Monaural Music Source Separation Using Convolutional Sparse Coding

Ping-Keng Jao, Li Su, *Member, IEEE*, Yi-Hsuan Yang, *Member, IEEE*, and Brendt Wohlberg, *Senior Member, IEEE*

***Abstract*—We present a comprehensive performance study of a new time-domain approach for estimating the components of an observed monaural audio mixture. Unlike existing time–frequency approaches that use the product of a set of spectral templates and their corresponding activation patterns to approximate the spectrogram of the mixture, the proposed approach uses the sum of a set of convolutions of estimated activations with prelearned dictionary filters to approximate the audio mixture directly in the time domain. The approximation problem can be solved by an efficient convolutional sparse coding algorithm. The effectiveness of this approach for source separation of musical audio has been demonstrated in our prior work, but under rather restricted and controlled conditions, requiring the musical score of the mixture being informed *a priori* and little mismatch between the dictionary filters and the source signals. In this paper, we report an evaluation that considers wider, and more practical, experimental settings. This includes the use of an audio-based multipitch estimation algorithm to replace the musical score, and an external dataset of audio single notes to construct the dictionary filters. Our result shows that the proposed approach remains effective with a larger dictionary, and compares favorably with the state-of-the-art nonnegative matrix factorization approach. However, in the absence of the score and in the case of a small dictionary, our approach may not be better.**

***Index Terms*—Convolutional sparse coding (CSC), multipitch estimation (MPE), monaural music source separation, nonnegative matrix factorization (NMF), phase, score-informed source separation.**

I. Introduction

THE goal of source separation is to recover the source signals that constitute an observed mixture. The observed mixture can be either single-channel or multi-channel, depending on the number of sensors employed to record the source signals. When there are fewer observed channels than sources, the separation problem is underdetermined and some prior knowledge about the source signals is needed to improve the separation. This is the case in many applications including musical audio, where there are usually no more than two channels.

A popular approach to source separation is to learn a model for each source (e.g. instrument) beforehand using a collection of clean source signals (i.e. a recording of the signal from a specific source with no interference from other sources). For audio source separation, this can be done by learning a set of spectral templates for each source, supposing that the spectral templates can cover the possible variations of a specific source in the frequency domain. Given a mixture, it is expected that a linear combination of the spectral templates corresponding to a specific source can be used to reconstruct the source signal that composes that mixture. Algorithms based on this idea, such as non-negative matrix factorization (NMF) [1]–[7], complex matrix factorization (CMF) [8]–[10] and probabilistic latent component analysis (PLCA) [11], [12], have been widely studied in the last decade. As the spectral templates are used as the basis to decompose the observed spectrogram, we also refer to them as the *dictionary vectors*.

In this paper, we instead study the decomposition of the mixture directly in the time domain, a less studied approach for source separation. Specifically, we propose to learn a set of time-domain filters from the clean source signals, and then use the sum of convolutions of *estimated* activations with the *dictionary filters* to approximate the audio mixture in the time domain. The potential advantages of this approach include:

1) While the frequency-domain approach needs to partition an input signal into successive frames with fixed-length short-time windows for Short-time Fourier Transform (STFT), the time-domain approach decomposes the mixture continuously without such fine-grained binning. Therefore, the time-domain approach may better capture the local temporal information of the signal.
2) While the frequency-domain approach decomposes the spectrum of each short-time frame independently and requires additional regularizers for ensuring the temporal continuity between adjacent frames in the recovered sources [5], the time-domain approach inherently accounts for temporal continuity.
3) While the frequency-domain approach requires specifically designed mechanisms to take care of phase [13]–[15], the phase information is implicitly considered when decomposing the signal in the time domain.
4) Many audio effects such as reverberation can usually be modeled in terms of convolution [16] and can therefore be easily taken into account by the proposed approach.

Mathematically, the time-domain approach to source separation can be formulated as follows. We are given a monaural time-domain audio signal $\mathbf{x} \in \mathbb{R}^n$ and a number of dictionary filters $\{\mathbf{d}_i\}_{i=1}^k$, where $\mathbf{d}_i \in \mathbb{R}^{t_i}, \forall i \in \mathcal{K} = \{1, 2, ..., k\}$, and n,

Manuscript received October 18, 2015; revised March 23, 2016 and June 27, 2016; accepted July 19, 2016. Date of publication August 4, 2016; date of current version September 2, 2016. This work was supported in part by a grant of the National Science Council of Taiwan under Contract MOST102-2221-E-001-004-MY3 and in part by the Academia Sinica Career Development Program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hirokazu Kameoka.

P.-K. Jao, L. Su, and Y.-H. Yang are with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 11564, Taiwan (e-mail: nafraw@citi.sinica.edu.tw; lisu@citi.sinica.edu.tw; yang@citi.sinica.edu.tw).

B. Wohlberg is with the 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.2016.2598323

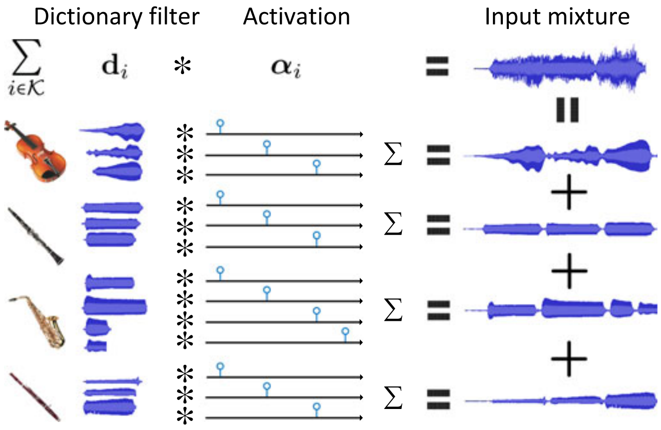


Fig. 1. Illustration of the proposed CSC approach to source separation. The dictionary filters are constructed in the training stage from the clean source signals, which are assumed to be similar with the unobserved source signals that compose an input mixture given in the testing stage. The activation patterns are estimated by CSC and are used to recover the unobserved source signals by summing up the corresponding convolutions.

t_i and k denote the length of the signal, the length of the ith dictionary filter, and number of filters, respectively. The goal is to find the activation patterns $\{\boldsymbol{\alpha}_i\}_{i=1}^{k}$ associated with each of the filters by solving the following optimization problem:

$$\underset{\{\boldsymbol{\alpha}_i\}_{i\in\mathcal{K}}}{\arg\min}\, f\left(\mathbf{x}-\sum_{i\in\mathcal{K}}\mathbf{d}_i*\boldsymbol{\alpha}_i\right)+\sum_{i\in\mathcal{K}}\lambda_i\, g(\boldsymbol{\alpha}_i)\,, \qquad (1)$$

where $f(\cdot)$ is a measurement of the fidelity of the approximation, $g(\cdot)$ is a regularizer weighted by the parameters λ_i, and $*$ is the convolution operator. As the activation patterns can be assumed to be sparse [5], a natural solution is to use the l_2 norm for $f(\cdot)$ and the sparsity-enforcing l_1 norm for $g(\cdot)$ [17], [18],[1] leading to the formulation of the so-called convolutional sparse coding (CSC) [20], [21].

The present paper is motivated by the development of an efficient solver of the CSC problem [22], [23] that reduces the time complexity of a previous alternating direction method of multipliers (ADMM) approach [24], [25] from $O(k^3n+kn\log n)$ to $O(kn\log n)$. This boost in efficiency makes it practical to employ CSC for source separation problems, which usually require a great many dictionary filters.

Fig. 1 illustrates the proposed CSC-based approach. The filters for different sources represent non-overlapping groups $\{\mathcal{G}_j\}_{j=1}^{p}$, where $\bigcup_{j=1}^{p}\mathcal{G}_j=\mathcal{K}$ and p denotes the number of sources. Therefore, the jth source can be recovered by:

$$\mathbf{e}_j=\sum_{i\in\mathcal{G}_j}\mathbf{d}_i*\boldsymbol{\alpha}_i\,. \qquad (2)$$

By virtue of Eq. (1), we have $\mathbf{e}_j\in\mathbb{R}^n$ and $\mathbf{x}\simeq\sum_{j=1}^{p}\mathbf{e}_j$.

The effectiveness of this approach for monaural music source separation was validated under rather restricted and controlled conditions in a prior work [26]. This paper extends and expands the prior work in the following respects:

1) In [26], we considered a simpler, *score-informed* setting to deactivate dictionary filters that do not match the score, i.e., mismatch of either pitch or instrument information associated to the filters. This simplifies that task and usually improves the separation quality, as these mismatched ones will likely not be used in the approximation of the input mixture when solving Eq. (1). However, the approach is not practical in real-world applications, because the musical score is not always available. Therefore, we propose to use an audio-based, automatic multi-pitch estimation (MPE) algorithm [27]–[30] to replace the role of the musical score.[2] As we do not assume the availability of scores in this setting, it can be considered as a *blind* source separation problem.[3]
2) In [26], we used a held-out set whose acoustic content is close to that of the test set for learning the dictionary filters, so as to minimize the possible mismatch between the filters and the test mixtures. In this paper, we consider a separate dataset comprising audio of single notes to learn the dictionary, thereby improving the generalizability of the experimental results.
3) We propose a heuristic to allow for arbitrary segmentation of the audio input for separation with alleviated boundary discontinuity, which will be explained in Section III-B, due to the use of circular convolution in Eqs. (1) and (2).
4) We report a comprehensive evaluation of the proposed time-domain approach against state-of-the-art frequency-domain approaches in a variety of experimental settings. A parameter sensitivity test for CSC is also reported.

Although our evaluation only concerns music, we expect that this methodology can be easily extended to speech.

The paper is organized as follows. Section II reviews related work. Section III describes the CSC algorithm and the proposed source separation approach. Then Sections IV–VI report the experiments and Section VII concludes the paper.

II. Related Work

Source separation is a fundamental signal processing problem. For many applications in music, such as audio remixing, remastering, and restoration, source separation is usually a required pre-processing step as the commercial music contents are mostly provided in only one or two channels. We might want to recover all the sources to manipulate and process them individually [31], [32], or to isolate a specific source (e.g. the singing voice or the lead guitar) from the mixture [33]–[36]. Source separation is also important for many other music information retrieval problems, such as instrument recognition [37], [38], singing voice analysis and editing [39], [40], beat tracking [41], drum pattern analysis [42], and automatic music transcription [43]–[45], amongst others.

[1] The l_2 and l_1 norms are defined as $\|\mathbf{a}\|_2=\sqrt{\sum_i a_i^2}$ and $\|\mathbf{a}\|_1=\sum_i|a_i|$, respectively, where $|\cdot|$ takes the absolute value. The term 'sparsity' suggests that only a few elements of the vector $\mathbf{a}$ are non-zero. It has been shown that the l_1 norm is the only norm that is both convex and sparsity-enforcing [19].

[2] We define MPE as a task that aims at automatically transcribing the notes of a polyphonic musical signal [28]. Therefore, MPE provides information about the pitch, onset and duration, but not the instruments.

[3] However, we do assume that we know the instruments presenting in the pieces (and strictly speaking this counts as a side information). We need this information for building the dictionary for source separation but not for MPE.

Since source separation for polyphonic music is typically an underdetermined problem, the use of prior knowledge or side information has been explored for better separation quality. First, it is usually assumed that the instruments contained in the mixture are known *a priori*. Clean source signals for the instruments are usually taken as prior knowledge to build dictionaries or models to characterize the sources [1]–[12]. Another important type of side information that has been increasingly used in the literature is the musical score [31], [45]–[50], which provides information about the notes (i.e. pitch, onset and duration) of a music piece. For example, the pitch and onset information can be employed to impose constraints on both the spectral templates and the activation patterns for NMF [46]. Beyond musical scores, other side information such as user input [51]–[53], automatically estimated pitch information [54], [55], and the cover version of a song [56] have also been exploited.

NMF is arguably the most widely studied technique for source separation over the past few years. The idea is to take the magnitude part of the STFT of an input, denoted as $\hat{\mathbf{X}} \in \mathbb{R}_{\succeq 0}^{r\times h}$, where r denotes the number of frequency bins and h the number of frames, and then approximate it by the product of two non-negative matrices $\mathbf{W} \in \mathbb{R}_{\succeq 0}^{r\times k}$ and $\mathbf{H} \in \mathbb{R}_{\succeq 0}^{k\times h}$:

$$\hat{\mathbf{X}} \approx \mathbf{W}\mathbf{H}\,, \tag{3}$$

where k is the number of dictionary vectors. A column in $\mathbf{W}$ can represent the frequency distribution of a specific pitch of an instrument, and its time (frame) activation is given in the corresponding row of $\mathbf{H}$. To ensure that different subsets of the dictionary are associated with different sources, an external dataset with clean source signals is usually employed to learn the dictionary offline [4]. Given $\mathbf{W}$ and $\hat{\mathbf{X}}$, the activation $\mathbf{H}$ can be computed with a multiplicative update algorithm that preserves the nonnegativity of the solution [57]. The jth source can then be recovered by taking the inverse STFT of the product of the corresponding dictionary vectors and activation patterns. In the score-informed setting, we can learn $\mathbf{W}$ directly from $\hat{\mathbf{X}}$ (i.e. without using an external dataset), by properly imposing constraints on different subsets of the dictionary for different sources [46].

It can be seen that NMF assumes the reconstruction can be done using the magnitude part of the STFT, leaving the phase information unaddressed. To estimate the time-domain source signals, the phase of the mixture is usually used directly in the inverse STFT. On one hand, it has been noted that magnitude additivity does not hold since the concurrent sources are typically not in-phase [58]. On the other hand, copying the phase of mixture to the individual source signals might lead to perceptual artifacts, as phase also carries important timbre information [59], [60]. This issue can be mitigated by using the complex variant of the NMF [8], [9], phase reconstruction methods such as multiple input spectrogram inversion (MISI) or consistent Wiener filtering [61]–[64], or other elaborated designs [13]–[15]. For example, Kameoka proposed a time-domain spectrogram factorization algorithm [15] that optimizes both $\mathbf{W}$ and $\mathbf{H}$ with a fidelity term defined in the time domain. Although such methods are promising and relevant, they do not share the advantages of the proposed approach outlined in Section I.

As musical audio is often stereo, it is possible to exploit the spatial information from the two channels for better separation [52], [56]. In this paper, we downmix the stereo signals into mono-channel ones and consider a monaural source separation problem. The development of the multi-channel version of the proposed approach is left as a future work.

CSC is conceptually closely related to shift-invariant sparse coding [21], [65]. We refer readers to [23] for a detailed review and an extensive discussion on the relations between CSC and shift-invariant sparse coding.

While CSC has been predominantly applied to computer vision problems thus far [20]–[25], shift-invariant sparse coding has been applied to audio source separation since a decade ago [66]–[68]. However, possibly due to the high computational complexity involved, these algorithms have not been evaluated with a dataset of reasonable size. For example, Mørup *et al.* [68] considered only a very simplified scenario of separating an organ from a piccolo in their experiment. To the best of our knowledge, the work reported in the present paper represents the first attempt to systematically evaluate the performance of a modern CSC algorithm for both score-informed and blind (i.e. in the case of MPE-informed) source separation of musical sources.

III. Algorithm

A. *Convolutional Sparse Coding*

Selecting the l_2 and l_1 norms respectively for $f(\cdot)$ and $g(\cdot)$ in Eq. (1) leads to the Convolutional Basis Pursuit DeNoising (CBPDN) [23] form of CSC:

$$\operatorname*{arg\,min}_{\{\boldsymbol{\alpha}_i\}_{i\in\mathcal{K}}} \frac{1}{2}\Big\|\mathbf{x} - \sum_{i\in\mathcal{K}} \mathbf{d}_i * \boldsymbol{\alpha}_i\Big\|_2^2 + \sum_{i\in\mathcal{K}} \lambda_i \left\|\boldsymbol{\alpha}_i\right\|_1 . \tag{4}$$

We can see that $\boldsymbol{\alpha}_i$ and $\mathbf{d}_i$ play similar roles as a row in $\mathbf{H}$ and a column of $\mathbf{W}$ in NMF do, respectively. The major difference is that CSC uses the convolution operator for l_1-regularized regression [69] and does not require the variables to be non-negative. The most efficient CBPDN algorithm proposed thus far [22], [23] is based on the ADMM algorithm [70], which introduces auxiliary variables $\{\boldsymbol{\beta}_i\}_{i=1}^{k}$ to Eq. (4):

$$\operatorname*{arg\,min}_{\{\boldsymbol{\alpha}_i\},\{\boldsymbol{\beta}_i\}} \frac{1}{2}\Big\|\mathbf{x} - \sum_{i\in\mathcal{K}} \mathbf{d}_i * \boldsymbol{\alpha}_i\Big\|_2^2 + \sum_{i\in\mathcal{K}} \lambda_i \left\|\boldsymbol{\beta}_i\right\|_1 \text{ s.t. } \boldsymbol{\alpha}_i = \boldsymbol{\beta}_i\,, \tag{5}$$

and then iteratively solves the following sub-problems:

$$\begin{aligned}\{\boldsymbol{\alpha}_i\}^{(\tau+1)} &= \operatorname*{arg\,min}_{\{\boldsymbol{\alpha}_i\}} \frac{1}{2}\Big\|\mathbf{x} - \sum_{i} \mathbf{d}_i * \boldsymbol{\alpha}_i\Big\|_2^2 \\ &\quad + \frac{\rho}{2}\sum_i \left\|\boldsymbol{\alpha}_i - \boldsymbol{\beta}_i^{(\tau)} + \boldsymbol{\gamma}_i^{(\tau)}\right\|_2^2 ,\end{aligned} \tag{6}$$

$$\begin{aligned}\{\boldsymbol{\beta}_i\}^{(\tau+1)} &= \operatorname*{arg\,min}_{\{\boldsymbol{\beta}_i\}} \sum_i \lambda_i \|\boldsymbol{\beta}_i\|_1 \\ &\quad + \frac{\rho}{2}\sum_i \left\|\boldsymbol{\alpha}_i^{(\tau+1)} - \boldsymbol{\beta}_i + \boldsymbol{\gamma}_i^{(\tau)}\right\|_2^2 ,\end{aligned} \tag{7}$$

$$\boldsymbol{\gamma}_i^{(\tau+1)} = \boldsymbol{\gamma}_i^{(\tau)} + \boldsymbol{\alpha}_i^{(\tau+1)} - \boldsymbol{\beta}_i^{(\tau+1)} , \tag{8}$$

where τ indexes the iteration numbers, ρ is a positive penalty parameter, and $\boldsymbol{\gamma}_i$ is the Lagrange multiplier [70] employed to enforce the equality constraint in Eq. (5). Sub-problem (7) is a classic l_1-minimization problem that can be efficiently solved in $O(kn)$ via the soft-thresholding (shrinkage) operator [19]. Sub-problem (6) is much more expensive as it involves a number of convolutions. An efficient solution is possible by exploiting the convolution theorem, transforming the problem to the frequency domain via the Fast Fourier Transform (FFT)

$$\arg\min_{\{\hat{\boldsymbol{\alpha}}_i\}} \frac{1}{2}\left\|\hat{\mathbf{x}} - \sum_i \hat{\mathbf{d}}_i \circ \hat{\boldsymbol{\alpha}}_i\right\|_2^2 + \frac{\rho}{2}\sum_i \|\hat{\boldsymbol{\alpha}}_i - \hat{\boldsymbol{\beta}}_i + \hat{\boldsymbol{\gamma}}_i\|_2^2\,, \quad (9)$$

where the variables $\hat{\mathbf{x}}$, $\hat{\mathbf{d}}_i$, $\hat{\boldsymbol{\alpha}}_i$, $\hat{\boldsymbol{\beta}}_i$ and $\hat{\boldsymbol{\gamma}}_i$ represent $\mathbf{x}$, $\mathbf{d}_i$, $\boldsymbol{\alpha}_i$, $\boldsymbol{\beta}_i$, $\boldsymbol{\gamma}_i$ in the Fourier domain, respectively, and $\circ$ represent the element-wise (Hadamard) product. The initial algorithm based on this approach exploited the structure of Eq. (9) to decompose it into n independent $k \times k$ linear systems, solving via Gaussian elimination with computational cost $O(k^3 n)$ [24]. In this paper, we adopt a more recent approach [22] that exploits additional structure in the n independent linear systems, giving $O(kn)$ computational cost for solving these systems, and $O(kn \log n + kn)$ cost, dominated by the computation of the FFT, for solving Eq. (6). This approach makes it possible to use a larger number of dictionary filters (k) in the context of source separation.

B. Segmentation and Temporal Continuity

It can be seen from the previous discussion that the computational bottleneck of CSC is now related to the FFT. Instead of using the full-length audio signal as the input, for better efficiency we may want to partition the signal into a number of shorter segments. This can be done either by uniformly segmenting the audio signal with fixed-length windows, or by using automatic, audio-based segmentation methods [71] to ensure the homogeneity of each segment. It is also possible to segment the audio by estimated onset times [72], [73], though this may lead to overly short segments (e.g. shorter than a note). Another advantage of partitioning the audio signal is that in such a segment level we can expect a small number of non-zero elements in the solution of CSC (as compared with the full-length signal), and therefore the CSC problem might be easier to solve.

No matter which method is adopted to segment the signal, we encounter a temporal discontinuity issue because of the circular convolution operation, which is used both in solving the CSC problem (4) and in reconstructing the individual source signals (2). If the reconstruction does not apply circular convolution, extra error would be introduced because of the discrepancy between Eq. (2) and the regression part of Eq. (4). However, when circular convolution is employed, one may find discontinuities in the boundaries of the segments due to the cyclic nature of circular convolution [74, Section 6.3], leading to perceptible impulses. This is a practical issue when applying CSC to the audio domain.

To circumvent this issue, we propose to consider a slightly extended window to perform CSC. Assume that the input audio signal is partitioned into a number of non-overlapping segments of the same length n. To recover the source signals for each of the segments, we take u extra samples before, and after, the segment and use the extended segment to perform CSC. In other words, $\mathbf{x}$, the *nominal window*, is replaced by $\mathbf{x}' \in \mathbb{R}^{n'}$, the *extended window*, in Eq. (4), where $n' = n + 2u$. After computing the estimated source $\mathbf{e}'_j \in \mathbb{R}^{n'}$ by (2), the first u and last u samples are discarded. This simple heuristic ensuring the temporal continuity across the segments is applicable regardless of the segmentation method.[4] In future work we will investigate the application of alternative boundary handling methods [75].

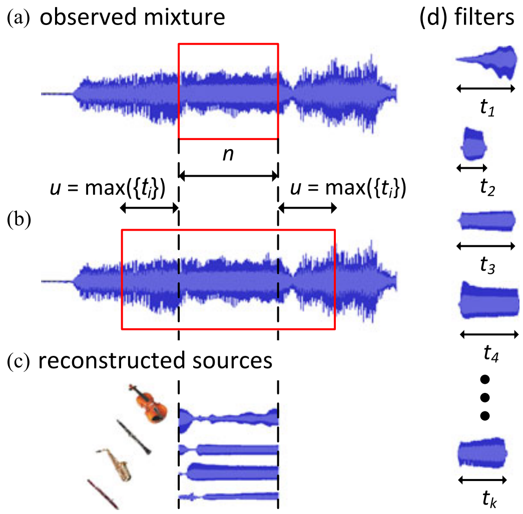


Fig. 2. Illustration of a heuristic that addresses the temporal discontinuity issue when CSC is performed on shorter segments of a musical signal. (a) An n-sample segment of a musical signal; (b) the extended window of length $n' = n + 2u$; (c) the extended n'-sample segment is used for CSC, but only the reconstructions for the middle n samples are taken; (d) the dictionary filters, which should be no longer than u.

In view of the convolution operation in Eq. (4), we can set $u = \max(\{t_i\}_{i\in\mathcal{K}})$, where t_i denotes the length of the ith filter. In other words, u should not be shorter than the longest dictionary filter. An illustration is shown in Fig. 2.

C. CSC-Based Monaural Source Separation

Given an audio segment $\mathbf{x}$, the proposed approach approximates it as the sum of convolutions with a set of pre-learned dictionary filters $\{\mathbf{d}_i\}_{i\in\mathcal{K}}$. As illustrated in Fig. 1, different subsets of the filters are designed to represent the p possible sources. Therefore, we can estimate the source signals $\{\mathbf{e}_j\}_{j=1}^{p}$ that compose the segment in the time domain by Eq. (2).

Given an audio signal that is a temporal concatenation of a number of audio segments, we use CSC to estimate the source signals for each segment, and then align them in time to get the estimated sources of the entire audio signal. The temporal continuity issue can be addressed by the heuristic described in Section III-B.

As we are using the dictionary filters to approximate the input, it is desirable to have the input $\mathbf{x}$ longer than any of the

[4] As the optimization considers the extended window, the errors in the extended parts may lead to errors in $\boldsymbol{\alpha}$. However, we empirically found that the benefit of temporal continuity across the segments outweighs such errors in perceptual quality of the separation result.

filters (i.e. $n \geq \max(\{t_i\}_{i\in\mathcal{K}})$). Consequently, the length of the input used in CSC is usually longer than the usual frame size used in NMF approaches, suggesting that the proposed approach takes into account longer temporal context while performing the separation.

We should also note that, unlike the case of the dictionary vectors in NMF, the dictionary filters in CSC do not need to have the same length [23], as illustrated in Fig. 2(d).

D. Dictionary Learning

In the proposed approach, each dictionary filter must be associated with only one instrument. Therefore, a training dataset of clean source signals for the instruments of interest is needed. In the literature of NMF, pre-training a dictionary is sometimes referred to as a supervised approach [53].

There are two approaches to build the dictionary. The first, *exemplar-based* approach uses the clean source signals directly as the filters [76], [77]. For example, we can treat recordings of each individual note of each instrument (e.g. from A0 to C8 for piano) directly as the dictionary filters [27]. Although it is easy to implement, this approach may not work well in practice because of the possible mismatch between the input signal and the dictionary. For example, a guitarist can express a pitch in different durations, dynamics, and timbres [78]. Exhausting all possible variations of each note is not possible.

The second approach, which is adopted in this paper, uses optimization algorithms to *learn* the dictionary filters using a formulation similar to the CSC formulation (4), thereby improving the generalizability of the dictionary to unseen data. Specifically, the following formulation can be used:

$$\underset{\{\mathbf{d}_i\},\{\boldsymbol{\alpha}_{i,l}\}}{\arg\min}\ \frac{1}{2}\sum_{l\in\mathcal{T}_{q,j}} \left\|\mathbf{z}_l - \sum_{i\in\mathcal{P}_q} \mathbf{d}_i * \boldsymbol{\alpha}_{i,l}\right\|_2^2 + \sum_{l\in\mathcal{T}_{q,j}}\sum_{i\in\mathcal{P}_q}\lambda_i\|\boldsymbol{\alpha}_{i,l}\|_1, \tag{10}$$

where $\mathbf{z}_l$ is a clean source signal in the training set $\mathcal{T}_{q,j}$ for the qth pitch of jth source, and $\{\mathbf{d}_i\}_{i\in\mathcal{P}_q\subset\mathcal{G}_j}$ is the corresponding dictionary filters to be learned. Additional constraint on the norm of the dictionary filters can be applied, for example to account for the possibly varying length of the filters. Formulation (10) can be solved via alternating minimization with respect to the coefficients $\{\boldsymbol{\alpha}_{i,l}\}_{i\in\mathcal{P}_q\subset\mathcal{G}_j, l\in\mathcal{T}_{q,j}}$ and the dictionary $\{\mathbf{d}_i\}_{i\in\mathcal{P}_q\subset\mathcal{G}_j}$, using the algorithm described in [23].

E. Subdictionary Selection in the Score-Informed and MPE-Informed Settings

In the score-informed setting, we can use the score information to choose the subset of dictionary filters $\{\mathbf{d}_i\}_{i\in\mathcal{P}}$, $\mathcal{P}\subseteq\mathcal{K}$ corresponding to the notes that are likely to present in a given segment of the input mixture. Using the smaller subdictionary avoids the possible confusion from similar yet irrelevant dictionary filters (e.g. octave notes from the same instrument or the same pitch from other instruments) and in turn simplifies the separation task [26].

Note that the musical score only helps reduce the size of the dictionary for each n-sample segment (i.e. we know *which* notes occur in the segment). It is still the job of CSC to estimate the activation pattern (i.e. *when* the dictionary filters activate) from the mixture.

In the MPE-informed setting, even if the result of MPE is perfect, we are given only the pitch, onset and duration of the notes. *We do not know which instrument plays which note*, unless additional algorithms for *instrument recognition* and *pitch streaming* are available [30], [37]. Therefore, although we can use the same idea of choosing a subdictionary, we need to consider the filters corresponding to the same pitch from all the instruments in the MPE-informed setting, and the resulting subdictionary can be at most p times larger than the one in the score-informed setting. This will significantly amplify the generalizability issue in separation, as elements of the same pitch from different instruments may work together to approximate the notes observed in the mixture, if the instruments in the testing mixture are not exactly the same as those used for training (e.g. in terms of timbre characteristics). This may cause significant *leakage* in source separation among instruments and lead to poor separation result, as we will see from the result of MPE-informed source separation using a *non-oracle* dictionary in Section VI-B.

In practice, however, there might be errors in MPE. It could be the presence of an extra pitch that does not actually exist (i.e. a false positive), or the absence of an actually presented pitch (i.e. a false negative). How the errors in MPE affect the performance of MPE-informed source separation will be empirically studied also in Section VI-B.

IV. Experimental Setup

This section describes the datasets and performance metrics employed in our evaluation. We will then evaluate source separation under the score-informed and MPE-informed settings in Sections V and VI, respectively.

A. Datasets

We evaluate the proposed algorithm on the Bach10 dataset compiled by Duan and Pardo [79]. The dataset consists of 10 pieces of 4-part J. S. Bach chorales, where the 4 parts (i.e. soprano, alto, tenor and bass) are performed respectively by violin, clarinet, saxophone and bassoon (i.e. $p = 4$), recorded with a sampling rate of 44.10 kHz. The length of the pieces ranges from 25 to 42 seconds, totaling 327 seconds. All the music pieces are composed of the same four instruments. The pitch, instrument, and onset/offset time for each note of the pieces can be obtained from the musical scores of the pieces, which have been aligned with the audio and been included in the Bach10 dataset. The dataset contains 18, 17, 18, and 24 unique pitches (from D2 to A5) for the four instruments, respectively. There are in total 1,957 notes, some of which overlap in time. To simplify our analysis, in case of temporal overlaps between successive notes, we revise the score and set the offset time of the earlier note to the onset time of the later note. The average duration of the notes is 0.68 second, with standard deviation being 0.56 second.

As an alternative source for building the dictionary, we also use the single notes of the four instruments found in the Real

World Computing (RWC) Musical Instrument Sound dataset [80]. Specifically, RWC contains the singe-note samples (i.e. only a series of single notes from an specific instrument is played in each recording) covering the full pitch range of a variety of instruments, recorded with different dynamic levels and playing techniques. We pick the samples corresponding to the 4 instruments used in the Bach10 dataset, with three dynamic levels (*forte*, *mezzo-forte*, and *piano*; meaning loud, moderately loud, and soft) and the normal playing technique. This amounts to 1,590 waveforms of single notes with variable length, totalling 4,310.50 seconds (i.e. 13 times longer than Bach10). There are 46, 40, 33, and 42 unique pitches (from `A#1` to `E7`) respectively for the four instruments, which cover all the pitches found in Bach10.

B. Parameters for Dictionary Learning and CSC

We use single notes segmented from the clean source signals of either Bach10 or RWC in learning the dictionary. Depending on the algorithms being used (see Section V-A for more details), different dictionaries will be learned. For CSC, the algorithm described in Section III-D is used. Unless otherwise specified, we learn $|\mathcal{P}_q| = \kappa = 4$ filters for each pitch per instrument, with the filter length being fixed to 0.10 second ($t_i = 4,410$), $\forall i \in \mathcal{K}$. Each dictionary filter $\mathbf{d}_i$ is normalized by the l_2 norm. The resulting dictionary size is $k = 308$ if Bach10 is used, and $k = 644$ for RWC. According to the instruments, the dictionary filters form 4 groups $\{\mathcal{G}_j\}_{j=1}^{4}$.

Given the dictionary and a test music piece (e.g. one of the 10 chorales), we partition the test piece into 0.25 second, half-overlapping nominal windows (i.e. $n = 11,025$) and then use the score-informed CSC approach described in Section III-E to recover the source signals. To deal with temporal discontinuity issues, an extended window size $n' = n + 2t_i = 19,845$ is used. The values of n and t_i are chosen such that each input segment contains a handful of notes. For Bach10, the partition leads to on average 268 segments per piece. Excluding the 3 segments that have no active notes, there are on average 5.45 active notes (min: 4, max: 9) per segment. Each segment must contain at least one active note from each of the four instruments.

Both dictionary learning and CSC (i.e. Eqs. (10) and (4)) are implemented based on a pre-release version of the SPORCO library [81]. Unless otherwise specified, we empirically set the maximum number of iterations to 250 and 500, and the regularization parameters $\lambda_i, \forall i \in \mathcal{K}$, to 0.05 and 0.01, for dictionary learning and CSC, respectively. For other parameters, we adopt the default setting of SPORCO.[5] We will report a parameter sensitivity test for CSC under the score-informed setting in Section V-C.

Fig. 3 shows the dictionary filters learned by CSC from the RWC dataset for `D4`, the only common pitch shared by the four instruments in both Bach10 and RWC. We can see that the filters corresponding to the same instrument have slightly different shapes in both time and frequency domains. The filters themselves do not reveal much information about the attack, decay, sustain and release parts of the sound, for that can be taken care of by varying amplitudes in the activation pattern $\{\boldsymbol{\alpha}_i\}_{i=1}^{k}$. We also find that the length of the filters can include from 7 (for `D2`) to 88 (for `A5`) periods of a pitch, which may provide sufficient temporal context to model the pitches. Although it is interesting to use varying filter lengths for different pitches, we leave this as a future work.

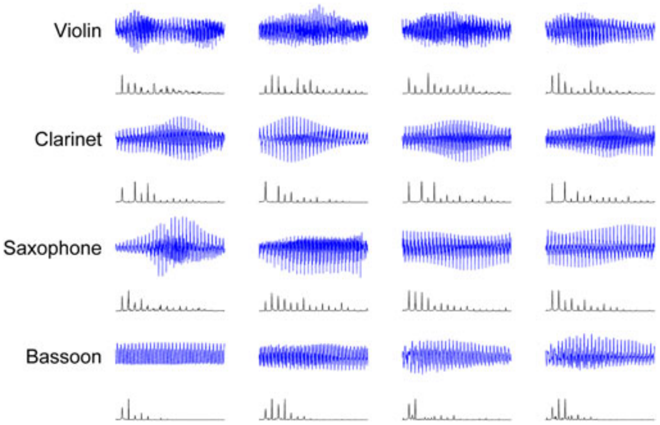


Fig. 3. The waveform (upper) and magnitude spectrum (lower) for the 4 dictionary filters (from left to right) learned by CSC from the RWC dataset for the pitch `D4` for 4 instruments. The filter length is fixed to 0.10 second.

C. Performance Metrics

We evaluate the separation for each of the 10 pieces of Bach10 and report the average result. The separation quality is measured in terms of the source to distortion ratio (SDR), source to interferences ratio (SIR), and source to artifact ratio (SAR), which account for all types of error, the interference error, and artifact error, respectively [82]. These three are standard metrics in source separation and can be computed by the Blind Source Separation Eval (BSS_Eval) toolbox v3.0 [83]. Larger values indicated better separation performance. As these metrics may not perfectly reflect the perceptual separation quality, we also provide in an accompanying website[6] the audio files of the separation results for subjective evaluation.

V. Evaluation of Score-Informed Source Separation

A large number of source separation algorithms have been proposed in the literature, such as the CMF [8], [9], convolutive NMF [84], and shift-invariant PLCA [85]. As the main purpose of our experiment is to evaluate the effectiveness of the time-domain approach CSC against frequency-domain approaches, we only consider two well-known frequency-domain approaches in our evaluation: score-informed NMF [46], [50] and Soundprism [45]. Please note that we do not intend to compare CSC against existing phase-aware methods such as CMF, because CSC does not explicitly consider phase as an objective in the optimization process.

[5] The ADMM parameter ρ is empirically initialized to 50, and is adjusted every 10 iterations to balance the associated primal and dual residuals [23].

[6] [Online] http://mac.citi.sinica.edu.tw/research/CSC_separation/

A. Methods for Score-Informed Source Separation

The following methods are considered:

1) *Score-informed CSC + oracle dictionary*: As a synthetic case implemented to test the possible upper bound performance of the CSC approach, we assume that the score is given and perfectly aligned, and that there is no mismatch between the test signals and the dictionary. This *oracle* dictionary is built from the clean source signals of the Bach10 dataset.
2) *Score-informed CSC + RWC dictionary*: A more realistic setting is to learn the dictionary from the RWC dataset. Due to the possible mismatch between the test signals and the dictionary, we experiment with using $\kappa = 4$, 8, and 12. As RWC includes the full pitch ranges of the instruments, the dictionary size is larger (e.g. $k = 644$ when $\kappa = 4$) and therefore we empirically use a smaller value for the regularization parameter $\lambda_i = 0.003, \forall i \in \mathcal{K}$.
3) *Score-informed NMF*: We adjust the score-informed NMF approach of Ewert and Müller [46] as the baseline. In this frequency-domain approach, the dictionary $\mathbf{W}$ is learned dynamically from the test music signal *itself* (i.e. not using a pre-learned dictionary), with score-informed constraints imposed on both the dictionary vectors and the segment-level (for the same basis point as CSC) activation patterns. Similar to score-informed CSC, we learn $\kappa = 4$ spectral templates for each pitch per instrument. The STFT uses Hamming window of 4,096 samples, with 50% overlapping. The Kullback-Leibler divergence is employed as the cost function, and the maximum number of iterations is 1,000.
4) *Score-informed NMF + oracle/RWC dictionary*: This variant of score-informed NMF uses a pre-learned dictionary $\mathbf{W}$ instead of learning it on-the-fly. Score information is only used to set constraints on the activation patterns. Similar to CSC, we learn a dictionary either from the Bach10 single notes or the RWC single notes. As this setting usually converges earlier, the maximum number of iterations is set to 500.
5) *Soundprism*: The last baseline we consider is the Soundprism approach proposed by Duan and Pardo [45]. It is also a frequency-domain approach, but it uses a parametric approach to model the magnitude spectrum of each pitch, taking into account possible overlapping harmonics across pitches. The source codes of Soundprism are available online [45].

For both Soundprism and NMF, we recover the time-domain signals from the separated spectrograms by using the Wiener filter [4]. For a fair comparison, the parameters of both NMF and Soundprism have been properly tuned.

B. Result of Score-Informed Source Separation

The result is shown in Table I, where we use score-CSC and score-NMF as a shorthand for score-informed CSC and NMF, respectively. We show the averaged SDR, SIR, and SAR over the ten chorales, for each of the 4 instruments (denoted as *Vln*, *Cla*, *Sax* and *Bsn*, respectively) and the average (i.e. *All*) of them. In case we need to test whether there is a significant performance difference between two methods, we perform a one-tailed non-parametric paired sign test between the per-piece average SDRs computed over the four instruments (i.e. *All*).[7] The following observations can be made:

1) From the first four rows, we see remarkably better CSC results are obtained using the oracle dictionary as opposed to the ones learned from RWC.
Increasing the number of dictionary filters improves the result, but using RWC dictionary with $\kappa = 12$ is still inferior to using the oracle dictionary with $\kappa = 4$, and the performance difference is significant (p-value < 0.001) for averaged SDR. This shows *the importance of reducing the mismatch between the dictionary and test signals*.
2) For NMF, the oracle dictionary (6th row) also performs consistently better than using a dictionary from RWC, for all the three metrics and all the instruments. We find that the oracle dictionary with $\kappa = 4$ (6th row) performs significantly better than the RWC dictionary with $\kappa = 12$ (7th row). Learning the dictionary on-the-fly (5th row) generally performs worse than using a pre-learned dictionary, *suggesting the benefit of learning a dictionary from a larger, external dataset*.
3) By comparing all the rows, we can see that CSC generally performs better than the prior arts NMF [46] and Soundprism [45] in most of the instruments and metrics. CSC using the oracle dictionary (1st row) performs the best, with significant performance difference over either NMF using the oracle dictionary (6th row) or Soundprism (10th row). In the realistic setting where the dictionary is learned from RWC, CSC still performs slightly better than NMF (i.e. considering either the (4th, 9th), (3rd, 8th) or (2nd, 7th) row-pair). The performance difference is significant (p-value < 0.05) for $\kappa = 12$. This probably suggests that the magnitude of frequency can be well presented by only a few templates already, so there is little gain using more templates. In contrast, *it is useful to use more dictionary filters in the time domain*. We also examine whether CSC+RWC or NMF+RWC is better than Soundprism and find that only CSC+RWC with $\kappa = 12$ is significantly better (p-value < 0.05). Overall, these comparisons *demonstrate the effectiveness of CSC*.
4) By comparing the columns, we see that all the considered methods perform the worst for the bassoon and the best for violin, among the four instruments. Noting that the bassoon plays the lowest part of the chorales and the violin plays the highest, it seems that the considered methods do not work well for low-frequency components. It might be possible to address this issue in CSC by setting different values of λ_i for different instruments, but we do not explore this further in this paper.

Fig. 4 shows the clean sources (ground truth) and the recovered ones from the mixture by score-informed CSC using the RWC dictionary ($\kappa = 4$), for a segment of Bach10 that contains

[7] The t-test is deemed inappropriate here as the distribution of data does not pass Kolmogorov-Smirnov test. i.e., the distribution is unlikely to be Gaussian.

TABLE I
RESULT OF SCORE-INFORMED SOURCE SEPARATION

	Method	SDR					SIR					SAR				
		Vln	Cla	Sax	Bsn	All	Vln	Cla	Sax	Bsn	All	Vln	Cla	Sax	Bsn	All
1	score-CSC + oracle dictionary ($\kappa = 4$)	10.49	8.66	9.30	5.37	8.45	23.87	18.32	16.31	9.23	16.94	10.72	9.34	10.45	8.24	9.69
2	score-CSC + RWC dictionary ($\kappa = 12$)	7.53	7.28	7.17	3.18	6.29	16.93	16.80	14.37	6.71	13.70	8.17	7.95	8.27	6.63	7.76
3	score-CSC + RWC dictionary ($\kappa = 8$)	7.22	7.05	7.15	3.09	6.13	16.73	16.78	14.44	6.69	13.66	7.85	7.70	8.24	6.48	7.57
4	score-CSC + RWC dictionary ($\kappa = 4$)	6.80	7.00	6.68	2.96	5.86	16.13	16.02	13.77	6.81	13.18	7.47	7.76	7.82	6.15	7.30
5	score-NMF [46]	5.41	7.63	6.55	2.62	5.55	16.47	15.05	13.51	7.27	13.07	5.88	8.77	7.76	5.23	6.91
6	score-NMF + oracle dictionary ($\kappa = 4$)	8.03	8.02	7.55	4.59	7.05	17.58	17.23	15.88	10.08	15.19	8.64	8.74	8.37	6.47	8.05
7	score-NMF + RWC dictionary ($\kappa = 12$)	6.74	6.86	6.73	3.10	5.86	15.66	15.91	15.17	7.86	13.65	7.50	7.57	7.54	5.59	7.05
8	score-NMF + RWC dictionary ($\kappa = 8$)	6.66	6.90	6.60	3.06	5.80	15.77	16.02	14.71	7.74	13.56	7.38	7.61	7.48	5.60	7.02
9	score-NMF + RWC dictionary ($\kappa = 4$)	6.25	7.03	6.32	3.04	5.66	15.14	15.34	14.32	7.91	13.18	7.02	7.91	7.24	5.49	6.91
10	Soundprism [45]	6.15	6.83	5.98	2.47	5.36	11.92	11.78	11.22	9.63	11.14	7.87	8.95	7.98	3.88	7.17

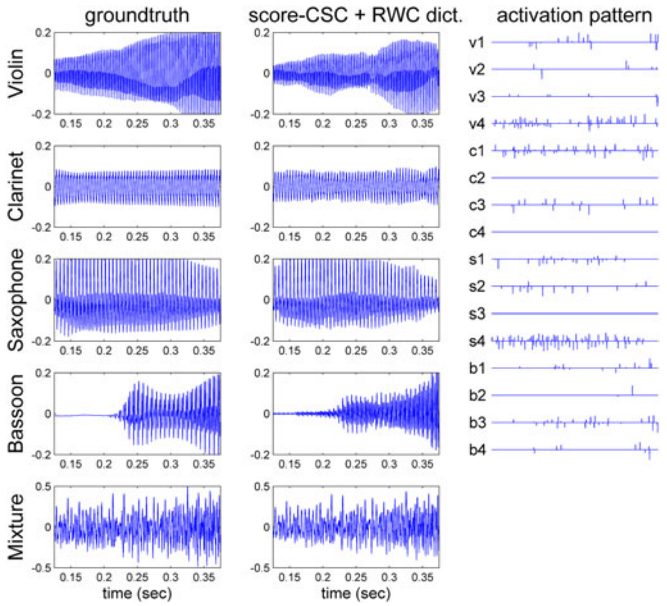


Fig. 4. The clean sources (left) and the recovered ones (middle) from the mixture by score-informed CSC using the RWC dictionary ($\kappa = 4$) for a segment of Bach10 that contains four notes, along with the corresponding activation patterns (right).

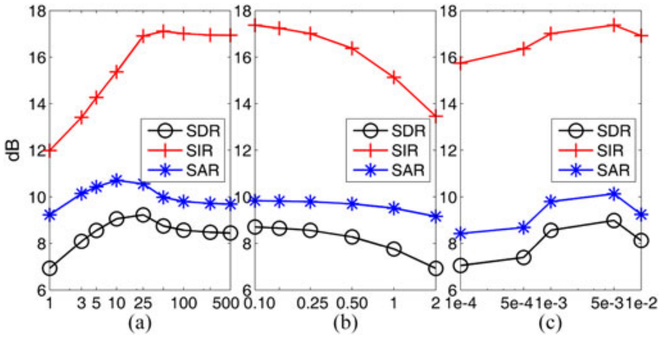


Fig. 5. Performance of 'score-informed CSC + oracle dictionary ($\kappa = 4$)' as we vary (a) the maximum number of iterations in CSC, (b) the length of the nominal window (in second), and (c) the regularization parameter λ.

a D4 from violin, an A3 from clarinet, an F3 from saxophone, and a D3 from bassoon.[8]

C. Parameter Sensitivity Test

Fig. 5 shows how the performance of score-informed CSC varies as a function of three parameters, fixing the dictionary to the oracle one ($\kappa = 4$). From the leftmost plot, we see that the performance of CSC actually reaches a plateau after 25 iterations, and there is little gain going further. Therefore, although we have set the maximum number of iterations to 500 in our previous experiments to ensure convergence, for shorter runtime one can use fewer iterations.[9] The plateauing of the performance of the algorithms at a low number of iterations has also been observed in NMF-based algorithms [86].

The middle plot of Fig. 5 investigates the effect of the length n of the nominal window. We can see improved result in all the three metrics (especially in SIR) as n decreases, but the result saturates after n is smaller than 0.25 second. We conjecture that a shorter nominal window gives better results because of the reduced complexity of the segment being processed and the more accurate constraint on the activation pattern (i.e. selection of the subdictionary). However, while the use of shorter nominal window is more accurate and the demand on memory is lower, it is more time consuming than using a long nominal window, due to the increased number of segments. Therefore, there is a trade-off.

Finally, the rightmost plot of Fig. 5 shows the effect of the regularization parameter and suggests we can have slightly better result in all the three performance metrics by setting λ to 0.005. Setting the value of λ too small degrades the result.

[8] It is the 2nd segment (i.e. 0.125–0.375 second) of the 4th piece in Bach10. Recovering the sources from the mixture (shown in the bottom left corner) might be challenging, due to the harmonic relations between the notes: e.g. a perfect fifth (D3–A3) and an octave (D3–D4). Despite the challenge, given the score, CSC can recover the sources quite well. We can for example observe the energy increase in the attack part of the bassoon note in the separation result. We also see from the activation patterns how CSC combines the dictionary filters, using both positive and negative weights, to create the separation result.

[9] We have included a naïve GPU implementation of CSC in SPORCO [81], which can further boost the time efficiency by an order of magnitude.

TABLE II
RESULT OF MPE-INFORMED SOURCE SEPARATION

		SDR					SIR					SAR				
	Oracle dictionary ($\kappa = 4$)	Vln	Cla	Sax	Bsn	All	Vln	Cla	Sax	Bsn	All	Vln	Cla	Sax	Bsn	All
1	oracle MPE + CSC	8.24	2.49	5.23	2.82	4.69	15.74	6.33	9.86	5.21	9.28	9.48	5.86	7.56	8.10	7.75
2	CFP-based MPE [28] + CSC	8.23	0.94	4.82	2.09	4.02	16.57	5.51	9.53	5.03	9.16	9.22	4.07	7.13	6.54	6.74
3	C-NMF-based MPE [29] + CSC	8.04	1.87	5.19	2.16	4.32	15.44	5.39	9.82	4.64	8.82	9.30	5.65	7.51	7.41	7.47
4	oracle MPE + NMF	7.22	4.80	5.40	3.76	5.29	15.19	9.63	12.50	8.54	11.46	8.12	7.10	6.62	6.14	7.00
5	CFP-based MPE [28] + NMF	7.05	2.96	4.70	2.67	4.35	14.65	8.26	11.89	7.66	10.62	8.05	5.24	5.96	5.06	6.08
6	C-NMF-based MPE [29] + NMF	6.83	3.90	4.92	3.41	4.76	13.15	7.59	11.64	8.36	10.18	8.23	7.12	6.31	5.73	6.85
	RWC dictionary ($\kappa = 4$)															
7	oracle MPE + CSC	0.69	−1.51	−3.31	−0.30	−1.11	4.65	1.26	−1.69	2.26	1.62	4.26	4.35	5.81	5.31	4.93
8	CFP-based MPE [28] + CSC	0.34	−2.55	−3.99	−1.20	−1.85	4.58	0.20	−2.29	1.33	0.95	3.73	3.85	5.30	4.83	4.43
9	oracle MPE + NMF	1.79	−1.19	−3.73	−1.17	−1.08	6.76	1.59	−1.49	1.40	2.07	4.32	4.52	4.23	4.82	4.47
10	CFP-based MPE [28] + NMF	1.49	−2.60	−3.80	−1.80	−1.68	6.54	0.33	−1.36	0.89	1.60	4.03	3.46	3.78	4.28	3.89

VI. EVALUATION OF MPE-INFORMED SOURCE SEPARATION

When the score is not given, MPE can be used as a preprocessing step to judiciously reduce the size of the dictionary. This section continues to evaluate the performance of CSC and NMF under the MPE-informed setting, using either the oracle or the RWC dictionary. For simplicity, we consider only the case $\kappa = 4$ in this evaluation.

A. Methods for MPE

We consider the following three methods for MPE:

1) *Oracle MPE*: This is a synthetic setting that assumes the result of MPE is perfect. The only difference between this case and the score-informed setting is that we are now not aware of the *instrument labels* of the notes (i.e. which instrument plays each note). As discussed in Section III-E, comparing with the score-informed setting, the subdictionary can be p times larger.
2) *CFP-based MPE*: We consider two audio-based MPE algorithms in this evaluation. The first one is the combined frequency and periodicity (CFP) method proposed by Su and Yang [28]. The basic idea is to exploit the commonality between a harmonic series formed in the frequency domain and a sub-harmonic series formed in the quefrency (i.e. lag) domain to identify pitches. The frame-level estimate of MPE is aggregated to the note-level by using a moving median filter of 0.25 s in length [28]. Using the parameter settings suggested in [28] can already perform well for MPE in the Bach10 dataset, reaching 85.51%, 85.80% and 85.22% for frame-level F-score, precision and recall, respectively.
3) *C-NMF-based MPE*: The other one is the constrained NMF (C-NMF) algorithm proposed by Vincent *et al.* [29].[10] If the parameters of this algorithm is properly tuned, the frame-level F-score can reach 79.78%, which is not that inferior to the result of CFP. However, to make the result of CFP and C-NMF more different, we deliberately tune the parameters of C-NMF[11] toward *high recall rate*, i.e., allowing for false positives (extra notes detected) rather than false negatives (miss of true notes). The resulting frame-level F-score, precision, and recall are 64.47%, 49.30%, and 93.14% respectively. Please note that a precision rate close to 50% means that half of the MPE estimates are redundant and do not correspond to real notes.

B. Result of MPE-Informed Source Separation

Result shown in Table II leads to the following observations:

1) Comparing with Table I, we see that using the MPE estimates instead of the scores largely degrades the separation quality, which may have been expected. The performance difference between score-informed CSC (1st row in Table I) and MPE-informed CSC (first 3 rows in Table II), when using the oracle dictionary, is around 4 dB in average SDR (significant difference; p-value < 0.001). As the main difference between oracle MPE and score is the availability of note-level instrument labels, this result *suggests that automatic instrument recognition is needed to close the performance gap*. When the instrument labels of the notes are not informed, it is possible that CSC will reconstruct a source using dictionary filters corresponding to other instruments. This can be seen from Fig. 6, which shows that the bassoon note recovered by MPE-informed CSC resembles the ground truth violin note, which is actually one octave higher.
2) Comparing the first 3 rows of Table II shows that, despite having different precision rates, MPE-informed CSC using any of the three MPE methods leads to similar performance. As they all have high recall rates, this may suggest that *the recall rate of MPE is more important than the precision rate for MPE-informed CSC*. However, the sign test reveals that the result of oracle MPE (1st row) is significantly better (p-value < 0.001) than the result of the other two methods (2nd and 3rd rows), suggesting that *the precision rate of MPE also matters*.

[10] Since we can also group the dictionary vectors according to pitch rather than instrument, we can also use NMF for MPE.

[11] Specifically, we set the peak-picking threshold to 32 and $\beta = 0.5$ [29].

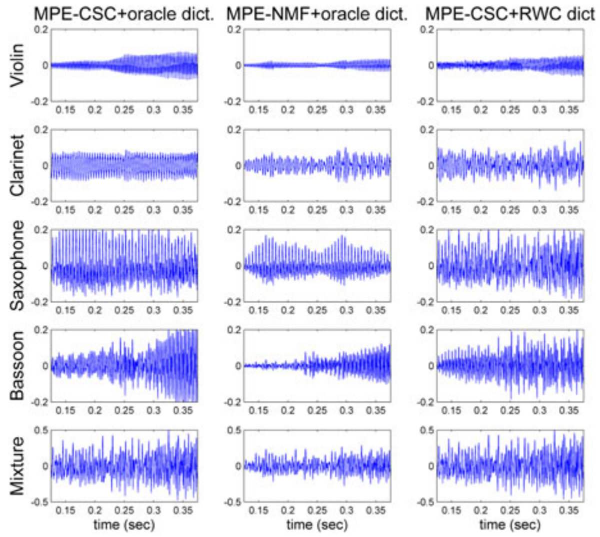


Fig. 6. Separation results of (from left to right) MPE-informed CSC with the oracle dictionary, MPE-informed NMF with the oracle dictionary, and MPE-informed CSC with the RWC dictionary for the same segment of Bach10 as Fig. 4. The MPE method considered here is CFP.

3) From rows 4–6 in Table II, we see that, when the oracle dictionary is used, MPE-informed NMF performs slightly better than MPE-informed CSC in SDR and SIR, but not in SAR. With oracle MPE, MPE-informed NMF has 0.60 dB higher average SDR than MPE-informed CSC, though the performance difference is not significant. This finding suggests that *NMF-based methods may not be inferior to CSC in non-score informed scenarios*.
4) Interestingly, the lower SAR of MPE-informed NMF (rows 4–6), as compared with MPE-informed CSC (rows 1–3), suggests *the advantage of the time-domain approach CSC in avoiding perceptual artifacts*.
5) Among the four instruments, clarinet and bassoon suffer remarkably due to the transition from the score-informed to the MPE-informed setting for CSC. In contrast, we *do not see such remarkable degradation for particular instruments in the result of NMF*.
6) Finally, from rows 7–10 in Table II, we see the separation quality becomes even poorer when using the RWC dictionary, instead of the oracle one. Both CSC and NMF yield negative SDR values, but the performance difference between them is not significant.[12] This result shows that *source separation for Bach10 is in general challenging* and proper prior knowledge (e.g. score or training data with similar timbre characteristics) is crucial.

Fig. 6 shows the separation result for three different MPE-informed methods for the same segment of Bach10 used in Fig. 4. From the result of MPE-informed CSC with the RWC dictionary (i.e. the rightmost ones), we see how hard it is to recover the sources. Concurrent notes with strong harmonic relations are common in Bach10.

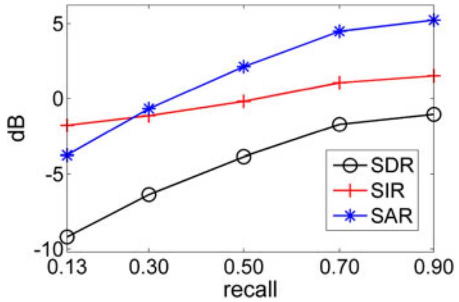


Fig. 7. Performance of MPE-informed CSC as a function of recall rates in MPE, using the RWC dictionary for CSC and the C-NMF-based method for MPE [29]. The F-scores of these five points are 0.23, 0.46, 0.65, 0.77, and 0.67; the one with the highest F-score does not lead to the best separation result.

We further test the importance of the recall rate of MPE by again varying the parameters of C-NMF-based MPE [29]. Result shown in Fig. 7 indicates that *the separation quality in all the three performance metrics improves almost linearly with the recall rate*. The best result is obtained when the recall rate of MPE is the highest, which may not correspond to the parameter setting yielding the highest F-score of MPE.

VII. Conclusion

In this paper, we have presented a novel time-domain approach for monaural music source separation, based on CSC. We reported a comprehensive performance study that assesses the strength of CSC, in both the score-informed and the more practical MPE-informed settings. We also evaluated the case where there is a mismatch in the acoustic content between the dictionary and the test signals. Our results show that, in many cases, the time-domain approach CSC compares favorably with the classic frequency-domain approach NMF. However, when the score is not given, NMF can sometimes perform better than CSC.

Unlike NMF-based methods, the use of CSC in the audio domain is still new and we believe there are many possible ways to further improve the result. For example, we have not optimized the filter length t_i and regularization parameter λ_i for different notes. The dictionary size can be further expanded for better separation result. We can study other cost functions, such as the Itakura Saito divergence [6] for the fidelity term $f(\cdot)$, and the group Lasso [87] or some other temporal constraints for the regularization term $g(\cdot)$ in Eq. (1). As suggested by our experiments, a reliable instrument recognizer for per-note instrument labels can be very helpful to MPE-informed source separation. Another interesting research direction is to train a dictionary only for a specific instrument (e.g. the piano) for automatic transcription of a solo performance [88]. We note that CSC is a generic approach and might be applied to other audio problems such as melody transcription [44]. We also plan to deeply investigate the role of phase in the future. This can be done by systematically perturbing the phase of a source to measure the effect of using the true phase, or by comparing the performance of CSC with existing phase-aware methods.

[12] We do not show in Table II the result of C-NMF-based MPE for the RWC dictionary because that is close to the result of CFP-based MPE.

References

[1] 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.

[2] D. FitzGerald, M. Cranitch, and E. Coyle, "Shifted 2D nonnegative tensor factorisation," in *Proc. IET Irish Signals Syst. Conf.*, 2006, pp. 509–513.

[3] M. N. Schmidt and M. Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Signal Separation*, 2006, pp. 700–707.

[4] B. Wang and M. D. Plumbley, "Investigating single-channel audio source separation methods based on non-negative matrix factorization," in *Proc. ICA Res. Netw. Int. Workshop*, 2006, pp. 17–20.

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

[6] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, 2009.

[7] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1118–1133, May 2012.

[8] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 3437–3440.

[9] J. Bronson and P. Depalle, "Phase constrained complex NMF: Separating overlapping partials in mixtures of harmonic musical sources," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 7475–7479.

[10] F. J. Rodriguez-Serrano, S. Ewert, P. Vera-Candeas, and M. Sandler, "A score-informed shift-invariant extension of complex matrix factorization for improving the separation of overlapped partials in music recordings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 61–65.

[11] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *Proc. Workshop Adv. Models Acoust. Process, NIPS*, 2006. [Online]. Available: http://paris.cs.illinois.edu/pubs/index.html

[12] P. Smaragdis, C. Fèvotte, G. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 66–75, May 2014.

[13] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto, "Beyond NMF: Time-domain audio source separation without phase reconstruction," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2013, pp. 369–374.

[14] S. Ewert, M. D. Plumbley, and M. Sandler, "Accounting for phase cancellations in non-negative matrix factorization using weighted distances," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 649–653.

[15] H. Kameoka, "Multi-resolution signal decomposition with time-domain spectrogram factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 86–90.

[16] T. Yoshioka *et al.*, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, Nov. 2012.

[17] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution," *Commun. Pure Appl. Math.*, vol. 59, pp. 797–829, 2006.

[18] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[19] I. Rish and G. Grabarnik, *Sparse Modeling: Theory, Algorithms, and Applications*, 1st ed. Boca Raton, FL, USA: CRC, 2014.

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

[21] J. Yang, K. Yu, and T. S. Huang, "Supervised translation-invariant sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 3517–3524.

[22] B. Wohlberg, "Efficient convolutional sparse coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 7223–7227.

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

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

[25] B. Kong and C. C. Fowlkes, "Fast convolutional sparse coding (FCSC)," Univ. California, Irvine, CA, USA, Tech. Rep., 2014. [Online]. Available: https://github.com/bkong/FCSC

[26] 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.*, 2015, pp. 236–240.

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

[28] 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.

[29] 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.

[30] Z. Duan, J. Han, and B. Pardo, "Multi-pitch streaming of harmonic sound mixtures," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 138–150, Jan. 2014.

[31] J. F. Woodruff, B. Pardo, and R. B. Dannenberg, "Remixing stereo music with score-informed source separation," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2006, pp. 314–319.

[32] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrument equalizer for query-by-example retrieval: Improving sound source separation based on integrated harmonic and inharmonic models," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2008, pp. 133–138.

[33] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2005, pp. 337–344.

[34] J.-L. Durrieu, G. Richard, and B. David, "An iterative approach to monaural musical mixture de-soloing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 105–108.

[35] T.-S. Chan *et al.*, "Vocal activity informed singing voice separation with the iKala dataset," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 718–722.

[36] T.-S. Chan and Y.-H. Yang, "Complex and quaternionic principal component pursuit and its application to audio separation," *IEEE Signal Process. Lett.*, vol. 23, no. 2, pp. 287–291, Feb. 2016.

[37] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2009, pp. 327–332.

[38] J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera, "A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2012, pp. 559–564.

[39] C.-Y. Sha, Y.-H. Yang, Y.-C. Lin, and H. H. Chen, "Singing voice timbre classification of chinese popular music," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 734–738.

[40] Y. Ikemiya, K. Yoshii, and K. Itoyama, "Singing voice analysis and editing based on mutually dependent F0 estimation and source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 574–578.

[41] J. R. Zapata and E. Gómez, "Using voice suppression algorithms to improve beat tracking in the presence of highly predominant vocals," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2013, pp. 51–55.

[42] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 529–540, Mar. 2008.

[43] E. Gómez, F. J. Cañadas Quesada, J. Salamon, J. Bonada, P. V. Candea, and P. C. n. Molero, "Predominant fundamental frequency estimation vs singing voice separation for the automatic transcription of accompanied flamenco singing," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2012, pp. 601–606.

[44] 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.

[45] Z. Duan and B. Pardo, "Soundprism: An online system for score-informed source separation of music audio," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1205–1215, Oct. 2011. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5887382&tag=1

[46] S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 129–132.

[47] U. Simsekli, Y. K. Yilmaz, and A. T. Cemgil, "Score guided audio restoration via generalised coupled tensor factorisation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 5369–5372.

[48] C. Joder and B. Schuller, "Score-informed leading voice separation from monaural audio," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2012, pp. 277–282.

[49] J. Fritsch and M. D. Plumbley, “Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 888–891.

[50] S. Ewert, B. Pardo, M. Müller, and M. Plumbley, “Score-informed source separation for musical audio recordings: An overview,” *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 116–124, May 2014.

[51] P. Smaragdis and G. J. Mysore, “Separation by humming: User-guided sound extraction from monophonic mixtures,” in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2009, pp. 69–72.

[52] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, “Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 257–260.

[53] N. J. Bryan and G. J. Mysore, “Interactive refinement of supervised and semi-supervised sound source separation estimates,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 883–887.

[54] Y. Li, J. Woodruff, and D. Wang, “Monaural musical sound separation based on pitch and common amplitude modulation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1361–1371, Sept. 2009.

[55] E. C. Cerón, “Pitch-informed solo and accompaniment separation,” Ph.D. dissertation, Fakultät für Elektrotechnik und Inform.technik, Technische Universität Ilmenau, Ilmenau, Germany, 2014.

[56] N. Souviraa-Labastie, A. Olivero, E. Vincent, and F. Bimbot, “Multichannel audio source separation using multiple deformed references,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 11, pp. 1775–1787, Nov. 2015.

[57] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.

[58] B. King and L. Atlas, “Single-channel source separation using simplified-training complex matrix factorization,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4206–4209.

[59] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan,“The delta-phase spectrum with application to voice activity detection and speaker recognition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2026–2038, Sep. 2011.

[60] L. Su, H.-M. Lin, and Y.-H. Yang, “Sparse modeling of magnitude and phase-derived spectra for playing technique classification,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 2, no. 12, pp. 2122–2132, Dec. 2014.

[61] J. L. Roux, H. Kameoka, N. Ono, and S. Sagayama, “Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency,” *Proc. 13th Int. Conf. Digital Audio Effects*, Graz, Austria, 2010.

[62] D. Gunawan and D. Sen, “Iterative phase estimation for the synthesis of separated sources from single-channel mixtures,” *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 421–424, May 2010.

[63] V. Gnann and M. Spiertz, “Multiresolution STFT phase estimation with frame-wise posterior window length decision,” in *Proc. Int. Conf. Digit. Audio Effects*, 2011, pp. 101–106.

[64] N. Sturmel, L. Daudet, and L. Girin, “Phase-based informed source separation for active listening of music,” *Proc. Int. Conf. Digital Audio Effects*, York, U.K., 2012.

[65] M. S. Lewicki and T. J. Sejnowski, “Coding time-varying signals using sparse, shift-invariant representations,” *Adv. Neural Inf. Process. Syst.*, vol. 11, pp. 730–736, 1999.

[66] 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.

[67] 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.

[68] M. Mørup, M. N. Schmidt, and L. K. Hansen, “Shift invariant sparse coding of image and music data,” Tech. Univ. Denmark, Kongens Lyngby, Denmark, Tech. Rep. IMM2008-04659, 2008.

[69] 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.

[70] 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. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[71] J. Paulus, M. Müller, and A. Klapuri, “State of the art report: Audio-based music structure analysis,” in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2010, pp. 625–636.

[72] J. Schlüter and S. Böck, “Improved musical onset detection with convolutional neural networks,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 6979–6983.

[73] C.-Y. Liang, L. Su, and Y.-H. Yang, “Musical onset detection using constrained linear reconstruction,” *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 2142–2146, Nov. 2015.

[74] M. Pedersen, J. Larsen, U. Kjems, and L. Parra, “A survey of convolutive blind source separation methods,” in *Springer Handbook of Speech Processing*. Berlin, Germany: Springer-Verlag, 2007, pp. 1065–1084.

[75] B. Wohlberg, “Boundary handling for convolutional sparse representations,” *Proc. IEEE Int. Conf. Image Processing*, Phoenix, AZ, USA, Sep. 2016.

[76] T. N. Sainath *et al.*, “Exemplar-based processing for speech recognition: An overview,” *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 98–113, Nov. 2012.

[77] P.-K. Jao and Y.-H. Yang, “Music annotation and retrieval using unlabeled exemplars: correlation and sparse codes,” *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1771–1775, Oct. 2015.

[78] Y.-P. Chen, L. Su, and Y.-H. Yang, “Electric guitar playing technique detection in real-world recordings based on F0 sequence pattern recognition,” in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2015, pp. 708–714.

[79] Bach10 dataset, 2011. [Online]. Available: http://music.cs.northwestern.edu/data/Bach10.html

[80] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “RWC music database: Music genre database and musical instrument sound database.” in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2003, vol. 3, pp. 229–230. [Online]. Available: https://staff.aist.go.jp/m.goto/RWC-MDB/

[81] B. Wohlberg, “SParse Optimization Research COde (SPORCO),” version (Matlab) 0.0.3, 2016. [Online]. Available: http://math.lanl.gov/brendt/Software/SPORCO/

[82] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[83] “BSS_Eval: A toolbox for performance measurement in (blind) source separation,” (2010). [Online]. Available: http://bass-db.gforge.inria.fr/bss_eval/

[84] P. Smaragdis, “Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs,” in *Independent Component Analysis and Blind Signal Separation*. Berlin, Germany: Springer-Verlag, 2004, pp. 494–499.

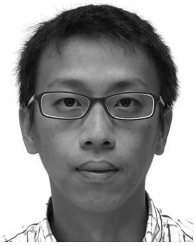
[85] E. Benetos and S. Dixon, “Multiple-instrument polyphonic music transcription using a convolutive probabilistic model,” in *Proc. 8th Sound Music Comput. Conf.*, 2011, pp. 19–24.

[86] D. Fitzgerald and R. Jaiswal, “On the use of masking filters in sound source separation,” *Proc. 15th Int. Conf. Digital Audio Effects*, York, U.K., 2012.

[87] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *J. Roy. Statist. Soc., B*, vol. 68, pp. 49–67, 2006.

[88] A. Cogliati, Z. Duan, and B. Wohlberg, “Piano music transcription with fast convolutional sparse coding,” in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2015, pp. 19–24.

Ping-Keng Jao received the B.Sc. and M.Sc. degrees in electrical engineering from National Cheng Kung University, Tainan, Taiwan, in 2009 and 2011, respectively. From 2013 to July 2015, he was as a Research Assistant in the Research Center for Information Technology Innovation, Academia Sinica, Taiwan. Since February 2016, he is working toward the Ph.D. degree at the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland and working on brain–computer interface.



Li Su (S'08–M'13) received the Ph.D. degree in communication engineering from National Taiwan University, Taipei, Taiwan, in 2012. Since 2012, he has been a Postdoctoral Research Fellow in the Research Center for Information Technology Innovation, Academia Sinica, Taiwan. In 2014, he was the Tutorial Chair of International Society for Music Information Retrieval Conference. In Spring 2016, he was an Adjunct Assistant Professor in the Department of Computer Science, National Tsing-Hua University, Hsinchu, Taiwan. His research interests include time–frequency analysis, machine learning, music information retrieval, biomedical signal processing, and microwave circuit design.

Yi-Hsuan Yang (M'11) received the Ph.D. degree in communication engineering from National Taiwan University, Taipei, Taiwan, in 2010. Since 2011, he has been affiliated with the Academia Sinica Research Center for IT Innovation, where he is currently an Associate Research Fellow. He served as a Technical Program Chair of the International Society for Music Information Retrieval Conference, in 2014, and an Associate Editor of the *IEICE Transactions on Information and Systems*, since 2016. His research interests include music information retrieval, machine learning, multimedia, and affective computing.

Brendt Wohlberg received the B.Sc.(Hons.) degree in applied mathematics, and the M.Sc. (Applied Science) and Ph.D. degrees in electrical engineering all from the University of Cape Town, 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. From 2010 to 2014, he was an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, and is currently the Chair of the Computational Imaging Special Interest Group of the IEEE Signal Processing Society and an Associate Editor for the IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING. His research interests include signal and image processing inverse problems, with an emphasis on sparse representations and exemplar-based methods.