Single Channel Polyphonic Music Separation Using Spectral Basis Selection and Overlapping NMF

Minje Kim and Seungjin Choi
Department of Computer Science and Engineering, POSTECH
{minjekim, seungjin}@postech.ac.kr

1. Introduction

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 [1], 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.

2. Overlapping NMF

Nonnegative matrix factorization (NMF) is a simple but efficient factorization method for decomposing multivariate data into a linear combination of basis vectors with nonnegativity constraints for both basis and encoding matrix [2]. Given a nonnegative data matrix $\mathbf{V} \in \mathbb{R}^{m \times N}$ (where $V_{ij} \geq 0$), NMF seeks a factorization

$$\mathbf{V} \approx \mathbf{WH}, \quad (1)$$

where $\mathbf{W} \in \mathbb{R}^{m \times n} (n \leq m)$ contains nonnegative basis vectors in its columns and $\mathbf{H} \in \mathbb{R}^{n \times N}$ represents the nonnegative encoding variable matrix. Appropriate objective functions and associated multiplicative updating algorithms for NMF can be found in [3].

The overlapping NMF is an interesting extension of the original NMF, where transform-invariant representation and a sparseness constraint are incorporated with NMF [1]. 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 such representative basis vectors such that fewer number of basis vectors could reconstruct observed data. Given a set of transformation matrices, $\mathcal{T} = \{\mathbf{T}^{(1)}, \mathbf{T}^{(2)}, \ldots, \mathbf{T}^{(K)}\}$, the overlapping NMF finds a nonnegative basis matrix $\mathbf{W}$ and a set of nonnegative encoding matrix $\{\mathbf{H}^{(k)}\}$ (for $k = 1, \ldots, K$) which minimizes

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

where $\|\cdot\|_F$ represents Frobenious norm. As in [3], the multiplicative updating rules for the overlapping NMF were derived in [1].

3. Spectral Basis Selection

The goal of spectral basis selection is to choose a few representative vectors from $\mathbf{V} = [\mathbf{v}_1 \cdots \mathbf{v}_N]$ where $\mathbf{V}$ is the data matrix associated with the spectrogram of mixed sound. In other words, each column vector of 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 $\mathbf{V}$ using a sparseness measure and a clustering technique. We use the sparseness measure proposed by Hoyer [4], described by

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

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

4. Numerical Experiments

We show a simulation results for monaural mixture of voice and cello. Experimental results are shown in Fig. 1.

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

$$\mathcal{T} = \left\{ \mathbf{T}^{(k)} \mid \mathbf{T}^{(k)} = \overrightarrow{\mathbf{I}}^{k-m}, \quad 1 \leq k \leq 2 \times m - 1 \right\}, \quad (6)$$

where $\mathbf{I}$ is the $m \times m$ identity matrix. For the case where $m = 3$ and $k = 2$, $\mathbf{T}^{(2)}$ is defined as

Table 1. Spectral basis selection procedure.

Calculate the sparseness value of every input vector, $\mathbf{v}_t$, using (5);
Normalize every input vector;
repeat until
the number of 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;

$$\mathbf{T}^{(2)} = \overrightarrow{\mathbf{I}}^{2-3} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \qquad (7)$$

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.

References

[1] Eggert, J., Wersing, H., K¨orner, E.: Transformation-invariant representation and NMF. In: Proc. Int'l Joint Conf. Neural Networks. (2004)
[2] Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401 (1999) 788-791
[3] Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems. Volume 13., MIT Press (2001)
[4] Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research 5 (2004) 1457-1469

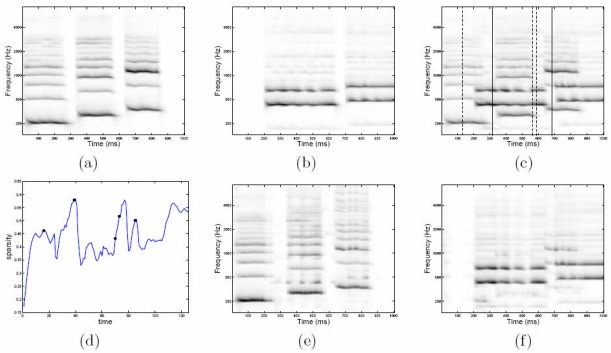

Fig. 1. Spectrograms of original sound of voice and a single string of a cello are shown in (a) and (b), respectively. Horizontal bars reflect the structure of harmonics. One can see that every note is the vertically-shifted version of each other if their musical instrument sources are the same. Monaural 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.