BRIDGING AUDIO ANALYSIS, PERCEPTION AND SYNTHESIS WITH PERCEPTUALLY-REGULARIZED VARIATIONAL TIMBRE SPACES

Philippe Esling, Axel Chemla–Romeu-Santos, Adrien Bitton
Institut de Recherche et Coordination Acoustique-Musique (IRCAM)
CNRS - UMR 9912, UPMC - Sorbonne Universite
1 Place Igor Stravinsky, F-75004 Paris, France
{esling, chemla, bitton}@ircam.fr

ABSTRACT

Generative models aim to understand the properties of data, through the construction of *latent spaces* that allow classification and generation. However, as the learning is unsupervised, the latent dimensions are not related to perceptual properties. In parallel, music perception research has aimed to understand timbre based on human dissimilarity ratings. These lead to timbre spaces which exhibit perceptual similarities between sounds. However, they do not generalize to novel examples and do not provide an invertible mapping, preventing audio synthesis. Here, we show that Variational Auto-Encoders (VAE) can bridge these lines of research and alleviate their weaknesses by regularizing the latent spaces to match perceptual distances collected from timbre studies. Hence, we propose three types of regularization and show that they lead to spaces that are simultaneously coherent with signal properties and perceptual similarities. We show that these spaces can be used for efficient audio classification. We study how audio descriptors are organized along the latent dimensions and show that even though descriptors behave in a non-linear way across the space, they still exhibit a locally smooth evolution. We also show that, as this space generalizes to novel samples, it can be used to predict perceptual similarities of novel instruments. Finally, we exhibit the generative capabilities of our spaces, that can directly synthesize sounds with continuous evolution of timbre perception.

1. INTRODUCTION

Generative models aim to understand the underlying distribution of data based on the observation of examples, in order to generate novel content. Recently, audio synthesis using these models has seen great improvements through efficient waveform models, such as *WaveNet* [19] and *SampleRNN* [17]. These models are able to generate high-quality audio matching the properties of the corpus they have been trained on. However, these models give little control over the output or the hidden features it results from. More recently, *NSynth* [4] has been proposed to generate instrumental notes, while allowing to morph between specific instruments. However, these models remain highly complex, requiring very large number of parameters, long training times and a large number of examples.

Amongst recent generative models, a key proposal is the *Variational Auto-Encoder* (*VAE*) [11]. In these models, encoder and decoder networks are jointly trained through the construction of a *latent space*, that allow both analysis and generation. VAEs address all the limitations of control and analysis through this latent space, while remaining simple and fast to learn without requiring large sets of examples. Furthermore, the VAE seems able to disentangle underlying variation factors by learning independent latent variables [7]. However, these unsupervised dimensions are not related to perceptual properties, which might hamper the control and use of these spaces for analysis and synthesis. The potential of VAEs for audio applications has only been scarcely investigated and mostly for speech source separation [13] and transformation [8]. However, the use of variational latent spaces specifically for musical audio synthesis is yet to be investigated.

In parallel, music perception research has tried to understand the mechanisms behind the perception of instrumental *timbre*. Several studies [15] collected dissimilarity ratings between pairs of instrumental samples. Then, Multi-Dimensional Scaling (MDS) is applied to these ratings to obtain *timbre spaces*, which exhibit the perceptual similarities between instruments. Although these spaces provide interesting analyses, they are inherently limited by the fact that MDS produces a fixed discrete space, which has to be recomputed for any new sample. Therefore, these spaces do not generalize to novel examples and do not provide an invertible mapping, preventing audio synthesis.

Here, we show that we can bridge analysis, synthesis and perceptual audio research by regularizing the learning of latent spaces so that they match the perceptual distances from timbre studies. Our overall approach is depicted in Figure 1. First, we adapt the VAE to analyze musical audio content, by relying on the Non-Stationary Gabor Transform (NSGT) with a Constant-Q scale. This transform allows us to obtain a log-frequency scale while remaining invertible, which is critical to perform audio synthesis. Even with a simple model on a small training set, we show that this provides a generative model with an interesting latent space, able to synthesize novel instrumental sounds.

Then, we propose three regularizations to the learning objective, aiming to enforce that the latent space exhibits the same topology as the topology of timbre spaces. We build a model of perceptual relationships by normalizing dissimilarity ratings from five timbre space studies [5, 9, 12, 14, 16]. We show that perceptually-regularized latent spaces are both coherent with perceptual dissimilarities, while being able to reconstruct audio samples with a high accuracy. Hence, we can drive the learning of the latent space to match the topology of any given target space.

We demonstrate that these spaces can be used for audio classification by training low-capacity classifiers on the spaces. We obtain high accuracy for *family* and *instrument* labels, but also for the *pitch* and *dynamics*, even though the model had no information on these during training. We exhibit the generative capabilities of our spaces, by assessing the reconstruction quality of the model on a test dataset. We show that the latent spaces can be directly used to synthesize sounds with continuous evolution of timbre perception. We also show that these spaces generalize to novel samples, by encoding instruments that were not part of the training set. Therefore, the spaces could be used to predict the perceptual similarities of novel instruments. Finally, we study how audio descriptors behave along the latent dimensions, by generating audio samples on a grid across space. We show that even though descriptors behave in a non-linear way across the space, they still follow a locally smooth evolution. Our source code, audio examples and additional figures and animations are available online [1].

2. STATE-OF-ART

2.1 Variational auto-encoders

Generative models are a flourishing class of machine learning approaches, aiming to find the underlying probability distribution of the data $p(\mathbf{x})$ [2]. Formally, based on a set of examples in $\mathbf{x} \in \mathbb{R}^{d_x}$, we assume that these follow an unknown probability distribution $p\left(\mathbf{x}\right)$. Furthermore, we consider a set of *latent variables* defined in a lower-dimensional space $\mathbf{z} \in \mathbb{R}^{d_z}$ ($d_z \ll d_x$), a higher-level representation that could have led to generate a given example. These latent variables help govern the distribution of the data and enhance the *expressivity* of the model. The complete model is defined by the joint probability distribution $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$. We could find $p(\mathbf{x})$ by marginalizing $\mathbf{z}$ from the joint probability. However, for most models, this integral can not be found in closed form.

Recently, *variational inference* (VI) has been proposed to solve this problem through *optimization*. VI assumes that if the distribution is too complex to find, we could find a simpler approximate distribution that still models the data, while trying to minimize its difference to the real distribution. VI specifies a family $\mathcal{Q}$ of approximate densities, where each member $q(\mathbf{z}|\mathbf{x}) \in \mathcal{Q}$ is a candidate approximation to the exact conditional $p\left(\mathbf{z} \mid \mathbf{x}\right)$. Hence, the inference can be transformed into an optimization problem by minimizing the Kullback-Leibler (KL) divergence between the approximation and the original density

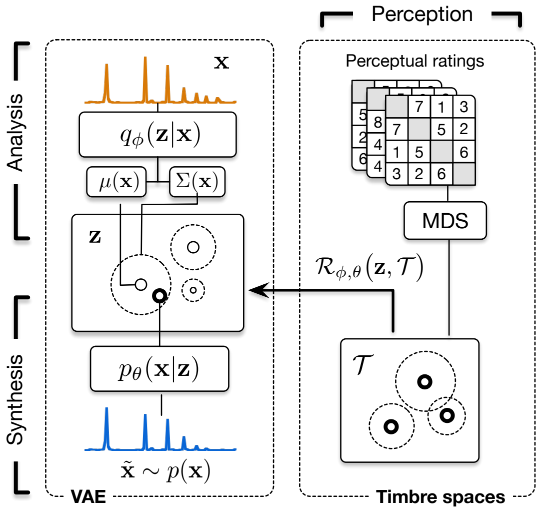

Figure 1. The VAE models audio samples $\mathbf{x}$ by learning an encoder $q_\phi(\mathbf{z} \mid \mathbf{x})$ which maps them to a Gaussian $\mathcal{N}(\mu(\mathbf{x}), \sigma(\mathbf{x}))$ in latent space $\mathbf{z}$. The decoder $p_\theta(\mathbf{x}|\mathbf{z})$ samples from this Gaussian to generate a reconstruction $\tilde{x}$. Perception studies use dissimilarity ratings to construct a *timbre space* that exhibits the perceptual distances between instruments. Here, we develop regularizations methods $\mathcal{R}(\mathbf{z}, \mathcal{T})$, to enforce that the variational model finds a topology of latent space $\mathbf{z}$ that matches the topology of the timbre space $\mathcal{T}$.

$$q^*(\mathbf{z} \mid \mathbf{x}) = \underset{q(\mathbf{z} \mid \mathbf{x}) \in \mathcal{Q}}{\arg\min} \ \mathcal{D}_{KL}\left[q\left(\mathbf{z} \mid \mathbf{x}\right) \parallel p\left(\mathbf{z} \mid \mathbf{x}\right)\right] \quad (1)$$

By developing this KL divergence and re-arranging terms (the detailed development can be found in [11]), we obtain

$$\log p(\mathbf{x}) - D_{KL}\left[q(\mathbf{z} \mid \mathbf{x}) \parallel p(\mathbf{z} \mid \mathbf{x})\right] = \\ \mathbb{E}_{\mathbf{z}}\left[\log p(\mathbf{x} \mid \mathbf{z})\right] - D_{KL}\left[q(\mathbf{z} \mid \mathbf{x}) \parallel p(\mathbf{z})\right] \quad (2)$$

This formulation describes the quantity we want to maximize $\log p(\mathbf{x})$ minus the error we make by using an approximate q instead of p. Therefore, we can optimize this alternative objective, called the *evidence lower bound* (ELBO). Now, to optimize this objective, we will rely on parametric distributions $q_\phi(\mathbf{z})$ and $p_\theta(\mathbf{z})$. Optimizing our generative model will amount to optimizing these parameters $\{\theta, \phi\}$ of these distributions with

$$\mathcal{L}_{\theta,\phi} = \mathbb{E}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \beta \cdot D_{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z})\right] \quad (3)$$

We can see that this equation involves $q_\phi(\mathbf{z} \mid \mathbf{x})$ which *encodes* the data $\mathbf{x}$ into the latent representation $\mathbf{z}$ and a *decoder* $p(\mathbf{x} \mid \mathbf{z})$, which allows generating a data $\mathbf{x}$ given a latent configuration $\mathbf{z}$. Hence, this structure defines the *Variational Auto-Encoder* (VAE), depicted in Figure 1 (Left).

[1] `https://github.com/acids-ircam/ismir2018`

The VAE objective can be interpreted intuitively. The first term increases the likelihood of the data generated given a configuration of the latent, which amounts to minimize the *reconstruction error*. The second term represents the error made by using a simpler distribution $q_\phi(\mathbf{z} \mid \mathbf{x})$ rather than the true distribution $p_\theta(\mathbf{z})$. Therefore, this *regularizes* the choice of approximation q so that it remains close to the true posterior distribution [11]. Here, we also introduced a weight β on the KL divergence, which has been shown to improve the capacity of the model to disentangle factors of variations in the data [7].

VAEs are powerful representation learning frameworks, while remaining simple and fast to learn without requiring large sets of examples [18]. Their potential for audio applications have been only scarcely investigated yet and mostly in topics related to speech processing such as blind source separation [13] and speech transformation [8]. However, to our best knowledge, the use of VAE to perform musical audio analysis and generation has yet to be investigated.

2.2 Timbre spaces and auditory perception

For decades, researchers have tried to understand the mechanisms of *timbre* perception. Timbre is the set of properties that distinguishes two instruments playing the same note at the same intensity. Several studies tried to understand this phenomenon by relying on *timbre spaces* [6], a model that aims to organize audio samples based on human dissimilarity ratings. The experimental protocol consists of presenting pairs of sounds to subjects. Each subject has to rate the perceptual dissimilarity of all pairs of samples inside a selected set of instruments. Then, these ratings are compiled into a set of dissimilarity matrices that are analyzed with Multi-Dimensional Scaling (MDS). The MDS algorithm provides a timbre space that exhibits the perceptual distances between different instruments. This process is depicted in Figure 1 (Right). Here, we briefly detail the studies and redirect the interested readers to the full articles for more details.

In his seminal paper, Grey [5] performed a study with 16 instrumental sound samples in which 22 subjects had to rate their dissimilarities on a continuous scale from 0 (most similar) to 1 (most dissimilar), leading to the first construction of a timbre space. Following this study, Krumhansl [12] used 21 instruments with 9 subjects on a discrete scale from 1 to 9, Iverson et al. [9] with 16 samples and 10 subjects on a continuous scale from 0 to 1, McAdams et al. [16] with 18 instruments and 24 subjects on a discrete scale from 1 to 16 and, finally, Lakatos [14] with 17 subjects and different instrument sets on a continuous scale from 0 to 1. Each of these studies shed light on different aspects of audio perception, depending on the interpretation of the dimensions. However, all studies produced different spaces with different dimensions, preventing a generalization on the acoustic cues that might correspond to timbre dimensions. Furthermore, these studies are inherently limited by the fact that ordination techniques (e.g. MDS) produce fixed spaces that must be recomputed for any new data point [16]. Hence, these spaces are unable to generalize nor can we generate data from these as they do not provide an invertible mapping. Here, we show that learning latent spaces, while regularizing their topology to fit perceptual ratings can alleviate these limitations.

3. REGULARIZING THE TOPOLOGY OF LATENT SPACES

We show that we can influence the learning of the latent space $\mathbf{z}$ so that it follows the topology of a given target space $\mathcal{T}$. Here, we rely on timbre spaces based on perceptual ratings as a target space. However, it should be noted that this idea can be applied to any target space. Here, we consider a set of audio samples x_i where each have relations in both latent space $\mathbf{z}_i$ and target space $\mathcal{T}_i$. In order to relate the elements of the audio set to the perceptual space, we consider that each sample is labeled with its instrumental class $\mathcal{C}_i$, that has an equivalent in the timbre space.

3.1 Penalty regularization

First, we define an additive *penalty* regularization $\mathcal{R}\left(\mathbf{z}, \mathcal{T}\right)$ that imposes that the properties of the latent $\mathbf{z}$ are similar to that of the target $\mathcal{T}$. Our objective becomes

$$\mathbb{E}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] - \beta D_{KL}\big[q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z})\big] + \alpha\mathcal{R}\big(\mathbf{z}, \mathcal{T}\big)$$

Hence, amongst two otherwise equal solutions, the model is pushed to select the one that comply with the penalty. The weight α allows us to control the influence of this regularization. In our case, we want the distances between instruments to follow the perceptual distances. Therefore, we need to minimize the differences between the distances in latent space $\mathcal{D}^{\mathbf{z}}_{i,j} = \mathcal{D}(\mathbf{z}_i, \mathbf{z}_j)$ and in target space $\mathcal{D}^{\mathcal{T}}_{i,j} = \mathcal{D}(\mathcal{T}_i, \mathcal{T}_j)$. The regularization criterion will minimize the differences between these sets of distances

$$\mathcal{R}\big(\mathbf{z}, \mathcal{T}\big) = \sum_{i \neq j} \mathcal{R}_{i,j}\big(\mathbf{z}, \mathcal{T}\big) = \sum_{i \neq j} \mathcal{R}\big(\mathcal{D}^{\mathbf{z}}_{i,j}, \mathcal{D}^{\mathcal{T}}_{i,j}\big) \tag{4}$$

Euclidean. First, we rely on the Euclidean distance to compute the distance between points in both spaces with $\mathcal{D}^{\mathcal{S}}_{i,j} = \|\mathcal{S}_i - \mathcal{S}_j\|^2$ and also to compare distance matrices

$$\mathcal{R}_{i,j}\big(\mathbf{z}, \mathcal{T}\big) = \big\|\mathcal{D}^{\mathbf{z}}_{i,j} - \mathcal{D}^{\mathcal{T}}_{i,j}\big\|^2 \tag{5}$$

This regularization provides an incentive to the model to obtain the Euclidean metric properties of the target space. *Gaussian.* Here, we model the fact that perceptual ratings are subjective assessments. Therefore, we consider that each perceptual rating between instruments i and j is drawn from a univariate Gaussian $d_{i,j} \sim \mathcal{N}\big(\mu_{i,j}, \sigma_{i,j}\big)$. As we can see, we define a different distribution for each *pair* of instruments. When evaluating the regularization, we draw a different distance at each iteration for all pairs

$$\big\{\mathcal{D}^{\mathcal{T}}_{i,j}\big\}_{it} \sim \mathcal{N}\big(\mu_{i,j}, \sigma_{i,j}\big)$$

Hence, this regularization models the uncertainty present in the set of perceptual ratings.

3.2 Prior regularization

In the VAE objective, we observe that the prior $p(\mathbf{z})$ already carries information on the organization of the latent space. Therefore, we can inject the desired topology of the latent space inside that term. Here, we propose to introduce a class-based prior

$$p(\mathbf{z}_i) = \mathcal{N}(\mu_{\mathcal{T}}(\mathcal{C}_i), \sigma_{\mathcal{T}}(\mathcal{C}_i))$$

where $\mathcal{C}_i$ is the class of element i. Therefore, this prior pushes the VAE to find a configuration of the samples in latent space so that they follow the same distribution as their class in our target timbre space. The computation of class means $\mu_{\mathcal{T}}(\mathcal{C}_i)$ and covariances $\sigma_{\mathcal{T}}(\mathcal{C}_i)$ based on the perceptual ratings is detailed in Section 4.1.

4. EXPERIMENTS

4.1 Datasets

Timbre studies. We rely on perceptual dissimilarity ratings collected across five independent timbre studies [5, 9, 12, 14, 16], detailed globally in [3, 15]. As discussed earlier (Section 2.2), even though all studies follow the same protocol, there are some discrepancies in the instruments, number of participants and rating scales.

Hence, we normalize the dissimilarity ratings so that all studies map to a common scale from 0 to 1. Then, we compute the maximal set of instruments for which we had pairwise ratings for all pairs by counting co-occurences in studies. This leads to a set of 11 instruments (Piano, Cello, Violin, Flute, Clarinet, Trombone, Horn, Oboe, Saxophone, Trumpet, Tuba). Finally, we extract the set of ratings that corresponds to our selected instruments, amounting to a total of 11845 pairwise ratings. Based on this set of ratings, we compute an MDS space to obtain the positions in target space of each instrument (which also corresponds to the mean $\mu_{\mathcal{T}}$) and to ensure the consistency of our normalized perceptual space. For all pairs of instruments, we also fit a Gaussian distribution to the pairwise dissimilarity ratings in order to obtain the mean $\mu_{i,j}$ and variance $\sigma_{i,j}$ of that pair for the Gaussian regularization. We derive the global variance $\sigma_{\mathcal{T}}$ for each instrument, by taking the mean of their pairwise variances. Results of this analysis are displayed in Figure 2. Even though ratings come from different studies, the resulting space appears very coherent with clusters of families and the distances between individual instruments correlated to previous perceptual studies.

Audio datasets. In order to learn the distribution of instrumental audio, we rely on the Studio On Line (SOL) database [1]. We selected 2,200 samples to represent the 11 instruments for which we extracted perceptual ratings. These represent the whole tessitura and dynamics available (to remove effects from the pitch and loudness). All recordings were resampled to a sampling rate of 22050Hz. For each audio sample, we compute the Non-Stationary Gabor Transform (NSGT) mapped on a Constant-Q scale of 24 bins per octave. We only keep the magnitude of the NSGT to train our models. Then, we perform a corpus-wide normalization to preserve the relative intensities of the samples and extract a single temporal frame to represent the given audio sample. Finally, the dataset is randomly split between a training (90%) and test (10%) set.

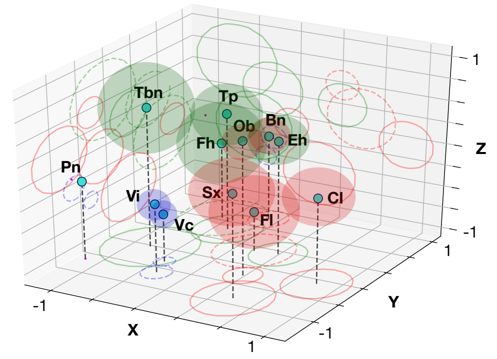

Figure 2. Multi-dimensional scaling (MDS) applied to the combined normalized set of perceptual dissimilarity ratings (strings in blue, brasses in green and winds in red).

4.2 Models

In order to evaluate our proposal, we rely on a very simple VAE architecture to show its efficiency. The encoder is defined as a 3-layers feed-forward network with ReLU activations and 3000 units per layer. The last layer maps to a latent space of 64 dimensions. The decoder is defined with the same architecture, mapping back to the dimensionality of the input. For learning the model, we use a value of β, which is linearly increased from 0 to 2 during the first 100 epochs (*warmup* procedure [18]). In order to train the model, we rely on the ADAM optimizer [10] with an initial learning rate of 0.00001, and a Xavier weight initialization [18]. In a first stage, we train the model without perceptual regularization ($\alpha = 0$) for 5000 epochs. Then, we introduce the perceptual regularization ($\alpha = 1$) and train for another 1000 epochs. This allows the model to first focus on the quality of the reconstruction with its own unsupervised regularization, and then to converge towards a solution with perceptual space properties. This leads to a training time of one hour on a NVIDIA Titan X GPU.

5. RESULTS

5.1 Latent spaces properties

In order to visualize the latent spaces, we apply a Principal Component Analysis (PCA) to obtain a 3d representation. Using a PCA ensures that the representation is a linear transform that preserves the distances inside the original space. This also provides an exploitable control space for audio synthesis. Results are displayed in Figure 3.

As we can see, the VAE without regularization is already able to dissociate instrumental distributions, while providing almost perfect reconstruction of audio samples from the low-dimensional space. This confirms that VAEs can provide interesting latent spaces for analysis and synthesis. However, the relationships between instruments are

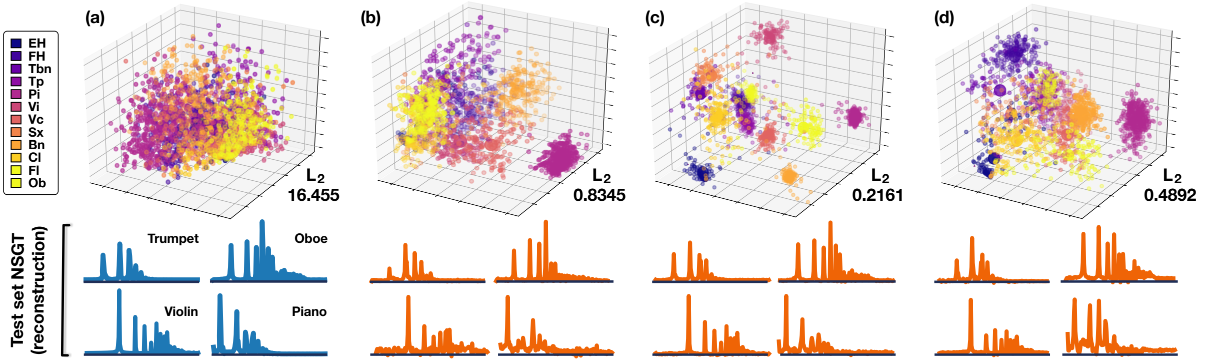

Figure 3. Comparing the latent spaces for the VAE unregularized (a) or with Prior (b), Euclidean (c) and Gaussian (d) regularization. The mean L_2 differences between latent and timbre spaces distances is indicated on each graph. We show under each space the reconstruction of NSGT distributions from the test set directly from these spaces.

entirely different from perceptual ratings. Furthermore, the large variance of the distributions seem to indicate that the model rather tries to spread the information across the latent space to help the reconstruction.

In the case of all regularizations (b-d), we can clearly see the enhancement on the dissociation of instrumental distributions. Furthermore, the overall distances between instruments match well the distances based on perceptual ratings (Figure 2). This similarity is particularly striking for the L_2 regularization (c), which provides the lowest overall differences to our combined timbre space. This might come from the fact that MDS spaces have an Euclidean metric topology. However, this might also indicate an effect of *over-regularization*, which might impact generalization. For all regularized latent spaces, the instrumental distributions are shuffled around the space in order to comply with the reconstruction objective. However, the pairwise distances reflecting perceptual relations are well matched as indicated by their respective L_2 differences to the timbre space. Finally, by looking at the reconstructions of the NSGT distributions from the test set, we can see that enforcing the perceptual topology to the latent spaces does not impact the quality of audio reconstruction (this evaluation is quantified in Section 5.3). However, we note an occasional addition of low-amplitude noise, which might indicate that the model focuses on optimizing the partials rather than the low-amplitude tail of the distribution.

5.2 Discriminative capabilities

We evaluate the discriminative capabilities of the latent spaces through a classification task. We use a very low-capacity classifier composed of a single-layer network of 512 ReLU units with batch normalization and softmax regression. The low-capacity classifier ensures that the latent space needs to be well organized to obtain a good accuracy. In order to evaluate the impact of our proposal, we also compare these results to a simple PCA with softmax regression and an Auto-Encoder (AE) with the same capacity as the VAE. Results are presented in Table 1.

We can see that all models perform an excellent classification of instrumental properties. However, a very interesting observation comes from the vanilla VAE providing the best accuracy on *instrument* classification, even though we regularized other models with distances highly relevant to these categories. This might underline the fact that perceptual information could blur discrimination of highly similar instruments (such as violin and violoncello). Interestingly, the symmetric results on *pitch* and *dynamics* categories might indicate that regularized model are pushed to focus on timbre properties. Therefore, they need to more clearly separate the variations coming from pitch and loudness to understand the variability of timbre.

Method	Family	Instrument	Pitch	Dynam.
PCA	0.790	0.697	0.167	0.527
AE	0.973	0.957	0.936	0.597
VAE	0.978	**0.993**	0.963	0.941
Prior	0.975	0.991	**0.993**	0.936
Euclidean	0.972	0.990	0.990	0.943
Gaussian	**0.982**	0.991	0.989	**0.948**

Table 1. Discriminative capabilities in classifying *family*, *instrument*, *pitch* and *dynamics* of the test set.

Method	$\log p(\mathbf{x})$	$\|\mathbf{x} - \tilde{\mathbf{x}}\|^2$
PCA	-	2.2570
AE	-1.2008	1.6223
VAE	-2.3443	0.1593
Prior	**-2.7143**	0.1883
Euclidean	17.8960	**0.1223**
Gaussian	0.2894	0.1749

Table 2. Generative capabilities evaluated on the log likelihood and reconstruction error over the test set.

5.3 Generative capabilities

We quantify generative capabilities by evaluating reconstructions from the latent space, through the log likelihood and mean difference between original and reconstructed audio on the test set. The results are presented in Table 2.

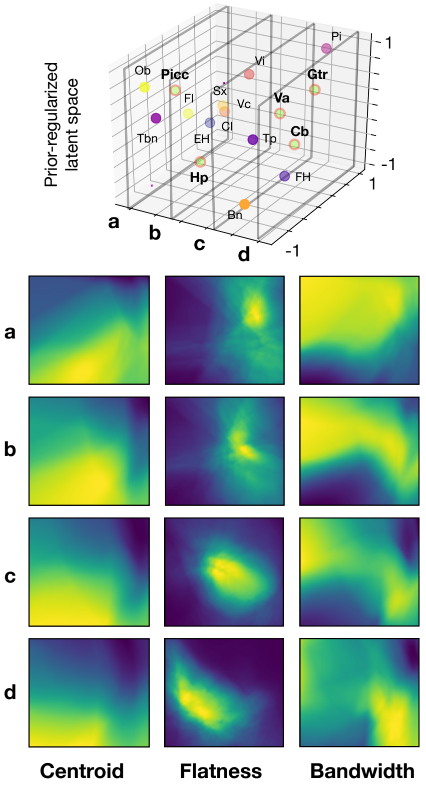

Figure 4. (Top) We encode instruments that were not part of timbre studies to show the *out-of-domain* capabilities of latent spaces. (Bottom) *Topology of descriptors*. We define 4 projection planes equally spaced across the x axis. We sample points at these positions on a 50x50 grid and reconstruct their audio distribution to compute their spectral *centroid*, *flatness* and *bandwidth*.

Overall, the regularizations do not impact the reconstruction quality of the model. Furthermore, we can now sample directly from the spaces to obtain novel sounds that remain perceptually relevant, which allows us to turn our spaces into generative timbre synthesizers. However, as previously hypothesized, the L_2 regularization seems to have a too strong effect on the latent space, disrupting the generalization of the model. We provide generated audio clips representing paths between different instruments in the latent space on the supporting repository for subjective evaluation of the latent space audio synthesis.

5.4 Perceptual inference

The encoder of our perceptually-regularized spaces is able to analyze new instruments that were not part of the original timbre studies. Hence, we could hope that it is able to predict perceptual relationships between new instruments, to feed further timbre studies. To evaluate this, we extracted instruments outside of our perceptual set (Contrabass, Guitar, Harp, Piccolo, Viola) and encode these samples in the latent space to study the *out-of-domain* generalization capabilities of our model. Results are presented in Figure 4 (Top, only the centroid of distributions are shown for clarity). Here, the Piccolo and Viola seem to group in a coherent way with their families. However, the Guitar and Harp do not provide such straightforward relationships. Therefore, perceptual inference from these spaces would require more extensive perception experiments.

5.5 The topology of audio descriptors

We analyze the behavior of signal descriptors across the latent space in order to study their topology. As the space is continuous, we do so by sampling uniformly the PCA space and then using the decoder to generate all audio samples on this grid. Then, we compute the audio descriptors of these samples. In order to provide a visualization here, we select equally-distant planes across the x dimension (at positions {-.75, -.25, .25, .75}) in Figure 4 for the spectral *flatness*, *centroid* and *bandwidth*. Videos of continuous traversals of the latent space for different descriptors are available on the supporting repository.

Audio descriptors seem to be organized in a non-linear way across our spaces. However, they still exhibit both locally smooth evolution and an overall logical organization. This shows that our model is able to organize audio variations. A very interesting observation comes from the topology of the centroid. Indeed, all perceptual studies underline its linear correlation to timbre perception, which is partly confirmed by our model (see Figure 4). This confirms the perceptual relevance of these latent spaces. However, this also shows that the relation between centroid and timbre perception might not be entirely linear.

6. CONCLUSION

We have shown that VAEs can learn a latent space allowing for high-level audio analysis and synthesis directly from these spaces. We proposed different methods for regularizing these spaces to follow the metric properties of timbre spaces. These regularized models provide a control space from which the generation of perceptually relevant audio content is straightforward. By analyzing the behavior of audio descriptors across the latent space, we have shown that, while following a non-linear evolution, they still exhibit some locally smooth properties. Future works on these spaces include perceptual experiments to confirm their perceptual topology and also to thrive on the smoothness of audio descriptors to develop a descriptor-based synthesizer.

7. ACKNOWLEDGEMENTS

This work was supported by the MAKIMOno project 17-CE38-0015-01 funded by the French ANR and the Canadian Natural Sciences and Engineering Reserch Council (STPG 507004-17) and the ACTOR Partnership funded by the Canadian Social Sciences and Humanities Research Council (895-2018-1023).

8. REFERENCES

[1] Guillaume Ballet, Riccardo Borghesi, Peter Hoffmann, and Fabien Lévy. Studio online 3.0: An internet" killer application" for remote access to ircam sounds and processing tools. *Journée dInformatique Musicale (JIM)*, 1999.

[2] Christopher M. Bishop and Tom M. Mitchell. Pattern recognition and machine learning. 2014.

[3] John A. Burgoyne and Stephen McAdams. A meta-analysis of timbre perception using nonlinear extensions to clascal. In *International Symposium on Computer Music Modeling and Retrieval*, pages 181–202. Springer, 2007.

[4] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural audio synthesis of musical notes with wavenet autoencoders. *arXiv preprint arXiv:1704.01279*, 2017.

[5] John M Grey. Multidimensional perceptual scaling of musical timbres. *the Journal of the Acoustical Society of America*, 61(5):1270–1277, 1977.

[6] John M. Grey and John W. Gordon. Perceptual effects of spectral modifications on musical timbres. *The Journal of the Acoustical Society of America*, 63(5):1493–1500, 1978.

[7] 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. 2016.

[8] Wei-Ning Hsu, Yu Zhang, and James Glass. Learning latent representations for speech generation and transformation. *arXiv preprint arXiv:1704.04222*, 2017.

[9] Paul Iverson and Carol L. Krumhansl. Isolating the dynamic attributes of musical timbrea. *The Journal of the Acoustical Society of America*, 94(5):2595–2603, 1993.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[12] Carol L. Krumhansl. Why is musical timbre so hard to understand. *Structure and perception of electroacoustic sound and music*, 9:43–53, 1989.

[13] Jen-Tzung Kuo and Kuan-Ting Chien. Variational recurrent neural networks for speech separation. *INTERSPEECH 2017*.

[14] Stephen Lakatos. A common perceptual space for harmonic and percussive timbres. *Perception & psychophysics*, 62(7):1426–1439, 2000.

[15] Stephen McAdams, Bruno L. Giordano, Patrick Susini, Geoffroy Peeters, and Vincent Rioux. A meta-analysis of acoustic correlates of timbre dimensions. *Journal of the Acoustical Society of America*, 120(5):3275, 2006.

[16] Stephen McAdams, Suzanne Winsberg, Sophie Donnadieu, Geert De Soete, and Jochen Krimphoff. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological research*, 58(3):177–192, 1995.

[17] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. Samplernn: An unconditional end-to-end neural audio generation model. *arXiv preprint arXiv:1612.07837*, 2016.

[18] Casper K. Sønderby, Tapani Raiko, Lars Maaløe, Søren K. Sønderby, and Ole Winther. How to train deep variational autoencoders and probabilistic ladder networks. *arXiv preprint arXiv:1602.02282*, 2016.

[19] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.