Timbre transfer between orchestral instruments with semi-supervised learning

Adrien Bitton, Axel Chemla-Romeu-Santos, Philippe Esling
bitton@ircam.fr - IRCAM, 1 Place Igor Stravinsky, 75004 Paris, France

Motivations

We aim to provide **new ways of synthesizing timbres** by high-level interaction and **transfer of properties** between instruments.
Our hypothesis is that each instrument defines a **timbral domain**.

Challenges in prior timbre studies
non-invertible analysis spaces of perceptual ratings
⇨ does not generalize nor synthesize audio
audio descriptors with limited correlation to timbre spaces
⇨ little predictive power
DSP techniques with complex sound decompositions ⇨ no knowledge model and high number of parameters / analysis dimensions

Our proposal
. variational learning for finding high-level structured representations
. joint optimization of **analysis and generation processes**
. dimensionality reduction onto latent spaces of higher abstraction
. no need of annotations nor ratings
. building a **common latent space** between instruments

I. Machine Learning Background

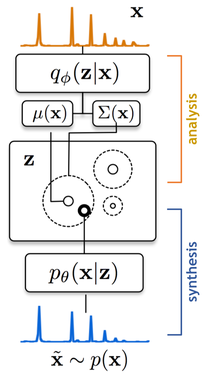
Variational Auto-Encoder **(VAE)**
Modeling the data distribution $p(x)$ based on lower-dimensional latent representation z that retrieves x so $p(x,z) = p(x|z)p(z)$

Approximate solution through variational inference over a parametric family of candidate distributions ⇨ evidence lower bound optimization of encoder q_φ mirrored with decoder p_θ

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

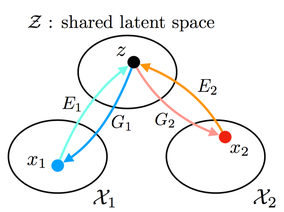
data reconstruction/generation
Negative Log-Likelihood (NLL)

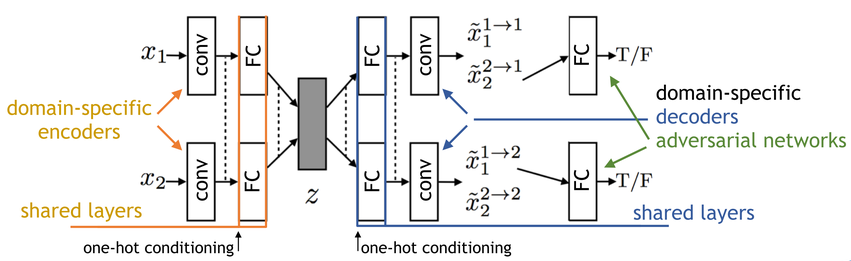
latent structuration
Kullback-Leibler Divergence (KLD)



- such models are light and fast to train
- variational learning is effective on small datasets (around 10^3 samples)
- latent space is **structured, generative and disentangles data variations**

Unsupervised Translation Networks **(UNIT)**
Paired data domains that map onto a common knowledge space
hypothesis: a **shared latent space allows transfer** from one to the other
no matched samples across domains, instead, an additional adversarial objective pushes separate decoder layers to match their domain attributes from any latent coordinates





⇨ competitive optimization: the generator produces more realistic samples while the discriminator becomes more accurate
⇨ **domain translation by switching decoders**
⇨ reinforced by a circle-consistency objective so that the reverse transfer should reconstruct the original sample

II. Timbre transfer with generative models

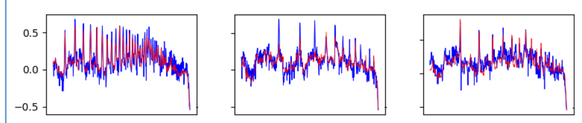
Studio On Line dataset **(SOL)**
. note recordings of 12 orchestral instruments (winds, strings, keyboard, brass) in several **dynamics, playing styles and all pitches**
. Non-Stationary Gabor Transform **(NSGT)** as input signal transform for its beneficial representation and invertibility properties
. 2D-(de)convolutional layers on the data domains to process blocks of NSGT-Mel frames ⇨ **spectro-temporal features** of ~120ms context

Timbre transfer strategies

Instrument-conditional VAEs switching the encoding condition to any decoding target condition
⇨ indexed latent subspaces w.r.t. instrument conditions
UNIT-like translators paired domains with single instruments, switching decoder from one to the other
⇨ shared latent space but one-to-one transfer
possible semitone-conditioning ⇨ subspaces abstracted from pitch
⇨ control over (un)transposed transfers
Across instrument families domains are groups of instruments
⇨ VAEs are instrument-conditional w.r.t. their family
switching decoders and selecting target instrument condition
⇨ more versatile one-to-many transfers

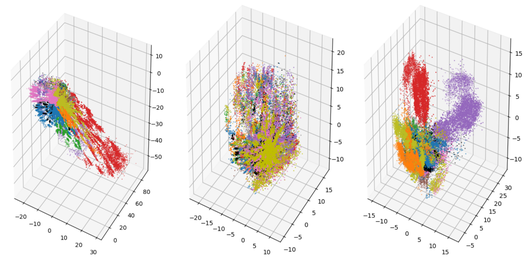
III. Results

Validating a base framework for further experiments (e.g.)
. warmup procedures of the different training objectives
. layer capacities, initializations, activations, batch-normalization
. optimizer, back-propagation and learning rate decays
⇨ good generative and generalization power with **down to 3 latent dimensions** (16*500 input dimensions) under the UNIT assumptions
⇨ 3D representation & synthesis spaces easing high-level interaction
note: generated audio samples are available for listening evaluation



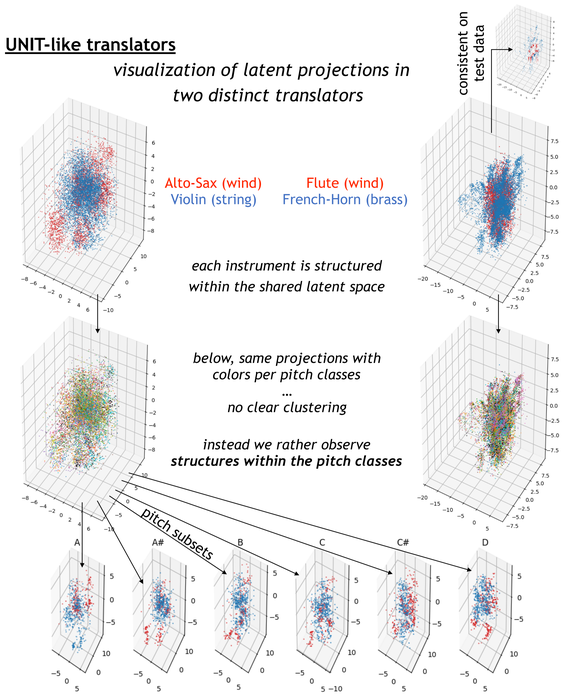
samples of spectral frames (scaled magnitude)
that were successfully auto-encoded (reconstruction in red)

Conditional VAEs
comparing latent auto-organizations of SOL

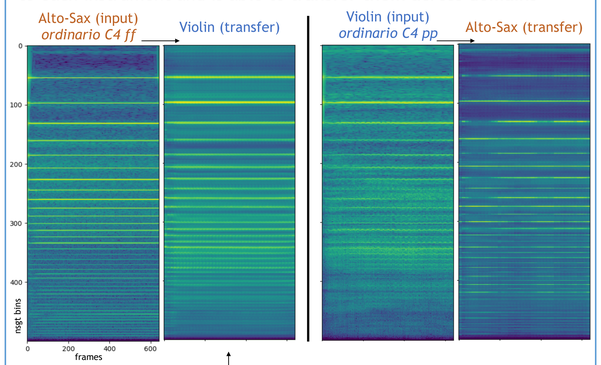


unconditioned VAE | instrument-conditional VAE (allows timbre transfer) | pitch-conditional VAE

UNIT-like translators
visualization of latent projections in two distinct translators

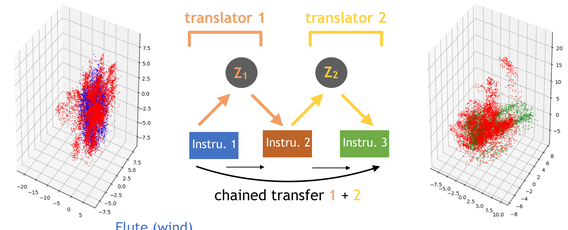


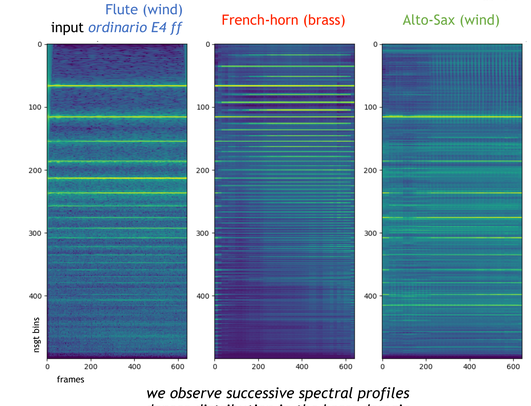
Hence the network jointly models pitch-independent features related to each instrument and is able to **transfer them across domains**



spectro-temporal structure clearly mimics violin, e.g.
higher inharmonicity and tristimulus (odd partials are louder than even ones)
bowing modulations

Chaining translator pairs
two separate trainings but as translators share a common instrument it enables a **chained transfer across that common data domain**





we observe successive spectral profiles
denser distribution in the brass domain
and sparser when transferring to the other wind instrument

Further experiments
Instrument-conditional family translators
Timbre trajectories as geodesic/constrained latent paths
Timbre interpolations and morphings in latent space

Conclusions and references

Our study provides **novel tools for interacting with timbre** data that do not suffer from previous limitations
⇨ 3D continuous representations that generalize and synthesize audio
⇨ techniques for timbre transfer/morphing and exploring novel tones that **efficiently visualize information and directly render sound**

Our framework can be applied on melodies, on sampled notes from synthesizers and could be adapted to non-musical sounds (e.g. tone-to-noise) which encourages **creative sonic applications**.

It also motivates further investigations, including:
. modeling timbre paths under constraints and along manifolds
. analyzing latent space correlations to signal descriptors
. developing methods for mapping across several translator pairs or adapting one translator to an other

Main references
Kingma et al. **Auto-Encoding Variational Bayes** arXiv:1312.6114 2013
Liu et al. **Unsupervised Image-to-Image Translation Networks** NIPS 2017
Balazs et al. **Theory, implementation and applications of Nonstationary Gabor frames** Journal of computational and applied mathematics 2011
Ballet et al. **Studio online 3.0: An internet "killer application" for remote access to IRCAM sounds and processing tools** JIM 1999