POLYPHONIC PIANO NOTE TRANSCRIPTION WITH NON-NEGATIVE MATRIX FACTORIZATION OF DIFFERENTIAL SPECTROGRAM

Lufei Gao[*], *Li Su*[†], *Yi-Hsuan Yang*[†], *Tan Lee*[*]

[*]Department of Electronic Engineering, The Chinese University of Hong Kong
[†]Academia Sinica, Taiwan
Emails: [*]{lgao, tanlee}@ee.cuhk.edu.hk, [†]{lisu, yang}@citi.sinica.edu.tw

ABSTRACT

Automatic music transcription is usually approached by using a time-frequency (TF) representation such as the short-time Fourier transform (STFT) spectrogram or the constant-Q transform. In this paper, we propose a novel yet simple TF representation that capitalizes the effectiveness of spectral flux features in highlighting note onset times. We refer to this representation as the differential spectrogram and investigate its usefulness for note-level piano transcription using two different non-negative matrix factorization (NMF) algorithms. Experiments on the MAPS ENSTDkCl dataset validate the advantages of the differential spectrogram over the STFT spectrogram for this task. Moreover, by adapting a state-of-the-art convolutional NMF algorithm with the differential spectrogram, we can achieve even better accuracy than the state-of-the-art on this dataset. Our analysis shows that the new representation suppresses unwanted TF patterns and performs particularly well in improving the recall rate.

Index Terms— Music information retrieval, spectral flux, differential spectrogram, non-negative matrix factorization

1. INTRODUCTION

Automatic music transcription (AMT) aims at transcribing a musical audio signal into a symbolic representation akin to the form of a musical score. A great number of algorithms have been proposed for AMT since the pioneering work of Moorer [1]. Some researchers focus on audio signal processing and the design of scoring function for pitch detection [2–5], while others employ machine learning algorithms such as non-negative matrix factorization (NMF) [6–9], sparse coding [10, 11], probabilistic models [12, 13] or classification-based models [14–17] to tackle the problem.

The audio representation adopted in these endeavors is usually a time-frequency (TF) representation such as the short-time Fourier transform (STFT) spectrogram or the constant-Q transform. For example, a widely studied approach is to use NMF or its variants to decompose a given TF representation into two non-negative components: a template of atoms usually formed by the spectra of musical notes, and an activation matrix indicating the temporal evolution of the notes [6]. NMF is an attractive solution partly due to the flexibility and expressivity of its model structure. For example, it is easy to add regularizers informed by musical knowledge [18–20]. The non-negative assumption also works well for many TF representations. Although recent years have witnessed an increasing interest in learning features for AMT by deep neural networks [17], to date NMF-based methods still represent the state-of-the-art in many subtasks of AMT, such as the note-level transcription of piano music [9].

Note-level music transcription requires accurate estimate of the pitches, onset and offset times of the musical notes [21]. For transcription of pitched percussive intruments such as piano, it has been found beneficial to employ instrument-specific acoustics to model the attack and decay characteristics of the music signal [9, 22, 23]. However, due to the rich acoustic variation seen in real world performances, parametric models of musical acoustics may not always work well. Moreover, as partials of different pitches overlap in the given TF representation, notes that are softly played can be easily missed in the resulting transcription.

Being inspired by the dedicated efforts on musical onset detection [24, 25], we investigate in this paper a simple yet relatively less explored idea of highlighting local energy increase in the TF representation for more reliable note-level transcription. Specifically, we propose a new TF representation, named *differential spectrogram* by using the idea of spectral flux (SF) [24] to emphasize positive energy changes in the spectrogram, thereby suppressing unwanted energy fluctuations due to partials, noises or room acoustics. As the differential spectrogram is non-negative, we also propose algorithms for note transcription based on existing NMF-based methods. We validate the advantages of the differential spectrogram over conventional STFT spectrogram through experiments with a piano dataset, and discuss its performance from a signal-level perspective. The experimental results show that using the proposed approach leads to an onset-aware ($\pm$50ms) F-measure 85.6%, which appears to be the

This research is partially supported by a GRF project grant (Ref: 14204014) from Hong Kong Research Grants Council.

best one reported for the MAPS ENSTDkCl dataset [12].

A similar idea is presented in [26], which learned feature representations for piano note transcription from the rectified first-order difference of semitone-filtered spectrograms by deep learning techniques. But by virtue of the neural network model it is hard to gain signal-level insights. Moreover, evaluation on the same dataset suggests that our proposed method leads to more accurate note-level piano transcription.

In what follows, we review two existing NMF methods in Section 2, and present the proposed feature representation and the adapted models in Section 3. Experimental results are reported in Section 4, followed by conclusions in Secion 5.

2. BASELINE METHODS

Given an input audio signal, most existing NMF-based methods use the STFT spectrogram as the feature representation and perform factorization using a pre-learned template from single-note recordings. Variants of NMF algorithms differ mainly in the formulation of the factorization model and the objective function. We consider here the standard NMF algorithm for its popularity and a more advanced convolutional NMF algorithm designed for note transcription.

2.1. Standard NMF (NMF)

Assuming that the spectrum is a linear combination of some single-note spectra, NMF tries to approximate the STFT spectrogram X_{ft} as the product of two non-negative matrices:

$$X_{ft} \simeq V_{ft} = \sum_{k=1}^{K} W_{fk} H_{kt}\,, \tag{1}$$

where $\mathbf{W}$ is the template of single-note spectra, $\mathbf{H}$ is the time-varying activation, K is the number of notes set to 88 in this work, $f \in [1, F]$ and $t \in [1, T]$ denote frequency bin and time frame index, respectively. The distortion $D(\mathbf{X}|\mathbf{V})$ is measured by the β-divergence, which encompasses the Itakura-Saito (IS) divergence (when $\beta \to 0$), the Kullback-Leibler (KL) divergence (when $\beta = 1$) and the Euclidean distance (when $\beta = 2$). The parameters $\mathbf{W}$ and $\mathbf{H}$ are estimated according to the multiplicative update rules [7].

NMF can be performed in an unsupervised way, where both W and H are directly computed from the input spectrogram. However, to facilitate pitch estimation based on the activation patterns, a supervised approach which incorporates a pre-learned W is preferred [6].

2.2. Attack/Decay Convolutional NMF (CNMF-AD)

A drawback of NMF is that a large number of template atoms may be needed to account for the rich variation in note intensity and recording environment. This can be circumvented by employing instrument-specific acoustics with the convolutional NMF (CNMF) model [27] to capture the attack and decay characteristics of the musical audio, as recently demonstrated by Cheng *et al.* for piano music [9]. In this model, the STFT spectrogram is assumed to be the summation of two parts: the attack phase and the decay phase. Mathematically, it is defined as

$$\begin{aligned} V_{ft} = &\sum_{k=1}^{K} W^a_{fk} \sum_{\tau=t-T_t}^{t+T_t} H_{k\tau} P(t-\tau) \\ &+ \sum_{k=1}^{K} W^d_{fk} \sum_{\tau=1}^{t} H_{k\tau} e^{-(t-\tau)\alpha_k}\,, \end{aligned} \tag{2}$$

where $\mathbf{W}^a$ is the percussive template for the attack phase, $\mathbf{W}^d$ is the harmonic template for the decay phase, $\mathbf{P}$ and α_k are the transient pattern and the exponential decay rate, respectively, and T_t determines the range of the transient pattern. Convolving $\mathbf{H}$ with $\mathbf{P}$ (or the exponential function), the attack (or decay) activation is obtained and denoted by $\mathbf{H}^a$ (or $\mathbf{H}^d$). Using the KL divergence for measuring distortion, the objective is to minimize $D(\mathbf{X}|\mathbf{V}) = \sum_{f,t} d(X_{ft}, V_{ft})$, where $d(x, y) = x \cdot \log(\frac{x}{y}) - x + y$, for $x, y > 0$. The parameters $\{\mathbf{W}^a, \mathbf{W}^d, \mathbf{H}, \mathbf{P}, \alpha\}$ are estimated by the multiplicative update rules derived in [9].

3. PROPOSED FEATURE AND MODEL ADAPTATIONS

In this section, we describe the proposed feature representation and its variants, and the adaptations of the aforementioned NMF models using the new feature representation.

3.1. Differential Spectrogram

Assuming that the intrument exhibits harmonics with locally stable frqeuencies, the differential spectrogram $\widehat{X}_L(f,t)$ is defined as:

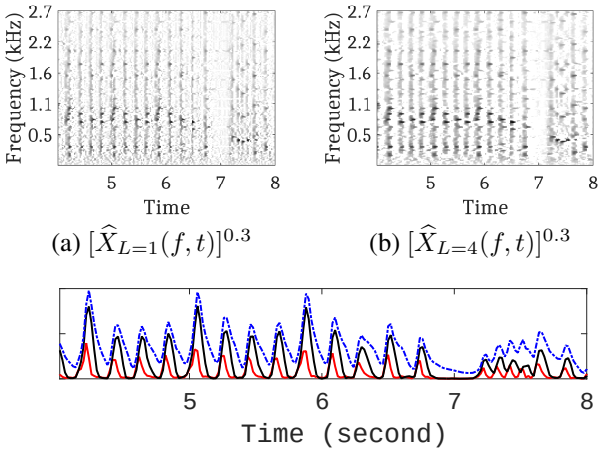
$$\widehat{X}_L(f,t) = \text{HWR}(|X(f,t+L)| - |X(f,t)|), \tag{3}$$

where HWR stands for the half-wave rectification ($\text{HWR}(x) = \frac{x+|x|}{2}$) and L is a positive integer determining the distance from the present frame to a preceding one. Figs. 1(a) and 1(b) illustrate two examples where $L = 1$ and $L = 4$, respectively. We can see that the differential spectrogram with larger L is less spotted and the TF patterns around onsets are emphasized. Fig. 1(c) shows the spectral flux of the same signal, defined as $\text{SF}_L(t) = \sum_{f=1}^{F} \widehat{X}_L(f,t)$. We can see that by increasing the distance, the SF peaks shift towards the peaks of the mixture signal. These are desirable properties for capturing the onset charateristics in the note transcription. For the instruments with oscillatory harmonic frequencies, a semitone filterbank can be applied prior to the difference operation to suppress the frequency modulations.

3.2. Model Adaptations

3.2.1. Standard NMF adaptation (NMF-Δ)

To incorporate the new feature into the standard NMF model, we can directly replace the spectrogram with the differential



(a) $[\widehat{X}_{L=1}(f,t)]^{0.3}$ (b) $[\widehat{X}_{L=4}(f,t)]^{0.3}$

(c) Red: $\mathrm{SF}_{L=1}(t)$; black: $\mathrm{SF}_{L=4}(t)$; blue: $\sum_f |X(f,t)|$.

Fig. 1: Illustration of the proposed differential spectrogram and the resulting spectral flux curves.

spectrogram. However, there are still many undesirable spots in the differential spectrogram, as shown in Figs. 1(a)(b). This significantly deteriorates our attempt to improving note transcription performance. Therefore, for standard NMF, we use the following feature to replace $\mathbf{X}$ and then follow Eq. (1) to get the decomposition,

$$\widetilde{X}_L(f,t) = c_1 X(f,t) + c_2 \widehat{X}_L(f,t), \tag{4}$$

where $0 \leq c_1, c_2 \leq 1$ are two scalars to weight the two terms.

3.2.2. Convolutional NMF adaptation (CNMF-Δ)

To approximate the differential spectrogram, we concentrate on the attack phase of Eq. (2) where the recurring pattern is theoretically stable. Specifically, the following model is utilized to estimate the note activation.

$$\widehat{X}_L(f,t) \simeq \widehat{V}_{ft} = \sum_{k=1}^{K} \widehat{W}_{fk} \sum_{\tau=t-T_t}^{t+T_t} \widehat{H}_{k\tau} \widehat{P}(t-\tau). \tag{5}$$

Convolving $\widehat{\mathbf{H}}$ with $\widehat{\mathbf{P}}$ yields the attack activation denoted by $\widehat{\mathbf{H}}^a$. The parameters $\{\widehat{\mathbf{W}}, \widehat{\mathbf{P}}, \widehat{\mathbf{H}}\}$ are estimated by the multiplicative update rules derived from Eq. (5), in a similar way as for (2).

3.2.3. Model initialization (CNMF-AD-Δ)

Note activation $\widehat{\mathbf{H}}$ can be estimated using the adapted model (5) with random initialization. But in this way, the information contained in the decay phase would be completely ignored. Another approach is to intialize it by $\mathbf{H}$ estimated using (2). It is expected that the activation values of some softly played notes are boosted. In the following, we use CNMF-Δ to represent the model (5) with random initialization of $\widehat{\mathbf{H}}$ and CNMF-AD-Δ to represent (5) with $\widehat{\mathbf{H}}$ intialized by $\mathbf{H}$.

4. EXPERIMENTS

In this section, we first elaborate the experimental settings, and then analyze the performance of the proposed approach with a dataset recorded on a Disklavier piano, using three state-of-the-art transcription methods for comparison.

4.1. Experimental Settings

As input, the system takes an audio signal with a sampling rate of 44.1 kHz. We segment frames by a Hamming window of 4096 samples and a hop-size of 882 samples. With 2-fold zero-padding, 8192-point discrete Fourier transform is computed on each frame. The spectrogram is smoothed with a median filter covering 100ms. The update algorithms are iterated for 50 times. T_t equal to 4 frames. After estimating $\mathbf{H}$, we employ the strategies proposed in [9] to detect onsets from $\mathbf{H}$. The threshold $\Theta_k(t)$ for peak picking is adapted to each music piece, expressed as $\Theta_k(t) = \frac{1}{M}\sum_{m=0}^{M-1} H^a_{k,t+m} + \delta \max_{k,t} H^a_{k,t}$. In this work, $M = 20$, $\delta = -23$dB, $c_1 = 1$, $c_2 = 1$ for NMF and -29dB for CNMF.

The training set contains the 88 forte isolated note recordings in the subset "ENSTDkCl" of MAPS [12]. The test dataset includes the 30 music pieces from the same subset. Only the first 30-second excerpt of each piece is used. For each model, the note activation is fixed according to the ground-truth and the other parameters are updated in the training stage. During testing, only the note activation is updated.

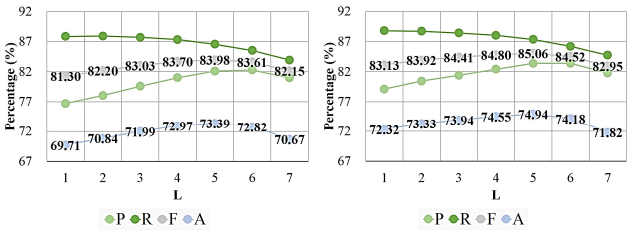
The following evaluation measures are employed: precision ($P = \frac{N_{tp}}{N_{tp}+N_{fp}}$), recall ($R = \frac{N_{tp}}{N_{tp}+N_{fn}}$), F-measure ($F = \frac{2PR}{P+R}$) and accuracy ($A = \frac{N_{tp}}{N_{tp}+N_{fp}+N_{fn}}$), where N_{tp}, N_{fp} and N_{fn} are the numbers of true positives, false positives and false negatives, respectively. We count a note estimate as a true positive if the pitch is correct and its onset time is within 50ms of the ground-truth time.

4.2. Result Analysis

4.2.1. System settings

We first investigate the effect of L in Eq. (3). Fig. 2 shows the results using different distances. We can see that there is a trade-off between precision and recall when increasing the distance until it reaches a certain value. Both systems achieve the best F-measure and accuracy when $L = 5$.

Comparing Fig. 2(a) with 2(b) demonstrates the effectiveness of the strategy proposed in Sec. 3.2.3. We can see that initializing CNMF-Δ with the estimated $\mathbf{H}$ of CNMF-AD can increase both F-measure and accuracy by 1 to 2 percent. However, it is suspected whether the performance improvement is due to more iterations for updating $\mathbf{H}$. A simple test is conducted by initializing $\mathbf{H}$ with the first-round estimate and updating it using CNMF-AD for 50 times again. It



(a) CNMF-Δ (b) CNMF-AD-Δ

Fig. 2: Results of our methods with different values of L.

Table 1: Performance comparison on "ENSTDkCl"

Method	P	R	F	A
NMF ($\beta = 0.5$)	59.70	34.51	41.81	27.24
NMF-Δ ($\beta = 0.5$)	71.04	42.48	50.70	35.13
NMF ($\beta = 2$)	51.67	43.11	46.34	30.54
NMF-Δ ($\beta = 2$)	67.83	58.21	61.76	45.10
CNMF-Δ	82.11	86.57	83.98	73.39
CNMF-AD-Δ	83.38	**87.34**	**85.06**	**74.94**
CNMF-AD [9]	**89.22**	78.35	82.91	71.55
Böck [26]	–	–	–	68.70
Berg-Kirkpatrick [13]	78.10	74.70	76.40	–

is verified that using only CNMF-AD does not improve the performance even with more updating iterations.

4.2.2. *Comparison with existing methods*

We compare our systems to three state-of-the-art systems for the note-level transcription: the attack/decay model (CNMF-AD) [9], the bidirectional Long Short-Term Memory (BLSTM) recurrent neural network [26], and an unsupervised probabilistic model [13][1]. To our knowledge, the attack/decay model reports the best F-measure and accuracy on the test dataset thus far.

The results are shown in Table 1[2]. Both CNMF-Δ and CNMF-AD-Δ achieve better F-measure and accuracy rate than the other systems. Although the standard NMF models do not yield good performances, it is obvious that replacing the spectrogram with the feature representation defined in Eq. (4) significantly increases F-measure and accuracy when either $\beta = 0.5$ or $\beta = 2$.

To understand the performance enhancement, the attack activations, i.e. $\mathbf{H}^a$ of CNMF-AD and $\widehat{\mathbf{H}}^a$ of CNMF-Δ, are plotted as in Fig. 3. These are the attack activations from 20 to 28 second of note E4 of the file "MAPS_MUS-mz_331_3_ENSTDkCl". We see that the attack activation of

[1]The training and testing data are from the same piano for our methods and CNMF-AD, which does not hold for the other two methods.

[2]In this paper, the NMF-based method is implemented to illustrate the effectiveness of the proposed feature. Parameters are not fully tuned.

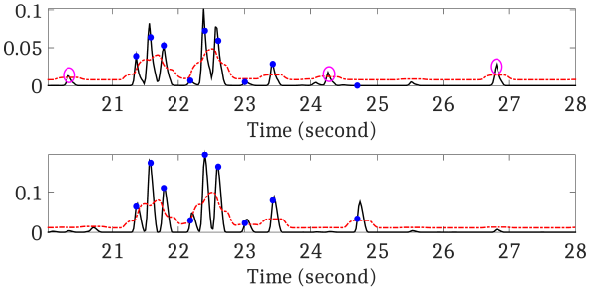
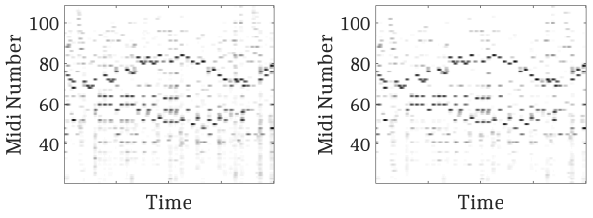


Fig. 3: Above is $H^a(44,t)$ of CNMF-AD; below is $\widehat{H}^a(44,t)$ of CNMF-Δ. Blue dots: true onsets; red dash lines: $\Theta_{44}(t)$ where $\delta = -34$dB.



(a) $(\widehat{\mathbf{H}}^a)^{0.3}$ of CNMF-Δ (b) $(\widehat{\mathbf{H}}^a)^{0.3}$ of CNMF-AD-Δ

Fig. 4: The rolls of attack activations of our methods.

CNMF-Δ is more clean and prominent. The false alarms indicated by the purple circles in CNMF-AD are suppressed in CNMF-Δ as well. This illustrates the benefits of using the differential spectrogram.

To illustrate that exploiting estimated $\mathbf{H}$ of CNMF-AD in the initialization of CNMF-Δ benefits the estimation, the rolls of $(\widehat{\mathbf{H}}^a)^{0.3}$ of the two models are shown as in Fig. 4. We can observe that the roll of CNMF-AD-Δ contains milder and less spots than that of CNMF-Δ, which could be the underlying reason that the accuracy rate can be further refined.

5. CONCLUSION

In this paper, we have proposed a new time-frequency representation called differential spectrogram for polyphonic piano note transcription. We adapt the standard NMF model and the attack/decay CNMF model to employ the proposed feature as their inputs. Evaluations on a piano dataset validate the effectiveness of our methods. In the future, differential spectrogram will be further developed to suppress the undesirable components in order to remove false alarms. We also plan to validate the effectiveness of the proposed approach with datasets of other instruments.

6. ACKNOWLEDGEMENT

We would like to thank Ms. Tian Cheng for her excellent work [9] and for generously providing her codes.

7. REFERENCES

[1] J. A. Moorer, "On the transcription of musical sound by computer," *Computer Music Journal*, pp. 32–38, 1977.

[2] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 2, pp. 255–266, Feb 2008.

[3] C. Yeh, *Multiple Fundamental Frequency Estimation of Polyphonic Recordings*, Thèse de doctorat, University Paris 6 (UPMC), Paris, 2008.

[4] K. Dressler, "Multiple fundamental frequency extraction for mirex 2012," in *MIREX*, 2012.

[5] A. Pertusa and J .M. Iñesta, "Multiple fundamental frequency estimation using gaussian smoothness," in *ICASSP*, March 2008, pp. 105–108.

[6] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *WASPAA*, Oct 2003, pp. 177–180.

[7] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 528–537, March 2010.

[8] B. Fuentes, R. Badeau, and G. Richard, "Harmonic adaptive latent component analysis of audio and application to music transcription," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 9, pp. 1854–1866, 2013.

[9] T. Cheng, M. Mauch, E. Benetos, and S. Dixon, "An attack/decay model for piano transcription," ISMIR, 2016.

[10] C.-T. Lee, Y.-H. Yang, and H.-H. Chen, "Multipitch estimation of piano music by exemplar-based sparse representation," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 608–618, June 2012.

[11] L. Gao and T. Lee, "Multi-pitch estimation based on sparse representation with pre-screened dictionary," in *MMSP*, 2015.

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

[13] T. Berg-Kirkpatrick, J. Andreas, and D. Klein, "Unsupervised transcription of piano music," in *NIPS*, 2014, pp. 1538–1546.

[14] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *IEEE Trans. Multimedia,*, vol. 6, no. 3, pp. 439–449, 2004.

[15] G. E. Poliner and D. PW Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP*, vol. 2007, no. 1, pp. 154–154, 2007.

[16] J. Nam, J. Ngiam, H. Lee, and M. Slaney, "A classification-based polyphonic piano transcription approach using learned feature representations.," in *ISMIR*, 2011, pp. 175–180.

[17] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic music transcription," *ArXiv e-prints*, Aug. 2015.

[18] A. Cont, "Realtime multiple pitch observation using sparse non-negative constraints," in *ISMIR*, 2006, pp. 206–211.

[19] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1066–1074, 2007.

[20] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 538–549, 2010.

[21] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of multiple-f0 estimation and tracking systems.," in *ISMIR*, 2009, pp. 315–320.

[22] J. Wu, E. Vincent, S. A. Raczynski, T. Nishimoto, N. Ono, and S. Sagayama, "Polyphonic pitch estimation and instrument identification by joint modeling of sustained and attack sounds," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1124–1132, 2011.

[23] W.-M. Szeto and K.-H. Wong, "A hierarchical bayesian framework for score-informed source separation of piano music signals," in *ISMIR*, 2015, pp. 155–161.

[24] Miguel A Alonso, Gaël Richard, and Bertrand David, "Tempo and beat estimation of musical signals.," in *ISMIR*, 2004.

[25] L. Su and Y.-H. Yang, *Power-scaled spectral flux and peak-valley group-delay methods for robust musical onset detection*, Ann Arbor, MI: Michigan Publishing, University of Michigan Library, 2014.

[26] S. Böck and M. Schedl, "Polyphonic piano note transcription with recurrent neural networks," in *ICASSP*, 2012, pp. 121–124.

[27] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 1–12, 2007.