Mixtures of Local Dictionaries for Unsupervised Speech Enhancement

Minje Kim, *Student Member, IEEE,* and Paris Smaragdis, *Senior Member, IEEE*

***Abstract*—We propose a novel extension of Nonnegative Matrix Factorization (NMF) that models a signal with multiple local dictionaries activated sparsely. This set of local dictionaries for a source, e.g. speech, disjointly constitute a superset that is more discriminative than an ordinary NMF dictionary, because its local structures represent the source's manifold better. A block sparsity constraint is used to regularize the NMF solutions so that only one or a small number of blocks are active at a given time. Moreover, a concentration prior further regularizes each block of bases to be close to each other for better locality preservation. We test the proposed Mixture of Local Dictionaries (MLD) on single-channel speech enhancement tasks and show that it outperforms the state of the art technology by up to 2dB in signal-to-distortion ratio, especially in the unsupervised environment where neither the speaker identity nor the type of noise is known in advance.**

***Index Terms*—Nonnegative Matrix Factorization, Speech Enhancement, Manifold Learning**

I. Introduction

Nonnegative Matrix Factorization (NMF) [1], [2] has been widely used in audio research, e.g. automatic music transcription [3], musical source separation [4], speech enhancement [5], etc. Among this work a key strategy for applying NMF towards single channel speech enhancement is to model a source's training set (usually magnitude spectra) with a dictionary that consists of a small number of basis vectors. These basis vectors should be able to approximate all of the source's produced spectra. Since both bases and weights are nonnegative, the dictionary learned from NMF eventually models the input magnitude spectra with a convex cone (more precisely, a simplicial cone [6]). Then, the main goal of the single channel speech enhancement becomes that of representing the noisy input spectrum as a weighted sum of speech bases and estimated noise bases. The source estimates are forced to lie inside their respective convex cones. Therefore, the source-specific convex cone concept is critical for the denoising performance since the less the convex cones overlap each other, the more discriminative they are.

However, this linear decomposition model can be limited when it comes to modeling a complex source manifold. Although NMF has been found to be a suitable model for analyzing audio spectra because of its flexible additive nature, this flexibility sometimes hinders the ability to discriminate between different sources. If all the data points on a complex manifold structure should belong to a convex set defined by the basis vectors, it is inevitable that this convex cone will include some unnecessary regions where source spectra cannot reside.

We propose a Mixture of Local Dictionaries (MLD) model along the lines of recent attempts to preserve the manifold of the audio spectra during the NMF-like analysis. The basic idea is to learn dictionaries using the sparse coding concept to better approximate the input [7]. Particularly, in [8] a non-parametric overcomplete dictionary model was proposed that fully makes use of the entire training spectra instead of discarding them after learning their convex model. This method encourages the source estimates to lie on the manifold by approximating them with the sum of only a very small number of training samples. A succeeding work did this job in a more direct way by encompassing only the nearest neighbors for the source estimation [9].

Another relevant work is the Universal Speech Model (USM) [10]. USM tackles the case where the identity of the speaker is unknown and clean speech signals from anonymous speakers are available for training instead. Since the anonymous training spectra can have too much variance, a naïve NMF approach that learns a single convex cone from the entire set of training signals can produce less discriminative results than the hypothetical ones learned from the ideal speaker. In order to address that, USM first uses regular NMF on each speaker to learn a speaker-specific convex cone, and then during the separation stage it only activates a very small number of speakers' dictionaries at a time, the ones that best fit the observed data. To this end, USM involves a block sparsity constraint that was also used in [11], [12], so that irrelevant speakers' basis sets are turned off in the group-wise manner.

The proposed MLD model intends to preserve the manifold of the source data in a more controlled way. The benefit of using MLD comes from the following points:

- During training MLD discovers several convex cones per a source, each of which covers a chunk of similar spectra across all speakers rather than one per a speaker.
- For each convex cone, MLD penalizes the difference between the local dictionary and its *a priori*, such as in the Maximum A Posteriori (MAP) estimation. Therefore, the learned bases are eventually more concentrated on the prior. As a result, each convex cone covers a smaller area than USM or NMF cases.
- During denoising MLD activates only a small number of dictionaries for a given noisy input spectrum. Because MLD makes this decision in the frame-by-frame way, the model dynamically finds an optimal fit while USM approach does this in a global sense over time.

M. Kim is with the Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 61801 USA (e-mail: minje@illinois.edu).

P. Smaragdis is with the Department of Computer Science and the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign. He is also with Adobe Systems, Inc. (e-mail: paris@illinois.edu)

Manuscript received April XX, 2014; revised XXX.

II. Dictionary-based Denosing Using NMF

This section reviews a common speech enhancement procedure that uses NMF basis vectors as a dictionary.

A. Dictionary Learning

For each source c, either $c = S$ for speech or $c = N$ for noise, we first perform the Short Time Fourier Transform (STFT) and take the magnitude to build a source specific nonnegative training matrix $V^c_{dic} \in \mathbb{R}_+^{M\times N_c}$. NMF then finds a pair of factor matrices $W^c_{dic} \in \mathbb{R}_+^{M\times R_c}$ and $H^c_{dic} \in \mathbb{R}_+^{R_c\times N_c}$ that define a convex cone to approximate the input: $V^c_{dic} \approx W^c_{dic}H^c_{dic}$ [1], [2]. Among all the possible choices of β-divergences to measure the approximation error as proposed in [13], we focus on the case $\beta = 1$, or a generalized KL-divergence as follows:

$$\mathcal{D}(x|y) = x(\log x - \log y) + (y - x). \tag{1}$$

The parameters W^c_{dic} and H^c_{dic} that minimize the error $\mathcal{D}(V^c_{dic}|W^c_{dic}H^c_{dic})$ are estimated by changing the step size of the gradient descent optimization so that they are updated in a multiplicative way:

$$\begin{aligned} W^c_{dic} &\leftarrow W^c_{dic} \odot \left\{ \left(\frac{V^c_{dic}}{W^c_{dic}H^c_{dic}} \right) {H^c_{dic}}^\top \right\} \Big/ \left\{ \mathbf{1}{H^c_{dic}}^\top \right\}, \\ H^c_{dic} &\leftarrow H^c_{dic} \odot \left\{ {W^c_{dic}}^\top \left(\frac{V^c_{dic}}{W^c_{dic}H^c_{dic}} \right) \right\} \Big/ \left\{ {W^c_{dic}}^\top \mathbf{1} \right\}, \end{aligned} \tag{2}$$

where the Hadamard product $\odot$ and division are carried out in the element-wise fashion. Once the parameters are initialized with nonnegative random numbers, their sign stays the same after the updates. The learned basis vectors W^c_{dic} per each source represent the source as a dictionary.

B. Speech Enhancement

The speech enhancement procedure on the unseen noisy signals is done by learning the activations of the corresponding dictionaries learned from the procedure in section II-A. For a noisy spectrogram $V_{test} \in \mathbb{R}_+^{M\times N_{test}}$, the activation per each source c is estimated as follows:

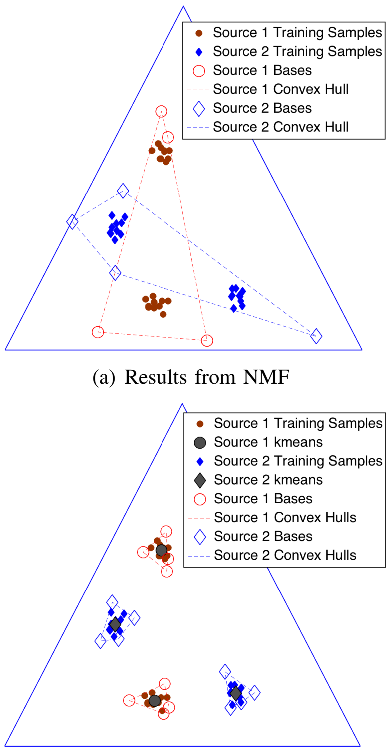
$$H^c_{test} \leftarrow H^c_{test} \odot \left\{ {W^c_{dic}}^\top \left(\frac{V_{test}}{W_{dic}H_{test}} \right) \right\} \Big/ \left\{ {W^c_{dic}}^\top \mathbf{1} \right\}, \tag{3}$$

where $W_{dic} = [W^S_{dic}, W^N_{dic}]$ and $H_{test} = \begin{bmatrix} H^S_{test} \\ H^N_{test} \end{bmatrix}$. Note that we call this case the *supervised* separation since both the speech and noise dictionaries are known. We do not usually update the learned dictionaries during the supervised separation. Finally, the speech part of the test spectrogram is recovered by masking the mixture matrix by the proportion of the speech estimate in the total reconstruction:

$$V^S_{test} \approx V_{test} \odot (W^S_{dic}H^S_{test}) \Big/ (W_{dic}H_{test}). \tag{4}$$

In the semi-supervised separation either the speech or noise training set is not available [14]. If the noise dictionary W^N_{dic} is unknown, it has to be learned from the mixture signal, calling for an update of the dictionary in addition to (3):

$$W^N_{dic} \leftarrow W^N_{dic} \odot \left\{ \left(\frac{V_{test}}{W_{dic}H_{test}} \right) {H^N_{test}}^\top \right\} \Big/ \left\{ \mathbf{1}{H^N_{test}}^\top \right\}. \tag{5}$$



(a) Results from NMF

(b) The proposed MLD model

Fig. 1: A comparison of convex hulls learned from a toy dataset.

III. The Mixture of Local Dictionaries Model

In this section we first address some downsides of the conventional NMF model with respect to source manifold preservation, and introduce the proposed MLD approach.

A. Locality in Data Manifolds

Fig. 1 (a) depicts the behavior of NMF dictionaries. For illustrational convenience we project the input vectors onto the simplex as if they are normalized. In this toy example, there are three spectral features that correspond to the three corners of the simplex. In Fig. 1 there are two different kinds of sources represented with small red dots and blue diamonds respectively. If we learn a set of four basis vectors that describe each source, they define the corners of a convex hull[1] (empty diamonds and circles), which surrounds the data points.

Each source has a manifold structure that consists of two distinct clusters, but NMF does not take it into account and results in a convex dictionary that wraps both intrinsic clusters. Hence, each convex hull can reconstruct not only the training data, but also spurious cases in the areas where the data is very unlikely to exist. On top of that, since NMF does not guarantee that the convex hull will surround the data tightly, the dictionary can include even more unnecessary regions. Meanwhile, when we unmix them, we use only the learned bases after discarding the training samples. Since

[1] Since the data points are normalized for the simplex representation, the convex cone learned by an NMF run reduces to a hull.

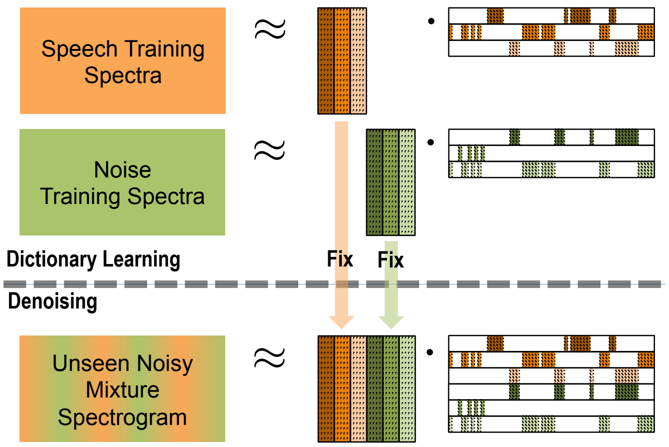


Fig. 2: A block diagram for the full speech enhancement procedure including source specific dictionary learning and its (block) sparse coding with the learned dictionaries.

the bases from the sources do not precisely represent the individual source's structure, we end up with overlapping convex hulls. This is problematic, because when an estimated source spectrum falls in the overlapped region of the red and blue convex hulls, it is impossible to identify which source it belongs to.

On the other hand, MLD learns a set of small convex dictionaries per each source, similarly to the way that a mixture model approximates an arbitrary distribution. The underlying latent modality of a mixture model corresponds to each convex cone in MLD. In Fig. 1 (b) we can see that MLD successfully tracks the data manifold with two disjoint convex hulls. Moreover, the hull wraps the data points tightly enough to include as small empty space as possible. Hence, we can expect a better preserved source manifolds in the learned dictionaries.

B. Mixture of Local Dictionaries: Algorithms

Fig. 2 describes the entire speech enhancement procedure using MLD. We first have to learn basis vectors from each source (orange and green) as in the dictionary learning procedure introduced in II-A. In MLD however, the basis vectors for a source are grouped into a pre-defined number of blocks, e.g. in the figure five bases per a block and three blocks per a source, each of which corresponds to a convex cone, or a local dictionary. At the same time the activation matrix H is learned to be block-wise sparse, so that a training sample belongs to only one or a very small number of local dictionaries. We use these learned dictionaries as they are when we perform the denoising job on some unseen noisy signals while we newly learn a block-wise sparse encoding matrix, H.

The objective function $\mathcal{J}$ is defined as follows:

$$\mathcal{J} = \mathcal{D}(V|WH) + \lambda \sum_t \Omega(h_t) + \eta \sum_g \mathcal{D}(\mu^{(g)}|W^{(g)}), \tag{6}$$

where $W = [W^{(1)}, \cdots, W^{(G)}]$, $H = [h_1, \cdots, h_N]$ and $h_t = [{h_t^{(1)}}^\top, \cdots, {h_t^{(G)}}^\top]^\top$. t and g are for the frame and group indices. The first term stands for KL-divergence in (1) from the original NMF algorithm. The function Ω on g-th block of each frame t is to give penalty to the solutions that are not sparse. In particular, we use $\log/l_1$ penalty, $\Omega(h_t) = \sum_g \log(\epsilon + ||h_t^{(g)}||_1)$, which was also used in [12], [10], for its monotonicity and induced multiplicative updates. The third term governs basis vectors in each block so that they are similar to each other and the resulting convex cones are compact. λ and η control the amount of the regularization.

Algorithm 1 The dictionary learning algorithm using MLD

```
Input: $V \in \mathbb{R}_+^{M \times N}, G, R$
Output: $W$
Find $G$ cluster means, $\mu^{(g)}$, by using K-means
Initialize $W^{(g)} \in \mathbb{R}_+^{M \times R}$ with $\mu^{(g)}$ and $H \in \mathbb{R}_+^{GR \times N}$ with random numbers
repeat
    Update $W$ and $H$ using (7), (8), and (9)
until Convergence
```

The main difference between USM and the proposed MLD model comes from the fact that the former sets block sparsity on speakers. It selects some relevant speakers in a global fashion, so the chosen ones are always active regardless of the time index t whereas the proposed method selects the participating blocks dynamically. Therefore, USM does not have the index t in the second term. Since it is not guaranteed that each speaker is assigned to a cluster, the data-driven way we choose is more suitable for modeling the general human speech with the limited number of clusters. After all, we expect that MLD can sort out the similar sound components into the same block regardless of who speaks them, and then eventually approximate the data more precisely.

Another thing that makes MLD unique is the newly introduced third term. For each block g we can have *a priori* knowledge about the bases, which can be learned beforehand by using any clustering techniques, e.g. K-means clustering. When the algorithm tries to reduce the error in the third term, the bases are more likely to be similar to the *a priori* information. This will also result in more concentrated solutions. Note that the regularization works as a conjugate prior in the corresponding probabilistic models, such as Dirichlet priors in Probabilistic Latent Semantic Indexing (PLSI) [15].

After majorizing the second term [10] we can derive some multiplicative update rules similarly to the NMF case for all $g \in \{1, \cdots, G\}$ and $t \in \{1, \cdots, N\}$:

$$W^{(g)} \leftarrow W^{(g)} \odot \frac{\left(\frac{V}{WH}\right){H^{(g)}}^\top + \eta \frac{\mu^{(g)}}{W^{(g)}}}{\mathbf{1}{H^{(g)}}^\top + \eta \mathbf{1}}, \tag{7}$$

$$H \leftarrow H \odot \left\{W^\top\left(\frac{V}{WH}\right)\right\} \Big/ \left\{W^\top \mathbf{1}\right\}, \tag{8}$$

$$h_t^{(g)} \leftarrow h_t^{(g)} \Big/ \left\{1 + \lambda/\left(\epsilon + ||h_t^{(g)}||_1\right)\right\}. \tag{9}$$

The MLD update rules are used to learn dictionaries from the source-specific training signals and to denoise an unseen noisy signal as in Section II-A and II-B. Algorithm 1 shows how we learn the dictionaries. First, we prepare a big set of speech training spectra V_{dic}^S recorded by anonymous speakers,

Algorithm 2 The speech enhancement algorithm using MLD

```
Input: $V \in \mathbb{R}_+^{M\times N}, \{W_{dic}^{S\,(g)} \in \mathbb{R}_+^{M\times R_S} | 1 \le g \le G_S\}$,
  and $\{W_{dic}^{N\,(g)} \in \mathbb{R}_+^{M\times R_N} | 1 \le g \le G_N\}$ (optional, semi-supervised) or $R_N$ (optional, unsupervised)
Output: $H^S, H^N$
Initialize $H^S$ and $H^N$ with random numbers
repeat
    Update $H$ using (8) and $h^{S\,(g)}_t$ using (9)
    if Unsupervised then
        Update $W^N_{dic}$ using (5)
    else
        Update $h^{N\,(g)}_t$ using (9)
    end if
until Convergence
```

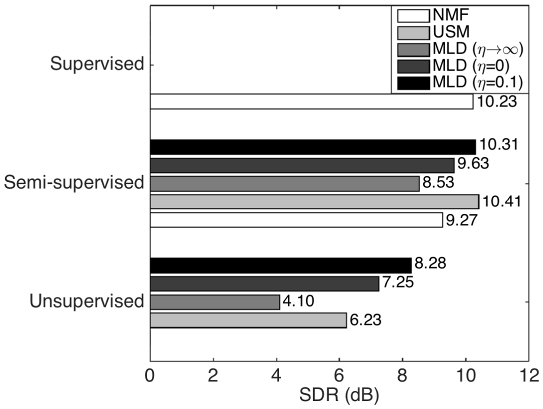


Fig. 3: Average SDR results from three models and cases.

and define the number of blocks G and the number of bases R in each block. If a noise training set is available, we prepare V^N_{dic} as well. Then, we use each matrix as the input to Algorithm 1, respectively, to get W^S_{dic} and W^N_{dic}.

There are three test scenarios depending on the availability of dictionaries:

- Unsupervised: the case when neither the speaker identity nor the type of noise is known. Therefore, we learn a suboptimal dictionary from someone else' clean speech and apply the semi-supervised technique.
- Semi-supervised: in this case either speech or noise dictionary is missing. The missing dictionary is learned during the test phase.
- Supervised: both kinds of information are known and their training data are available.

MLD is mainly for the unsupervised and semi-supervised cases. In the unsupervised scenario, since we have no speaker information, Algorithm 1 learns the speech dictionary W^S_{dic} using third-party speech signals as an alternative training set. Then, the noise dictionary is learned from the mixture signal (the "Unsupervised" option in Algorithm 2) using the ordinary NMF update in (5). In the semi-supervised case where only the type of noise is known, we solve this as if it was a supervised case with the noise dictionary and the suboptimal speech dictionary.

IV. Experimental Results

We compare the results from the MLD algorithms with USM and NMF in the three denoising scenarios. All signals used are sampled at 16kHz. As for STFT, we use 1024 samples for Hann windowing and FFT with 256 pt hop size. A small value for $\epsilon = 10^{-5}$ worked fine. We randomly select 20 speakers from the training set of TIMIT corpus as our anonymous speakers, and mix 5 test speakers' speech with 10 different non-stationary noise signals that were proposed in [14] to build 50 test sequences as in USM experiments. Instead of learning 10 NMF bases from each of the 20 speakers, we learn a set of $G_S = 20$ local dictionaries, each of which holds $R_S = 10$ bases. This number of speech bases is eventually same with the original USM setting while holding different information about the source. All USM and NMF results in this section are from the experiments in [10].

In the unsupervised case, the number of noise bases is fixed with the optimal ones investigated in [14], i.e. one of {20, 10, 200, 20, 20, 10, 20, 10, 10, 10}, depending on the noise type used. $\lambda = 64$ is big enough to yield sparse solutions. MLD outperforms USM by 2dB in terms of the average Signal-to-Distortion Ratio (SDR) [16] with an adequate $\eta = 0.1$ (the bottom bars in Fig. 3). Note that when $\eta \to \infty$ we get worse results, representing the case where each local dictionary is completely concentrated on its mean. Another case of interest is when $\eta = 0$, which boils down to a variation of USM where the speaker selection is done at every frame, but without the third regularization term in (6). We can say that another 1dB improvement is explained by the third term.

Next, we learn noise dictionaries as well with the same parameters ($G_N = G_S, R_N = R_S$), and then run Algorithm 2 without the unsupervised option. $\lambda = 2$ gives less sparse noise coding, but provides best separation results. Compared to the other semi-supervised algorithms (middle bars in the figure), the result is significantly better than NMF (1dB improvement) and comparable to USM. In this case, learning 200 basis vectors for each individual noise type is a difficult problem, and eventually does not outperform USM.

In Fig. 3 the result from the fully supervised NMF algorithm is provided for a further comparison. Once we are allowed to learn dictionaries for noise (semi-supervised), both USM and MLD models produce comparable results to the fully supervised NMF case, but without knowing the exact speaker.

V. Conclusion

In this paper we proposed the Mixture of Local Dictionaries (MLD) model that preserves the underlying manifold of the data. With extensions to the Universal Speech Model (USM), such as temporally relaxed block sparsity and concentration of basis vectors, the near-disjoint combination of local dictionaries provided substantial improvement over NMF and USM in the unsupervised single-channel speech enhancement tasks with as little *a priori* information about the sources as possible. Since this assumption about the source is general enough, i.e. an English speech, while being robust to many possible variations, such as dialects, speaker identities, and genders, we believe that MLD could be a practical solution to the single-channel unsupervised speech enhancement problem.

References

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

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

[3] P. Smaragdis and J. C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2003, pp. 177–180.

[4] T. O. Virtanen, “Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[5] N. Mohammadiha, J. Taghia, and A. Leijon, “Single channel speech enhancement using Bayesian NMF with recursive temporal updates of prior distributions,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.

[6] D. L. Donoho and V. Stodden, “When does nonnegative matrix facotrization give a correct decomposition into parts?” in *Advances in Neural Information Processing Systems (NIPS)*, vol. 16. MIT Press, 2004.

[7] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online learning for matrix factorization and sparse coding,” *Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[8] P. Smaragdis, M. Shashanka, and B. Raj, “A sparse non-parametric approach for single channel separation of known sounds,” in *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, BC, Canada, 2009.

[9] M. Kim and P. Smaragdis, “Manifold preserving hierarchical topic models for quantization and approximation,” in *Proceedings of the International Conference on Machine Learning (ICML)*, Atlanta, Georgia, 2013.

[10] D. L. Sun and G. J. Mysore, “Universal speech models for speaker independent single channel source separation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013.

[11] A. Hurmalainen, R. Saeidi, and T. Virtanen, “Group sparsity for speaker identity discrimination in factorisation-based speech recognition,” in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2012.

[12] A. Lefèvre, F. Bach, and C. Févotte, “Itakura-Saito non-negative matrix factorization with group sparsity,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[13] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis,” *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[14] Z. Duan, G. J. Mysore, and P. Smaragdis, “Online plca for real-time semi-supervised source separation,” in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2012, pp. 34–41.

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

[16] 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.