Division of Healthcare Simulation Sciences, University of Washington

Background

Project Motivation

- Medical simulation manikins are used to practice clinical skills without interaction with live patients *(Image 1)*
- Current methods for training lung auscultation rely on pre-recorded breathing sounds from limited sources
- We hypothesize that deep learning methods can be used to generate dynamic breathing sounds that can:
  - Align with a customized patient physiology and demographic profile
  - Improve the generalizability and efficacy of auscultation training simulations

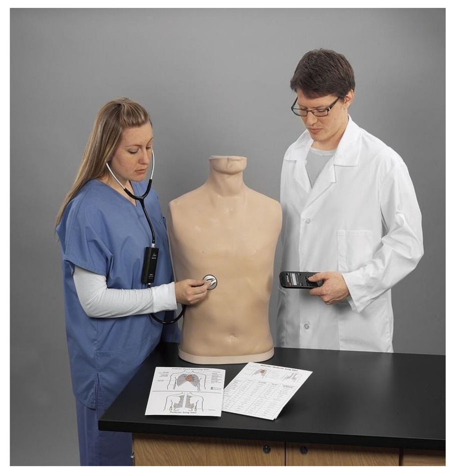

Image 1. *Medical trainees using a simulation manikin (1).*

Audio Synthesis

- Audio waveforms can be decomposed into their frequency components over time, visualized as a spectrogram, using a short-time Fourier transform (*Figure 1*)[1]
- Mel spectrograms use a frequency scale based on human pitch perception, with more detail in lower, perceivable frequencies
  - Optimal for breath sound analysis and generation[2]
- Both magnitude and phase of each frequency component are important for accurate signal reconstruction
- The Griffin-Lim algorithm converts Mel spectrograms back to waveforms by iteratively estimating phase from magnitude

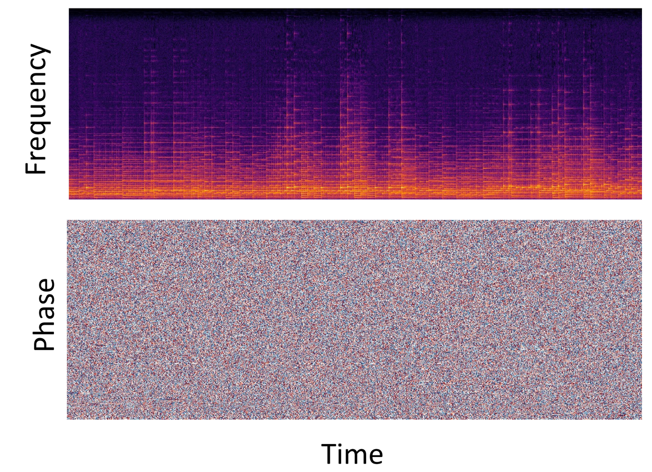

Figure 1. *Magnitude (top) and phase (bottom) spectrograms of a piano recording.*

Deep Learning Approach

- Deep learning (DL) approaches for sound generation can capture more complexity than other modeling methods
- Types of generative models include autoregressive models, variational autoencoders, and adversarial networks
- The black box nature of DL models may not be optimal for clinical applications
- Conditional variational autoencoders (CVAEs):
  - Follow an encoder-decoder structure during training
  - During generation, the decoder takes the learned latent vector and conditional labels to generate new samples

Methods

Dataset Information and Preprocessing Techniques

- The data was collated from two open-source lung sound databases[3,4], which contained:
  1. 1025 breathing audios from 219 patients
  2. Age, sex, location of sound acquisition (anterior/posterior & left/right), disease state information for each patient
- The model was trained on the Mel spectrograms of available audio samples and patient metadata
- Model Inputs (per sample):
  1. Normalized (min-max normalization) Mel spectrogram of the sample represented as a 2D matrix
  2. Demographic and clinical attributes (5 total features) of the corresponding patient encoded into a single vector

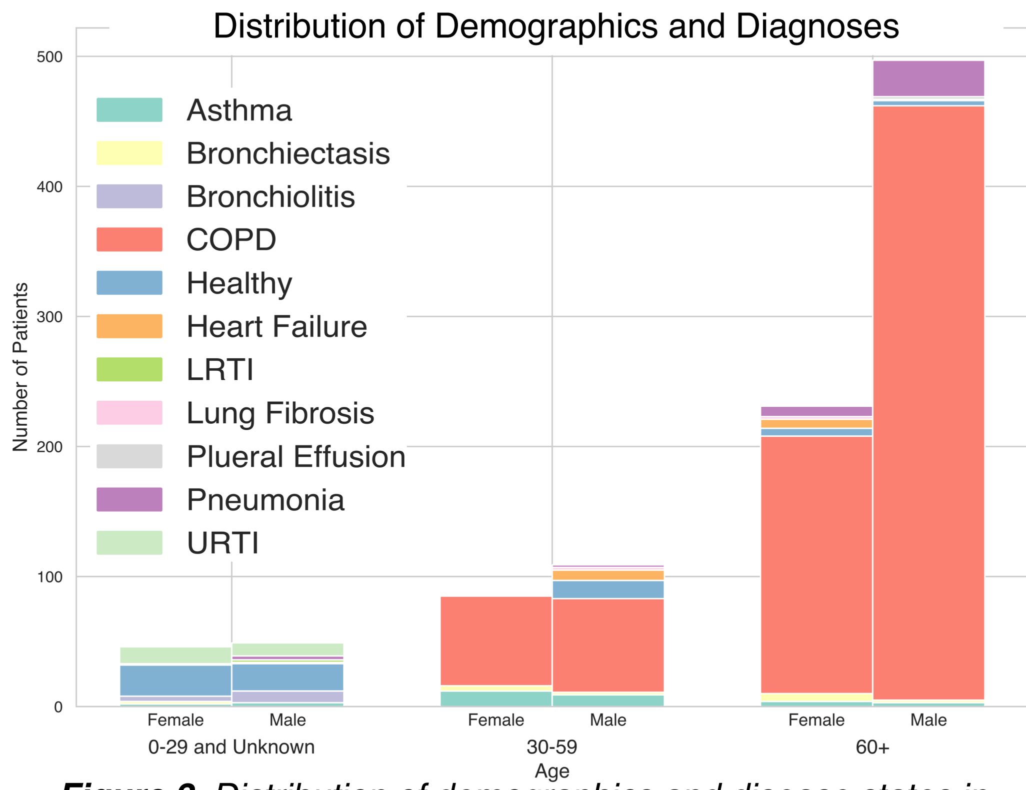

Figure 2. *Distribution of demographics and disease states in the dataset used for model training.*

Model Architecture: CVAE

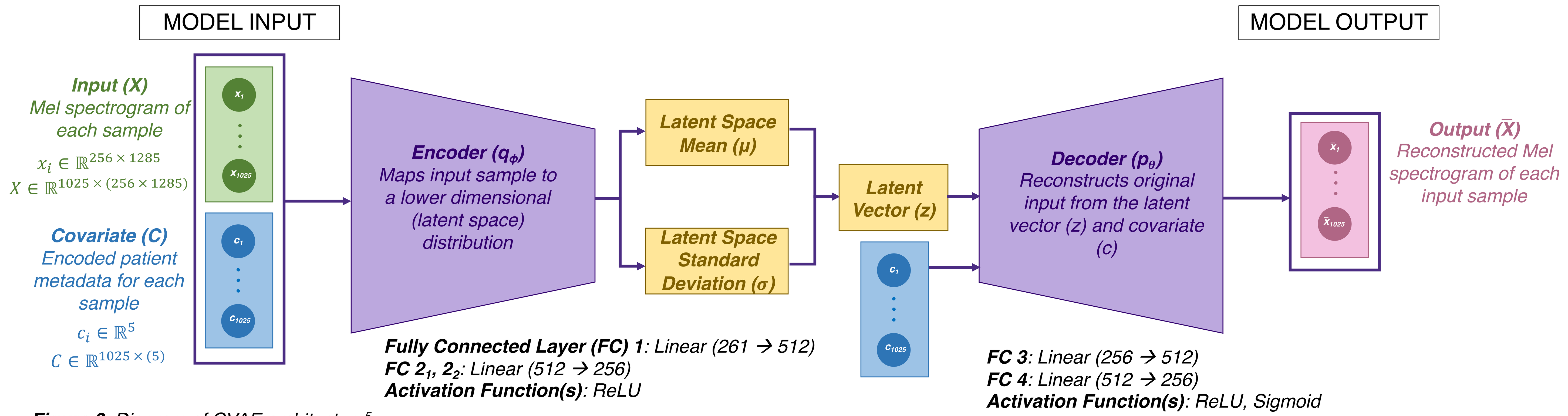

Figure 3. *Diagram of CVAE architecture[5].*

Results

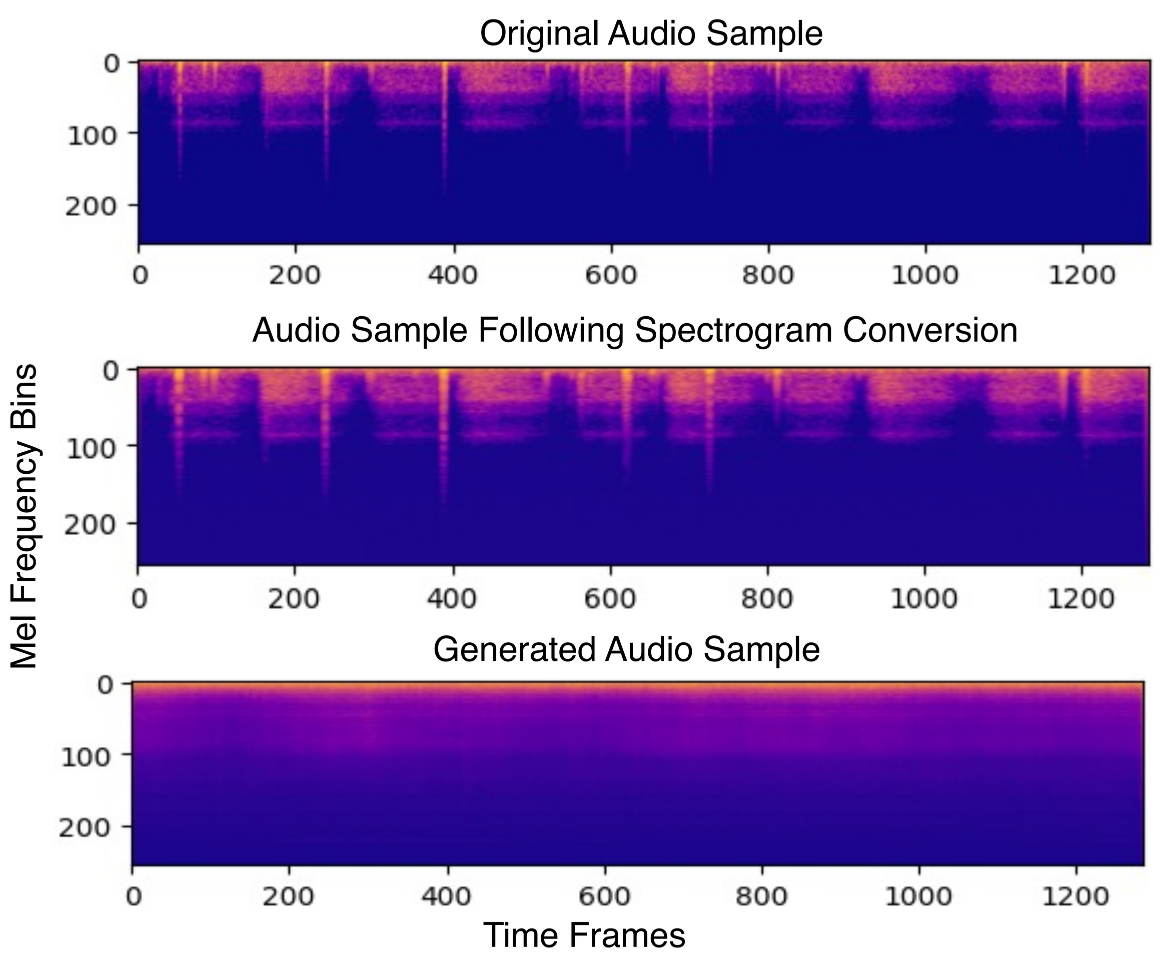

Figure 4. *Mel spectrograms of a) an original audio sample (top), b) the same audio sample after a cycle of spectrogram to audio conversion, and c) a generated sample, all corresponding to same patient profile (70-year-old female with COPD, left-side sample acquisition).*

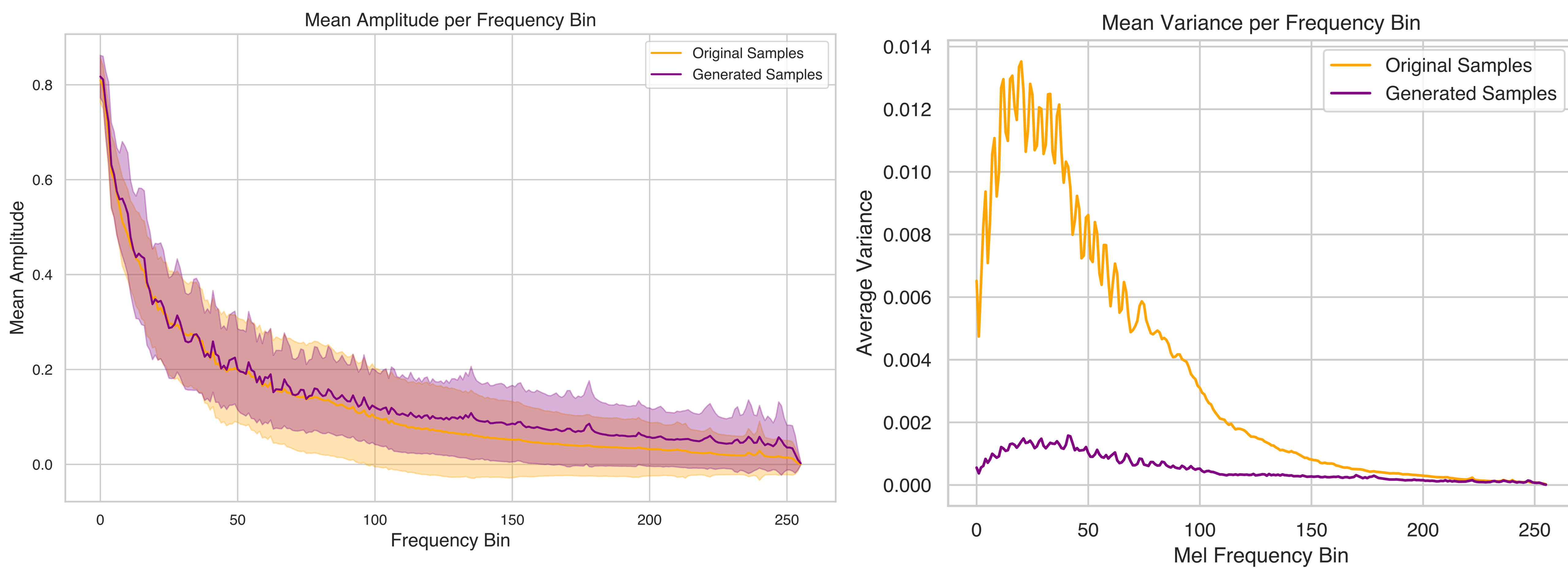

Figure 5. *a) Mean Mel frequency amplitude (left) and b) mean Mel temporal frequency variance (right) across held-out test samples (original) versus generated samples conditioned on the same metadata.*

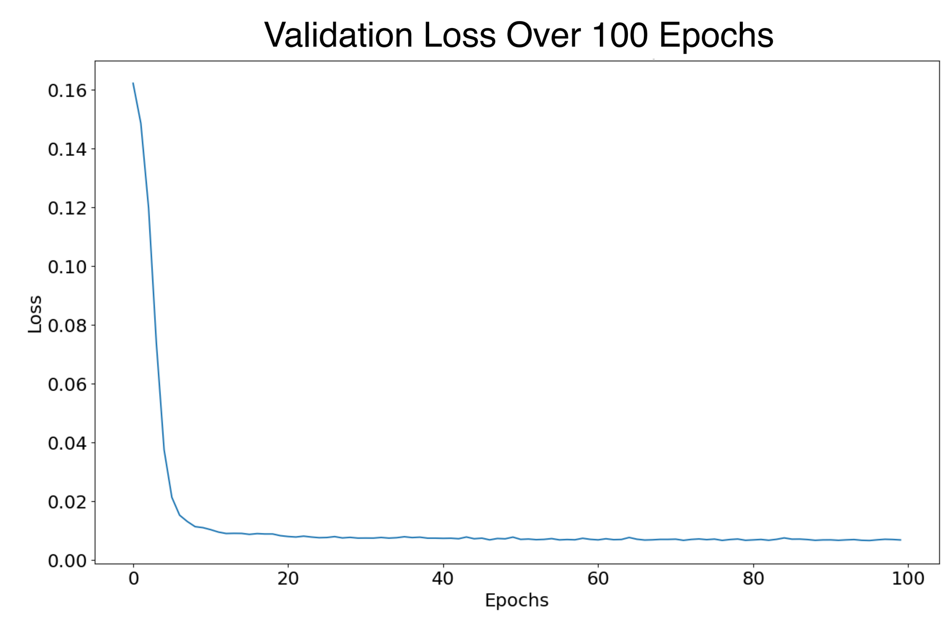

Figure 6. *Validation loss over the first 100, out of 10,000 training epochs, computed using mean squared error (MSE).*

- Noisy artifacts are introduced following the spectrogram to audio conversion cycle *(Figures 4a, 4b)*
- Generated spectrograms are smoother (lower resolution) than the original audio spectrograms (*Figure 4c*)
- There is consistent overlap between the generated and original audio samples for the mean amplitude (*Figure 5a*)
- The validation loss plateaus around 0.01 before 10 epochs (*Figure 6*)

Conclusions

- Mel spectrogram conversions compress magnitude data and omit phase information, limiting accurate audio reconstruction and model performance
- The CVAE captures lower Mel frequency components of the original samples but fails to accurately capture high-resolution features
- Learning phase information is challenging for the model due to the inherent randomness of phase data

Future Work

- Experiment with waveform-based models to reduce dependence on phase data for audio reconstruction
- Include data from diverse sources to improve model generalizability
- Design an accessible and scalable user interface
- Integrate the software with a training manikin system to provide a realistic training experience

References and Acknowledgements

[1] Sander Dieleman, "Generating music in the waveform domain", Mar. 2020
[2] J. Park *et al.*, *Scientific Reports*, vol. 13, Jan. 2023
[3] M. Fraiwan *et al*, *Data in Brief*, vol. 35, Apr. 2021
[4] B. M. Rocha *et al.*, *Precision Medicine Powered by pHealth and Connected Health*, vol. 66, Nov. 2017
[5] C. De Luca, "CVAE for Timbre Manipulation and Sound Generation", Dec. 2023

Thank you to Dr. Austin Baird and Dr. David Hananel for their guidance and mentorship throughout this project, and to Dr. Christopher Neils and Dr. Rupak Rajachar for their ongoing support.