POLYPHONIC MUSIC TRANSCRIPTION WITH SEMANTIC SEGMENTATION

[1]*Yu-Te Wu,* [1]*Berlin Chen, and* [2]*Li Su*

[1]Department of Computer Science, National Taiwan Normal University, Taipei, Taiwan
[2]Institute of Information Science, Academia Sinica, Taipei, Taiwan

ABSTRACT

The multi-instrument transcription task refers to joint recognition of instrument and pitch of every event in polyphonic music signals generated by one or more classes of music instruments. In this paper, we leverage multi-object semantic segmentation techniques to solve this problem. We design a time-frequency representation, which has multiple channels to jointly represent the harmonic structure and pitch saliency of a pitch activation. The transcription task therefore becomes a pixel-wise multi-task classification problem including pitch activity detection and instrument recognition. Experiments on both single- and multi-instrument data verify the competitiveness of the proposed method.

Index Terms— Automatic music transcription, multi-pitch estimation, semantic segmentation.

1. INTRODUCTION

Deep learning techniques have brought great success in image semantic segmentation in recent years, by providing full-fledged solutions in segmenting multiple classes of objects, where each class of object may have multiple instances on a single image [1]. Such an achievement is accumulated by a number of inventions, such as the fully convolutional networks (FCN) [2], the U-net structure [3], pooling of dilated convolution [4], novel loss function [5], to name but a few. It would be intriguing to see if these novel methods can leverage the analysis of music content. In particular, we consider the semantic segmentation on the time-frequency image of a polyphonic signal generated by multiple musical instruments. Here, an instrument is equivalent to a class of object, and each class typically has multiple instances (i.e., pitch activation) on the time-frequency plane. This problem is part of the automatic music transcription (AMT) problem, and its goal is to jointly transcribe pitch and instrument from audio. We refer to this task as multi-instrument transcription in this paper.

Instrument-level information has been widely discussed in AMT. For example, previous studies including non-negative matrix factorization (NMF) [6] and probabilistic latent component analysis (PLCA) [7, 8] adopt instrument-wise templates to better fit the spectral patterns lying in the signals. This implicitly allows multi-instrument transcription. However, most studies took such instrument labels merely for improving the performance of multi-pitch estimation (MPE), and the benchmarks made simultaneously on both instrument recognition and MPE are few. Multi-instrument transcription still remains an rarely investigated problem.

Recently, deep learning-based semantic segmentation models are catching increasing attention in various music signal processing problems, such as singing voice separation [9], MPE [10], and melody extraction [11]. Furthermore, the high flexibility of deep learning models also allows us to build a multi-task learning (MTL) model with less effort. For example, [12] proposed a piano transcription model to transcribe note onset, offset, pitch, and velocity at the same time. [13] also considered an MTL notwork for MPE, melody, vocal, and bass transcription. Novel loss functions such as the focal loss is also applied for melody extraction to emphasize pitch objects that are typically thin and are hard to capture using conventional loss function [11].

To learn multiple attributes from music, the input data representation is also critical for semantic segmentation models. [10] proposed the harmonic constant-Q transform (HCQT), which is a multi-channel feature allowing the harmonic sequence of a pitch to be 'visible' on one single pixel at the same time. This is analogue to the RGB channels in colored images. Similarly, [11] paralleled the spectral and temporal characteristics of a pitch object jointly, based on the combined frequency and periodicity (CFP) approach. Experiments have also shown that the performance can benefit from such multi-view data representations. In this work, we consider combining the HCQT and CFP methods with semantic segmentation modeling for multi-instrument transcription. Experiments show that, first, the proposed segmentation model achieves state-of-the-art performance in the MPE problem of piano music, and second, the proposed data representation achieves better performance than the baseline method in multi-instrument transcription.

2. METHOD

2.1. Data representation

The data representation adopted in this work is derived from two previous studies: the CFP representation in [14] and the

Harmonic Constant-Q Transform (HCQT) proposed in [10]. According to [14], leveraging multiple features in both the time and frequency domains can lead to better performance in AMT. In addition, to better capture harmonic information, we also incorporate the idea of stacking harmonic information along the channel axis as in [10].

Let $\mathbf{X} \in \mathbb{R}^{2F \times T}$ be the magnitude part of the short-time Fourier transform (STFT), where F is the dimension of the spectrum in the positive frequency range, and T is the dimension of time. Consider the following two data representations:

$$\mathbf{Z}_f[k,n] := \mathbf{Q}_f \left|\mathbf{W}_f \mathbf{X}\right|^{\gamma_f} , \tag{1}$$

$$\mathbf{Z}_q[k,n] := \mathbf{Q}_q \left|\mathbf{W}_q \mathbf{F}^{-1} \mathbf{Z}_f\right|^{\gamma_q} , \tag{2}$$

where $\mathbf{Z}_f$, $\mathbf{Z}_q \in \mathbb{R}^{F \times T}$, k and n are the frequency and time index respectively, $\mathbf{F}$ is the DFT matrix, $\mathbf{W}_f$ and $\mathbf{W}_t$ are high-pass filters to discard low-varying parts [14], and the element-wise power-scaled nonlinear activation function $|\cdot|^\gamma$ is defined as: $|x|^\gamma := |\max(0,x)|^\gamma$. We follow [15] to set the parameters $(\gamma_f, \gamma_q) = (0.24, 0.6)$. $\mathbf{Q}_f, \mathbf{Q}_q \in \mathbb{R}^{F \times 2F}$ are two triangular filterbanks: they respectively map a feature from the frequency or time domain to the positive log-frequency domain. Both filterbanks have 352 triangular filters, with 48 semitones per octave, ranging from 27.5 Hz (A0) to 4487 Hz (C8). Therefore, we have $F = 352$, and T is dependent on the length of input, In brief, $\mathbf{Z}_f$ represents a *power-scaled spectrogram* showing the fundamental frequencies and their harmonics in a signal, and $\mathbf{Z}_q$ represents a *generalized cepstrum* showing the fundamental frequencies and their sub-harmonics [16, 17]. [14, 16, 17] pointed out that combining $\mathbf{Z}_f$ and $\mathbf{Z}_q$ for the MPE models can give significantly better MPE performance than using the spectrogram feature.

Next, we consider the harmonic information. One key idea of the HCQT is the multi-channel data representation combining a pre-computed time-frequency representation and its pitch-shifted version such that the harmonic peaks of a pitched component are aligned to the same pixel [10]. The main purpose of doing so is to make a local convolutional kernel cover the global pitch profile (i.e., the whole harmonic pattern) of a component. In this paper, we extend this idea to both the power spectrogram and the generalized cepstrum. Consider the mth harmonic frequency mf_0 of a fundamental frequency f_0. According to equal temperament, the pitch number of mf_0 is $\eta(m) := \text{round}(12 \log_2 m)$ semitones higher than f_0. For example, the 2nd, 3rd, and 4th harmonics of f_0 are 12, 19, and 24 semitones higher than f_0 respectively. Similarly, the 2nd, 3rd, and 4th sub-harmonics are lower than f_0 by 12, 19, and 24 semitones. We therefore consider the following two data representations:

$$\mathbf{Z}_f^{(m)}[k,n] := \mathbf{Z}_f\left[k + \eta(m) \cdot \delta, n\right] \tag{3}$$

$$\mathbf{Z}_q^{(m)}[k,n] := \mathbf{Z}_q\left[k - \eta(m) \cdot \delta, n\right] \tag{4}$$

where δ is the number of bins per semitone. Notice that $\mathbf{Z}_f^{(1)} = \mathbf{Z}_f$ and $\mathbf{Z}_q^{(1)} = \mathbf{Z}_q$. In this work, we consider $\mathbf{Z}_f^{(m)}$ and $\mathbf{Z}_q^{(m)}$ for $m = 1, 2, \cdots, 6$. We align these representations along the channel axis, which therefore form a 12-channel representation $\mathbf{Z}_{\text{HCFP}} := [\mathbf{Z}_f^{(1:m)}, \mathbf{Z}_q^{(1:m)}]$. This will be referred to as the HCFP representation in the following sections. In this multi-channel data representation, every pixel contains not only the spectral component but also the harmonic and sub-harmonic peaks. Since the frequency scale is 48 semitones per octave, so we have $\delta = 4$. Besides HCFP, in this paper, we also discuss the CFP representation without high-order harmonics (i.e., $m = 1$), namely $\mathbf{Z}_{\text{CFP}} := [\mathbf{Z}_f^{(1)}, \mathbf{Z}_q^{(1)}]$.

The input audio recordings are mono-channel and are re-sampled to 16 kHz. The STFT is computed with a Blackman-Harris window which size is 0.128 second, namely 2,048 samples. The hop size of STFT is 0.02 second.

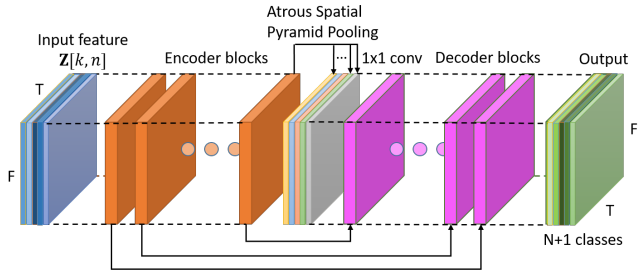

Fig. 1. The conceptual diagram of the proposed system.

2.2. Model

Fig. 1 illustrates a schematic diagram of the proposed multi-instrument transcription system. The proposed model is based on our prior work used in [11], but with some modification to fit our task requirement. This model is originally based on the DeepLabV3 and its improved version, DeepLabV3+ [4], both which are fully convolution neural networks with an encoder-decoder architecture. The original model output was only the binary prediction on one channel, which is in a single-task mode. To predict multiple instrument activation, we extend the output to $(N+1)$ channels to recognize N classes of instrument plus one non-instrument channel. That means, each channel represents one of the transcription of one class instrument. This allows the model to predict multiple instruments at the same time. The output value of each pixel is between 0 and 1, which represents whether the pitch at that moment is off or on.

The input of the model is the $2m$-channel data representation $\mathbf{Z}_{\text{HCFP}} \in \mathbb{R}^{2m \times F \times T}$. The model input is then fed through a sequence of encoder blocks, each of which is constructed by a convolution, a ReLU function, a stride convolution and a skip connection [11]. The decoder block is constructed by the stack of convolution and transpose convolution for better resolution in the decoded image [18]. We also adopt the U-net structure to our model, in which the output from the encoder layers and the corresponding decoder layers are concatenated.

We use padding for each convolution such that the output of each block has the same dimension $F \times T$.

One important feature of the model is the use of atrous spatial pyramid pooling (AASP) in between the encoding and decoding processes [19]. It introduces convolution operations with several dilation sizes, and pooling them together to capture objects in various scales. We adopt the focal loss [5] as the loss function to solve the class imbalance problem. Since most of the pixels in the output are silence (i.e., no activation at that time-frequency position), the model would tend to predict all the pixels to be silence when using conventional loss function such as binary cross entropy. The focal loss provides a weighting factor for balancing the importance between active and silence examples, and such a loss function has been found useful in vocal melody extraction [11].

3. EXPERIMENTS

3.1. Data

We evaluate our proposed methods on two datasets, MAPS [20] and MusicNet [21]. The MAPS dataset is one of the most widely used dataset in AMT. The dataset contains 270 annotated piano solo pieces generated from nine different piano sources, two of which are real-world and seven of which are synthesized pianos. Each source has 30 pieces. The MusicNet dataset contains totally 330 pieces of classical music recordings in solo or chamber music. It contains 11 kinds of instruments, including piano, harpsichord, violin, viola, cello, contrabass, horn, oboe, bassoon, clarinet, and flute. All of the pitch annotations in this dataset have instrument labels.

We follow the Configuration II [22] to split the MAPS dataset into the training and testing parts: 210 pieces from the seven synthesized piano are for training (180 pieces) and validation (30 pieces), and the remaining 60 pieces from the real piano, ENSTDkAm and ENSTDkCl, are set aside for testing. The MusicNet dataset is split into training and testing subsets according to the default setting provided by the authors of the dataset: 320 pieces are the training set and the remaining 10 pieces are the testing set. The IDs of the audio files contained in the test set are: 1759, 1819, 2106, 2191, 2298, 2303, 2382, 2416, 2556, and 2628.

3.2. Experimental settings

We consider two frame-level MPE evaluation schemes: 1) conventional MPE (denoted as C-MPE), which only counts the accuracy of pitch, as discussed in the previous studies [7, 10, 17], 2) multi-instrument MPE (denoted as MI-MPE), which takes a prediction as a true positive only when both the pitch and the instrument class are correctly predicted. For data representation, we compare CFP to HCFP, the latter is expected more suitable for MI-MPE as it incorporates harmonic information. To verify the performance of the proposed semantic segmentation model and the HCFP feature on both C-MPE and MI-MPE tasks, We perform the following two experiments:

- First, we take the CFP representation as the input of the proposed model, and evaluate C-MPE performance on the MAPS and MusicNet dataset. Here the instrument information of the MusicNet dataset is not used; that means, the model output has only two channels representing no matter an event is activated or inactivated.
- Second, we take the HCFP representation as the model input and evaluate MI-MPE performance on the MusicNet dataset. Since the MusicNet dataset has 11 instruments, the model output has 12 channels for all instrument classes and one non-instrument class.

The metrics we used for evaluation are precision, recall, and F-measure, which can be computed by counting the number of true positive (TP), false positive (FP), and false negative (FN) over all frames in the testing data: $\mathrm{P} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, $\mathrm{R} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, and $\mathrm{F} = 2\mathrm{PR}/(\mathrm{P} + \mathrm{R})$. We use the `mir_eval.multipitch` function with default parameter setting in the `mir_eval` library to evaluate all experiments. For the evaluation on MI-MPE, we compute the instrument-wise F-measure for each individual output channel. Since a music piece does not have all the 11 instruments, we compute the F-measures only for the instruments that really appear in that piece. The resulting P, R, F are the average of these results of all pieces over the full dataset.

Finally, we need a process to tune the threshold on the model output to obtain the final transcription result. The optimal threshold for each output channel is obtained by a grid search from 0 to 1 over the validation set. The target of this search is to get the highest F-measure in the validation set. The test data are not used in such a threshold tuning process.

All experiments are run on an Ubuntu 16.04 computer with i7-7700 CPU and a 64G RAM. We also use an GTX 1080 Ti GPU card to accelerate our training. The model is implemented using the Keras library with TensorFlow backend. The parameters are updated with stochastic gradient descent (SGD) using the ADAM optimizer. The initial learning rate is set to 0.001. For each training epoch, we fix the number of updating step to 6,000 mini-batches, and every mini-batch has 12 ramdomly selected input segments. Each segment has 128 frames, equivalent to 2.56 seconds. Our source code will be announced online for reproducibility. [1]

3.3. Results

Table 1 lists the C-MPE results on the two experimental datasets. We compare our results to the recently-published state-of-the-art methods: for the MAPS dataset, we list the result of the Onset and Frame model [12] trained with the

[1] https://github.com/BreezeWhite/Music-Transcription-with-Semantic-Segmentation

Dataset	Method	P	R	F1
MAPS	[23]	**92.86**	78.46	84.91
	Proposed ($\mathbf{Z}_{\text{CFP}}$)	87.48	**86.29**	**86.73**
MusicNet	[21]	68.71	77.30	72.75
	Proposed ($\mathbf{Z}_{\text{CFP}}$)	**69.34**	**79.29**	**73.70**
	Proposed ($\mathbf{Z}_{\text{HCFP}}$)	68.98	78.85	73.20

Table 1. Results (in %) of frame-level C-MPE.

	CFP ($\mathbf{Z}_{\text{CFP}}$)			HCFP ($\mathbf{Z}_{\text{HCFP}}$)		
	P	R	F1	P	R	F1
pf	53.39	64.98	58.59	54.25	78.36	**63.17**
vn	37.18	70.84	**48.20**	39.84	65.14	47.28
va	26.13	46.00	32.90	27.04	42.42	**32.98**
vc	35.38	48.03	40.67	30.19	70.29	**41.72**
hn	64.66	64.12	**64.34**	52.19	52.73	52.46
bn	26.49	43.74	32.98	27.41	70.87	**39.52**
cl	40.62	77.96	53.41	42.45	77.89	**54.95**

Table 2. Results (in %) of MI-MPE. Instrument labels are: pf = piano, vn = violin, va = viola, vc = cello, hn = horn, bn = bassoon, and cl = cello.

MAESTRO dataset [23]; and for MusicNet we list the result of a novel CNN-based network proposed in [21]. Since [21] only reported the results on the first 90 seconds of three testing pieces, we take a re-implemented code and obtain the results on all the test data after confirming its performance on the setting used in [21] without data augmentation.[2]

The F-measure of our proposed model on the MAPS dataset achieves 86.73%, which outperforms [23] by 1.8%. To our best knowledge, this is the recorded high performance of C-MPE on the MAPS dataset under Configuration II. For MusicNet, the F-measure of the proposed method also outperforms the state-of-the-art method by 1%. The proposed semantic segmentation model together with the CFP feature is shown highly competitive in terms of C-MPE.

Fig. 2 illustrates two selected transcription results on MAPS (a piano solo) and MusicNet (a string quartet) using $\mathbf{Z}_{\text{CFP}}$. The result of piano solo clearly demonstrate the power of the proposed model: even when the sustain pedal is performed through out the piece, most music notes are correctly recognized, and their onset and offset time are well captured. This indicates the high potential of the proposed model in note-level transcription. For the result on the string quartet, there are some more errors in the inner parts possibly due to the diversity of instruments. However, it still captures the pitches in the highest and lowest parts well.

The last row of Table 1 shows that using the HCFP feature can achieve performance comparable to, but still lower than the one of CFP, possibly because the harmonic information is not that relevant to the fundamental frequencies.

[2]https://github.com/yoyololicon/translation-invariant

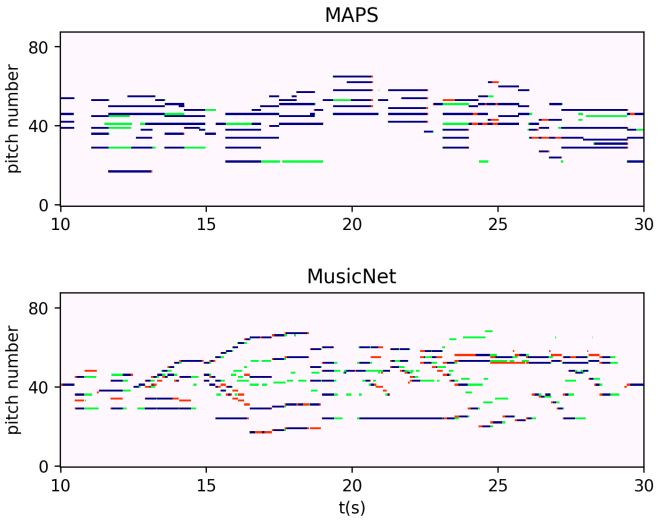

Fig. 2. The transcription results (in piano roll) of two excerpts from MAPS and MusicNet respectively. Top: the result of 'MAPS_MUS-bk_xmas5_ENSTDkCl.wav' in the 'MAPS/ENSTDkCl' subset, which achieves an F-measure of 0.9102. Bottom: the result of '2106.wav' in 'MusicNet/test_data', which achieves an F-measure of 0.6861. Blue line: TP. Green line: FP. Red line: FN.

The effectiveness of HCFP is demonstrated in MI-MPE. Table 2 compares the MI-MPE results on MusicNet using the CFP and HCFP representations. HCFP outperforms CFP except for the cases of violin and horn. This indicates that HCFP can generally improve multi-instrument transcription with the incorporation of harmonic feature. As expected, the harmonic information in HCQT could contribute more on specifying the spectral patterns of different instruments.

4. CONCLUSION

We have demonstrated the effectiveness of semantic segmentation models in joint transcription of pitch and instrument, by reporting new state-of-the-art performance values on conventional MPE tasks, and a new benchmark on multi-instrument MPE. We also found that designing data representations revealing pitch and instrument information from the aspect of signal processing is a critical step to harness the power of semantic segmentation models. These findings indicate further research directions on more practical AMT tasks, such as note-level transcription and multipitch streaming.

5. ACKNOWLEDGEMENT

This work is partially supported by MOST Taiwan, under the contract MOST 106-2218-E-001-003-MY3. The authors would like to thank Chin-Yun Yu for providing his implementation of the algorithm of [21].

6. REFERENCES

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE ICCV*, 2017, pp. 2980–2988.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE CVPR*, 2015, pp. 3431–3440.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015, pp. 234–241.

[4] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *eprint arXiv:1706.05587*, 2017.

[5] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar., "Focal loss for dense object detection," *eprint arXiv:1708.02002*, 2017.

[6] J. J. Carabias-Orti, T. Virtanen, P. Vera-Candeas, N. Ruiz-Reyes, and F. J. Cañadas-Quesada, "Musical instrument sound multi-excitation model for non-negative spectrogram factorization," *IEEE J. Sel. Topics Signal Proc.*, vol. 5, no. 6, pp. 1144–1158, 2011.

[7] E. Benetos, S. Cherla, and T. Weyde, "An effcient shift-invariant model for polyphonic music transcription," in *Proc. Int. Workshop on Machine Learning and Music*, 2013.

[8] E. Benetos and S. Dixon, "A shift-invariant latent variable model for automatic music transcription," *Computer Music Journal*, vol. 36, no. 4, pp. 81–94, 2012.

[9] A. Jansson, E. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proc. ISMIR*, Oct. 2017.

[10] R. M. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello, "Deep salience representations for f0 estimation in polyphonic music.," in *Proc. ISMIR*, 2017, pp. 63–70.

[11] W.-T. Lu and L. Su, "Vocal melody extraction with semantic segmentation and audio-symbolic domain transfer learning," in *Proc. ISMIR*, 2018.

[12] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," in *Proc. ISMIR*, 2018.

[13] R. M. Bittner, B. McFee, and J. P. Bello, "Multitask learning for fundamental frequency estimation in music," *arXiv preprint arXiv:1809.00381*, 2018.

[14] Y.-T. Wu, B. Chen, and L. Su, "Automatic music transcription leveraging generalized cepstral features and deep learning," in *Proc. IEEE ICASSP*, 2018, pp. 401–405.

[15] L. Su, T.-Y. Chuang, and Y.-H. Yang, "Exploiting frequency, periodicity and harmonicity using advanced time-frequency concentration techniques for multipitch estimation of choir and symphony," in *Proc. ISMIR*, 2016, pp. 393–399.

[16] G. Peeters, "Music pitch representation by periodicity measures based on combined temporal and spectral representations," in *Proc. IEEE ICASSP*, 2006.

[17] L. Su and Y.-H. Yang, "Combining spectral and temporal representations for multipitch estimation of polyphonic music," *IEEE/ACM Trans. Audio, Speech, Language Processing*, vol. 23, no. 10, pp. 1600–1612, 2015.

[18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *arXiv preprint arXiv:1802.02611*, 2018.

[19] L.-C. Chen, Y. Zhu, P. George, S. Florian, and A. Hartwig, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *eprint arXiv:1802.02611*, 2018.

[20] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 18, no. 6, pp. 1643–1654, 2010.

[21] J. Thickstun, Z. Harchaoui, D. P. Foster, and S. M. Kakade, "Invariances and data augmentation for supervised music transcription," in *Proc. IEEE ICASSP*, 2018, pp. 2241–2245.

[22] R. Kelz, M. Dorfer, F. Korzeniowski, S. Böck, A. Arzt, and G. Widmer, "On the potential of simple framewise approaches to piano transcription," *.arXiv:1612.05153*, 2016.

[23] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the maestro dataset," *arXiv preprint arXiv:1810.12247*, 2018.