# Continuous Mood Conversion for Music

Anis, Kilian, Ondřej, Giorgia, Carlos

September 22, 2018

**Abstract**

This project's aim is to develop a system for imposing a particular mood to a given audio content. The work was inspired by style transfer and domain adaptation methods. In particular, we extended existing architectures for translation between discrete (instrument) domains to a continuous mood space (valence/arousal). We use a WaveNet autoencoder model. The encoder extracts the content of an audio signal, which is a mood-agnostic latent representation thanks to the applied confusion loss. The latent representation is transformed back to an audio signal by a single decoder that is globally conditioned on the mood.

# 1 Introduction

## 1.1 Task proposal

In this work, we introduce the problem of mood translation on music as a domain transfer task aiming at producing a musical rendition inspired from an input piece, that exhibits predefined mood characteristics. We propose to treat the mood of a song as some kind of texture parametrized by its coordinates in a certain mood space, and apply techniques coming from style transfer, and domain translation fields, to produce a rendering of a musical pieces with different mood textures. An interesting application would be a counterpart to music recommendation systems. Instead of recommending certain songs based on the user's music taste, our system adapts the mood of the songs to the user's current mood. The mood could be manually chosen by the user or retrieved through affective computing techniques (Poria et al., 2017) using data such as heart rate or temperature measurements and skin conductance. Due to the wide spread of smart devices, those data are easily obtained.

The valence/arousal plane (see 1) is a widespread way of representing a mood, the valence being closely related to the positivity of an emotion and the arousal to a notion of energy. This way of representing mood is convenient for a statistical approach to mood transfer, for it provides us with a continuous, vector representation that can be used as a conditioning variable in generative models.

## 1.2 Background

Music, being a form of art, is inherently linked to a notion of emotion, or mood. For this reason, a lot of efforts from the MIR community has been directed

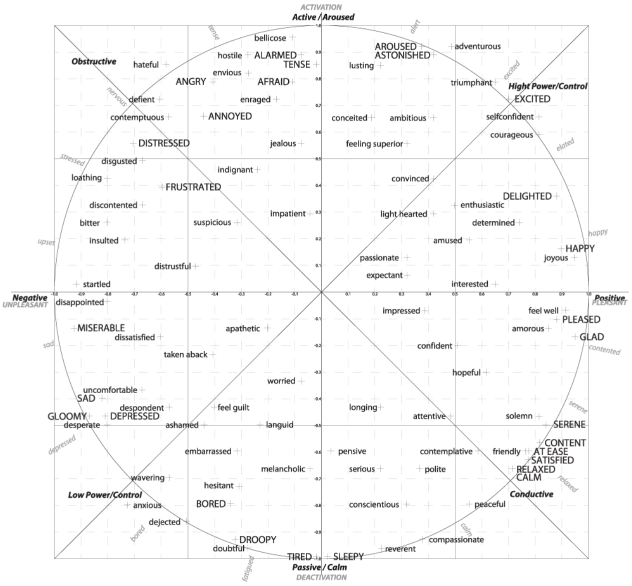

Figure 1: Example representation in the arousal/valence plane of English adjectives.

toward apprehending this link, whether it is by being able to predict it, or to control it. In (Remi Delbouys, 2018), the authors focused on the problems predicting a mood annotation from the audio content and lyrics of a song. Expressive rendering for automatic piano performances has been addressed in works like (Canazza et al., 2015), which aims at providing the user with a way to control the emotional expressiveness of the performed music by selecting a point or drawing a trajectory in a Valence-Arousal emotional space. This is quite similar to our idea, although is applied to a MIDI score, that is to a symbolic representation, rather than directly on an audio signal. As far as we know, parametrized *mood translation* has never been addressed as an end-to-end problem, very much akin to style transfer.

On the other side, there are few works on *audio style transfer*, i.e. transferring a given style to a specific audio content.

The problem of style transfer was firstly addressed in Computer Vision using Convolution Neural Networks (CNNs) Grinstein et al. (2017). In fact, the abstract features provided by a CNN optimized for object recognition, allowed the author of Gatys et al. (2016) to transfer the style of a particular target painting to the content of a photograph preserving his layout and structure. In general, it is not clear what exactly defines the style of an image, but in this case it is mainly defined as texture at various scales and color map. Following this approach, style transfer with Deep Learning had a big increase and it has become a trending topic (Grinstein et al., 2017).

The proposed approaches for audio style transfer use CNNs, or deep models which are more suitable for audio signals, to construct the target style signal starting from his properties. Recently, the authors of Grinstein et al. (2017) investigated the use of different CNNs architectures as well as of a human auditory system-inspired model with really promising results. Some work has been done on transferring sound textures (Grinstein et al., 2017), using techniques similar to those proposed for image style transfer by Gatys et al. (2016).

A recent paper by van den Oord et al. (2017) has shown impressive results[1] on speaker conversion for speech (although it lacks empirical evaluation). The authors train a variational autoencoder (VAE) with a WaveNet-like (van den Oord et al., 2016) architecture and a discrete latent code (representing content), with the decoder conditioned on a speaker ID (representing style). Then, they can encode speech from one speaker and perform reconstruction using a different speaker ID, obtaining the same speech rendered in a different voice.

A similar idea has been applied to music signals by Mor et al. (2018). They use WaveNets in an encoder-decoder structure to achieve domain translation between music pieces. The encoder takes an audio signal as input and outputs its latent representation, which is style-agnostic. Then, one of several decoders is applied to produce a music signal with the desired style. Each style (domain) has a dedicated decoder. In the paper that introduced the WaveNet (van den Oord et al., 2016), the authors show how local and global conditioning can be performed when doing WaveNet decoding. They take the example of the speech synthesis task, where WaveNet decoding without any local conditioning would only produce babbling, whereas conditioning on the speech textual content allow to produce articulate audio content. Global conditioning can be used for example to render speech with a specific person's voice.

This natural way of introducing local and global conditioning in the decoding step makes us very optimistic on WaveNets being the right tool to achieve convincing mood translation. In the following we present an architecture that aims, similarly to Mor et al. (2018), at extracting a mood-agnostic representation, which will act as a local condition, and which would use the position of the track in the valence arousal space as a global conditioning for the generation of the output signal. In the next section we give more detail about the architecture and the way we plan on training it.

# 2 Architecture

Our proposed model (depicted in section 2) is a WaveNet auto-encoder based on the work of Engel et al. (2017) and Mor et al. (2018). As in Mor et al., the model is trained to make the latent representation independent of the variable that we wish to control (i.e. mood). The main difference is that in our case, this variable is continuous. For this reason, we use only one decoder (as opposed to one decoder per target domain in Mor et al.), which is globally conditioned on the mood variable.

## 2.1 Auto-encoder

The first step is use a $\mu$-law coded version of the audio input file $X$ and input it to a wavenet encoder that gives us the signal $X_c$, which has the same dimension as the input itself. The reason why we have chosen a wavenet encoder is because it has been used on Mor et al. (2018) successfully for style transfer and because it is possible to apply a global and a local conditioning on a wave-net encoder-decoder using a straight forward way.

[1] `https://avdnoord.github.io/homepage/vqvae/`

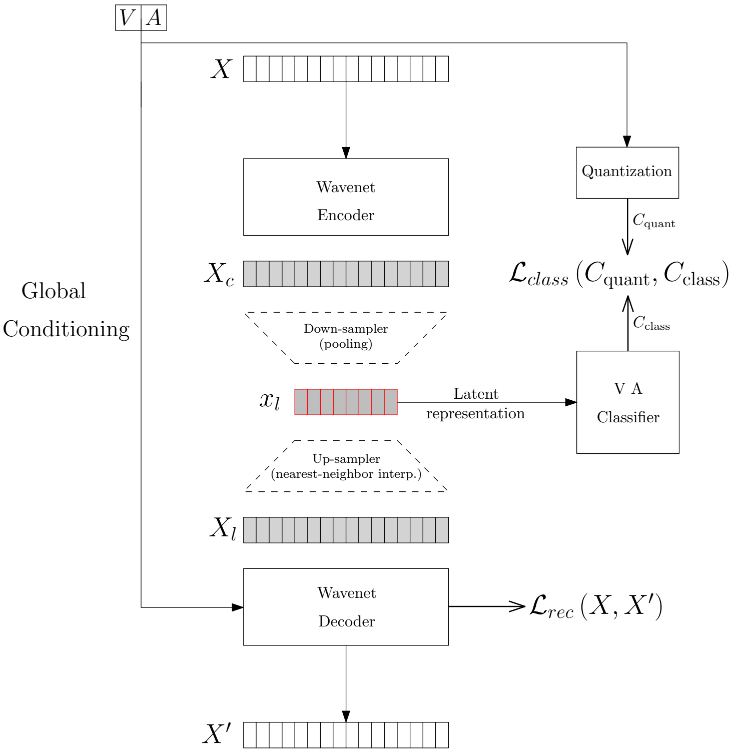

Figure 2: Architecture of the mood transfer system

After being down-sampled by a pooling technique and having its dimension reduced, we are able to obtain a latent representation $x_l$ for our music. The latent representation can be decoded after being up-sampled by a nearest neighbor interpolation. The decoder architecture is also a generative wave-net system, which is able to generate different versions $X'$ of the original song depending on the global conditioning value of Valence & Arousal.

During the training stage, the decoder is conditioned using the correct valence/arousal values given by the annotations on the musics on the Million Song Dataset (MSD) Remi Delbouys (2018).

## 2.2 Confusion network and training

To enforce the independence between the latent representation and the mood variable, we employ a confusion network, identical to that of Mor et al.. This network is an adversarial classifier trained to predict the mood of the source audio:

$$\min_{\theta_C} \mathcal{L}_{\text{class}} \tag{1}$$

Since our mood variable is continuous, we propose to quantize the valence-arousal space (using a vector quantization method such as $k$-means) and use the quantized values as labels for the classifier.

A confusion term is added to the reconstruction loss of the autoencoder, leading it to produce latent codes that are hard for the classifier to classify. The

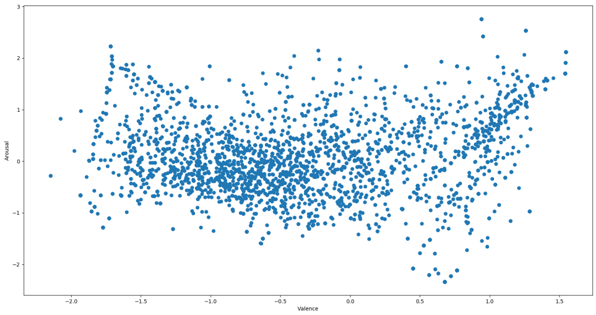

Figure 3: Distribution of the valence/arousal values in the mood MSD dataset.

auto-encoder objective then becomes

$$\min_{\theta_{\text{AE}}} \mathcal{L}_{\text{rec}} - \lambda \mathcal{L}_{\text{class}} \tag{2}$$

where $\lambda$ is a balancing hyperparameter associated with the confusion loss.

### 2.3 Training

The model will be trained on the Million Songs Dataset (MSD) with the additional annotations assigning mood coordinates to each song by Remi Delbouys (2018).

### 2.4 Evaluation metrics

A possible evaluation can be done using listening tests with humans to answer the following questions:

- How natural sounds the resulting music?
- Are all the type of music suitable for that application?
- Is the generated song somehow similar to the original one?

## 3 Research Questions

### 3.1 Valence/arousal distribution in the dataset

It is likely that being able to generate convincing mood from any input requires some care to be taken on the dataset creation, or at least how we sample it during training. In particular, we think that sampling songs that uniformly distributed in the valence/arousal space is highly important. In figure 3 is represented the valence/arousal distribution of the mood msd dataset Remi Delbouys (2018).

### 3.2 Training stability

To achieve mood-agnostic representation, we are simultaneously training the encoder and a classifier in an adverse fashion, much like the generator and discriminator of a GAN would be. It is known that such training schemes come with critical stability issues.

## References

Sergio Canazza, Giovanni De Poli, and Antonio Rodà. Caro 2.0: an interactive system for expressive music rendering. *Advances in Human-Computer Interaction*, 2015:2, 2015.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with WaveNet autoencoders. In *ICML*, 2017.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.

Eric Grinstein, Ngoc Q. K. Duong, Alexey Ozerov, and Patrick Pérez. Audio style transfer. *CoRR*, abs/1710.11385, 2017.

Noam Mor, Lior Wolf, Adam Polyak, and Yaniv Taigman. A universal music translation network. *arXiv preprint arXiv:1805.07848*, 2018.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

Francesco Piccoli Jimena Royo-Letelier Manuel Moussallam Remi Delbouys, Romain Hennequin. Music mood detection based on audio and lyrics with deep neural net. *ISMIR2018*, 2018.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. In *SSW*, 2016.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NIPS*, 2017.