INFO8010: Project Proposal

Corentin Jemine[1] and **Mathias Beguin**[2]

[1] *cjemine@student.uliege.be (s123578)*
[2] *Mathias.Beguin@student.uliege.be (s140309)*

I. TASK

We wish to experiment with source separation in polyphonic music. Given an audio waveform of a musical piece and the identity of an instrument present in the piece, our model is expected to generate an audio waveform of the performance of that instrument alone as if it had been recorded on its own.

One variation of this task, called "blind source separation", is to separate the instruments in the piece without having any information on the identity of the instruments or their number. We will not work with this blind context, and instead will either provide instrument identities from ground truth labels during training or through manual extraction at inference time.

The instrument identities can be represented in several ways. A fixed-size table of the instruments can be learned and instruments can then be referenced by their index in that table. A standalone numerical representation can be used instead, such as a set of expertly tuned parameters or a machine learned embedding. The generalization ability of the model on the task of source separation depends greatly on this choice of representation. We have not yet decided on which to use.

II. IMPLEMENTATION

We wish to train the model using a dataset of both separate and joint audio tracks. Generation of such data is feasible without much effort: one can automatically generate audio tracks for different instruments and mix them to create songs. Arbitrary music generation can be achieved this way, or more advanced approaches can be used as to achieve musically pleasant songs. Alternatively, one can use existing datasets containing multitrack audio files (e.g. [1]). We also suggest the use of MIDI files, for which the audio tracks are already separated and a waveform can be easily generated.

We want to work on the raw waveform rather than on a spectrogram, as is becoming increasingly common in sound-related deep learning applications. The model should be conditioned on an instrument identity and either convolve over the audio sequence or process it in a recurrent fashion while generating the output audio sequence. The model thus acts a transformer of the input sequence, and the input and output domains are identical.

III. RELATED WORKS

We have found several related papers in the literature.

[2], [3], [4] and [5] propose approaches specific to separating the singing voice from the instruments.

[6] investigates melody extraction.

[7] and [8] demonstrate source separation paired with a corresponding video. [7] disentangles speakers talking over each other while [8] disentangles musical instruments playing together.

[9] proposes an approach to identify which instruments are present in a musical piece.

[10], [11], [12], [13], [14], [15], [16] and [17] work on the same task as ours and [18] additionally works in a blind setting as described before. Some of these methods operate on the raw waveform (e.g. [11], [12]) and some on a spectrogram representation (e.g. [13], [17]). The approach described in [11] uses embeddings for instrument identities.

[1] Rachel M. Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. Medleydb: A multitrack dataset for annotation-intensive mir research. In *ISMIR*, 2014.

[2] Emad M. Grais, Dominic Ward, and Mark D. Plumbley. Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders, 2018.

[3] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, Derry FitzGerald, and Bryan Pardo. An overview of lead and accompaniment separation in music, 2018.

[4] Li Su. Vocal melody extraction using patch-based cnn. 2018.

[5] Andrew J. R. Simpson, Gerard Roma, and Mark D. Plumbley. Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network, 2015.

[6] Tsung-Han Hsieh, Li Su, and Yi-Hsuan Yang. A streamlined encoder/decoder architecture for melody extraction, 2018.

[7] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels, 2018.

[8] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement, 2018.

[9] Peter Li, Jiyuan Qian, and Tian Wang. Automatic instrument recognition in polyphonic music using convolutional neural networks, 2015.

[10] Abhimanyu Sahai, Romann Weber, and Brian McWilliams. Spectrogram feature losses for music source separation, 2019.

[11] Prem Seetharaman, Gordon Wichern, Shrikant Venkataramani, and Jonathan Le Roux. Class-conditional embeddings for music source separation, 2018.

[12] Francesc Lluís, Jordi Pons, and Xavier Serra. End-to-end music source separation: is it possible in the waveform domain?, 2018.

[13] Joachim Muth, Stefan Uhlich, Nathanael Perraudin, Thomas Kemp, Fabien Cardinaux, and Yuki Mitsufuji. Improving dnn-based music source separation using phase features, 2018.

[14] Jen-Yu Liu and Yi-Hsuan Yang. Denoising auto-encoder with recurrent skip connections and residual regression for music source separation, 2018.

[15] Shinichi Mogami, Hayato Sumino, Daichi Kitamura, Norihiro Takamune, Shinnosuke Takamichi, Hiroshi Saruwatari, and Nobutaka Ono. Independent deeply learned matrix analysis for multichannel audio source separation, 2018.

[16] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. 2018.

[17] Sungheon Park, Taehoon Kim, Kyogu Lee, and Nojun Kwak. Music source separation using stacked hourglass networks, 2018.

[18] Sören Schulze and Emily J. King. Musical instrument separation on shift-invariant spectrograms via stochastic dictionary learning, 2018.