SUPERVISED NON-EUCLIDEAN SPARSE NMF VIA BILEVEL OPTIMIZATION WITH APPLICATIONS TO SPEECH ENHANCEMENT

Pablo Sprechmann,[1] *Alex M. Bronstein,*[2] *and Guillermo Sapiro*[1]
[1] Duke University, USA; [2] Tel Aviv University, Israel.

ABSTRACT

Traditionally, NMF algorithms consist of two separate stages: a training stage, in which a generative model is learned; and a testing stage in which the pre-learned model is used in a high level task such as enhancement, separation, or classification. As an alternative, we propose a task-supervised NMF method for the adaptation of the basis spectra learned in the first stage to enhance the performance on the specific task used in the second stage. We cast this problem as a bilevel optimization program that can be efficiently solved via stochastic gradient descent. The proposed approach is general enough to handle sparsity priors of the activations, and allow non-Euclidean data terms such as β-divergences. The framework is evaluated on single-channel speech enhancement tasks.

Index Terms— Supervised learning, tast-specific learning, bilevel, NMF, speech enhancement.

1. INTRODUCTION

The problem of isolating or enhancing a speech signal recorded in a noisy environment has been widely studied in the audio processing community [1, 2]. It becomes particularly challenging in the presences of non-stationary background noise, which is a very common situation in many applications encountered, e.g., in telephony. We approach this problem as a monaural source separation method by modeling the speech as one source, and the noise as the other. This is a natural approach when the characteristics of both the signal of interest and the noise vary throughout time [3, 4, 5, 6].

The decomposition of time-frequency representations, such as the power or magnitude spectrogram in terms of elementary atoms of a dictionary, has become a popular tool in audio processing. In particular, non-negative matrix factorization (NMF) [7], and its probabilistic counterpart, the probabilistic latent component analysis (PLCA) [8], were shown effective for various speech processing tasks such as speech separation [9, 10], denoising [4, 6, 11], and robust automatic speech recognition [12, 13], among many others.

NMF and PLCA produce high quality separation results when the dictionaries for different sources are sufficiently distinct. There is naturally a compromise between the approximation of the training data and tightness of the model: the more general is the dictionary the higher is the chance it will include elements that match spectral patterns in the competing sources. In order to mitigate this problem, recent approaches have proposed alternative models constraining the solution in meaningful ways, as for example, by imposing sparsity of the activations [10, 14].

Particularly good performance of NMF-based speech enhancement is achieved in supervised regimes, that include an offline training stage with access to examples of the clean speech signal and, sometimes, of the noise. Separate dictionaries for the speech and the noise are constructed during the training stage. However, the mismatch between the optimization objective used to train the dictionaries and that used to perform the actual estimation at testing time results in suboptimal performance, especially when the speech and the noise signals are of similar nature. In this paper, we propose a supervised dictionary learning scheme that is tailored for the specific task of signal denoising or separation. Following recent ideas proposed in the sparse coding [15], our training scheme is formulated as a bilevel optimization problem, which can be efficiently solved using standard stochastic optimization techniques. These ideas were recently used for enhancing the performance of NMF based music transcription systems [16]. In this work, we adapt them to the speech enhancement context and extend to more general $\beta-$divergences as the fitting cost.

It is worth mentioning that there is much additional structure in speech (as well as in the noise) which is not sufficiently (or at all) exploited in the method discussed in this work. At testing, the proposed method, like standard NMF approaches, treats different time-frames independently, ignoring the temporal dynamics of speech signals. Recent studies have proposed regularized variants of NMF or PLCA trying to overcome this limitation, including co-occurrence statistics of the basis functions [3], smoothness of the activation coefficients [17], and learned temporal dynamics [5, 18, 19]. In all these methods the model is expressed as the minimization of a cost with a data fitting term and some structure-promoting penalties. We argue that many of these models could also benefit from the approach discussed in this paper, since they also share the mismatch between the optimization objective used to train the models and that used at estimation.

Work partially supported by BSE, ONR, NGA, NSF, ARO, and AFOSR. AB is supported by the ERC grant 335491 (RAPID).

2. NMF SPEECH ENHANCEMENT

NMF-based denoising techniques typically operate on the (non-negative) magnitude or the power spectrum. Given the noisy signal $\mathbf{V} \in \mathbb{R}^{m \times n}$ comprising m frequency bins and n temporal frames, NMF attempts to find the non-negative activations $\mathbf{H}_\mathrm{s} \in \mathbb{R}^{q \times n}$ and $\mathbf{H}_\mathrm{n} \in \mathbb{R}^{r \times n}$ best representing the speech and the noise components, respectively, in two fixed dictionaries $\mathbf{W}_\mathrm{s} \in \mathbb{R}^{n \times q}$ and $\mathbf{W}_\mathrm{n} \in \mathbb{R}^{n \times r}$. This task is achieved through the solution of the minimization problem

$$\min_{\mathbf{H}_\mathrm{s}, \mathbf{H}_\mathrm{n} \geq \mathbf{0}} D(\mathbf{V} | \mathbf{W}_\mathrm{s}\mathbf{H}_\mathrm{s} + \mathbf{W}_\mathrm{n}\mathbf{H}_\mathrm{n}) + \lambda\, \psi(\mathbf{H}_\mathrm{s}, \mathbf{H}_\mathrm{n}). \tag{1}$$

The first term in the optimization objective is a divergence measuring the dissimilarity between the input data and the estimated channels. Typically, this data fitting term is assumed to be separable,

$$D(\mathbf{A}|\mathbf{B}) = \sum_{i,j} D(a_{ij}|b_{ij}).$$

Significant attention has been devoted in the literature to the case in which the scalar divergence D belongs to the family of the so-called β-divergences [20],

$$D_\beta(a|b) = \begin{cases} \frac{a}{b} - \log\frac{a}{b} - 1 & : \beta = 0, \\ a \log a/b + (a-b) & : \beta = 1, \\ \frac{1}{\beta(\beta-1)}(a^\beta + (\beta-1)b^\beta - \beta a b^{\beta-1}) & : \text{otherwise.} \end{cases}$$

This family includes the three most widely used cost functions in NMF: the squared Euclidean distance ($\beta = 2$), the Kullback-Leibler divergence ($\beta = 1$), and the Itakura-Saito divergence ($\beta = 0$). For $\beta \geq 1$, the divergence is convex. The case of $\beta = 0$ is attractive despite the lack of convexity, due to the scale-invariance of the Itakura-Saito divergence, which makes the NMF procedure insensitive to volume changes

The second (optional) term in the minimization objective is included to promote some desired structure of the activations. This is done using a designed regularization function ψ and its relative importance is controlled by the parameters λ.

Once the optimal activations are solved for, the spectral envelopes of the speech and the noise are estimated as $\mathbf{W}_\mathrm{s}\mathbf{H}_\mathrm{s}$ and $\mathbf{W}_\mathrm{n}\mathbf{H}_\mathrm{n}$, respectively. Since these estimated speech spectrum envelope contains no phase information, speech signal is estimated from the mixture by Wiener filtering.

In supervised NMF, the speech and noise dictionaries are trained independently from available training data. The underlying assumption of this approach is that the speech and the noise signals forming the mixture are sufficiently distinct to be unambiguously decomposed into $\mathbf{V} \approx \mathbf{W}_\mathrm{s}\mathbf{H}_\mathrm{s} + \mathbf{W}_\mathrm{n}\mathbf{W}_\mathrm{n}$. However, this assumption is often violated, e.g., in the presence of multitalker babble noise, when the learned speech and noise dictionaries might be very similar (or coherent). In other words, the independently trained dictionaries do not ensure that the solutions $\mathbf{W}_\mathrm{s}\mathbf{H}_\mathrm{s}$ and $\mathbf{W}_\mathrm{n}\mathbf{H}_\mathrm{n}$ obtained from (2) will resemble the original components of the mixture.

2.1. Case study

The method proposed in this paper, described in Section 3, can be applied to a large family of approaches following the supervised NMF paradigm. In this paper, we opted to use a sparsity-regularized version of NMF as a case study. In this case, the regularizer is given by the ℓ_1 norm,

$$\begin{aligned} \min_{\mathbf{H}_\mathrm{s}, \mathbf{H}_\mathrm{n} \geq \mathbf{0}} \quad & D(\mathbf{V} | \mathbf{W}_\mathrm{s}\mathbf{H}_\mathrm{s} + \mathbf{W}_\mathrm{n}\mathbf{H}_\mathrm{n}) + \lambda_\mathrm{s}\|\mathbf{H}_\mathrm{s}\|_1 \\ & + \lambda_\mathrm{n}\|\mathbf{H}_\mathrm{n}\|_1 + \tfrac{\mu}{2}(\|\mathbf{H}_\mathrm{s}\|_2^2 + \|\mathbf{H}_\mathrm{n}\|_2^2). \end{aligned} \tag{2}$$

For technical reasons, that will be clear in Section 4, we also include an ℓ_2 regularizer on the activations. The speech dictionary is trained in the supervised regime by solving

$$\min_{\mathbf{H}_\mathrm{s}, \mathbf{W}_\mathrm{s} \geq \mathbf{0}} D(\mathbf{V}_\mathrm{s} | \mathbf{W}_\mathrm{s}\mathbf{H}_\mathrm{s}) + \lambda_\mathrm{s}\|\mathbf{H}_\mathrm{s}\|_1 + \mu\|\mathbf{H}_\mathrm{s}\|_2^2 \tag{3}$$

on a training set $\mathbf{V}_\mathrm{s}$ of clean speech signals. A similar procedure is performed independently for the noise dictionary.

3. TASK-SPECIFIC SUPERVISED NMF

The main motivation of our paper is the realization that the optimization problem (2) is merely a proxy to the estimation of the speech signal, and that the standard dictionary learning does not guarantee that its solutions will produce the best speech estimate even on mixtures created from the training data. Ideally, we would like to train the dictionaries that explicitly maximize the performance of the specific task at hand, namely, the separation of speech from the background noise.

In what follows, we denote the solutions of (2) by $\mathbf{H}^*_\mathrm{s}(\mathbf{V}, \mathbf{W})$ and $\mathbf{H}^*_\mathrm{n}(\mathbf{V}, \mathbf{W})$, where $\mathbf{V} \approx \mathbf{V}_\mathrm{s} + \mathbf{V}_\mathrm{n}$ (the sum is approximate due to the non-linear effects of the phase), $\mathbf{W} = [\mathbf{W}_\mathrm{s}, \mathbf{W}_\mathrm{n}]$ is a matrix concatenating both speech and noise dictionaries, and $\mathbf{H}^*$ is the vertical concatenation of the optimal activations such that the product $\mathbf{W}\mathbf{H}^*$ is well defined. We cast the *task-specific* NMF problem as the minimization of

$$\min_{\mathbf{W}_\mathrm{s}, \mathbf{W}_\mathrm{n} \geq \mathbf{0}} \ell(\mathbf{V}_\mathrm{s}, \mathbf{V}_\mathrm{n}, \mathbf{W}_\mathrm{s}, \mathbf{W}_\mathrm{n}, \mathbf{H}^*(\mathbf{V}, \mathbf{W})), \tag{4}$$

where ℓ is a cost function measuring how well the speech and the noise signals are separated. In this work we use a cost of the form

$$D(\mathbf{V}_\mathrm{s} | \mathbf{W}_\mathrm{s}\mathbf{H}^*_\mathrm{s}(\mathbf{V}, \mathbf{W})) + \alpha D(\mathbf{V}_\mathrm{n} | \mathbf{W}_\mathrm{n}\mathbf{H}^*_\mathrm{n}(\mathbf{V}, \mathbf{W})), \tag{5}$$

where α is a parameter controlling the relative importance of background recovery; typically, one would set $\alpha = 0$ in a denoising application, and $\alpha = 1$ in a source separation application. Naturally, this is just an example, one could consider other types of cost functions for evaluating the quality of the separation (e.g., perceptually motivated measures). We also imposed the norm of the columns of $\mathbf{W}_\mathrm{s}$ and $\mathbf{W}_\mathrm{n}$ to be smaller or equal than one as standard in sparse dictionary learning [21].

Note that the objective of (4) depends on the minimizers $\mathbf{H}_{\mathrm{s}}^*$ and $\mathbf{H}_{\mathrm{n}}^*$ of the estimation problem (2). Such optimization problems are referred to as *bilevel*. In what follows, we leverage the recent development in bilevel optimization techniques for supervised dictionary learning [15] to formulate a practical numerical scheme for the solution of (4).

4. OPTIMIZATION

As NMF, the bilevel optimization problem (4) is non-convex. Hence, we aim at finding a good local minimizer. Bellow we describe the general optimization algorithm used for this purpose.

4.1. Stochastic gradient descent

Problem (2) has a unique solution when $\beta \geq 1$ and $\mu > 0$, due to the strict convexity of the objective. In this situation, a local minimizer of (4) can be found via (projected) stochastic gradient descent (SGD) [22]. SGD is a gradient descent optimization algorithm for minimizing an objective function expressed as a sum or average of some training data of an almost-everywhere differentiable function. At each iteration, the gradient of the objective function is approximated using a randomly picked sub-sample.

For solving (2) the process goes as follows: we first randomly draw a speech sample and a noise sample from the training data and sum them together to obtain a mixture sample. We denote by $\mathbf{v}^i$, $\mathbf{v}_{\mathrm{n}}^i$ and $\mathbf{v}_{\mathrm{s}}^i$ the spectral samples of the mixture, noise, and speech at iteration i respectively. Then the combined dictionary at iteration $i+1$, $\mathbf{W}^{i+1} = [\mathbf{W}_{\mathrm{s}}^{i+1}, \mathbf{W}_{\mathrm{n}}^{i+1}]$, is obtained by

$$\mathbf{W}^{i+1} \leftarrow \mathcal{P}(\mathbf{W}^i - \eta_i \nabla_{\mathbf{W}} \ell(\mathbf{v}_{\mathrm{s}}^i, \mathbf{v}_{\mathrm{n}}^i, \mathbf{W}_{\mathrm{s}}^i, \mathbf{W}_{\mathrm{n}}^i, \mathbf{h}^*(\mathbf{v}^i, \mathbf{W}^i))),$$

where $0 \leq \eta_i \leq \eta$ is a decreasing sequence of step-sizes, and $\mathcal{P}$ is an operator that projects a matrix to be non-negativete and have columns with norm smaller or equal than one. Note that the learning requires the gradient $\nabla_{\mathbf{W}} \ell$, which in turn relies (via the chain rule) on the gradient of $\nabla_{\mathbf{W}} \mathbf{h}^*(\mathbf{v}, \mathbf{W})$ with respect to $\mathbf{W}$. Even though $\mathbf{h}^*(\mathbf{v}, \mathbf{W})$ is obtained by solving a non-smooth optimization problem, it is almost everywhere differentiable, and one can compute its gradient with respect to $\mathbf{W}$ in a closed form. In the next section, we briefly summarize the derivation of the gradients $\nabla_{\mathbf{W}} \ell$.

Following [15], we use a step size of the form $\eta_i = \eta \min(1, i_0/i)$ in all our experiments, which means that a fixed step size is used during the first i_0 iterations, after which it decays according to the $1/i$ annealing strategy. We set in all our experiments i_0 to be half of the total number of iterations. A common heuristic used in practice for accelerating the convergence speed of SGD algorithms consists randomly drawing several samples (a minibatch) at each iteration instead of a single one. A natural initialization of the speech and noise dictionaries is the individual training via the solution of (3), as in standard supervised NMF denoising.

4.2. Gradient computation

We consider (2) in vectorial form, i.e., matrices with a single column. Let us denote by Λ the active set of the solution of (2), this is, the indexes of the non-zero coefficients of $\mathbf{h}^*$. We will use the sub-index Λ to indicate the sub-vector restricted to the active set, e.g., $\mathbf{h}_\Lambda^*$. The first-order optimality conditions of (2) require the derivatives with respect to $\mathbf{h}_\Lambda$ to be zero,

$$\mathbf{W}_\Lambda^{\mathrm{T}} \boldsymbol{\phi} + \mathbf{p}_\Lambda + \mu\, \mathbf{h}_\Lambda^* = \mathbf{0}, \tag{6}$$

where $\mathbf{W}_\Lambda$ is the matrix retaining only the columns of the dictionary associated with the active set, $\mathbf{p}$ is a vector which takes the value λ_{s} and λ_{n} in the coefficients of Λ belonging to $\mathbf{h}_{\mathrm{s}}^*$ and $\mathbf{h}_{\mathrm{n}}^*$, respectively, and zero otherwise, and $\boldsymbol{\phi} = (\mathbf{W}_\Lambda \mathbf{h}_\Lambda^*)^{\beta-2} \odot (\mathbf{W}_\Lambda \mathbf{h}_\Lambda^* - \mathbf{v})$, where the product $\odot$ and the exponential are element-wise operations.

A key observation is that, almost surely, the set of active constraints in the solution of (2) remains constant on a local neighborhood of $\mathbf{v}$ and $\mathbf{W}$ [23]. That is, for small changes in the dictionary, the active set Λ remains constant. Based on this property, we know that only the gradient $\nabla_{\mathbf{W}_\Lambda} \mathbf{h}^*$ will be non-zero. Changes in the columns of $\mathbf{W}$ that do not affect the coefficients in Λ do not change the cost function.

Taking the derivative in (6) with respect to $\mathbf{W}_\Lambda$ we obtain,

$$d\mathbf{W}_\Lambda^{\mathrm{T}} \boldsymbol{\phi} + \mathbf{W}_\Lambda^{\mathrm{T}} \boldsymbol{\Phi}(d\mathbf{W}_\Lambda \mathbf{h}_\Lambda^* + \mathbf{W}_\Lambda d\mathbf{h}_\Lambda^*) + \mu\, d\mathbf{h}_\Lambda^* = \mathbf{0}, \tag{7}$$

where we used d to denote the differentials, and

$$\boldsymbol{\Phi} = \operatorname{diag}\big((\mathbf{W}_\Lambda \mathbf{h}_\Lambda^*)^{\beta-2} + (\beta-2)(\mathbf{W}_\Lambda \mathbf{h}_\Lambda^*)^{\beta-3} \odot (\mathbf{W}_\Lambda \mathbf{h}_\Lambda^* - \mathbf{v})\big).$$

Invoking the chain rule, we have

$$\nabla_{\mathbf{W}} \ell = \operatorname{trace}(\nabla_{\mathbf{h}^*} \ell^{\mathrm{T}} d\mathbf{h}^*) + \nabla_{\mathbf{W}} \hat{\ell}, \tag{8}$$

where $\nabla_{\mathbf{W}} \hat{\ell}$ represents the gradient of ℓ with respect to $\mathbf{W}$ assuming $\mathbf{h}^*$ fixed. Combining (7) and (8) follows that

$$\nabla_{\mathbf{W}} \ell = \boldsymbol{\phi}\, \boldsymbol{\xi}^{\mathrm{T}} + \boldsymbol{\Phi} \mathbf{W}_\Lambda \boldsymbol{\xi} {\mathbf{h}_\Lambda^*}^{\mathrm{T}}, \tag{9}$$

where $\boldsymbol{\xi} = \mathbf{Q} \nabla_{\mathbf{h}^*} \ell$, and $\mathbf{Q} = (\mathbf{W}_\Lambda^{\mathrm{T}} \boldsymbol{\Phi} \mathbf{W}_\Lambda + \mu \mathbf{I})^{-1}$. Note that the size of the matrix being inverted is given by the sparsity level of the representation.

5. EXPERIMENTAL RESULTS

Data sets. We evaluated the separation performance of the proposed methods on a subset of the GRID dataset [24]. Three randomly chosen sets of distinct clips each were used

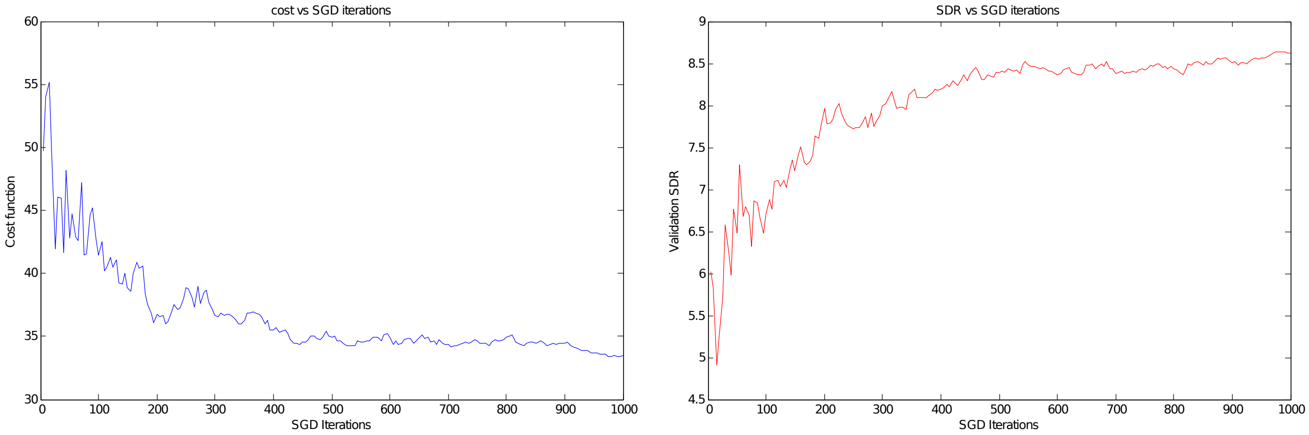

Fig. 1: Evolution of the average high level cost function (left) and the average SDR (in dB) on the validation set with the SGD iterations.

for training (500 clips), validation (10 clips), and testing (50 clips). The clips were resampled to 8 KHz. For the noise signals we used the AURORA corpus [25], which contains six categories of noise recorded from different real environments (street, restaurant, car, exhibition, train, and airport). As before, three sets of distinct clips each were used for training (15 clips), validation (3 clips), and testing (15 clips).

Evaluation measures. As the evaluation criteria, we used the *source-to-distortion ratio* (SDR), *source-to-interference ratio* (SIR), and *source-to-artifact ratio* (SAR) from the BSS-EVAL metrics [26]. We also computed the standard *signal-to-noise ratio* (SNR). When dealing with several frames, we computed a global score (GSDR, GSIR, GSAR and GSNR) by averaging the metrics over all test clips from the same speaker and noise weighted by the clip duration.

Training setting. The same training settings were used in all experiments. We used dictionaries of size 60 and 10 atoms for representing the speech and noise, respectively. These values were obtained using cross-validation. We used $\lambda_{\rm s} = 0.1$ and $\lambda_{\rm n} = 0$ (which means that no sparsity was promoted in the representation of the noise) and $\mu = 0.001$. As the example, we chose $\beta = 1$, which corresponds to the Kullback-Leibler divergence, and $\alpha = 0$ in the high level cost (4). For the SGD algorithm we used $\eta = 0.1$ and minibatch of size 50. These were obtained by trying several values of during a small number of iterations, keeping those producing the lowest error on a small validation set. All training signals where mixed at $5\ dB$.

Results. Figure 1 shows the evolution of the high level cost (4) and the SDR on the validation set with the SGD iterations. The algorithm converges to a dictionary that achieves about 2 dB better SDR on the validation set. Tables 1 and 2 show some initial results for the proposed approach. We compare the performance of standard supervised sparse-NMF (referred simply as NMF) against the performance of the same sparse-NMF model trained on a task-specific manner (referred as TS-NMF). Observe that the task-specific supervision leads to improvements in performance, maintaining (at $5dB$ SNR) the improvements observed on the validation set. In future work we plan to analyze what happens when a non-speaker specific dictionaries are trained. We expect to observe similar improvements, if the training data is diverse enough.

Table 1: Average performance (in dB) for NMF and proposed supervised NMF methods measured in terms of SDR, SIR, SAR and SNR. Speech and noise were mixed at $5dB$ of SNR. The standard deviation of each result is shown between brackets.

	SDR	SIR	SAR	SNR
NMF	7.5 [1.5]	13.7 [0.9]	8.9 [1.7]	8.2 [1.3]
TS-NMF	9.3 [1.1]	13.3 [0.5]	11.8 [1.6]	9.7 [0.9]

Table 2: See description of Table 1. In this case, speech and noise were mixed at $0dB$ of SNR.

	SDR	SIR	SAR	SNR
NMF	4.6 [1.1]	9.3 [0.9]	6.8 [1.2]	5.8 [0.8]
TS-NMF	5.5 [0.8]	9.1 [0.5]	8.6 [1.0]	6.2 [0.5]

6. CONCLUSION

In this work we presented an algorithm for the task-supervised training of NMF models. Unlike standard supervised NMF, the proposed approach matches the optimization objective used at the train and testing stages. In this way, the dictionaries can be trained in a task-specific manner. We cast this problem as bilevel optimization that can be efficiently solved via stochastic gradient descent. The proposed approach allows non-Euclidean data terms such as β-divergences. A limited case study of sparse NMF with task specific supervision demonstrates promising results. Including temporal dynamics into this model is the subject of ongoing research.

7. REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, vol. 30, CRC, 2007.

[2] E. Hänsler and G. Schmidt, *Speech and Audio Processing in Adverse Environments*, Springer, 2008.

[3] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *ICASSP*, 2008, pp. 4029–4032.

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

[5] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *ICASSP*, 2011, pp. 17–20.

[6] Z. Duan, G. J. Mysore, and P. Smaragdis, "Online plca for real-time semi-supervised source separation," in *LVA/ICA*, 2012, pp. 34–41.

[7] D.D. Lee and H.S. Seung, "Learning parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[8] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," *Advances in models for acoustic processing, NIPS*, vol. 148, 2006.

[9] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *INTERSPEECH*, Sep 2006.

[10] M. V. S. Shashanka, B. Raj, and P. Smaragdis, "Sparse Overcomplete Decomposition for Single Channel Speaker Separation," in *ICASSP*, 2007.

[11] M. N. Schmidt, J. Larsen, and F.-T. Hsiao, "Wind noise reduction using non-negative sparse coding," in *MLSP*, Aug 2007, pp. 431–436.

[12] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, no. 7, pp. 2067–2080, 2011.

[13] F. Weninger, M. Wöllmer, J. T. Geiger, B. Schuller, J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and G. Rigoll, "Non-negative matrix factorization for highly noise-robust asr: To enhance or to recognize?," in *ICASSP*, 2012, pp. 4681–4684.

[14] P.O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[15] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 4, pp. 791–804, 2012.

[16] T. BenYakar, R. Litman, P. Sprechmann, A. Bronstein, and G. Sapiro, "Bilevel sparse models for polyphonic music transcription," in *ISMIR*, 2013.

[17] C. Févotte, "Majorization-minimization algorithm for smooth itakura-saito nonnegative matrix factorization," in *ICASSP*. IEEE, 2011, pp. 1980–1983.

[18] J. Han, G. J. Mysore, and B. Pardo, "Audio imputation using the non-negative hidden markov model," in *LVA/ICA*, 2012, pp. 347–355.

[19] C. Févotte, J. Le Roux, and J. R. Hershey, "Non-negative dynamical system with application to speech and audio," in *ICASSP*, 2013.

[20] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[21] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *ICML*, 2009, pp. 689–696.

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

[23] Gopal Vasudevan, Layne Terry Watson, and FH Lutze, "Homotopy approach for solving constrained optimization problems," *Automatic Control, IEEE Transactions on*, vol. 36, no. 4, pp. 494–498, 1991.

[24] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. of the Acoustical Society of America*, vol. 120, pp. 2421, 2006.

[25] D. Pearce and H.-G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *INTERSPEECH*, 2000, pp. 29–32.

[26] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 14, no. 4, pp. 1462–1469, 2006.