COLLABORATIVE AUDIO ENHANCEMENT USING PROBABILISTIC LATENT COMPONENT SHARING

Minje Kim *

University of Illinois at Urbana-Champaign
Department of Computer Science
IL, USA

Paris Smaragdis

University of Illinois at Urbana-Champaign
Adobe Systems Inc.

ABSTRACT

This paper presents a collaborative audio enhancement system that aims to recover common audio sources from multiple recordings of a given audio scene. We do so in the context where each recording is uniquely corrupted. To this end, we propose a method of simultaneous probabilistic latent component analyses on synchronized inputs. In the proposed model, some of the parameters are fixed to be same during and after the learning process to capture common audio content while the rest models unwanted recording-specific interferences and artifacts. Our model also allows for prior knowledge about the parameters of the model, e.g. representative spectra of the components, to be incorporated in the factorization. A post processing scheme that consolidates the extracted sources from the set of inputs is also proposed to handle the possible loss of certain frequency regions. Experiments on commercial music signals with various artifacts show the merit of the proposed method.

Index Terms— Probabilistic Latent Component Analysis, Nonnegative Matrix Partial Co-Factorization, Convolutive Common Nonnegative Matrix Factorization, Crowdsourcing

1. INTRODUCTION

Because of widespread use of hand-held devices, we often find many overlapping recordings of an audio scene. Our goal in this paper is to fully utilize these low cost noisy data by extracting common audio sources from them so as to produce a higher quality rendering of the recorded event. Hence, it can be seen as a collaborative approach to audio enhancement sharing some similar concepts with crowdsourcing methods [1, 2]. The first step towards unifying these recordings is to synchronize them, something we can easily achieve using one of the efficient and robust synchronization methods proposed in the past [3, 4]. Once this is done, one could simply use the best available recording at any point in time, assuming there is an automated way of quality-ranking the signals. This can be the simplest implementation of *collaborative audio enhancement*, where we can take advantage of other people's recordings to improve ours. However, such simple reasoning does not work for many common cases, so we will address this problem using a different approach.

Fig. 1 shows a case where the obvious approach might fail. Between the two synchronized recordings, we cannot simply choose one because both are deficient, albeit in a different way. The bottom recording has a poor high frequency response, which could be the effect of a low-cost microphone or aggressive audio coding. On the other hand, the full bandwidth recording at the top has some interference in the 3 – 4.3 second region, which is however not present in the bottom one.

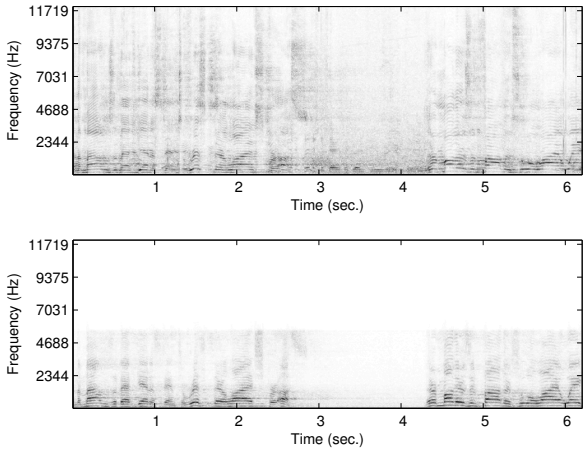

Fig. 1: An example of a difficult scenario, when a synchronization and selection method can easily fail to produce a good recording. In this case we observe unwanted interference (top) and the other is band-limited (bottom).

As the number of input recordings increases, the unique distortions in each recording make choosing a single best recording difficult, if not impossible. One could encounter various types of nonlinear artifacts or interferences, e.g., the audience chatter near the microphone, lens zooming noises, button clicks, clipping, band-pass filtering, etc. Eventually we would like to solve this problem by using information from all recordings and combine it appropriately in order to produce a higher quality render.

Nonnegative Matrix Partial Co-Factorization (NMPCF) was proposed to extract common spectral components out of multiple music excerpts in the past. Its several versions focussed on various characteristics of drum sounds that are expected to be common across multiple signals: spectral similarity between the drum solo signals and the drum source components in the music mixture [5], repeatability of drum source components across all the chunks of the song [6], and their unified version [7]. Convolutive Common Nonnegative Matrix Factorization (CCNMF) was recently introduced to recover the common music and effect parts from multiple soundtracks with different languages [8]. CCNMF differs from

*This work was performed while at Adobe Systems Inc.

NMPCF in that it shares both basis vectors and corresponding encodings of the common components to extract the music and effects while the set-aside track-specific ones capture dialogues in particular languages.

The proposed method, Probabilistic Latent Components Sharing (PLCS), is based on the probabilistic counterparts of Nonnegative Matrix Factorization (NMF) [9, 10], such as Probabilistic Latent Semantic Indexing (PLSI) [11, 12] and Probabilistic Latent Component Analysis (PLCA) [13]. PLCS extends PLCA with the common component sharing concept. PLCS differs from the NMPCF-based methods in that it decomposes each input matrix into three parts, rather than just two, so that we can share both bases and encoding matrices while providing slack in the model by letting the weights of the components to not be shared. Because PLCS controls the contribution of the latent components with probabilistic weights, $P^{(l)}(z)$, it gives more intuitive interpretation of the roles of components in the reconstruction whereas in the CCNMF model [8] they are absorbed in the filtering factor. Moreover, because the whole process is based on the probabilistic model, we could explicitly take advantage of Bayesian approaches, which is not straightforward in either NMPCF or CCNMF. The Bayesian approach provides a straightforward way to involve a certain prior knowledge about the bases, which we can get in advance from the cleaner, but different versions of the similar sources. Finally, we propose an additional post processing method to efficiently consolidate recording-specific reconstructions.

This paper consists of the following sections. Section 2 is an introduction to PLCA from which the proposed model is originated. Section 3 and 4 provide the basic PLCS model and its version using prior information, respectively. They are followed by Section 5 where the post processing procedure is discussed. Section 6 shows the experimental results on real signals. Lastly, Section 7 summarizes the work.

2. SYMMETRIC PLCA

Given the magnitude of an input spectrogram, $V = |X|$, with elements $V_{f,t}$ indexed by the frequency bin f and the time frame t, symmetric PLCA tries to maximize the log-likelihood $\mathcal{P}$ of observing the energy quanta of $V_{f,t}$,

$$\mathcal{P} = \sum_{f,t} V_{f,t} \log P(f,t) = \sum_{f,t} V_{f,t} \log \sum_z P(f,t|z)P(z)$$
$$= \sum_{f,t} V_{f,t} \log \sum_z P(f|z)P(t|z)P(z).$$

To get the second equality, the component-specific distributions $P(f,t|z)$ is further factorized into three factors: the frequency distribution $P(f|z)$, its temporal activations $P(t|z)$, and the component specific weights $P(z)$. Note that the term "symmetric" came from this tri-factorization [11], which eventually let us have control over additional temporal distributions of components as well as frequency distributions. This being a latent variable model, we use the Expectation-Maximization (EM) algorithm to estimate its parameters. In the E-step we find a posterior probability of the latent variable z given the time and frequency indices,

$$P(z|f,t) = \frac{P(f|z)P(t|z)P(z)}{\sum_z P(f|z)P(t|z)P(z)}.$$

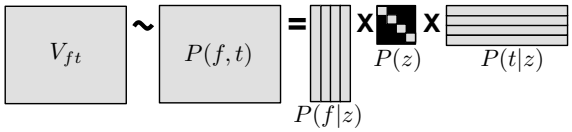

Fig. 2: A matrix representation of the PLCA with four components. Note that the weights $P(z)$ are represented as a diagonal matrix.

In the M-step the expected complete data log-likelihood is maximized, which yields to the following update rules:

$$P(f|z) = \frac{\sum_t V_{f,t}P(z|f,t)}{\sum_{f,t} V_{f,t}P(z|f,t)}, \quad P(t|z) = \frac{\sum_f V_{f,t}P(z|f,t)}{\sum_{f,t} V_{f,t}P(z|f,t)},$$
$$P(z) = \frac{\sum_{f,t} V_{f,t}P(z|f,t)}{\sum_{f,t,z} V_{f,t}P(z|f,t)}. \quad (1)$$

Fig. 2 depicts the relationship between the input matrix V and the estimated joint distribution $P(f,t)$ from which the observations were drawn.

3. PROBABILISTIC LATENT COMPONENTS SHARING

Let us assume that there are L input magnitude spectrogram matrices, corresponding to L available recordings in the collaborative audio enhancement application. We partition the values of the latent components in the l-th recording $z^{(l)}$ into two disjoint subsets, $z^{(l)} = z_C \cup z_I^{(l)}$, where z_C is the subset that contains indices of the common components shared across all recordings, and $z_I^{(l)}$ contains those of all the other components present only in the l-th recording. Now, the log-likelihood $\mathcal{P}$ of observing L given recordings can be written as:

$$\mathcal{P} = \sum_l \sum_{f,t} V_{f,t}^{(l)} \log \Bigg\{ \sum_{z \in z_C} P_C(f|z)P_C(t|z)P^{(l)}(z)$$
$$+ \sum_{z \in z_I^{(l)}} P_I^{(l)}(f|z)P_I^{(l)}(t|z)P^{(l)}(z) \Bigg\}. \quad (2)$$

The main new feature in (2) is to fix both the spectral and the temporal distributions to be same across all inputs for $z \in z_C$, which are specified as the common variables $P_C(f|z)$ and $P_C(t|z)$. On the other hand, components indicated by $z \in z_I^{(l)}$ represent recording-specific sound components, such as interferences, characterized by parameters $P_I^{(l)}(f|z)$ and $P_I^{(l)}(t|z)$. We refer to this model as PLCS, for which the E-step is:

$$P^{(l)}(z|f,t) = \frac{P^{(l)}(f|z)P^{(l)}(t|z)P^{(l)}(z)}{\sum_{z \in z^{(l)}} P^{(l)}(f|z)P^{(l)}(t|z)P^{(l)}(z)}, \forall z \in z^{(l)}.$$

Note that the parameters $P^{(l)}(f|z)$ and $P^{(l)}(t|z)$ can either refer to the common parameters $P_C(f|z)$ and $P_C(t|z)$ when $z \in z_C$ or $P_I^{(l)}(f|z)$ and $P_I^{(l)}(t|z)$ when $z \in z_I^{(l)}$, respectively.

Using Lagrange multipliers to ensure that the probability distributions sum to one, we maximize the expected complete data log-likelihood with following update rules as the M-step:

For $z \in z_I^{(l)}$

$$P_I^{(l)}(f|z) = \frac{\sum_t V_{f,t}^{(l)}P^{(l)}(z|f,t)}{\sum_{f,t} V_{f,t}^{(l)}P^{(l)}(z|f,t)}, \; P_I^{(l)}(t|z) = \frac{\sum_f V_{f,t}^{(l)}P^{(l)}(z|f,t)}{\sum_{f,t} V_{f,t}^{(l)}P^{(l)}(z|f,t)}, \quad (3)$$

For $z \in z_C$

$$P_C(f|z) = \frac{\sum_{l,t} V_{f,t}^{(l)} P^{(l)}(z|f,t)}{\sum_{l,f,t} V_{f,t}^{(l)} P^{(l)}(z|f,t)}, \quad P_C(t|z) = \frac{\sum_{l,f} V_{f,t}^{(l)} P^{(l)}(z|f,t)}{\sum_{l,f,t} V_{f,t}^{(l)} P^{(l)}(z|f,t)}, \tag{4}$$

For $z \in z^{(l)}$

$$P^{(l)}(z) = \frac{\sum_{f,t} V_{f,t}^{(l)} P^{(l)}(z|f,t)}{\sum_{z,f,t} V_{f,t}^{(l)} P^{(l)}(z|f,t)}.$$

Note that the updates for $P_C(f|z)$ and $P_C(t|z)$ include summation over l to involve all the reconstructions of common components.

4. INCORPORATING PRIORS

It is often useful to involve prior knowledge about the parameters in probabilistic models. For instance, we can have a clean recording of the same content as in the provided inputs, albeit recorded at a different time (e.g. a studio recording of a song whose recordings we obtain from a concert). Or, it is also possible to assume that the interferences are a certain kind of sources, e.g. human voice. On the other hand, we cannot simply learn the bases of those a priori signals and fix them as our target parameters, $P_C(f|z)$ or $P_I^{(l)}(f|z)$, as there is no guarantee that the a priori known signals have exactly the same spectral characteristics with the target sources. To address this problem we follow a Bayesian approach to derive a maximum a posteriori (MAP) estimator of the parameters.

First, we learn the bases of the magnitude spectrograms of the similar sources and interferences by directly applying PLCA update rules in (1). The learned bases vectors $P_{\text{source}}(f|z)$ and $P_{\text{interf}}^{(l)}(f|z)$ are used in the PLCS model to construct a new expected complete data log-likelihood

$$\begin{aligned}\langle \mathcal{P} \rangle = \sum_{l,f,t} V_{f,t}^{(l)} \Bigg\{ & \sum_{z \in z_C} \Big(P^{(l)}(z|f,t) \log P_C(f|z) P_C(t|z) P^{(l)}(z) \\ & + \alpha P_{\text{source}}(f|z) \log P_C(f|z) \Big) \\ & + \sum_{z \in z_I^{(l)}} \Big(P^{(l)}(z|f,t) \log P_I^{(l)}(f|z) P_I^{(l)}(t|z) P^{(l)}(z) \\ & + \beta P_{\text{interf}}^{(l)}(f|z) \log P_I^{(l)}(f|z) \Big) \Bigg\},\end{aligned}$$

where α and β controls the amount of influence of the prior bases. Once again, by using proper Lagrange multipliers, we can derive the final M-step with priors as follow:

For $z \in z_I^{(l)}$

$$P_I^{(l)}(f|z) = \frac{\sum_t V_{f,t}^{(l)} P^{(l)}(z|f,t) + \beta P_{\text{interf}}^{(l)}(f|z)}{\sum_{f,t} V_{f,t}^{(l)} P^{(l)}(z|f,t) + \beta P_{\text{interf}}^{(l)}(f|z)}, \tag{5}$$

For $z \in z_C$

$$P_C(f|z) = \frac{\sum_{l,t} V_{f,t}^{(l)} P^{(l)}(z|f,t) + \alpha P_{\text{source}}(f|z)}{\sum_{l,f,t} V_{f,t}^{(l)} P^{(l)}(z|f,t) + \alpha P_{\text{source}}(f|z)}. \tag{6}$$

E-step and the other M-step update rules are not changed from the original PLCS model. Fig. 3 summarizes the whole PLCS process on three different inputs: low-pass filtered, high-pass filtered, and mid-pass filtered inputs. All three inputs also contain unique distortions represented with different noise patterns in the figure.

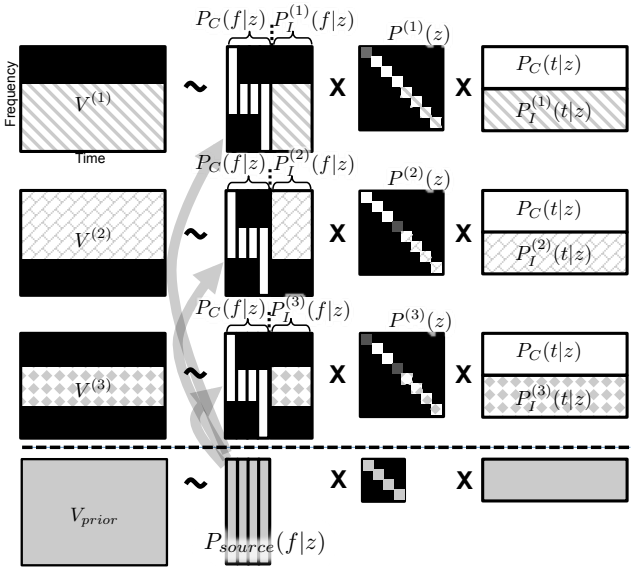

Fig. 3: An example of common source separation process using PLCS on three defected input matrices and prior information.

Note that the first common component of $l = 1$ case (first row) degrades the reconstruction as its basis vector has high frequency energy while $V^{(1)}$ was low-pass filtered. Therefore, the first weight in the diagonal matrix $P^{(1)}(z = 1)$ has a very low (dark) value. Similarly, $P^{(2)}(z = 4)$, $P^{(3)}(z = 1)$, and $P^{(3)}(z = 4)$ are also those weights that *turn off* inactive common components. Note also that the a priori learned bases $P_{\text{source}}(f|z)$ are full-banded and have somewhat different spectral shapes from the common bases, so they cannot replace the common bases as they are. To recover the magnitudes of the desired sources, we multiply the sum of the posterior probabilities of $z \in z_C$ to the input complex-valued spectrograms $X_{f,t}$,

$$\hat{S}_{f,t}^{(l)} = X_{f,t}^{(l)} \sum_{z \in z_C} P^{(l)}(z|f,t),$$

where $\hat{S}^{(l)}$ is the spectrogram of the separated sources from the l-th input.

5. POST PROCESSING

It is possible that the recorded signals exhibit non-uniform frequency responses due to recording device and format specifications. The PLCS method can identify the isolated common sources, but it is not expected to ameliorate effects like frequency response losses, since that information will be coded in the basis vectors and is not readily accessible as an artifact. We propose a collaborative post processing step to address this issue. Our approach is motivated by the fact that even if most of the recordings are filtered in some way, one recording that did not go through such filtering can give us enough information to recover the full-banded reconstruction. Suppose that we get L recovered spectrograms $\hat{S}_{f,t}^{(l)}$ using PLCS. The post processing begins with getting the normalized average magnitude spectrum of those reconstructions $y_f^{(l)} = \frac{\sum_t |\hat{S}_{f,t}^{(l)}|}{\sum_{f,t} |\hat{S}_{f,t}^{(l)}|}$. Now, we can obtain some global weights by considering the balance among different recordings in

each particular frequency bin, $w_f = \frac{\sum_l y_f^{(l)}}{\max_l y_f^{(l)}}$.

Then, the final complex spectrogram of the desired output is obtained by dividing the sum of the band-limited reconstructions with the corresponding elements of the global weight:

$$\hat{S}_{f,t} = \frac{\sum_l \hat{S}_{f,t}^{(l)}}{w_f}. \quad (7)$$

6. EXPERIMENTAL RESULTS

In this section, we compare the proposed PLCS models and other relevant methods in terms of signal-to-distortion ratio (SDR) [14], because we desire to reduce all the artifacts, noises, and interferences. To this end, we use five different single channel songs with 44.1kHz sampling rate and 16 bit encoding, each of which has a pair of versions: a 15 seconds-long clean live recording S as the source and a 30 seconds-long clean studio recording S_{prior} as the prior information of the source. The professional live recording S goes through three different sets of artificial deformations to simulate usual recording scenarios. The resulting three mixture spectrograms are:

- $X^{(1)}$: Low-pass filtering at 8kHz (a recording with a low sampling rate) / additional female speech as an interference
- $X^{(2)}$: High-pass filtering at 500Hz / additional female speech different from $X^{(1)}$ as an interference
- $X^{(3)}$: Low-pass filtering at 11.5kHz / high-pass filtering at 500Hz / clipping.

Short-time Fourier transform was applied to the signals with following settings: 1024 sample frame-length and 512 sample of hop size. For the priors, we get 100 bases for the source prior $P_{\text{source}}(f|z)$ from the studio recording S_{prior} while 50 interference prior bases $P_{\text{interf}}(f|z)$ are learned from anonymous female speeches [15].

We compare five different models, including:

- Median: Pixel-wise medians of L input magnitude spectrograms are calculated as follows

  $$|\hat{S}_{f,t}| = \text{median}(V_{f,t}^{(1)}, V_{f,t}^{(2)}, \cdots, V_{f,t}^{(L)}).$$

  The phase information of the sum of inputs is used to invert the reconstruction to the time domain.
- Oracle PLCA with ideal bases: We run PLCA on the source $|S|$ to get the ideal bases $P_{\text{ideal}}^{(l)}(f|z)$ that perfectly represent spectral characteristics of the desired source. And then, we apply PLCA again on each $V^{(l)}$ by initializing and fixing some of the bases with $P_{\text{ideal}}^{(l)}(f|z)$ while learning the others to capture interferences and artifacts. We get 100 bases for $P_{\text{ideal}}^{(l)}(f|z)$ from the first PLCA, and learn 50 individual components in the second round. Note that we also use the compensation process from Section 5 for this model.
- PLCS with initialization: This model learns the common bases and encodings using the PLCS model and the post processing. It initializes its bases with $P_{\text{source}}(f|z)$ learned from the studio recording, but learn the bases $P_C(f|z)$ using the usual update rules (3) and (4) rather than (5) and (6). We randomly initialize 50 individual bases for $P_I^{(l)}(f|z)$.
- PLCS with the source prior: This PLCS model uses the source priors $P_{\text{source}}(f|z)$ both to initialize and to learn $P_C(f|z)$ using (6). 50 individual bases are randomly initialized for $P_I^{(l)}(f|z)$ and learned using (3).

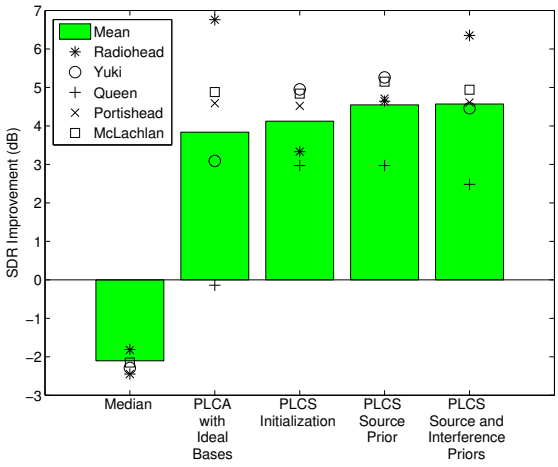

Fig. 4: The mean and song-specific improvements of SDR by each model for the consolidated reconstruction.

- PLCS with the source and interference priors: This full PLCS model uses both the source and interference priors, $P_{\text{source}}(f|z)$ and $P_{\text{interf}}(f|z)$ to initialize them and learn them using both (5) and (6). Note that we do not assume the interference prior for $X^{(3)}$, because it is riddled with clipping, not an additional interference.

Fig. 4 shows the SDR improvements caused by the proposed systems. The most noticeable observation is that the medians of the three mixture spectrograms do not provide good results as there is no guarantee that the median of the contaminated pixels is from the common source. For the given five songs, we can also see that the proposed PLCS models outperform the maximal performance bound of the PLCA model with ideal bases, mainly because of the strong sharing of both spectral and temporal aspects of the common components. Moreover, PLCS models exhibit less variance than PLCA. On top of that, PLCS with priors can provide better performance than the usual initialization method by gently reducing the impact of prior information as the iteration i increases ($\alpha = \beta \propto e^{-i}$). Although it is not observable in this objective quality measurements, adding the interference priors improves the perceptual sound quality of the result.

7. CONCLUSION

We propose the PLCS model for collaborative audio enhancement, where common audio sources are constructed out of multiple noisy recordings of the same audio scene. The model is characterized by its ability to share both spectral and temporal marginal factors while providing a level of flexibility by not sharing the weight probabilities $P^{(l)}(z)$. PLCS also models individual artifacts from the common sources by not sharing some components. The advantage of this sharing concept was shown by experiments on commercial music signals by outperforming the ordinary PLCA model even with ideal bases which are not available in realistic cases. The proposed model is also equipped with a consolidation process that can harmonize the recording-specific reconstructions. Finally, prior information can be easily use in this model, which further improves the performance in our simulations.

8. REFERENCES

[1] D. C. Brabham, "Crowdsourcing as a model for problem solving: An introduction and cases," *Convergence: The International Journal of Research into New Media Technologies*, vol. 14, no. 1, pp. 75–90, 2008.

[2] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, Aug. 2010.

[3] N. J. Bryan, P. Smaragdis, and G. J. Mysore, "Clustering and synchronizing multicamera video via landmark cross-correlation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012.

[4] P. Shrestha, M. Barbieri, H. Weda, and D. Sekulovski, "Synchronization of multiple camera videos using audio-visual features," *IEEE Transactions on Multimedia*, vol. 12, no. 1, pp. 79–92, 2010.

[5] J. Yoo, M. Kim, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for drum source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010.

[6] M. Kim, J. Yoo, K. Kang, and S. Choi, "Blind rhythmic source separation: Nonnegativity and repeatability," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010.

[7] ——, "Nonnegative matrix partial co-factorization for spectral and temporal drum source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1192–1204, 2011.

[8] P. Leveau, S. Maller, J. Burred, and X. Jaureguiberry, "Convolutive common audio signal extraction," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2011, pp. 165–168.

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[10] ——, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 13. MIT Press, 2001.

[11] T. Hofmann, "Probablistic latent semantic indexing," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1999.

[12] ——, "Probablistic latent semantic analysis," in *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*, 1999.

[13] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *Neural Information Processing Systems Workshop on Advances in Models for Acoustic Processing*, 2006.

[14] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium, Philadelphia*, 1993.