Make That Sound More Metallic: Towards a Perceptually Relevant Control of the Timbre of Synthesizer Sounds Using a Variational Autoencoder

Fanny Roche, Thomas Hueber, Maëva Garnier, Samuel Limier, Laurent Girin

TISMIR

Roche, F., et al. (2021). Make That Sound More *Metallic*: Towards a Perceptually Relevant Control of the Timbre of Synthesizer Sounds Using a Variational Autoencoder. *Transactions of the International Society for Music Information Retrieval,* 4(1), pp. 52–66. DOI: https://doi.org/10.5334/tismir.76

RESEARCH

Make That Sound More *Metallic*: Towards a Perceptually Relevant Control of the Timbre of Synthesizer Sounds Using a Variational Autoencoder

Fanny Roche*, Thomas Hueber†, Maëva Garnier†, Samuel Limier* and Laurent Girin†

In this article, we propose a new method of sound transformation based on control parameters that are intuitive and relevant for musicians. This method uses a variational autoencoder (VAE) model that is first trained in an unsupervised manner on a large dataset of synthesizer sounds. Then, a perceptual regularization term is added to the loss function to be optimized, and a supervised fine-tuning of the model is carried out using a small subset of perceptually labeled sounds. The labels were obtained from a perceptual test of Verbal Attribute Magnitude Estimation in which listeners rated this training sound dataset along eight perceptual dimensions (French equivalents of *metallic*, *warm*, *breathy*, *vibrating*, *percussive*, *resonating*, *evolving*, *aggressive*). These dimensions were identified as relevant for the description of synthesizer sounds in a first Free Verbalization test. The resulting VAE model was evaluated by objective reconstruction measures and a perceptual test. Both showed that the model was able, to a certain extent, to capture the acoustic properties of most of the perceptual dimensions and to transform sound timbre along at least two of them (*aggressive* and *vibrating*) in a perceptually relevant manner. Moreover, it was able to generalize to unseen samples even though a small set of labeled sounds was used.

Keywords: Synthesizer sounds; timbre perception and verbal description; variational autoencoders; machine learning; audio synthesis

1. Introduction

Synthesizers are powerful instruments that offer musicians a large palette of possibilities for creating sounds. However, the most common synthesis methods (additive synthesis, subtractive synthesis, frequency modulation and physical modeling (Miranda, 2002)) are controlled by low-level parameters that are often numerous and not easily correlated with musical intent. Consequently, musicians often need technical expertise, or even assistance, to generate interesting sounds. To broaden the range of possible sounds and improve synthesizers' ergonomics, it might be better for musicians to control the sound synthesis from a reduced number of higher level dimensions that are more intuitive and directly related to timbre perception.

A first issue when searching for these control dimensions is that musical timbre is neither unidimensional nor uniparametric (von Bismarck, 1974). Controlling timbre with a synthesizer therefore involves manipulating several perceptual dimensions, resulting in the joint variation of multiple acoustic descriptors.

* ARTURIA, Montbonnot Saint-Martin, FR

† Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, Grenoble, FR

Corresponding author: Fanny Roche (fanny.roche@arturia.com)

A second issue is to choose the angle from which to approach the problem: either by studying the consequences of parameterized acoustic variations on supposedly relevant perceptual dimensions (psychoacoustic approach) (Grey and Moorer, 1977; McAdams et al., 1999), or by identifying the perceptual dimensions on which listeners rely to evaluate timbre and by searching for their acoustic correlates (semioacoustic approach) (Faure 2000; Traube, 2004).

A third problem is to find a reduced number of dimensions organizing the timbre space. Multidimensional scaling (MDS) was used to organize, by perceptual similarity, the timbre of different orchestral musical instruments in a geometric space with a reduced number of dimensions (usually three for visualization but there may be more – see Wedin and Goude (1972); Grey (1977); Grey and Moorer (1977); Wessel (1979); Krumhansl (1989); Iverson and Krumhansl (1993); Krimphoff et al. (1994); McAdams et al. (1995); Faure (2000); Marozeau et al. (2003); Zacharakis (2013); McAdams (2019) for a review). Varying acoustic correlates to these dimensions were found: descriptors of the long-term average spectral envelope (e.g. spectral centroid or degree of harmonicity), of the temporal envelope (e.g. logarithm of the onset time) and of the spectral variations over time (e.g. spectral flux or vibrato).

However, the semantic interpretation of these dimensions remains relatively unclear. Following an alternative approach based on free categorization/verbalization, other studies identified a greater, though limited, number of main semantic dimensions (from two to twenty) to describe the timbre of musical instruments such as "brightness" or "warmth" (Traube 2004; Garnier et al., 2007; Fritz et al., 2012; Reymore and Huron, 2020). However, (i) the number and label of these semantic dimensions strongly depend on the sound category (Dubois, 2000; McAdams, 2019), (ii) they depend on the culture and expertise of the listeners as well as their listening goal (Dubois, 2000), (iii) they are not always strictly orthogonal but may present some form of semantic overlap, inclusion or opposition, (iv) they can be relatively objective (e.g. "vibrato") and closely related to a low-level acoustic parameter, or more subjective and related to the combined variation of several acoustic parameters (e.g. "strident"), (v) they can hardly be conceptualized without using language. This consequently requires semantic analyses to identify the most shared and non-polysemic terms to represent each perceptual dimension in a given language.

A last complexity of this mapping comes from the non-linear relationship between acoustic and perceptual spaces. After decades of linear models and regressions to correlate both spaces (Garnier et al., 2007; Fritz et al., 2012), more recent studies have started using machine learning methods, in particular deep neural networks (DNNs) such as autoencoders (AEs) or generative adversarial networks (GANs) (Goodfellow et al., 2016), to model and synthesize audio (Colonel et al., 2017; Engel et al., 2017; Roche et al., 2019; Donahue et al., 2019; Engel et al., 2019). In particular, AE-based methods identify a limited number of latent dimensions underlying the physical space of a training dataset. The extracted latent space is then used as a control space for creating or hybridizing new sounds (Engel et al., 2017). As an alternative to the classic AE, the variational autoencoder (VAE) introduced by Kingma and Welling (2014) and Rezende et al. (2014) has also been applied to musical sound modeling (e.g. Çakir and Virtanen, 2018; Esling et al., 2018; Roche et al., 2019; Girin et al., 2019). It can be seen as a probabilistic extension of the (deterministic) AE. Specifically, a prior distribution is used to structure (or regularize) the extracted latent coefficients and thus encourage dimensions to be mutually orthogonal. Since however they do not easily relate to perception (Esling et al., 2018; Roche et al., 2019), Esling et al. (2018) investigated how to force the VAE latent space to match the topology of a perceptual timbre space using a fully-labeled dataset of orchestral instruments and an extra regularization term in the VAE objective function. Esling et al. (2020) also experimented mapping the VAE latent space into the parameter space of a synthesizer using an invertible transform (the normalizing flows introduced by Rezende and Mohamed (2015)). This mapping was constrained by arbitrarily-chosen binary semantic tags linked to the synthesizer presets (e.g. "aggressive" vs. "calm").

The objective of the present study is to propose a prototype of an audio synthesizer that can transform the timbre of musical synthetic sounds, by controlling a limited number of perceptual dimensions. The main contributions of this study are:

(i) Identifying, with a free verbalization experiment, the most important perceptual dimensions on which musicians organize their perception of purely synthetic sounds, as well as the most typical and shared verbal descriptors of these dimensions in French. This is presented in Section 3.1. So far, very few previous timbre studies have focused on purely synthetic sounds that do not imitate orchestral instruments (Lichte 1941; von Bismarck, 1974; Miller and Carterette, 1975; Grey and Moorer, 1977; Samson et al., 1997; McAdams et al., 1999; Kendall et al., 1999; Zacharakis, 2013) and they were not necessarily interested in their verbal description, or they mostly relied on English labels selected *a priori*.

(ii) Creating a dataset of 80 synthetic sounds rated by human listeners along these main perceptual dimensions (this is presented in Section 3.2), and using these perceptual ratings to regularize a VAE model and force its latent space to follow the identified perceptual dimensions. To do so, we used an extra term to perceptually regularize our model, in line with Esling et al. (2018). However, in contrast to this previous study, our regularization aims at encouraging each individual latent dimension to drive one of the perceptual dimensions. Furthermore, due to the moderate size of our dataset (of purely synthetic sounds), our study relies on a weakly-supervised method whereas Esling et al. (2018) used bigger sound datasets (of orchestral instruments coming from some of the above-listed MDS studies), which enabled them to use a fully-supervised method instead. Also, compared to Esling et al. (2020), our perceptual regularization relies on continuous values and does not depend on the synthesizer's engine (since sounds remain synthesized directly from the VAE latent space as by Esling et al. (2018)). The overall methodology of sound transformation with a VAE model is presented in Sections 2.1 and 2.2. The proposed perceptual regularization of the VAE is presented in Section 2.3 and implemented in Section 4.2.

(iii) Evaluating the proposed method both objectively and perceptually. This evaluation is presented in Section 4. In particular, we modified the latent coefficients along different perceptual dimensions, resynthesized new corresponding signals, and conducted a new perceptual test to assess the effectiveness of the overall methodology.

2. General Methodology for Sound Transformation with VAEs

2.1 Analysis-transformation-synthesis process

In line with previous studies applying (V)AE models to sound synthesis (Colonel et al., 2017; Blaauw and Bonada, 2016; Hsu et al., 2017a; Esling et al., 2018; Roche et al., 2019), our study follows an analysis-transformation-synthesis approach as illustrated in **Figure 1**.

The first step of the process is to convert the original time-domain signal into the time-frequency domain using the short-term Fourier transform (STFT). The STFT *magnitude* spectrogram is given to the VAE encoder frame by frame, i.e. each column of the magnitude spectrogram is encoded into a latent vector.[1] A complete spectrogram is thus encoded into a sequence of latent vectors. Then, this latent vector sequence can be modified by the musician. For example, a sequence can be shifted with an offset, or two latent vector sequences encoding two different sounds can be interpolated to generate a hybrid sound. The final step of the process consists in decoding the sequence of (possibly modified) latent vectors in order to reconstruct a magnitude spectrogram. The output audio signal is then synthesized by combining the decoded magnitude spectrogram with the phase spectrogram of the input signal and applying inverse STFT (ISTFT). If the latent coefficients are not modified in between encoding and decoding, the decoded magnitude spectrogram is close to the original one and the original phase spectrogram can be directly used for good-quality waveform reconstruction. Otherwise, if the latent coefficients are modified so that the decoded magnitude spectrogram becomes too different from the original, the Griffin & Lim algorithm (Griffin and Lim, 1984) is used to reconstruct the waveform with a more consistent phase spectrogram.

2.2 Variational autoencoders

The proposed approach is based on the VAE model (Kingma and Welling, 2014; Rezende et al., 2014), which can be seen as a probabilistic AE. It delivers a parametric model of the data distribution:

$$p_\theta(\mathbf{x},\mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z}), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^F$ is the input vector, $\mathbf{z} \in \mathbb{R}^L$ is its corresponding low-dimensional latent representation (in general we have $L \ll F$) and θ denotes the set of distribution parameters. The likelihood function $p_\theta(\mathbf{x}|\mathbf{z})$ plays the role of a probabilistic decoder, modeling how the generation of observed data $\mathbf{x}$ is conditioned on the latent data $\mathbf{z}$. The prior distribution $p_\theta(\mathbf{z})$ is used to structure (or regularize) the latent space. Typically a standard Gaussian distribution is used: $p_\theta(\mathbf{z}) = p(\mathbf{z}) = N(\mathbf{z}; \mathbf{0}, \mathbf{I}_L)$, where $\mathbf{I}_L$ is the identity matrix of size L (Kingma and Welling, 2014). This encourages the entries of the latent vector $\mathbf{z}$ to be mutually orthogonal and to lie in a similar range. The likelihood function $p_\theta(\mathbf{x}|\mathbf{z})$ is usually defined as:

$$p_\theta(\mathbf{x}|\mathbf{z}) = N\left(\mathbf{x}; \boldsymbol{\mu}_\theta(\mathbf{z}), \mathrm{diag}\{\boldsymbol{\sigma}^2_\theta(\mathbf{z})\}\right), \tag{2}$$

where $N(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes the multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, diag{·} is the operator that forms a diagonal matrix from a vector by putting the vector entries on the diagonal, and $\boldsymbol{\mu}_\theta(\mathrm{z}) \in \mathbb{R}^F$ and $\boldsymbol{\sigma}^2_\theta(\mathbf{z}) \in \mathbb{R}^F_+$ are non-linear functions of $\mathbf{z}$ implemented with the so-called *decoder network*, which is a feed-forward DNN. θ is thus the set of weights and biases of this network. The VAE decoder is illustrated in the right part of **Figure 2**.

Due to the highly non-linear relationship between $\mathbf{z}$ and $\mathbf{x}$, the exact posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ corresponding to the above generative model is intractable. In the VAE methodology, it is approximated by a tractable parametric inference model $q_\phi(\mathbf{z}|\mathbf{x})$ which acts as the probabilistic encoder. This model is generally similar in form to the decoder:

$$q_\phi(\mathbf{z}|\mathbf{x}) = N\left(\mathbf{z}; \boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}\{\boldsymbol{\sigma}^2_\phi(\mathbf{x})\}\right), \tag{3}$$

where $\boldsymbol{\mu}_\phi(\mathbf{x}) \in \mathbb{R}^L$ and $\boldsymbol{\sigma}^2_\phi(\mathbf{x}) \in \mathbb{R}^L_+$ are non-linear functions of $\mathbf{x}$ implemented as the output of the so-called *encoder network*. The encoder network is also a feed-forward DNN, here parameterized by ϕ. It is illustrated in the left part of **Figure 2**.

The marginal log-likelihood of a data vector $\log p_\theta(\mathbf{x})$ is also intractable. The training of the VAE model, i.e. the estimation of θ and ϕ, is therefore done by maximizing a tractable lower-bound of $\log p_\theta(\mathbf{x})$ over a large dataset of vectors $\mathbf{x}$. It is shown by Kingma and Welling (2014) that this lower bound, called the variational lower bound (VLB), is given by (for an individual vector $\mathbf{x}$):

$$\mathcal{L}(\phi,\theta,\mathbf{x}) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{reconstruction accuracy}} \underbrace{- D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \,||\, p(\mathbf{z})\right)}_{\text{regularization}}, \tag{4}$$

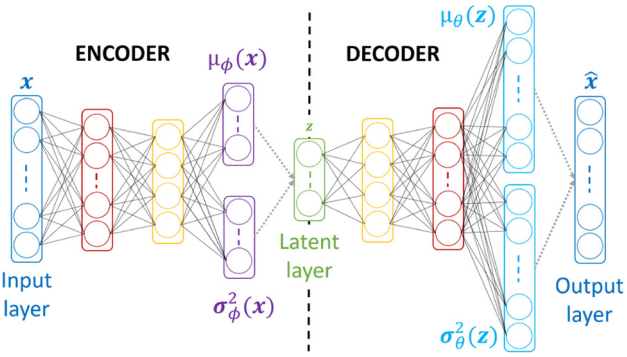


Figure 2: General architecture of a VAE. Grey dotted arrows represent sampling processes.

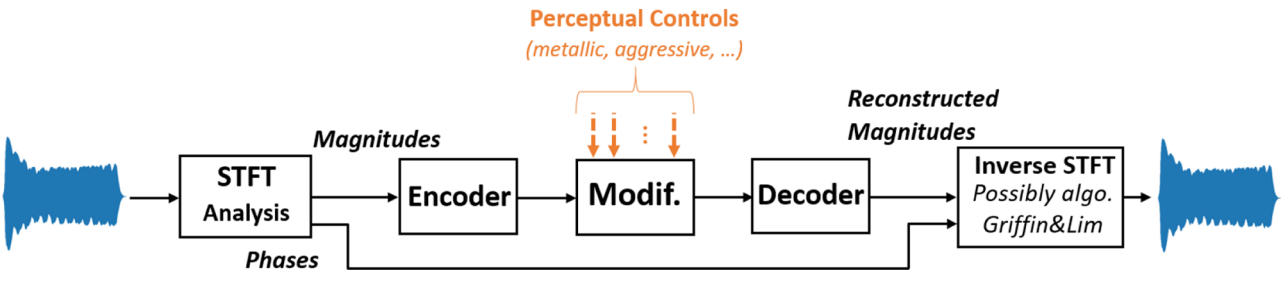


Figure 1: Global diagram of (V)AE-based sound analysis-transformation-synthesis.

where $D_{\mathrm{KL}}(\cdot\|\cdot) \geq 0$ denotes the Kullback-Leibler (KL) divergence between two distributions. In practice, the model is trained by maximizing $\mathcal{L}(\phi,\theta,\mathbf{x})$ with respect to the parameters ϕ and θ on a set of training data vectors. As we can see in Eq. (4), the lower-bound is composed of two terms: the first term represents the average reconstruction accuracy and the second term acts as a regularizer, encouraging $q_\phi(\mathbf{z}|\mathbf{x})$ to be close to the prior $p(\mathbf{z})$. The maximization of the VLB involves an iterative combination of sampling, stochastic gradient ascent (applied in practice on mini-batches of data) with error backpropagation through the decoder and encoder layers, and parameter updating. For more technical details about VAE training, the reader is referred to Kingma and Welling (2014).

The above "conventional VAE" was later extended to a β-VAE where β is a weighting coefficient introduced in the VLB to arbitrarily control the balance between the reconstruction and the regularization terms (Blaauw and Bonada, 2016; Higgins et al., 2017):

$$\mathcal{L}(\phi,\theta,\beta,\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \beta D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})\right). \quad (5)$$

Indeed, for some applications it is important to control the tradeoff between the quality of the generated/reconstructed signal and the organization of the latent space. In practice, β is set empirically so that the values of the reconstruction and regularization terms are in the same range.

2.3 Perceptual regularization

Although VAEs allow to extract an interesting high-level representation space for speech and audio signals with good interpolation properties (Blaauw and Bonada, 2016; Hsu et al., 2017a; Roche et al., 2019), the extracted dimensions may not be perceptually meaningful (Esling et al., 2018). In this subsection, we describe the method used to train the VAE where the latent space is forced to match the perceptual dimensions identified in our listening tests.

2.3.1 Perceptual score vectors

For each sound of a subset $\mathcal{X}_L$ of our dataset (and thus for each vector $\mathbf{x}$ extracted from this sound), a *perceptual score vector* (PSV) $\mathbf{d}(\mathbf{x})$ was defined, in which each entry represents the magnitude of a perceptual dimension, as rated by human listeners in the continuous range [−1,1]. See Section 3 for more details on the perceptual test and the eight dimensions considered.

2.3.2 Perceptually regularized VLB

Following the approach of Esling et al. (2018) and Pati and Lerch (2020), we inserted an additional regularization term in the VLB of Eq. (5). In these studies, this additional regularization term aimed at encouraging the properties of the latent space to match those of the perceptual/attribute space by minimizing the difference between pairwise distances in the latent space and in the perceptual space. In our case however, the additional term is intended to encourage the latent space to match the perceptual dimensions identified in the listening test. Formally, our perceptually regularized VLB is written as:

$$\begin{aligned}\mathcal{L}(\phi,\theta,\beta,\alpha,\mathbf{x}) = {} & \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] \\ & - \beta\, \mathrm{D}_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})\right) \\ & + \alpha\, \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\mathcal{R}\left(\mathbf{z},\mathbf{d}(\mathbf{x})\right)\right],\end{aligned} \quad (6)$$

where $\mathcal{R}\left(\mathbf{z},\mathbf{d}(\mathbf{x})\right)$ is the perceptual regularization term based on the distance between the latent vector $\mathbf{z}$ and the perceptual score vector $\mathbf{d}(\mathbf{x})$ of the sample sound from which $\mathbf{x}$ is extracted. α is a corresponding weighting factor which has a similar role as β for the classic KL divergence term.

The number of perceptual dimensions P is limited (P = 8 in our experiment, see Section 3). However, the number of latent dimensions (i.e. the size of $\mathbf{z}$) must be large enough to maintain good quality of the whole encoding-transformation-decoding process. We therefore chose to force the first P entries of $\mathbf{z}$, i.e. the subvector $\mathbf{z}_{1:P}$, to match the perceptual score vector. The other entries of $\mathbf{z}$ are left free to encode other aspects of the sounds (although nothing prevents them from also encoding some aspects of the considered perceptual dimensions). Regarding the function $\mathcal{R}$, no assessment was made on the relationship between the PSV values and the actual perception of potential users. We therefore chose a simple metric (square error) in the present study:

$$\mathcal{R}\left(\mathbf{z},\mathbf{d}(\mathbf{x})\right) = \|\mathbf{z}_{1:P} - \mathbf{d}(\mathbf{x})\|^2. \quad (7)$$

2.3.3 2-step learning procedure

Since the labeled subset of sounds resulting from our listening test is limited in size (see Section 3), we could not use supervised training. We therefore focused on a semi-supervised approach. Hinton and Salakhutdinov (2007) presented a semi-supervised learning method operating in two steps: the first step is to train the model in an unsupervised manner, using all available data to extract "sensible, high-level features." The second step consists in refining the model (i.e. fine-tuning) using only the labeled data. Following both Hinton and Salakhutdinov (2007) and Esling et al. (2018), we investigated the use of a 2-step learning procedure to add perceptual regularization to a VAE model. In a first step, the model is trained in an unsupervised manner, using both the unlabeled dataset $\mathcal{X}_U$ and the labeled dataset $\mathcal{X}_L$, and maximizing the weighted VLB of Eq. (5). Then, the VAE is fine-tuned, using the labeled dataset $\mathcal{X}_L$ only, and maximizing the regularized VLB of Eq. (6). Thus, the proposed methodology can be summarized as:

1. **Unsupervised pre-training:**
   Maximize $\mathcal{L}(\phi,\theta,\beta,\mathbf{x})$ from Eq. (5) on $\mathcal{X}_U \cup \mathcal{X}_L$,
2. **Supervised fine-tuning:**
   Maximize $\mathcal{L}(\phi,\theta,\beta,\alpha,\mathbf{x})$ from Eq. (6) on $\mathcal{X}_L$.

3 Perceptual Description of Synthesizer Sounds

3.1 First perceptual test: Free verbalization

A first perceptual test was conducted in order to identify the most relevant perceptual dimensions underlying the perception of synthesizer sounds, and their most shared verbal descriptors.

3.1.1 Stimuli

First, a large audio dataset (referred to as the *ARTURIA dataset*)[2] was created, consisting of 1,233 audio samples generated with varying ARTURIA software applications,[3] having the same pitch (E3, 165 Hz), a similar duration (2 to 2.5 seconds) and normalized in loudness.

Fifty stimuli were selected from the *ARTURIA dataset* to cover as broadly as possible the range of acoustic variation of these sounds. To this end, a k-means algorithm was applied from the acoustic characterization of the dataset, using 12 classical audio descriptors.[4] The 50 stimuli were then randomly chosen from the 50 clusters obtained.

Finally, a subset of 20 stimuli was assigned to each participant, half of which was shared with the preceding listener. Thus, each of the 50 stimuli received a comparable number of evaluations (21.5 ± 2.82 evaluations on average).

3.1.2 Participants and task

This first perceptual test was undertaken by 101 French-speaking listeners. None of them reported any hearing disorder.[5] The test was conducted online, using a self-developed web interface based on the Web Audio Evaluation Tool (Jillings et al., 2015). Twenty stimuli were successively presented. Participants could listen to each sound as many times as they wanted. They were asked to give verbal descriptions of each sound (at least one, at maximum five) using preferably isolated words or very short sentences, and trying to avoid aesthetic judgments (like "beautiful" or "ugly"). Five input fields were displayed on the web page, in which participants could type in these descriptions. The test lasted approximately 20 minutes.

3.1.3 Clustering of verbal expressions by semantic proximity

A first step of data "cleaning" and reduction consisted in correcting typos, and grouping verbal expressions with a shared lexical root (e.g. *brillant* and *qui brille*). This resulted in a set of 784 verbal expressions. Most of them were "classical" descriptors, such as *métallique*, *chaud* or *brillant*[6] that have already been reported in previous studies to distinguish the timbre of orchestral instruments (Reymore and Huron, 2020; Faure, 2000; Zacharakis, 2013), to qualify timbre variations within some instrument categories (piano, violin, guitar, voice, etc.) (Cheminée et al., 2005; Traube, 2004; Fritz et al., 2012; Garnier et al., 2007), or to qualify the timbre of purely synthetic sounds (von Helmholtz, 1875). However, almost half of them were also new expressions such as *spatial*, *robotique* or *saccadé*[7] that had, to our knowledge, never been reported in the literature.

We then evaluated the semantic proximity of these expressions, both within individuals (i.e. when two expressions were used together more than twice by a participant to describe a sound) and between individuals (i.e. when two expressions were used by two different participants to describe the same set of sounds). These analyses were performed from the 3D occurrence matrix **M** of the collected expressions, of size (number of expressions) × (number of participants) × (number of stimuli). Then, for each participant (denoted S_k), we evaluated the co-occurrence of pairs of expressions within the same listener ($\mathbf{J}^{\text{intra}}$) and between pairs of listeners ($\mathbf{J}^{\text{inter}}$), based on the Jaccard distance metric (Jaccard, 1912):

$$\begin{cases} J^{\text{intra}}_{i,j,S_k} = d_J\left(\mathbf{M}(i,S_k,:),\mathbf{M}(j,S_k,:)\right), \\ J^{\text{inter}}_{i,j,S_k} = \dfrac{1}{K-1}\displaystyle\sum_{\substack{l=1 \\ l\neq k}}^{K} d_J\left(\mathbf{M}(i,S_k,:),\mathbf{M}(j,S_l,:)\right), \end{cases}$$

where d_J denotes the Jaccard distance, i and j are the indices of the pair of expressions considered, and ":" is shorthand for "all entries in that dimension". This metric is particularly well adapted for our study as it does not take zeros into account. The above matrices were then averaged across participants, and combined into a single final similarity matrix $\mathbf{J} = 0.5(\mathbf{J}^{\text{intra}} + \mathbf{J}^{\text{inter}})$. Finally, we applied a hierarchical agglomerative clustering algorithm (HAC) (Day and Edelsbrunner, 1984) on **J** to group the verbal expressions into categories. We used Ward's aggregation index for the clustering (i.e. intra-cluster variance minimization). The final number of groups was chosen by manually detecting the largest decrease in homogeneity (i.e. the intra-category inertia) between two successive clusters. This resulted in 98 final perceptual categories, containing eight expressions on average (from 2 to 21). The complete process is illustrated in **Figure 3**.

3.1.4 Identification of perceptual dimensions and corresponding verbal labels

The "strength", or relevance, of each semantic cluster was then evaluated, based on (i) its occurrence frequency (defined as the total number of occurrences of every expression in the cluster) and (ii) its transversality (defined as the percentage of listeners who used at least once one of the cluster's expressions). This allowed us to

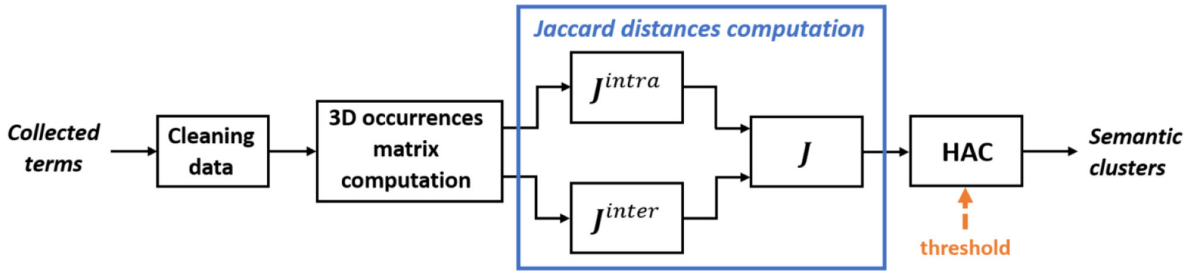


Figure 3: Global diagram of the semantic proximity analysis.

select the eight most frequent and transverse semantic clusters, supposed to correspond to the most relevant perceptual dimensions to describe and control the timbre of synthesized sounds. Finally, a last step was to select the most representative verbal descriptor for each of these eight perceptual dimensions. We only considered qualifiers and ignored nouns (designating objects or acoustic sources), and selected again the most frequent and transverse verbal descriptor within each semantic cluster: *métallique*, *chaud*, *soufflé*, *qui vibre*, *percussif*, *qui résonne*, *qui évolue* and *agressif* (whose equivalent expressions in English could be: *metallic*, *warm*, *breathy*, *vibrating*, *percussive*, *resonating*, *evolving* and *aggressive*). The results are summarized in **Table 1**.

In the end, the verbal descriptors selected for the eight main perceptual dimensions had all previously been reported in the literature of musical timbre. Two of them: *metallic* and *warm* are quite unanimously considered as main perceptual dimensions of musical timbre in previous studies (Zacharakis, 2013; Reymore and Huron, 2020; Kendall et al., 1999; Faure, 2000; Fritz et al., 2012). Three others: *breathy*, *resonating* and *vibrating* were also considered in some studies, but not in others (*resonating* and *vibrating* were even combined as one and the same dimension by Reymore and Huron (2020)). Finally, the two remaining descriptors: *evolving* and *aggressive* did not appear in the list of main perceptual dimensions of previous studies, although *aggressive* could be compared to *harsh*, *hard* or *strident*.

3.2 Second perceptual test: Verbal Attribute Magnitude Estimation

A second perceptual test was conducted, based on the Verbal Attribute Magnitude Estimation (VAME) method, to get quantitative evaluations of timbre for a subset of synthesizer sounds, along the eight perceptual dimensions previously identified, and use these ratings as perceptual score vectors for the regularization of our VAE.

3.2.1 Stimuli

Eighty stimuli were selected from the *ARTURIA* dataset so as to be as representative as possible of its acoustic space (by applying again a k-means algorithm to its acoustic characterization, as described in Section 3.1.1). These stimuli were split into a training subset (10 stimuli) and a main subset (70 stimuli). The 10 training stimuli were always presented at the beginning of the test so that participants could become familiar with the perceptual dimensions. They were followed by 30 additional stimuli, 25 of which were randomly selected from the main subset and five from the training subset (in order to evaluate the intra-listener agreement). On average, the samples of the main subset were evaluated by 26 different participants and samples from the training subset received 36 different evaluations.

3.2.2 Participants and task

This second perceptual test was fully completed by 71 French-speaking participants. None of them reported any hearing disorder.[5] The test was also conducted online, using a self-developed web interface based on the Web Audio Evaluation Tool (Jillings et al., 2015). Forty stimuli were successively presented without informing the participants that the 10 first examples were considered as training, and that five of them would be reintroduced among the 30 following examples. Participants could listen to each sound as many times as necessary. They were asked to evaluate its timbre according to the eight perceptual dimensions *métallique*, *chaud*, *soufflé*, *qui vibre*, *percussif*, *qui résonne*, *qui évolue* and *agressif*, using continuous scales implemented with horizontal sliders starting from *pas du tout* (*not at all*, scored as −1) to *extrêmement* (*extremely*, scored as 1). The slider was initially positioned in the middle (score 0). No definition of the eight perceptual dimensions was given to participants. However, they were asked to give their own understanding of each verbal descriptor in written form afterward. The test lasted approximately 20 minutes.

Table 1: Frequency and transversality measures of the eight most frequent and transverse semantic clusters, represented by the most frequent and transverse verbal descriptor within each cluster. The frequency of the semantic clusters is expressed as percentages of the evaluated sounds for each participant and their transversality as percentages of the total number of participants. The frequency and transversality of the verbal descriptors are expressed as percentages of the expressions within each cluster.

Semantic Cluster	**Frequency** (in %)		**Transversality** (in %)	
	Semantic Cluster	Isolated Verbal Desc.	Semantic Cluster	Isolated Verbal Desc.
Qui résonne (cluster of 8 expressions)	13.5	25.7	47.5	37.5
Métallique (cluster of 4 expressions)	10.6	52.6	43.6	75.0
Agressif (cluster of 4 expressions)	9.9	47.6	40.6	48.8
Qui vibre (cluster of 7 expressions)	7.8	43.5	46.5	40.4
Chaud (cluster of 4 expressions)	7.7	45.8	36.6	40.5
Qui évolue (cluster of 8 expressions)	5.7	27.0	29.7	33.3
Soufflé (cluster of 5 expressions)	4.5	43.5	25.7	57.7
Percussif (cluster of 4 expressions)	3.6	37.8	25.7	26.9

3.2.3 Analysis and results

Intra-listener agreement was explored for each participant and each perceptual dimension by computing the Pearson correlation between the ratings of the five training sounds that were presented twice. The different dimensions showed an average intra-rater agreement ranging from R = 0.38 to R = 0.81 (see **Table 2**), indicating that the two dimensions *qui vibre* and *qui résonne* were intrinsically more difficult to evaluate, compared to the others. Participants who demonstrated a correlation lower than 0.5 for a given perceptual dimension were not considered as reliable enough and were excluded from further analysis of that dimension. The number of remaining participants for each perceptual dimension is reported in the second column of **Table 2**.

Inter-listener agreement was then analyzed for each perceptual dimension by computing, for each pair of participants, the Pearson correlation between the ratings of the stimuli that they both evaluated (from 4 to 29 common stimuli, depending on the listeners' pair). The average inter-listener agreement for each perceptual dimension is given in the third column of **Table 2**. Only the dimensions *percussif* and *agressif* showed an acceptable degree of inter-listener agreement (greater than 0.5). The lower levels of inter-listener agreement observed for the other dimensions may indicate that the verbal descriptors were understood differently by the listeners. To explore this further, we performed, for each dimension, a hierarchical agglomerative clustering (HAC) in order to distinguish groups of participants who may share a common conception of the verbal descriptors.[8] This resulted in two or three groups for each perceptual dimension. Most of them showed a higher degree of inter-listener agreement (see last three columns of **Table 2**). Finally, for each perceptual dimension, we selected the largest group of participants corresponding, from an application point of view, to the majority of users of our synthesizer, and considered the average evaluation score given to each sound by the participants in that group. As a result, each of the 80 stimuli was described by a eight-value vector corresponding to the average evaluation of the sound timbre along the perceptual dimensions.[9]

4. Experiments with Perceptually Regularized VAEs

4.1 Data pre- and post-processing

For magnitude and phase spectrogram extraction, we applied a 1024-point STFT to each input waveform (sampled at 44.1 kHz) using a sliding Hamming window with 50% overlap. Silent portions at the beginnings and ends of signals were removed. The resulting 513-point positive-frequency magnitudes were converted to log-scale and normalized in energy: we set the maximum of each log-magnitude spectrogram to 0 dB (the corresponding scale factor was stored to be used for signal reconstruction). Then, every log-magnitude below a fixed threshold value of −100 dB was set to −100 dB, i.e. the spectrogram range was clipped to [−100, 0] dB. Finally, the spectrogram was linearly rescaled to [−1, 1], which is a usual procedure for DNN inputs. Corresponding denormalization, rescaling and log-to-linear conversion were applied to the decoded magnitude spectrogram. Waveform reconstruction was achieved by combining the resulting magnitude spectrogram with the original phase spectrogram, and then applying inverse STFT with overlap-add and optionally the Griffin & Lim algorithm (see Section 2.1).

Our experiments were conducted using the entire *ARTURIA* dataset described in Section 3.1.1 and containing 1,233 synthesizer sound samples. This dataset was split into a training set (80%) and a testing set (20%). The unlabeled dataset $\mathcal{X}_U$ consisted of the normalized

Table 2: Intra- and inter-listener agreement on the eight perceptual dimensions, for the second perceptual test. The first two columns give information on the intra-listener agreement: the average Pearson's coefficient R and the percentage of participants showing an $R > 0.5$. The four last columns report the levels of inter-listener agreement observed over the whole group of participants and for each group of participants identified from the HAC analysis (the percentage of participants in each group being reported in brackets).

Perceptual dimension	Intra-listener agreement		Inter-listener agreement			
	Average Pearson's R	% of part. for whom $R > 0.5$	Average Pearson's R (% of participants)			
			All	1st cluster (selected)	2nd cluster	3rd cluster
Métallique	0.59	69.0%	0.38	0.47 (59.2%)	0.40 (40.8%)	
Chaud	0.50	64.8%	0.36	0.48 (43.5%)	0.39 (37.0%)	0.45 (19.5%)
Soufflé	0.58	66.2%	0.31	0.40 (53.2%)	0.35 (46.8%)	
Qui vibre	0.38	49.3%	0.23	0.42 (62.9%)	0.30 (37.1%)	
Percussif	0.81	87.3%	0.56	0.62 (85.5%)	0.40 (14.5%)	
Qui résonne	0.41	57.7%	0.23	0.33 (56.1%)	0.27 (43.9%)	
Qui évolue	0.54	67.6%	0.42	0.47 (70.8%)	0.49 (29.2%)	
Agressif	0.68	81.7%	0.51	0.58 (60.3%)	0.60 (27.6%)	0.57 (12.1%)

log-magnitude spectra of the unrated sounds, computed as described above. The labeled dataset $\mathcal{X}_L$ was composed of the normalized log-magnitude spectra of the 80 sounds rated in the VAME test (see Section 3), associated to the 80 corresponding perceptual score vectors (PSVs) (note that the PSV of a sound is used to label all the successive spectral vectors that compose the normalized log-magnitude spectrogram of that sound). To maximize the use of the labeled data, when dividing the dataset into training and test sets, we ensured that all the 80 labeled sounds were contained in the training set and not in the test set. Following the methodology explained in Section 2.3, the unsupervised pre-training was made using $\mathcal{X}_U \cup \mathcal{X}_L$, composed of 243,861 513-dimensional spectral vectors and the supervised fine-tuning was performed using $\mathcal{X}_L$, containing 16,011 vectors.

4.2 Regularized VAE implementation

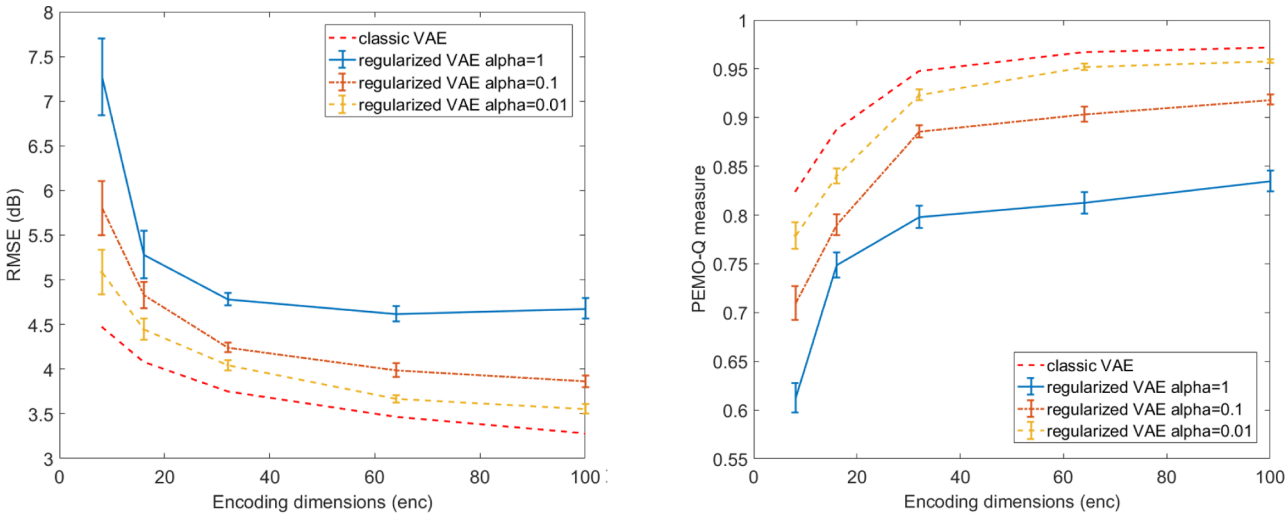
Considering the results reported by Roche et al. (2019) and other similar experiments that we conducted with the *ARTURIA* dataset, we focused in this experiment on a VAE model of the form [513, 128, *enc*, 128, 513] (this vector contains the number of neuron units on the successive layers). We investigated different values for the encoding dimension *enc* ranging from 8 to 100. Note that 8 is the size of the perceptual space evidenced in Section 3, so when $enc = P = 8$ we have $\mathbf{z}_{1:P} = \mathbf{z}$ in Eq. (7). We used tanh and linear activation functions for the hidden and output units respectively. β was empirically set to 0.25 so that the reconstruction term and the conventional KL regularization term were in the same range. Concerning the weighting coefficient α for the proposed perceptual regularization, we tested three values: 0.01, 0.1 and 1.

All models were implemented in Python using the *Keras* toolkit[10] with *tensorflow* backend (Abadi et al., 2016). The implementation of the perceptually-regularized VAE is mainly based on the implementation of the VAE provided in the Keras VAE tutorial, for which we added the perceptual regularization loss described in Eq. (6). The training was performed using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 10^{-3} and with a batch size of 512. For the unsupervised pre-training phase, 20% of the training set was kept for validation and an early stopping criterion with a patience of 30 epochs was used on the validation loss to avoid overfitting. During the fine-tuning stage, the model was forced to train for 600 epochs.

4.3 Analysis-resynthesis experiments

First, we evaluated the perceptually-regularized VAE in an objective way, by comparing the different versions of the model (including the baseline VAE model without perceptual regularization) in an analysis-synthesis framework, without modification of the latent vector **z**. This was done by computing the root-mean squared error (RMSE) between original and reconstructed spectrograms from the test set, as well as the PEMO-Q score between the original and reconstructed waveforms (Huber and Kollmeier, 2006). The PEMO-Q is an objective measure of audio quality, based on an auditory perception model, defined in the range [0,1] (the higher the better). Note that the proposed perceptual regularization is expected to deteriorate the quality of the reconstructed signal, especially in terms of RMSE. Indeed, the perceptual regularization term in Eq. (6) is balanced with the reconstruction error term that is directly related to the RMSE within the current Gaussian data model. Note also that the effects of the perceptual regularization term and the conventional KL regularization term are cumulative. We wish to quantify this degradation and also assess it perceptually with the PEMO-Q score.

Figure 4 shows that the RMSE and PEMO-Q results present similar (opposite) behaviors, with a performance increasing logarithmically with the encoding size. As expected, increasing α deteriorates the quality of the reconstructed signal. In particular, for *enc* = 8, the



(a) RMSE (in dB) **(b)** PEMO-Q scores

Figure 4: Performance of the classic VAE and the proposed perceptually-regularized VAE in terms of **(a)** RMSE (in dB) and **(b)** PEMO-Q scores, for three values of α (error bars represent 95% confidence intervals calculated with paired t-tests considering the classic VAE as the reference).

regularization strongly deteriorates the quality of the signal, with variations in RMSE and PEMO-Q reaching respectively +62% and –25% for $\alpha = 1$. This result is not surprising, since this dimension corresponds to the size of the perceptual space and therefore, during fine-tuning, the model is encouraged to encode the successive spectra of a given sound (most often non-stationary) with constant values (the entries of the PSV) for *all* latent dimensions. In other words, the dynamics of the input signal are encoded through the dynamics of the latent trajectories, and for *enc* = 8, the signal dynamics were encouraged to be constant, which severely affects the quality of the output signal. Allowing the additional latent dimensions to freely encode the signal dynamics results in a rapid decrease in RMSE and increase in PEMO-Q score from *enc* = 8 to *32*. The evolution becomes slower after 32, and the results are statistically very consistent. For *enc* = 64, an α value varying from 0.01 up to 1 causes an increase in RMSE over the baseline of 6% up to 33% (respectively 1.6% to 16% decrease in PEMO-Q score). This shows that the setting of α can have a significant impact on the quality of the reconstructed signals.

4.4 Analysis of the latent space

The next step in evaluating the effectiveness of the proposed approach was to investigate how the structure of the latent space was modified by the perceptual regularization. We focused here on a VAE model where *enc* = 64 and $\alpha = 0.1$, since this setting appeared to achieve a good tradeoff between regularization and reconstruction accuracy.

4.4.1 Organization of the latent space

To analyze the organization of the latent space, we first computed the Spearman correlation coefficients (SCC) between the extracted latent dimensions obtained using samples from the labeled dataset. The first column of **Figure 5** shows that the classic VAE presents low correlations between the different dimensions extracted. The second column shows that the perceptual regularization significantly affected the eight constrained dimensions, while preserving the others. The bottom chart of **Figure 5** shows the SCC for the perceptual score vectors collected during the perceptual test of Section 3.2. We can see that the structure of the eight constrained VAE latent dimensions (second row) closely matches that of the PSVs, which shows the effectiveness of the perceptual regularization. Moreover, a closer observation of the correlation coefficients between the perceptually-constrained dimensions and all the other unconstrained dimensions (third row) shows that these latter dimensions were not impacted by the perceptual regularization and that they remained mostly uncorrelated from the perceptual dimensions.

4.4.2 Mapping and disentanglement evaluation

Some recent studies have proposed metrics to evaluate the mapping and disentanglement of extracted latent dimensions (Adel et al., 2018; Locatello et al., 2020; Pati and Lerch, 2020). In the present study, we are not particularly interested in the disentanglement of the constrained latent dimensions as this would assume that the perceptual dimensions themselves are uncorrelated, which is not the case. Indeed, if we focus on the dimension *métallique* for example, the bottom chart of **Figure 5** clearly shows that it is strongly correlated with the dimensions *qui vibre* and *agressif* while being negatively correlated with the dimension *chaud*, and uncorrelated with the dimensions *qui résonne*, *percussif* and *qui évolue*. Instead, we therefore rather considered four of the metrics presented by Pati and Lerch (2020) – interpretability, mutual information gap (MIG), separated attribute predictability (SAP) and maximum SCC between the extracted dimensions and the perceptual ratings – to compare our perceptually-regularized model with the classic VAE baseline.

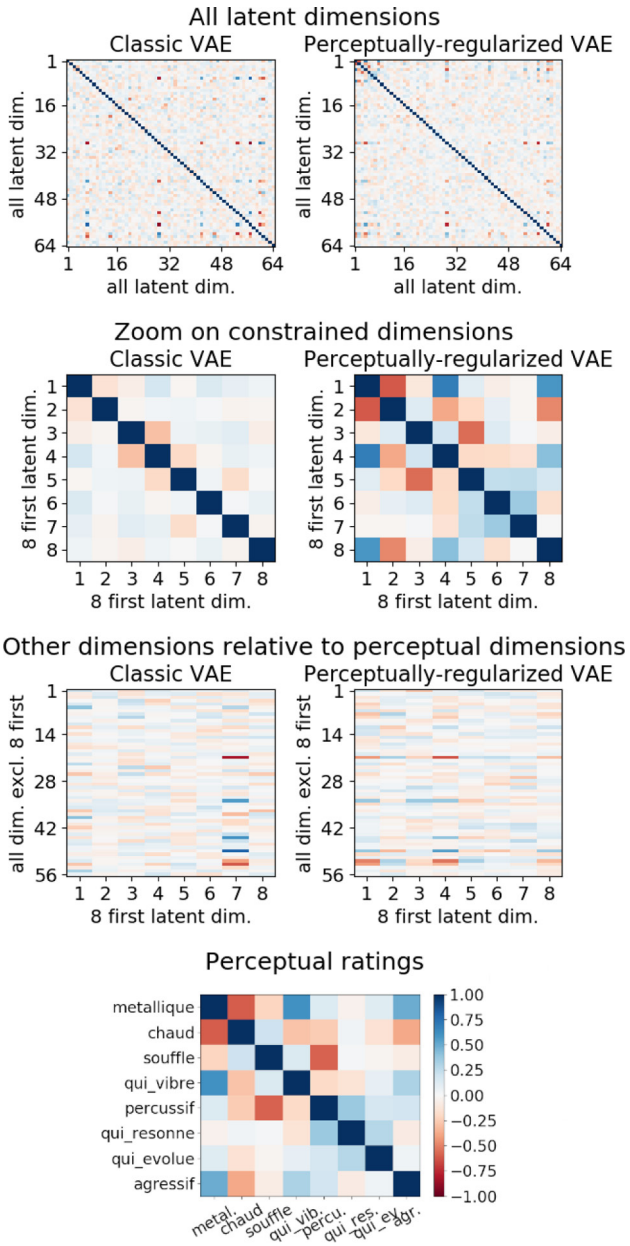


Figure 5: Spearman correlation coefficients between extracted latent dimensions (first three rows) and perceptual ratings (last row).

We first computed the interpretability measure for each of the perceptual dimensions. We compared the ratings obtained during the VAME test with the extracted latent dimensions obtained for both the classic VAE and the perceptually-regularized VAE. The results are illustrated in **Figure 6**. They clearly show the effectiveness of the perceptual regularization, increasing significantly the interpretability of the extracted dimensions.

We then computed the average scores obtained by the two models with the four mapping and disentanglement metrics (see **Table 3**). The results show that the perceptual regularization has a clear impact on the structure of the latent space and that the obtained perceptually-regularized latent space significantly outperforms the baseline for all the metrics (the higher the better according to Pati and Lerch (2020)).

4.5 Perceptual evaluation of the regularized model

Finally, to assess the impact of the proposed method at the perceptual level, we conducted a final A/B listening test. The main goal of this test was to evaluate the perceptual impact of a basic modification (offset) of the VAE latent coefficients.

4.5.1 Stimuli

For this test, we focused on the five perceptual dimensions that are more related to spectral characteristics: *métallique*, *chaud*, *soufflé*, *qui vibre* and *agressif*. We did not consider the dimensions *percussif*, *qui résonne* and *qui évolue* because our first attempts at controlling these dimensions with a VAE were not convincing, certainly due to the fact that they were more related to sound temporal dynamics and therefore poorly captured by the static VAE model.

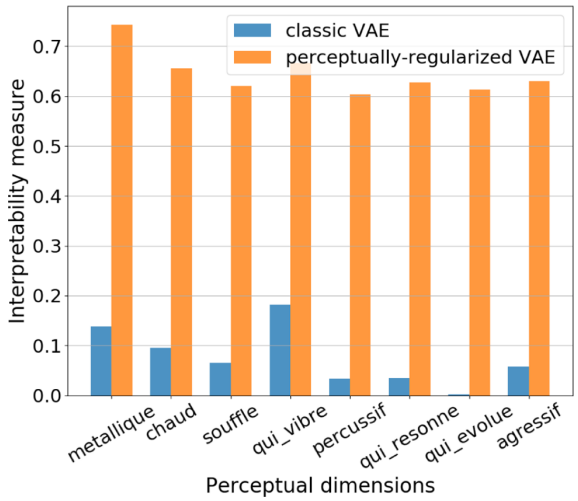


Figure 6: Interpretability measure (Pati and Lerch, 2020) for the first eight dimensions of the latent space.

For each of these five dimensions, we selected 12 sound samples. Six of these were selected from the labeled dataset (train): the three samples that received the highest mean rating (close to 1) on this perceptual dimension, with the lowest standard deviation (thus considered as very representative of this dimension), and the three samples that received the lowest mean rating (close to −1), with the lowest standard deviation (thus considered as unrepresentative of that dimension). Six other samples were randomly selected from the test dataset (unlabeled) to evaluate the generalization ability of the model.

The 60 selected sound samples (12 stimuli per perceptual dimension × 5 dimensions), were then transformed applying the analysis-modification-synthesis method described in Section 2.1. The modification consisted in adding a predefined constant offset to the trajectory of the latent coefficient corresponding to the perceptual dimension to be modified. From the results presented in Section 4.3, we decided to use the perceptually-regularized VAE with a [513, 128, 64, 128, 513] architecture, $\beta = 0.25$ and $\alpha = 0.1$. Since the modification of the latent coefficients can lead to significant changes in the decoded spectrogram, the waveform signal was reconstructed with the Griffin & Lim algorithm (Griffin and Lim, 1984).

Independently from the (targeted) modification of perceptual dimensions, the overall sound transformation process (analysis-transformation-synthesis) can produce audible artifacts in the reconstructed signal. To prevent these artifacts from biasing the test, we applied the same transformation process (with different offset values) to each pair of samples compared during the test: we presented two modified versions of the same source samples using two different offset values for the transformation instead of presenting the original source sample and a single modified version. The reference stimulus was set to the lowest offset for which we could notice a perceptual difference with basic encoding-decoding (without modifying the latent vectors). The "accentuated stimulus" was modified with a larger offset: we first searched for the threshold value for which all the reconstructed signals were perceptually identical (somehow saturating the decoder). Then, we set the actual offset value to 50% of this threshold (the resulting offset was always significantly larger than the small offset applied for the reference signal).

Examples of transformed sounds obtained with different offset values are available at the companion webpage.[11]

Table 3: Averaged mapping and disentanglement metrics (Pati and Lerch, 2020) obtained for the classic and perceptually-regularized VAE models.

	SCC	Interpretability	MIG	SAP
Classic VAE	0.3216	0.0762	0.0035	0.0264
Perceptually-regularized VAE	0.7895	0.6448	0.0513	0.4275

4.5.2 Participants and task

Thirty listeners participated in this third perceptual test.[5] The test was conducted again online, using a self-developed web interface based on the Web Audio Evaluation Tool (Jillings et al., 2015). Participants were presented with 60 successive pairs of samples, in a random order. Each pair was resynthesized from the same original sound sample modified with the two different offset values on the corresponding dimension. They were asked to choose, by clicking on the corresponding "A" or "B" button (each sample of the pair being randomly assigned to these buttons) which of the two samples sounded the most *métallique* for example. The test lasted about 20 minutes.

4.5.3 Analysis and results

The test outcome was encoded as a binary variable y (equal to 1 if the participant evaluation followed the targeted VAE transformation, and 0 otherwise). The statistical analysis was performed independently for each perceptual dimension. The experimental data was modeled using a logistic random effects regression, considering the listener and the pair of stimuli as random effects, and considering as a fixed effect the data subset from which the sound was selected (train vs. test).[12]

First, the data origin (train vs. test) was not found to influence the perception of any of the five perceptual dimensions significantly (see the first two columns of **Table 4**), meaning that the perceptually-regularized model was able to generalize and modify the acoustic characteristics of a sound sample even if this sample was not used for training our model ($p > 0.05$).[13] As a result, the data model was simplified to account for only the two random effects.

Figure 7 and the last three columns of **Table 4** summarize how well the perception of each dimension followed the intended sound transformation. Thus, the perceptual dimensions *agressif* and *qui vibre* turned out to be well captured by the model, with respectively 80.8% and 72.8% of "correct" answers (i.e. perceptual evaluations that follow the intended sound transformation), which was significantly above chance level ($p < 0.0001$). Conversely, the results cannot reject the hypothesis that the participants randomly evaluated the dimensions *chaud* and *soufflé*, with a percentage of "correct" answers of 51.9% and 52.8% respectively ($p > 0.05$). These results, combined with the fair degree of inter-listener agreement obtained on the dimension *soufflé* (0.36) may indicate that the participants of this test understood this verbal descriptor in a different way to the majority group of the second VAME test, whose ratings were used to train the VAE. For the dimension *chaud*, the low level of both "correct" answers (51.9%, $p = 0.36$) and inter-rater agreement (0.08) indicate that the model failed to capture the perceptual dimension as intended. Finally, the perception of the dimension *métallique* tended to follow relatively well the intended VAE transformation with a percentage of "correct" answers greater than chance (60.8%) and a fair inter-rater agreement (0.29), although this tendency was not statistically significant ($p = 0.07$).

5. Conclusions and Perspectives

5.1 Conclusions

In this study, we proposed a new method to perceptually regularize a variational autoencoder (VAE) model, using a semi-supervised learning procedure. This method allowed

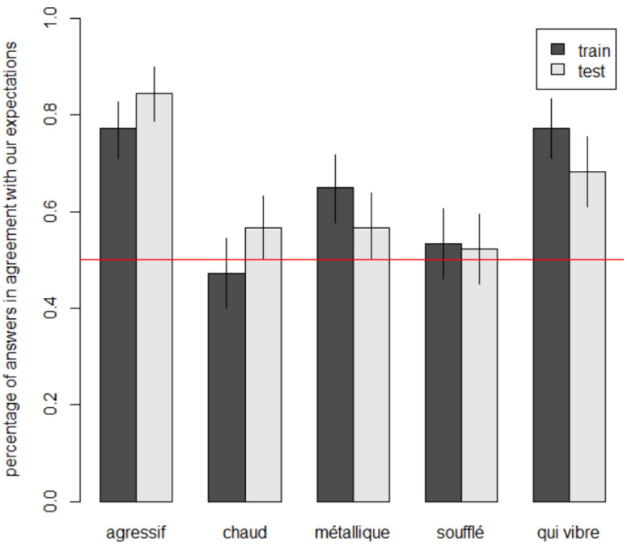


Figure 7: Results of the A/B perceptual test for five perceptual dimensions and for labeled stimuli (train) and new unknown ones (test). Bars represent mean values, error bars represent 95% confidence interval, and the red line indicates chance level.

Table 4: Statistical results of the perceptual A/B test results for the five selected perceptual dimensions. Effect of the train/test data origin factor (first two columns). Comparison of the perceptual choice (A/B) with chance level (third and fourth columns). Inter-listener agreement using Randolph's free-marginal multi-rater kappa (Randolph, 2005) (last column).

Dimension	**"train/test dataset" factor**		**Chance threshold comparison**		**Inter-listener agreement**
	χ^2	p-value	z	p-value	Randolph's κ
Agressif	0.026	0.87	3.39	≪0.0001	0.50
Chaud	1.07	0.30	0.37	0.36	0.08
Métallique	0.43	0.61	1.47	0.07	0.29
Soufflé	0.002	0.96	0.25	0.40	0.36
Qui vibre	0.76	0.38	3.93	≪0.0001	0.27

us to obtain perceptually relevant control parameters for the transformation of synthesizer sound timbre. To our knowledge, no clear definition of relevant perceptual dimensions and corresponding verbal descriptors has already been provided in the literature for synthesizer sounds. We therefore conducted a first perceptual test to identify these relevant perceptual dimensions and their associated labels in French: *métallique*, *chaud*, *soufflé*, *qui vibre*, *percussif*, *qui résonne*, *qui évolue*, *agressif*. We then conducted a second VAME test to get an entire sound dataset rated along these perceptual dimensions. We then used these ratings as perceptual score vectors for the regularization of our VAE.

Our experiments allowed us to draw several conclusions. As expected, we first observed that the additional regularization slightly degraded the quality of the audio signals generated by the model but that this quality remained acceptable when choosing an appropriate weighting factor α. This issue may be overcome in the future by considering bigger datasets (both labeled and unlabeled $\mathcal{X}_L$ and $\mathcal{X}_U$). We also observed that using this extra regularization increased the interpretability of the constrained dimensions and modified their behavior to relate closely to the perceptual ratings obtained from our listening test. Finally, we conducted a last perceptual test to get a preliminary evaluation of how well the perceptually-regularized model performed sound transformation. This experiment validated the proposed methodology, showing that the model was relatively good at capturing and modifying the acoustic properties of the dimensions *agressif* and *qui vibre*. It was not efficient for the dimensions *chaud* and *soufflé* and the results on the dimension *métallique* are not so clear and would deserve further investigation. Furthermore, the model was able to generalize well to unseen sounds, even though the labeled dataset was very small.

5.2 Perspectives

A first perspective is to further study the semantic relationships between the eight perceptual dimensions, in order to better understand their potential redundancy, or inclusion. Then, in the present study, we implemented perceptual regularization by computing the MSE between the VAE latent vectors and the perceptual score vectors collected during the second perceptual test. However, the relationship between these vectors may not be linear and several metrics may possibly be more appropriate for this application. It might also be interesting to investigate further how normalizing flows (Rezende and Mohamed, 2015) can be used to organize the latent space of our model, as done by Adel et al. (2018) and Esling et al. (2020), to benefit from a more complex and flexible posterior distribution model. Finally, one of the limitations of the present VAE model is that it is a static model, i.e. the input vectors (extracted from the input sound spectrogram) are processed one-by-one independently. Recently, the VAE model and corresponding variational training methodology have been extended to dynamic models, including the modeling of temporal dependencies between consecutive observed and/or latent vectors (Chung et al., 2015; Fraccaro et al., 2016; Hsu et al., 2017b; Krishnan et al., 2017). Our future work will consider extending the proposed study to those dynamic models, which should improve the model's ability to capture and manipulate dynamic perceptual dimensions such as *qui évolue* or *percussif*.

Notes

[1] The choice of the exact data representation to feed the VAE model will be detailed further in Section 4.1.

[2] This dataset is publicly available at http://doi.org/10.5281/zenodo.4680486.

[3] Sounds were generated from all factory presets of the ARTURIA software applications resulting in single-pitched sounds with a sample rate of 44.1 kHz.

[4] Attack time, attack slope, decay time, jitter, shimmer, MFCC, spectral centroid, spectral bandwidth, spectral contrast, spectral flatness, spectral roll-off and zero-crossing rate; see Peeters et al. (2011) for a review of commonly used audio descriptors and their implementations.

[5] The participants freely agreed to participate in the three online perceptual tests. No identifying information was asked or stored.

[6] Which could be translated to English as *metallic*, *warm* and *bright*.

[7] The closest expressions in English would be *space*, *robotic* and *jerky*.

[8] Clustering was performed on the correlation matrices (converted into distance matrices beforehand using $D = 1 - C$ where D and C are the distance and correlation matrices respectively) and the method to obtain the final clusters was similar to that presented in Section 3.1.3.

[9] These vectors were then used as perceptual score vectors when regularizing our VAE.

[10] https://keras.io.

[11] http://synthsounds.eu/article_companion.html.

[12] We used the *glmer* function of the *lme4* package of the R software; https://CRAN.R-project.org.

[13] This analysis was conducted using the *anova* function of the R software.

Funding Information

This work was supported in part by a CIFRE PhD grant funded by the ANRT (Association Nationale de la Recherche et de la Technologie).

Competing Interests

The authors have no competing interests to declare.

Author Contributions

Fanny Roche contributed to the design of the method, made the implementation and experiments, and contributed to the analysis of the results. All other authors contributed equally to scientific supervision in the design of the method and in the analysis of the results. All authors contributed equally to the writing of the paper and agreed to the published version of the manuscript.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *Proceedings of the USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, pages 265–283. Savannah, GA, USA.

Adel, T., Ghahramani, Z., and **Weller, A.** (2018). Discovering interpretable representations for both deep generative and discriminative models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 50–59. Stockholm, Sweden.

Blaauw, M., and **Bonada, J.** (2016). Modeling and transforming speech using variational autoencoders. In *Proceedings of the Conference of the International Speech Communication Association (Interspeech)*. San Francisco, CA, USA. DOI: https://doi.org/10.21437/Interspeech.2016-1183

Çakir, E., and **Virtanen, T.** (2018). Musical instrument synthesis and morphing in multidimensional latent space using variational convolutional recurrent autoencoders. In *Proceedings of the Audio Engineering Society Convention*, New York, NY, USA.

Cheminée, P., Gherghinoiu, C., and **Besnainou, C.** (2005). Analyses des verbalisations libres sur le son du piano versus analyses acoustiques. In *Proceedings of the Conference on Interdisciplinary Musicology (CIM05)*, Montréal, Canada.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A., and **Bengio, Y.** (2015). A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2980–2988, Montréal, Canada.

Colonel, J., Curro, C., and **Keene, S.** (2017). Improving neural net auto-encoders for music synthesis. In *Proceedings of the Audio Engineering Society Convention*, New York, NY, USA.

Day, W., and **Edelsbrunner, H.** (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1(1): 7–24. DOI: https://doi.org/10.1007/BF01890115

Donahue, C., McAuley, J., and **Puckette, M.** (2019). Adversarial audio synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA.

Dubois, D. (2000). Categories as acts of meaning: The case of categories in olfaction and audition. *Cognitive science quarterly*, 1(1): 35–68.

Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., and **Roberts, A.** (2019). GANSynth: Adversarial neural audio synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA.

Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., and **Simonyan, K.** (2017). Neural audio synthesis of musical notes with Wavenet autoencoders. In *Proceedings of the International Conference on Machine Learning (ICML)*, Sydney, Australia.

Esling, P., Chemla-Romeu-Santos, A., and **Bitton, A.** (2018). Generative timbre spaces with variational audio synthesis. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Aveiro, Portugal.

Esling, P., Masuda, N., Bardet, A., Despres, R., and **Chemla-Romeu-Santos, A.** (2020). Flow synthesizer: Universal audio synthesizer control with normalizing flows. *Applied Sciences*, 10(1): 302. DOI: https://doi.org/10.3390/app10010302

Faure, A. (2000). *Des sons aux mots, comment parle-t-on du timbre musical ?* PhD thesis, Ecole des Hautes Etudes en Sciences Sociales (EHESS).

Fraccaro, M., Sønderby, S. K., Paquet, U., and **Winther, O.** (2016). Sequential neural models with stochastic layers. In *Advances in Neural Information Processing Systems (NIPS)*, Barcelona, Spain.

Fritz, C., Blackwell, A., Cross, I., Woodhouse, J., and **Moore, B.** (2012). Exploring violin sound quality: Investigating english timbre descriptors and correlating resynthesized acoustical modifications with perceptual properties. *The Journal of the Acoustical Society of America*, 131(1): 783–794. DOI: https://doi.org/10.1121/1.3651790

Garnier, M., Henrich, N., Castellengo, M., Sotiropoulos, D., and **Dubois, D.** (2007). Characterisation of voice quality in western lyrical singing: From teachers' judgements to acoustic descriptions. *Journal of Interdisciplinary Music Studies*, 1(2): 62–91.

Girin, L., Hueber, T., Roche, F., and **Leglaive, S.** (2019). Notes on the use of variational autoencoders for speech and audio spectrogram modeling. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Birmingham, UK.

Goodfellow, I., Bengio, Y., and **Courville, A.** (2016). *Deep learning*. MIT Press. http://www.deeplearningbook.org.

Grey, J. (1977). Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America*, 61(5): 1270–1277. DOI: https://doi.org/10.1121/1.381428

Grey, J., and **Moorer, J.** (1977). Perceptual evaluations of synthesized musical instrument tones. *The Journal of the Acoustical Society of America*, 62(2): 454–462. DOI: https://doi.org/10.1121/1.381508

Griffin, D., and **Lim, J.** (1984). Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2): 236–243. DOI: https://doi.org/10.1109/TASSP.1984.1164317

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and **Lerchner, A.** (2017). β-vae: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France.

Hinton, G., and **Salakhutdinov, R.** (2007). Using deep belief nets to learn covariance kernels for Gaussian processes. In *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada.

Hsu, W.-N., Zhang, Y., and **Glass, J.** (2017a). Learning latent representations for speech generation and transformation. In *Proceedings of the Conference of the International Speech Communication Association (Interspeech)*, Stockholm, Sweden. DOI: https://doi.org/10.21437/Interspeech.2017-349

Hsu, W.-N., Zhang, Y., and **Glass, J.** (2017b). Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1878–1889, Long Beach, CA, USA.

Huber, R., and **Kollmeier, B.** (2006). PEMO-Q: A new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6): 1902–1911. DOI: https://doi.org/10.1109/TASL.2006.883259

Iverson, P., and **Krumhansl, C.** (1993). Isolating the dynamic attributes of musical timbre. *The Journal of the Acoustical Society of America*, 94(5): 2595–2603. DOI: https://doi.org/10.1121/1.407371

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, 11(2): 37–50. DOI: https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Jillings, N., Moffat, D., De Man, B., and **Reiss, J.** (2015). Web Audio Evaluation Tool: A browser-based listening test environment. In *Proceedings of the Sound and Music Computing Conference (SMC)*, Maynooth, Ireland.

Kendall, R. A., Carterette, E. C., and **Hajda, J. M.** (1999). Perceptual and acoustical features of natural and synthetic orchestral instrument tones. *Music Perception*, 16(3): 327–363. DOI: https://doi.org/10.2307/40285796

Kingma, D., and **Ba, J.** (2015). Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

Kingma, D., and **Welling, M.** (2014). Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Banff, Canada.

Krimphoff, J., McAdams, S., and **Winsberg, S.** (1994). Caractérisation du timbre des sons complexes. ii. analyses acoustiques et quantification psychophysique. *Le Journal de Physique IV*, 4(C5): C5–625. DOI: https://doi.org/10.1051/jp4:19945134

Krishnan, R., Shalit, U., and **Sontag, D.** (2017). Structured inference networks for nonlinear state space models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, San Francisco, CA, USA.

Krumhansl, C. (1989). Why is musical timbre so hard to understand? *Structure and Perception of Electroacoustic Sound and Music*, 9: 43–53.

Lichte, W. (1941). Attributes of complex tones. *Journal of Experimental Psychology*, 28(6): 455. DOI: https://doi.org/10.1037/h0053526

Locatello, F., Tschannen, M., Bauer, S., Rätsch, G., Schölkopf, B., and **Bachem, O.** (2020). Disentangling factors of variation using few labels. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Marozeau, J., de Cheveigné, A., McAdams, S., and **Winsberg, S.** (2003). The dependency of timbre on fundamental frequency. *The Journal of the Acoustical Society of America*, 114(5): 2946–2957. DOI: https://doi.org/10.1121/1.1618239

McAdams, S. (2019). The perceptual representation of timbre. In *Timbre: Acoustics, Perception, and Cognition*, pages 23–57. Springer. DOI: https://doi.org/10.1007/978-3-030-14832-4_2

McAdams, S., Beauchamp, J., and **Meneguzzi, S.** (1999). Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. *The Journal of the Acoustical Society of America*, 105(2): 882–897. DOI: https://doi.org/10.1121/1.426277

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and **Krimphoff, J.** (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58(3): 177–192. DOI: https://doi.org/10.1007/BF00419633

Miller, J., and **Carterette, E.** (1975). Perceptual space for musical structures. *The Journal of the Acoustical Society of America*, 58(3): 711–720. DOI: https://doi.org/10.1121/1.380719

Miranda, E. (2002). *Computer sound design: Synthesis techniques and programming*. Music Technology series. Focal Press.

Pati, A., and **Lerch, A.** (2020). Attribute-based regularization of latent spaces for variational autoencoders. *Neural Computing and Applications*, pages 1–16. DOI: https://doi.org/10.1007/s00521-020-05270-2

Peeters, G., Giordano, B., Susini, P., Misdariis, N., and **McAdams, S.** (2011). The Timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5): 2902–2916. DOI: https://doi.org/10.1121/1.3642604

Randolph, J. (2005). Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. In *Joensuu Learning and Instruction Symposium*, Joensuu, Finland.

Reymore, L., and **Huron, D.** (2020). Using auditory imagery tasks to map the cognitive linguistic dimensions of musical instrument timbre qualia. *Psychomusicology: Music, Mind, and Brain*. DOI: https://doi.org/10.1037/pmu0000263

Rezende, D., Mohamed, S., and **Wierstra, D.** (2014). Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the International Conference on Machine Learning (ICML)*, Beijing, China.

Rezende, J., and **Mohamed, S.** (2015). Variational inference with normalizing flows. In *Proceedings of the International Conference on Machine Learning (ICML)*, Lille, France.

Roche, F., Hueber, T., Limier, S., and **Girin, L.** (2019). Autoencoders for music sound modeling: A comparison of linear, shallow, deep, recurrent and variational models. In *Proceedings of the Sound and Music Computing Conference (SMC)*, Málaga, Spain.

Samson, S., Zatorre, R., and **Ramsay, J.** (1997). Multidimensional scaling of synthetic musical timbre: Perception of spectral and temporal characteristics. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 51(4): 307. DOI: https://doi.org/10.1037/1196-1961.51.4.307

Traube, C. (2004). *An interdisciplinary study of the timbre of the classical guitar*. PhD thesis, McGill University.

von Bismarck, G. (1974). Timbre of steady sounds: A factorial investigation of its verbal attributes. *Acta Acustica united with Acustica*, 30(3): 146–159.

von Helmholtz, H. (1875). *On the sensations of tone as a physiological basis for the theory of music*. Longmans, Green. DOI: https://doi.org/10.1037/10838-000

Wedin, L., and **Goude, G.** (1972). Dimension analysis of the perception of instrumental timbre. *Scandinavian Journal of Psychology*, 13(1): 228–240. DOI: https://doi.org/10.1111/j.1467-9450.1972.tb00071.x

Wessel, D. (1979). Timbre space as a musical control structure. *Computer Music Journal*, 3: 45. DOI: https://doi.org/10.2307/3680283

Zacharakis, A. (2013). *Musical timbre: Bridging perception with semantics*. PhD thesis, Queen Mary University of London.

How to cite this article: Roche, F., Hueber, T., Garnier, M., Limier, S., & Girin, L. (2021). Make That Sound More Metallic: Towards a Perceptually Relevant Control of the Timbre of Synthesizer Sounds Using a Variational Autoencoder. *Transactions of the International Society for Music Information Retrieval,* 4(1), pp. 52–66. DOI: https://doi.org/10.5334/tismir.76

Submitted: 21 September 2020 **Accepted:** 29 March 2021 **Published:** 18 May 2021

]u[*Transactions of the International Society for Music Information Retrieval* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS