2023-05-22

An Exploration of the Latent Space of a Convolutional Variational Autoencoder for the Generation of Musical Instrument Tones

Anastasia Natsiou
Technological University Dublin, anastasia.natsiou@tudublin.ie

Sean O'Leary
Technological University Dublin, sean.oleary@tudublin.ie

Luca Longo
Technological University Dublin, luca.longo@tudublin.ie

An Exploration of the Latent Space of a Convolutional Variational Autoencoder for the Generation of Musical Instrument Tones

Anastasia Natsiou[1], Seán O'Leary[1], and Luca Longo[1]

School of Computer Science,
Artificial Intelligence and Cognitive Load Research Lab,
Technological University Dublin,
Dublin, Republic of Ireland
{anastasia.natsiou, sean.oleary, luca.longo}@tudublin.ie

Abstract. Variational Autoencoders (VAEs) constitute one of the most significant deep generative models for the creation of synthetic samples. In the field of audio synthesis, VAEs have been widely used for the generation of natural and expressive sounds, such as music or speech. However, VAEs are often considered black boxes and the attributes that contribute to the synthesis of a sound are yet unsolved. Existing research focused on the way input data can influence the generation of latent space, and how this latent space can create synthetic data, is still insufficient. In this manuscript, we investigate the interpretability of the latent space of VAEs and the impact of each attribute of this space on the generation of synthetic instrumental notes. The contribution to the body of knowledge of this research is to offer, for both the XAI and sound community, an approach for interpreting how the latent space generates new samples. This is based on sensitivity and feature ablation analyses, and descriptive statistics.

Keywords: Explainable Artificial Intelligence (XAI) · Variational Autoencoders (VAE) · Audio Representations · Audio Synthesis · Latent Feature Importance.

1 Introduction

Generative models in the field of sound synthesis have enabled musicians, and sound designers to create and manipulate sounds in new and innovative ways [20, 32]. Variational Autoencoder (VAE) is a type of deep generative model that has been widely used in the field of audio generation [12, 15, 26]. VAEs can be trained on a dataset of sound recordings to learn a compressed representation of the sounds in the latent space. The latent space can then be manipulated to generate new, similar sounds. The design of VAEs is considered successful when samples with similar principles map closer to each other in the latent space, and new sounds can be generated by interpolating between previous sounds with specific properties. However, this is not always the case. The model extracts

information from the audio samples forming a latent space that does not always match human perception. Explainable artificial intelligence (XAI) is an emerging area of research that aims to develop techniques to make deep learning models transparent and interpretable to humans [4, 37, 38]. Explainability can be used in generative models for sound synthesis for the creation of audio samples in a more understandable and controllable way. However, according to [29], XAI for arts, and more specifically generative music, is still in its early stages. Most of the existing publications focus on controlling the synthesis of the generative model by regularizing its latent space to specific characteristics [23, 36, 40].

In this manuscript, we attempt to explain how latent dimensions are linked to the generation of discreet instrumental musical tones for monophonic, and harmonic audio samples. First, we provide explanation techniques for understanding the impact of different parameters on the resulting sound using VAEs, and then we explore the contribution of each attribute of the latent space to the synthesis of a specific instrument. Section 2 overviews the existing methodology used in this manuscript while briefly providing a literature review on the most prominent techniques. The chapter provides a concise overview of the prevalent audio representations commonly used and our proposed representation, along with an explanation of the architecture of VAE. Additionally, it delves into the examination of statistical methods and visualization techniques employed to uncover the significance of latent attributes. Through careful analysis, these approaches enable the identification and assessment of the importance of latent variables in the context of audio synthesis. Section 3 describes the dataset and hyper-parameterization of the VAE, along with evaluation methods for sound reconstruction. Finally, the results are reported in Section 4 and the conclusion and future directions in Section 5.

2 Methodology and Related Work

Our goal is to permit the manipulation of latent representations of discreet instrumental notes created by deep generative models and provide an explanation for the synthesis of musical tones. In Chapter 2.1, we demonstrate the importance of input representations for sound synthesis followed by the development of a new representation for monophonic and harmonic sounds. Chapter 2.2 outlines the benefits of Variational Autoencoders compared to the classic autoencoders and introduces ways to create the latent space. The proposed methodology for the generation of the latent space is illustrated in Fig. 2. Finally, Chapter 2.3 provides techniques for the interpretation of VAEs based on methods for latent feature importance.

2.1 Audio Representations

Audio representations are mathematical representations of acoustic signals that are used to analyze, process, and manipulate sound. The fundamental form of an acoustic signal is the discretized version of the waveform which is created by

sampling the continuous wave in time and amplitude. In deep learning applications, this waveform is called raw audio and it represents the acoustic wave as a sequence of numbers, each number representing an amplitude sample at a chosen sampling frequency. Although raw audio is the most accurate representation of sound, it is often considered unsuitable for deep learning models due to its high dimensionality and lack of interpretability.

To overcome these obstacles, recent studies propose high-level forms that offer more meaningful descriptions [27]. Time-frequency representations such as spectrograms, mel-spectrograms, or Constant-Q Transformations (CQT) have been proven beneficial for deep generative models [2, 3, 35]. Their success is mainly achieved because of their ease to be stored and processed. In deep learning models, where memory and computational limitations can slow down the training process, time-frequency representations provide an efficient solution. Finally, spectrograms provide a visual representation that captures important characteristics such as frequency content, harmonics, timbre, and temporal dynamics. They provide a physically and perceptually meaningful representation making them more useful for deep generative models [13]. An overview of spectrograms is depicted in Fig. 1. In an attempt to reduce even more the wealth of acoustic information, various studies extract perceptual features from the original signal. Compact representations such as acoustic features [11] or spectral coefficients [14] capture the essential spectral information of the audio signal while reducing the amount of data required for analysis and processing. Spectral coefficients can be used to represent the spectral envelope of an audio signal, which contains information about the shape of the frequency spectrum and the relative amplitudes of the various frequency components.

In this work, we design a new audio representation based on acoustic features that is able to represent monophonic, and harmonic instrumental notes. The proposed representation is created by the fundamental frequency and the logarithm of the amplitudes of the first 7 harmonics in overlapping frames of sound. The first 7 harmonics are able to capture the most perceptually significant part of a note providing information about the spectral shape of the acoustic signal. Since the dataset is monophonic and harmonic, the waveform can be synthesized with the given amplitudes and the integer multiples of the fundamental frequency:

$$f_n = nf_0 \tag{1}$$

where $n \in [1, 7]$ represents the number of the harmonic for every frame i of the sound. The suggested representation offers a lower-dimensional alternative to spectrograms, enabling further compression by autoencoders. Moreover, its meaningful structure is expected to result in a latent space that is more interpretable, allowing for a better understanding and analysis of the learned representations.

2.2 From Autoencoders to Variational Autoencoders

An autoencoder [5] is a type of neural architecture used for unsupervised learning. The fundamental idea behind an autoencoder is to learn a compressed rep-

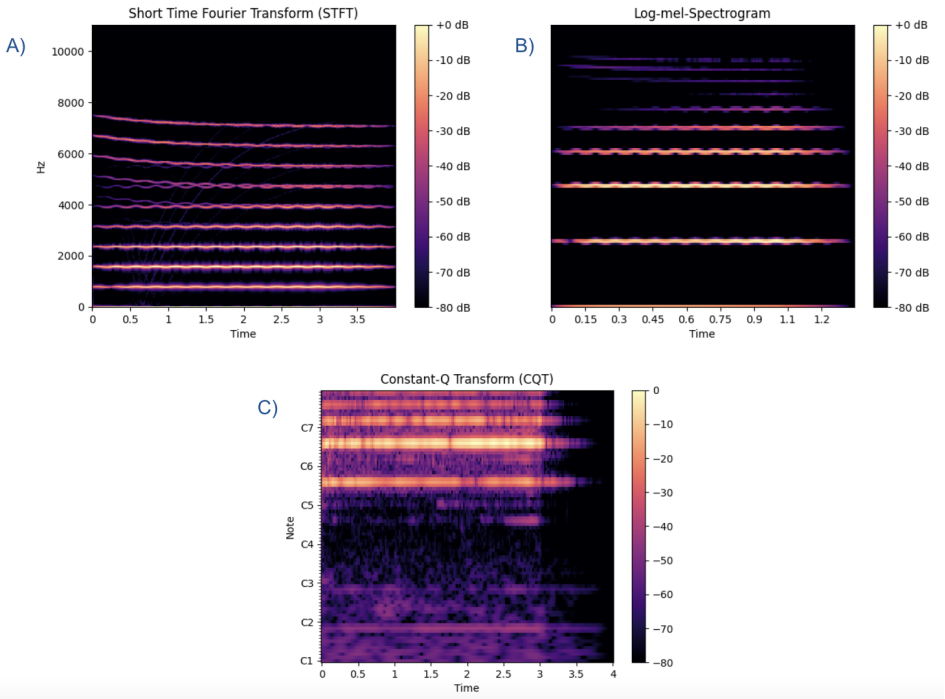

Fig. 1. Demonstration of time-frequency representations of sound: A) Magnitude spectrogram (using STFT) B) Log-mel-spectrogram and C) Constant-Q Transform (CQT)

resentation of the input data that captures its most important features, in order to reconstruct the original input data as accurately as possible. Autoencoders consist of two main parts: an encoder that attempts to reduce the dimensionality of the original samples producing a compressed or *latent representation*, and a decoder that aims to reconstruct the original input data given the latent representation. For a feedforward single layer model, the encoder maps an input vector $x \in \mathbb{R}^d$ to an encoding $z \in \mathbb{R}^e$ where $d > e$ using a non-linear activation function $f(.)$

$$y = f(Wx + b) \tag{2}$$

where $W \in \mathbb{R}^{(e \times d)}$ represents the weights of the connections between the neurons and $b \in \mathbb{R}^e$ accounts for the bias term. The decoder maps z back to the reconstructed $\widehat{x} \in \mathbb{R}^d$ using a similar approach

$$\widehat{x} = f(W_{out}y + b_{out}) \tag{3}$$

where $W_{out} \in \mathbb{R}^{(d \times e)}$ and $b_{out} \in \mathbb{R}^d$. The reconstruction is achieved by a training procedure where the autoencoder attempts to minimize the difference between

the input data and the reconstructed output data, using a loss function such as mean squared error or binary cross-entropy.

Although autoencoders can achieve a low reconstruction error, these types of architectures do not promise a meaningful learned representation. The latent space created by an autoencoder often lacks interpretability and similar samples can be mapped at different regions in the space. To solve this problem, Variational autoencoders (VAEs) [18] convert the input data to a latent space through a stochastic distribution, making it more "smooth" and interpretable. In the vanilla VAE, the encoder maps a latent variable $z \in \mathbb{R}^e$ with an input variable $x \in \mathbb{R}^d$ where $d > e$ by using a distribution $q(z|x)$ to approximate $p(z|x)$. This approximation is named *Variational Inference*. The prior distribution of z is $p(z)$ and therefore the decoder is parametrized to approximate the distribution $p(x|z)$. More specifically, the encoder outputs the mean μ_M and the covariance σ_M as the inputs of the Gaussian distribution function $N(z; \mu_M, \sigma_M^2 I)$ over a latent space with M number of dimensions. The objective of the network is to minimize the KL divergence between $q(z|x)$ and $p(z)$ by maximizing the evidence lower bound:

$$\mathbb{E}[\log p(x|z)] - KL(q(z|x) \parallel p(z)) \leq \log p(x) \tag{4}$$

In the above equation, the first term measures how well the reconstructed data matches the original, while the second term measures the difference between the approximate posterior distribution $q(z|x)$ and the prior distribution $p(z)$. The KL divergence term encourages the learned posterior distribution to be close to the prior distribution, which in turn regularizes the learned latent space representation. By sampling from a distribution, VAEs have the ability to generate new data samples that are similar to the input data. Furthermore, the generated latent space is regularized by a specific distribution making it more structured and continuous. This means that similar input data will be closer in the latent space and small changes in the latent space correspond to small changes in the generated data. Finally, interpolation between two points in the latent space is feasible. This way, a new sample can be created by combining the properties of the multiple original data. VAEs have demonstrated their potential in multiple audio applications including real-time synthesis [7], polyphonic synthesis [21], or instrumental tones generation controlled over their pitch [33].

2.3 Latent Feature Importance

In deep generative models, such as VAEs, latent variables can play a critical role in capturing complex relationships in the data and generating high-quality outputs. However, these variables are not directly observable and the task to interpret their role in the model can be challenging [1]. Latent feature importance refers to the relative importance of hidden or latent variables, as opposed to the input or output data [8]. For the investigation of latent feature importance, a variety of techniques have been proposed. *Sensitivity analysis* involves the process of perturbing the values of the latent features and observing the resulting

changes in the output variables. By measuring the reconstruction error of the generated data when a slight variation of the latent feature is applied, we gain insights into which latent features are more important for producing the correct output. Gradient-based investigation of each feature of the latent space [28] constitutes a prominent technique for sensitivity analysis of the VAEs.

Feature ablation involves the procedure of removing individual latent features from the model and measuring the resulting changes in the output variables. In VAEs, feature ablation studies can be conducted to calculate and visualize the encoded samples for the most significant parts of the latent space that lead to the generation of synthetic data [19]. In a similar approach, *feature attribution techniques* involves the process of assigning a score to each latent feature based on its contribution to the final output. Several methods have been proposed for computing feature attribution scores, including Local Interpretable Model-Agnostic Explanations (LIME) [30], SHapley Additive exPlanations (SHAP) [22], and Integrated Gradients [34]. A more contemporary method for quantifying the reliance of a model on each feature is the Shapley Additive Global importancE (SAGE) [9]. This approach assigns a score to each feature based on five desirable properties, namely efficiency, symmetry, dummy, monotonicity, and linearity.

Other methods for experimenting with latent feature importance include a variety of clustering techniques. *Clustering* involves grouping similar points in the latent space together based on their output variables. Clustering has been used as an integrated method of the architecture of VAEs to regularize the latent space based on existing distributions [16, 31, 41] or as a regularization method based on specific characteristics of the training data [6, 29]. Finally, valuable insights into the importance of the latent features can be gained through visualization of the latent variables. However, since the latent space usually indicates a high dimensionality, visualization can be achieved by projecting the high-dimensional latent space onto a lower-dimensional space that can be easily plotted and interpreted. Two of the most popular visualization techniques are the Principle Component Analysis (PCA) [25] and the two- or three-dimensional Stochastic Neighbor Embedding (t-SNE) [24]. In this manuscript, we investigate many of the above methods to analyze and understand the way latent features contribute to the generation of synthetic instrumental notes. Our experimentation includes a sensitivity analysis along with a feature ablation analysis for the understanding of the influence of each feature of the latent space on the properties of the generated sound. Furthermore, we provide statistical analysis and visualization of the latent space for identifying the latent attributes related to each instrument.

3 Experimental setup

In this section, we describe the experimental setup, including details of the dataset and its pre-processing, model configuration, and evaluation methods for the reconstruction. Fig. 2 illustrates the overall reconstruction schema for

the extraction of the latent features of audio samples and the reconstruction of instrumental notes.

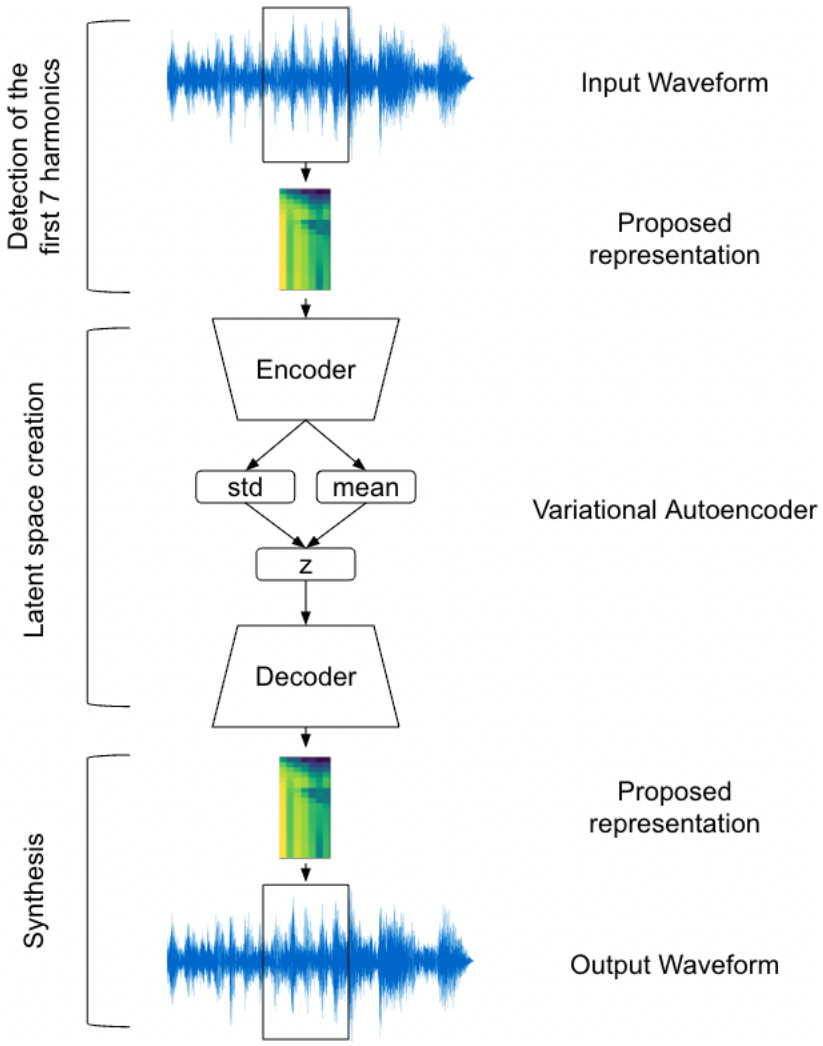

Fig. 2. Proposed architecture for the generation of the latent space using Variational Autoencoders and an audio representation based on the first 7 harmonics.

3.1 Dataset

The experiments were conducted using the NSynth dataset[1], a collection of four-second monophonic notes from a variety of instruments in acoustic, electronic, or synthetic form in different categories as per their velocity or acoustic quality. The training and testing data had a pitch in the range of 80Hz-2100Hz. Furthermore, an analysis of the samples revealed that many of the data were not harmonic and they were excluded from the experiments along with some samples that had variations in the fundamental frequency or amplitude. The remaining dataset is composed of 101911 training samples and 1324 testing samples of guitar, bass, brass, keyboard, flute, organ, mallet, reed, and string.

The waveform of these samples was pre-processed to generate a representation that includes the fundamental frequency and the logarithm of the amplitude of the first 7 harmonics for overlapping segments of sound. The fundamental frequency was computed using the YIN algorithm [10] with a post-processing step to ensure that the pitch will not vary more than 3% between consecutive frames. The amplitudes of the first 7 harmonics were calculated by a peak detection technique in the time-frequency domain. For the conversion of the waveform to the time-frequency domain, we used the Short Time Fourier Transform (STFT) with a normalized Blackman window of 690 samples, an FFT window of 1024, and a hop size of 172. The final representation was later normalized using the min-max scaling to be transformed into the range $[0, 1]$.

3.2 Model configuration

For the audio reconstruction, we used convolutional VAEs with a mirrored encoder and decoder as it is presented in Fig. 3. The two components are composed of two 2D convolutional layers with 32 filters each, a kernel size of 3, a stride of 2, and the same padding. Two dense layers are used to calculate the mean and variation and the latent space is the vector created after sampling from the distribution. Our experiments found an optimum of 8, for the dimensionality of the latent space, after which the performance of the reconstruction decreased. The ReLU is used as an activation function for the convolutional layers while the softmax function is applied to the output layer to form the generated normalized representation. The network is trained using the ADAM optimizer [17] with an initial learning rate of 0.001 in batches of size 128. For the reconstruction loss, we use binary cross-entropy, and an early stopping patience limit is set equal to 20 to avoid wasting resources during training. Additional regularization techniques such as dropout, or L1 and L2 regularization did not improve the quality of the generated samples. All models were implemented using the TensorFlow library[2] on a Tesla P100 GPU.

[1] https://magenta.tensorflow.org/datasets/nsynth
[2] https://www.tensorflow.org/

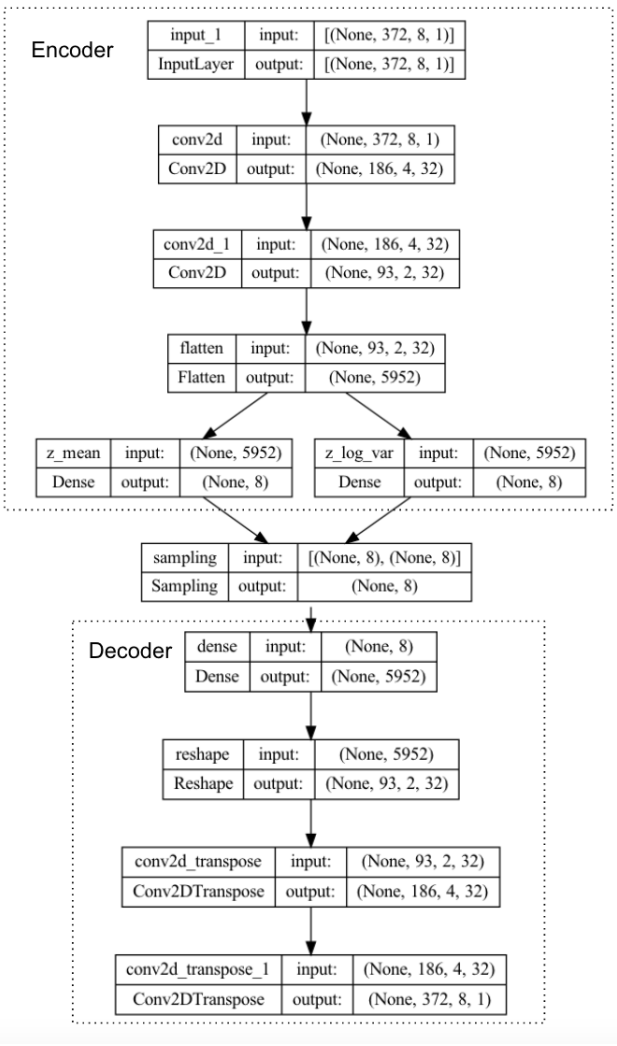

Fig. 3. The architecture of the Convolutional Variational Autoencoder VAE. Parameters and dimensions of the audio representation for every step of dimensionality reduction.

3.3 Reconstruction Quality

We evaluated the reconstruction capacity of the VAE using the Mean Squared Error (MSE) between the original and generated audio representation of the fundamental frequency plus the first 7 harmonics. However, MSE is not always sufficient as a metric for the reconstructed samples since it measures the pixel-wise difference between the original and reconstructed images, without taking into account the high-level structure and semantics of the data. Therefore, we

additionally computed the Structural SIMilarity (SSIM) [39] between the two representations. SSIM is a method for computing the structural similarity between two images, which takes into account the luminance, contrast, and structural information. The SSIM index ranges between -1 and 1, with values closer to 1 indicating higher similarity between the two images.

4 Results and Discussion

To interpret the way VAEs generate new samples, we conducted two types of experiments. In the first set of experiments, we performed a global analysis measuring the reconstructive capabilities of each attribute of the latent space. The second type of experiment is an attempt of interpreting the contribution of each latent feature to the synthesis of musical notes from a specific instrument.

4.1 Global Analysis

The goal of this section is to provide experimental results on the generation of instrumental notes using VAEs. It is also an attempt to analyze and interpret the latent space of the generative network. As it is illustrated in Fig. 3, the encoder projects the audio samples into a high-level representation by sampling from the distribution using the mean and the standard deviation predicted for each attribute. Then, the decoder uses as input the generated latent features trying to reconstruct the original samples. The ability of the network to generalize and create new samples that are similar to the original is affected by many parameters. One of the most important parameters is the size of the latent space. A relatively small latent space can increase computational efficiency while improving the ability to interpret its results. In our architecture, by decreasing the number of latent parameters, we concluded to a latent space with a dimensionality of 8 since it provided the right balance between the size of the latent space and the reconstruction error. The trained network achieved an average MSE of 0.039 and an average SSIM of 0.948 across all the samples of the testing dataset. To gain some knowledge of the latent features, we initially depicted the distribution of a testing dataset of notes for a variety of instruments along with some statistical information. Fig. 4 illustrates the boxplot of all possible values for the 8 attributes of the latent space. According to this figure, the second and third attributes have more clustered data around the median obtaining fewer possible values for the instrumental notes. Continuing with the interpretation of the latent space, we conducted a feature ablation analysis. In this set of experiments, we investigated the importance of each attribute of the latent space by enabling each time a single feature and measuring the SSIM with the original samples. The results from this analysis are demonstrated in Fig. 5 while the reconstructed images by permitting only one feature at a time are illustrated in Fig. 6. Based on these experiments, the first three attributes of the latent space demonstrate higher similarity with the original samples implying higher importance for the reconstruction of new samples.

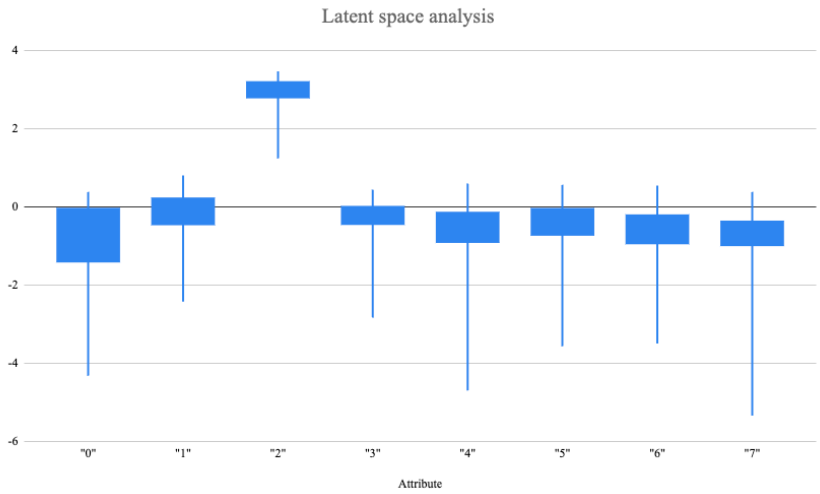

Fig. 4. Boxplot of the attribute analysis of the latent space. It illustrates all possible values of the 8 latent attributes.

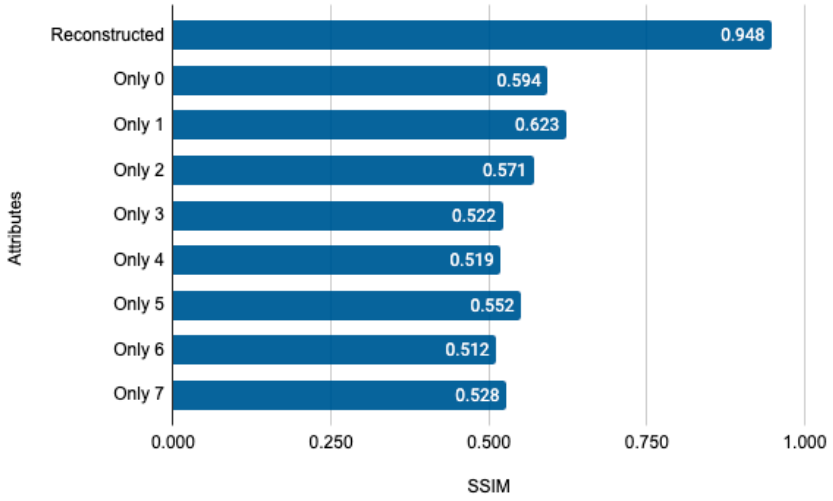

Fig. 5. SSIM between the original samples and the samples generated by only one enabled attribute.

A final examination of the latent space of the VAE covers a sensitivity analysis by applying the method of perturbation and observing the resulting changes in the synthesized samples. More specifically, we modified each attribute by 10%, 20%, 30%, 40%, 50%, and 60% respectively, and measured the percentage

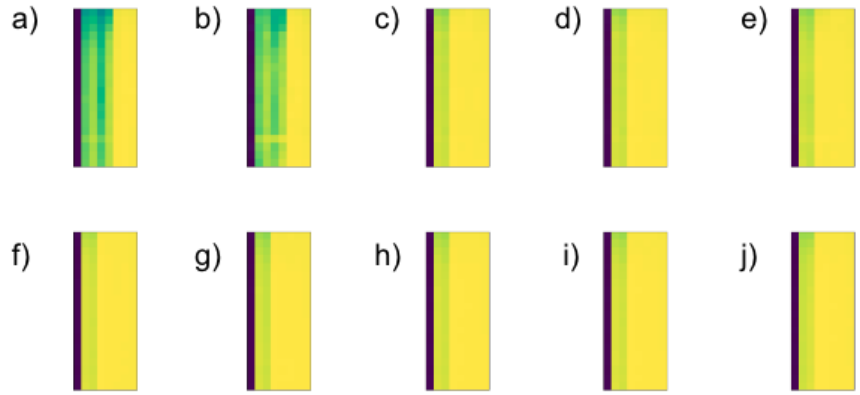

Fig. 6. Visualization of the reconstructed representation using one feature at a time. a) Original representation. b) Reconstructed representation. c) Only the "0" attribute. d) Only the "1" attribute. e) Only the "2" attribute. f) Only the "3" attribute. g) Only the "4" attribute. h) Only the "5" attribute. i) Only the "6" attribute. j) Only the "7" attribute.

of SSIM decrease. The results of these experiments are depicted in Fig. 7. This analysis provides information about the sensitivity of each attribute, pointing out that features one, two, and five are less resilient to change. The conducted global analysis provided information about the overall behavior of the model and the importance of the features of the latent space. It showed that the attributes do not have the same importance and do not contribute equally to the generation of new samples. Finally, it proved that the latent features do not have the same resilience to change. Most of the analyses conducted resulted in similar results proving that the attributes "0", "1", "2", and "5" play the most significant role in the reconstruction of the sound.

4.2 Instrument-based Analysis

In this section, we provide an instrument-based analysis for analyzing sound signals to identify the underlying musical instrument or instruments that produced the sound from the latent representation. Notably, we investigate the association of latent features with musical instruments. In order to do that, we statistically analyzed the latent space separately for every instrument class to address the significance of each latent feature for each instrument. Table 1 provides the mean and the standard deviation of every attribute for each instrument. It presents the variable range demonstrating that a specific range of some attribute can imply a particular instrument. For example, if the value of the feature "2" of the latent space is 4, the instrument will be more probably a flute or have some properties of flutes.

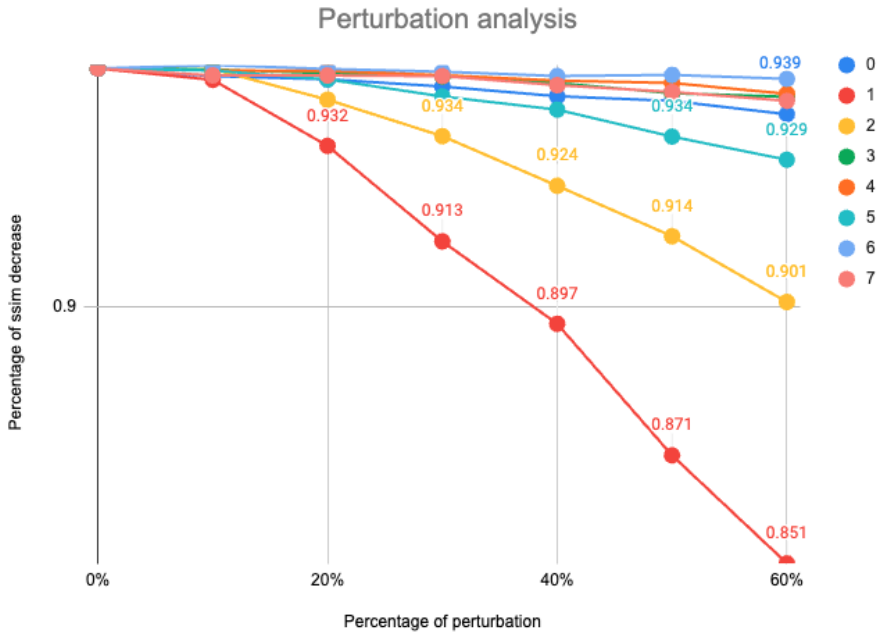

Fig. 7. Perturbation analysis of the latent attributes. Each line shows the percentage of the SSIM decrease by modifying each attribute from 0% to 60%.

Another way to investigate instruments from their latent features is by visualizing the latent space. Fig. 8 depicts the projection of the 8-dimensional latent space in the 2-dimensional space indicating every instrument with a specific annotation. For the dimensionality reduction, we used the t-SNE algorithm with PCA initialization and perplexity of 49. The visualization indicates that the model is able to learn high-level representations with respect to instruments since sounds from a specific instrument present a smaller distance in the latent space. However, some clusters, such as string and brass, are not completely disentangled. This could be due to either reducing the dimensionality of the data or the model lacking sufficient information to identify instrument-based features. Enhancing the precision of the generative model for synthesizing sound from a lower-dimensional space, or incorporating additional regularization methods that offer timbre information, may result in a latent space with clusters that are more spread out.

4.3 Contribution to the body of knowledge

The research presented in this manuscript has yielded a notable dual contribution to the fields of explainable AI and audio synthesis. Through a comprehensive series of experiments, the study introduces an analysis protocol that offers the

Table 1. Mean and standard deviation of each latent feature for every instrument class.

Class	"0"	"1"	"2"	"3"	"4"	"5"	"6"	"7"
Guitar	0.23 ± 0.56	-0.39 ± 0.74	3.32 ± 0.24	0.04 ± 0.58	0.3 ± 0.74	0.41 ± 0.72	-0.84 ± 1.05	0.14 ± 0.69
Keyboard	-0.26 ± 0.84	-0.34 ± 0.76	3.26 ± 0.24	-0.17 ± 0.86	0.29 ± 0.96	0.75 ± 0.73	-0.31 ± 0.91	-0.12 ± 0.91
Vocal	-2.49 ± 1.59	0.54 ± 0.89	2.71 ± 0.61	-0.12 ± 0.68	-0.3 ± 0.67	-0.73 ± 1.03	0.16 ± 1.05	-1.16 ± 1.39
Organ	-0.56 ± 1.37	0.32 ± 0.96	3.09 ± 0.65	0.05 ± 0.56	0.06 ± 1.06	-0.18 ± 0.77	0.15 ± 1.44	-0.03 ± 1.12
Bass	-1.11 ± 0.97	0.31 ± 1.48	2.77 ± 0.72	-0.38 ± 1.09	-0.3 ± 1.19	-0.25 ± 1.43	-0.31 ± 0.82	-0.53 ± 1.01
Brass	-0.55 ± 0.78	0.29 ± 0.85	2.98 ± 0.41	0.11 ± 0.62	-0.88 ± 1.01	0.21 ± 0.96	0.11 ± 0.9	-0.81 ± 0.89
Flute	-0.34 ± 0.73	1.01 ± 1.09	3.81 ± 0.86	0 ± 0.38	0.4 ± 1.07	0.06 ± 0.45	-0.69 ± 0.67	-0.13 ± 0.94
Reed	-0.1 ± 0.83	0.35 ± 0.78	3.14 ± 0.36	0.12 ± 0.73	-0.76 ± 1.02	-0.39 ± 1.2	-0.07 ± 0.83	-0.42 ± 0.87
String	-0.92 ± 0.8	0.14 ± 1.39	3.02 ± 0.22	0.36 ± 0.7	-0.65 ± 0.83	-0.45 ± 0.73	-0.17 ± 0.77	0.04 ± 1.06

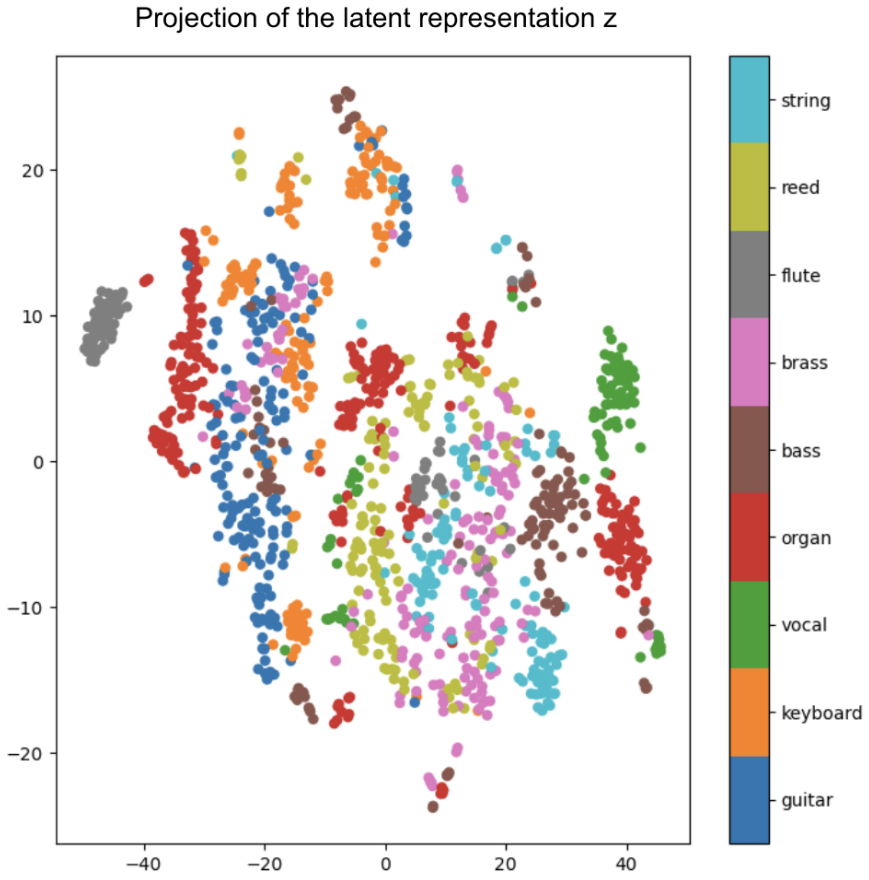

Fig. 8. Visualization of the projection of the latent representation in a 2-dimensional space using t-SNE.

XAI community a systematic approach to assess the significance of individual latent attributes in the generation of new samples using Convolutional Varia-

tional Autoencoders. Furthermore, it extends its impact to the realm of sound synthesis by investigating the underlying mechanisms employed by deep generative networks to create novel audio outputs. Explainability techniques such as sensitivity analysis and feature ablation methods are useful for comprehending how latent features contribute to generating synthetic samples. Moreover, statistical measures and visualization techniques can aid in distinguishing the way different classes of samples are represented in the latent space. Overall, this research provides valuable insights and practical knowledge, enriching the existing body of literature in both domains and advancing our understanding of explainable AI and audio synthesis.

5 Conclusion

Explainable artificial intelligence (XAI) methods offer promising tools for analyzing and interpreting the complex processes underlying sound synthesis. By providing greater transparency into the inner workings of deep learning models, XAI techniques can help researchers and musicians better understand the factors that contribute to the creation of sound, and how to optimize models to produce high-quality, diverse, and musically meaningful sounds. In this manuscript, we used XAI methods such as feature importance analysis, latent feature visualization, and sensitivity analysis to interpret the latent space of Variational Autoencoders (VAEs) for the synthesis of new musical notes produced by a specific instrument, which gives rise to its unique timbre and tonal quality. The conducted study pointed out that the attributes of the latent space do not contribute equally to the process of generating new samples, and demonstrated that the features of the latent space retain information about the instrument. Furthermore, visualizing the latent space indicated that sounds generated from a specific instrument exhibit a shorter distance between them. However, some instruments presented partial entanglement of clusters that could be attributed to the dimensionality reduction or inadequate information in the model to discern features based on instruments. Therefore, potential future work would include augmenting the precision of the generative model to synthesize sound in a lower-dimensional space or introducing further regularization techniques that provide timbre information. Finally, additional explainability methods can be adopted to interpret the synthesis of new samples.

Acknowledgement

This work was funded by Science Foundation Ireland and its Centre for Research Training in Machine Learning (18/CRT/6183)

References

1. Ahmed, T., Longo, L.: Examining the size of the latent space of convolutional variational autoencoders trained with spectral topographic maps of eeg frequency bands. IEEE Access **10**, 107575–107586 (2022). https://doi.org/10.1109/ACCESS.2022.3212777
2. Aouameur, C., Esling, P., Hadjeres, G.: Neural drum machine: An interactive system for real-time synthesis of drum sounds. In: International Conference on Computational Creativity (2019)
3. Arık, S.Ö., Jun, H., Diamos, G.: Fast spectrogram inversion using multi-head convolutional neural networks. IEEE Signal Processing Letters **26**(1), 94–98 (2018)
4. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion **58**, 82–115 (2020)
5. Baldi, P., Hornik, K.: Neural networks and principal component analysis: Learning from examples without local minima. Neural networks **2**(1), 53–58 (1989)
6. Caillon, A., Bitton, A., Gatinet, B., Esling, P.: Timbre latent space: exploration and creative aspects. In: Timbre International Conference (2020)
7. Caillon, A., Esling, P.: RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. In: International Conference on Learning Representations (2022)
8. Chikkankod, A.V., Longo, L.: On the dimensionality and utility of convolutional autoencoder's latent space trained with topology-preserving spectral eeg head-maps. Machine Learning and Knowledge Extraction **4**(4), 1042–1064 (2022). https://doi.org/10.3390/make4040053, https://www.mdpi.com/2504-4990/4/4/53
9. Covert, I., Lundberg, S.M., Lee, S.I.: Understanding global feature contributions with additive importance measures. Advances in Neural Information Processing Systems **33**, 17212–17223 (2020)
10. De Cheveigné, A., Kawahara, H.: Yin, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America **111**(4), 1917–1930 (2002)
11. Défossez, A., Zeghidour, N., Usunier, N., Bottou, L., Bach, F.: Sing: Symbol-to-instrument neural generator. Advances in neural information processing systems **31** (2018)
12. Dhariwal, P., Jun, H., Payne, C., Kim, J.W., Radford, A., Sutskever, I.: Jukebox: A generative model for music. arXiv e-prints pp. arXiv–2005 (2020)
13. Engel, J., Agrawal, K.K., Chen, S., Gulrajani, I., Donahue, C., Roberts, A.: Gansynth: Adversarial neural audio synthesis. In: International Conference on Learning Representations (2019)
14. Engel, J., Gu, C., Roberts, A., et al.: DDSP: Differentiable Digital Signal Processing. In: International Conference on Learning Representations (2019)
15. Franzson, D.B., Shepardsson, V., Magnusson, T.: Autocoder: a variational autoencoder for spectral synthesis (2022)
16. Graving, J., Couzin, I.: VAE-SNE: a deep generative model for simultaneous dimensionality reduction and clustering. BioRxiv pp. 2020–07 (2020)
17. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization (2017), http://arxiv.org/abs/1412.6980, number: arXiv:1412.6980 arXiv:1412.6980 [cs]
18. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

19. Kobayashi, K., Miyake, M., Takahashi, M., Hamamoto, R.: Observing deep radiomics for the classification of glioma grades. Scientific Reports **11**(1), 10942 (2021)
20. Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W.Z., Sotelo, J., de Brebisson, A., Bengio, Y., Courville, A.: MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. arXiv:1910.06711 [cs, eess] (2019), http://arxiv.org/abs/1910.06711, arXiv: 1910.06711
21. Lee, S., Kim, M., Shin, S., Lee, D., Jang, I., Lim, W.: Conditional variational autoencoder to improve neural audio synthesis for polyphonic music sound. arXiv preprint arXiv:2211.08715 (2022)
22. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
23. Luo, Y.J., Agres, K., Herremans, D.: Learning disentangled representations of timbre and pitch for musical instrument sounds using gaussian mixture variational autoencoders. arXiv preprint arXiv:1906.08152 (2019)
24. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of machine learning research **9**(11) (2008)
25. Maćkiewicz, A., Ratajczak, W.: Principal Components Analysis (PCA). Computers & Geosciences **19**(3), 303–342 (1993)
26. Natsiou, A., Longo, L., O'Leary, S.: An investigation of the reconstruction capacity of stacked convolutional autoencoders for log-mel-spectrograms. In: 2022 16th International Conference on Signal-Image Technology Internet-Based Systems (SITIS). pp. 155–162 (2022). https://doi.org/10.1109/SITIS57111.2022.00038
27. Natsiou, A., O'Leary, S.: Audio representations for deep learning in sound synthesis: A review. In: 2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA). pp. 1–8. IEEE (2021)
28. Nguyen, Q.P., Lim, K.W., Divakaran, D.M., Low, K.H., Chan, M.C.: Gee: A gradient-based explainable variational autoencoder for network anomaly detection. In: 2019 IEEE Conference on Communications and Network Security (CNS). pp. 91–99. IEEE (2019)
29. Reed, C., et al.: Exploring XAI for the arts: Explaining latent space in generative music (2022)
30. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
31. Saseendran, A., Skubch, K., Falkner, S., Keuper, M.: Shape your space: A gaussian mixture regularization approach to deterministic autoencoders. Advances in Neural Information Processing Systems **34**, 7319–7332 (2021)
32. Shan, S., Hantrakul, L., Chen, J., Avent, M., Trevelyan, D.: Differentiable wavetable synthesis. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4598–4602. IEEE (2022)
33. Subramani, K., Rao, P., D'Hooge, A.: Vapar synth-a variational parametric model for audio synthesis. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 796–800. IEEE (2020)
34. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
35. Tatar, K., Bisig, D., Pasquier, P.: Latent timbre synthesis: Audio-based variational auto-encoders for music composition and sound design applications. Neural Computing and Applications **33**, 67–84 (2021)

36. Vigliensoni, G., McCallum, L., Fiebrink, R.: Creating latent spaces for modern music genre rhythms using minimal training data. In: Conference on Computational Creativity (2020)
37. Vilone, G., Longo, L.: A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods. Frontiers in Artificial Intelligence **4**, 160 (2021). https://doi.org/10.3389/frai.2021.717899
38. Vilone, G., Rizzo, L., Longo, L.: A comparative analysis of rule-based, model-agnostic methods for explainable artificial intelligence. In: Longo, L., Rizzo, L., Hunter, E., Pakrashi, A. (eds.) Proceedings of The 28th Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Republic of Ireland, December 7-8, 2020. CEUR Workshop Proceedings, vol. 2771, pp. 85–96. CEUR-WS.org (2020)
39. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
40. Watcharasupat, K.N., Lerch, A.: Evaluation of latent space disentanglement in the presence of interdependent attributes. In: International Society for Music and Information Retrieval Conference (ISMIR) (2021)
41. Xu, J., Ren, Y., Tang, H., Pu, X., Zhu, X., Zeng, M., He, L.: Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9234–9243 (2021)