MIREX 2016 Submission for Singing Voice Separation

Yi-Chun Huang
Master Program of SMIT
National Chiao Tung University, Taiwan
hunejaap@gmail.com

Tai-Shih Chi
Dept. of Elec. & Comp. Engineering
National Chiao Tung University, Taiwan
tschi@mail.nctu.edu.tw

ABSTRACT

The nonnegative matrix factorization (NMF), which learns dictionaries from source spectra and uses the learned dictionaries to decompose the mixture in the test phase, is a widely used tool for audio source separation. However, the standard NMF does not consider temporal continuity of the signals when learning dictionaries. Besides, the learned dictionaries should be partitioned into subgroups to account for sources with different spectro-temporal properties, such as speech signals from different speakers or music signals from different instruments. Therefore, we propose a method by combine two extensions of NMF to address the requirements of continuity and partitioning for singing voice separation. For continuity, our method adopts a post-filtering technique [1], which derives a source specific vector autoregressive (VAR) model to smooth the NMF coefficients in the test phase. For partitioning, we make use of the mixture of local dictionaries (MLD) [2] technique to divide dictionaries into subgroups by considering intra- and inter- group distances. To sum up, our NMF-extended singing voice separation method put additional considerations on the temporal continuity of each subgroup.

1. INTRODUCTION

The nonnegative matrix factorization (NMF), which learns dictionaries from source spectra during training and then uses the learned dictionaries to decompose the mixture in the test phase, is widely used for audio source separation. However, the standard NMF only considers spectral patterns but not temporal dynamics of the signals. The temporal continuity is critical in decoding/encoding a sound signal with high fidelity. Therefore, our goal is to develop a NMF-based separation method with more smoothed temporal responses. For this matter, we adopt a post-filtering technique, which uses the source specific vector autoregressive (VAR) model [1] to smooth the NMF coefficients during test.

In singing voice separation, the music accompaniments comprise of percussive and harmonic music signals, which have totally different spectro-temporal patterns [3]. Therefore, we think the dictionaries should be arranged in subgroups during training to account for singing voice, percussive and harmonic signals respectively. For this matter, we adopt the mixture of local dictionaries (MLD) technique [2] to divide dictionaries into subgroups while preserving group sparsity on NMF coefficients.

The rationale of our approach is that singing voice, percussive and harmonic signals should have different degrees of temporal smoothness, hence, should have their own group-wise VAR models. In our method, we first use the MLD dictionary learning algorithm [2] to extract local dictionaries from source signals while considering group level sparsity and the diversity within each group. The individual VAR model is then estimated from NMF coefficients of each subgroup. In the test phase, the learned dictionaries are first used to decompose mixtures. Then for each subgroup, the post-smoothing filter [1] derived from its VAR model is applied on NMF coefficients.

2. METHOD

2.1 Mixture of Local Dictionaries

In order to preserve local properties of each subgroup's dictionary, the source spectrogram is first partitioned into G clusters by K-means to find their spectral centroids. The dictionary of each subgroup is then initialized with one spectral centroid as *a priori*. Then the MLD algorithm [2] is adopted to learn local dictionaries while considering the l_1 penalty on their distances away from the corresponding centroid. The objective function is defined as follows:

$$D(\mathrm{V}|\mathrm{WH})+\lambda\sum_t \Omega(h_t)+\eta\sum_g D\left(\mu^{(g)}|W^{(g)}\right) \qquad (1)$$

where $W^{(g)}$ and $\mu^{(g)}$ arethe local dictionary and the centroid of the g-th cluster, and h_t are the NMF coefficients. The λ and η parameters are the weights of group-sparsity and difference penalty between dictionaries and their centroids. Detailed descriptions can be found in [2].

2.2 Prediction Based Filtering

In [1], the NMF was first applied on each source to learn the spectral bases and the coefficients for each spectrogram. These coefficients were used to estimate the VAR

model of each source in a global manner. In the test phase, the mixture was first decomposed into pairs of bases and coefficients with learned bases. These VAR models were then used to smooth the NMF coefficients of the mixture.

2.3 Proposed Method

In our method, we use the MLD to obtain local groups and we estimate the VAR model for each subgroup of each source. The coefficients of source *s* and subgroup *g* at time *t* are smoothed as follows:

$$h_t^{s,g} \leftarrow h_t^{s,g} \odot \left(h_{t|t-1}^{s,g}\right)^{\beta^s} \quad (2)$$

where β^s is the filtering weight for source *s*, and $h_{t|t-1}^{s,g}$ is the prediction based on coefficients at past time. The prediction is derived as follows:

$$h_{t|t-1}^{s,g} = A^{s,g} h_{t-1}^{s,g} \quad (3)$$

where $A^{s,g}$ is the source-specific and group-wise VAR model and can be directly estimated from NMF coefficients during training as in [1].

3. EXPERIMENTS

Our model was trained using the public 252-clip subset of iKala dataset.[1] For scalability reasons, we used the first 500 frames of each clip to compute the short-time Fourier transform (STFT) spectrogram. For the spectrogram, we used the window length of 2048 samples and the frame hop length of 512 samples. We set $G^{singing} = 4$ groups for singing voice, $G^{accompany} = 4$ groups for accompaniments, and each group had 25 bases. The λ and η were set to 256 and 0.01.The filtering weights for singing voice and accompaniments should be $\beta^{singing} = 0.6$ and $\beta^{accompany} = 0.1$.

4. REFERENCES

[1] N. Mohammadiha, P. Smaragdis, and A. Leijon: "Prediction based filtering and smoothing to exploit temporal dependencies in NMF,"*Proc. ICASSP*, pp. 873-877, 2013.

[2] K. Minje, and P. Smaragdis: "Mixtures of local dictionaries for unsupervised speech enhancement," *IEEE Signal Processing Letters*, 22.3, pp. 293-297, 2015.

[3] F. Yen, Y.-J. Luo, and T.-S. Chi, "Singing voice separation using spectro-temporal modulation features," *Proc. ISMIR*, pp. 617-622, 2014.

[1] http://mac.citi.sinica.edu.tw/ikala/