TastyPiano: A Subjective Sound–Taste Synesthetic Machine

Cédric Colas
cdric.colas@gmail.com

Abstract

TastyPiano is a subjective sound–taste synesthetic machine that takes in an audio piano piece and imagines a matching cocktail recipe — thereby achieving a meaningful cross-modal domain transfer, from ears to taste buds.

1 Introduction

Could you ever see the color of happiness, feel the texture of the letter A, taste Duke Ellington's piano piece *Take the A Train*? These cross-modal experiences are called *synesthesia*, a rather strange perceptual phenomenon in which the stimulation of one sensory domain induces an involuntary experience in a different domain [1–3].[1] Poetry, literature and even everyday expressions often describe such experiences through synesthetic metaphors — "*bright sound*," "*warm color*," "*sweet voice*" [4, 5].

In his book *L'Écume des Jours*, Boris Vian imagines the *Pianocktail*, a magical machine turning any piano piece played on it into a sophisticated cocktail which sublimes the emotional experience induced by the piece — a sound–taste synesthetic machine. Can such a machine be called to life? With *TastyPiano* I use machine learning to democratize synesthesia, to *let you taste music*.

2 TastyPiano

A pianocktail implements a specific type of cross-modal domain transfer, from a source domain (piano pieces) to a target domain (cocktail recipes). Any good domain transfer requires: 1) *locality*: local variations in the source domain (similar piano pieces) must correspond to local variations in the target domain (similar cocktails); 2) *semantic alignment*: a piece of music must be translated into the *corresponding* cocktail [6]. Because synesthetic experiences are *subjective* in essence, semantic alignment refers to the *subjectivity* of a particular pianocktail implementation.

Pipeline. An audio piano piece first enters the *TastyPiano*, gets transcribed to the symbolic *midi* format, summarized into a *piece representation*, mapped to a corresponding *cocktail representation* before it is finally decoded into a *cocktail recipe* (Fig. 1).

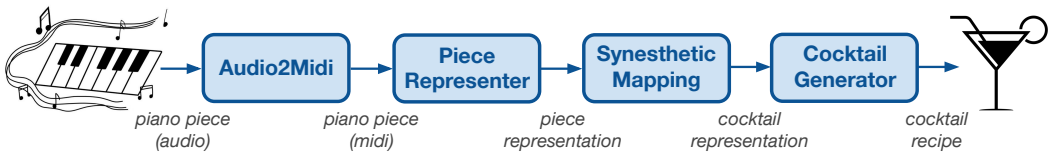

Figure 1: Main components of the TastyPiano system.

[1] To get a sense of the diversity of synesthetic experiences, visit reddit's *r/synesthesia* or the *synesthesia tree*.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Datasets. I curated a dataset of 30k piano midi files from existing sources [7–13] and by transcribing YouTube and Spotify playlists with the model from [14]. For cocktails, I curated a dataset of 600 recipes with ingredients, preparation type and serving glass.

Learning Piece Representations. I first encode midi pieces into sequences using the *structured encoding* [15, 16]: — 4 tokens per note (pitch, velocity, duration, time shift). Then, I learn piano piece representations in two steps. Step 1 – BERT pretraining [17]: this optimizes token embeddings to maximize the architecture's ability to retrieve masked notes, see also MidiBERT [18] and MusicBERT [19]. Step 2 – SentenceBERT finetuning [20]: this finetunes piece representations computed as the average of token embeddings to improve the capture of perceptual distances between pieces. More precisely, it uses a *contrastive loss* to get the embeddings of two sub-sequences from the same piece closer while getting embeddings of two sub-sequences from different pieces further away [21].

Defining Cocktail Representations. I first define taste representations for each ingredient: *sour, sweet, alcohol, bitter, fruit, herb, complex, spicy, oaky, fizzy, colorful, eggy*, then obtain cocktail representations as the proportion-weighted average of their ingredient representations. This accounts for the dilution expected from the cocktail preparation type and the alcoholic content [22].

Learning the Piece–Cocktail Synesthetic Mapping. I trained a *translation* variational auto-encoder (VAE) from [6]. This VAE takes in either cocktail or piece representations, maps them to an inter-modal latent space and decodes either cocktail or piece representations. Using standard VAE losses, the architecture is trained to represent the latent distributions of both cocktail and piece representations separately (using either cocktail input and outputs or piano pieces inputs and outputs respectively). Then, I use a sliced-Wasserstein distance loss (SWD) to maximize the overlap between the latent distributions in the two modalities. This allows *translation*, as a representation encoded from one modality can now be decoded into the other. Finally, I use *alignment classification losses* to nudge the semantic alignment to respect some grounding categories that I define. For instance, cuban latin jazz and cuban cocktails (containing cuban rum, mint and lime) must be semantically aligned to the *cuban* label, i.e. their respective latent embeddings must be classified as *cuban* by a one-layer classification head placed on top of the inter-modal latent layer. At test time, the translation VAE encodes a piece representation to parameters of the inter-modal latent distribution, samples an inter-modal latent code and decodes it into a cocktail representation.

Generating Cocktails. A genetic algorithm (GA) optimization process is run to obtain the final recipe from the generated cocktail representation. The GA initializes a population of random cocktails characterized by their recipe (volume between 0 and *max volume* for each ingredient). The score of each candidate recipe is computed as the negative distance between its cocktail representation and the music-induced target cocktail representation. The parents of the next generation are then selected proportionally to the ranks of the children's scores and mutated to obtain the new generation. Further constrains can disqualify candidate recipes: too much or too little alcohol, egg white without citrus, more than one fizzy ingredients, more than 2 liqueurs, more than 3 liquors. After 50 generations, the process converges on the final cocktail recipe, the one that matches best the target cocktail representation. A separate network predicts the preparation type and the service glass such that the system can output full instructions for the cocktail preparation.

Evaluation. Looking at t-SNE representations and nearest neighbor tests, it seems that the piece and cocktail representation spaces capture well *perceptual distances* such that pieces/cocktails that sound/taste closer *perceptually* (pieces from the same genre, the same composer; cocktails from the same type, using similar ingredients) are indeed found to be close in their respective representation spaces using the Euclidean metric. The translation VAE generates similar cocktails for similar piano pieces. It respects the alignment constraints I set and demonstrates specific biases in the cocktail generation for clusters of similar piano pieces beyond these constraints. In short, the *TastyPiano* system demonstrates *perceptual locality* and *subjectivity* while respecting the alignment points I wanted.

Figure 2: *Take the A Train (Duke Ellington)* Shaken.

```
Ingredients:
   15.0 mL - cointreau
   15.0 mL - honey syrup
   20.0 mL - lemon juice
   45.0 mL - rye whiskey
   30.0 mL - soda
Serve in:
   Collins glass.
```

Conclusion. Non-synesthetes can now taste music thanks to *TastyPiano*! I hope to build the physical machine in the near future. Figure 2 shows what Duke Ellington's *Take the A Train* could taste like. Try it for yourself using the online demo: https://huggingface.co/spaces/ccolas/TastyPiano!

References

[1] Richard E Cytowic. *Synesthesia: A union of the senses*. MIT press, 2002.

[2] Jamie Ward and Jason B Mattingley. Synaesthesia: an overview of contemporary findings and controversies. *Cortex*, 2006.

[3] Richard E Cytowic and David M Eagleman. *Wednesday is indigo blue: Discovering the brain of synesthesia*. Mit Press, 2011.

[4] June E Downey. Literary synesthesia. *The Journal of Philosophy, Psychology and Scientific Methods*, 1912.

[5] Patricia Lynne Duffy. Synesthesia in literature. *Simner and Hubbard*, 2013.

[6] Yingtao Tian and Jesse Engel. Latent Translation: Crossing Modalities by Bridging Generative Models. *ArXiv - id/1902.08261*, 2019.

[7] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. *ArXiv - id/1810.12247*, 2019.

[8] Qiuqiang Kong, Bochen Li, Jitong Chen, and Yuxuan Wang. Giantmidi-piano: A large-scale midi dataset for classical piano music. *ArXiv - id/2010.07061*, 2020.

[9] Zhengshan Shi, Craig Stuart Sapp, Kumaran Arul, Jerry McBride, and Julius O. Smith. Supra: Digitizing the stanford university piano roll archive. In *Proceedings of the 20th International Society for Music Information Retrieval*, 2019.

[10] Meinard Müller, Verena Konz, Wolfgang Bogler, and Vlora Arifi-Müller. Saarland music data (SMD). 2011.

[11] Bernd Krueger. http://www.piano-midi.de.

[12] Matthias Dorfer, jr. Hajič, Jan, Andreas Arzt, Harald Frostel, and Gerhard Widmer. Msmd - multimodal sheet music dataset, 2019.

[13] Valentin Emiya, Roland Badeau, and Bertrand David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1643–1654, 2009.

[14] Qiuqiang Kong, Bochen Li, Xuchen Song, Yuan Wan, and Yuxuan Wang. High-resolution piano transcription with pedals by regressing onset and offset times. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[15] Gaëtan Hadjeres and Léopold Crestel. The piano inpainting application. *ArXiv - id/2107.05944*, 2021.

[16] Nathan Fradet, Jean-Pierre Briot, Fabien Chhel, Amal El Fallah-Seghrouchni, and Nicolas Gutowski. Miditok: A python package for midi file tokenization. In *22nd International Society for Music Information Retrieval Conference*, 2021.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proc. of ACL*, 2019.

[18] Yi-Hui Chou, I.-Chun Chen, Chin-Jui Chang, Joann Ching, and Yi-Hsuan Yang. MidiBERT-Piano: Large-scale Pre-training for Symbolic Music Understanding. *ArXiv - id/2107.05223*, 2021.

[19] Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training. *Proc. of ACL*, 2021.

[20] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *Proc. of ACL*, 2019.

[21] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *Proc. of ICLR*, 2020.

[22] Dave Arnold. *Liquid Intelligence: The Art and Science of the Perfect Cocktail*. WW Norton & Company, 2014.