Context-Dependent Piano Music Transcription With Convolutional Sparse Coding

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

***Abstract*—This paper presents a novel approach to automatic transcription of piano music in a context-dependent setting. This approach employs convolutional sparse coding to approximate the music waveform as the summation of piano note waveforms (dictionary elements) convolved with their temporal activations (onset transcription). The piano note waveforms are pre-recorded for the specific piano to be transcribed in the specific environment. During transcription, the note waveforms are fixed and their temporal activations are estimated and post-processed to obtain the pitch and onset transcription. This approach works in the time domain, models temporal evolution of piano notes, and estimates pitches and onsets simultaneously in the same framework. Experiments show that it significantly outperforms a state-of-the-art music transcription method trained in the same context-dependent setting, in both transcription accuracy and time precision, in various scenarios including synthetic, anechoic, noisy, and reverberant environments.**

***Index Terms*—Automatic music transcription, convolutional sparse coding, piano transcription, reverberation.**

I. Introduction

AUTOMATIC music transcription (AMT) is the process of automatically inferring a high-level symbolic representation, such as music notation or piano-roll, from a music performance [1]. It has several 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 non-notated music (e.g., Jazz improvisations and most non-Western music), and music enjoyment (e.g., visualizing the music content).

Music transcription of polyphonic music is a challenging task even for humans. It is related to *ear training*, a required course for professional musicians on identifying pitches, intervals, chords, melodies, rhythms, and instruments of music solely by hearing. AMT for polyphonic music was first proposed in 1977 by Moorer [2], and Piszczalski and Galler [3]. Despite almost four decades of active research, it is still an open problem and current AMT systems cannot match human performance in either accuracy or robustness [1].

A core problem of music transcription is figuring out *which* notes are played and *when* they are played in a piece of music. This is also called *note-level transcription* [4]. A note produced by a pitched musical instrument has five basic attributes: pitch, onset, offset, timbre and dynamic. Pitch is a perceptual attribute but can be reliably related to the fundamental frequency (F0) of a harmonic or quasi-harmonic sound [5]. Onset refers to the beginning time of a note, in which the amplitude of that note instance increases from zero to an audible level. This increase is very sharp for percussive pitched instruments such as piano. Offset refers to the ending time of a note, i.e., when the waveform of the note vanishes. Compared to pitch and onset, offset is often ambiguous [4]. Timbre is the quality of a sound that allows listeners to distinguish two sounds of the same pitch and loudness [5]. Dynamic refers to the player's control over the loudness of the sound; e.g., a piano player can strike a key with different forces, causing notes to be soft or loud. The dynamic can also change the timbre of a note; e.g., on a piano, notes played *forte* have a richer spectral content than notes played *piano*[6]. In this paper we focus on *pitch estimation* and *onset detection* of notes from polyphonic piano performances.

In the literature, these two problems are often addressed separately and then combined to achieve note-level transcription (see Section II). For onset detection, commonly used methods are based on spectral energy changes in successive frames [7]. They do not model the harmonic relation of frequencies that exhibit this change, nor the temporal evolution of partial energy of notes. Therefore, they tend to miss onsets of soft notes in polyphonic pieces and to detect false positives due to local partial amplitude fluctuations caused by overlapping harmonics, reverberation or beats [8].

Pitch estimation in monophonic music is considered a solved problem [9]. In contrast, polyphonic pitch estimation is much more challenging because of the complex interaction (e.g., the overlapping harmonics) of multiple simultaneous notes. To properly identify all the concurrent pitches, the partials of the mixture must be separated and grouped into clusters belonging to different notes. Most multi-pitch analysis methods operate in the frequency domain with a time-frequency magnitude representation [1]. This approach has two fundamental limitations: it introduces the time-frequency resolution trade-off due to the Gabor limit [10], and it discards the phase, which contains useful cues for the harmonic fusing of partials [5]. Current state-of-the-art results are below 70% in F-measure, which is too low for practical purposes, as evaluated by MIREX 2015 on orchestral pieces with up to 5 instruments and piano pieces [11].

In this paper, we propose a novel time-domain approach to transcribe polyphonic piano performances at the note-level. More specifically, we model the piano audio waveform as a

Manuscript received December 21, 2015; revised April 28, 2016 and July 8, 2016; accepted July 31, 2016. Date of publication August 4, 2016; date of current version September 19, 2016. This research was supported by the U.S. Department of Energy through the LANL/LDRD Program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hiroshi Saruwatari.

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 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 paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2598305

convolution of note waveforms (i.e., dictionary templates) and their activation weights (i.e., transcription of note onsets). We pre-learn the dictionary by recording the audio waveform of each note of the piano, and then employ a recently proposed efficient convolutional sparse coding algorithm to estimate the activations. Compared to current state-of-the-art AMT approaches, the proposed method has the following advantages:

1) The transcription is performed in the time domain and avoids the time-frequency resolution trade-off by imposing structural constraints on the analyzed signal – i.e., a context specific dictionary and sparsity on the atom activations – resulting in better performance, especially for low-pitched notes;
2) It models temporal evolution of piano notes and estimates pitch and onset simultaneously in the same framework;
3) It achieves much higher transcription accuracy and time precision compared to a state-of-the-art AMT approach;
4) It works in reverberant environments and is robust to stationary noise to a certain degree.

One important limitation of the proposed approach is that it only works in a context-dependent setting, i.e., the dictionary needs to be trained for each specific piano and acoustic environment. While transcription of professionally recorded performances is not possible, as the training data is not generally available, the method is still useful for musicians, both professionals and amateurs, to transcribe their performances with much higher accuracy than state-of-the-art approaches. In fact, the training process takes less than 3 minutes to record all 88 notes of a piano (each played for about 1 second). In most scenarios, such as piano practices at home or in a studio, the acoustic environment of the piano does not change, i.e., the piano is not moved and the recording device, such as a smartphone, can be placed in the same spot, and the trained dictionary can be re-used. Even for a piano concert in a new acoustic environment, taking 3 minutes to train the dictionary in addition to stage setup is acceptable for highly accurate transcription of the performance throughout the concert.

A preliminary version of the proposed approach has been presented in [12]. In this paper, we describe this approach in more detail, conduct systematic experiments to evaluate its key parameters, and show its superior performance against a state-of-the-art method in various conditions. The rest of the paper is structured as follows: Section II reviews note-level AMT approaches and puts the proposed approach in context. Section III reviews the basics of convolutional sparse coding and its efficient implementation. Section IV describes the proposed approach and Section V conducts experiments. Finally, Section VI concludes the paper.

II. Related Work

There are in general three approaches to note-level music transcription. *Frame-based* approaches estimate pitches in each individual time frame and then form notes in a post-processing stage. *Onset-based* approaches first detect onsets and then estimate pitches within each inter-onset interval. *Note-based* approaches estimate notes including pitches and onsets directly. The proposed method uses the third approach. In the following, we will review methods of all these approaches and discuss their advantages and limitations.

A. Frame-Based Approach

Frame-level multi-pitch estimation (MPE) is the key component of this approach. The majority of recently proposed MPE methods operate in the frequency domain. One group of methods analyze or classify features extracted from the time-frequency representation of the audio input [1]. Raphael [13] used a Hidden Markov Model (HMM) in which the states represent pitch combinations and the observations are spectral features, such as energy, spectral flux, and mean and variance of each frequency band. Klapuri [14] used an iterative spectral subtraction approach to estimate a predominant pitch and subtract its harmonics from the mixture in each iteration. Yeh *et al.* [15] jointly estimated pitches based on three physical principles – harmonicity, spectral smoothness and synchronous amplitude evolution. More recently, Dressler [16] used a multi-resolution Short Time Fourier Transform (STFT) in which the magnitude of each bin is weighted by the bin's instantaneous frequency. The pitch estimation is done by detecting peaks in the weighted spectrum and scoring them by harmonicity, spectral smoothness, presence of intermediate peaks and harmonic number. Poliner and Ellis [17] used Support Vector Machines (SVM) to classify the presence of pitches from the audio spectrum. Pertusa and Iñesta [18] identified pitch candidates from spectral analysis of each frame, then selected the best combinations by applying a set of rules based on harmonic amplitudes and spectral smoothness. Saito *et al.*[19] applied a specmurt analysis by assuming a common harmonic structure of all the pitches in each frame. Finally, methods based on deep neural networks are beginning to appear [20]–[23].

Another group of MPE methods are based on statistical frameworks. Goto [24] viewed the mixture spectrum as a probability distribution and modeled it with a mixture of tied-Gaussian mixture models. Duan *et al.* [25] and Emiya *et al.* [26] proposed Maximum-Likelihood (ML) approaches to model spectral peaks and non-peak regions of the spectrum. Peeling and Godsill [27] used non-homogenous Poisson processes to model the number of partials in the spectrum.

A popular group of MPE methods in recent years are based on *spectrogram factorization* techniques, such as Non-negative Matrix Factorization (NMF) [28] or Probabilistic Latent Component Analysis (PLCA) [29]; the two methods are mathematically equivalent when the approximation is measured by Kullback-Leibler (KL) divergence. The first application of spectrogram factorization techniques to AMT was performed by Smaragdis and Brown [30]. Since then, many extensions and improvements have been proposed. Grindlay *et al.* [31] used the notion of *eigeninstruments* to model spectral templates as a linear combination of basic instrument models. Benetos *et al.* [32] extended PLCA by incorporating shifting across log-frequency to account for vibrato, i.e., frequency modulation. Abdallah *et al.* [33] imposed sparsity on the activation weights. O'Hanlon *et al.* [34], [35] used structured sparsity, also called group sparsity, to enforce harmonicity of the spectral bases.

Time domain methods are far less common than frequency domain methods for multi-pitch estimation. Early AMT methods operating in the time domain attempted to simulate the human auditory system with bandpass filters and autocorrelations [36], [37]. More recently, other researchers proposed time-domain probabilistic approaches based on Bayesian models [38]–[40]. Bello *et al.* [41] proposed a hybrid approach exploiting both frequency and time-domain information. More recently, Su and Yang [42] also combined information from spectral (harmonic series) and temporal (subharmonic series) representations.

The closest work in the literature to our approach was proposed by Plumbley *et al.* [43]. In that paper, the authors proposed and compared two approaches for sparse decomposition of polyphonic music, one in the time domain and the other in the frequency domain. The time domain approach adopted a similar shift-invariant (i.e., convolutional) sparse coding formulation to ours. However, they used an unsupervised approach and a complete transcription system was not demonstrated due to the necessity of manual annotation of atoms. The correct number of individual pitches in the piece was also required in their approach. In addition, the sparse coding was performed in 256-ms long windows using 128-ms long atoms, thus not modeling the temporal evolution of notes. As we will show in Section V-A, this length is not sufficient to achieve good accuracy in transcription. Furthermore, the system was only evaluated on very short music excerpts, possibly because of the high computational requirements.

To obtain a note-level transcription from frame-level pitch estimates, a post-processing step, such as a median filter [42] or an HMM [44], is often employed to connect pitch estimates across frames into notes and remove isolated spurious pitches. These operations are performed on each note independently. To consider interactions of simultaneous notes, Duan and Temperley [45] proposed a maximum likelihood sampling approach to refine note-level transcription results.

B. Onset-Based Approach

In onset-based approaches, a separate onset detection stage is used during the transcription process. This approach is often adopted for transcribing piano music, given the relative prominence of onsets compared to other types of instruments. SONIC, a piano music transcription by Marolt *et al.*, used an onset detection stage to refine the results of neural network classifiers [46]. Costantini *et al.*[47] proposed a piano music transcription method with an initial onset detection stage to detect note onsets; a single CQT window of the 64 ms following the note attack is used to estimate the pitches with a multi-class SVM classification. Cogliati and Duan [48] proposed a piano music transcription method with an initial onset detection stage followed by a greedy search algorithm to estimate the pitches between two successive onsets. This method models the entire temporal evolution of piano notes.

C. Note-Based Approach

Note-based approaches combine the estimation of pitches and onsets (and possibly offsets) into a single framework. While this increases the complexity of the model, it has the benefit of integrating the pitch information and the onset information for both tasks. As an extension to Goto's statistical method [24], Kameoka *et al.* [49] used so-called harmonic temporal structured clustering to jointly estimate pitches, onsets, offsets and dynamics. Berg-Kirkpatrick *et al.* [50] combined an NMF-like approach in which each note is modeled by a spectral profile and an activation envelope with a two-state HMM to estimate play and rest states. Ewert *et al.* [51] modeled each note as a series of states, each state being a log-magnitude frame, and used a greedy algorithm to estimate the activations of the states. In this paper, we propose a note-based approach to simultaneously estimate pitches and onsets within a convolutional sparse coding framework. A preliminary version of this work was published in [12].

III. Background

In this section, we present the background material for convolutional sparse coding and its recently proposed efficient algorithm to prepare the reader for its application to automatic music transcription in Section IV.

A. Convolutional Sparse Coding

Sparse coding – the inverse problem of sparse representation of a particular signal – has been approached in several ways. One of the most widely used is Basis Pursuit DeNoising (BPDN) [52]:

$$\arg\min_{x} \frac{1}{2}\|D\boldsymbol{x} - \boldsymbol{s}\|_2^2 + \lambda\|\boldsymbol{x}\|_1, \tag{1}$$

where $\boldsymbol{s}$ is a signal to approximate, D is a dictionary matrix, $\boldsymbol{x}$ is the vector of activations of dictionary elements, and λ is a regularization parameter controlling the sparsity of $\boldsymbol{x}$.

Convolutional Sparse Coding (CSC), also called shift-invariant sparse coding, extends the idea of sparse representation by using convolution instead of multiplication. Replacing the multiplication operator with convolution in Eq. (1) we obtain Convolutional Basis Pursuit DeNoising (CBPDN) [53]:

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

where $\{\boldsymbol{d}_m\}$ is a set of dictionary elements, also called filters; $\{\boldsymbol{x}_m\}$ is a set of activations, also called coefficient maps; and λ controls the sparsity penalty on the coefficient maps $\boldsymbol{x}_m$. Higher values of λ lead to sparser coefficient maps and lower fidelity approximation to the signal s.

CSC has been widely applied to various image processing problems, including classification, reconstruction, denoising and coding [54]. In the audio domain, $\boldsymbol{s}$ represents the audio waveform for analysis, $\{\boldsymbol{d}_m\}$ represents a set of audio atoms, and $\{\boldsymbol{x}_m\}$ represents their activations. Its applications to audio signals include music representations [43], [55] and audio classification [56]. However, its adoption has been limited by its computational complexity in favor of faster factorization techniques, such as NMF or PLCA.

CSC is computationally very expensive, due to the presence of the convolution operator. A straightforward implementation in the time-domain [57] has a complexity of $\mathcal{O}(M^2N^2L)$, where M is the number of atoms in the dictionary, N is the size of the signal and L is the length of the atoms.

B. *Efficient Convolutional Sparse Coding*

An efficient algorithm for CSC has recently been proposed [54], [58]. This algorithm is based on the Alternating Direction Method of Multipliers (ADMM) for convex optimization [59]. The algorithm iterates over updates on three sets of variables. One of these updates is trivial, and the other can be computed in closed form with low computational cost. The additional update consists of a computationally expensive optimization due to the presence of the convolution operator. A natural way to reduce the computational complexity of convolution is to use the Fast Fourier Transform (FFT), as proposed by Bristow *et al.* [60] with a computational complexity of $\mathcal{O}(M^3N)$. The computational cost of this subproblem has been further reduced to $\mathcal{O}(MN)$ by exploiting the particular structure of the linear systems resulting from the transformation into the spectral domain [54], [58]. The overall complexity of the resulting algorithm is $\mathcal{O}(MN \log N)$ since it is dominated by the cost of FFTs. The complexity does not depend on the length of the atoms L as the atoms are zero-padded to the length of the signal N.

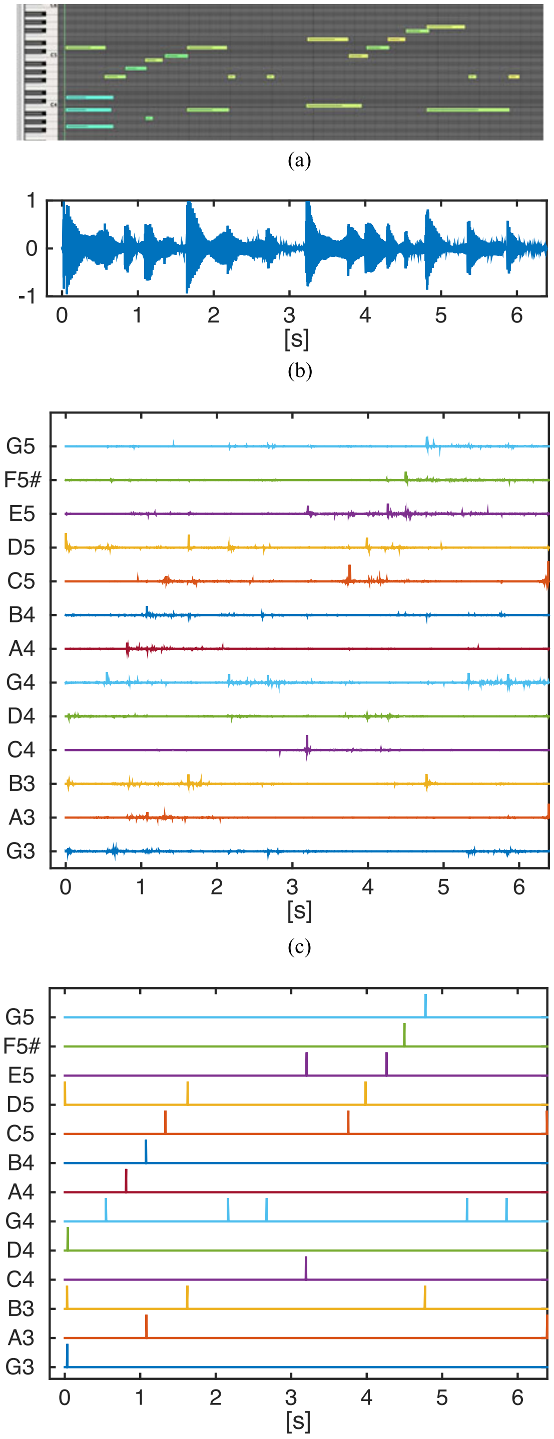
IV. Proposed Method

In this section, we describe how we model the piano transcription problem as a convolutional sparse coding problem in the time domain, and how we apply the efficient CSC algorithm [54], [58] to solve the problem.

A. *Transcription Process*

The whole transcription process is illustrated with an example in Fig. 1. Taking a monaural, polyphonic piano audio recording $\boldsymbol{s}(t)$ as input (Fig. 1(b)), we approximate it with a sum of dictionary elements $\boldsymbol{d}_m(t)$, representing a typical, amplitude-normalized waveform of each individual pitch of the piano, convolved with their activation vectors $\boldsymbol{x}_m(t)$:

$$\boldsymbol{s}(t) \simeq \sum_m \boldsymbol{d}_m(t) * \boldsymbol{x}_m(t). \tag{3}$$



(d)

Fig. 1. Piano roll (a), waveform produced by an acoustic piano (b), raw activation vectors (c), and the final detected note onsets (d) of Bach's Minuet in G major, BWV Anh 114, from the Notebook for Anna Magdalena Bach.

The dictionary elements $\boldsymbol{d}_m(t)$ are pre-set by sampling all the individual notes of a piano (see Section IV-A1) and are fixed during transcription. The activations $\boldsymbol{x}_m(t)$ are estimated using the efficient convolutional sparse coding algorithm [54], [58]. Note that the model is based on an assumption that the waveforms of the same pitch do not vary much with dynamic and duration. This assumption seems to be over-simplified, yet we will show that it is effective in the experiments. We will also discuss its limitations and how to improve the model in Section IV-B. Ideally, these activation vectors are impulse trains, with each impulse indicating the onset of the corresponding note at a certain time. In practice, the estimated activations contain some noise (Fig. 1(c)). After post-processing, however, they look like impulse trains (Fig. 1(d)), and recover the underlying ground-truth note-level transcription of the piece (Fig. 1(a)). Details of these steps are explained below.

1) Training: The dictionary elements are pre-learned in a supervised manner by sampling each individual note of a piano

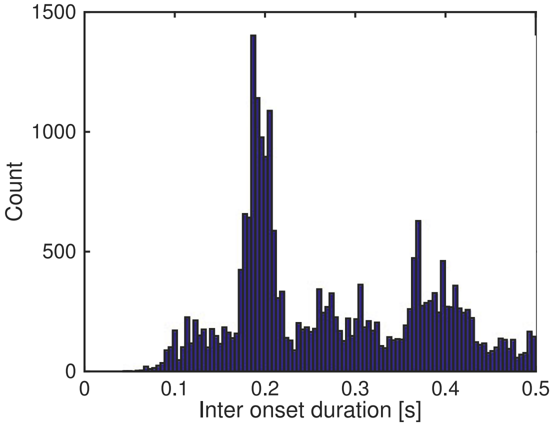


Fig. 2. Distribution of the time intervals between two consecutive activations of the same note in the ENSTDkCl collection of the MAPS dataset [26]. The distribution has been truncated to 0.5 s for visualization.

at a certain dynamic level, e.g., *forte*, for 1 s. We used a sampling frequency of 11,025 Hz to reduce the computational workload during the experiments. The length was selected by a parameter search (see Section V-A). The choice of the dynamic level is not critical, even though we observed that louder dynamics produce better results than softer dynamics.

2) Convolutional Sparse Coding: The activation vectors are estimated from the audio signal using an open source implementation [61] of the efficient convolutional sparse coding algorithm described in Section III-B. The sampling frequency of the audio mixture to be transcribed must match the sampling frequency used for the training stage, so we downsampled the audio mixtures to 11,025 Hz. As described in Section V-A, we investigated the dependency of the performance on the parameter λ on an acoustic piano dataset and selected the best value, $\lambda = 0.005$. We then used the same value for all experiments covering synthetic, anechoic, noisy and reverberant scenarios. We used 500 iterations in our experiments, even though we observed that the algorithm usually converges after approximately 200 iterations.

The result of this step is a set of raw activation vectors, which can be noisy due to the mismatch between the atoms in the dictionary and notes in the audio mixture (see Fig. 1(c)). Note that no non-negativity constraints are applied in the formulation, so the activations can contain negative values. Negative activations can appear in order to correct mismatches in loudness and duration between the dictionary element and the actual note in the sound mixture. However, because the waveform of each note is quite consistent across different instances (see Section IV-B), the strongest activations are generally positive.

3) Post-Processing: We perform peak picking by detecting local maxima from the raw activation vectors to infer note onsets. However, because the activations are noisy, multiple closely located peaks are often detected from the activation of one note. To deal with this problem, we only keep the earliest peak within a 50 ms window and discard the others. This enforces local sparsity of each activation vector. We choose 50 ms because it represents a realistic limit on how fast a performer can play the same note repeatedly. In fact, Fig. 2 shows the distribution of the time intervals between two consecutive activations of the same note in the ENSTDkCl collection of the MAPS dataset [26]. No interval is shorter than 50 ms.

4) Binarization: The resulting peaks are also binarized to keep only peaks that are higher than 10% of the highest peak in the entire activation matrix. This step is necessary to reduce ghost notes, i.e., false positives, and to increase the precision of the transcription. The value was chosen by comparing the RMS of each note played *forte* with the RMS of the corresponding note played *piano* in the isolated note collection of MAPS (ENSTDkCl set). The average ratio is 6.96, with most of the ratios below 10. This threshold is not tuned and is kept fixed throughout our experiments.

B. Discussion

The proposed model is based on the assumption that the waveform of a note of the piano is consistent when the note is played at different times at the same dynamic. This assumption is valid, thanks to the mechanism of piano note production [6]. Each piano key is associated with a hammer, one to three strings, and a damper that touches the string(s) by default. When the key is pressed, the hammer strikes the string(s) while the damper is raised from the string(s). The string(s) vibrate freely to produce the note waveform until the damper returns to the string(s), when the key is released. The frequency of the note is determined by the string(s); it is stable and cannot be changed by the performer (e.g., vibrato is impossible). The loudness of the note is determined by the velocity of the hammer strike, which is affected by how hard the key is pressed. The force applied to the key is the only control that the player has over the onset articulation. Modern pianos generally have three foot pedals: sustain, sostenuto, and soft pedals; some models omit the sostenuto pedal. The sustain pedal is commonly used. When it is pressed, all dampers of all notes are released from all strings, regardless whether a key is pressed or released. Therefore, its usage only affects the offset of a note, if we ignore the sympathetic vibration of strings across notes.

Fig. 3 shows the waveforms of four different instances of the C4 note played on an acoustic piano at two dynamic levels. We can see that the three *f* notes are very similar, even in the transient region of the initial 20 ms. The waveform of the the *mf* note is slightly different, but still resembles the other waveforms after applying a global scaling factor. Our assumption is that softer dynamics excite fewer modes in the vibration of the strings, resulting in less rich spectral content compared to louder dynamics. However, because the spectral envelope of piano notes is monotonically decreasing, higher partials have less energy compared to lower partials, so softer notes can still be approximated with notes played at louder dynamics. To prove the last assertion, we compared an instance of a C4 note played *forte* with different instances of the same pitch played at different dynamics and also with different pitches. As we can see from Table I, different instances of the same pitch are highly correlated, regardless of the dynamic, while the correlation between different pitches is low.

As discussed in Section II, Plumbley *et al.* [43] suggested a model similar to the one proposed here. The efficient CSC

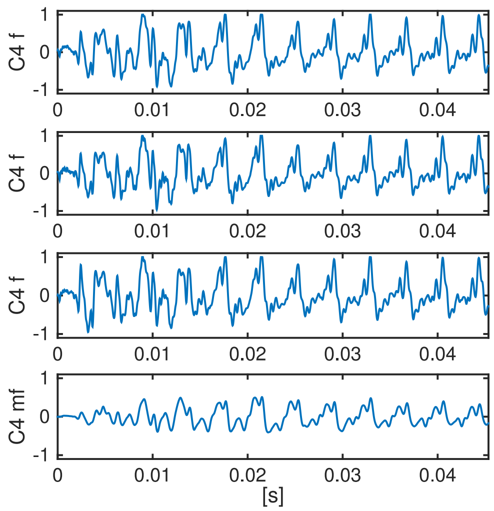


Fig. 3. Waveforms of four different instances of note C4 played manually on an acoustic piano, three at *forte (f)* and one at *mezzo forte (mf)*. Their waveforms are very similar, after appropriate scaling.

TABLE I
PEARSON CORRELATION COEFFICIENTS OF A SINGLE C4 NOTE PLAYED *forte* WITH THE SAME PITCH PLAYED AT DIFFERENT DYNAMIC LEVELS AND WITH DIFFERENT PITCHES. VALUES SHOWN ARE THE MAXIMA IN ABSOLUTE VALUE OVER ALL THE POSSIBLE ALIGNMENTS.

Note	Correlation Coefficient
C4 *f* #1	0.989
C4 *f* #2	0.969
C4 *f* #3	0.977
C4 *mf* #1	0.835
C4 *mf* #2	0.851
C4 *mf* #3	0.837
C4 *p* #1	0.608
C4 *p* #2	0.602
C4 *p* #3	0.606
C5 *f* #1	– 0.144
C5 *f* #2	– 0.146
C5 *f* #3	– 0.143
G4 *f* #1	– 0.016
G4 *f* #2	– 0.019
D4 *f* #1	0.042
D4 *f* #2	– 0.042

algorithm has also been applied to a score-informed source separation problem by Jao *et al.* in [62]. This method used very short atoms (100 ms), which might be a limiting factor as we prove in Section V, however this limitation may be mitigated, especially for sustaining instruments, by including 4 templates per pitch.

The proposed method can operate online by segmenting the audio input into 2 s windows, and retaining the activations for the first second. The additional second of audio is necessary to avoid the border effects of the circular convolution. Initial experiments show that the performance of the algorithm is unaffected by online processing, with the exception of silent frames. As the binarization step is performed in each window, silent frames introduce spurious activations in the final transcription, so an additional step to detect silent frames, either with a global thresholding or an adaptive filter, is required. Since the computation time of the algorithm is linear in the length of the signal, a shorter signal does not make the algorithm run in real-time in our current CPU-based implementation, which runs in about 5.9 times the length of the signal, but initial experiments with a GPU-based implementation of the CSC algorithm suggest that real-time processing is achievable.

V. EXPERIMENTS

We conduct experiments to answer two questions: (1) How sensitive is the proposed method to key parameters such as the sparsity parameter λ, and the length and loudness of the dictionary elements? (2) How does the proposed method compare with state-of-the-art piano transcription methods in different settings such as anechoic, noisy, and reverberant environments?

For the experiments we used three different datasets: the ENSTDkCl (close-mic acoustic recordings) and the SptkBGCl (synthetic recordings) collections from the MAPS dataset [26], and another synthetic dataset we created specially for this paper, using MIDI files in the ENSTDkCl collection. We will call this dataset ENSTGaSt.

The ENSTDkCl dataset is used to validate the proposed method in a realistic scenario. This collection contains 30 pieces of different styles and genres generated from high quality MIDI files that were manually edited to achieve realistic and expressive performances. The MIDI files will be used as the ground-truth for the transcription. The pieces were played on a Disklavier, which is an acoustic piano with mechanical actuators that can be controlled via MIDI input, and recorded in a close microphone setting to minimize the effects of reverb. The SptkBGCl dataset uses a virtual piano, the Steinway D from The Black Grand by Sampletekk. For both datasets, MAPS also provides the 88 isolated notes, each 1 s long, played at three different dynamics: *piano* (MIDI velocity 29), *mezzo-forte* (MIDI velocity 57) and *forte* (MIDI velocity 104). We always use the *forte* templates for all the experiments, except for the experiment investigating the effect of the dynamic level of the dictionary atoms. The synthetic dataset is also useful to set a baseline of the performance in an ideal scenario, i.e., absence of noise and reverb.

The ENSTGaSt dataset was created to investigate the dependency of the proposed method on the length of the dictionary elements, as note templates provided in MAPS are only 1 s long. The dataset was also used to verify some alignment issues that we discovered in the ground truth transcriptions of the ENSTDkCl and SptkBGCl collections of MAPS. The ENSTGaSt dataset was created from the same 30 pieces in the ENSTDkCl dataset and re-rendered from the MIDI files using a digital audio workstation (Logic Pro 9) with a sampled virtual piano plug-in (Steinway Concert Grand Piano from the Garritan Personal Orchestra); no reverb was used at any stage. The details of the synthesis model, i.e., the number of different samples per pitch and the scaling of the samples with respect to the MIDI velocity, are not publicly available. To gain some insight on the synthesis model we generated 127 different instances of the same pitch, i.e., C4, one for each value of the valid MIDI velocities, each 1 s long. We then compared the instances with cross correlation and

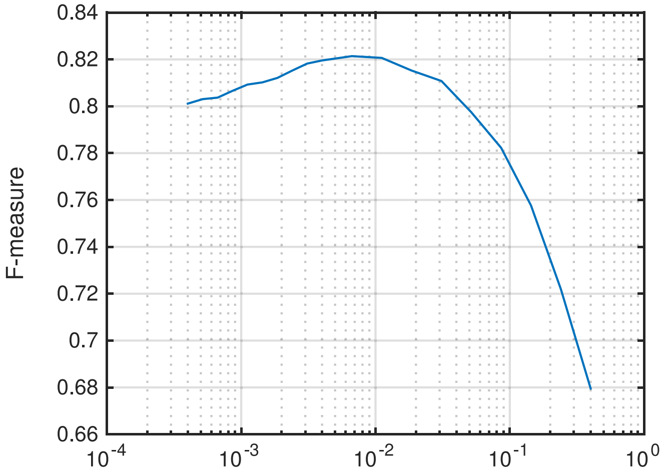


Fig. 4. Average F-measure on the 30 pieces in the ENSTDkCl collection (close-mic acoustic recordings) of the MAPS dataset for different values of λ, using 1 s long atoms.

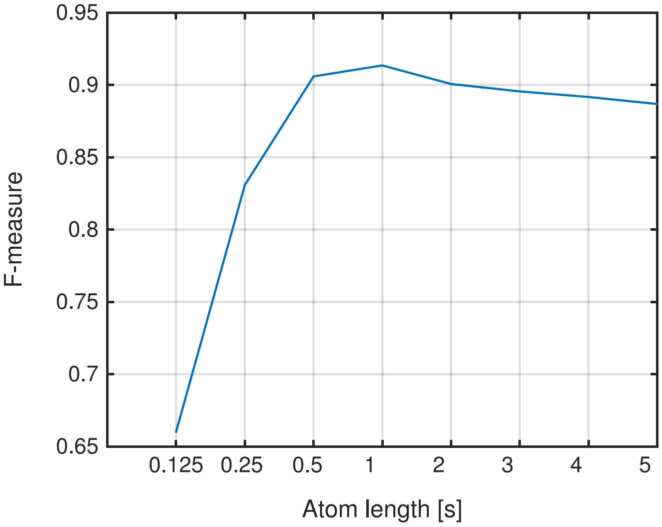


Fig. 5. Average F-measure on the 30 pieces in the ENSTGaSt dataset versus dictionary atom length, with λ fixed at 0.005.

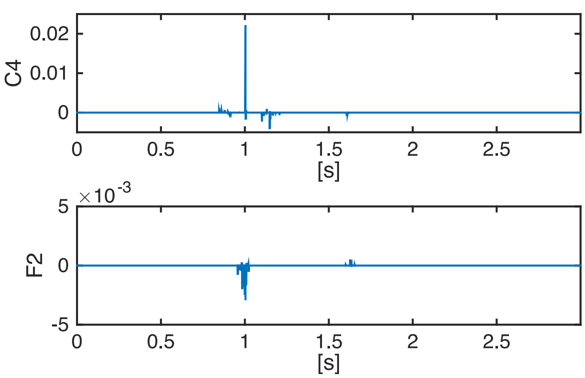


Fig. 6. Raw activations of the two most active note templates when transcribing a *piano* C4 note with 88 *forte* note templates. Note that the activation of the wrong note template is mostly negative.

determined that the virtual instrument uses 4 different samples per pitch, and that the amplitude of each sample is exponentially scaled based on the MIDI velocity. To ensure the replicability of this set of experiments, the dataset is available on the first author's website[1].

We use F-measure to evaluate the note-level transcription [4]. It is defined as the harmonic mean of precision and recall, where precision is defined as the percentage of correctly transcribed notes among all transcribed notes, and recall is defined as the percentage of correctly transcribed notes among all ground-truth notes. A note is considered correctly transcribed if its estimated discretized pitch is the same as a reference note in the ground-truth and the estimated onset is within a given tolerance value (e.g., ± 50 ms) of the reference note. We do not consider offsets in deciding the correctness.

A. Parameter Dependency

To investigate the dependency of the performance on the parameter λ, we performed a grid search with values of λ logarithmically spaced from 0.4 to 0.0004 on the ENSTDkCl collection in the MAPS dataset [26]. The dictionary elements were 1 s long. The results are shown in Fig. 4. As we can observe from Fig. 4, the method is not very sensitive to the value of λ. For a wide range of values, from 0.0004 to about 0.03, the average F-measure is always above 80%.

We also investigated the performance of the method with respect to the length of the dictionary elements, using the ENSTGaSt dataset. The average F-measure versus the length over all the pieces is shown in Fig. 5. The sparsity parameter λ is fixed at 0.005. The highest F-measure is achieved when the dictionary elements are 1 second long. The MAPS dataset contains pieces of very different styles, from slow pieces with long chords, to virtuoso pieces with fast runs of short notes. Our intuition suggested that longer dictionary elements would provide better results for the former, and shorter elements would be more appropriate for the latter, but we discovered that longer dictionary elements generally give better results for all the pieces.

Finally, we investigated the effect of the dynamic level of the dictionary atoms, using the ENSTDkCl collection. In general we found the proposed method to be very robust to differences in dynamic levels, but we obtained better results when louder dynamics were used during training. A possible explanation can be seen in Figs. 6 and 7. In Fig. 6 we transcribed a signal consisting of a single C4 note played *piano* with a dictionary of *forte* notes. The second most active note shows strong negative activations, which do not influence the transcription, as we only consider positive peaks. The negative activations might be due to the partials with greater amplitude contained in the *forte* dictionary element but not present in the *piano* note; i.e., CSC tries to achieve a better reconstruction by subtracting some frequency content. On the other side, in Fig. 7 we tested the opposite scenario, a single C4 note reconstructed *forte* with a dictionary of *piano* notes. The second most active note shows both positive and negative activations; positive activations might potentially

[1] http://www.ece.rochester.edu/˜acogliat/repository.html.

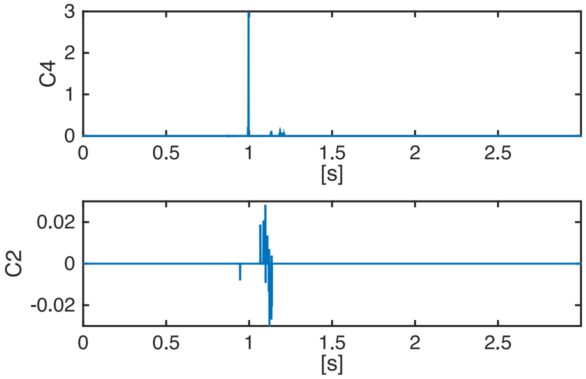


Fig. 7. Raw activations of the two most active note templates when transcribing a *forte* C4 note with 88 *piano* note templates. Note that the activation of the wrong note template contains a strong positive portion, which may lead to false positives in the final transcription.

lead to false positives. In this case, the *forte* note contains some spectral content not present in the *piano* template, so CSC improves the signal reconstruction by adding other note templates. Negative activations also appear when there is a mismatch between the length of a note in the audio signal and the length of the dictionary element. Using multiple templates per pitch, with different dynamics and different lengths, might reduce the occurrence of negative activations at the expense of increased computational time.

B. Comparison to State of the Art

We compared our method with a state-of-the-art AMT method proposed by Benetos and Dixon [32], which was submitted for evaluation to MIREX 2013 as BW3 [63]. The method will be referred to as BW3-MIREX13. This method is based on probabilistic latent component analysis of a log-spectrogram energy and uses pre-extracted note templates from isolated notes. The templates are also pre-shifted along the log-frequency in order to support vibrato and frequency deviations, which are not an issue for piano music in the considered scenario. The method is frame-based and does not model the temporal evolution of notes. To make a fair comparison, dictionary templates of both BW3-MIREX13 and the proposed method were learned on individual notes of the piano that was used for the test pieces. We used the implementation provided by the author along with the provided parameters, with the only exception of the hop size, which was reduced to 5 ms to test the onset detection accuracy.

1) Anechoic Settings: For this set of experiments we tested multiple onset tolerance values to show the highest onset precision achieved by the proposed method. The dictionary elements were 1 s long. We used the *forte* templates. The sparsity parameter λ was fixed at 0.005. The results are shown in Figs. 8–10. From the figures, we can notice that the proposed method outperforms BW3-MIREX13 by at least 20% in median F-measure for onset tolerance of 50 ms and 25 ms (50 ms is the standard onset tolerance used in MIREX [4]). When using dictionary elements played at *piano* dynamic, the median F-measure on the ENSTDkCl collection of the MAPS dataset drops to 70% (onset

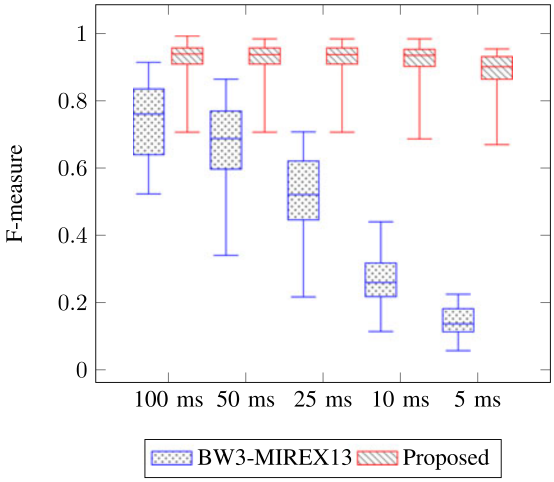


Fig. 8. F-measure for 30 pieces in the ENSTGaSt dataset (synthetic recordings). Each box contains 30 data points.

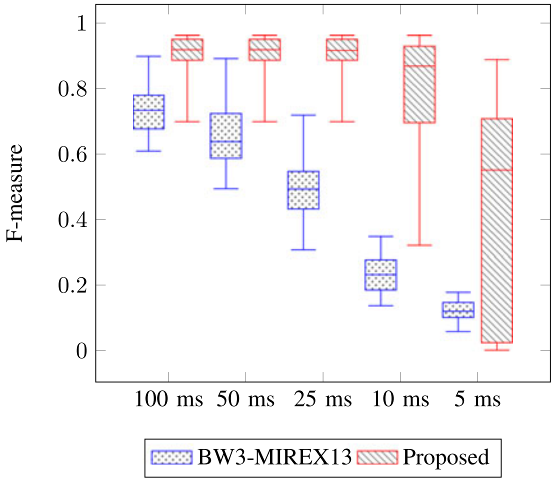


Fig. 9. F-measure for 30 pieces in the SptkBGCl dataset (synthetic recordings). Each box contains 30 data points.

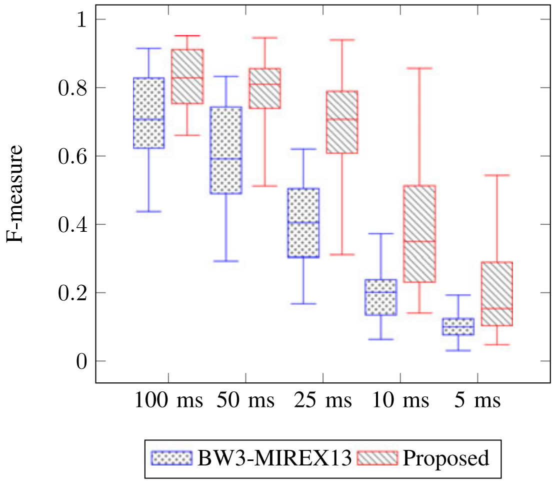


Fig. 10. F-measure for the 30 pieces in the ENSTDkCl collection (close-mic acoustic recordings) of the MAPS dataset. Each box contains 30 data points.

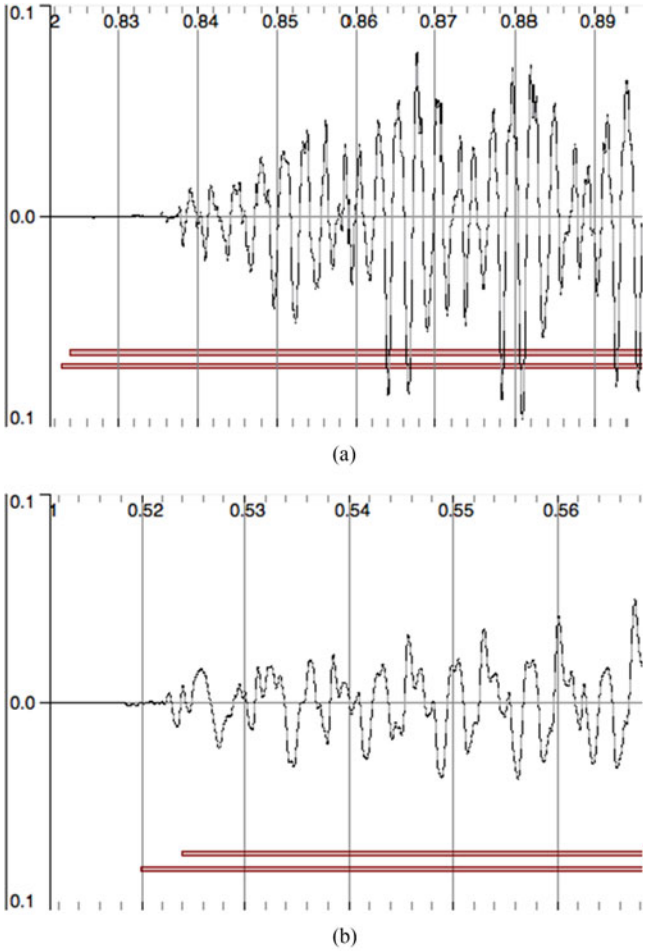


Fig. 11. Two pieces from the ENSTDkCl collection in MAPS showing different alignments between audio and ground truth MIDI notes (each red bar represents a note, as in a MIDI pianoroll). The figures show the beginning of the two pieces. The audio files are downmixed to mono for visualization. The time axis is in seconds.

tolerance set at 50 ms). In the experiment with the ENSTGaSt dataset, shown in Fig. 8, the proposed method exhibits consistent accuracy of over 90% regardless of the onset tolerance, while the performance of BW3-MIREX13 degrades quickly as the tolerance decreases under 50 ms. The proposed method maintains a median F-measure of 90% even with an onset tolerance of 5 ms. In the experiment on acoustic piano, both the proposed method and BW3-MIREX13 show a degradation of the performances with small tolerance values of 10 ms and 5 ms.

The degradation of performance on ENSTDkCl and SptkBgCl with small tolerance values, especially the increased support in the distribution of F-measure at 10 ms and 5 ms, drove us to further inspect the algorithm and the ground truth. We noticed that the audio and the ground truth transcription in the MAPS database are in fact not consistently lined up, i.e., different pieces show a different delay between the activation of the note in the MIDI file and the corresponding onset in the audio file. Fig. 11 shows two files from the ENSTDkCl collection of MAPS. Fig. 11(b) shows a good alignment between the audio and MIDI onsets, but in Fig. 11(a) the MIDI onsets occur 15 ms earlier than audio onsets. This inconsistency may be responsible for the poor results with small tolerance values.

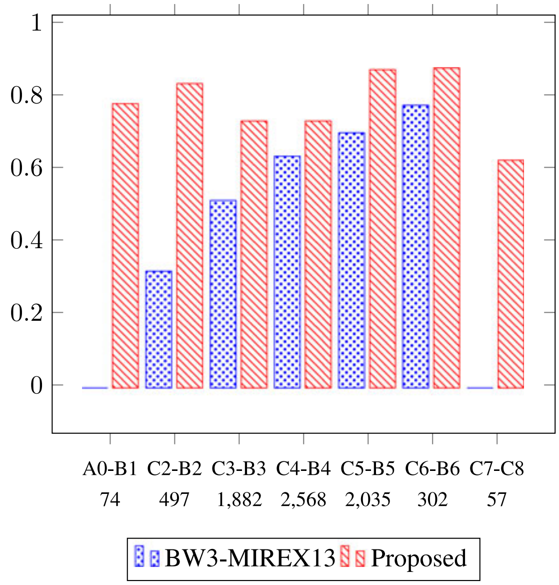


Fig. 12. Average F-measure per octave for the 30 pieces in the ENSTDkCl collection (close-mic acoustic recordings) of the MAPS dataset. Onset tolerance 50 ms. λ set to 0.005. The letters on the horizontal axis indicate the pitch range, the numbers show the total number of notes in the ground truth for the corresponding octave.

To test this hypothesis we re-aligned the ground truth with the audio by picking the mode of the onset differences for the correctly identified notes by the proposed method per piece. With the aligned ground truth, the results on the SptkBgCl dataset for 10 ms of tolerance are similar to the ones on the ENSTGaSt dataset; for 5 ms, the minimum F-measure is increased to 52.7% and the median is increased to 80.2%. On the ENSTDkCl dataset, the proposed method increases the median F-measure by about 15% at 10 ms and 5 ms. It might be argued that the improvement might be due to a systematic timing bias in the proposed method. However, as shown in Fig. 8, the transcription performance of the proposed method on the ENSTGaSt dataset does not show clear degradation when the onset tolerance becomes smaller. This suggests that there are some alignment problems between the audio and ground-truth MIDI transcription in the SptkBGCl and ENSTDkCl collections of MAPS. This potential misalignment issue only becomes prominent when evaluating transcription methods with small onset tolerance values, which are rarely used in the literature. Therefore, we believe that this issue requires additional investigations from the research community before our modified ground-truth can be accepted as the correct one. We thus make the modified ground-truth public on the first author's website, but still use the original non-modified ground truth in all experiments in this paper.

2) Robustness to Pitch Range and Polyphony: Fig. 12 compares the average F-measure achieved by the two methods along the different octaves of a piano keyboard. The figure clearly shows that the results of BW3-MIREX13 depend on the fundamental frequencies of the notes; the results are very poor

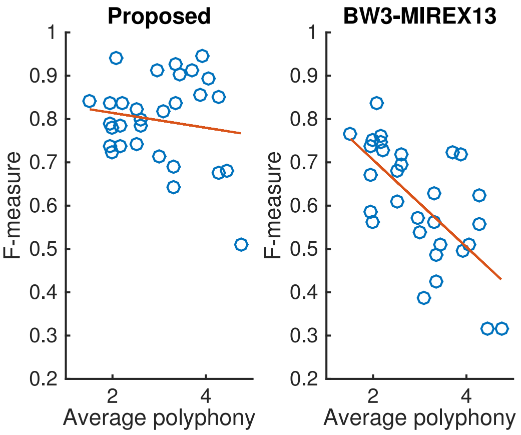


Fig. 13. F-measure of the 30 pieces in the ENSTDkCl collection (close-mic acoustic recordings) of MAPS versus average instantaneous polyphony. The orange line shows the linear regression of the data points.

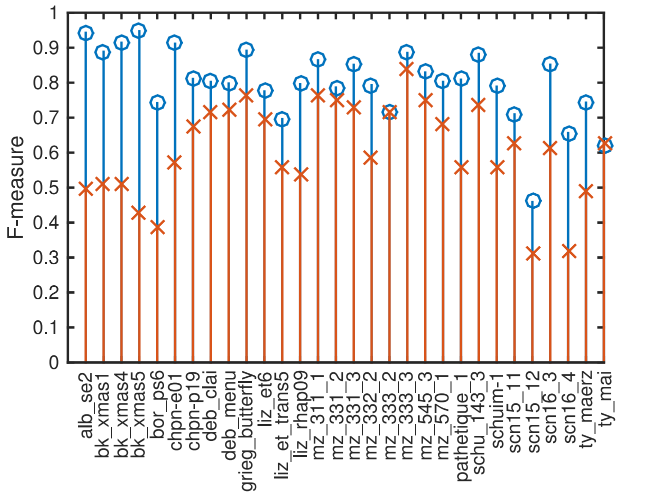


Fig. 14. Individual F-measures of the 30 pieces in the ENSTDkCl collection (close-mic acoustic recordings) of MAPS. Proposed method in blue circles, BW-MIREX13 in orange crosses.

for the first two octaves, and increase monotonically for higher octaves, except for the highest octave, which is not statistically significant. The proposed method shows a more balanced distribution. This suggests the advantage of our time-domain approach in avoiding the time-frequency resolution trade-off. We do not claim that operating in the time domain automatically overcomes the time-frequency trade-off, and explain the high accuracy of the proposed method as follows. Each dictionary atom contains multiple partials spanning a wide spectral range, and the relative phase and magnitude of the partials for a given note have low variability across instances of that pitch. This, together with the sparsity penalty, which limits the model complexity, allows for good performance without violating the fundamental time-frequency resolution limitations.

The proposed algorithm is less sensitive to the polyphony of the pieces compared to BW3-MIREX13. For each piece in the ENSTDkCl collection of MAPS we calculated the average polyphony by sampling the number of concurrently sounding notes every 50 ms. The results are shown in Fig. 13. BW3-MIREX13 shows a pronounced degradation in performance for denser polyphony, while the proposed method only shows minimal degradation.

Fig. 14 shows the results on the individual pieces of the ENSTDkCl collection of MAPS. The proposed method outperforms BW13-MIREX13 for all pieces except for two, for which the two methods achieve the same F-measure – Mozart's *Sonata 333*, second movement (mz_333_2) and Tchaikovsky's *May - Starlight Nights* (ty_mai) from *The Seasons*. The definite outlier is Schuman's *In Slumberland* (scn15_12), which is the piece with the worst accuracy for both the proposed method and BW13-MIREX13; it is a slow piece with the highest average polyphony in the dataset (see Fig. 13). The piece with the second worst score is Tchaikovsky's *May - Starlight Nights* (ty_mai); again a slow piece but with a lower average polyphony. A very different piece with an F-measure still under 70% is Listz's *Transcendental Étude no. 5* (liz_et5); it is a very fast piece with many short notes and high average polyphony. Further research is needed to investigate why a lower accuracy resulted from these pieces.

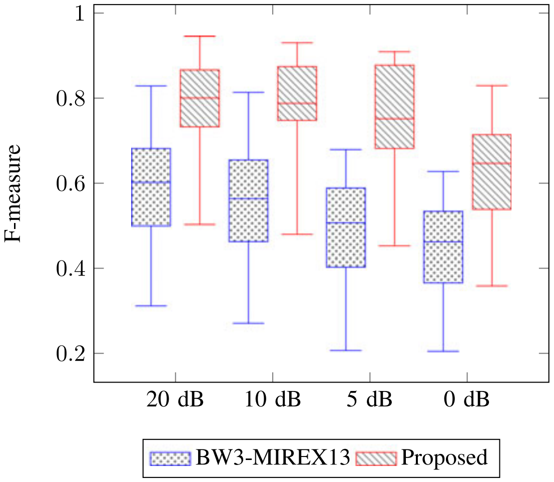


Fig. 15. F-measure for the 30 pieces in the ENSTDkCl collection (close-mic acoustic recordings) of MAPS with white noise at different SNR levels. Each box contains 30 data points.

3) Robustness to Noise: In this section, we investigate the robustness of the proposed method to noise, and compare the results with BW3-MIREX13. We used the original noiseless dictionary elements with length of 1 second and tested both white and pink additive noisy versions of the ENSTDkCl collection of MAPS. White and pink noises can represent typical background noises (e.g., air conditioning) in houses or practice rooms. We used the same parameter settings: $\lambda = 0.005$ and 1 s long, *forte* templates. The results are shown in Figs. 15 and 16. As we can notice from the plots, the proposed method shows great robustness to white noise, even at very low SNRs, always having a definite advantage over BW3-MIREX13. The proposed method consistently outperforms BW3-MIREX13 by about 20% in median F-measure, regardless of the level of noise. The proposed method is also very tolerant to pink noise and outperforms BW3-MIREX13 with low and medium levels of noise, up to an SNR of 5 dB.

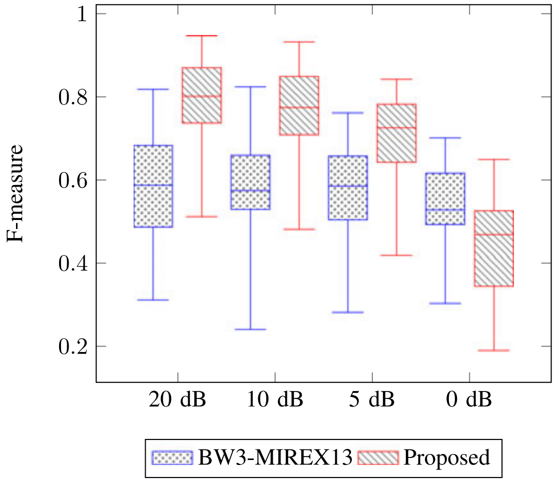


Fig. 16. F-measure for the 30 pieces in the ENSTDkCl collection (close-mic acoustic recordings) of MAPS with pink noise at different SNR levels. Each box contains 30 data points.

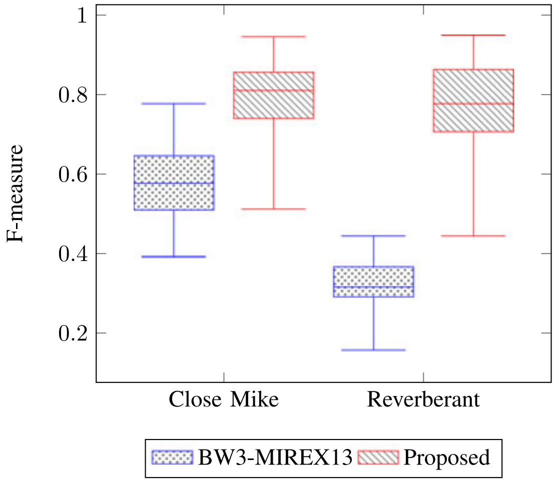


Fig. 17. F-measure for the 30 pieces in the ENSTDkCl collection (close-mic acoustic recordings) of MAPS with reverb. Each box contains 30 data points.

4) Robustness to Reverberation: In the third set of experiments we tested the performance of the proposed method in the presence of reverberation. Reverberation exists in nearly all real-world performing and recording environments, however, few systems have been designed and evaluated in reverberant environments in the literature. Reverberation is not even mentioned in recent surveys [1], [64]. We used a real impulse response of an untreated recording space[2] with an RT60 of about 2.5 s, and convolved it with both the dictionary elements and the audio files. The results are shown in Fig. 17. As we can notice, the median F-measure is reduced by about 3% for the proposed method in presence of reverb, showing a high robustness to reverb. The performance of BW3-MIREX13, however, degrades significantly, even though it was trained on the same reverberant piano notes. This further shows the advantage of the proposed method in real acoustic environments.

5) Sensitivity to Environment Mismatch: To illustrate the sensitivity of the method to the acoustic environment, we generated two synthetic impulse responses with RIR Generator [65], one with RT60 equal to 500 ms and the other with RT60 equal to 250 ms. These two values were picked to simulate an empty concert hall, and the same hall with an audience, whose presence reduces the reverberation time by adding absorption to the acoustic environment. We applied the longer impulse response to the dictionary and the shorter one to the 30 pieces in the ENSTDkCl collection of MAPS. The median F-measure for the experiment decreases from 82.7%, as in Fig. 10, to 75.2%. It should be noted that this is an extreme scenario, as a typical application would use a close mic setup, reducing the influence of the room acoustics.

6) Runtime: We ran all the experiments on an iMac equipped with a 3.2 GHz Intel Core i5 processor and 16 GB of memory. The code was implemented in MATLAB. For the 30 pieces in the ENSTDkCl collection of MAPS, the median runtime was 174 s, with a maximum of 186 s. Considering that we transcribed the first 30 s of each piece, the entire process takes about 5.9 times the length of the signal to be transcribed. Initial experiments with GPU implementation of the CSC algorithm show an average speedup of 10 times with respect to the CPU implementation.

VI. Discussion and Conclusions

In this paper we presented an automatic music transcription algorithm based on convolutional sparse coding in the time-domain. The proposed algorithm consistently outperforms a state-of-the-art algorithm trained in the same scenario in all synthetic, anechoic, noisy, and reverberant settings, except for the case of pink noise at 0 dB SNR. The proposed method achieves high transcription accuracy and time precision in a variety of different scenarios, and is highly robust to moderate amounts of noise. It is also highly insensitive to reverb, as long as the training session is performed in the same environment used for recording the audio to be transcribed. However, a limited generalization to a different room acoustic has been shown in the experiments.

While in this specific context the proposed method is clearly superior to the state-of-the-art algorithm used for comparison (BW3-MIREX13 [32]), it must be noted that our method cannot, at the moment, generalize to different contexts. In particular, it cannot transcribe performances played on different pianos not used for the training. Preliminary experiments with transcribing the ENSTDkCl dataset using the dictionary from the SptkBGCl dataset show a dramatic drop in precision resulting in an average F-measure of 16.9%; average recall remains relatively high at 64.7%. BW3-MIREX13 and, typically, other spectral domain-based methods are capable of being trained on multiple instruments and generalize to different instruments of the same kind. Nonetheless, the proposed context-dependent approach is useful in many realistic scenarios, considering that pianos are usually fixed in homes or studios. Moreover, the training procedure is simple and fast, in case the context changes. Future research is needed to adapt the dictionary to different pianos.

[2] WNIU Studio Untreated from the Open AIR Library http://www.openairlib.net/auralizationdb/content/wniu-studio-untreated.

The proposed method cannot estimate note offsets or dynamics, even though the amplitude of the raw activations (before binarization) is proportional to the loudness of the estimated notes. A dictionary containing notes of different lengths and different dynamics could be used in order to estimate those two additional parameters, even though group sparsity constraints should probably be introduced in order to avoid concurrent activations of multiple templates for the same pitch.

Another interesting future research direction is to evaluate the model on other percussive and plucked pitched instruments, such as harpsichord, marimba, bells and carillon, given the consistent nature of their notes and the model's ability to capture temporal evolution.

Acknowledgements

We would like to thank Dr. Emmanouil Benetos for kindly providing the code of his transcription system to compare the performance. We also thank the three anonymous reviewers for their thorough and constructive comments.

References

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

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

[3] M. Piszczalski and B. A. Galler, "Automatic music transcription," *Comput. Music J.*, vol. 1, no. 4, pp. 24–31, 1977.

[4] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of multiple-F0 estimation and tracking systems," in *Proc. Int. Soc. Music Inform. Retrieval Conf.*, 2009, pp. 315–320.

[5] P. R. Cook, *Music, Cognition, and Computerized Sound*. Cambridge, MA, USA: MIT Press, 1999.

[6] H. Suzuki and I. Nakamura, "Acoustics of pianos," *Appl. Acoust.*, vol. 30, no. 2, pp. 147–205, 1990.

[7] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047, Sep. 2005.

[8] S. Böck and G. Widmer, "Local group delay based vibrato and tremolo suppression for onset detection," in *Proc. Int. Soc. Music Inform. Retrieval Conf.*, 2013, pp. 361–366.

[9] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[10] D. Gabor, "Theory of communication. Part 1: The analysis of information," *J. Inst. Elect. Eng.—Part III: Radio Commun. Eng.*, vol. 93, no. 26, pp. 429–441, 1946.

[11] "MIREX2015 Results," (2015). [Online]. Available: http://www.music-ir.org/mirex/wiki/2015:Multiple_Fundamental_Frequency_Estimation_%26 Tracking Results - MIREX Dataset

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

[13] C. Raphael, "Automatic transcription of piano music," in *Proc. Int. Soc. Music Inform. Retrieval Conf.*, 2002.

[14] A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 804–816, Nov. 2003.

[15] C. Yeh, A. Röbel, and X. Rodet, "Multiple fundamental frequency estimation of polyphonic music signals," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 3, 2005, pp. iii-225–iii-228.

[16] K. Dressler, "Multiple fundamental frequency extraction for MIREX 2012," in *Proc. 8th Music Inform. Retrieval Eval. eXchange*, 2012.

[17] G. Poliner and D. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 8, pp. 154–162, Jan. 2007.

[18] A. Pertusa and J. M. Iñesta, "Multiple fundamental frequency estimation using Gaussian smoothness," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, Apr. 2008, pp. 105–108.

[19] S. Saito, H. Kameoka, K. Takahashi, T. Nishimoto, and S. Sagayama, "Specmurt analysis of polyphonic music signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 639–650, Mar. 2008.

[20] 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 Inform. Retrieval Conf.*, 2011, pp. 175–180.

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

[22] 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. Learning*, Scotland, U.K., 2012, pp. 1159–1166.

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

[24] 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.

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

[26] 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.

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

[28] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–91, 1999.

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

[30] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2003, pp. 177–180.

[31] 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.

[32] 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.

[33] S. A. Abdallah and M. D. Plumbley, "Polyphonic music transcription by non-negative sparse coding of power spectra," in *Proc. 5th Int. Conf. Music Inform. Retrieval*, 2004, pp. 318–325.

[34] 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.

[35] 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.

[36] R. Meddis and M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," *J. Acoust. Soc. Amer.*, vol. 89, pp. 2866–2882, 1991.

[37] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 708–716, Nov. 2000.

[38] P. J. Walmsley, S. J. Godsill, and P. J. Rayner, "Polyphonic pitch tracking using joint bayesian estimation of multiple frame parameters," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 1999, pp. 119–122.

[39] A. T. Cemgil, H. J. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 679–694, Mar. 2006.

[40] M. Davy, S. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *J. Acoust. Soc. Amer.*, vol. 119, no. 4, pp. 2498–2517, 2006.

[41] J. P. Bello, L. Daudet, and M. B. Sandler, "Automatic piano transcription using frequency and time-domain information," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2242–2251, Nov. 2006.

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

[43] M. D. Plumbley, S. A. Abdallah, T. Blumensath, and M. E. Davies, "Sparse representations of polyphonic music," *Signal Process.*, vol. 86, no. 3, pp. 417–431, 2006.

[44] M. Ryynänen and A. Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Comput. Music J.*, vol. 32, no. 3, pp. 72–86, Fall 2008.

[45] Z. Duan and D. Temperley, "Note-level music transcription by maximum likelihood sampling," in *Proc. Int. Symp. Music Inform. Retrieval Conf.*, Oct. 2014, pp. 181–186.

[46] M. Marolt and A. Kavcic and M. Privosnik and S. Divjak, "On detecting note onsets in piano music," in *Proc. 11th Medit Electrotech. Conf., 2002—MELECON 2002*, pp. 385–389, doi: 10.1109/MELECON.2002.1014600.

[47] G. Costantini, R. Perfetti, and M. Todisco, "Event based transcription system for polyphonic piano music," *Signal Process.*, vol. 89, no. 9, pp. 1798–1811, 2009.

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

[49] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 982–994, Mar. 2007.

[50] T. Berg-Kirkpatrick, J. Andreas, and D. Klein, "Unsupervised transcription of piano music," in *Proc. Adv. Neural Inform. Process. Syst.*, 2014, pp. 1538–1546.

[51] 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, Australia, 2015, pp. 569–573.

[52] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[53] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2528–2535.

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

[55] T. Blumensath and M. Davies, "Sparse and shift-invariant representations of music," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 50–57, Jan. 2006.

[56] R. Grosse, R. Raina, H. Kwong, and A. Y. Ng, "Shift-invariance sparse coding for audio classification," in UAI 2007, Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, Vancouver, Canada, 2007, pp. 149–158.

[57] M. Zeiler, D. Krishnan, G. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2528–2535.

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

[59] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[60] H. Bristow, A. Eriksson, and S. Lucey, "Fast convolutional sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 391–398.

[61] B. Wohlberg. *MATLAB Library SParse Optimization Research COde (SPORCO) version 0.0.2*. (2015) [Online]. Available: http://math.lanl.gov/brendt/Software/SPORCO/

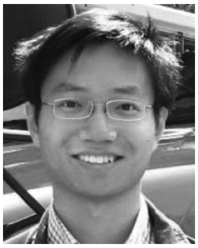
[62] P.-K. Jao, Y.-H. Yang, and B. Wohlberg, "Informed monaural source separation of music based on convolutional sparse coding," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Brisbane, Australia, Apr. 2015, pp. 236–240.

[63] E. Benetos and T. Weyde, "BW3—MSSIPLCA_fast_NoteTracking2," (2013). [Online]. Available: http://www.music-ir.org/mirex/wiki/2013:Multiple_Fundamental_Frequency_Estimation_%26 Tracking Results

[64] M. Davy and A. Klapuri, *Signal Processing Methods for Music Transcription*. New York, NY, USA: Springer, 2006.

[65] E. Habets, "RIR Generator," (2003). [Online]. Available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator

Andrea Cogliati received the B.S. (Laurea) and M.S. (Diploma) degrees in mathematics from University of Pisa, Pisa, Italy and Scuola Normale Superiore, Pisa, Italy, respectively. He is currently working toward the Ph.D. degree in the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA. He is working in the AIR Lab under the supervision of Dr. Zhiyao Duan. He spent almost 20 years working in the IT industry as a Consultant and Trainer. His research interests include computer audition, in particular automatic music transcription and melody extraction.



Zhiyao Duan (S'09–M'13) received the B.S. and M.S. degrees in automation from Tsinghua University, Beijing, China, in 2004 and 2008, respectively, and received the Ph.D. degree in computer Science from Northwestern University, Evanston, IL, USA, in 2013. He is currently an Assistant Professor in the Electrical and Computer Engineering Department, University of Rochester, Rochester, NY, USA. His research interests include the broad area of computer audition, i.e. designing computational systems that are capable of analyzing and processing sounds, including music, speech, and environmental sounds. Specific problems that he has been working on include automatic music transcription, multi-pitch analysis, music audio-score alignment, sound source separation, and speech enhancement.

Brendt Wohlberg received the B.Sc.(Hons.) degree in applied mathematics and the M.Sc. in applied science and Ph.D. degrees in electrical engineering from the University of Cape Town, South Africa, in 1990, 1993 and 1996, respectively. He is currently a Staff Scientist in Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, USA. His primary research interests include signal and image processing inverse problems, with an emphasis on sparse representations and exemplar-based methods. From 2010 to 2014, he was an Associate Editor of IEEE Transactions on Image Processing, and is currently Chair of the Computational Imaging Special Interest Group of the IEEE Signal Processing Society and an Associate Editor of IEEE Transactions on Computational Imaging.