Piano Transcription With Convolutional Sparse Lateral Inhibition

Andrea Cogliati, *Student Member, IEEE*, Zhiyao Duan, *Member, IEEE*, and Brendt Wohlberg, *Senior Member, IEEE*

***Abstract*—This letter extends our prior work on context-dependent piano transcription to estimate the length of the notes in addition to their pitch and onset. This approach employs convolutional sparse coding along with lateral inhibition constraints to approximate a musical signal as the sum of piano note waveforms (dictionary elements) convolved with their temporal activations. The waveforms are pre-recorded for the specific piano to be transcribed in the specific environment. A dictionary containing multiple waveforms per pitch is generated by truncating a long waveform for each pitch to different lengths. During transcription, the dictionary elements are fixed and their temporal activations are estimated and postprocessed to obtain the pitch, onset, and note length estimation. A sparsity penalty promotes globally sparse activations of the dictionary elements, and a lateral inhibition term penalizes concurrent activations of different waveforms corresponding to the same pitch within a temporal neighborhood, to achieve note length estimation. Experiments on the MIDI aligned piano sounds dataset show that the proposed approach significantly outperforms a state-of-the-art music transcription method trained in the same context-dependent setting in transcription accuracy.**

***Index Terms*—Automatic music transcription (AMT), convolutional sparse coding (CSC), lateral inhibition, offset detection.**

I. Introduction

Automatic music transcription (AMT) is the process of inferring a symbolic representation from an audio signal [1]. It has applications in music education (e.g., providing feedback to a piano learner), content-based music search (e.g., searching songs with a similar bassline), musicological analysis of nonnotated music (e.g., Jazz improvisations), and music enjoyment (e.g., visualizing the music content).

Pitch, onset, and offset (or, equivalently, note length), are the three main basic parameters of a musical note. AMT systems that aim to achieve note-level transcription must estimate these parameters. Most existing research has focused on pitch and onset detection, while considerably less attention has been devoted to offset detection [1]. However, for many applications, especially those requiring music notation transcription [2], relatively accurate note length estimation is essential.

Many note-level music transcription methods are *frame based*, i.e., they attempt to identify pitches in each time frame, then determine note onsets and offsets through postprocessing [1]. The most popular approaches in this category are based on spectrogram decomposition, and use either nonnegative matrix factorization (NMF) [3], [4] or probabilistic latent component analysis (PLCA) [5], which are numerically equivalent. To obtain note-level transcription results, a postprocessing step, such as a median filter or a hidden Markov model (HMM), is required to estimate note onsets and offsets from frame-level pitch estimates [6]. Other frame-based methods include deep neural networks [7]–[10], and probabilistic methods, such as [11]–[15].

In contrast to frame-based methods, *note-based* methods attempt to directly identify full notes. Piano notes are characterized by significant temporal evolution, in both the waveform and the spectral content. In particular, different partials decay at different rates, i.e., higher frequency partials decay faster than lower frequency ones [16]–[18]. Grindlay and Ellis [19] proposed a generalization of PLCA to account for the temporal evolution of each note. Mysore *et al.* [20] introduced a variant of NMF called nonnegative factorial HMM (N-FHMM) to learn multiple spectral templates for each note and a Markov chain describing the temporal transition between them. Ewert *et al.* [21] recently proposed a dynamic programming variation of N-FHMM to reduce its high computational cost. This method has been extended and adapted to piano music transcription by Cheng *et al.* [22]. Nonnegative matrix de-convolution (NMD) as introduced in [23] is capable of modeling the temporal evolution of nonstationary sounds. All these methods are capable of estimating the note length, but they are generally evaluated on onset-only estimation [24]. Even the most recent MIREX contest shows that most algorithms cannot achieve good results in both onset detection and length estimation (see MIREX 2016 [25]).

In [26] and [27], we proposed a time-domain approach, which we will refer to as CDW-15 in the following, to address piano music transcription in a context-dependent setting. CDW-15 approximates the music signal $\boldsymbol{s}$ with the summation of note waveforms $\{\boldsymbol{d}_m\}$ convolved with their temporal activation coefficients $\{\mathbf{x}_m\}$:

$$\underset{\{\mathbf{x}_m\}}{\arg\min}\ \frac{1}{2}\left\|\sum_m \mathbf{d}_m * \mathbf{x}_m - \mathbf{s}\right\|_2^2 + \lambda \sum_m \|\mathbf{x}_m\|_1 \tag{1}$$

where λ is a regularization constant. The waveform $\boldsymbol{d}_m$ of each individual pitch m is pre-recorded on the same piano in the same environment as the music signal to be transcribed, and its length is always truncated to 1s. The ℓ_1 regularization term encourages sparse activations of notes, higher values of λ result in sparser activations. This approach achieves higher accuracy

Manuscript received November 10, 2016; revised January 12, 2017; accepted January 27, 2017. Date of publication February 8, 2017; date of current version March 1, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Balazs Bank. *(Corresponding author: Andrea Cogliati.)*

A. Cogliati and Z. Duan are with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627, USA (e-mail: andrea.cogliati@rochester.edu; zhiyao.duan@rochester.edu).

B. Wohlberg is with the Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA (e-mail: brendt@lanl.gov).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2017.2666183

in pitch and onset estimation than the state-of-the-art [6], but it does not estimate note offsets.

In this letter, we extend CDW-15 to estimate the note length by using a dictionary containing multiple atoms with different lengths per pitch, thus creating *pitch groups* of atoms corresponding to the same pitch. When using multiple atoms per pitch, we need to avoid concurrent activations of multiple atoms in the same pitch group. In order to achieve this result, we impose a lateral inhibition [28] regularization term on the activation coefficients of atoms in the same pitch group, in addition to the ℓ_1 regularization on all atoms. The lateral inhibition regularization prevents concurrent activation of multiple atoms in the same pitch group within a temporal neighborhood. We can call this property *within-group sparsity*.

II. Structured Sparsity

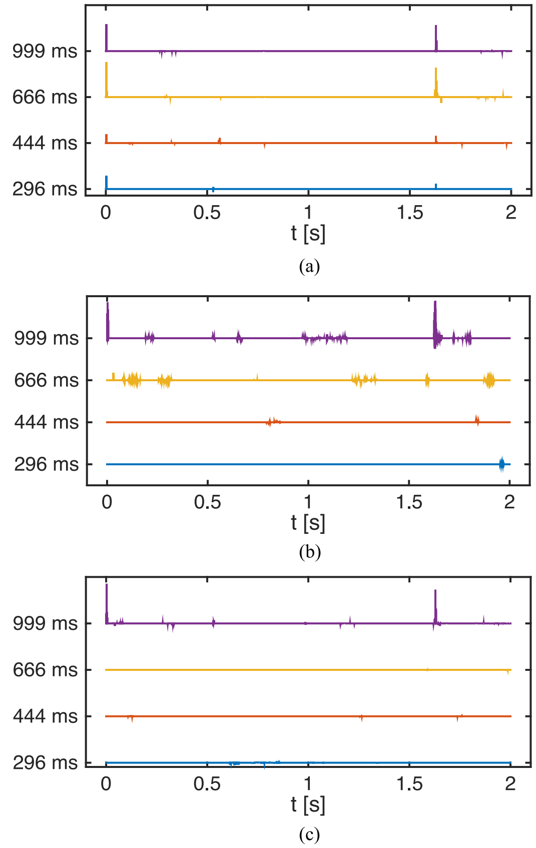
Standard sparsity assumes a representation that has only a few nonzero coefficients, but makes no additional assumptions on how these nonzero coefficients are distributed within the coefficient vector or matrix. Structured sparsity, in contrast, is based on the assumption that there is some sort of identifiable structure to the distribution of these coefficients. This structure can take many forms, the most common being group sparsity and joint sparsity [29]. The former requires the assignment of dictionary atoms to distinct groups, and assumes that only a few groups are active, but does not require sparse activations within each group. The latter is defined within a multiple measurement vector context [30], and assumes that the representations of different signal vectors share the same or similar pattern of activations. Both of these types of structure can be promoted by the use of the $\ell_{2,1}$ norm [29].

Structured sparsity has previously been applied to AMT. For example, in an NMF framework, a dictionary with multiple atoms per pitch can be learned, in which each atom in the same group represents a different frame of a long note of a particular pitch. Group sparsity can be introduced to promote multiple atoms in the same group to be activated contiguously, i.e., one after the other. An example of such structured sparsity was introduced by O'Hanlon *et al.* [31], who used a modified nonnegative basis-pursuit greedy approach. Another example of group sparsity in an NMF framework was proposed by O'Hanlon and Plumbley [32] to promote the co-activation of harmonically related narrow-band atoms. In this case, each group still represents a single pitch, but each pitch is sliced harmonically, not temporally as in the previous method.

In this letter, we are interested in limiting the number of concurrently active atoms inside each group, as each atom represents a full note. We call this property within-group sparsity. However, this property alone is not sufficient to achieve a clean activation matrix and, thus, a good transcription. In order to obtain a good transcription, global sparsity on the activations must also be promoted.

III. Proposed Method

The key idea of the proposed method is to jointly estimate pitch, onset, and duration of notes by using a dictionary containing multiple atoms with different length for each pitch in the convolutional sparse coding (CSC) framework of (1). To create the dictionary we truncate the 1-s long template trained as in CDW-15 [27] to different lengths. This approach is easier and faster than sampling the same pitch played with different durations. However, expanding the dictionary does not lead to satisfying results, as multiple templates in the same pitch group are activated concurrently, as we can see in Fig. 1(a) at the beginning of the signal and slightly after $t = 1.5$ s. The reason is that the ℓ_1 regularization in (1) only promotes sparse activations of all templates across all times, but does not distinguish activations of templates in the same pitch group from activations of templates in different pitch groups; moreover, it does not distinguish activations that are temporally close from activations that are temporally apart. While it is possible for a player to play different notes in a rapid pace, it is unlikely to play the same note repeatedly too quickly [27]. Therefore, we need a regularization term that distinguishes these activations and penalizes close activations of templates in the same pitch group.



(a)
(b)
(c)

Fig. 1. Activations of the atoms for pitch D5 for the opening of Bach's Minuet in G. D5 should be activated twice, at $t = 0$ and $t \simeq 1.6$ s. The numbers on the vertical axis indicate the length of each template. (a) ℓ_1 regularization only. (b) Lateral inhibition regularization only. (c) Combined lateral inhibition and ℓ_1 regularizations.

We propose to use a lateral inhibition [28] regularization term on the activations of templates in the same pitch group within a temporal window. The cost of activating atom m at time t is given by

$$\Gamma(\{\mathbf{x}_m\}) = |\boldsymbol{x}_m(t)| \left[\left(\sum_{\substack{n \in G(m) \\ |t-\tau|<T}} |\boldsymbol{x}_n(\tau)| \right) - |\boldsymbol{x}_m(t)| \right] \qquad (2)$$

where $G(m)$ is the pitch group to which atom m belongs, and T is the length of the temporal window of inhibition. The activation of atom m at time t will inhibit the activation of all the other atoms in the same pitch group within the temporal window around t. The term $|\boldsymbol{x}_m(t)|$ needs to be subtracted from the summation to avoid self-inhibition.

The full regularization term is the summation of all the costs over all atoms and all time instants, multiplied by a constant μ. The objective function becomes

$$\underset{\{\boldsymbol{x}_m\}}{\arg\min} \frac{1}{2} \left\| \sum_m \boldsymbol{d}_m * \boldsymbol{x}_m - \boldsymbol{s} \right\|_2^2 + \mu \sum_m \Gamma(\{\mathbf{x}_m\}). \quad (3)$$

As we can see in Fig. 1(b), this objective function minimizes the concurrent activations of atoms in the same pitch group and inside the inhibition time window (50 ms), but the activations are not globally sparse over time. Moreover, not shown in the figures, the activations of other groups are also nonsparse. Global sparsity is a key component of CDW-15, and has been successfully applied to AMT for a long time [1]. In order to promote global sparsity on all activations of all templates, we added a global ℓ_1 norm to the basic lateral inhibition model in (10). The objective function with both the global ℓ_1-norm regularization and lateral inhibition regularization is

$$\underset{\{\boldsymbol{x}_m\}}{\arg\min} \frac{1}{2} \left\| \sum_m \boldsymbol{d}_m * \boldsymbol{x}_m - \boldsymbol{s} \right\|_2^2 + \lambda \sum_m \|\boldsymbol{x}_m\|_1 + \mu \sum_m \Gamma(\{\mathbf{x}_m\}). \quad (4)$$

Using this regularization, as we can see from Fig. 1(c), the activation vectors are now sparser and less noisy, and also globally sparse, as we will show in the experimental section.

IV. Algorithm

The simplest form of lateral inhibition structured sparse coding problem [28] is

$$\frac{1}{2}\|D\mathbf{x} - \mathbf{s}\|_2^2 + |\mathbf{x}|^T \Omega |\mathbf{x}| \quad (5)$$

where D is a dictionary matrix, and Ω is a matrix encoding the pattern of desired mutual inhibitions. As was pointed out in [28], if the entries of Ω are nonnegative, we can define $\mathbf{w} = |\mathbf{x}|^T \Omega$, and write (5) as a weighted basis pursuit denoising (BPDN) problem

$$\frac{1}{2}\|D\mathbf{x} - \mathbf{s}\|_2^2 + \|\mathbf{w} \odot \mathbf{x}\|_1 \quad (6)$$

where $\odot$ is the Hadamard product, allowing the problem to be tackled by modifying a standard algorithm for the BPDN problem to include iteratively updating the weight vector $\mathbf{w}$, which depends on the solution variable $\mathbf{x}$. Szlam *et al.* reported [28] that good performance was obtained with a fast iterative shrinkage-thresholding algorithm (FISTA) algorithm. They also proposed a convolutional form of (5), but applied it to a sufficiently small $\mathbf{s}$ to make it feasible to retain an explicit weighting matrix Ω in the formulation.

Our innovation with respect to the algorithm is twofold. First, since we wish to apply the model to a signal $\mathbf{s}$ that is far too large for an explicit weighting matrix Ω to be practical, we have modified the regularization term so that the lateral inhibition is specified by the product of a convolution filter determining the inhibition in time, and a small matrix that determines the inhibition within and between groups of dictionary atoms. Second, since alternating direction method of multipliers (ADMM) has been shown to be more effective than FISTA for the convolutional BPDN (CBPDN) problem [33], we modify the ADMM algorithm proposed in [34] to include the necessary iterative reweighting. We found experimentally that good results were obtained by updating the new weight vector $\mathbf{w}$ from the primary variable $\mathbf{x}$ rather than from the auxiliary variable introduced in the variable splitting, and by smoothing this weight vector update by defining it as a convex linear combination of the previous and new values.

The lateral inhibition regularization terms in (3) and (4) are rewritten in terms of convolution as

$$\Gamma(\{\mathbf{x}_m\}) = \sum_m \sum_n c_{m,n} (|\boldsymbol{x}_n| * \boldsymbol{h})^T |\boldsymbol{x}_m| \quad (7)$$

where $\boldsymbol{h}$ is the time inhibition window, which is equal to 1 around the origin within a radius of $T/2$, and $c_{m,n}$ is defined as

$$c_{m,n} = \begin{cases} 1 & \text{if } m \neq n \text{ and } G(m) = G(n) \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

If we define

$$\boldsymbol{\omega}_m^T = \sum_n c_{m,n} \left(|\boldsymbol{x}_n| * \boldsymbol{h}\right)^T \quad (9)$$

then (3) can be rewritten as

$$\underset{\{\boldsymbol{x}_m\}}{\arg\min} \frac{1}{2} \left\| \sum_m \boldsymbol{d}_m * \boldsymbol{x}_m - \boldsymbol{s} \right\|_2^2 + \mu \sum_m \boldsymbol{\omega}_m^T |\boldsymbol{x}_m| \quad (10)$$

which immediately shows that the regularization is a weighted ℓ_1-norm on $\boldsymbol{x}_m$. Similarly, (4) can be written as

$$\underset{\{\boldsymbol{x}_m\}}{\arg\min} \frac{1}{2} \left\| \sum_m \boldsymbol{d}_m * \boldsymbol{x}_m - \boldsymbol{s} \right\|_2^2 + \lambda \sum_m \|\boldsymbol{x}_m\|_1 + \mu \sum_m \boldsymbol{\omega}_m^T |\boldsymbol{x}_m| \,. \quad (11)$$

Finally, the two regularization terms can be combined into a single term as

$$\underset{\{\boldsymbol{x}_m\}}{\arg\min} \frac{1}{2} \left\| \sum_m \boldsymbol{d}_m * \boldsymbol{x}_m - \boldsymbol{s} \right\|_2^2 + \sum_m \left(\lambda \mathbf{1} + \mu \boldsymbol{\omega}_m^T\right) |\boldsymbol{x}_m| \,, \quad (12)$$

where $\mathbf{1}$ is a row vector comprised of all ones.

The resulting ADMM algorithm[1] is very similar to the efficient ADMM algorithm for the CBPDN problem [33], except for the use of a weighted ℓ_1 norm, which requires a minor modification to the soft-thresholding step [36], and in the need for recomputing the weight vector at every iteration, as described above.

[1] An implementation will be included in a future release of the SPORCO library [35].

The raw activation vectors thus obtained must be postprocessed to detect peaks, which correspond to note onsets. This step is a refinement of the method described in [27], generalized to the extended dictionary. We start by setting all the activations below a global threshold, currently set at 10% of the maximum value across the activation matrix $\boldsymbol{X}$, to 0. Then we determine all the local peaks in each activation vector. Finally, we iterate over all the peaks, in order of magnitude starting from the largest one, and we set to 0 all the activations in the same pitch group and inside the inhibition window, currently set at 50 ms.

The complexity of the algorithm is dominated by the calculation of the cost vectors $\boldsymbol{\omega}_m$ and is $\mathcal{O}(M^2 N \log N)$, where M is the number of atoms and N is the length of the signal $\boldsymbol{s}$.

V. Experiment

We applied the different models described in Section III to the first 10 s of the 30 pieces in the ENSTDkCl dataset of MIDI Aligned Piano Sounds (MAPS) [13]. The limit of 10 s was determined by the amount of graphics processing unit (GPU) memory required by the current MATLAB implementation of the algorithm, however, a longer piece could be transcribed by segmenting it into 10 s long chunks, as described in our previous paper [27]. We used a value of $\lambda = 0.05$ and $\mu = 0.5$. These values were empirically tuned on a single piece and then fixed for the entire dataset. For each piece we calculated precision, recall, and F-measure with both onset-only and onset-offset criteria [24], with the standard MIREX parameters: Onset tolerance of 50 ms and offset tolerance of 20% of the correct note length or 50 ms, whichever is longer. The lengths of the different atoms in the dictionary for each pitch were chosen to approximate the distribution of note lengths in MAPS, i.e., higher density for shorter notes around 100 ms and lower density for longer notes; we also spaced the durations exponentially in order to maximize the likelihood of estimating the correct length according to the onset–offset criterion. The durations were: 39, 58, 88, 132, 197, 297, 444, 666, and 999 ms. We also calculated the average overlap ratio (AOR) [24]. AOR gives a measure of how much a correctly returned note overlaps with the corresponding ground-truth note. We compared the proposed method with several baseline methods: CDW-15, with note lengths fixed at 100 ms; BW-13, a state-of-the-art frame-based method based on PLCA proposed by Benetos and Weyde [37]; SONIC, a piano music transcription system based on neural networks [38]; DT-14, a generic music transcription system based on maximum likelihood by Duan and Temperley [15]; and VBB-10, an NMF-based transcription system by Vincent *et al.* [39]. For all the baseline methods we used the original authors' implementation. BW-13 was also trained in the same context of the proposed method on the isolated notes in the ENSTDkCl dataset of MAPS. It must be noted that SONIC, VBB-10, and DT-14 cannot be trained in the same context, so the comparison is biased against these methods.

TABLE I
AVERAGE RESULTS ON THE FIRST 10 S OF THE 30 PIECES IN THE ENSTDKCL DATASET OF MAPS (HIGHER VALUES ARE BETTER)

	Onset only				Onset–offset			
Method	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	AOR	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	AOR
BW-13	64.1	59.8	61.2	55.8	19.2	18.2	18.5	81.4
SONIC	78.0	72.0	74.5	**58.7**	**28.5**	25.7	**26.9**	83.4
DT-14	55.2	34.7	41.3	51.7	15.3	9.4	11.3	82.0
VBB-10	52.5	75.7	60.9	38.8	11.5	15.0	12.8	63.6
CDW-15	**79.7**	83.5	**80.8**	40.1	17.8	18.0	17.8	68.8
ℓ_1	55.4	**88.7**	65.4	54.6	16.8	25.8	19.5	84.4
LI	42.2	83.7	53.3	55.9	12.5	**27.3**	16.3	**84.8**
ℓ_1 + LI	77.7	79.6	77.5	54.6	22.3	23.0	22.3	84.5

Bold font indicates the best value in each column.

The average results for the entire dataset are shown in Table I. We can observe that almost all variants of the CSC-based methods, except LI (Lateral Inhibition), outperform BW-13, VBB-10 and DT-14 on F-measure for the onset-only criterion; CDW-15 and ℓ_1+LI also outperform SONIC, showing the advantage of the time-domain approach over frequency-domain methods in this setting. Moreover, ℓ_1+LI significantly outperforms both LI and ℓ_1 on F-measure. This supports our analysis that both within-group and global sparsity are needed. From CDW-15 to ℓ_1 F-measure drops significantly for the onset-only criterion but increases slightly for the onset–offset criterion. The only difference between these two methods is that ℓ_1 uses nine templates per pitch while CDW-15 uses only one template. As noted, multiple templates can be activated simultaneously in ℓ_1 resulting in a lower precision but higher recall, and when onset-offset criterion is used, the improvement on recall dominates the decrease on precision. Similarly, from CDW-15 to LI, precision drops significantly, while recall increases slightly under the onset-only criterion and significantly under the onset–offset criterion. However, the drop of precision is due to the false activation of wrong notes instead of the false activation of multiple templates of the correct note. Finally, when onset-only criterion is used, LI+ℓ_1 slightly under-performs CDW-15 on F-measure, but significantly outperforms CDW-15 on AOR; when onset–offset criterion is used, LI+ℓ_1 falls behind SONIC on F-measure but significantly outperforms CDW-15 on both F-measure and AOR. Overall, the proposed method with both lateral inhibition and global sparsity regularization brings the CSC-based approach to the highest level of performance.

VI. Conclusion

In this letter, we extended our prior work on CSC for time-domain piano transcription in a context-dependent setting. The proposed method uses multiple templates with different lengths per pitch to achieve note length estimation. Lateral inhibition regularization is introduced to ensure that at most one template per pitch is activated within an inhibition window. Global sparsity is achieved through ℓ_1 regularization to reduce false activations of wrong notes. Experiments show that the proposed method significantly outperforms our prior work and another state-of-the-art frequency-domain method trained in the same context.

Acknowledgment

The authors would like to thank LighTopTech Corporation for kindly providing the GPU-based system to run the experiments.

REFERENCES

[1] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: Challenges and future directions," *J. Intell. Inf. Syst.*, vol. 41, no. 3, pp. 407–434, 2013.

[2] A. Cogliati, D. Temperley, and Z. Duan, "Transcribing human piano performances into music notation," in *Proc. Int. Soc. Music Inf. Retrieval*, 2016, pp. 758–764.

[3] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[4] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 556–562.

[5] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," *in Proc. Workshop Adv. Models Acoust. Process.*, 2006.

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

[7] J. Nam, J. Ngiam, H. Lee, and M. Slaney, "A classification-based polyphonic piano transcription approach using learned feature representations," in *Proc. Int. Soc. Music Inf. Retrieval*, 2011, pp. 175–180.

[8] S. Böck and M. Schedl, "Polyphonic piano note transcription with recurrent neural networks," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, 2012, pp. 121–124.

[9] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription," in *Proc. 29th Int. Conf. Mach. Learn.*, Edinburgh, Scotland, U.K., 2012, pp. 1159–1166.

[10] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 5, pp. 927–939, May 2016.

[11] M. Goto, "A real-time music-scene-description system: Predominant-f0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Commun.*, vol. 43, no. 4, pp. 311–329, 2004.

[12] Z. Duan, B. Pardo, and C. Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 8, pp. 2121–2133, Nov. 2010.

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

[14] P. H. Peeling and S. J. Godsill, "Multiple pitch estimation using non-homogeneous Poisson processes," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1133–1143, Oct. 2011.

[15] Z. Duan and D. Temperley, "Note-level music transcription by maximum likelihood sampling," in *Proc. Int. Soc. Music Inf. Retrieval*, 2014, pp. 181–186.

[16] A. Cogliati and Z. Duan, "Piano music transcription modeling note temporal evolution," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Brisbane, Qld, Australia, 2015, pp. 429–433.

[17] M. Campbell and C. Greated, *The Musician's Guide to Acoustics*. London, U.K.: Oxford Univ. Press, 1994.

[18] T. Cheng, S. Dixon, and M. Mauch, "Modelling the decay of piano sounds," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Brisbane, Qld, Australia, 2015, pp. 594–598.

[19] G. C. Grindlay and D. P. W. Ellis, "Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1159–1169, Oct. 2011.

[20] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden Markov modeling of audio with application to source separation," in *Proc. Latent Variable Anal. Signal Separation*, 2010, pp. 140–148.

[21] S. Ewert, M. D. Plumbley, and M. Sandler, "A dynamic programming variant of non-negative matrix deconvolution for the transcription of struck string instruments," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Brisbane, Qld, Australia, 2015, pp. 569–573.

[22] T. Cheng, M. Mauch, E. Benetos, and S. Dixon, "An attack/decay model for piano transcription," in *Proc. Int. Soc. Music Inf. Retrieval*, 2016, pp. 584–590.

[23] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Proc. Ind. Compon. Anal. Blind Signal Separation*, 2004, pp. 494–499.

[24] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of multiple-f0 estimation and tracking systems," in *Proc. Int. Soc. Music Inf. Retrieval*, 2009, pp. 315–320.

[25] MIREX2016 results. 2016. [Online]. Available: http://www.music-ir.org/mirex/wiki/2016:Multiple_Fundamental_Frequency_Estimation_%26_Tracking_Results_-_MIREX_Dataset

[26] A. Cogliati, Z. Duan, and B. Wohlberg, "Piano music transcription with fast convolutional sparse coding," in *Proc. IEEE 25th Int. Workshop Mach. Learn. Signal Process*, 2015, pp. 1–6.

[27] A. Cogliati, Z. Duan, and B. Wohlberg, "Context-dependent piano music transcription with convolutional sparse coding," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 12, pp. 2218–2230, Dec. 2016.

[28] A. Szlam, K. Gregor, and Y. LeCun, "Structured sparse coding via lateral inhibition," in *Proc. 24th Int. Conf. Adv. Neural Inf.*, 2011, pp. 1116–1124.

[29] P. Sprechmann, I. Ramírez, G. Sapiro, and Y. C. Eldar, "C-HiLasso: A collaborative hierarchical sparse modeling framework," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4183–4198, Sep. 2011.

[30] E. van den Berg and M. P. Friedlander, "Joint-sparse recovery from multiple measurements," Dept. Comput. Sci., Univ.British Columbia, Vancouver, BC, Canada, *Tech. Rep. TR-2009–07*, Apr. 2009.

[31] K. O'Hanlon, H. Nagano, and M. D. Plumbley, "Structured sparsity for automatic music transcription," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2012, pp. 441–444.

[32] K. O'Hanlon and M. D. Plumbley, "Polyphonic piano transcription using non-negative matrix factorisation with group sparsity," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2014, pp. 3112–3116.

[33] B. Wohlberg, "Efficient algorithms for convolutional sparse representations," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 301–315, Jan. 2016.

[34] B. Wohlberg, "Efficient convolutional sparse coding," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Florence, Italy, 2014, pp. 7173–7177.

[35] B. Wohlberg, "SParse Optimization Research COde (SPORCO)," 2016. Software library available from http://purl.org/brendt/software/sporco

[36] B. Wohlberg, "Convolutional sparse representations as an image model for impulse noise restoration," in *Proc. IEEE Image, Video, Multidimensional Signal Process. Workshop*, Bordeaux, France, Jul. 2016, pp. 1–5.

[37] E. Benetos and T. Weyde, "An efficient temporally-constrained probabilistic model for multiple-instrument music transcription," in *Proc. Int. Soc. Music Inf. Retrieval*, 2015, pp. 701–707.

[38] M. Marolt, "SONIC: Transcription of polyphonic piano music with neural networks," in *Proc. Workshop Current Direction Comput. Music Res.*, Audiovisual Inst., Pompeu Fabra Univ., 2001, pp. 217–224.

[39] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 3, pp. 528–537, Mar. 2010. [Online]. Available: https://hal.inria.fr/inria-00544094