INFORMED MONAURAL SOURCE SEPARATION OF MUSIC BASED ON CONVOLUTIONAL SPARSE CODING

Ping-Keng Jao[†] *Yi-Hsuan Yang*[†] *Brendt Wohlberg*[*]

[†]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[*]Theoretical Division, Los Alamos National Laboratory, USA

ABSTRACT

Monaural source separation is a challenging problem that has many important applications in music information retrieval. In this paper, we focus on the score-informed variant of this problem. While non-negative matrix factorization and some other approaches have been shown effective, few existing approaches have properly taken the phase information into account. There are unnatural sound in the separation result, as the phase of each source signal is considered equivalent to the phase of the mixed signal. To remedy this, we propose to perform source separation directly in the time domain using a convolutional sparse coding (CSC) approach. Evaluation on the Bach10 dataset shows that, when the instrument, pitch and onset/offset time are informed, the source to distortion ratio of the separation result reaches 8.59 dB, which is 2.02 dB higher than a state-of-the-art system called Soundprism.

Index Terms— Convolutional sparse coding, dictionary learning, score-informed monaural source separation

1. INTRODUCTION

The problem of source separation has many variants in music, such as the separation of the leading instrument from the accompaniments [1], isolating the singing voice [2–4], and the separation of all the instruments involved in a music piece from one another [5]. Sometimes source separation is considered as a pre-processing step of the subsequent music signal analysis, sometimes the result of separation is directly used in applications such as automatic Karaoke.

As monaural source separation is challenging, the use of side information is considered necessary for better separation result. Even a user-guided separation system with intense user feedback [6, 7] is an acceptable and emerging solution. We consider in this paper the case where the score of the target music piece is given, but without additional user involvement. Such a score-informed scenario has received increasing attention in recent years, because the scores for some music performances may not be difficult to obtain [8–10]. However, existing approaches usually operate on the magnitude part of the spectrogram, leaving phase information unexploited. To generate the separated time-domain signals for each source, these approaches simply use the phase of the mixture as the phase of each source, leading to perceptible unnatural sound. Some approaches have been proposed to alleviate this "phase copy" issue, but were not tested in a real recording with the score informed [11–13]. Our proposal is to perform separation directly in the time domain using convolutional sparse coding (CSC) [14, 15], which assumes that the mixture can be reconstructed by the sum of a set of convolutions with time-domain dictionary filters. As phase information is preserved, this approach can result in better perceptual quality.

In what follows, we first review several related works, and then define the problem. We study how to perform source separation with CSC by building a semantically meaningful dictionary. Finally, we report experimental results illustrating the phase copy issue and validating the effectiveness of CSC.

2. RELATED WORKS

Many prior works on informed source separation only consider the magnitude of short time Fourier transform (STFT). For example, Ewert and Müller employed non-negative matrix factorization (NMF) to separate the left/right hand of a piano performance with score information [8]. The formulation is $\mathbf{X}_F \stackrel{def}{=} |\mathbf{STFT}(\mathbf{x})| \approx \mathbf{WH} \stackrel{def}{=} \mathbf{Y}$, where $\mathbf{X}_F \in \mathbb{R}_{\geq 0}^{f \times n}$ is the magnitude of STFT on the observed waveform $\mathbf{x} \in \mathbb{R}^n$, and $\mathbf{X}$ is approximated by the product of $\mathbf{W} \in \mathbb{R}_{\geq 0}^{f \times q}$ and $\mathbf{H} \in \mathbb{R}_{\geq 0}^{q \times n}$, and $q \ll \min(f, n)$. $\mathbf{W}$ stands for the q templates appeared in the observation. For example, a column stands for the frequency distribution of a pitch of the piano, and its time activation is given in the corresponding row of $\mathbf{H}$. By initializing $\mathbf{W}$ and $\mathbf{H}$ properly according to a well-annotated MIDI file (which served as the score), they achieved decent separation quality. The separation process is

$$\hat{X}^j_{F,lm} = \frac{W_{lj}H_{jm}}{\sum_j W_{lj}H_{jm}} X_{F,lm} = \frac{Y^j_{lm}}{Y_{lm}} X_{F,lm}, \qquad (1)$$

where the subscripts, l, j, m, index the elements of the corresponding matrices. Eq. 1 is a Wiener filter that redistributes each time-frequency bin to the j^{th} component (i.e. source).

This work was supported by the Academia Sinica Career Development Program and the Ministry of Science and Technology of Taiwan.

As the second example, Duan and Pardo proposed an online system called *Soundprism* that performs source separation based on a score follower, with score given by a MIDI file [5]. Similar to, but slightly different from an NMF approach, Soundprism also reconstructs each separated signal by properly splitting the magnitude of each time-frequency bin of $\mathbf{X}_F$ to corresponding channels according to found pitches.

Although both approaches yield good quality in reconstruction, they suffer from the phase copy issue as only the magnitude part of the spectrogram is considered. Both approaches require copying the phase from $\mathbf{STFT}(\mathbf{x})$ for reconstructing all the sources, but this is evidently sub-optimal for 1) the phase of different sources would not be identical, and 2) phase usually carries important timbre information [16]. The resulting unnatural sound, despite being subtle at times, may not be acceptable to a critical ear.

Recently, Yoshii *et al.* [11] proposed an extension of NMF called log-determinant positive semidefinite tensor factorization (PSDTF). While NMF is not applicable to time-domain signals because of the presence of negative values, PSDTF is applicable because the outer product of a time-domain signal is a rank-1 PSD matrix that can be used as the input. Accordingly, the phase copy issue is circumvented. They obtained superior quality in simple synthetic data, but they did not report the result on real data due to extremely high computational cost [11]. There are some other NMF-based approaches that consider phase, but they were either tested on simple synthetic waveforms or synthetic piano recordings [12, 13].

Sparse coding (SC) is another component analysis tool akin to NMF [17, 18]. We will consider the specific SC form

$$\underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{x}_{F,i} - \mathbf{D}_F\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1\,, \tag{2}$$

where $\mathbf{x}_{F,i}$ is the i^{th} column of $\mathbf{X}_F$, $\mathbf{D}_F \in \mathbb{R}^{f\times k}$ is referred to as the *dictionary*, which plays a similar role as $\mathbf{W}$, and $\boldsymbol{\alpha} \in \mathbb{R}^k$ is akin to a row of $\mathbf{H}$. SC assumes that $\mathbf{x}_{F,i}$ can be approximated by a sparse linear combination of a set of dictionary vectors, whose sparsity is controlled by the parameter λ. Although it is possible to use time-domain signals as input to SC, for audio signals it is more plausible to replace the matrix multiplication by convolution, in view of the source-filter model of sound production [19]. This gives rise to the idea of CSC [14, 15], whose details will be given in Section 4.

To our best knowledge, however, there have been few applications of SC to source separation. Blumensath and Davies employed SC for note extraction from polyphonic piano and monaural blind source separation, without side information [20]. In addition to using a simple dictionary, they considered building a shift-invariant dictionary with shifted *codewords* (i.e. columns of the dictionary) in time domain. As the convolution operator has commutative property, the idea is equivalent to CSC. However, they only discussed how to discard redundant codewords in the dictionary when solving Eq. 2, rather than solving with the full dictionary. On the other hand, Mørup *et al.* studied shift-invariant SC, which has the same form as CSC, in image and music [21]. For music, they applied a non-negativity constraint as the amplitude of the spectrum is used, but they only presented a preliminary result on separating an organ and a piccolo, rather than an objective evaluation. More importantly, this approach would also suffer from the phase copy issue.

3. PROBLEM DEFINITION

In this work, we consider the separation of a monaural and polyphonic music with p different instruments (i.e. sources) with certain side information. For a monaural time-domain audio signal $\mathbf{x}_u \in \mathbb{R}^n$ with length n, we can express it as

$$\mathbf{x}_u = \mathbf{S}_u\mathbf{1}_p + \epsilon\,, \tag{3}$$

where $\mathbf{S}_u \in \mathbb{R}^{n\times p}$ is the source matrix with p channels, $\mathbf{1}_p$ is a p-dimensional mixing vector with all elements equal to 1, and ϵ is a noise term. For convenience, we further segment $\mathbf{x}_u^T = [\mathbf{x}_{u,1}^T, \mathbf{x}_{u,2}^T, ..., \mathbf{x}_{u,o}^T]$. The ultimate goal is recovering $\mathbf{S}_u$ with only $\mathbf{x}_u$ given. As the problem is clearly ill-posed, most approaches, this work included, require other information for high-quality separation. Specifically, we consider the following two types of side information in this paper.

type 1 We assume the musical instruments and the pitches of each part that will be presented in the audio signal (but not their orderings and onsets/offsets) are known.

type 2 Besides the type 1 information, we are further given the onset and offset time of each note. In practice, this information can be obtained by a score follower [5], a multi-pitch estimator [22], or the score, assuming that tempo is given and players are sufficiently professional.

4. PROPOSED ALGORITHM

SC has been applied with success in many music classification problems [23, 24]. Nevertheless, SC-based source separation requires the solution $\boldsymbol{\alpha}$ to be truly sparse, which is not a major concern for classification problems. In addition, SC-based source separation demands either the $\mathbf{D}_F$ to be semantically meaningful or availability of appropriate side information, so that we can reconstruct the sources by using different parts of $\mathbf{D}_F$. For example, each column of $\mathbf{D}_F$ may correspond to a pitch of an instrument. To increase the chance of finding the exact solution, we implicitly put constraints in the time domain by using a convolutional approach. Moreover, with a convolutional approach, it is easy to directly process the time-domain signal to circumvent the phase copy issue.

4.1. Convolutional Sparse Coding

We first assume a time-domain dictionary superset $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_k]$ is known, where $\mathbf{d}_i \in \mathbb{R}^{t_i}, \forall i \in \mathcal{K} = \{1, 2, ..., k\}$

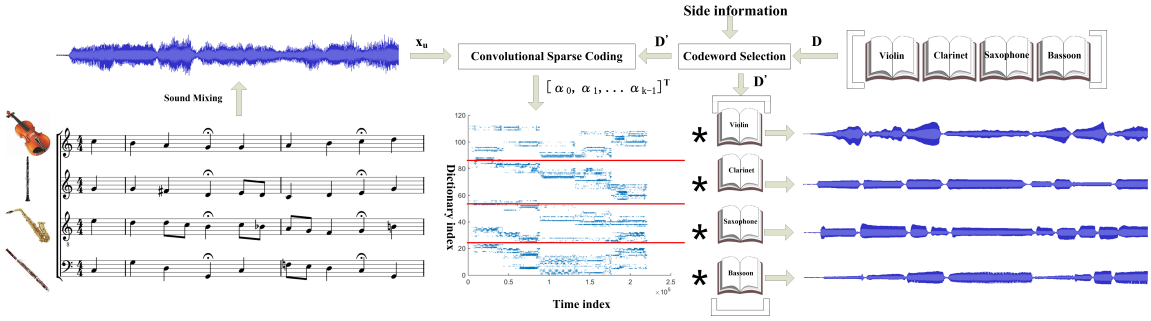

Fig. 1. Illustration of the source separation system based on convolutional sparse coding.

Table 1. The chorale names and indices of the Bach10

#	name	#	name
1	Ach Gott und Herr	6	Die Sonne
2	Ach, Lieben Christen	7	Herr Gott
3	Christe der du bist Tag und Licht	8	Für Deinen Thron
4	Christe, Du Beistand	9	Jesus
5	Die Nacht	10	Nun Bitten

is a codeword with arbitrary finite length t_i,which can be different for different codewords. $\mathbf{D}$ is assumed to cover all the pitch ranges of all the instruments of interest. For an input $\mathbf{x}_u$, CSC aims to solve the following optimization problem

$$\underset{\{\boldsymbol{\alpha}_i\}_{i\in\mathcal{S}_{u,w}}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{x}_{u,w} - \sum_{i\in\mathcal{S}_{u,w}} \mathbf{d}_i \star \boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{i\in\mathcal{S}_{u,w}} \|\boldsymbol{\alpha}_i\|_1 , \quad (4)$$

where $\mathbf{x}_{u,w}$ denotes the w-th segment of the input, $\boldsymbol{\alpha}_i \in \mathbb{R}^n$ is the sparse coefficient vector of the i-th codeword, and $\mathcal{S}_{u,w} \subset \mathcal{K}$ is a dictionary subset selected by side information (i.e. type 1 or 2 described in Section 3). Different $\mathcal{S}_{u,w}$ would be used for different inputs, and $\mathcal{S}_{u,w}$ is selected according to whether a pitch presents in $\mathbf{x}_{u,w}$. When onset and offset are informed (i.e. type 2), $\mathcal{S}_{u,w}$ can be precisely determined, and the segmentation of the input can be arbitrary. In contrast, for type 1 we solve for $\mathbf{x}_u$ directly in Eq. 4 without segmentation.

Comparing the formulation to NMF, we see that $\boldsymbol{\alpha}_i$ is similar to the i-th row of $\mathbf{H}$ (the activation of a pattern), and $\mathbf{d}_i$ is the i-th column of $\mathbf{W}$. The difference is that CSC adopts the convolution operator, $\star$, for reconstruction, and that a sparsity constraint on $\boldsymbol{\alpha}_i$ is enforced.

Figure 1 shows the overview of our source separation system. As the dictionary is composed of multiple instruments, we can recover each instrument signal by simply summing the convolved result of codewords corresponding to that instrument. That is, $\hat{\mathbf{s}}_{j,u,w} = \sum_{i\in\mathcal{I}_{j,u,w}} \mathbf{d}_i \star \boldsymbol{\alpha}_i$, where $\mathcal{I}_j$ is the set of codewords for the j^{th} instrument and $\biguplus_{j=1}^{p} \mathcal{I}_{j,u,w} = \mathcal{S}_{u,w}$.

4.2. Supervised Dictionary Learning

The dictionary can be built by an exemplar or a learning approach. Using an *exemplar* dictionary copied from the real separated source should be a near perfect approach. However, the problem is that the separated source is not known *a priori*. Simply using an exemplar dictionary built from other music may not perform well due to the mismatch between the input signal and the dictionary. For example, a violist can express a pitch in different durations, dynamics, and timbre. As an alternative, we can use a *learned* dictionary by learning a number of codewords to capture the variations of each pitch in each instrument. This dictionary learning process entails

$$\underset{\{\mathbf{d}_i\},\{\boldsymbol{\alpha}_{i,j}\}}{\operatorname{argmin}} \frac{1}{2} \sum_{j\in\mathcal{T}_m} \|\mathbf{z}_j - \sum_{i\in\mathcal{P}_m} \mathbf{d}_i \star \boldsymbol{\alpha}_{i,j}\|_2^2 + \lambda \sum_{j\in\mathcal{T}_m} \sum_{i\in\mathcal{P}_m} \|\boldsymbol{\alpha}_{i,j}\|_1 , \quad (5)$$

where $\mathbf{z}_j \ \forall j \in \mathcal{T}_m$ are the training data for m-th set of codewords, $\mathcal{P}_m$ is an index set such that $\biguplus_{m=1}^{r} \mathcal{P}_m = \mathcal{K}$, and r is the number of trained sub-dictionaries.

In our implementation, we use an efficient CSC solver [15] for solving Eqs. 4 and 5.

5. EXPERIMENTS

5.1. Dataset & Evaluation Criteria

We performed objective evaluation of the separation quality using the Bach10 dataset [5], since it is one of the few datasets recorded with real instruments evaluation. Table 1 lists the ten J. S. Bach chorales in Bach10. Each of them consists of four parts: violin, clarinet, saxophone and bassoon. The sampling rate is 44,100 Hz and duration ranges from 25 to 42 seconds. From the annotated text files of Bach10, we extracted all the side information needed for both type 1 and 2 for each chorale. In our experiment, we considered a 2-fold cross validation evaluation protocol, using the full chorales of one fold for dictionary learning and the first 5 seconds of the other fold for source separation, and repeating the experiment again by interchanging the roles of the two folds. We took a random partition and used {[1, 2, 3, 6, 10]} and {[4, 5, 7, 8, 9]} as the two folds. The performance was measured over the average of sources in terms of source to distortion ratio (SDR), source to interferences ratio (SIR), and source to artifact ratio (SAR), calculated by the Blind Source Separation (BSS) Eval toolbox v3.0 [25]. As these are standard metrics in source separation, we refer readers to [25] for the definitions and details thereof.

Table 2. Objective performance evaluation

Method	Dictionary	SDR	SIR	SAR
type 1+$\mathbf{S}_u$ CSC	exemplar (oracle)	13.48	30.89	13.61
type 1 CSC	learned (0.1 sec)	7.01	10.45	10.66
type 2 CSC	learned (0.1 sec)	8.59	13.54	10.95
Soundprism [5]	N/A	6.57	10.66	9.41

5.2. Setting and Result

Table 2 compares the performance of different settings of CSC against Soundprism [5], which achieved the best known performance for Bach10 so far. The first row gives an approximate upper bound of CSC-based approach with an exemplar dictionary. This can be viewed as an *oracle* method because $\mathbf{D}$ was built from segmentations of $\mathbf{S}_u$ of the ten chorales, where the segmentation was done according to the onset and offset times. Moreover, we used the type 1 side information to inform CSC. It can be found that this oracle method outperforms Soundprism by almost 7dB in SDR, exhibiting extremely high quality separation.

The second and third rows, on the other hand, used the same dictionary superset $\mathbf{D}$ learned from the training fold. Specifically, we set $t_i = 4,410$, $\forall i \in \mathcal{K}$ (i.e. the length of the dictionary filters is 0.1 second) and $|\mathcal{P}_m| = 4, \forall m = 1, ..., r$ (i.e. four dictionary filters for each pitch in each instrument). The parameter λ is set to 0.05 for solving Eq. 5, and λ in Eq. 4 is empirically determined and fixed for all the ten chorales. The two rows differ in the side information given. For type 1, we solve Eq. 4 using the full-length 5-second $\mathbf{x}_u$ as the input; for type 2, we solve Eq. 4 using the five one-second segments $\mathbf{x}_{u,w}$ segmented uniformly from $\mathbf{x}_u$ as the inputs one-by-one. It can be seen that using more side information (i.e. type 2) performs better in all the three metrics. Our hypothesis is that too many codewords might introduce ambiguity among the codewords and thereby impede stable recovery [26, 27], but further work is needed to validate this hypothesis. However, although type 2 performs better in objective evaluation, there might be audible artifacts near the boundary of the segments due to discontinuities caused by independently solving each segment. While a comprehensive subjective evaluation of the separation quality is left as a future work, we do provide the audio files of the separation result as an online supplementary material that can be accessed at the link `http://mac.citi.sinica.edu.tw/research/CSC_separation/`.

By comparing the second to fourth rows of Table 2, we see that for either type of side information the score-informed CSC outperforms Soundprism in most metrics. The performance difference between the type 2 CSC is around 1.5 dB for SDR and 3 dB for SIR. Moreover, even without the onset/offset information, the type 1 CSC still compares favorably with Soundprism. While Soundprism requires a MIDI file that contains both the pitch and the corresponding order of the pitch, the type 1 method only needs pitch information.

To illustrate the phase copy issue, we show five waveforms in Figure 2 and present a qualitative analysis. The waveforms are excerpts of the third pitch from the saxophone with midi number 57 in chorale 9 of Bach10. From the top to bottom, the waveforms are ground truth, Soundprism (phase copied from mixed source), Soundprism with phase copied from ground truth (the window length is different in both cases), our approach with onset informed, and the last one combined the magnitude of our approach and the phase from ground truth. By observing the shape of first three waveforms from top, it can be found that, with phased copied from the ground truth, Soundprism has a relatively similar shape to the ground truth. As a much different envelope gives a perceptible auditory difference, we can infer that using a phase from the mixed observation may easily cause unnatural sound. On the contrary, the differences among the ground truth and the two waveforms of CSC are relatively subtle, suggesting that the timbre information is well preserved by CSC. This finding is consistent with our subjective experience in listening to the separation result, although it is difficult to quantify the subjective evaluation here.

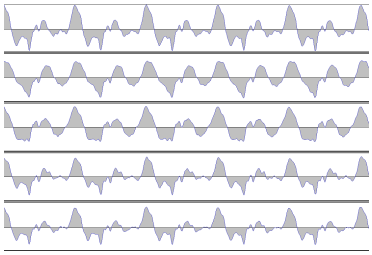

Fig. 2. Analysis of the phase copy issue using real signals.

6. CONCLUSION

In this paper, we have demonstrated that CSC is an adequate method to exploit side information for time-domain monaural source separation. We have proposed and evaluated two variants of the CSC approach, using different levels of side information, and showed that pitch informed CSC compares favorably with a state-of-the-art method Soundprism that requires the exact ordering of the pitches. Moreover, when onset and offset times of the pitch is informed and employed to segment the input signals, the proposed method outperforms Soundprism remarkably in three objective performance metrics. Although not quantitatively reported here, we found that CSC can tolerate some errors in the onset/offset time. Based on our findings, we envision that one can combine CSC with state-of-the-art multi-pitch and onset detectors to develop a practical source separation system [28, 29].

That being said, we need to point out that the performance of CSC is sensitive to the quality of the dictionary. If the dictionary does not match the input signal, the separation quality can be poor. Therefore, further work is needed to improve the mismatch tolerance of CSC or to improve its scalability.

7. REFERENCES

[1] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-Level representation for pitch estimation and musical audio source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1180–1191, 2011.

[2] P. Sprechmann, Alexander M. B., and Guillermo S., "Real-time online singing voice separation from monaural recordings using robust low-rank modeling," in *Proc. Int. Soc. Music Info. Retrieval Conf.*, 2012, pp. 67–72.

[3] H. Tachibana, N. Ono, and S. Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 228–237, 2014.

[4] Z. Rafii, Z. Duan, and B. Pardo, "Combining rhythm-based and pitch-based methods for background and melody separation," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1884–1893, 2014.

[5] Z. Duan and B. Pardo, "Soundprism: an online system for score-informed source separation of music audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1205–1215, 2011.

[6] B. Fuentes, R. Badeau, and G. Richard, "Blind harmonic adaptive decomposition applied to supervised source separation," in *Proc. European Signal Processing Conf.*, 2012, pp. 2654–2658.

[7] A. Lefevre, F. Bach, and C. Févotte, "Semi-supervised NMF with time-frequency annotations for single-channel source separation," in *Proc. Int. Soc. Music Info. Retrieval Conf.*, 2012, pp. 115–120.

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

[9] J. Ganseman, P. Scheunders, G. J. Mysore, and J. S Abel, "Evaluation of a score-informed source separation system," in *Proc. Int. Soc. Music Info. Retrieval Conf.*, 2010, pp. 219–224.

[10] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2013, pp. 888–891.

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

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

[13] 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. Acoustics, Speech and Signal Processing*, 2014, pp. 649–653.

[14] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010, pp. 2528–2535.

[15] B. Wohlberg, "Efficient convolutional sparse coding," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2014, pp. 7223–7227.

[16] 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, and Language Processing*, vol. 19, no. 7, pp. 2026–2038, 2011.

[17] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1?," *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[18] P. O. Hoyer, "Non-negative sparse coding," in *Proc. IEEE Workshop on Neural Networks for Signal Processing*, 2002, pp. 557–565.

[19] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.

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

[21] M. Mørup, M. N. Schmidt, and L. K. Hansen, "Shift invariant sparse coding of image and music data," Tech. Rep. IMM2008-04659, Technical University of Denmark, 2008.

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

[23] J. Nam, J. Herrera, M. Slaney, and J. O Smith, "Learning sparse feature representations for music annotation and retrieval," in *Proc. Int. Soc. Music Info. Retrieval Conf.*, 2012.

[24] M. Henaff, K. Jarrett, K. Kavukcuoglu, and Y. LeCun, "Unsupervised learning of sparse features for scalable audio classification," in *Proc. Int. Soc. Music Info. Retrieval Conf.*, 2011.

[25] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[26] B. Wohlberg, "Noise sensitivity of sparse signal representations: Reconstruction error bounds for the inverse problem," *IEEE Trans. Signal Processing*, vol. 51, no. 12, pp. 3053–3060, Dec. 2003.

[27] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Information Theory*, vol. 52, no. 1, pp. 6–18, 2006.

[28] E. Benetos and S. Dixon, "Joint multi-pitch detection using harmonic envelope estimation for polyphonic music transcription," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1111–1123, Oct 2011.

[29] J. Schlüter and S. Böck, "Improved musical onset detection with convolutional neural networks," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2014, pp. 6979–6983.