2013 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 22–25, 2013, SOUTHAMPTON, UK

SINGLE CHANNEL SOURCE SEPARATION USING SMOOTH NONNEGATIVE MATRIX FACTORIZATION WITH MARKOV RANDOM FIELDS

Minje Kim

University of Illinois at Urbana-Champaign
Department of Computer Science
Urbana, IL, USA 61801

Paris Smaragdis

University of Illinois at Urbana-Champaign
and
Adobe Systems, Inc.

ABSTRACT

This paper presents a single channel source separation method based on an extension of Nonnegative Matrix Factorization (NMF) algorithm by smoothing the original posterior probabilities with an additional Markov Random Fields (MRF) structure. Our method is based on the alternative interpretation of NMF with β-divergence as latent variable models. By doing so, we can redefine NMF-based separation procedure as a Bayesian labeling problem where each label stands for the mask for a specific source. This understanding leads us to intervene in the calculation of posterior probabilities, so that the priors from MRF's neighboring structure can smooth out isolated masking values that have different labeling results from their neighbors. Experiments on several dictionary-based source separation tasks show sensible performance gains.

Index Terms— Markov Random Fields, Nonnegative Matrix Factorization, Probabilistic Latent Component Analysis, Probabilistic Latent Semantic Indexing, Informed Source Separation

1. INTRODUCTION

Nonnegative Matrix Factorization (NMF) [1, 2] has been widely used to analyze magnitude or power spectrograms to reveal underlying latent components of the input audio signals. Its parts-based representation gives intuitive analysis results for music signals [3], where each latent variable corresponds to a note. The NMF-based music transcription task also initiated audio source separation works based on spectrogram decomposition using NMF, such as NMF with additional temporal continuity of sources and sparseness constraints [4], speech enhancement using source and noise statistics [5], drum source separation with Nonnegative Matrix Partial Co-Factorization (NMPCF) [6], etc.

Most of those separation algorithms eventually use the NMF results to calculate the posterior probabilities of latent variables given the time-frequency observation, so that they are multiplied to the mixture spectrogram as a soft mask such as in Wiener filtering. Therefore, it is straightforward to regard the NMF's multiplicative update rules as an EM-like algorithm once they are using the family of β-divergence [7, 8] to measure the error between the input and the reconstruction. In this way we do not have to change the multiplicative updates, but can reformulate the system to be readily harmonized with MRF structures.

Although there have been a lot of effort to regularize NMF with some amount of prior information, such as modeling temporal dependencies using a Kalman-like prediction [9], harmonic-temporal clustering [10], nonngeative factorial hidden Markov models [11], etc, each specific variant requires a particular inference method that hinders generalization.

However, as underdetermined source separation problems wind up seeking a proper masking of mixture spectrograms, we can consider the NMF-based single channel spectrogram masking as a Bayesian labeling problem, too. It is important to extract posterior probabilities from NMF's ordinary multiplicative update rules, because then we can conveniently take our prior knowledge about the latent components into account.

Based on the previous masking-based source separation method in [12], binaural cues, such as Interchannel Level Differences (ILD), of multi-channel inputs can be tackled by this kind of clustering approach, too. In [13], Markov Random Fields (MRF) was successfully harmonized into the ILD clustering problem, so that the neighboring structure of latent variables can be involved into the process in a controlled manner. As shown in many computer vision tasks, such as image denoising, segmentation, and stereo matching, MRF can provide standardized way to communicate with locally relevant labeling results and smooth out the results.

The system in [13], however, deals with ILD features as its input that limits its application to the cases where more than one mixture signals are available. The proposed smooth NMF in this paper also harmonizes the MRF structure into the source separation problem, but now considers a single mag-

This work was supported by the Intel Science and Technology Center for Embedded Computing (ISTC-EC).

978-1-4799-1180-6/13/$31.00

nitude spectrogram as the input matrix. At the same time, the input matrix is decomposed with the help of NMF in place of Gaussian mixture models underlying the referred multichannel work.

In this paper, we first briefly introduce NMF and MRF in section 2, and then reformulate the NMF algorithm as a Bayesian labeling problem in section 3 to derive the proposed smooth NMF, which is defined by a new Maximum A Posteriori (MAP) estimation approach with MRF's smoothness costs. We compare the smooth NMF with conventional one in terms of single channel convolutive source separation performances on a few mixtures with different types of interferences in section 4. Section 5 concludes the work.

2. BACKGROUND

2.1. NMF with β-divergence

NMF takes a nonnegative matrix $\mathbf{V} \in \mathbb{R}_+^{M\times N}$ as input and tries to approximate it with a pair of factor matrices $\mathbf{W} \in \mathbb{R}_+^{M\times R}$ and $\mathbf{H} \in \mathbb{R}_+^{R\times N}$, where the set $\mathbb{R}_+$ stands for nonnegative real numbers, and R is for the number of latent components [1, 2]. A generalized way to measure the approximation error between the input $\mathbf{V}$ and the reconstruction $\mathbf{WH} = \sum_{\mathbf{z}=1}^{R} \mathbf{w}_\mathbf{z}\mathbf{h}_\mathbf{z}$ can be the β-divergence, which is defined by

$$\mathcal{D}_\beta(x|y) = \begin{cases} \frac{x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}}{\beta(\beta-1)} & \beta \in \mathbb{R}\backslash\{0,1\} \\ x(\log x - \log y) + (y-x) & \beta = 1 \\ \frac{x}{y} - \log\frac{x}{y} - 1 & \beta = 0 \end{cases} \tag{1}$$

for any pair of elements x and y in the input and the reconstruction, respectively. Note that (1) reduces to Frobenius norm, unnormalized Kullback-Leibler divergence, and Itakura-Saito divergence [14] when β equals to 2, 1, and 0, respectively. Therefore, the objective function of NMF can be defined as follows:

$$\mathcal{J}_\beta = \mathcal{D}_\beta\Big(\mathbf{V}\Big|\sum_{\mathbf{z}=1}^{R}\mathbf{w}_\mathbf{z}\mathbf{h}_\mathbf{z}\Big). \tag{2}$$

Using the fact that the derivative of the $\mathcal{D}_\beta(x|y)$ with respect to y is

$$\frac{\partial \mathcal{D}_\beta(x|y)}{\partial y} = y^{\beta-2}(y-x), \tag{3}$$

we can calculate the derivatives of the objective function (2) as follows:

$$\begin{aligned} \frac{\partial \mathcal{J}_\beta}{\partial \mathbf{w}_\mathbf{z}} &= \Big\{(\mathbf{WH})^{(\beta-2)} \odot (\mathbf{WH}-\mathbf{V})\Big\}\mathbf{h}_\mathbf{z}^\top, \\ \frac{\partial \mathcal{J}_\beta}{\partial \mathbf{h}_\mathbf{z}} &= \mathbf{w}_\mathbf{z}^\top\Big\{(\mathbf{WH})^{(\beta-2)} \odot (\mathbf{WH}-\mathbf{V})\Big\}, \end{aligned} \tag{4}$$

where $\odot$ is for Hadamard products and exponentiations are carried in the element-wise manner as well.

We can derive the multiplicative update rules of NMF by selecting the step size of the gradient descent method in such a way that it turns the update into a multiplicative form. An alternative view of this process is to simply choose the negative and positive terms of the derivative as the numerator and the denominator, respectively, which in turn produces following update rules:

$$\begin{aligned} \mathbf{w}_\mathbf{z} &\leftarrow \mathbf{w}_\mathbf{z} \odot \frac{\Big\{(\mathbf{WH})^{(\beta-2)} \odot \mathbf{V}\Big\}\mathbf{h}_\mathbf{z}^\top}{(\mathbf{WH})^{(\beta-1)}\mathbf{h}_\mathbf{z}^\top}, \\ \mathbf{h}_\mathbf{z} &\leftarrow \mathbf{h}_\mathbf{z} \odot \frac{\mathbf{w}_\mathbf{z}^\top\Big\{(\mathbf{WH})^{(\beta-2)} \odot \mathbf{V}\Big\}}{\mathbf{w}_\mathbf{z}^\top(\mathbf{WH})^{(\beta-1)}}. \end{aligned} \tag{5}$$

2.2. Single-channel source separation using NMF

With some training signals from clean sources, we can learn the bases in advance and fix them to learn their activations only in the mixture. By using the update rules (5) we can learn the desired source' and interference' spectrum bases $\mathbf{W}^s \in \mathbb{R}_+^{M\times R^s}$ and $\mathbf{W}^n \in \mathbb{R}_+^{M\times R^n}$, respectively, from the training signals that contain either the source or interference. Now that we have fixed basis vectors from two training signals, we can use them to separate mixture of unseen signals $\mathbf{V}$, but of the same kind of sources and interferences. Usually the number of components R^s and R^n are unknown, so we need to heuristically guess or investigate them as in [15].

In the separation step, we define the fixed $\mathbf{W} = [\mathbf{W}^s, \mathbf{W}^n]$, and the number of component $R = R^s + R^n$. Then, the activation matrix $\mathbf{H}$ we get by using (5), but skipping the update for $\mathbf{w}_\mathbf{z}$, consists of two source groups, $\mathbf{h}_{\mathbf{z}\in\{1:R^s\}}$ and $\mathbf{h}_{\mathbf{z}\in\{R^s+1:R\}}$. Therefore, the desired source spectrogram can be separated from the mixture by masking the mixture spectrogram $\mathbf{V}$ with the proportion of the corresponding components in the total reconstruction as follows:

$$\begin{aligned} \mathbf{V} &\approx \mathbf{V} \odot \frac{\sum_{\mathbf{z}=1}^{R^s}\mathbf{w}_\mathbf{z}\mathbf{h}_\mathbf{z}}{\mathbf{WH}} + \mathbf{V} \odot \frac{\sum_{\mathbf{z}=R^s+1}^{R}\mathbf{w}_\mathbf{z}\mathbf{h}_\mathbf{z}}{\mathbf{WH}} \\ &= \mathbf{V} \odot \mathbf{P}_{\mathbf{z}\in\{1:R^s\}} + \mathbf{V} \odot \mathbf{P}_{\mathbf{z}\in\{R^s+1:R\}}. \end{aligned} \tag{6}$$

A soft mask matrix for a particular component $\mathbf{z}$ can be seen as posterior probabilities $\mathbf{P}_\mathbf{z} = P(\mathbf{z}|\mathbf{V})$ of the latent variable $\mathbf{z}$ given the observed input magnitude spectrogram $\mathbf{V}$. However, note that the masks for the two groups of components, $\mathbf{P}_{\mathbf{z}\in\{1:R^s\}}$ and $\mathbf{P}_{\mathbf{z}\in\{R^s+1:R\}}$, are actually sums of component-wise masks:

$$\mathbf{P}_{\mathbf{z}\in\{1:R^s\}} = \sum_{\mathbf{z}=1}^{R^s}\mathbf{P}_\mathbf{z}, \quad \mathbf{P}_{\mathbf{z}\in\{R^s+1:R\}} = \sum_{\mathbf{z}=R^s+1}^{R}\mathbf{P}_\mathbf{z}. \tag{7}$$

2.3. MRF for clustering

As undirected graphical models, MRF try to model the observations with two sets of relationships: edge potentials be-

tween latent variables and node potentials between latent variables and observations. Any clustering problem with class labels as a latent variable can be represented with MRF if we have proper prior knowledge among local pixels' labels. It is often convenient to use these concepts for processing grid structured data, e.g. digital images and Short-Time Fourier Transformed (STFT) signals.

Figure 1 represents an MRF structure that assumes similarity between a node and its four surrounding neighbors. The circled nodes of the MRF correspond to a set of R labels that make up the mask we seek to find while the filled squares $\mathbf{V}_{f,t}$ stand for deterministic observed pixel values. As aforementioned, each node is linked to its observation with a node potential $\phi(\mathbf{z}_{f,t}, \mathbf{V}_{f,t})$ and to its neighbors $\mathbf{z}_{k,l}$ with edge potentials $\phi(\mathbf{z}_{f,t}, \mathbf{z}_{k,l})$, where (k,l) indexes the neighboring pixels of (f,t)-th pixel, $\mathcal{N}_{f,t}$.

Therefore, the clustering problem boils down to a MAP estimation where we minimize the posterior probabilities of getting $\mathbf{z}_{f,t}$ given the observation and knowing about its neighbors:

$$P(\mathbf{z}|\mathbf{V}) \propto \prod_{f,t} P(\mathbf{V}_{f,t}|\mathbf{z}_{f,t}) \prod_{k,l \in \mathcal{N}_{f,t}} P(\mathbf{z}_{f,t}|\mathbf{z}_{k,l}) \qquad (8)$$
$$= \frac{1}{Z} \prod_{f,t} \phi(\mathbf{z}_{f,t}, \mathbf{V}_{f,t}) \prod_{k,l \in \mathcal{N}_{f,t}} \phi(\mathbf{z}_{f,t}, \mathbf{z}_{k,l}),$$

where Z is to normalize the unnormalized probabilities ϕ.

MRFs are different from each other by how we define the potentials and neighbors. For instance, they reduce to a Gaussian mixture model when we define a Gaussian generative model per a component for $P(\mathbf{V}_{f,t}|\mathbf{z}_{f,t})$ without additional smoothing.

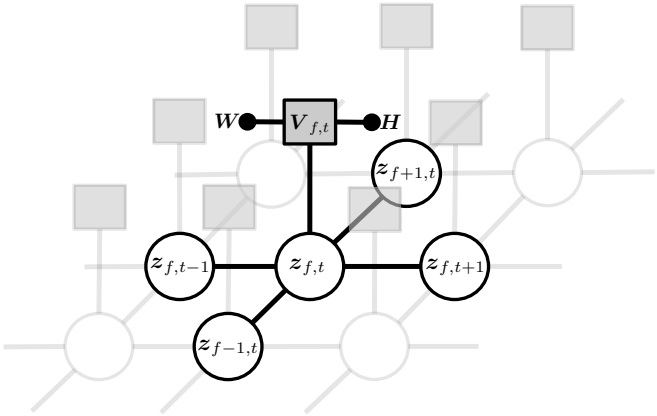


Fig. 1. The proposed pairwise MRF structure with four neighbors and $\mathbf{W}$ and $\mathbf{H}$ as deterministic parameters.

3. NMF AS A BAYESIAN LABELING PROBLEM

NMFs can have different generative models depending on their assumption about the divergence measure. For example, Poisson distributions when $\beta = 1$ [16] and complex Gaussians [14] when $\beta = 0$ are known underlying distributions. In this paper, we adopt $\mathbf{W}$ and $\mathbf{H}$ as deterministic parameters of node potentials (solid circles in Figure 1). As pointed out in [14] the Wiener filtering-like posterior probabilities that are common in audio community, such as $\mathbf{P_z}$ in (6), do not always match with the underlying generative models that vary with β. However, it is also true that the representation works like a mask in practice and generally shows good separation performances.

By using this insight, we stick to use the $\mathbf{P_z}$ representation as our posterior. Then, we can rearrange the NMF update rules into an EM-like ones. First, we explicitly calculate the posterior probabilities $\mathbf{P_z}$ (E-step), and then plug it into the parameter update rules (M-step) as follow:

$$\mathbf{P_z} \leftarrow \frac{\mathbf{w_z h_z}}{\mathbf{WH}}, \qquad (9)$$

$$\mathbf{w_z} \leftarrow \frac{\left\{(\mathbf{WH})^{(\beta-1)} \odot \mathbf{V} \odot \mathbf{P_z}\right\}\mathbf{1}^{N\times 1}}{(\mathbf{WH})^{(\beta-1)}\mathbf{h_z}^\top}, \qquad (10)$$

$$\mathbf{h_z} \leftarrow \frac{\mathbf{1}^{1\times M}\left\{(\mathbf{WH})^{(\beta-1)} \odot \mathbf{V} \odot \mathbf{P_z}\right\}}{\mathbf{w_z}^\top(\mathbf{WH})^{(\beta-1)}}. \qquad (11)$$

Note that these update rules are equivalent to those of Probabilistic Latent Component Analysis (PLCA) [17] when we set $\beta = 1$ and with proper normalizations.

If we ignore edge potentials that are needed for the MRF structure for now, the calculation of posterior probabilities $\mathbf{P_z}$ can be thought of as a special case of (8), where the node potential $\phi(\mathbf{V}_{f,t}, \mathbf{z}_{f,t})$ is defined by

$$\phi(\mathbf{V}_{f,t}, \mathbf{z}_{f,t} = z) = \mathbf{w}_z(f)\mathbf{h}_z(t), \qquad (12)$$

which needs to be normalized by $\sum_{\mathbf{z}=1}^{R} \mathbf{w_z}(f)\mathbf{h_z}(t)$ as posterior probabilities.

3.1. Smooth NMF

If we want to smooth the posterior probabilities with the concept of MRF, we need to define its neighbor structure and edge potentials in between them. Suppose that we are using simple four neighbors as in Figure 1. We define the edge potential with a simple Gaussian-like relationship as follows:

$$\phi(\mathbf{z}_{f,t}, \mathbf{z}_{k,l}) = \mathrm{e}^{-\left\{f(\mathbf{z}_{f,t}, \mathbf{z}_{k,l})\right\}^2/\sigma^2_{\mathcal{N}}}, \qquad (13)$$

where $\sigma^2_{\mathcal{N}}$ is a pre-defined variance. The distance function f sets the distance of two latent variable values in the same source group 0, and 1 otherwise, as follows:

$$f(\mathbf{z}_{f,t}, \mathbf{z}_{k,l}) = \begin{cases} 0 & \text{if } \mathbf{z}_{f,t}, \mathbf{z}_{k,l} \in S^i \\ 1 & \text{otherwise} \end{cases}, \qquad (14)$$

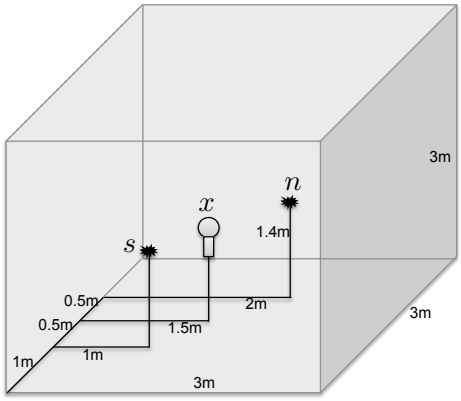


Fig. 2. The mixing environment

where the subset $\mathbf{S}^i$ holds indices of components that belong to i-th source. For instance, in the single channel source separation problem discussed in section 2.2, the $R = R^s + R^n$ components can be divided into two exclusive subsets S^1 and S^2, which represent the indices for the main source s and the interference n, respectively. As an extreme case, we could allow each value of latent variable $\mathbf{z}$ can solely define a subset, too, so that the number of subset equals to the that of components.

In the proposed separation system, we adopt the former relaxed definition of "agreement between neighbors" as the straightforward full relationships between all the latent components $\mathbf{z} = \{1, \cdots, R\}$ are often over-specific when it comes to tens of components per a source.

With some estimation of parameters $\mathbf{W}$ and $\mathbf{H}$ from the previous iteration, the E-step of smooth NMF that substitutes (9) is now defined as follow:

$$\begin{aligned}\max_{\mathbf{z}} \mathbf{P}_{\mathbf{z}} &= \max_{\mathbf{z}} \frac{1}{Z} \prod_{f,t} \phi(\mathbf{V}_{f,t}, \mathbf{z}_{f,t}) \prod_{k,l \in \mathcal{N}_{f,t}} \phi(\mathbf{z}_{f,t}, \mathbf{z}_{k,l}) \\ &= \max_{\mathbf{z}} \frac{1}{Z} \prod_{f,t} \mathbf{w}_{\mathbf{z}}(f)\mathbf{h}_{\mathbf{z}}(t) \prod_{k,l \in \mathcal{N}_{f,t}} \phi(\mathbf{z}_{f,t}, \mathbf{z}_{k,l}),\end{aligned} \tag{15}$$

where the second equation comes from our assumption about the node potentials in (12) and edge potentials in (13).

The new E-step now requires an inference algorithm as the change for a pixel should be propagated to the other nodes. Among many well-known inference algorithms, for instance, this labeling problem can be understood as a hard decision procedure to assign posterior probability 1 to the most probable component at a pixel if we use a certain inference algorithms, such as graph cuts. Otherwise, we can say that the edge potential adjust the current posterior probabilities at (f, t) a little bit to encourage them to be similar to its neighbors' current posterior probabilities. Followed by a proper normalization subject to $\mathbf{z}$, now the rest of the process is to update the parameters $\mathbf{w}_{\mathbf{z}}$ and $\mathbf{h}_{\mathbf{z}}$ using (10) and (11).

In particular, we use Gibbs sampling to solve the MAP problem in the E-step. It is however not confined to use a specific inference algorithm. Instead, we can use any other graphical model inference algorithms to meet the application-specific needs, and it is an advantage of using MRF.

4. EXPERIMENTAL RESULTS

The proposed smooth NMFs is compared with ordinary NMFs in terms of the speech source separation performance. Figure 2 illustrates the room simulation environment with only one microphone in the middle, which captures both the reverberant speech source s and the interference n. It is assumed that the surface of the walls absorbs 50% of the incoming sound waves.

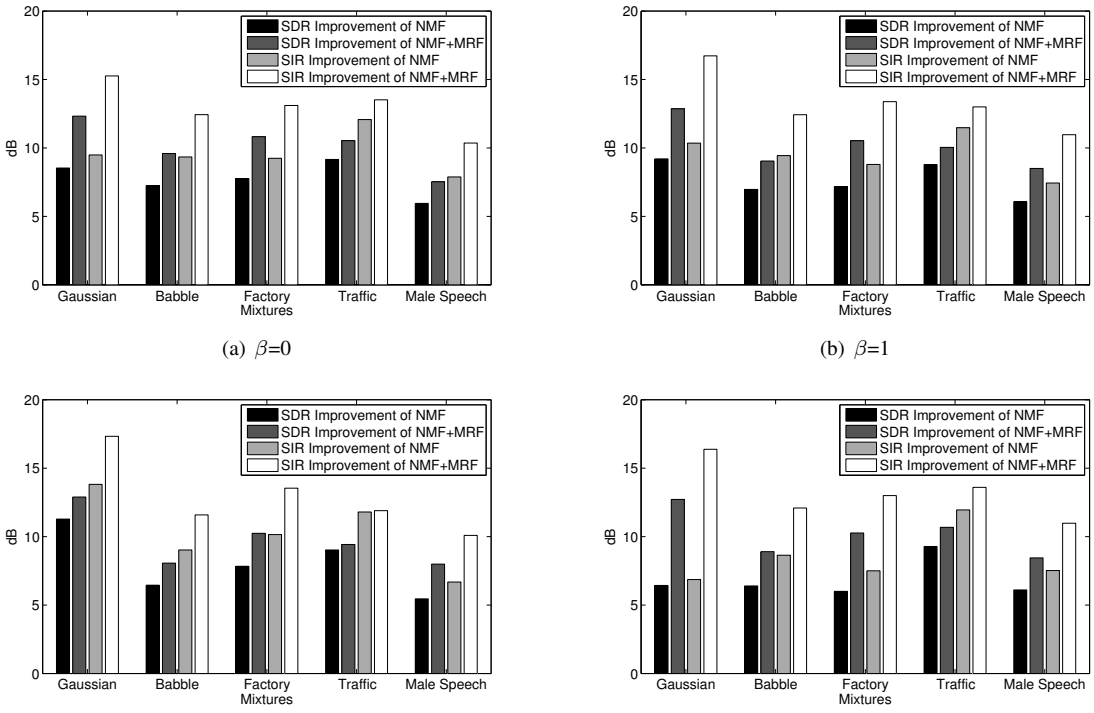
On top of a female speech from TIMIT dataset [18] as the desired source to be enhanced, we add five different kinds of interferences to it: a Gaussian white noise, a babble noise, a factory noise, a traffic noise, and a male speech. The reverberant mixture signal with the 16kHz sampling rate is converted into STFT domain with 1024 points window with 50% overlap to construct the input magnitude spectrogram.

We priorly learn 50 and 30 bases respectively from clean speech signals of the same speaker and the same kind of interferences to fix $\mathbf{W}^s$ and $\mathbf{W}^n$ during the updates. Therefore, there are four different sets of fixed $\mathbf{W}$'s to cover three different choices of $\beta = \{0, 1, 2\}$ plus PLCA. During separation we infer the posterior probabilities using (15) instead of (9), skip (10), and update $\mathbf{h}_{\mathbf{z}}$ using (11) at every iteration.

Figure 3 shows the results when input mixture is -6 dB, which means that n is a lot louder than s. The Signal-to-Interference Ratio (SIR) improvements show that the proposed MRF structure further enhances the plain NMF-based system almost always except the marginal improvement in the traffic noise case with $\beta = 2$ (compare the third and forth bars of each group at every setting of β).

After considering the artifacts that are added during separation procedures the SDR improvements are usually smaller than SIR. However, we can still observe that the proposed MRF structure provides better SDR numbers than NMF only (compare the first and second bars).

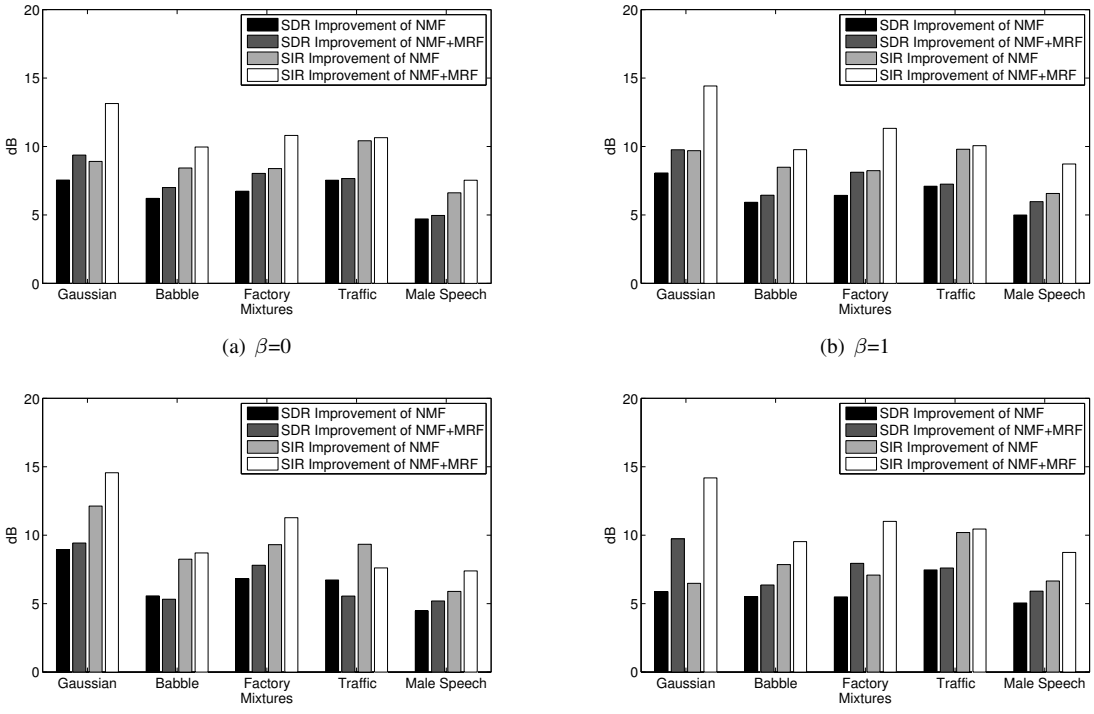
We can check the separation results on the same kinds of mixtures, but with different input mixture SDR in Figure 4. In this second experiment we use 0 dB input where the female speech and the interferences have roughly same amount of energy. First of all, we can see that both NMF and the smooth NMF do their jobs as well, because the improvement bars are positive. In the meantime the merit of introducing MRF structure is mitigated a little especially in the case of the traffic noise with $\beta = 2$. Yet, the proposed method generally outperforms the ordinary NMF.



(a) β=0 (b) β=1

(c) β=2 (d) PLCA (β=1 and normalization)

Fig. 3. Average SDR and SIR Improvements of -6 dB input mixture with 5 different random parameter initializations.



(a) β=0 (b) β=1

(c) β=2 (d) PLCA (β=1 and normalization)

Fig. 4. Average SDR and SIR Improvements of 0 dB input mixture with 5 different random parameter initializations.

5. CONCLUSION

In this paper we provided a harmonization of MRF and NMF based on the understanding of NMF as a Bayesian labeling problem. By isolating the posterior probabilities as such, we can intervene in the calculation in order to make sure the labels are locally associated. The remainder of the NMF update rules can be seen as parameter update procedures that are assumed to be fixed during the MRF inference as an E-step. We checked that the proposed smoothing on NMF could improve the separation results for several different kinds of mixtures.

In the future, we plan to elaborate the proposed model with more audio and speech-centric knowledge. For instance, instead of the four simple neighbors, more complicated neighbor sets that can cover the harmonics structure could maximize the benefit of MRF structures. Furthermore, it is also promising that some of the MRF inference algorithms are ready to be parallelized and then can be accelerated by custom hardwares.

6. 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] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 4029–4032.

[6] M. Kim, J. Yoo, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for spectral and temporal drum source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1192–1204, 2011.

[7] A. Cichocki and S. Amari, "Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities," *Entropy*, vol. 2010, no. 12, pp. 1532–1568, 2010.

[8] M. Minami and S. Eguchi, "Robust blind source separation by beta-divergence," *Neural Computation*, vol. 14, pp. 1859–1886, 2002.

[9] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Prediction based filtering and smoothing to exploit temporal dependencies in nmf," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[10] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 982–994, 2007.

[11] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden markov modeling of audio with application to source separation," in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 140–148.

[12] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[13] M. Kim, P. Smaragdis, G. Ko, and R. Rutenbar, "Stereophonic spectrogram segmentation using markov random fields," in *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Santander, Spain, 2012.

[14] 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.

[15] V. Y. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization," in *SPARS'09-Signal Processing with Adaptive Sparse Structured Representations*, 2009.

[16] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, 2009.

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

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