AUDIO SOURCE SEPARATION BY PROBABILISTIC LATENT COMPONENT ANALYSIS

Yinyi Guo, Mofei Zhu

Center for Computer Research in Music and Acoustics
Stanford University

ABSTRACT

The problem of audio source separation from a monophonic sound mixture having known instrument types but unknown timbres is presented. An improvement to the Probabilistic Latent Component Analysis (PLCA) source separation method is proposed. The technique uses a basis function dictionary to produce a first round PLCA source separation. The PLCA weights are then refined by incorporating note onset information. The source separation is then performed using a second round PLCA in which the refined weights are held fixed, and the basis functions are updated. Preliminary experimental results on mixtures of two instruments are quite promising, showing a 6 dB improvement in SIR over standard PLCA.

Index Terms— Audio Source Separation, Probabilistic Latent Component Analysis, Basis Function Adaptation, Onset Detection

1. INTRODUCTION

Source separation of audio has been an active research topic in audio signal processing. In recent years, a great deal of the work in this area is based on spectral decomposition methods. Probabilistic latent component analysis (PLCA) [1] is one such method. To achieve good separation with few artifacts, it has a high demand of the prior knowledge of the target sound in the mixture. Not only do the sound types need to be known, but precise basis functions of the target timbres should be pre-trained as well. [2]

In this paper, we extend the PLCA based supervised separation method described in [2]. We use pre-trained spectral basis of general instruments as the spectral prior knowledge to lead the algorithm to give preliminary temporal results, which can be optimized further. We then perform onset detection based post-processing to adjust the temporal information given by the previous stage decomposition. Eventually, the source separation is performed and the basis functions of target sources in the mixture are updated with the refined temporal weights held fixed using a second round PLCA. An overview of the proposed approach is shown in Fig. 1.

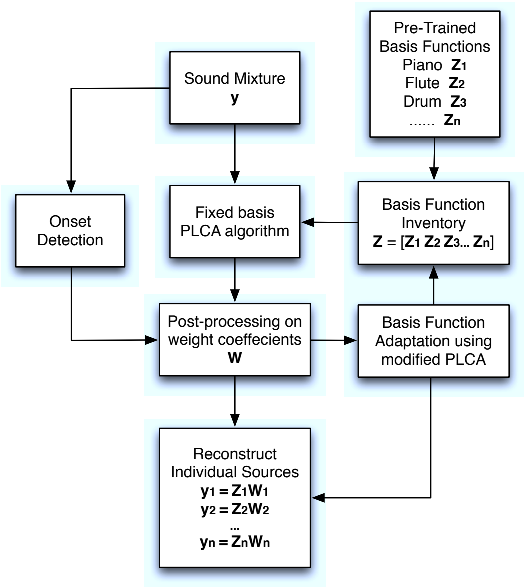
Pre-Trained Basis Functions
Piano Z_1
Flute Z_2
Drum Z_3
...... Z_n
Sound Mixture
y
Onset Detection
Fixed basis PLCA algorithm
Basis Function Inventory
$Z = [Z_1\ Z_2\ Z_3 ...\ Z_n]$
Post-processing on weight coeffecients
W
Basis Function Adaptation using modified PLCA
Reconstruct Individual Sources
$y_1 = Z_1W_1$
$y_2 = Z_2W_2$
...
$y_n = Z_nW_n$

Fig. 1. System overview flowchart

2. PROPOSED METHOD

2.1. PLCA Model

The Probabilistic Latent Component Analysis (PLCA) algorithm is the most important element of our system, which performs the source separation. PLCA is a probabilistic model, whose components are all non-negative. It interprets the spectrogram as a histogram and the spectral and temporal components as distributions along time and frequency. Numerically, this probabilistic latent spectrogram factorization method is similar to the non-negative matrix factorization (NMF) approach [3][4]. The magnitude spectrogram of audio signals can be treated as a 2D distribution of "sound quanta" over time and frequency. In this respect, the non-negative factorization is used to model the spectrogram as a linear combination of spectral basis vectors, which are regarded as the latent variables. The model is defined as follows,

1. $P(f \mid z)$ is defined as a multinomial probability distribution of frequency f given certain latent variable z . Each distribution is actually a spectral basis vector of

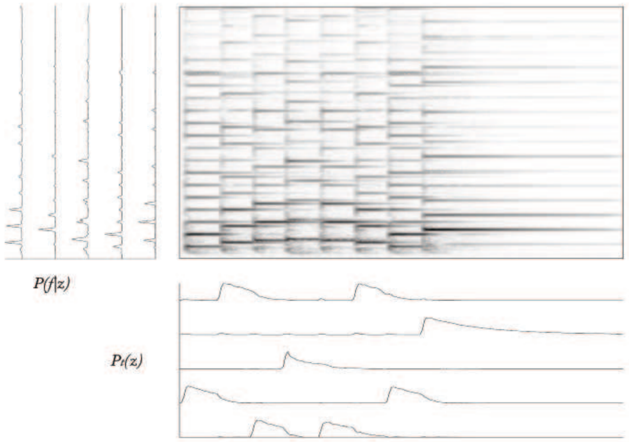

Fig. 2. An illustration of applying PLCA to a spectrogram of eight notes that are by a piano.

sound component.

2. $P_t\,(z)$ is defined as a multinomial probability distribution of weights for each latent component z at time t .

3. $P\,(t)$ is a distribution of the relative magnitudes at different time frame t.

4. $P_t\,(f)$ is defined as normalized spectrogram vector at time frame t

5. $S_t\,(f)$ is defined the original spectrogram vector at time frame t

$$S_t\,(f) \;=\; P\,(t)\,P_t\,(f) \tag{1}$$

$$P_t\,(f) \;=\; \sum_z P\,(f \mid z)\,P_t\,(z) \tag{2}$$

Equation 1 is to normalize S_t into $P_t\,(f)$, Equation 2 is the magnitude decomposition of the spectrogram into spectral basis vectors and weights vectors. An example of applying this decomposition is shown in Fig 2.

We can see five distinct notes in Fig 2. Each of the notes is represented by a latent variable z. The spectral basis $P\,(f \mid z)$ of each component z can be seen as the harmonic series of each note.

The expectation-maximization (EM) algorithm can be used to decompose the spectrogram $P_t\,(f)$ into spectral basis vectors $P\,(f \mid z)$ and corresponding weights $P_t\,(z)$. Given the mixture spectrogram, we randomly initialize all the unknown parameters, and estimate the model parameters by iterating between E step and M step until convergence.

- E Step - Estimate the posterior given spectral basis vectors and weights vectors

$$P_t\,(z \mid f) \;=\; \frac{P_t\,(f \mid z)\,P\,(z)}{\sum_z P_t\,(f \mid z)\,P\,(z)} \tag{3}$$

- M Step - Estimate spectral basis vectors and weights vectors given the posterior

$$P_t\,(z) \;=\; \frac{\sum_f P_t\,(f)\,P_t\,(z \mid f)}{\sum_z \sum_f P_t\,(f)\,P_t\,(z \mid f)} \tag{4}$$

$$P\,(f \mid z) \;=\; \frac{\sum_t P_t\,(f)\,P_t\,(z \mid f)}{\sum_t \sum_f P_t\,(f)\,P_t\,(z \mid f)} \tag{5}$$

The source separation applying the standard PLCA model is conducted as follows: [2][5]

- Calculate the spectrogram $S_t(f)$ of a sound mixture.

- Learn the basis function $P\,(f \mid z_m)$ of the sources Z_m in the mixture using PLCA.

- Initialize the prior-based PLCA algorithm with the learned basis functions of the active components, and estimate the model parameters $P\,(f \mid z_m)$ and $P_t\,(z)$.

- Reconstruct each instrument in the mixture by only using their corresponding optimized $P\,(z_m)$, $P\,(f \mid z_m)$ The complete process of our proposed approach is described in the following sections.

2.2. Fixed-Basis PLCA

We start with a fixed-basis PLCA which is first performed to get weights of individual sources. Given that the types of instrument in the mixture are known, we choose accordingly the basis function of target sources from the pre-trained typical spectral basis dictionary and fix them as spectral prior information for the PLCA algorithm. Since the timbre characteristics are very similar among the same type of instrument, this fixed-basis PLCA gives us more robust weights estimation comparing with using the traditional PLCA, in which the spectral basis and weights are not fixed. In order to separate the target sources in the mixture precisely, the spectral basis should be updated for reconstruction. To achieve this, more reliable temporal prior knowledge has to be prepared for the next stage PLCA: first, we need to further refine the current weight coefficients that may contain unwanted crossover between notes or instruments. Second, the right timing when should update the spectral basis function has to be found based on the refined weights. We therefore perform an onset detection based post-processing on the weights as a preparation for the second stage PLCA.

2.3. Onset Detection

An amplitude envelope function is calculated from the Short Time Fourier Transformation of the input mixture signal. Based on algorithms from previous literature [6, 7], the well-known first order difference function of the spectral energy envelopes is computed, which gives prominent peaks from which onsets can be located. A threshold was generated in order to determine if a peak is an onset candidate [8]. Every value above this threshold in the first order difference function of the energy envelope should be a potential onset peak.

2.4. Post-processing on Weight Coefficients

An onset implies something new happened in a signal. In the case of a musical signal, this could be a change in pitch or instrument, i.e. the precise time when a new note is produced by an instrument. Knowing the location of the onsets allows us to easily determine which instruments are playing. This can be accomplished by comparing the intensities of their weights envelopes in the frames following the onset frame, since the weights envelopes could be considered as activation indicators of sources. Thus it is more reliable to make decisions to remove residual cross-talk components right after onset frames. After we detect each onset, the weight coefficients of different notes are compared against their sums of magnitude during the following 0.15 seconds. For each instrument, the note that has the maximum weight[1] coefficient among all the weights is taken as an active note, while the weights of the other notes are removed. This refinement on the weights is operated after each onset. Eventually, the weights coefficients of active notes and their corresponding basis functions are restored for the further processing.

2.5. Update Basis Function

We record those appropriate time frames in which the weights coefficients after post-processing are relatively reliable to update corresponding basis functions of certain source. If the sum of the weights of active notes are dominant in some time frames between two successive onsets, another modified PLCA algorithm based on EM learning rule is performed to update the basis functions of the active notes, as follows:

- For the initialization of the EM algorithm, the basis functions $P(f \mid z)$ are constructed as the basis functions of active notes $P(f \mid z_a)$ cascaded with a set of randomly distributed basis functions $P(f \mid z_r)$ for modeling residual sources other than the dominant active notes. The new weight vector $P_t(z)$ consists of weights of active notes $P_t(z_a)$ as well as the weights of random basis $P_t(z_r)$ that share the rest of the weights.
- The expectation step remains the same as Equation 3.
- The maximization step equations are given by the following update equations:

$$P_t(z_r) = \frac{\sum_f P_t(f) P_t(z_r \mid f)}{\sum_z \sum_f P_t(f) P_t(z_r \mid f)} \quad (6)$$

$$P(f \mid z) = \frac{\sum_t P_t(f) P_t(z \mid f)}{\sum_t \sum_f P_t(f) P_t(z \mid f)} \quad (7)$$

The above equations are iterated until convergence. We only update the weights for random sources $P_t(z_r)$ while the weights for active latent variables $P_t(z_a)$ are fixed in the M-step in that they are ensured to be reliable temporal prior information given by the first round PLCA and post-processing. The chosen of the number of random basis depends mainly on the number of sources in the mixture, since those extra randomly initialized basis functions $P(f \mid z_r)$ are supposed to explain the artifacts introduced by the algorithm as well as the sustain or release potions of any source before. The basis functions of dominant components $P(f \mid z_a)$ will therefore be biased to update towards the target source timbres eventually based on the fixed weight coefficients.

3. EXPERIMENTS AND RESULTS

Our algorithm has been tested on synthetic examples of 20 distinct data sets (mixture of two instruments per set), which include 5 distinct instruments (piano, flute, strings, saxophone and guitar). All the examples are generated from Logic Pro. The source timbres of the mixture are different from those in the pre-trained basis functions in the inventory. For example, in the basis functions inventory an octave of piano basis functions from C4 to C5 are trained from piano sound source of Steinway; An octave of flute basis functions from C6 to C7 are trained from Super Air Flute . The corresponding test data is mixture sound sources of Yamaha piano and Thin flute. In all these experiments we measured the Signal to Distortion Ratio (SDR), the Signal to Interference Ratio (SIR), and the Signal to Artifacts Ratio (SAR) as defined in [9]. The results are shown in Table 1,2. We can see the proposed algorithm enjoys around 6dB improvement in SIR over traditional PLCA.

In these experiments, we used 5 basis functions per note, ran 10 times of 20 distinct experiments, each spectral basis and spectrogram of 2048-point FFT with 75% overlap, 80 iterations of the PLCA algorithm. The approach works on magnitude spectrograms; for the phase we copy the phase of the original mixture to each of the separated sources. The resulting separated spectrograms of the piano-flute mixture can be seen in Fig 3.

[1] In this project we work on cases of single note concurrently for each instrument

Piano	SIR(dB)	SAR(dB)	SDR(dB)
Traditional PLCA	25.86	4.90	4.85
Proposed Algorithm	32.96	8.46	8.45

Flute	SIR(dB)	SAR(dB)	SDR(dB)
Traditional PLCA	11.21	11.79	8.33
Proposed Algorithm	15.60	15.52	12.16

Table 1. Performance metrics for synthetic mixture example of piano-flute

	SIR(dB)	SAR(dB)	SDR(dB)
Improvement	6.05	1.71	1.86

Table 2. Averaged metric improvement over traditional PLCA for all the 20 experimental datasets

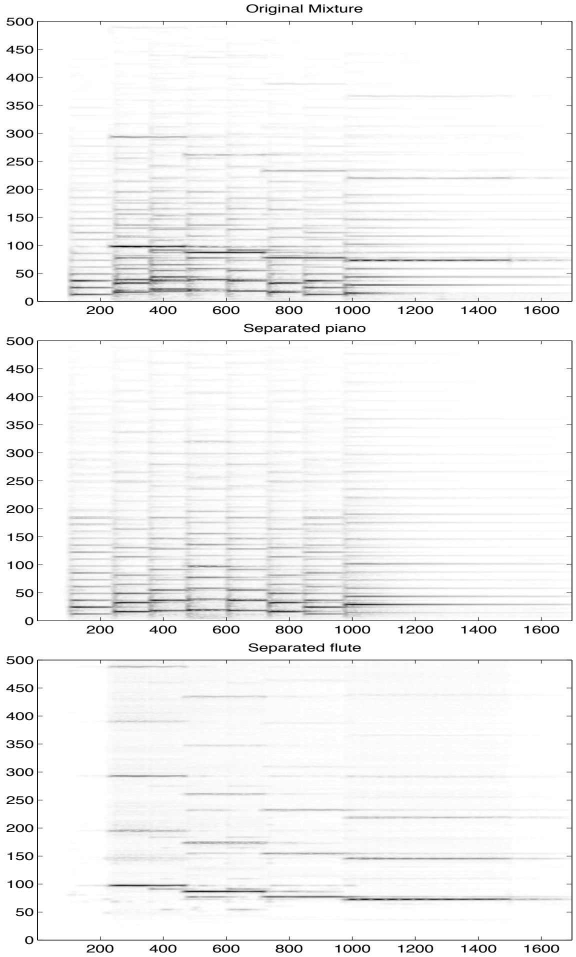

Fig. 3. Example result of separated piano and flute spectrograms from a mixture. The top plot shows the input mixture spectrogram, the middle one shows the separated piano spectrogram, the bottom one shows the separated flute spectrogram. x-axis is the time frames. y-axis is the frequency bins using 2048-point FFT with 75% overlap

4. CONCLUSION

We have presented a learning scheme based on PLCA algorithm that is used for the source separation of multiple concurrent instruments knowing the type of instruments in the mixture but without a prior knowledge of the precise basis functions of target sources beforehand. The proposed algorithm has been demonstrated on mixtures of two instruments. The use of prior generic basis functions inventory, post-processing on weight coefficients as well as the basis function adaptation have been shown to improve the performance of the original PLCA algorithm. This method shows promising results and in future work, we plan to improve the performance by studying perceptual auditory model.

5. REFERENCES

[1] M. Shashanka P. Smaragdis, B. Raj, "A probabilistic latent variable model for acoustic modeling," *Advances in models for acoustic processing, NIPS*, 2006.

[2] M. Shashanka P. Smaragdis, B. Raj, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *International Conference on Independent Component Analysis and Signal Separation*, 2007.

[3] H. Seung D. Lee, "Algorithms for non-negative matrix factorization," *NIPS*, 2001.

[4] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as non-negative factorizations," *Computational Intelligence and Neuoscience Journal, special issue on Advances in Non-negative Matrix and Tensor Factorization*, 2008.

[5] P. Smaragdis and G. J. Mysore, "Separation by humming: User-guided sound extraction from monophonic mixtures," *Proc. of IEEE Workshop on Applications Signal Processing to Audio and Acoustics*, 2009.

[6] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Am., vol. 104, pp. 588–601*, 1998.

[7] M. Goto and Y. Muraoka, "Beat tracking based on multiple-agent architecture - a real-time beat tracking system for audio signals," *Proceedings of The Second International Conference on Multiagent Systems, pp.103–110,*, 1996.

[8] C. Duxbury, M. Sandler, and M. Davies, "A hybrid approach to musical note onset detection," *Proceedings of the Digital Audio Effects Conference, Hamburg*, 2002.

[9] C. Fevotte, R. Gribonval, and E. Vincent., "Bss eval toolbox user guide," *IRISA Technical Report 1706, Rennes, France*, 2005.