On Spectral Basis Selection for Single Channel Polyphonic Music Separation

Minje Kim and Seungjin Choi

Department of Computer Science,
Pohang University of Science and Technology,
San 31 Hyoja-dong, Nam-gu, Pohang 790-784, Korea
{minjekim, seungjin}@postech.ac.kr

Abstract. In this paper we present a method of separating musical instrument sound sources from their monaural mixture, where we take the harmonic structure of music into account and use the sparseness and the overlapping NMF to select representative spectral basis vectors which are used to reconstruct unmixed sound. A method of spectral basis selection is illustrated and experimental results with monaural instantaneous mixtures of voice/cello and saxophone/viola, are shown to confirm the validity of our proposed method.

1 Introduction

The nonnegative matrix factorization (NMF) [1] or its extended version, nonnegative matrix deconvolution (NMD) [2], was shown to be useful in polyphonic music description [3], in the extraction of multiple music sound sources [2], and in general sound classification [4]. On the other hand, a method based on multiple cause models and sparse coding was successfully applied to automatic music transcription [5]. Some of these methods regard each note as a source, which might be appropriate for music transcription and work for source separation in a very limited case.

In this paper we present a method for single channel polyphonic music separation, the main idea of which is to select a few representative spectral basis vectors using the sparseness and the overlapping NMF [6], which are used to reconstruct unmixed sound signals. We assume that the structure of harmonics of a musical instrument approximately remains the same, even if it is played at different pitches. This view allows us to reconstruct original sound using only a few representative spectral basis, through the overlapping NMF. We illustrate a method of spectral basis selection from the spectrogram of mixed sound and show how these basis vectors are used to restore unmixed sound. Promising results with monaural instantaneous mixtures of voice/cello and saxophone/viola, are shown to confirm the validity of our proposed method.

2 Overlapping NMF

Nonnegative matrix factorization (NMF) is a simple but efficient factorization method for decomposing multivariate data into a linear combination of basis

W. Duch et al. (Eds.): ICANN 2005, LNCS 3697, pp. 157–162, 2005.

vectors with nonnegativity constraints for both basis and encoding matrix [1]. Given a nonnegative data matrix $\boldsymbol{V} \in \mathbb{R}^{m \times N}$ (where $V_{ij} \geq 0$), NMF seeks a factorization

$$\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H}, \tag{1}$$

where $\boldsymbol{W} \in \mathbb{R}^{m \times n}$ $(n \leq m)$ contains nonnegative basis vectors in its columns and $\boldsymbol{H} \in \mathbb{R}^{n \times N}$ represents the nonnegative encoding variable matrix. Appropriate objective functions and associated multiplicative updating algorithms for NMF can be found in [7].

The overlapping NMF is an interesting extension of the original NMF, where transform-invariant representation and a sparseness constraint are incorporated with NMF [6]. Some of basis vectors computed by NMF could correspond to the transformed versions of a single representative basis vector. The basic idea of the overlapping NMF is to find transformation-invariant basis vectors such that fewer number of basis vectors could reconstruct observed data. Given a set of transformation matrices, $\mathcal{T} = \left\{\boldsymbol{T}^{(1)}, \boldsymbol{T}^{(2)}, \ldots, \boldsymbol{T}^{(K)}\right\}$, the overlapping NMF finds a nonnegative basis matrix $\boldsymbol{W}$ and a set of nonnegative encoding matrix $\left\{\boldsymbol{H}^{(k)}\right\}$ (for $k = 1, \ldots, K$) which minimizes

$$\mathcal{J}(\boldsymbol{W}, \boldsymbol{H}) = \frac{1}{2}\left\|\boldsymbol{V} - \sum_{k=1}^{K} \boldsymbol{T}^{(k)}\boldsymbol{W}\boldsymbol{H}^{(k)}\right\|_F^2, \tag{2}$$

where $\|\cdot\|_F$ represents Frobenious norm. As in [7], the multiplicative updating rules for the overlapping NMF were derived in [6], which are summarized below.

Algorithm Outline: Overlapping NMF [6].

Step 1 Calculate the reconstruction: $\boldsymbol{R} = \sum_{k=1}^{K} \boldsymbol{T}^{(k)}\boldsymbol{W}\boldsymbol{H}^{(k)}$.

Step 2 Update the encoding matrix by

$$\boldsymbol{H}^{(k)} \leftarrow \boldsymbol{H}^{(k)} \odot \frac{\boldsymbol{W}^T \left[\boldsymbol{T}^{(k)}\right]^T \boldsymbol{V}}{\boldsymbol{W}^T \left[\boldsymbol{T}^{(k)}\right]^T \boldsymbol{R}}, \quad k = 1, \ldots, K, \tag{3}$$

where $\odot$ denotes the Hadamard product and the division is carried out in an element-wise fashion.

Step 3 Calculate the reconstruction $\boldsymbol{R}$ again using the encoding matrix $\boldsymbol{H}^{(k)}$ updated in Step 2, as in Step 1.

Step 4 Update the basis matrix by

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \odot \frac{\sum_{k=1}^{K} \left[\boldsymbol{T}^{(k)}\right]^T \boldsymbol{V} \left[\boldsymbol{H}^{(k)}\right]^T}{\sum_{k=1}^{K} \left[\boldsymbol{T}^{(k)}\right]^T \boldsymbol{R} \left[\boldsymbol{H}^{(k)}\right]^T}. \tag{4}$$

3 Spectral Basis Selection

The goal of spectral basis selection is to choose a few representative vectors from $\boldsymbol{V} = [\boldsymbol{v}_1 \cdots \boldsymbol{v}_N]$ where $\boldsymbol{V}$ is the data matrix associated with the spectrogram of mixed sound. In other words, each column vector of $\boldsymbol{V}$ corresponds to the power spectrum of the mixed sound at time $t = 1, \ldots, N$. Selected representative vectors are fixed as basis vectors that are used to learn an associated encoding matrix through the overlapping NMF with the sparseness constraint, in order to reconstruct unmixed sound.

Our spectral basis selection method consists of two parts, which is summarized in Table 1. The first part is to select several candidate vectors from $\boldsymbol{V}$ using a sparseness measure and a clustering technique. We use the sparseness measure proposed by Hoyer [8], described by

$$\text{sparseness}(\boldsymbol{v}) = \frac{\sqrt{m} - (\sum |v_i|)/\sqrt{\sum v_i^2}}{\sqrt{m} - 1}, \tag{5}$$

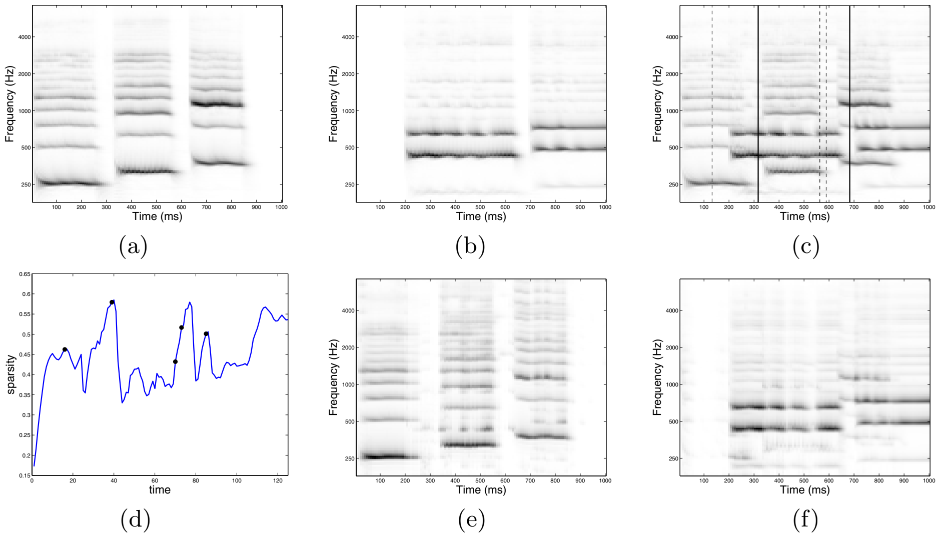
where v_i is the ith element of the m-dimensional vector $\boldsymbol{v}$.

Table 1. Spectral basis selection procedure

```
Calculate the sparseness value of every input vector, $\boldsymbol{v}_t$, using (5);
Normalize every input vector;
repeat until the number candidates $<$ threshold or all input vectors are eliminated
  Select a candidate with the highest sparseness value;
  Estimate the fundamental frequency bin for each input vector;
  Align each input vector such that its frequency bin location is the same as the candidate;
  Calculate Euclidean distances between the candidate and every input vector;
  Cluster input vectors using Euclidean distances;
  Eliminate input vectors in the cluster which the candidate belongs to;
end (repeat)
repeat for every possible combination of candidates
  Set all candidate vectors as input vectors;
  Select a combination of candidates;
  Learn a encoding matrix, through the overlapping NMF,
    with fixing these selected candidates as basis vectors;
  Compute the reconstruction error of the overlapping NMF;
end (repeat)
Select the combination of candidates with the lowest reconstruction error;
```

The first part of our spectral basis selection method starts with choosing a candidate vector that has the largest sparseness values (see Fig. 1 (d)). Then we estimate the location of fundamental frequency bin for each input vector, which corresponds to the lowest frequency bin above the mean value. Each input vector is aligned to the candidate vector such that the fundamental frequency bin appears at the same location as the candidate vector. Euclidean distances between these aligned input vectors and the candidate vectors are calculated and a hierarchical clustering method (or any other clustering methods) is applied to eliminate whatever vectors belong to a group which the candidate vector belongs to. This procedure is repeated until we choose a pre-specified number of candidate vectors. Increasing this pre-specified number provides more feasible candidate vectors, however, the computational complexity in the second part increases. The repetition procedure produces several candidates, some of which are expected to represent a original musical instrument sound in such a way that a set of vertically-shifted basis restores the original sound.

The second part of our method is devoted for the final selection of representative spectral basis vectors from candidates obtained in the first part. Candidate vectors are regarded as input vectors for the overlapping NMF. For every possible combination of candidates (for the case of 2 sources, 2 out of the number of candidates), we learn an associated encoding matrix with selected candidates fixed as basis vectors, and calculate the reconstruction error. Final representative spectral basis vectors are the ones which give the lowest reconstruction error.



(a) (b) (c)

(d) (e) (f)

Fig. 1. Spectrograms of original sound of voice and a single string of a cello are shown in (a) and (b), respectively. Horizontal bars reflect the structure of harmonics. One can see that every note is the vertically-shifted version of each other if their musical instrument sources are the same. Monaural mixture of voice and cello is shown in (c) where 5 candidate vectors selected by our algorithm in Table 1 are denoted by dotted or solid vertical lines. Two solid lines represent final representative spectral basis vectors which give the smallest reconstruction error in the overlapping NMF. Each of these two basis vectors is a representative one for voice and a string of cello. Associated sparseness values are shown in (d) where black dots on a graph are associated with the candidate vectors. Unmixed sound is shown in (e) and (f) for voice and cello, respectively.

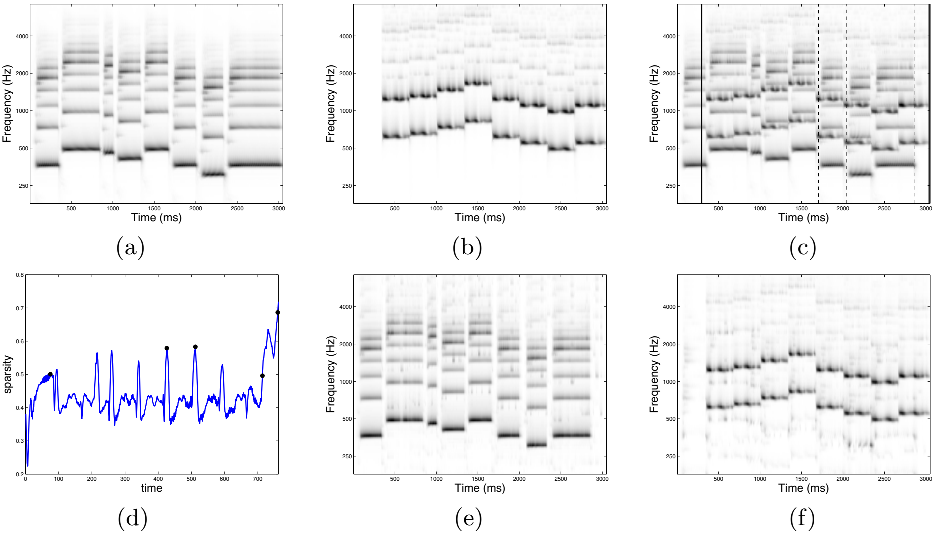
4 Numerical Experiments

We present two simulation results for monaural instantaneous mixtures of: (1) voice and cello; (2) saxophone and viola. We apply our spectral basis selection method with the overlapping NMF to these two data sets. Experimental results are shown in Fig. 1 and 2 where figure captions describe detailed results.

The set of transformation matrices, $\mathcal{T}$, that we used, is

$$\mathcal{T} = \left\{ \boldsymbol{T}^{(k)} \,\middle|\, \boldsymbol{T}^{(k)} = \overset{\overset{k-m}{\longmapsto}}{\boldsymbol{I}}, \quad 1 \le k \le 2m-1 \right\}, \tag{6}$$

where $\boldsymbol{I} \in \mathbb{R}^{m \times m}$ is the identity matrix and $\overset{\overset{j}{\longmapsto}}{\boldsymbol{I}}$ leads to the shift-up or shift-down of row vectors of $\boldsymbol{I}$ by j, if j is positive or negative, respectively. After shift-up or -down, empty elements are zero-padded.



(a) (b) (c)

(d) (e) (f)

Fig. 2. Spectrograms of original sound of saxophone and viola are shown in (a) and (b), respectively. Every note is artificially generated by changing the frequency of a real sample sound, so that the spectral character of each instrument is constant in all the variations of notes. Monaural mixture is shown in (c) where 5 selected candidate vectors are denoted by vertical lines. Each of two solid lines among them represents final representative spectral basis vector of each instrument. Associated sparseness values are shown in (d) where black dots associated with the candidate vectors are marked. Unmixed sound is shown in (e) and (f) for saxophone and viola, respectively.

For the case where $m = 3$ and $k = 2$, $\boldsymbol{T}^{(2)}$ and $\boldsymbol{T}^{(5)}$ are defined as

$$\boldsymbol{T}^{(2)} = \overset{\overset{2-3}{\longmapsto}}{\boldsymbol{I}} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \boldsymbol{T}^{(5)} = \overset{\overset{5-3}{\longmapsto}}{\boldsymbol{I}} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{7}$$

Multiplying a vector by these transformation matrices, leads to a set of vertically-shifted vectors.

5 Discussions

We have presented a method of spectral basis selection for single channel polyphonic music separation, where the harmonics, sparseness, clustering, and the overlapping NMF were used. Rather than learning spectral basis vectors from the data, our approach is to select a few representative spectral vectors among given data and fix them as basis vectors to learn associated encoding variables through the overlapping NMF, in order to restore unmixed sound. The success of our approach lies in the assumption that the distinguished timbre of a given musical instrument can be expressed by a transform-invariant time-frequency representation, even though their pitches are varying. A string instrument has multiple distinguished harmonic structures. In such a case, it is reasonable to assign a basis vector for each string.

Acknowledgments. This work was supported by ITEP Brain Neuroinformatics Program and ETRI.

References

1. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401** (1999) 788–791
2. Smaragdis, P.: Non-negative matrix factor deconvolution: Extraction of multiple sound sources from monophonic inputs. In: Proc. Int'l Conf. Independent Component Analysis and Blind Signal Separation, Granada, Spain (2004) 494–499
3. Smaragdis, P., Brown, J.C.: Non-negative matrix factorization for polyphonic music transcription. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY (2003) 177–180
4. Cho, Y.C., Choi, S.: Nonnegative features of spectro-temporal sounds for classfication. Pattern Recognition Letters **26** (2005) 1327–1336
5. Plumbley, M.D., Abdallah, S.A., Bello, J.P., Davies, M.E., Monti, G., Sandler, M.B.: Automatic transcription and audio source separation. Cybernetics and Systems (2002) 603–627
6. Eggert, J., Wersing, H., Körner, E.: Transformation-invariant representation and NMF. In: Proc. Int'l Joint Conf. Neural Networks. (2004)
7. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems. Volume 13., MIT Press (2001)
8. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research **5** (2004) 1457–1469