FLOWSYNTH: SEMANTIC AND VOCAL SYNTHESIS CONTROL

Philippe Esling, Naotake Masuda, Adrien Bardet, Romeo Despres, Axel Chemla
IRCAM - CNRS UMR 9912, Sorbonne Université, Paris, France
esling@ircam.fr, axel.chemla@unimi.it

EXTENDED ABSTRACT

Since their commercial beginnings, audio synthesizers have flourished in music production, and now even entirely defining new music genres. While becoming widely accessible and pervasive in music, their increasing complexity and number of parameters also rendered their use concomitantly difficult. Hence, the development of methods allowing to easily create and explore with synthesizers is a crucial need. Unfortunately, no synthesizer provides intuitive or semantic controls related to the *perception* of the synthesized audio. A solution proposed by some manufacturers is *macro-parameters*, which control multiple parameters through a single knob, but still need to be programmed manually. An alternative would be to infer the set of parameters that could best reproduce a given *target sound*. This *parameter inference* task has been studied in the past years using various optimization techniques such as genetic programming [2] or deep learning [4]. However, all previous approaches share the same flaws that they do not account for non-linear relationships between parameters, do not learn explanatory parameters, nor provide *semantic* controls over synthesizers.

Recently, we introduced a novel framework and generic formalization of the problem of audio synthesizers control [1]. We cast this problem as finding an invertible mapping between an *auditory latent space* $\mathbf{z}$ that represents the synthesizer capabilities (based on audio examples $\mathbf{x}$) and the *space of its parameters* $\mathbf{v}$. Hence, this defines a generative model $p(\mathbf{x}, \mathbf{v}, \mathbf{z}) = p(\mathbf{x}|\mathbf{v}, \mathbf{z})p(\mathbf{v}|\mathbf{z})p(\mathbf{z})$. To solve this new formulation, we relied on normalizing flows [3], by introducing the idea of *regression flow* to map one latent space to another one. This model performs the parameter regression either as *posterior parameterization* or *conditional amortization*). We showed that this proposal is able to address simultaneously the tasks of *parameter inference*, and *unsupervised macro-control learning* within a single model [1]. However, the unsupervised dimensions do not relate to specific semantic controls and rather define some principal directions of audio variations.

Here, we address this learning of semantic controls, where we aim to use categorical tags to organize latent dimensions in a supervised way. We introduce *disentangling flows*, which steer the organization of some dimensions to match target distributions. Hence, this enforces a form of supervised disentanglement, providing latent dimensions explicitly linked to target semantic tags, smoothly mapping to the synthesizer parameters. By using this formulation, we show that we can address *semantic dimension learning*. Finally, we introduce an experimental Max4Live and VST interface that implements all of these mechanisms[1].

First, we define a model that accounts for *semantic tags* by expanding latent factors $\mathbf{z}$ with a categorical variable $\mathbf{t}$. Hence, we define $p_\theta(\mathbf{x}|\mathbf{t}, \mathbf{z})$ where $p(\mathbf{t}) = Cat(\mathbf{t}|\pi)$ and $p(\pi)$ is the prior distribution of the tags. We define the inference model $q_\phi(\mathbf{z}, \mathbf{t}|\mathbf{x})$ and assume that it factorizes as $q_\phi(\mathbf{z}, \mathbf{t}|\mathbf{x}) = q_\phi(\mathbf{z}|\mathbf{x})q_\phi(\mathbf{t}|\mathbf{x})$. When $\mathbf{t}$ is unknown, it is considered as a latent variable over which we can perform posterior inference

$$\mathcal{L}_u = -\mathbb{E}\left[\log p_\theta(\mathbf{x}|\mathbf{t}, \mathbf{z}) + \log p_\theta(\mathbf{t}) + \log p_\theta(\mathbf{z})\right] - \mathbb{E}\left[\log q_\phi(\mathbf{t}, \mathbf{z}|\mathbf{x})\right] \quad (1)$$

When $\mathbf{t}$ is known, we take a rather unusual approach through the idea of *disentangling flows*. As we seek to obtain a latent dimension with continuous semantic control, we define a tag pair as a set of negative $\mathbf{t}_-$ and positive $\mathbf{t}_+$ samples. We define two *target* distributions $p(z_{\mathbf{t}_-}) \sim \mathcal{N}(-\mu_*, \sigma_-)$ and $p(z_{\mathbf{t}_+}) \sim \mathcal{N}(+\mu_*, \sigma_+)$ that model samples of a semantic pair as opposite sides of a latent dimension. Hence, we turn the treatment of tags into a *density estimation* problem, where we aim to match tagged samples densities $\mathbf{t}_* \in \{\mathbf{t}_-, \mathbf{t}_+\}$ to our

[1] Source code results and plugins are available on a supporting webpage: https://acids-ircam.github.io/flow_synthesizer/

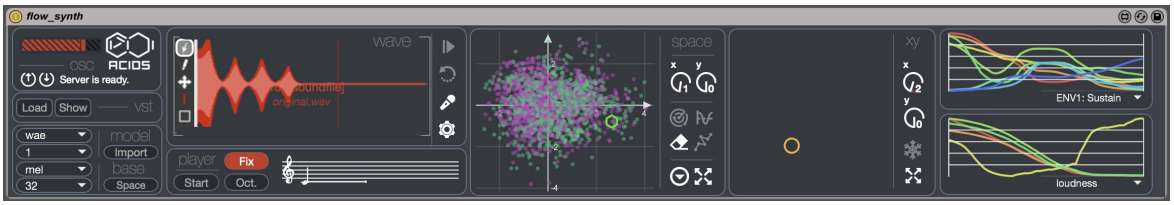

Figure 1. *FlowSynth* interface for audio synthesizer control in Ableton Live.

targets by minimizing $\mathcal{D}_{\text{KL}}\left[q_\phi(z_{\mathbf{t}_*}|\mathbf{x})\|p(z_{\mathbf{t}_*})\right]$. To solve this, we consider that $q_\phi(z_{\mathbf{t}_*}|\mathbf{x})$ is parameterized by a normalizing flow f_k applied to the latent $\mathbf{z}$, leading to our observed objective

$$\mathcal{L}_o = \mathcal{D}_{\text{KL}}\left[q_\phi(z_{\mathbf{t}_*})\|p(z_{\mathbf{t}_*})\right] = \mathbb{E}\left[\log p(\mathbf{z}) - \sum_{i=1}^{k} \log\left|\det\frac{\partial f_i}{\partial \mathbf{z}_{i-1}}\right| - \log p(z_{\mathbf{t}_*})\right] \tag{2}$$

This formulation enforces a form of *supervised* disentanglement, where latent $\mathbf{z}$ are transformed to provide dimensions with explicit target properties, and the full loss optimized is $\mathcal{L} = \mathcal{L}_u + \mathcal{L}_o$.

We constructed a dataset of synthesizer sounds and parameters for *Diva* [2], as detailed in [1], but our model can work for any synthesizer. Here, we briefly summarize results that are detailed on the supporting webpage. Compared to the unsupervised model [1], this semantic model appears to slightly impair the audio accuracy. However, the model still outperform all baseline models, while providing the huge benefit of explicit semantic macro-controls. We also evaluated the robustness to increasing sets of parameters (from 16 to 64) and observed that, while our proposal also suffers from larger sets of parameters, it appears as the most resilient and manages this higher complexity. We also evaluate *out-of-domain generalization* with a set of audio samples produced by other synthesizers, orchestral instruments and direct vocal imitations, where the overall distribution of scores remains consistent with previous observations. However, it seems that the average error is quite high, indicating a potentially distant reconstruction of some examples. Upon closer listening, it seems that the models accurately reproduces the temporal shape of target sounds, but the local timbre structure is somewhat distant. Finally, our proposed *disentangling flows* can steer the organization of selected latent dimensions so that these dimensions provide *macro-parameters* with a clearly defined semantic meaning. We studied the traversal of these semantic dimensions and observed that, while the parameters evolution is smooth, it exhibits non-linear relationships between different parameters. This correlates with the intuition that there are complex interplays in parameters of a synthesizer. Regarding the effect of different semantic tags, it appears that some pairs provides very intuitive results, while others are harder to interpret. In any case, our proposal appears successful to uncover *semantic macro-parameters* for a given synthesizer.

We implemented both the unsupervised model [1] and the supervised semantic model detailed here in an experimental Max4Live and VST interface that is displayed in Figure 1. This interface wraps the Diva VST and allows to provide control based on both proposed models. Hence, this interface allows to input a wave file or direct vocal recording to perform *parameter inference*, to perform pathes across the audio space to both rely on *unsupervised macro-control* and also to explore *supervised semantic dimensions*.

REFERENCES

[1] Philippe Esling, Naotake Masuda, Adrien Bardet, Romeo Despres, et al. Universal audio synthesizer control with normalizing flows. In *International Conference on Digital Audio Effects (DaFX 2019)*, 2019.

[2] Ricardo A Garcia. Automatic design of sound synthesis techniques by means of genetic programming. In *Audio Engineering Society Convention 113*, 2002.

[3] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning (ICML)*, 2015.

[4] Matthew John Yee-King, Leon Fedden, and Mark d'Inverno. Automatic programming of vst sound synthesizers using deep networks and other techniques. *IEEE Transactions on ETCI*, 2(2), 2018.

[2] https://u-he.com/products/diva/