LEARNING DISENTANGLED AUDIO REPRESENTATIONS THROUGH CONTROLLED SYNTHESIS

Yusuf Brima, Ulf Krumnack, Simone Pika & Gunther Heidemann
Institute of Cognitive Science
Universität Osnabrück
Wachsbleiche 27, D-49090 Osnabrück, Germany
{ybrima,krumnack,spika,gheidema}@uos.de

ABSTRACT

This paper tackles the scarcity of benchmarking data in disentangled auditory representation learning. We introduce *SynTone*, a synthetic dataset with explicit ground truth explanatory factors for evaluating disentanglement techniques. Benchmarking state-of-the-art methods on SynTone highlights its utility for method evaluation. Our results underscore strengths and limitations in audio disentanglement, motivating future research.

1 INTRODUCTION

Disentangled representation learning shows promise for creating useful data encodings. However, it remains underexplored for auditory data, despite potential benefits (Locatello et al., 2019; Higgins et al., 2017; Kim & Mnih, 2018). Disentangled representations refer to learned encodings (*codes* for short) where explanatory factors are encoded in *independent*, *compact*, and *informative* dimensions. These *factorized* representations offer an improved generalization, explainability, reduced sample complexity, and avoidance of shortcut learning (Carbonneau et al., 2020).

Progress in auditory disentanglement is hampered by the lack of suitable benchmarking datasets for evaluation (Hsu et al., 2017; Mo et al., 2019). Auditory signals present unique challenges due to their *high dimensionality*, *temporal dynamics*, and *rich hierarchical* structure. To help address this evaluation gap, we introduce *SynTone*: a basic dataset of synthetic tones curated for quantitatively assessing disentangled representation learning techniques in the acoustic domain. We used this new dataset to conduct a benchmarking study comparing state-of-the-art (SOTA) disentanglement frameworks. Performance was measured using *supervised disentanglement metrics*. Our goals were to (i) demonstrate SynTone's utility for method evaluation/comparison, and (ii) highlight the strengths and limitations of audio disentanglement algorithms to help guide future research.

2 METHODOLOGY

2.1 SYNTONE DATASET

Following ideas from disentanglement in the vision domain (Matthey et al., 2017; Burgess & Kim, 2018), we develop SynTone - a simple synthetic audio dataset with explicit ground truth factors to facilitate disentanglement analysis. SynTone comprises $32,000$ 1-second audio samples (16kHz sampling rate) synthesized by systematically varying three generative parameters: timbre $\mathbb{T}$ (sine, triangle, square, sawtooth waveforms), amplitude $\mathbb{A}$ (20 discrete levels from 0 to 1), and frequency $\mathbb{F}$ (400 discrete steps from 440 to 8000Hz). The dataset structure is formally: $\mathbb{T} \times \mathbb{A} \times \mathbb{F}$, with each sample having a unique $(T \in \mathbb{T}, A \in \mathbb{A}, F \in \mathbb{F})$ tuple defining its generative factors. Having full knowledge of the data generative process allows quantitative assessment of models' effectiveness at learning the ground truth factors from the audio samples. This dataset aims to provide a benchmark for the analysis of disentanglement techniques on a perceptual signal domain where the factorized generative sources are specified.

Model	Compactness		Holistic		Modularity
	MIG	SAP	DCIMIG	JEMMIG	Mod. Score
Vanilla VAE	**0.280 ± 0.0**	**0.045 ± 0.000**	**0.005 ± 0.000**	0.300 ± 0.0	0.651 ± 0.031
β-VAE	0.184 ± 0.0	0.035 ± 0.000	0.004 ± 0.000	0.201 ± 0.0	0.750 ± 0.022
Factor-VAE	0.118 ± 0.0	0.035 ± 0.000	0.003 ± 0.000	**0.126 ± 0.0**	**0.762 ± 0.014**
β-TCVAE	0.166 ± 0.0	0.023 ± 0.000	**0.005 ± 0.000**	0.183 ± 0.0	0.747 ± 0.016

Table 1: Disentanglement metrics with standard deviation for different models over 10 evaluation runs each. Values are presented as mean ± standard deviation.

2.2 Modeling

We employ four VAE-based models to disentangle SynTone's ground truth factors: vanilla VAE (Kingma & Welling, 2013), β-VAE (Higgins et al., 2017), Factor-VAE (Kim & Mnih, 2018), and β-TCVAE (Chen et al., 2018). The use of VAEs offers interpretable latent spaces, and these model variants introduce objectives specifically targeting greater independence among latent factors. SynTone's rich ground truth provides a basis for quantifying the success of recovering explicitly defined variations. Our focus lies on testing the explanatory disentanglement capabilities afforded by VAE techniques, in contrast to assessing the sample quality of less interpretable GANs (Chen et al., 2016).

To evaluate the disentanglement achieved by these different approaches, we apply various metrics that quantify each model's ability to recover ground truth factors, emphasizing *different desired properties* of disentanglement. These properties include *modularity*, where variation in one factor does not affect other factors (i.e., there is no causal effect between them); *compactness*, where a factor should be explained by a minimal subset of the latent space, ideally a single dimension; and *explicitness*, where learned codes should be semantically meaningful. Specifically, we use the Mutual Information Gap (MIG) (Chen et al., 2018) and Attribute Predictability Score (SAP) (Kumar et al., 2017) to measure compactness. We also use Joint Entropy Minus Mutual Information Gap (JEMMIG) (Do & Tran, 2019) and Disentanglement, Completeness, and Informativeness MIG (DCIMIG) (Sepliarskaia et al., 2019) to assess *holisticness*, as they capture modularity, compactness, and explicitness properties in a single score. Moreover, the code modularity is measured with the Modularity score (Ridgeway & Mozer, 2018). For a more in-depth analysis of disentanglement metrics, refer to Carbonneau et al. (2020).

3 Experiments

We trained four model architectures on SynTone: vanilla VAE, β-VAE, Factor-VAE, and β-TCVAE. To facilitate training, the audio data is transformed into a time-frequency domain, resulting in a 2D input for our networks (refer to appendix A.1). All variational autoencoders used the same convolutional architecture (detailed in appendix A.2), with variations only in the objective function. Extensive exploration of hyperparameter settings for the different models was conducted, and only the best-performing configuration was adopted for the final experiment. Table 1 provides a summary of these metrics for each model.

In our analysis, the vanilla VAE's superior performance in compactness metrics (MIG, SAP) and comparable DCIMIG to β-TCVAE was unexpected, especially as compactness is of lesser practical importance. Its simpler structure may better isolate factors in the SynTone dataset. Factor-VAE led in JEMMIG and Modularity, as expected from its training approach. The underperformance of β-VAE and β-TCVAE in certain metrics suggests their loss functions might not be ideally suited for SynTone, emphasizing the need for dataset-specific approaches in audio-disentangled representation learning. For deeper analysis, see the qualitative results in the appendix A.3.

4 Conclusion

Our work addresses the unexplored realm of disentangled representation learning in audio by introducing SynTone, a dataset with explicit ground truth factors. We evaluated state-of-the-art meth-

ods, revealing the challenges in achieving comprehensive disentanglement in audio representations. While the vanilla VAE demonstrates robust compactness, the Factor-VAE excels in modularity. However, limitations persist, emphasizing the need for further research in audio disentanglement. Our dataset proves valuable for method evaluation, though potential extensions to more diverse datasets and real-world scenarios warrant exploration. Our findings provide insights for future efforts in enhancing disentangled audio representation learning.

ACKNOWLEDGEMENTS

This work was funds of the research training group (RTG) in "Computational Cognition" (GRK2340) provided by the Deutsche Forschungsgemeinschaft (DFG), Germany, and an EU-Consolidator grant (772000, TurnTaking).

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of the ICLR 2024 Tiny Papers Track.

REFERENCES

Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Marc-André Carbonneau, Julian Zaidi, Jonathan Boilard, and Ghyslain Gagnon. Measuring disentanglement: A review of metrics. *arXiv preprint arXiv:2012.09276*, 2020.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*, pp. 2615–2625, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations. *arXiv preprint arXiv:1908.09961*, 2019.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations (ICLR)*, 2017.

Wei-Ning Hsu, Bernd Bolte, Yao-Hung Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Unsupervised learning for sequential speech separation in noisy environments. *IEEE/ACM transactions on audio, speech, and language processing*, 25(5):1098–1112, 2017.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658. PMLR, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

Francesco Locatello, Gabriele Abbati, Tom Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *International conference on machine learning*, pp. 4114–4124, 2019.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Kaizhi Mo, Jimmy Ba, Xuchen Zheng, Zhenyu Liu, Wei Xie, and Takuya Yoshioka. Semantic speech separation using an unsupervised spatial feature. In *ICASSP*, pp. 7205–7209, 2019.

Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. *Advances in neural information processing systems*, 31, 2018.

Anna Sepliarskaia, Julia Kiseleva, and Maarten de Rijke. How to not measure disentanglement. *arXiv preprint arXiv:1910.05587*, 2019.

A APPENDIX

A.1 MODEL INPUT PREPROCESSING

The input representation to each model is a mel-spectrogram transformation with a window size of 2024 and a hop size of 512, aligning with speech processing best practices. Mel-spectrograms efficiently encapsulate audio details into perceptually relevant frequency bands. The encoder outputs parameterize a multivariate Gaussian distribution over latent variables, guided by a β hyperparameter during training for incentivizing disentangled dimensions. Supervised by $\mathbb{T}$, $\mathbb{A}$, and $\mathbb{F}$ generative factors, designated latent dimensions are manipulated post-convergence to control amplitude, timbre, and frequency in generated samples. Performance assessment includes quantitative and qualitative evaluations of disentanglement metrics, reconstructions, and sample quality respectively.

A.2 MODEL ARCHITECTURE

Our VAE model applies four convolutional layers to extract hierarchical spatial features from input mel-spectrograms, followed by two fully connected layers to parameterize a Gaussian latent code capturing explanatory factors of variation. The decoder mirrors this pathway using transpose convolutions to recover the original input dimensions from sampled codes. All layers use leaky ReLU activations (slope=0.01) for faster convergence, with no batch norm for clearer disentanglement analysis. We strategically upsample in the decoder to balance reconstruction quality without overparameterization. This convolutional VAE structure provides an interpretable latent space to analyze captured generative factors while modeling complex mel-spectrogram dependencies. By parameterizing the distributions, we additionally obtain useful uncertainty information.

Building upon this, the full implementation details, including the architecture, training procedures, and hyperparameters used in our VAE models, are meticulously documented. This comprehensive resource is designed to assist in replicating or building upon our work. The provided link offers access to our codebase VAE Models Full Implementation.

A.3 RECONSTRUCTION AND SAMPLING

We visually assess models' performance through reconstructions of held-out mel-spectrograms and novel samples synthesized by decoding random codes. Reconstructions indicate how effectively the data compression is captured. Sampling assesses generative modeling abilities when guided solely by the learned latent distributions. Comparing the similarity of reconstructions and coherence of samples provides intuitive insight into encoding quality between models, complementing metric-based evaluation.

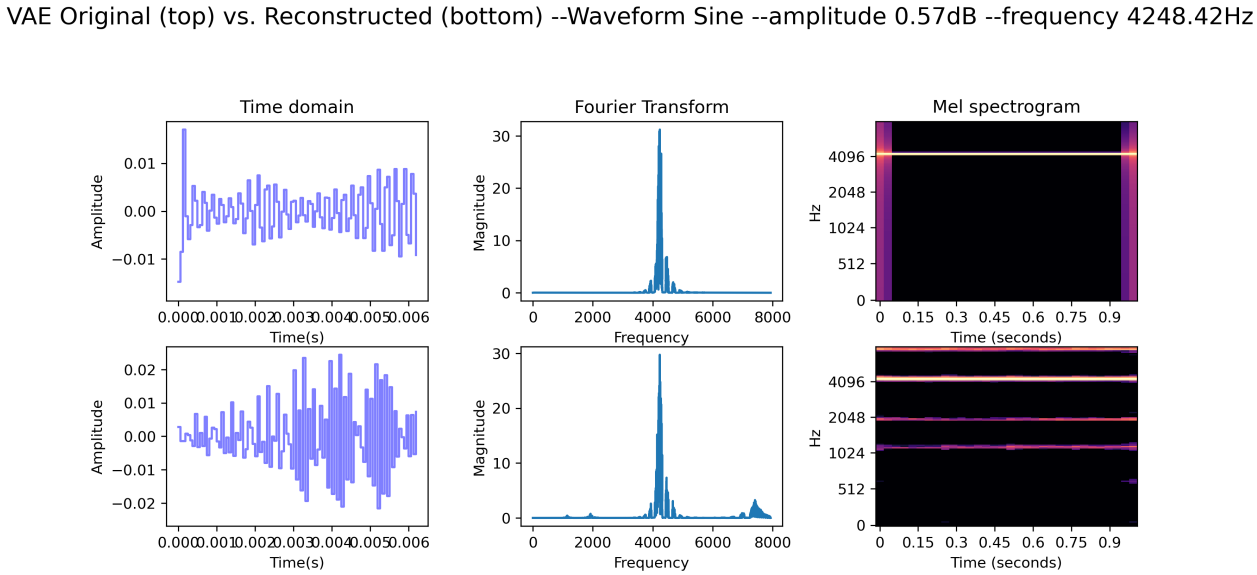

Figure 1: Time-domain representation (left), Fourier transforms (middle), and Time-Frequency (Mel-spectrogram) representation (right) for original and reconstructed samples using the VAE model. This model shows a lower reconstruction quality as indicated in all plots due to the addition of noise frequencies around 7000Hz as well as lower harmonic partials as portrayed in the mel-spectrogram.

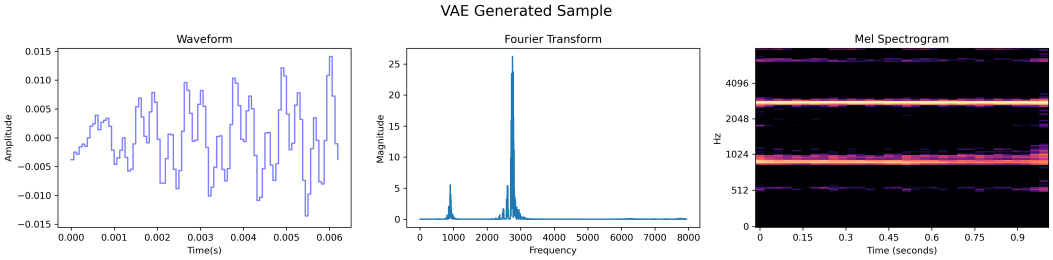

Figure 2: Visualization of generated audio samples using the VAE model. The left column depicts the time-domain representation, the middle column illustrates the Fourier transforms, and the right column showcases the Time-Frequency (Mel-spectrogram) representation. This plot shows that the model can isolate is single major frequency band close to 3000Hz as illustrated by the Fourier domain periodogram.

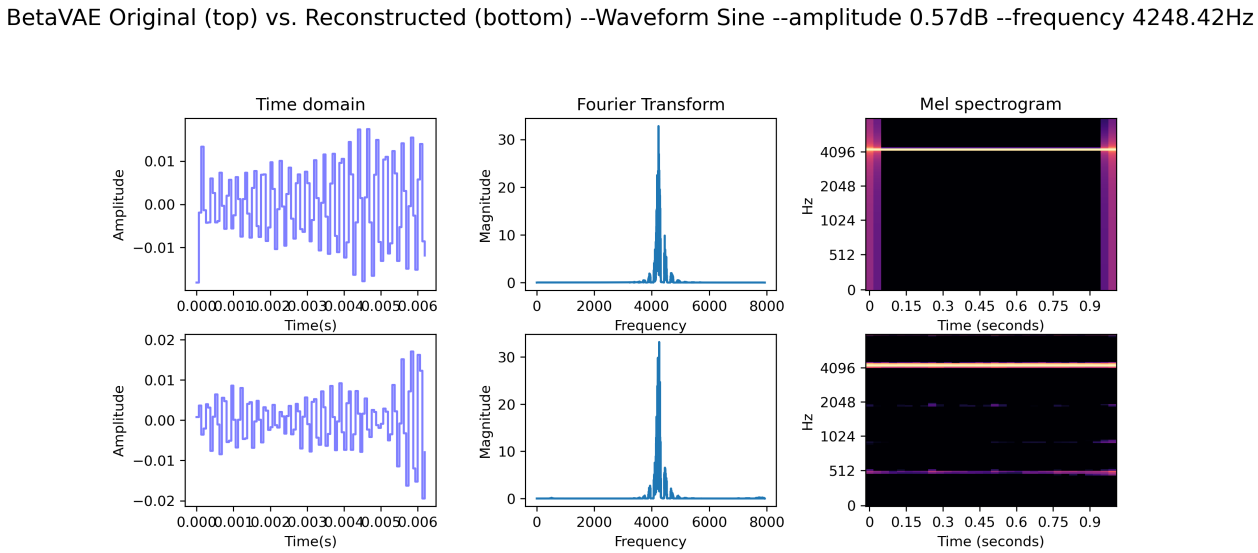

Figure 3: Time-domain representation (left), Fourier transforms (middle), and Time-Frequency (Mel-spectrogram) representation (right) for original and reconstructed samples using the β-VAE model. This model achieved a suitable reconstruction quality compared to the vanilla VAE as shown in both the Fourier space where the frequency factor of 4248.42Hz is correctly isolated.

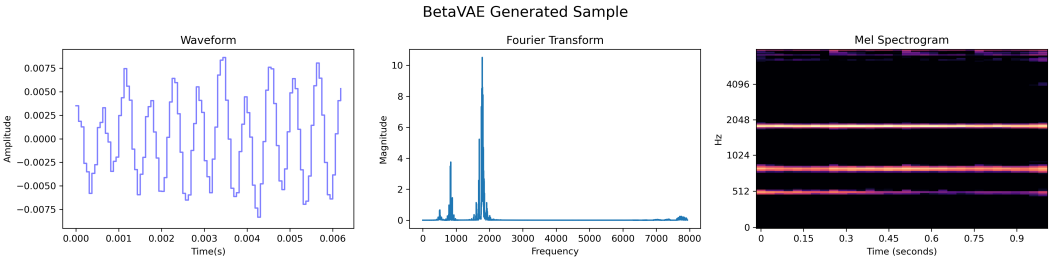

Figure 4: Visualization of generated audio samples using the β-VAE model. The left column depicts the time-domain representation, the middle column illustrates the Fourier transforms, and the right column showcases the Time-Frequency (Mel-spectrogram) representation. This figure, however, shows minor harmonic partials close to 1000Hz while the major frequency component is also predominant in both the periodogram and mel-spectrogram plots.

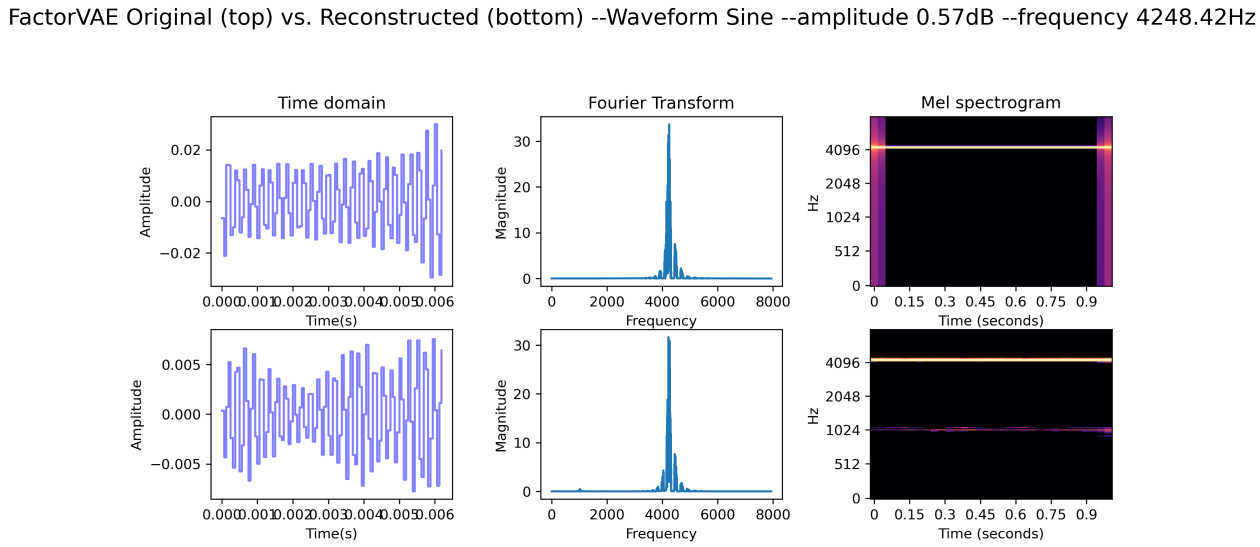

Figure 5: Time-domain representation (left), Fourier transform (middle), and Time-Frequency (Mel-spectrogram) representation (right) for original and reconstructed samples using the FactorVAE model. This model achieved the best reconstruction quality for an assessment of audio representation in both the Fourier space where the frequency factor of 4248.42Hz is correctly isolated.

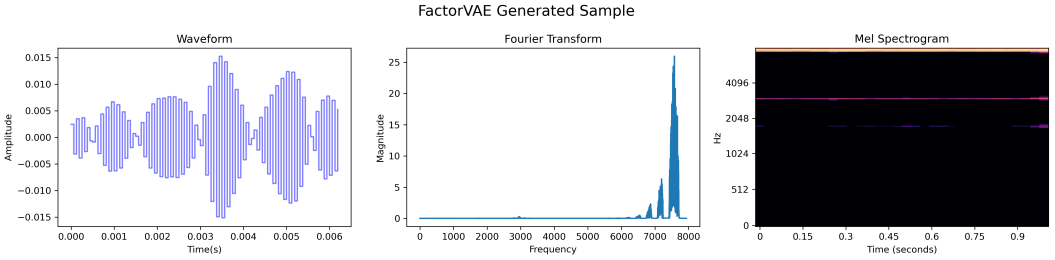

Figure 6: Visualization of generated audio samples using the Factor-VAE model. The left column depicts the time-domain representation, the middle column illustrates the Fourier transforms, and the right column showcases the Time-Frequency (Mel-spectrogram) representation. The model can decode a complex latent code that is smoothly represented in both the Fourier plot and mel-spectrograms with minimal noise levels.

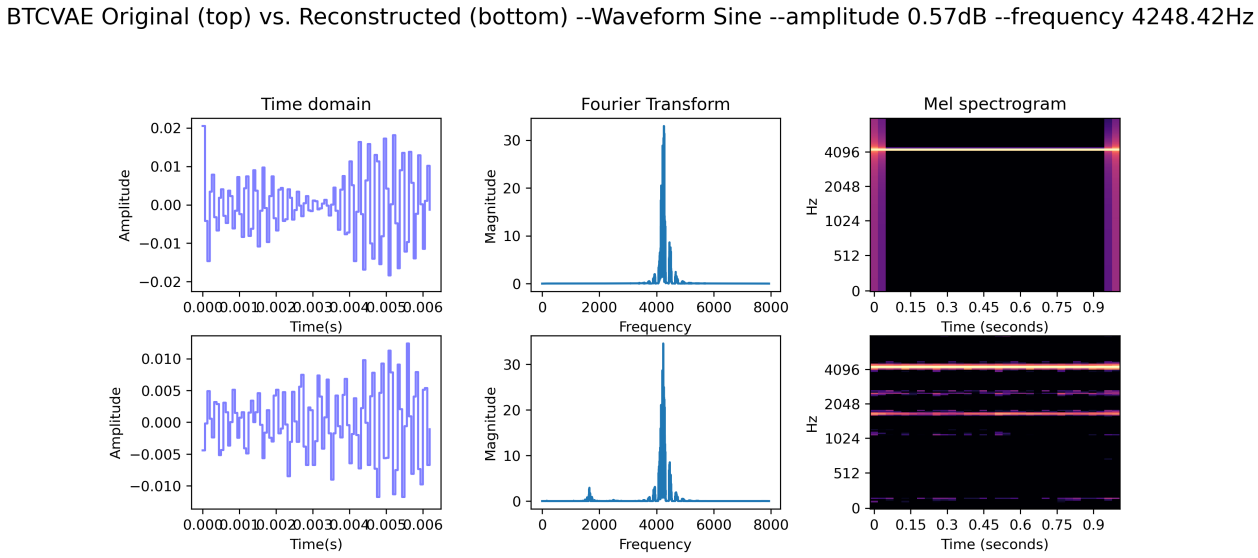

Figure 7: Time-domain representation (left), Fourier transforms (middle), and Time-Frequency (Mel-spectrogram) representation (right) for original and reconstructed samples using the β-TCVAE model. Like the vanilla VAE, this model also suffers from low-frequency noise in encoding and decoding the audio signal.

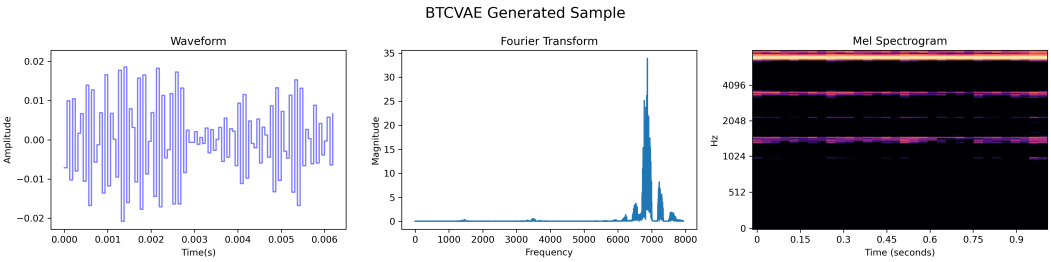

Figure 8: Visualization of generated audio samples using the β-TCVAE model. The left column depicts the time-domain representation, the middle column illustrates the Fourier transforms, and the right column showcases the Time-Frequency (Mel-spectrogram) representation. Like the FactorVAE decoder, this model is also able to decode complex waveforms however, it has more predominate noise levels as indicated in the plots where minor frequency contributions are visible.

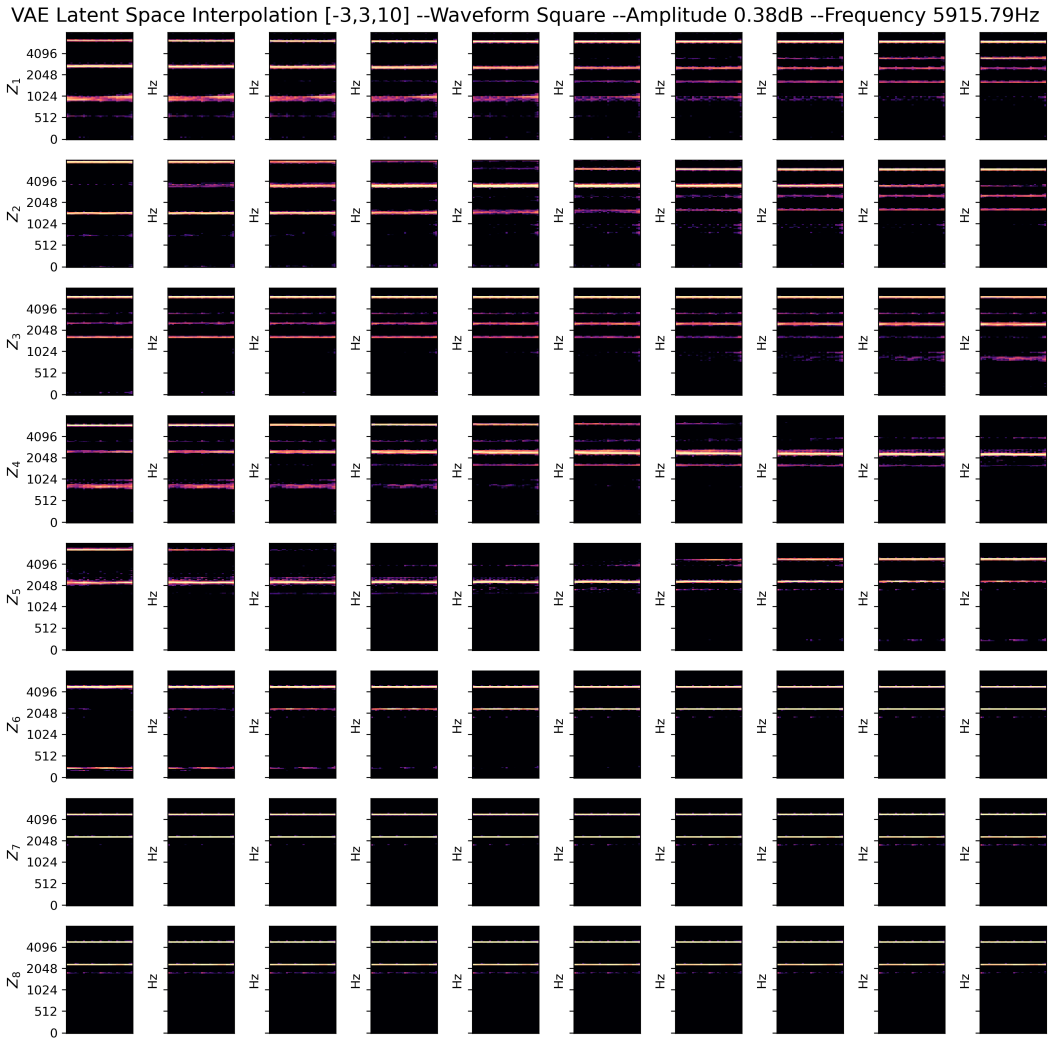

Figure 9: Latent space interpolation for the vanilla VAE model, showcasing variations in a single latent code while others are fixed. Each row represents 10 discrete samples per latent code.

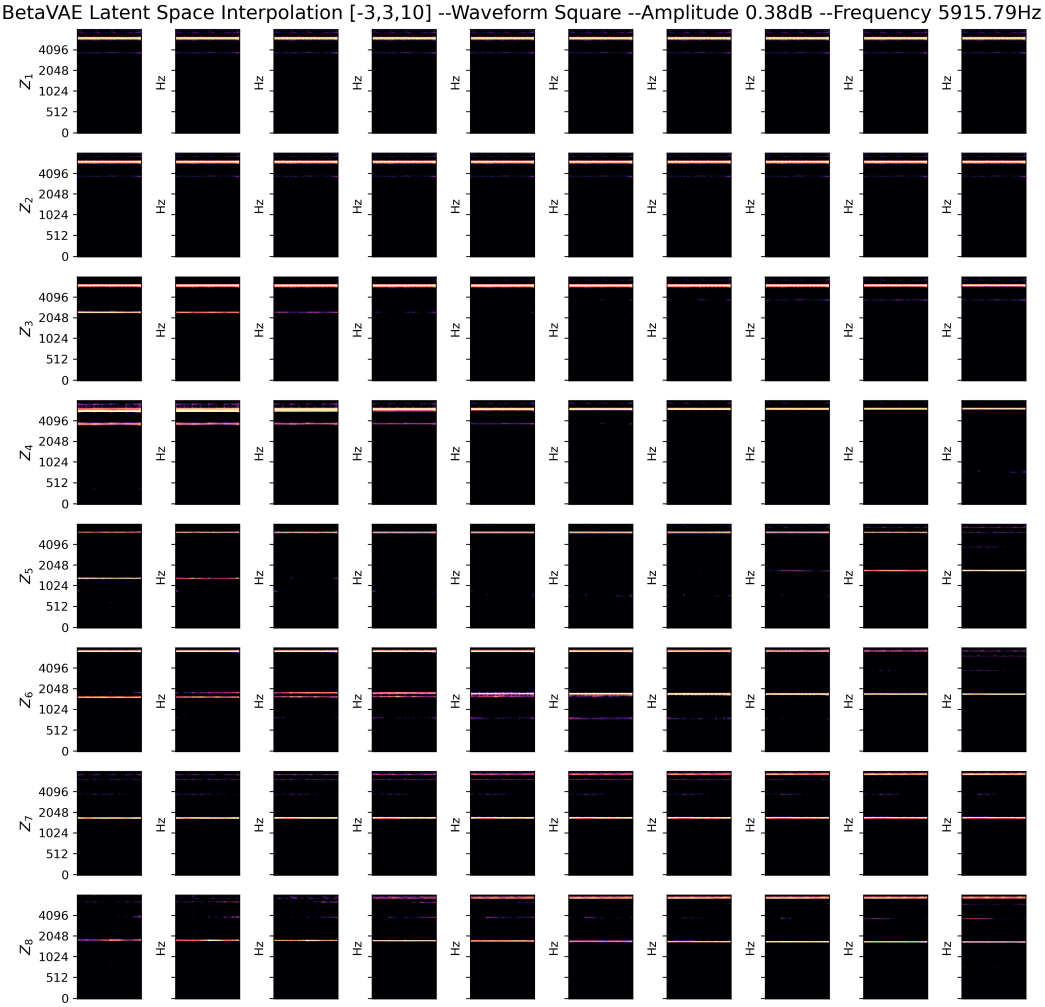

Figure 10: Latent space interpolation for the β-VAE model, showcasing variations in a single latent code while others are fixed. Each row represents 10 discrete samples per latent code.

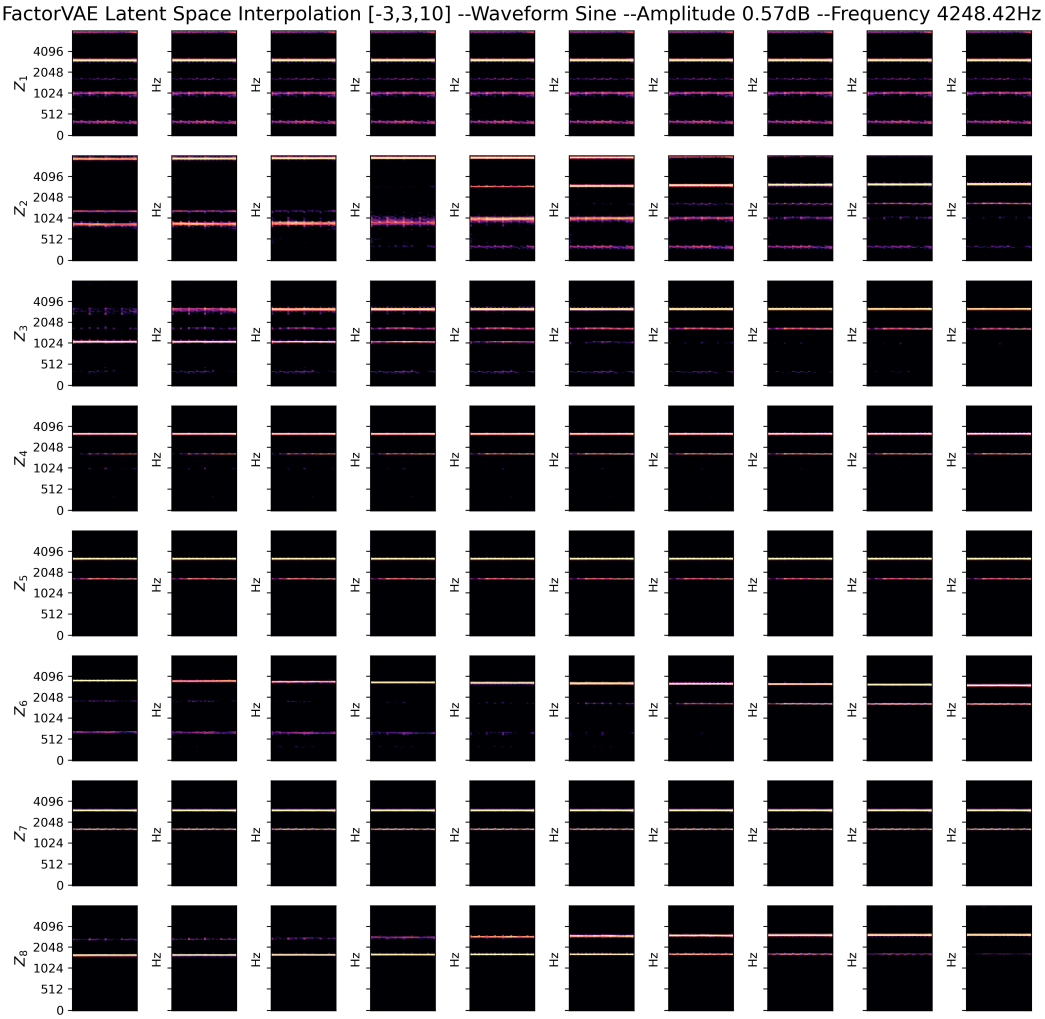

Figure 11: Latent space interpolation for the Factor-VAE model, showcasing variations in a single latent code while others are fixed. Each row represents 10 discrete samples per latent code.

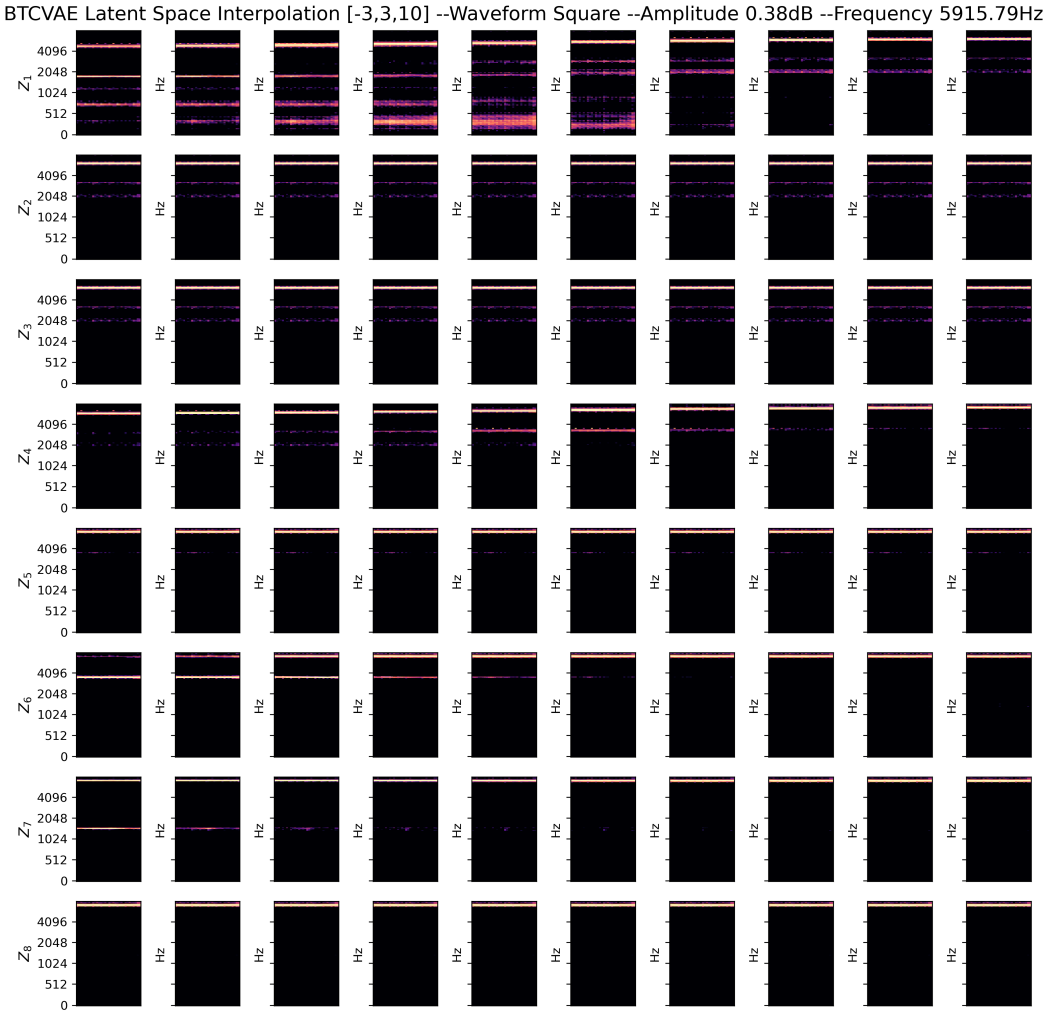

Figure 12: Latent space interpolation for the β-TCVAE model, showcasing variations in a single latent code while others are fixed. Each row represents 10 discrete samples per latent code.