Discriminative NMF and its application to single-channel source separation

Felix Weninger, Jonathan Le Roux, John R. Hershey, and Shinji Watanabe

Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA

felix@weninger.de | {leroux,hershey,watanabe}@merl.com

Abstract

The objective of single-channel source separation is to accurately recover source signals from mixtures. Non-negative matrix factorization (NMF) is a popular approach for this task, yet previous NMF approaches have not optimized directly this objective, despite some efforts in this direction. Our paper introduces discriminative training of the NMF basis functions such that, given the coefficients obtained on a mixture, a desired source is optimally recovered. We approach this optimization by generalizing the model to have separate analysis and reconstruction basis functions. This generalization frees us to optimize reconstruction objectives that incorporate the filtering step and SNR performance criteria. A novel multiplicative update algorithm is presented for the optimization of the reconstruction basis functions according to the proposed discriminative objective functions. Results on the 2nd CHiME Speech Separation and Recognition Challenge task indicate significant gains in source-to-distortion ratio with respect to sparse NMF, exemplar-based NMF, as well as a previously proposed discriminative NMF criterion.

1. Background

Non-negative matrix factorization (NMF) is a popular algorithm commonly used for challenging single-channel audio source separation tasks, such as speech enhancement in the presence of non-stationary noises [1, 2]. In this context, the basic idea is to represent the features of the sources via sets of basis functions and their activation coefficients, one set per source. Mixtures of signals are then analyzed using the concatenated sets of basis functions, and each source is reconstructed using its corresponding activations and basis set.

NMF operates on a matrix of F-dimensional non-negative spectral features, usually the power or magnitude spectrogram of the mixture, $\mathbf{M} = [\mathbf{m}_1 \cdots \mathbf{m}_T]$, where T is the number of frames and $\mathbf{m}_t \in \mathbb{R}_+^F$, $t = 1, \ldots, T$ are obtained by short-time Fourier analysis of the time-domain signal. For the general case of separating S sources, a set of R_l non-negative basis vectors $\mathbf{w}_1^l, \cdots, \mathbf{w}_{R_l}^l$ is assumed for each source $l \in \{1, \ldots, S\}$, and concatenated into matrices $\mathbf{W}^l = [\mathbf{w}_1^l \cdots \mathbf{w}_{R_l}^l]$. From this, a factorization

$$\mathbf{M} \approx \mathbf{WH} = [\mathbf{W}^1 \cdots \mathbf{W}^S][\mathbf{H}^1; \cdots; \mathbf{H}^S] \quad (1)$$

is obtained[1]. An approach related to Wiener filtering is typically used to reconstruct each source while ensuring that the source estimates sum to the mixture:

$$\hat{\mathbf{S}}^l = \frac{\mathbf{W}^l\mathbf{H}^l}{\sum_l \mathbf{W}^l\mathbf{H}^l} \otimes \mathbf{M}, \quad (2)$$

where $\otimes$ denotes element-wise multiplication and the quotient line element-wise division. In our study, all $\mathbf{W}^l$ are learnt in advance from training data, and at run time only the activation matrices $\mathbf{H}^l = [\mathbf{h}_1^l \cdots \mathbf{h}_T^l]$, where $\mathbf{h}_t^l \in \mathbb{R}_+^{R_l}$, are estimated. This is called *supervised NMF* [3]. In the supervised case, the activations for each frame are independent from the other frames ($\mathbf{m}_t \approx \sum_l \mathbf{W}^l\mathbf{h}_t^l$). Thus, source separation can be performed on-line and with latency corresponding to the window length plus the computation time to obtain the activations for one frame [1].

At test time, supervised NMF finds the optimal activations $\hat{\mathbf{H}}$ such that

$$\hat{\mathbf{H}} = [\hat{\mathbf{H}}^1; \cdots; \hat{\mathbf{H}}^S] = \underset{\mathbf{H}}{\operatorname{argmin}}\, D(\mathbf{M} \mid \mathbf{WH}) + \mu|\mathbf{H}|_1, \quad (3)$$

where D is a cost function that is minimized when $\mathbf{M} = \mathbf{WH}$. Here we use the β-divergence, D_β, which for $\beta = 1$ yields the generalized Kullback-Liebler (KL) divergence, and for $\beta = 2$, yields the Euclidean distance, both of which are considered in this paper. An L_1 sparsity constraint with weight μ is added to favor solutions where few basis vectors are active at a time. A convenient algorithm [4] for minimizing (3) that preserves non-negativity of $\mathbf{H}$ by multiplicative updates is given by iterating

$$\mathbf{H}^{(q+1)} = \mathbf{H}^{(q)} \otimes \frac{\mathbf{W}^\intercal(\mathbf{M} \otimes (\mathbf{\Lambda}^{(q)})^{\beta-2})}{\mathbf{W}^\intercal(\mathbf{\Lambda}^{(q)})^{\beta-1} + \mu}, \quad 0 \le q < Q$$

until convergence, with $\mathbf{\Lambda}^{(q)} := \mathbf{WH}^{(q)}$, the superscripts (q) and $(q+1)$ indicating iterates, and $Q \ge 1$ giving the maximum number of iterations. $\mathbf{H}^0$ is initialized randomly.

Since sources often have similar characteristics in the short-term observations (such as unvoiced phonemes and broadband noise, or voiced phonemes and music), it seems beneficial to use information from multiple time frames. In our study, this is done by stacking features: the observation $\mathbf{m}'_t$ at time t corresponds to the observations $[\mathbf{m}_{t-T_\mathrm{L}}; \cdots; \mathbf{m}_t; \cdots; \mathbf{m}_{t+T_\mathrm{R}}]$ where T_L and T_R are the left and right context sizes. Analogously, each basis element $\mathbf{w'}_k^l$ will model a sequence of spectra, stacked into a column vector. For readability, we subsequently drop the $'$.

The main contribution of this paper is a new objective function and optimization methods for training the model. Our method trains an NMF model *discriminatively* to minimize the error in estimating $\hat{\mathbf{S}}^l$ by (2) based on the activations obtained by (3). To our knowledge, this is the first method that optimizes the model to minimize reconstruction error based on the inference algorithm that will be used at test time.

2. Obtaining NMF Bases

A common approach [5, 6] to obtaining bases $\mathbf{W}^l$ is to fit an NMF model $\mathbf{W}^l\mathbf{H}^l$ to the spectrograms of source signals, $\mathbf{S}^l$, by separately minimizing the objective $D_\beta(\mathbf{S}^l \mid \mathbf{W}^l\mathbf{H}^l)$ for each source. Sparse regularization is useful for many problems, so here we consider sparse NMF (SNMF), which has the objective,

$$\overline{\mathbf{W}}^l, \overline{\mathbf{H}}^l = \underset{\mathbf{W}^l, \mathbf{H}^l}{\operatorname{argmin}}\, D_\beta(\mathbf{S}^l \mid \widetilde{\mathbf{W}}^l\mathbf{H}^l) + \mu|\mathbf{H}^l|_1, \quad (4)$$

[1] For simplicity, we use the notation $[\mathbf{a}; \mathbf{b}]$ for $[\mathbf{a}^\intercal\mathbf{b}^\intercal]^\intercal$.

for each source, l, where $\widetilde{\mathbf{W}}^l = \left[\frac{\mathbf{w}_1^l}{\|\mathbf{w}_1^l\|} \cdots \frac{\mathbf{w}_{R_l}^l}{\|\mathbf{w}_{R_l}^l\|}\right]$ is the column-wise normalized version of $\mathbf{W}^l$. Since the L1 sparsity constraint on $\mathbf{H}$ is not scale-invariant, it can be trivially minimized by scaling of the factors; by including the normalization in the cost function, the scale indeterminacy can be avoided. Note that for the reasons pointed out by [7], this is not the same as performing standard NMF optimization and scaling one of the factors to unit norm after each iteration, which is often the way sparsity is implemented in NMF and which we shall denote by NMF+S. A multiplicative update algorithm to optimize (4) for arbitrary $\beta \geq 0$ is given by [6].

To avoid the peculiarities of sparse NMF training, in practice, exemplar-based approaches, where every basis function corresponds to an observation of the source l in the training data, have become popular for large-scale factorizations of audio signals [2, 8]. We will consider both sparse basis learning and exemplar bases as baselines for comparison with our proposed approach, which we present in the next section.

2.1. Discriminative approach to NMF

The model underlying the above source separation process can be called a *factorial* one, where separately trained source models are concatenated to yield a model of the mixture.This comes with the benefit of modularity: models of different sources can be substituted for one another without having to train the whole system. However, this type of model also has a fundamental flaw: the objectives (3) and (4) used at test and training time are considerably different. The test-time inference objective (3) operates on a mixture while the training objective (4) operates on separated sources.

It is easy to see that if there is spectral overlap in the bases of the different sources – which cannot be avoided in the general case, such as for speech/noise separation – the activations obtained using (3) will be different than those obtained using (4). It is clear that (4) cannot be used at test time, since $\mathbf{S}^l$ is unknown. Hence, our discriminative approach is based on taking into account the objective function from (3) at training time.

This involves having mixtures $\mathbf{M}$ along with their ground truth separated sources $\mathbf{S}^l$ available for training (parallel training). However, supervised NMF also assumes the availability of separated training signals for all sources, and assumes simple linear mixing of the sources at test time. Generating the mixtures from the training signals for parallel training requires no additional assumptions.

We propose the following optimization problem for training bases, termed *discriminative NMF* (DNMF):

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmin}} \sum_l \gamma_l D_\beta \left(\mathbf{S}^l \mid \mathbf{W}^l \hat{\mathbf{H}}^l(\mathbf{M}, \mathbf{W})\right), \quad (5)$$

$$\text{where } \hat{\mathbf{H}}(\mathbf{M}, \mathbf{W}) = \underset{\mathbf{H}}{\operatorname{argmin}}\, D_\beta(\mathbf{M} \mid \widetilde{\mathbf{W}}\mathbf{H}) + \mu|\mathbf{H}|_1, \quad (6)$$

and γ_l are weights accounting for the application-dependent importance of the source l; for example, in speech de-noising, we focus on reconstructing the speech signal. The first part (5) minimizes the reconstruction error given $\hat{\mathbf{H}}$. The second part ensures that $\hat{\mathbf{H}}$ are the activations that arise from the test-time inference objective. Note that, in (5), $\mathbf{W}$ does not need normalization. Given the bases $\mathbf{W}$, the activations $\hat{\mathbf{H}}(\mathbf{M}, \mathbf{W})$ are uniquely determined, due to the convexity of (6). Nonetheless, the above remains a difficult bi-level optimization problem [9], since the bases $\mathbf{W}$ occur in both levels[2].

[2] After this paper was submitted, an objective function similar to (5) and (6) was independently proposed in [10], which tackles the bi-level optimization but without generalizing the model as in (7).

Note that the bases $\mathbf{W}$ used for *analysis* in (6) are the same as the ones used for *reconstruction* in (5). However, there is no clear benefit to such a contraint, and we can generalize the problem by separating the reconstruction and analysis bases:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmin}} \sum_l \gamma_l D_\beta \left(\mathbf{S}^l \mid \mathbf{W}^l \hat{\mathbf{H}}^l\big(\mathbf{M}, \overline{\mathbf{W}}\big)\right), \quad (7)$$

where $\overline{\mathbf{W}}$ are the analysis bases. This *generalized DNMF* problem, in its full generality, is still bi-level, but it gives us the option of holding $\overline{\mathbf{W}}$ constant to alleviate the difficulty, as $\mathbf{W}$ can then be obtained using the classical NMF updates. Optimizing both $\mathbf{W}$ and $\overline{\mathbf{W}}$ jointly using (7) is interesting but challenging, and it is unclear how much benefit it would bring. We thus procede with $\overline{\mathbf{W}}$ trained separately on each source using (4), and introduce the shorthand notation $\hat{\mathbf{H}} = \hat{\mathbf{H}}\big(\mathbf{M}, \overline{\mathbf{W}}\big)$.

2.2. Optimizing bases for Wiener filtering and SNR

As an additional benefit, the proposed framework allows us to easily extend the optimization to the Wiener-filter reconstruction (2), yielding the optimization of a new training objective:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmin}} \sum_l \gamma_l D_\beta \left(\mathbf{S}^l \Big| \frac{\mathbf{W}^l \hat{\mathbf{H}}^l}{\mathbf{W}\hat{\mathbf{H}}} \otimes \mathbf{M}\right). \quad (8)$$

At test time, the source estimates $\hat{\mathbf{S}}$ for a mixture $\mathbf{M}$ are reconstructed using (2) with $\hat{\mathbf{W}}$ and the activations $\hat{\mathbf{H}}\big(\mathbf{M}, \overline{\mathbf{W}}\big)$.

The framework also allows us to optimize the bases to improve signal-to-noise ratio (SNR) in the case where features are magnitude spectra. Indeed, minimizing the Euclidean distance $D_2^l := D_2(\mathbf{S}^l|\hat{\mathbf{S}}^l)$ between the magnitude spectrum of the source l, $\mathbf{S}^l$, and that of its reconstruction, $\hat{\mathbf{S}}^l$, directly corresponds to maximizing the SNR, neglecting the difference between noisy and oracle phases (we thus optimize for an upper-bound of the actual SNR). Thus, training $\mathbf{W}$ using (8) with $\beta = 2$ amounts to optimizing the bases for maximum SNR. Note that this does not mean that the activations $\hat{\mathbf{H}}\big(\mathbf{M}, \overline{\mathbf{W}}\big)$ used in (8) necessarily have to be found with the same setting $\beta = 2$ in (6), as long as the same β is used at training time and test time. In fact, we found that best results were obtained by using $\beta = 1$ in (6) and $\beta = 2$ in (8). This might be due to the KL divergence being better suited to decomposing mixtures, as was shown in previous evaluations [11, 12].

2.3. Relation to Prior Work

Various prior methods incorporating some notion of discriminative training have previously been proposed. However none of them satisfies our criterion of training-time optimization of the reconstruction using the test-time inference method applied to mixed signals. A discriminatively trained classifier on $\mathbf{H}$ is used in [13] to improve pitch detection by NMF, but does not consider basis learning or reconstruction. A heuristic method to train bases is proposed in [14] to minimize spectral overlap; however, it is not clear what the effect of this is on test-time reconstruction, since test-time reconstruction is not directly taken into account. In a similar vein, [15] proposes to take into account an objective function based on a discriminatively trained classifier in optimizing the NMF bases for multiple pitch estimation. A new training objective is proposed in [16] that additively combines the sparse NMF training-time and test-time objective functions, i.e., the

right hand sides of (3) and (4). This gives rise in our setting to a cost function of the form[3]

$$\sum_l D_\beta(\mathbf{S}^l \mid \widetilde{\mathbf{W}}^l\mathbf{H}^l) + D_\beta(\mathbf{M} \mid \sum_l \widetilde{\mathbf{W}}^l\mathbf{H}^l) + 2\mu|\mathbf{H}|_1. \quad (9)$$

The above objective is described as a discriminative NMF; however, it suffers from the same fundamental flaw as plain NMF: the training-time optimization differs from the test-time inference and reconstruction procedure, so the objective is not truly discriminative. Moreover, our experimental results (shown in Table 1) indicate that it performs no better than sparse NMF.

3. Multiplicative Update Algorithms for Discriminative NMF with Wiener filtering

We now derive a multiplicative update algorithm to minimize the objective in (8) with respect to $\mathbf{W}$ for fixed $\mathbf{H}$, where our goal is to reconstruct a single source $l \in \{1, \ldots, S\}$. We set $\gamma_l = 1$, $\gamma_{l':l'\neq l} = 0$, and define $\mathbf{\Lambda} = \sum_l \mathbf{W}^l\mathbf{H}^l$, $\mathbf{\Lambda}^l = \mathbf{W}^l\mathbf{H}^l$ for $l \in \{1, \ldots, S\}$, $\mathbf{\Lambda}^{\bar{l}} = \mathbf{\Lambda} - \mathbf{\Lambda}^l$, and $\hat{\mathbf{S}}^l = \mathbf{\Lambda}^l/\mathbf{\Lambda} \otimes \mathbf{M}$.

KL objective with Wiener filtering (DNMF-W-KL): For the case where $\beta = 1$ (KL), the objective function in (8) becomes

$$D_1^l := D_1(\mathbf{S}^l \mid \hat{\mathbf{S}}^l) = \sum_{i,j} S^l_{i,j} \log \frac{S^l_{i,j}}{M_{i,j}\frac{\Lambda^l_{i,j}}{\Lambda_{i,j}}} + M_{i,j}\frac{\Lambda^l_{i,j}}{\Lambda_{i,j}} - S^l_{i,j}.$$

The partial derivative of D_1^l with respect to the i-th element of the k-th basis function of the desired source, $w^l_{i,k}$, is

$$\begin{aligned}\frac{\partial D_1^l}{\partial w^l_{i,k}} &= \sum_j S^l_{i,j}\left(\frac{h^l_{k,j}}{\Lambda_{i,j}} - \frac{h^l_{k,j}}{\Lambda^l_{i,j}}\right) + M_{i,j}\frac{h^l_{k,j}\Lambda_{i,j} - \Lambda^l_{i,j}h^l_{k,j}}{\Lambda^2_{i,j}} \\ &= \sum_j -\frac{S^l_{i,j}\Lambda^{\bar{l}}_{i,j}}{\Lambda_{i,j}\Lambda^l_{i,j}}h^l_{k,j} + \frac{M_{i,j}\Lambda^{\bar{l}}_{i,j}}{\Lambda^2_{i,j}}h^l_{k,j}, \quad (10)\end{aligned}$$

where we use in the second equality that, by definition, $\Lambda_{i,j} - \Lambda^l_{i,j} = \Lambda^{\bar{l}}_{i,j}$. Similarly, we obtain

$$\frac{\partial D_1^l}{\partial w^{l'}_{i,k}} = \sum_j \frac{S^l_{i,j}}{\Lambda_{i,j}}h^{l'}_{k,j} - \frac{M_{i,j}\Lambda^l_{i,j}}{\Lambda^2_{i,j}}h^{l'}_{k,j} \quad (11)$$

for any $l' \neq l$. Since all matrix elements are non-negative, we can derive multiplicative update rules by splitting (10) and (11) into positive and negative parts, as done in [4]:

$$\mathbf{W}^l \leftarrow \mathbf{W}^l \otimes \frac{\frac{\mathbf{S}^l\otimes\mathbf{\Lambda}^{\bar{l}}}{\mathbf{\Lambda}\otimes\mathbf{\Lambda}^{\bar{l}}}\mathbf{H}^{l\intercal}}{\frac{\mathbf{M}\otimes\mathbf{\Lambda}^{\bar{l}}}{\mathbf{\Lambda}^2}\mathbf{H}^{l\intercal}}$$

$$\mathbf{W}^{\bar{l}} \leftarrow \mathbf{W}^{\bar{l}} \otimes \frac{\frac{\mathbf{M}\otimes\mathbf{\Lambda}^{l}}{\mathbf{\Lambda}^2}\mathbf{H}^{\bar{l}\intercal}}{\frac{\mathbf{S}^l}{\mathbf{\Lambda}}\mathbf{H}^{\bar{l}\intercal}}$$

where $\mathbf{W}^{\bar{l}} := [\mathbf{W}^1 \cdots \mathbf{W}^{l-1}\mathbf{W}^{l+1} \cdots \mathbf{W}^S]$, i.e., the bases of all sources except l, and $\mathbf{H}^{\bar{l}}$ is defined accordingly. The general case of $\gamma_l \geq 0$ for all l is an easy extension due to the linearity of the gradient.

LS objective with Wiener filtering (DNMF-W-LS): For the case where $\beta = 2$ (LS: least-squares), the gradient of D_2^l leads to:

$$\mathbf{W}^l \leftarrow \mathbf{W}^l \otimes \frac{\frac{\mathbf{M}\otimes\mathbf{S}^l\otimes\mathbf{\Lambda}^{\bar{l}}}{\mathbf{\Lambda}^2}\mathbf{H}^{l\intercal}}{\frac{\mathbf{M}^2\otimes\mathbf{\Lambda}^l\otimes\mathbf{\Lambda}^{\bar{l}}}{\mathbf{\Lambda}^3}\mathbf{H}^{l\intercal}}$$

$$\mathbf{W}^{\bar{l}} \leftarrow \mathbf{W}^{\bar{l}} \otimes \frac{\frac{\mathbf{M}^2\otimes(\mathbf{\Lambda}^l)^2}{\mathbf{\Lambda}^3}\mathbf{H}^{\bar{l}\intercal}}{\frac{\mathbf{M}\otimes\mathbf{S}^l\otimes\mathbf{\Lambda}^l}{\mathbf{\Lambda}^2}\mathbf{H}^{\bar{l}\intercal}}$$

The extension to general γ_l is again straightforward.

4. Experiments and Results

Our methods are evaluated on the corpus of the 2nd CHiME Speech Separation and Recognition Challenge, which is publicly available[4]. The task is to separate speech from noisy and reverberated mixtures ($S = 2$, $l = 1$: speech, $l = 2$: noise). The noise was recorded in a home environment with mostly non-stationary noise sources such as children, household appliances, television, radio, etc. Training, development, and test sets of noisy mixtures along with noise-free reference signals are created from the Wall Street Journal (WSJ-0) corpus of read speech and a corpus of training noise recordings. The dry speech recordings are convolved with room impulse responses from the same environment where the noise corpus is recorded. The training set consists of 7 138 utterances at six SNRs from -6 to 9 dB, in steps of 3 dB. The development and test sets consist of 410 and 330 utterances at each of these SNRs, for a total of 2 460 / 1 980 utterances. Our evaluation measure for speech separation is source-to-distortion ratio (SDR) [17]. By construction of the WSJ-0 corpus, our evaluation is speaker-independent. Furthermore, the background noise in the development and test set is disjoint from the training noise, and a different room impulse response is used to convolve the dry utterances.

4.1. Feature extraction

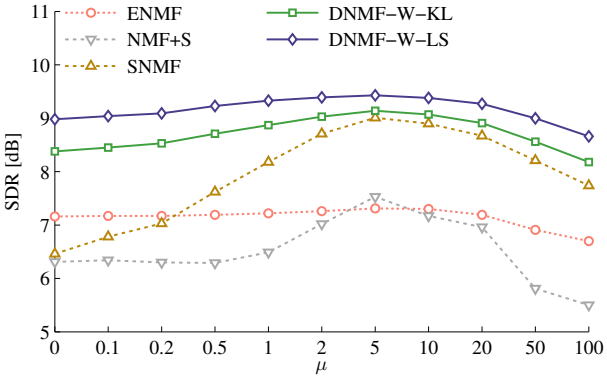
Each feature vector (in the mixture, source, and reconstructed source spectrograms as well as the basis vectors) covers nine consecutive frames ($T_\mathrm{L} = 8$, $T_\mathrm{R} = 0$) obtained as short-time Fourier spectral magnitudes, using 25 ms window size, 10 ms window shift, and the square root of the Hann window. Since no information from the future is used ($T_\mathrm{R} = 0$), the observation features ($\mathbf{m}_t$) can be extracted on-line. In analogy to the features in $\mathbf{M}$, each column of $\hat{\mathbf{S}}^l$ corresponds to a sliding window of consecutive reconstructed frames. To fulfill the on-line constraint, only the last frame in each sliding window is reconstructed.

4.2. Baselines: Exemplar-based and sparse NMF

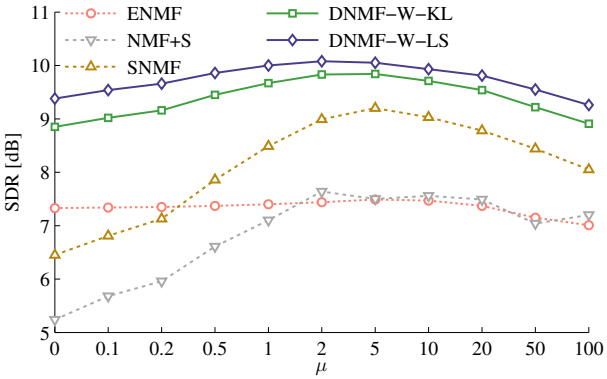
We use the same number R of basis vectors for speech and noise ($R^1 = R^2 = R$). We run an experiment for $R = 100$ and $R = 1000$. The maximum number of iterations at test time is set to $Q = 25$ based on the trade-off between SDR and complexity – running NMF until convergence increased SDR only by about .1 dB SDR. At training time, we use up to $Q = 100$ iterations. As a first baseline, we consider *exemplar-based NMF* (ENMF), where the analysis basis $\overline{\mathbf{W}}$ corresponds to $R^1 + R^2 = 2R$ randomly selected spectral patches of speech and noise, each spanning $T_\mathrm{L} + 1 = 9$ frames, from the isolated CHiME speech and background noise training sets. Next, we perform NMF basis training by SNMF according to (4), setting $\mathbf{S}^1$ and $\mathbf{S}^2$ to the spectrograms of the concatenated noise-free CHiME training set and the corresponding background noise in the multi-condition training set. This yields SNMF bases $\overline{\mathbf{W}}^l$, $l = 1, 2$. Due to space complexity, we use only 10 % of the training utterances. As initial solution for $\overline{\mathbf{W}}$, we use the exemplar bases. The sparsity weight μ is chosen from

[3] In [16], all combinations of isolated training signals are included as training data, but this is infeasible for the size of speech corpus we used.

[4] http://spandh.dcs.shef.ac.uk/chime_challenge/ – as of Feb. 2014



(a) Results for $R = 100$



(b) Results for $R = 1000$

Figure 1: Average SDR obtained for various sparsity weights μ on the CHiME Challenge (WSJ-0) development set.

$\{0, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100\}$. We found that using the exemplar bases as initialization provided fast convergence of the objective especially for large values of μ. Finally, we also consider NMF+S, an NMF where renormalization of the bases is done between iterations, as mentioned in Section 2. Surprisingly, to our knowledge, no comparisons between NMF+S and SNMF exist in the literature.

In the ENMF, NMF+S and SNMF experiments, the matrix $\overline{\mathbf{W}}$ is used both for determining $\hat{\mathbf{H}}$ according to (3) and for reconstruction using (2).

4.3. Discriminative NMF

In the discriminative NMF experiments, a discriminatively trained basis set $\hat{\mathbf{W}}$ is obtained using the DNMF-W-KL or DNMF-W-LS algorithm ($l = 1$, cf. Section 3), which is used for reconstruction using (2). $\mathbf{M}$ is set to the concatenated spectrogram of 10 % of the CHiME training set, so that we train DNMF and SNMF on the same data. The SNMF basis $\overline{\mathbf{W}} = [\overline{\mathbf{W}}^1 \overline{\mathbf{W}}^2]$ is used as initialization for $\hat{\mathbf{W}}$, as well as to get the activations $\hat{\mathbf{H}}$ using (3), again using $Q = 25$ iterations.

4.4. Evaluation on CHiME development and test sets

Figure 1 compares the average SDR obtained on the CHiME development set by using ENMF, NMF+S and SNMF, as well as DNMF-W-KL and DNMF-W-LS based on SNMF bases, for various sparsity parameters as well as basis sizes R. For a given R, all models have the same architecture and test time complexity. It can be seen that SNMF outperforms ENMF for higher sparsity values ($\mu \geq 0.5$). It is also interesting to see that SNMF consistently and significantly outperforms its ad hoc counterpart NMF+S. Furthermore, for $R = 100$, there is only a slight improvement by DNMF-W-KL over the best SNMF setting ($\mu = 5$, SDR=9.14 vs. 9.01 dB). However, DNMF-W-LS improves to 9.43 dB. For $R = 1000$, the improvement by DNMF is much more pronounced, while the SNMF performance increases only slightly (up to 10.1 dB for DNMF, 9.20 dB for SNMF). This could be attributed to the algorithm being able to exploit the increased set of trainable parameters more effectively than SNMF. It is also notable that the results by DNMF seem to be much less influenced by the choice of μ. This suggests that discriminative training can recover from the errors caused by using sub-optimal bases for analysis (such as those trained with sub-optimal μ).

SDR [dB]	Input SNR [dB]						
	-6	-3	0	3	6	9	Avg.
Noisy	-2.27	-0.58	1.66	3.40	5.20	6.60	2.34
ENMF	3.01	5.58	7.60	9.58	11.79	13.67	8.54
NMF+S	3.60	5.98	7.58	9.19	10.95	12.16	8.24
SNMF	5.48	7.53	9.19	10.88	12.89	14.61	10.10
d.NMF [16]	5.47	7.52	9.19	10.89	12.91	14.63	10.10
DNMF-W-KL	6.46	8.29	9.80	11.32	13.27	14.93	10.68
DNMF-W-LS	**6.61**	**8.40**	**9.97**	**11.47**	**13.51**	**15.17**	**10.86**

Table 1: Source separation performance on CHiME Challenge (WSJ-0) test set using $\mu = 5$ and $R = 1000$.

Table 1 shows the results on the CHiME test set by ENMF, NMF+S, SNMF, DNMF-W-KL, and DNMF-W-LS using $\mu = 5$ as tuned on the development set, for $R = 1000$. The results mirror those obtained on the development set. DNMF-W-LS improves over the ENMF baseline by 2.3 dB and over SNMF by .76 dB. There is a larger gain at low SNRs (DNMF-W-LS vs. SNMF: +1.1 dB SDR at -6 dB input SNR). The improvement by DNMF-W-LS over SNMF at each SNR is significant according to a Wilcoxon signed rank test [18]. The attempt at discriminative NMF from [16] does not improve over SNMF in our experiments. When optimizing (9) starting from an optimal SNMF solution, i.e., with $\mu = 5$, only the sparsity cost in the objective function (9) decreased, not the reconstruction terms. This is consistent with the slight SNR improvements reported in [16] because they did not consider $\mu > 0.1$ although performance increased monotonously for $\mu \in [0, 0.1]$.

5. Conclusions

We have presented an effective discriminative approach for training NMF reconstruction bases that provide the best reconstruction of the source given the activations obtained in conventional supervised NMF, yielding significant performance gains in the CHiME speech/noise separation task. While it does require parallel training data, it does not increase complexity at test time with respect to conventional NMF. The method is able to yield good performance even on top of sub-optimal analysis bases, such as bases trained with very low or very high sparsity. This suggests that our method could also improve in cases where sparse NMF is not an appropriate model due to the lack of sparseness, such as polyphonic sources in music. A caveat is that it is unclear how such a method may be used in the context of semi-supervised NMF. Finally, a remaining challenge is to perform joint discriminative training of both analysis and reconstruction bases.

6. References

[1] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-time speech separation by semi-supervised nonnegative matrix factorization," in *Proc. of LVA/ICA*, Mar. 2012, pp. 322–329.

[2] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *Proc. of Interspeech*, 2010, pp. 717–720.

[3] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. of ICA*, 2007, pp. 414–421.

[4] 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, Mar. 2009.

[5] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 1–14, 2007.

[6] P.D. O'Grady and B.A. Pearlmutter, "Discovering convolutive speech phones using sparseness and non-negativity," in *Proc. of ICA*, 2007, pp. 520–527.

[7] J. Eggert and E. Körner, "Sparse coding and NMF," in *Proc. of Neural Networks*, 2004, vol. 4, pp. 2529–2533.

[8] J.T. Geiger, F. Weninger, A. Hurmalainen, J.F. Gemmeke, M. Wöllmer, B. Schuller, G. Rigoll, and T. Virtanen, "The TUM+TUT+KUL approach to the CHiME Challenge 2013: Multi-stream ASR exploiting BLSTM networks and sparse NMF," in *Proc. of CHiME Workshop*, Mar. 2013, pp. 25–30.

[9] B. Colson, P. Marcotte, and G. Savard, "An overview of bilevel optimization," *Annals of operations research*, vol. 153, no. 1, pp. 235–256, 2007.

[10] P. Sprechmann, A.M. Bronstein, and G. Sapiro, "Supervised non-euclidean sparse NMF via bilevel optimization with applications to speech enhancement," in *Proc. of HSCMA*, May 2014.

[11] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[12] F. Weninger and B. Schuller, "Optimization and Parallelization of Monaural Source Separation Algorithms in the openBliSSART Toolkit," *Journal of Signal Processing Systems*, vol. 69, no. 3, pp. 267–277, 2012.

[13] F. Weninger, C. Kirst, B. Schuller, and H.J. Bungartz, "A discriminative approach to polyphonic piano note transcription using supervised non-negative matrix factorization," in *Proc. of ICASSP*, 2013, pp. 6–10.

[14] E.M. Grais and H. Erdogan, "Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation," in *Proc. of Interspeech*, 2013.

[15] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Discriminative non-negative matrix factorization for multiple pitch estimation," in *Proc. of ISMIR*, 2012, pp. 205–210.

[16] Z. Wang and F. Sha, "Discriminative non-negative matrix factorization for single-channel speech separation," in *Proc. of ICASSP*, 2014.

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

[18] M. Hollander and D.A. Wolfe, *Nonparametric Statistical Methods*, John Wiley & Sons, New York, NY, 1973.