Automatic Music Transcription and Instrument Transposition with Differentiable Rendering

Hayato Sumino[1], Adrien Bitton[2], Lisa Kawai[1], Philippe Esling[2], and Tatsuya Harada[1,3] ⋆

[1] The University of Tokyo
[2] IRCAM
[3] RIKEN
{sumino, kawai, harada}@mi.t.u-tokyo.ac.jp
{bitton, esling}@ircam.fr

Abstract. Automatic music transcription aims to extract a musical score from a given audio signal. Conventional machine learning frameworks usually address this task by relying solely on error back-propagation from annotated MIDI data, without consideration for acoustic similarities. In this study, we complement the onset and frames prediction objective with an acoustic distance, through *differentiable rendering* of the estimated piano-roll and approximate reconstruction of the analyzed signal. We apply our method to piano and show that this added reconstruction error improves the performance achieved with the usual supervised transcription loss. Moreover, using solely this acoustic criterion allows fully unsupervised training and results outperforming classical techniques. Finally, our method also enables performing *automatic instrument transposition* by using audio samples of a different instrument from the original sound source when reconstructing the input signal.

1 Introduction

Automatic Music Transcription (AMT) is the task of estimating a musical score from a given acoustic signal. Various methods have been studied for this problem (Benetos, Dixon, Duan, & Ewert, 2019), but recent deep learning techniques with Convolutional Neural Network (CNN) have reached unprecedented accuracy on this task (Kelz et al., 2016; Hawthorne et al., 2018, 2019). In these approaches, large datasets of acoustic signals and their corresponding score annotations need to be collected beforehand to train the models. Although some

⋆ This work was partially supported by JST AIP Acceleration Research Grant Number JPMJCR20U3, and partially supported by JSPS KAKENHI Grant Number JP19H01115. This work was also supported by the ANR:17-CE38-0015-01 MAKIMOno project, the SSHRC:895-2018-1023 ACTOR Partnership and Emergence(s) ACIDITEAM project from Ville de Paris and ACIMO project of Sorbonne Université.

annotated datasets exist for piano scores (Emiya, Badeau, & David, 2010), the amount of available data remains insufficient for the other instruments, and the cost for producing and annotating such large datasets is prohibitive. Oppositely, previous AMT methods such as Non-negative Matrix Factorization (NMF) (Lee & Seung, 1999) do not require such large datasets for optimization. Applied to AMT, the time-frequency spectrogram of an acoustic signal can be considered as a non-negative matrix, which is decomposed into the pitch bases and activation amplitudes. Hence, instrument sounds for every pitch can be used as a prior distribution for the basis matrix (Smaragdis & Brown., 2003; Vincent, Bertin, & Badeau, 2010; O'Hanlon & Plumbley., 2014). However, since this method is an unsupervised linear decomposition, it achieves lower accuracies than deep learning methods, which harness large amounts of annotated data to optimize the relationships between audio and scores.

Machine learning-based AMT usually optimizes the Binary Cross Entropy (BCE) between the estimated and the ground-truth piano-rolls (Hawthorne et al., 2018), a criterion which does not reflect acoustic perception. For instance, wrongfully adding a C$^\sharp$3 to a C3 note is a more crucial error than adding a C4, as this octave error is included in the overtone components of musical instruments. Hence, octave errors produce a less significant artifact than the strong auditory dissonance introduced by the concomitant note. Similarly, a prediction error is less perceptible if the note density is high than when the score is sparse. Nevertheless, the BCE criterion evaluates all of these situations equivalently without accounting for these differences. Recent works on unsupervised transcription (Choi & Cho, 2019) have proposed an alternative training objective based on rendering the estimated transcription and computing a reconstruction error with the analyzed input signal. In this study, we address the limitations of the supervised transcription loss with this rendering objective such that the model learns AMT with additional acoustic cues. To do so, we perform a *differentiable rendering* of the estimated piano-roll which enables end-to-end error backpropagation from the output signal. This loss function is defined as the L1 distance between Mel-spectrograms of the input and the reconstructed signal.

In the semi-supervised setting, large amounts of unlabelled signals can be used as supplementary training data to obtain the better performance, and if this acoustic distance is used in isolation, it is also possible to do purely unsupervised learning. Furthermore, since any waveform can be used when rendering the transcription, we introduce *automatic instrument transposition*. In this task, a signal from any source instrument(s) is transcribed into a score that is rendered with audio samples from another target instrument. As a result, the unsupervised model learns to estimate a score and an audio reconstruction sounding as similar as possible to the input source but played with another instrument. We apply this method to automatically transposing orchestral music to piano, result examples can be found in our project page (`https://acids-ircam.github.io/PianoTranscriptionTransposition/`).

2 Related Work

Automatic Music Transcription The current approaches for AMT can be roughly classified into two types, namely, methods based on spectrogram decomposition and deep learning.

In the decomposition methods, the input spectrogram is usually factorized into a *basis matrix* (a set of spectrum templates corresponding to each pitch) and an *activation matrix* (the intensity of each pitch component at given time) by using Non-negative Matrix Factorization (NMF) (Smaragdis & Brown., 2003) algorithm, that decomposes a non-negative matrix into the product of two low-rank non-negative matrices. In conventional NMF for AMT, since the model learns the timbre patterns of a single time frame as an individual basis, this can prove limited to musical instruments that have a clear temporal evolution, such as piano. Non-negative Matrix Factorization Decomposition (NMFD) (Smaragdis, 2004) is an extended version of NMF that can learn time-varying basis matrices with time width, at the expense of a higher complexity. This idea was expanded in (Berg-Kirkpatrick, Andreas, & Klein, 2014) by probabilistic modeling of piano music spectrograms through multiple spectral components and time envelopes. Parameters of each piano key are jointly optimized with the estimate of note events (duration and velocity), resulting in individual spectrogram components that are superimposed in order to render the complete spectrogram. The reconstruction error of the target spectrogram allows unsupervised optimization of each piano key parameters and events, yielding the final transcription estimate.

On the other hand, supervised methods based on deep learning with CNN have recently achieved excellent results. A comparative study (Kelz et al., 2016) across large sets of models and hyperparameters for frame-based music transcription showed that CNN was the most effective. For instance, the recently proposed *Onsets and Frames* (Hawthorne et al., 2018) approach has produced state-of-the-art performance. In this work, a logarithmic Mel-spectrogram input is converted into features by two independent CNN and Bi-LSTM layers. Each network predicts either onset times or pitch values, while the estimated onset is used as a supplementary input to predict the pitch. By explicitly predicting onsets in addition to the conventional transcription, the recognition performance for the piano was greatly improved, which is due to the characteristics of piano sounds that have large differences between the onset and sustained parts. Supervised learning of AMT requires large amounts of human-annotated audio performances, and the usual symbolic prediction loss does not account for the resulting acoustic errors. An alternative unsupervised learning framework was proposed (Choi & Cho, 2019) for drum transcription based on acoustic reconstruction. In this work, a trainable transcriber is paired with a fixed drum synthesizer so that a signal is generated from the estimated score, using a prerecorded sample library, and compared to the input signal. This differentiable rendering allows the backpropagation of the reconstruction error in the acoustic domain to the score estimate and alleviates the need for annotations to train the transcriber network. Our model relies on the optimization of the reconstructed sounds which are rendered through *Onsets and Frames* approach.

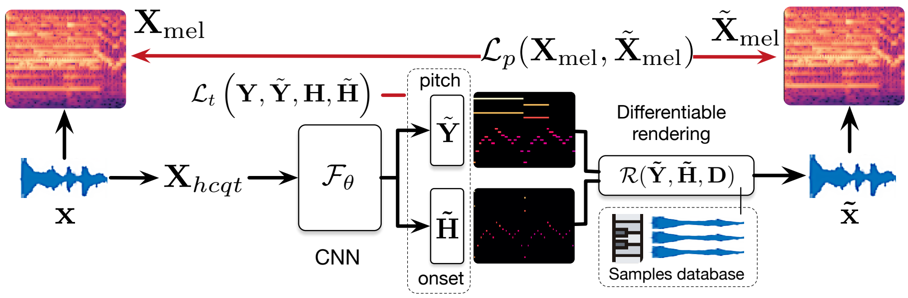

Fig. 1. Our proposed method for supervised transcription with differentiable rendering. A model $\mathcal{F}_\theta$ produces approximated pitch $\tilde{\boldsymbol{Y}}$ and onset $\tilde{\boldsymbol{H}}$ matrices. These are used along with reference samples $\boldsymbol{D}$ to produce a signal $\tilde{\mathbf{x}}$. This allows us to evaluate the loss $\mathcal{L}_p(\boldsymbol{X}_{\text{mel}}, \tilde{\boldsymbol{X}}_{\text{mel}})$ in the spectral domain, and also enables unsupervised training.

Automatic Instrument Transposition There are different approaches to music conversion, depending on whether the task is addressed in the audio signal or symbolic score domains. In this work, we focus on audio-based methods. In (Grinstein, Duong, Ozerov, & Pérez., 2018), an image style conversion algorithm (Johnson, Alahi, & Fei., 2016) is applied to spectrograms. The major issue with processing spectrograms as images is that the phase information is lost. Hence, the phase must be restored using the Griffin-Lim algorithm, but the direct relationship between the original and converted signals cannot be learned. This incurs limitations to the transposition quality as well as an undesired latency. In (Mital., 2018), conversion between waveforms is performed using pre-trained WaveNet encoder-decoder models (van den Oord et al., 2016). This idea is expanded in the Universal Music Translation Network (UMTN) (Mor, Wolf, Polyak, & Taigman., 2019), which addresses multiple conversion tasks using a specific decoder for each target domain. Although high-quality transposition is achieved by this model, its auto-regressive nature and the need to train a decoder for each domain incur large learning costs and inference times.

3 Proposed Method

In this study, we combine the advantages of NMF, through evaluation on acoustic features, with previous deep learning models for supervised AMT. A trainable transcription network analyzes an audio input to predict an estimated piano-roll, which is compared with the annotated ground-truth in the symbolic transcription loss function $\mathcal{L}_t$. Using pre-recorded audio samples, a differentiable rendering system generates a signal from the estimated score (Choi & Cho, 2019). The output signal can be compared with the input signal, yielding an additional acoustic reconstruction objective $\mathcal{L}_p$ that does not require any annotations.

System Overview Our system takes an input audio signal $\mathbf{x} \in \mathbb{R}^N$ of length N and computes Harmonic-CQT $\boldsymbol{X}_{\text{hcqt}} \in \mathbb{R}^{C \times F \times T}$ (Bittner, McFee, Salamon, Li, & Bello., 2017) input features, with F the number of bins for C octave shifts and T time frames. The AMT task aims at uncovering the corresponding piano-roll $\boldsymbol{Y} \in \mathbb{R}^{P \times T}_{[0,1]}$, where P is the number of pitches. We first define a learning model $\mathcal{F}_\theta$ that receives the features of the acoustic signal and outputs

$$\mathcal{F}_\theta(\boldsymbol{X}_{\text{hcqt}}) = [\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{H}}] \in \mathbb{R}^{2 \times P \times T}_{[0,1]}. \tag{1}$$

Similar to (Hawthorne et al., 2018), the model outputs an approximation to the transcription $\tilde{\mathbf{Y}}$, along with approximate onsets denoted as $\tilde{\boldsymbol{H}} \in \mathbb{R}^{P \times T}_{[0,1]}$. To obtain binary onset positions, we further use a threshold value γ for the onsets, where $h_{p,t} > \gamma$ at time t, denotes an onset for pitch p. In our experiments, we use $\gamma = 0.05$. Based on the approximated piano-roll $\tilde{\boldsymbol{Y}}$ and onset matrix $\tilde{\boldsymbol{H}}$, we aim to render the transcription into a signal $\tilde{\mathbf{x}}$ using a dataset of reference samples $\boldsymbol{D} \in \mathbb{R}^{P \times N}$ for each pitch. Hence we define a rendering function

$$\tilde{\mathbf{x}} = \mathcal{R}(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{H}}, \boldsymbol{D}) \tag{2}$$

where the function $\mathcal{R}$ should be differentiable with respect to $\tilde{\boldsymbol{Y}}$ and $\tilde{\boldsymbol{H}}$. Finally, the Mel-spectrogram $\tilde{\boldsymbol{X}}_{\text{mel}} \in \mathbb{R}^{F \times T}$ is computed from the waveform reconstruction $\tilde{\mathbf{x}}$ in order to assess the error $\mathcal{L}_p$ to the Mel-spectrogram $\boldsymbol{X}_{\text{mel}}$ computed from the input $\mathbf{x}$ (Fig. 1). Onsets are explicitly used in the rendering function $\mathcal{R}$ because some instruments such as piano include strong aperiodic components and tone variations along time. In the case of more stationary instrument sounds, it could be possible to simplify this rendering method by solely relying on the amplitude information without onset time.

Differentiable Rendering In order to perform backpropagation of the audio reconstruction error, the synthesis function from the piano-roll should be differentiable. We introduce a rendering technique loosely based on NMFD, along with a suited optimization criterion to complement the usual symbolic losses. Hence, based on a transcription $\boldsymbol{Y}$, onsets $\boldsymbol{H}$ and sample dataset $\boldsymbol{D}$

$$\mathcal{R}(\boldsymbol{Y}, \boldsymbol{H}, \boldsymbol{D})_i = \sum_s \sum_p (\vec{\boldsymbol{H}}^s \otimes \boldsymbol{Y})_{pj}\, d_{p,k+s\frac{T}{N}}, \tag{3}$$
$$j = \lfloor i \times \frac{\mathrm{T}}{\mathrm{N}} \rfloor,\ k = i - s \times j$$

where $\boldsymbol{d}_p = (d_{p,1}, d_{p,2}, \cdots d_{p,i}, \cdots d_{p,N})$ is the signal basis of pitch p, N is the waveform sample length and $\vec{\boldsymbol{H}}^s$ is a matrix obtained by shifting matrix $\boldsymbol{H}$ to the right by s frames and padding with zeros, such that

$$\vec{\boldsymbol{H}}^0 = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix},\ \vec{\boldsymbol{H}}^1 = \begin{pmatrix} 0 & 1 & 2 \\ 0 & 4 & 5 \end{pmatrix},\ \vec{\boldsymbol{H}}^2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 4 \end{pmatrix}. \tag{4}$$

Rendering with Multiple Sample Sources One issue with sample-based rendering is that the timbre of reference samples is different from the input. Hence, the value of $\mathcal{L}_p$ might not be able to become 0 while $\mathcal{L}_t = 0$. This problem could be alleviated if audio samples from the original sound source were available. Usually, this requirement cannot be fulfilled and as a workaround, we can use multiple recordings of a same instrument. This allows us to reduce the impact of small differences in timbre between the input and the reference instrument.

Loss Functions In the following, we consider both the traditional piano-roll transcription loss $\mathcal{L}_t$ and our proposed perceptual loss $\mathcal{L}_p$ on acoustic features.

$$\mathcal{L}_t(\boldsymbol{Y}, \tilde{\boldsymbol{Y}}, \boldsymbol{H}, \tilde{\boldsymbol{H}}) = \sum_{p,t} \mathrm{CE}(y_{p,t}, \tilde{y}_{p,t}) + \alpha \mathrm{CE}(h_{p,t}, \tilde{h}_{p,t}) \tag{5}$$

$$\mathcal{L}_p(\boldsymbol{X}_{\mathrm{mel}}, \tilde{\boldsymbol{X}}_{\mathrm{mel}}) = \sum_{f,t} \Big| \log\big(|x_{f,t}| + \epsilon\big) - \log\big(|\tilde{x}_{f,t}| + \epsilon\big) \Big| \tag{6}$$

where CE represents the cross-entropy function, and α balances the impact of the onsets in the loss. In the experiments, we evaluate the importance of the onsets by setting $\alpha = \{0, 1\}$. Finally, the complete training loss is defined as the weighted sum of these two functions with $\lambda_t = \{0, 1\}$ and $\lambda_p = 0.1$

$$\mathcal{L} = \lambda_t \mathcal{L}_t + \lambda_p \mathcal{L}_p. \tag{7}$$

4 Experiments on Automatic Music Transcription

Experimental Settings We evaluate our method for AMT on the MAPS Dataset (Emiya et al., 2010), which is composed of performance recordings and the corresponding ground-truth MIDI data. It contains 270 classic songs amounting to about 1,300 minutes, from which we randomly select 240 songs for the training and 30 songs for the test set. For the rendering, we use audio samples from the NSynth Dataset (Engel et al., 2017) (*keyboard ac. 001*) as the reference samples for piano.

Based on the input audio signal $\mathbf{x}$, we compute the Harmonic-CQT $\boldsymbol{X}_{\mathrm{hcqt}} \in \mathbb{R}^{C \times F \times T}$ as input features (Bittner et al., 2017). The Harmonic-CQT is obtained by computing multiple CQT-spectrograms at octave shifts of a given reference frequency and then concatenating them in the Z-axis direction. In our study, we use $F = 252$, CQT bins for frequencies ranging from C1 to B7, computed at $C = 3$ different octaves, window length 2048, and hop length 512.

We pre-process the training data by lengthening the notes according to the pedal information which our method cannot detect. We also rescale the estimated amplitude values in matrix $\mathbf{Y}$, by computing $Y_{i,j} \leftarrow a|Y_{i,j}|^2$ prior to audio rendering following (Dannenberg, 2006), where $a = 0.16$ in our experiments.

We rely on the same network architecture and hyperparameters as the best-performing CNN model detailed in (Kelz et al., 2016) for transcription.

supervised	λ_p	α	f-meas.	P (%)	R (%)
baseline	0	0	71.6	79.5	65.2
baseline	0	1	76.3	83.1	70.7
ours	0.1	0	75.3	77.3	71.9
ours	0.1	1	**78.2**	**84.5**	**74.0**
unsupervised					
NMF	-	-	37.7	32.8	45.9
ours	0.1	-	**54.3**	**57.8**	**52.6**

Table 1. F-measure, precision, and recall comparison on the AMT task.

method	num. of sounds	f-meas.	P (%)	R (%)
$\mathcal{L}_p$	1	54.3	57.8	52.6
$\mathcal{L}_p$	15	**56.9**	62.5	54.0
$\mathcal{L}_t + \mathcal{L}_p$	1	78.2	84.5	74.0
$\mathcal{L}_t + \mathcal{L}_p$	15	**80.4**	84.5	76.3

Table 2. Evaluation of multiple sound sources rendering.

Results We display in Table 1 the evaluation of different methods for the task of supervised (top) and unsupervised (bottom) transcription, tested with or without onset loss (α in Eq. (5)). The supervised baseline method ($\lambda_p = 0$) with onset loss corresponds to (Hawthorne et al., 2018) and the best results are obtained with our method, combining $\mathcal{L}_t$ with onsets and $\mathcal{L}_p$. Regarding unsupervised transcription, it can be seen that even without using $\mathcal{L}_t$, our method vastly outperforms the NMF approaches, as we can use the flexibility of neural networks to produce better estimates of the transcription.

We also include 15 acoustic piano sources from the NSynth Dataset (Engel et al., 2017) as reference samples for rendering and display the results in Table 2, which we report by the average value of the performance for each sound source. As we can see, the accuracy of our method is even higher when multiple sound sources are used for rendering. Hence, we can still improve the accuracy of our method with more sound sources, but this comes at a higher processing cost.

5 Automatic Instrument Transposition

The AMT experiments aimed at transcribing a given instrument, but if another instrument than the original one in the input sound is used as the rendering sound source, the framework of our proposed method can be applied to *automatic instrument transposition*. In this task, we aim to reproduce the input audio with a different instrument, while keeping the output as acoustically similar as possible (Fig. 2). In this case, the main output is a rendered waveform, but our method allows us to obtain a transposed score as a byproduct (without ground truth). The merit of using waveforms for both input and output is that it can take into account the information of musical expressions such as delicate tone changes, volume changes, and timing fluctuations.

Experimental Settings Here, we rely on MusicNet (Thickstun, Harchaoui, & Kakade., 2017), which is a dataset containing 330 piano songs and classical chamber music. In this experiment, 34 Beethoven string quartets were rendered using NSynth (Engel et al., 2017) piano sources. As a reference baseline, we used

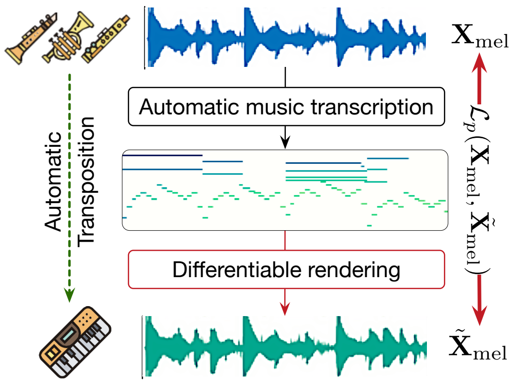

Fig. 2. Overview of our method for automatic instrument transposition. It produces a score for different instruments, while maximizing the acoustic similarity between input and transposed signals.

method	UMTN	ours
L1 loss	0.486	**0.362**
L2 loss	7.730	**4.432**
time (s)	178.8	**2.4**

Table 3. Performance comparison between UMTN and our method on different losses (top) and in computation time required to generate a 60s. waveform (bottom).

the publicly available trained model of Universal Music Translation Network (UMTN) (Mor et al., 2019), trained on the MusicNet Dataset.

Results Table 3 (top) shows the quantitative evaluation results between UMTN and our method for 10 pieces of test data. Both the L1 and L2 spectrogram distances have smaller errors in our method, which indicates that it can generate automatically transposed outputs that are acoustically closer to the given audio input. We as well report in Table 3 (bottom) the comparison of computation time required for inference when generating a 60-second waveform on a NVIDIA Titan V GPU. While UMTN relies on WaveNet encoder-decoder model that incurs a slow autoregressive sample prediction, our method only uses a light-weight CNN that can be rendered in real time.

6 Conclusion

In this study, we applied a new approach to AMT for piano by using *differentiable rendering* along with a loss function that ensures that the rendered transcription remains close to the original sound signal. To achieve this, we proposed a data structure based on pitch and onset, allowing us to define a NMFD-like differentiable rendering system. Our experiments show that the evaluation function on acoustic features is effective for strongly improving the performance of automatic transcription. In addition, we show that this method can be applied to automatic instrument transposition, where an instrument track is reproduced by another instrument while remaining acoustically close to the given performance. Future prospects include the subjective evaluation of the automatic transposition method, and a more flexible definition of the differentiable rendering system.

References

Benetos, E., Dixon, S., Duan, Z., & Ewert, S. (2019). Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, *36*(1), 20-30.

Berg-Kirkpatrick, T., Andreas, J., & Klein, D. (2014). Unsupervised transcription of piano music. In *Advances in neural information processing systems 27* (pp. 1538–1546).

Bittner, R., McFee, B., Salamon, J., Li, P., & Bello., J. (2017). Deep salience representations for f0 estimation in polyphonic music. In *Proceedings of the 18th international society for music information retrieval conference, ISMIR.*

Choi, K., & Cho, K. (2019). Deep unsupervised drum transcription. In *Proceedings of the 20th international society for music information retrieval conference, ISMIR.*

Dannenberg, R. B. (2006). The interpretation of midi velocity. In *Proceedings of the international computer music conference, ICMC.*

Emiya, V., Badeau, R., & David, B. (2010). Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech and Language Processing*, *18*(6), 1643–1654.

Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., & Simonyan, K. (2017). Neural audio synthesis of musical notes with wavenet autoencoders. In *Proceedings of machine learning research* (Vol. 70, pp. 1068–1077).

Grinstein, E., Duong, N., Ozerov, A., & Pérez., P. (2018). Audio style transfer. In *IEEE international conference on acoustics, speech and signal processing, ICASSP.*

Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., ... Eck, D. (2018). Onsets and frames: Dual-objective piano transcription. In *Proceedings of the 19th international society for music information retrieval conference, ISMIR.*

Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C. A., Dieleman, S., ... Eck., D. (2019). Enabling factorized piano music modeling and generation with the maestro dataset. In *International conference on learning representations, ICLR.*

Johnson, J., Alahi, A., & Fei., L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the european conference on computer vision, ECCV.*

Kelz, R., Dorfer, M., Korzeniowski, F., Bock, S., Arzt, A., & Widmer., G. (2016). On the potential of simple framewise approaches to piano transcription. In *Proceedings of the 17th international society for music information retrieval conference, ISMIR.*

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature*, *401*, 788–791.

Mital., P. K. (2018). Time domain neural audio style transfer. In *Advances in neural information processing systems 31.*

Mor, N., Wolf, L., Polyak, A., & Taigman., Y. (2019). A universal music translation network. In *International conference on learning representations, ICLR.*

O'Hanlon, K., & Plumbley., M. D. (2014). Polyphonic piano transcription using non-negative matrix factorisation with group sparsity. In *IEEE international conference on acoustics, speech and signal processing, ICASSP.*

Smaragdis, P. (2004). Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *Independent component analysis and blind signal separation, 5th international conference, ICA* (Vol. 3195, pp. 494–499). Springer.

Smaragdis, P., & Brown., J. C. (2003). Non-negative matrix factorization for polyphonic music transcription. In *IEEE workshop on applications of signal processing to audio and acoustics, WASPAA.*

Thickstun, J., Harchaoui, Z., & Kakade., S. M. (2017). Learning features of music from scratch. In *International conference on learning representations, ICLR.*

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. In *The 9th ISCA speech synthesis workshop* (p. 125).

Vincent, E., Bertin, N., & Badeau, R. (2010). Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech and Language Processing*, *18*(3), 528–537.