Abstract—We propose a novel extension of Nonnegative Matrix Factorization (NMF) that models a signal with multiple local dictionaries that are sparsely activated. 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 or 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 the nearest neighbors only 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 uses a single convex cone from the entire training signal set can be less discriminative than a convex cone learned from a single 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 just 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 DENOISING USING NMF

This section reviews the straightforward 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 \( V_d^c \in \mathbb{R}^{M \times N_c} \). NMF then finds a pair of factor matrices \( W_d^c \in \mathbb{R}^{M \times R_c} \) and \( H_d^c \in \mathbb{R}^{R_c \times N_c} \) that define a convex cone to approximate the input: \( V_d^c \approx W_d^c H_d^c \) [1], [2]. Among all the possible choices of \( \beta \)-divergences to measure the approximation error as proposed in [13], we focus on the case \( \beta = 1 \), or a generalized KL-divergence as follows:

\[
\mathcal{D}(x|y) = x \log x - \log y + (y - x).
\]  

The parameters \( W_d^c \) and \( H_d^c \) that minimize the error \( \mathcal{D}(V_d^c|W_d^c H_d^c) \) are estimated by changing the step size of the gradient descent optimization so that they are updated in a multiplicative way:

\[
W_d^c \leftarrow W_d^c \odot \left( \left( \frac{V_d^c}{W_d^c H_d^c} \right) H_d^c \right)^\tau,\quad H_d^c \leftarrow H_d^c \odot \left( \left( \frac{V_d^c}{W_d^c H_d^c} \right)^\tau \right)^{-1},
\]

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 \( W_d^c \) per each source work as a dictionary for the source.

B. Speech Enhancement

The speech enhancement procedure on the unseen noisy signals is done by learning the activations of the corresponding dictionaries as 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:

\[
H_{test}^c \leftarrow H_{test}^c \odot \left( \frac{V_{test}}{W_d^c H_{test}} \right)^\tau,\quad (W_d^c)_{\text{test}} \leftarrow \frac{W_d^c}{H_{test}^c}.
\]  

where \( W_{\text{dic}} = [W_S^c, W_N^c] \) and \( H_{\text{test}} = [H_{test}^S, H_{test}^N] \). 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:

\[
V_{test}^S \approx V_{test} \odot (W_d^S H_{test}^S)(W_d^N H_{test}^N).
\]

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

\[
W_d^N \leftarrow W_d^N \odot \left( \frac{V_{test}}{W_d^N H_{test}^N} \right)^\tau,\quad (H_{test}^N)_{\text{test}} \leftarrow \frac{H_{test}^N}{H_{test}^N}.
\]

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 construct a convex hull as its corners (empty diamonds and circles), which surrounds the data points. Note that each source has a manifold structure that consists of two distinct clusters.

NMF does not take this structure 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. During the unmixing phase, we use the learned bases only. Since the bases from two

1Since the data points are normalized for the simplex representation, the convex cone learned by an NMF run reduces to a hull.
sources do not precisely represent each source’s structure, we end up with overlapping convex hulls. This is a problem since 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 hulls wrap the data points tightly enough to exclude any surrounding empty space. In this case we can expect less ambiguity than with the ordinary NMF case, and this representation will be able to model more complex source manifolds than the simple case we present here.

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 $J$ is defined as follows:

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

where $W = [W^{(1)}, \ldots, W^{(G)}]$, $H = [h_1, \ldots, h_N]$ and $h_t = [h_t^{(1)} \, \ldots, h_t^{(G)}]^\top$. $t$ and $g$ are for the frame and group indices. The first term is for KL-divergence in (1) from the original NMF algorithm. The function $\Omega$ on $g$-th block of each frame $t$ is to give penalty to the solutions that are not sparse. In particular, we use $\log |l_t|$ 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 tight. $\lambda$ and $\eta$ control the amount of the regularization.

The main difference between USM [10] and the proposed MLD model comes from the fact that the former sets block sparsity on speakers. It selects relevant speakers in a global fashion, so the chosen one is 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 associated with a distinct cluster, the data-driven way we choose is more suitable for modeling the general human speech with the same 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 an 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 multiplicative update rules similarly to the NMF case for all $g \in \{1, \ldots, G\}$ and $t \in \{1, \ldots, N\}$:

$$W^{(g)} \leftarrow W^{(g)} \odot \left( \frac{V}{WH} \right) H^{(g)^\top} + \eta \mu^{(g)} / \Omega(h_t^{(g)}), \quad (7)$$

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

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

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 1** The dictionary learning algorithm using MLD

1: Input: $V \in \mathbb{R}_+^{M \times N}, G, R$
2: Output: $W$
3: Find $G$ cluster means, $\mu^{(g)}$, by using K-means
4: Initialize $W^{(g)} \in \mathbb{R}_+^{M \times K}$ with $\mu^{(g)}$ and $H \in \mathbb{R}_+^{GR \times N}$ with random numbers
5: repeat
6: Update $W$ and $H$ using (7), (8), and (9)
7: until Convergence
Algorithm 2 The speech enhancement algorithm using MLD
1: Input: $V \in \mathbb{R}_{+}^{M \times N}, \{W_{dic}^{S}(g) \in \mathbb{R}_{+}^{M \times R_S} | 1 \leq g \leq G_{S}\}$, and $\{W_{dic}^{N}(g) \in \mathbb{R}_{+}^{M \times R_N} | 1 \leq g \leq G_{N}\}$ (optional, semi-supervised) or $R_{N}$ (optional, unsupervised)
2: Output: $H^{S}, H^{N}$
3: Initialize $H^{S}$ and $H^{N}$ with random numbers
4: repeat
5: \hspace{6mm} Update $H$ using (8) and $h^{S(g)}_{t}$ using (9)
6: \hspace{6mm} if Unsupervised then
7: \hspace{12mm} Update $W_{dic}^{N}$ using (5)
8: \hspace{12mm} else
9: \hspace{18mm} Update $h^{N(g)}_{t}$ using (9)
10: \hspace{6mm} end if
11: 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 $V_{dic}^{N}$ as well. Then, we use each matrix as the input to Algorithm 1, respectively, to get $W_{dic}^{S}$ and $W_{dic}^{N}$.

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’s 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 $W_{dic}^{S}$ 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, 10, 10, 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 $\eta \rightarrow \infty$ 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 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, and 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