

HARMONIC VARIABLE-SIZE DICTIONARY LEARNING FOR MUSIC SOURCE SEPARATION

Steven K. Tjoa, Matthew C. Stamm, W. Sabrina Lin, and K. J. Ray Liu

Signals and Information Group, Dept. of Electrical and Computer Engineering
University of Maryland – College Park, MD 20742 USA
{kiemyang, mcstamm, wylin, kjrliu}@umd.edu

ABSTRACT

Dictionary learning through matrix factorization has become widely popular for performing music transcription and source separation. These methods learn a concise set of dictionary atoms which represent spectrograms of musical objects. However, there is no guarantee that the atoms learned will be perceptually meaningful, particularly when there exists significant spectral and temporal overlap among the musical sources. In this paper, we propose a novel dictionary learning method that imposes additional harmonic constraints upon the atoms of the learned dictionary while allowing the dictionary size to grow appropriately during the learning procedure. When there is significant spectral-temporal overlap among the musical sources, our method outperforms popular existing matrix factorization methods as measured by the recall and precision of learned dictionary atoms.

Index Terms— Nonnegative matrix factorization, pitch estimation, sparse coding, music transcription.

1. INTRODUCTION

Recently, researchers have proposed many approaches for performing music transcription and source separation. In particular, one such category of approaches – spectral decomposition through matrix factorization – has received plenty of attention. By first expressing a time-frequency representation of the musical signal as a matrix, these methods decompose each column of the matrix into a summation of individual vectors, each corresponding to one musical source or note [1, 2].

These methods commonly share two important steps: *dictionary learning* and *sparse coding*. Dictionary learning refers to the construction of a set of atoms – the dictionary – from which the input signal can be represented, and sparse coding is used to compute the contribution of each dictionary atom to the signal at each moment in time. Methods known as nonnegative matrix factorization (NMF) also impose a nonnegativity constraint on the dictionary and its coefficients in order to learn more meaningful atoms. The nonnegativity constraint makes sense considering that we only have the presence or absence of a source from a signal and never the “subtraction” of a source from a signal in which it is already absent.

Unfortunately, these methods also share a common limitation. When there is significant spectral-temporal overlap in the signal among the dictionary atoms, it becomes difficult for these methods to learn atoms properly. Often, information from multiple atoms is represented as a single atom by the learning procedure. If an atom in the output dictionary contains musical information from multiple sources, transcription and source separation cannot be accurately performed. Furthermore, if the dictionary atoms themselves are

highly correlated, as is common when harmonic frequencies between atoms overlap, accurate dictionary learning becomes even more difficult.

In this paper, we propose a novel dictionary learning method designed to perform well despite spectral-temporal overlap among the dictionary atoms. The dictionaries learned by this method contain atoms which accurately resemble the original notes and sources which comprise the input signal. While our method is based on matrix factorization, it imposes an additional harmonic constraint that restricts each atom to represent at most one pitch. Furthermore, our method is flexible by allowing the size of the dictionary to grow based upon the complexity of the input signal, unlike other methods which fix the dictionary size a priori. Our method consistently outperforms other dictionary learning methods such as nonnegative matrix factorization with multiplicative updates (NMF-MU) [3], K-SVD [4], nonnegative K-SVD (NN-K-SVD) [5], and the method of optimal directions (MOD) [6], as measured by the recall and precision of learned dictionary atoms.

2. PROBLEM FORMULATION

Dictionary learning methods based upon matrix factorization accept a time-frequency representation of the musical signal as the input. Although there exist many different time-frequency representations, we will simply use the magnitude spectrogram of the input signal.

Given a discrete-time single-channel music signal $x(n)$, the magnitude spectrogram of the input signal is a real-valued nonnegative matrix $\mathbf{X} \in \mathbb{R}_+^{M \times N}$, where $\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_N]$, whose N columns are the discrete Fourier transform (DFT) magnitudes of consecutive, possibly overlapping, frames of the input signal. Given the matrix \mathbf{X} , our primary goal is to find two matrices, the dictionary $\mathbf{A} \in \mathbb{R}_+^{M \times K}$ and gain matrix $\mathbf{S} \in \mathbb{R}_+^{K \times N}$, which minimize some distance between \mathbf{X} and \mathbf{AS} . If we denote $\|\mathbf{X}\|_F$ as the Frobenius norm of \mathbf{X} , where $\|\mathbf{X}\|_F^2 = \text{tr}(\mathbf{X}^T \mathbf{X}) = \sum_{i,j} x_{ij}^2$, then we can describe the problem as follows:

$$\min_{\mathbf{A}, \mathbf{S}} \|\mathbf{X} - \mathbf{AS}\|_F^2 \quad \text{s.t. } \mathbf{A} \in \mathbb{R}_+^{M \times K}, \mathbf{S} \in \mathbb{R}_+^{K \times N}. \quad (1)$$

The columns of the matrix $\mathbf{A} = [\mathbf{a}_1 \ \mathbf{a}_2 \ \dots \ \mathbf{a}_K]$ correspond to the individual atoms of the dictionary. In this musical context, these atoms resemble the spectra of individual sources or notes found in the musical mixture. The gain matrix $\mathbf{S} = [\mathbf{s}_1 \ \mathbf{s}_2 \ \dots \ \mathbf{s}_K]^T$ represents the contribution of each dictionary atom in the spectrogram \mathbf{X} , i.e., the element s_{kn} indicates the amount of atom \mathbf{a}_k present in observation \mathbf{x}_n . We refer to the row vector \mathbf{s}_k^T as the k^{th} row of \mathbf{S} , i.e., \mathbf{s}_k indicates the activity of atom \mathbf{a}_k across time.

3. DICTIONARY LEARNING: EXISTING METHODS

To motivate our proposed algorithm, we discuss existing dictionary learning procedures based upon singular value decomposition (SVD), including K-SVD and its nonnegative variant, NN-K-SVD [4, 5]. SVD computes the matrix factorization $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ where $\mathbf{U} = [\mathbf{u}_1 \mathbf{u}_2 \dots \mathbf{u}_M]$ and $\mathbf{V} = [\mathbf{v}_1 \mathbf{v}_2 \dots \mathbf{v}_N]$ are both orthonormal matrices and the diagonal matrix $\mathbf{\Sigma}$ is such that for any choice of K , the difference $\|\mathbf{X} - \sum_{k=1}^K \sigma_{kk} \mathbf{u}_k \mathbf{v}_k^T\|_F$ is minimized. In our context, the dictionary \mathbf{A} corresponds to the first K columns of \mathbf{U} , and the gain matrix \mathbf{S} corresponds to the first K rows of $\mathbf{\Sigma}\mathbf{V}^T$. Intuitively, through SVD, we find the K dictionary atoms and their associated gains which best represent the magnitude spectrogram \mathbf{X} .

However, SVD does not guarantee sparsity or nonnegativity of the factorization. On the other hand, K-SVD is an iterative algorithm that learns a dictionary that can be overcomplete and whose gain coefficients are sparse. Instead of immediately solving for \mathbf{A} and \mathbf{S} jointly, this algorithm solves the minimization in (1) one dictionary atom at a time, ignoring the nonnegativity constraints, while the other atoms remain constant. In other words, for a given k , each iteration of K-SVD solves the minimization

$$\min_{\mathbf{a}_k, \mathbf{s}_k} \|\mathbf{X} - \mathbf{A}\mathbf{S}\|_F^2. \quad (2)$$

Note that

$$\begin{aligned} \|\mathbf{X} - \mathbf{A}\mathbf{S}\|_F &= \left\| \mathbf{X} - \sum_{j=1}^K \mathbf{a}_j \mathbf{s}_j^T \right\|_F \\ &= \left\| \left(\mathbf{X} - \sum_{j \neq k} \mathbf{a}_j \mathbf{s}_j^T \right) - \mathbf{a}_k \mathbf{s}_k^T \right\|_F. \end{aligned} \quad (3)$$

For convenience, denote

$$\mathbf{E}_k = \mathbf{X} - \sum_{j \neq k} \mathbf{a}_j \mathbf{s}_j^T. \quad (4)$$

Then, the solution to (2) is the rank-one approximation of the SVD $\mathbf{E}_k = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, specifically, $\mathbf{a}_k = \mathbf{u}_1$ and $\mathbf{s}_k = \sigma_{11} \mathbf{v}_1$. K-SVD adjusts \mathbf{a}_k and \mathbf{s}_k accordingly in each iteration and moves on to the next dictionary atom in the next iteration. The entire process is repeated until convergence of the dictionary occurs. Sparse coding is applied to update the gain matrix before each set of K iterations.

While K-SVD encourages sparsity and accommodates overcompleteness, it still does not influence the nonnegativity of either the dictionary \mathbf{A} or the gain matrix \mathbf{S} . On the other hand, nonnegative K-SVD (NN-K-SVD) retains the same flavor of K-SVD while maintaining nonnegativity of the matrix elements. Consider the following constrained minimization:

$$\min_{\mathbf{a}_k} \|\mathbf{E}_k - \mathbf{a}_k \mathbf{s}_k^T\|_F^2 \quad \text{s.t. } \mathbf{a}_k \in \mathbb{R}_+^M. \quad (5)$$

Here, we keep \mathbf{s}_k constant and enforce the nonnegativity of \mathbf{a}_k . By differentiating the objective function, it can be shown that the optimal solution for \mathbf{a}_k (similarly, for \mathbf{s}_k by keeping \mathbf{a}_k constant) is

$$\mathbf{a}_k = \left[\frac{\mathbf{E}_k \mathbf{s}_k}{\mathbf{s}_k^T \mathbf{s}_k} \right]_+, \quad \mathbf{s}_k = \left[\frac{\mathbf{E}_k^T \mathbf{a}_k}{\mathbf{a}_k^T \mathbf{a}_k} \right]_+, \quad (6)$$

where $[\cdot]_+$ denotes a matrix or vector whose negative elements are set to zero. By observing that the Hessian of the objective function

with respect to \mathbf{a}_k is proportional to the identity matrix, the optimal projection from the unconstrained minimum to the constrained minimum is performed simply by setting all negative elements of the unconstrained solution to zero, hence the solution in (6). Each iteration of NN-K-SVD uses these rules to update \mathbf{a}_k and \mathbf{s}_k . While we no longer minimize \mathbf{a}_k and \mathbf{s}_k jointly, the updates still guarantee a decrease in the objective function while maintaining nonnegativity of the matrices \mathbf{A} and \mathbf{S} .

4. PROPOSED ALGORITHM

While NN-K-SVD can find numerically acceptable solutions to (1), some problems remain. First, there is no guarantee that the individual atoms of the learned dictionary will each correspond to only one musical source. In particular, when multiple atoms coincide in time (e.g., \mathbf{s}_1 and \mathbf{s}_2 are highly correlated), the aforementioned algorithms will learn a single atom that contains information from both \mathbf{a}_1 and \mathbf{a}_2 . For example, consider the learned atoms in Fig. 1. We fabricate a dictionary \mathbf{A} with two atoms whose gain coefficients \mathbf{S} have significant overlap in time, and then construct the spectrogram $\mathbf{X} = \mathbf{A}\mathbf{S}$. The dictionaries learned by K-SVD, NN-K-SVD, and NMF with multiplicative updates yield output dictionaries which do not match the input dictionary. However, the dictionary learned by our proposed method, discussed below, does accurately resemble the original dictionary.

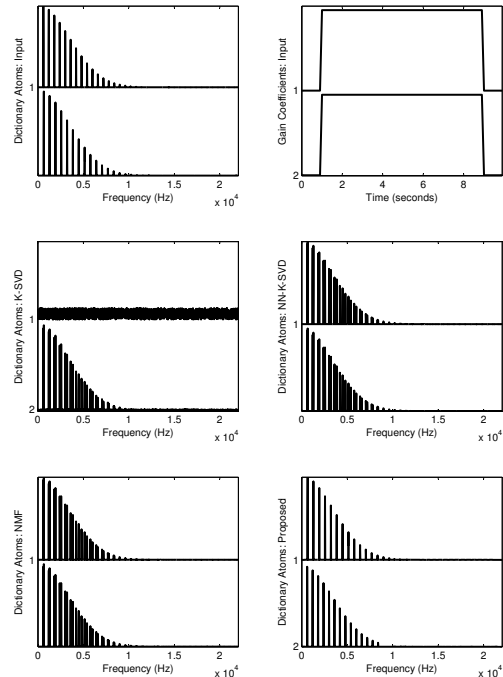


Fig. 1. Two dictionary atoms (top left) and their gain coefficients (top right) were used to construct a spectrogram. Using either K-SVD (middle left), NN-K-SVD (middle right), or NMF-MU (bottom left), the learned dictionary atoms do not resemble the original atoms. Using the proposed algorithm (bottom right), the original and learned atoms match.

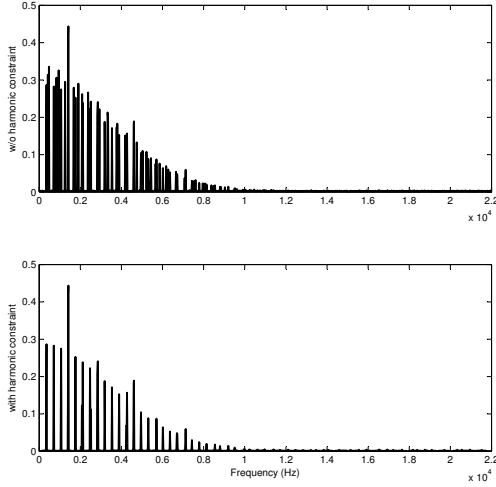


Fig. 2. Dictionary atom of original spectrum (top) and atom after filtering spectrum through a comb filter (bottom).

The second problem deals with the size of the dictionary, K . For the popular existing algorithms, the dictionary size K must be chosen before the algorithm begins. If the chosen value of K is too low, then the learned dictionary cannot accurately represent the input spectrogram. If K is too high, computation and memory requirements can increase dramatically and unnecessarily. When executing eigendecompositions and/or matrix multiplications, these requirements can become overwhelming.

In order to solve these problems, we propose a novel approach to dictionary learning that emphasizes the presence of at most *one pitch per dictionary atom*. Our method builds upon the technical foundations of NN-K-SVD mentioned earlier. As illustrated in Fig. 1, existing dictionary learning algorithms are intended for general purposes, i.e., they do not enforce any perceptual constraints on the learned dictionary atoms. Under the assumption that individual musical sources do not overlap in time or frequency, existing algorithms can learn dictionaries accurately. However, this assumption is not necessarily true for musical contexts where individual sources are highly correlated.

Motivated by the observation that music contains a series of pitched and unpitched sounds, we enforce a harmonic constraint on the learned atom by filtering the spectrum represented by \mathbf{a}_k through a comb filter, thus preserving the spectral energy around the harmonic frequencies and eliminating the energy at other frequencies as shown in Fig. 2. To estimate the fundamental frequency, we simply compute the harmonic product spectrum [7] from the first five harmonics for candidate pitches. Other frequency-domain pitch estimation algorithms can work, as well.

The other notable feature of our algorithm is the initialization and growth of the dictionary. For the best $\mathbf{A} \in \mathbb{R}^{M \times K}$, if $\|\mathbf{X} - \mathbf{A}\mathbf{S}\|_F$ is still not low enough, we increment K and add another column vector \mathbf{a}_K to \mathbf{A} and another row vector \mathbf{s}_K^T to \mathbf{S} . There are many reasonable ways to initialize \mathbf{a}_K . One could randomly generate \mathbf{a}_K , or \mathbf{a}_K could equal the mean of the columns of $\mathbf{X} - \mathbf{A}\mathbf{S}$. For this work, we simply set \mathbf{a}_K to equal a column of \mathbf{E} , \mathbf{e}_n , where n is chosen such that \mathbf{e}_n has high energy. Then, we initialize $\mathbf{s}_K = \begin{bmatrix} \mathbf{E}_K^T \mathbf{a}_K \\ \mathbf{a}_K^T \mathbf{a}_K \end{bmatrix}_+$ as shown in (6).

With each of the basic building blocks described, we now summarize the proposed algorithm.

1. Set the dictionary size K to equal 1.
2. Initialize \mathbf{a}_K and \mathbf{s}_K as desired.
3. For each $k \in \{K, K-1, \dots, 2, 1\}$,

(a) Compute \mathbf{E}_k :

$$\mathbf{E}_k = \mathbf{X} - \sum_{j \neq k} \mathbf{a}_j \mathbf{s}_j^T.$$

(b) Find \mathbf{a}_k :

$$\mathbf{a}_k = \begin{bmatrix} \mathbf{E}_k \mathbf{s}_k \\ \mathbf{s}_k^T \mathbf{s}_k \end{bmatrix}_+.$$

(c) Estimate the fundamental frequency, f_0 , for the spectrum \mathbf{a}_k using the harmonic product spectrum.

(d) Filter \mathbf{a}_k through a comb filter tuned to f_0 to emphasize its harmonicity.

(e) Find \mathbf{s}_k :

$$\mathbf{s}_k = \begin{bmatrix} \mathbf{E}_k^T \mathbf{a}_k \\ \mathbf{a}_k^T \mathbf{a}_k \end{bmatrix}_+.$$

Repeat step 3 until the dictionary \mathbf{A} converges.

4. If $\|\mathbf{X} - \mathbf{A}\mathbf{S}\|_F^2$ is low enough, stop. Otherwise, increment K , and go to step 2.

5. EXPERIMENTS

For our experiments, we synthesize a dictionary \mathbf{A}_{in} of harmonic atoms similar to the atoms in Fig. 1 having a fixed envelope on the order of $\exp(-m^2)$, where $m \in \{1, 2, \dots, M\}$ is the frequency bin index, and $M = 2048$ corresponds to the Nyquist frequency. We also synthesize the corresponding gain coefficients to be a $K \times N$ matrix with $N = 100$ and with L ones randomly placed in each column and zero otherwise. These two matrices are multiplied to obtain \mathbf{X} , the input to each dictionary learning algorithm. Six dictionary learning algorithms are tested: the proposed algorithm, NMF-MU [3], K-SVD and NN-K-SVD [8], the method of optimal directions [6, 8], and basic SVD.

The output dictionary \mathbf{A}_{out} from each algorithm is compared against the input dictionary in terms of hits, misses, and false alarms. A hit occurs if both dictionaries contain corresponding atoms whose normalized correlation exceeds 0.9. A miss occurs if an atom from \mathbf{A}_{in} does not correlate with any atom in \mathbf{A}_{out} , and a false alarm occurs if an atom from \mathbf{A}_{out} does not correlate with any atom in \mathbf{A}_{in} . Two measures are used to measure performance: recall and precision. Recall is equal to hits/(hits + misses), and precision is equal to hits/(hits + false alarms). These measures are averaged over ten trials of each experiment.

First, we illustrate the effects of the dictionary size K and the number of simultaneously active atoms L on dictionary learning. For each trial, we generate a dictionary with K harmonic atoms, each with a randomly-selected fundamental frequency that is uniformly distributed over the MIDI interval [48, 84]. Fig. 3 illustrates results for $K = 5$ and $L \in \{1, 2, 3, 4\}$, while Fig. 4 illustrates results for $K = 20$ and $L \in \{1, 2, \dots, 19\}$. Because the existing algorithms are initialized to strictly contain K atoms, each miss must accompany a false alarm, thus making their recall and precision is equal. On the other hand, the proposed algorithm must infer the proper value for

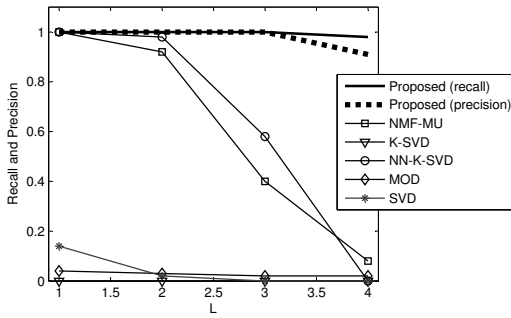


Fig. 3. Recall and precision when $K = 5$ for $L \in \{1, 2, 3, 4\}$. Ground-truth pitches are initialized randomly over ten trials.

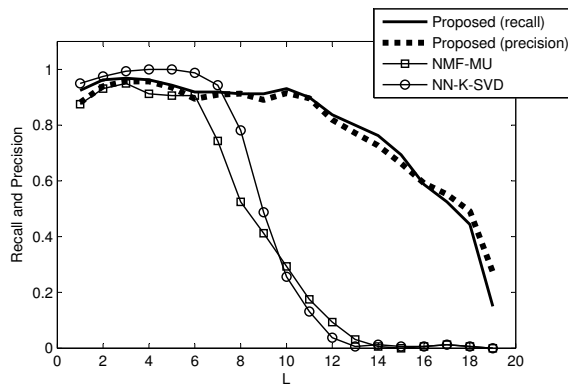


Fig. 4. Recall and precision when $K = 20$ for $L \in \{1, 2, \dots, 19\}$. Ground-truth pitches are initialized randomly over ten trials.

K as described earlier. When the estimated and true values of K differ, then misses and false alarms can occur independently.

As shown in Figs. 3 and 4, the recall and precision for the proposed algorithm is better than the other algorithms for most combinations of K and L , particularly when L is high. The performance of all methods degrades as L increases because the amount of spectral-temporal overlap also increases. However, the proposed method learns more accurate atoms when L is high because of the additional harmonic constraints. The two existing methods with nonnegative constraints, NMF-MU and NN-K-SVD, both perform well except when L is high because of their inability to resolve the spectral-temporal overlap. The remaining methods – K-SVD, MOD, and SVD – all fail because of the lack of a nonnegativity constraint.

Next, we show results when the pitches of the input dictionary atoms have overlapping harmonics. To ensure a high amount of overlap, we fix $K = 5$ and $L = 3$. The five chosen pitches are 200, 300, 400, 500, and 600 Hz. The gain matrix is once again randomly generated by assigning L ones to each column of \mathbf{S} as described earlier. Fig. 5 shows that the best recall and precision is achieved by the proposed algorithm. Again, recall and precision are equal for each of the existing methods because K is fixed to its correct value thus creating a one-to-one correspondence between misses and false alarms. Finally, we fix $K = 10$ and $L = 5$ where $f_0 \in \{200, 300, 400, 500, 600, 800, 900, 1000, 1200, 1500\}$. Fig. 5 again shows that the best recall and precision is achieved by the proposed algorithm.

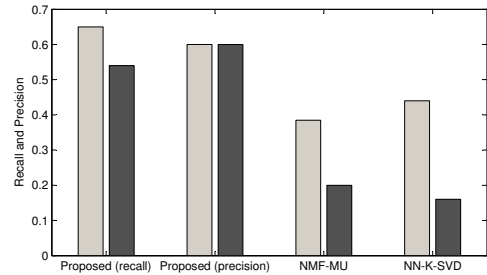


Fig. 5. Recall and precision when $K = 5$ and $L = 3$ (light gray) and when $K = 10$ and $L = 5$ (dark gray). Pitches are chosen such that large spectral-temporal overlap occurs.

6. CONCLUSIONS

We have presented a novel method of dictionary learning based upon nonnegative K-SVD which can separate sources that are otherwise inseparable using common methods. Despite the simplicity of our algorithm, it performs well for a variety of musical scenarios involving pitched sounds with spectral-temporal overlap. In the future, we plan to investigate the robustness proposed algorithm under different acoustic conditions, particularly for music that includes additive noise or unpitched sources, along with decomposition of time-frequency representations of natural music signals.

After learning a dictionary of perceptually meaningful atoms, the next stage involves clustering of the dictionary atoms according to their musical source. While some clustering methods already exist, difficulties remain when doing this in an unsupervised manner. If combined with a successful atom clustering method, we believe that the proposed algorithm can offer state-of-the-art accuracy and robustness in music transcription and source separation tasks.

7. REFERENCES

- [1] P. Smaragdis and J. C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 2003, pp. 177–180.
- [2] S. A. Abdallah and M. D. Plumbley, “Unsupervised analysis of polyphonic music using sparse coding,” *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 179–196, Jan. 2006.
- [3] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Advances in Neural Information Processing Systems*, Denver, 2001, pp. 556–562.
- [4] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
- [5] M. Aharon, M. Elad, and A. M. Bruckstein, “K-SVD and its non-negative variant for dictionary design,” in *Proc. SPIE conference wavelets*, July 2005, vol. 5914, pp. 327–339.
- [6] K. Engan, S. O. Aase, and J. H. Husoy, “Method of optimal directions for frame design,” in *Proc. of the IEEE Conf. Acoustics, Speech, and Signal Processing*, Mar. 1999, vol. 5, pp. 2443–2446.
- [7] M. R. Schroeder, “Period Histogram and Product Spectrum: New Methods for Fundamental-Frequency Measurement,” *The Journal of the Acoustical Society of America*, vol. 43, pp. 829, Apr. 1968.
- [8] M. Elad, “Software,” <http://www.cs.technion.ac.il/~elad/software/>.