LOW-ARTIFACT SOURCE SEPARATION USING PROBABILISTIC LATENT COMPONENT ANALYSIS

Nasser Mohammadiha[*†] *Paris Smaragdis*[‡] *Arne Leijon*[†]

[†] KTH Royal Institute of Technology
[‡]University of Illinois, Adobe Systems

ABSTRACT

We propose a method based on the probabilistic latent component analysis (PLCA) in which we use exponential distributions as priors to decrease the activity level of a given basis vector. A straightforward application of this method is when we try to extract a desired source from a mixture with low artifacts. For this purpose, we propose a maximum a posteriori (MAP) approach to identify the common basis vectors between two sources. A low-artifact estimate can now be obtained by using a constraint such that the common basis vectors in the interfering signal's dictionary tend to remain inactive. We discuss applications of this method in source separation with similar-gender speakers and in enhancing a speech signal that is contaminated with babble noise. Our simulations show that the proposed method not only reduces the artifacts but also increases the overall quality of the estimated signal.

***Index Terms*—** Source Separation, Nonnegative Matrix Factorization (NMF), PLCA, Dictionary Learning, Artifact Reduction

1. INTRODUCTION

A popular class of dictionary learning approaches is nonnegative matrix factorization (NMF) in which nonnegative dictionaries are learned from the magnitude or power spectrogram of the speech signals denoted by $\mathbf{X}$. This factorization is written as $\mathbf{X} \approx \mathbf{BV}$, where $\mathbf{B}$ and $\mathbf{V}$ are usually referred to as the basis and activation matrices. Since the basic NMF has many degrees of freedom, researchers have used different constraints to obtain more semantic factorizations with a better performance in a considered application [1, 2, 3, 4].

In the probabilistic formulations of NMF, some prior distributions are considered over the basis or the activation matrices. These prior distributions may be motivated, e.g., by the temporal dependencies of the audio signals. The goal of this prior information is to guide NMF by making some combination of the basis vectors more likely. For instance, to employ the time correlation of the audio signals, a constraint or a prior distribution is usually designed to govern the activations, e.g., [3]. Thus, having the result of the factorization at the current time instance, we put a prior over the activations for the next time instance that encourages the same pattern of activities as in the current time.

In this paper, we consider a source separation problem in which the underlying sources have some common basis vectors, i.e., some of the the basis vectors are shared between two sources. In practice this happens, e.g., when we try to separate speech signals from a mixture in which two speakers have the same gender, or when a speech signal is mixed with a multitalker babble noise [5]. Consequently, these problems are among the hardest ones in the NMF-based approaches. As a result of having a common set of basis vectors, NMF can not correctly separate the sources and depending on the initial conditions one of the sources will steal some parts of the other one. This will lead to artifacts in the separated signals. In a denoising problem, we may prefer to reduce the noise as far as it does not introduce artifacts in the speech. Similarly for a source separation problem, we may have a preference over one of the speakers and then our goal would be to separate one of the sources with low artifacts. One approach to achieve this is to learn the basis matrix of the interfering signal such that its similarity with the known basis matrix of the target signal is minimized [6].

Another solution to separate a desired source with low artifacts is that we discourage the activation of the common basis vectors in the basis matrix of the interfering source. By doing so, we let the basis vectors of the desired source to take over and explain the mixture signal. To the best of our knowledge, there is not any work in the NMF community that investigates this solution. In this paper, we consider probabilistic latent component analysis (PLCA) [7] and propose an algorithm that can be used for this purpose.

In this paper, we argue that the Dirichlet distribution is not suitable as a prior to estimate the nonnegative elements in PLCA, even though it is the conjugate distribution for this purpose. We instead propose to use an exponential distribution as the prior and show that it can be used to force some basis vectors to be *inactive*. Moreover, we derive a MAP approach to identify a set of the common basis vectors and use that to separate a desired source with arbitrarily low artifacts. We demonstrate the application of this method in a simple toy example and also in speech denoising and speech source separation for speakers with same and different genders. Our experiments show that the presented approach leads to a higher quality for the estimated signal by reducing the artifacts.

2. PROPOSED SOLUTION

In the following we first describe the basic PLCA approach. Then, we present our algorithm in which we use exponential distributions as priors for the activations. We discuss how this approach can be used to prevent (reduce) the activity of a given subset of the basis vectors. Additionally, we describe an approach to find a set of the common basis vectors between two underlying sources in Section 2.4. This information is then combined with the algorithm from Section 2.2 to design a source separation or speech enhancement algorithm in which we can recover a source with as low artifacts as desired.

2.1. PLCA: A Review

PLCA is a probabilistic nonnegative matrix factorization in which the speech magnitude spectrogram is modeled as a count data and

[*] This work was performed while at the University of Illinois.

is assumed to have a multinomial distribution:

$$\begin{aligned} \mathbf{x}_t &\sim \text{Mult}\left(\boldsymbol{\theta}_t\right), \\ \theta_{ft} &= p_t\left(f\right) = \sum_{z=1}^{I} p\left(f \mid z\right) p_t\left(z\right), \end{aligned}$$

where $\mathbf{x}_t$ is the vector of the DFT magnitudes at time frame t, f is the frequency index, θ_{ft} is the f-th element of $\boldsymbol{\theta}_t$, and z is the hidden variable that can take an integer value from $\{1 \ldots I\}$. An NMF approximation of $\mathbf{x}_t$ can be obtained as the expected value of its distribution:

$$\mathbf{x}_t \approx \hat{\mathbf{x}}_t = g_t \boldsymbol{\theta}_t, \tag{1}$$

where $g_t = \sum_f x_{ft}$. The set of the I probability vectors $p\left(f \,|z\right)$ are the basis vectors and can be found using an expectation-maximization approach. In the E step of the algorithm, the posterior probabilities of the hidden variables (z) are computed as:

$$p_t(z \mid f) = \frac{p(f \mid z) p_t(z)}{\sum_{z'=1}^{I} p(f \mid z') p_t(z')}. \tag{2}$$

In the M step of the algorithm the basis vectors and the weights are updated as:

$$p_t\left(z\right) = \frac{\sum_f x_{ft} p_t\left(z \mid f\right)}{\sum_{f,z'} x_{ft} p_t\left(z' \mid f\right)}, \tag{3}$$

$$p(f \mid z) = \frac{\sum_t x_{ft} p_t\left(z \mid f\right)}{\sum_{f',t} x_{f't} p_t\left(z \mid f'\right)}. \tag{4}$$

2.2. PLCA with Exponential Priors

We can impose constraints on PLCA to use our a-priori knowledge. In this paper, we focus on prior distributions over the activations. Since Dirichlet distribution is the conjugate prior of the multinomial, we first give the update rules for this case. Let $\boldsymbol{\beta}_t$ be an I-dimensional vector with elements $\beta_{zt} = p_t(z)$. The Dirichlet prior for $\boldsymbol{\beta}_t$ is given as:

$$p\left(\boldsymbol{\beta}_t\right) \propto \prod_{z=1}^{I} p_t\left(z\right)^{\alpha_z - 1},$$

where $\alpha_z > 0$, $z \in \{1 \ldots I\}$ are the parameters of the Dirichlet distribution. The E step of the algorithm (2) and the update rule of the basis vectors (4) remain the same as before. The update of the weights however changes to:

$$p_t\left(z\right) = \frac{\sum_f x_{ft} p_t\left(z \mid f\right) + \alpha_z - 1}{\lambda_t}, \tag{5}$$

where λ_t is a Lagrange multiplier and is used to ensure that $p_t\left(z\right)$ is a probability vector. Computing λ_t is trivial in this case and is given as the sum of the numerator of (5) over z.

The problem of the Dirichlet prior is that it does not naturally fit to the estimation of the nonnegative elements $p_t\left(z\right)$ and can lead to a negative value in the right hand side of (5). One way to avoid this problem is to put a threshold on (5) such that its minimum value is limited to be a very small positive number. Here, we propose to use an exponential distribution as the prior that does not suffer from this problem, and at the same time provides a single parameter to control the activity of each basis vector individually. The form of this prior is given by:

$$p\left(\boldsymbol{\beta}_t\right) \propto \prod_{z=1}^{I} e^{-p_t(z)/\alpha_z}, \tag{6}$$

where $\boldsymbol{\alpha} = \{\alpha_z\}$ is the vector of scale or inverse rate parameters. The update rule of the activations can be obtained by using the EM algorithm in which the M step is given by:

$$p_t\left(z\right) = \frac{\sum_t x_{ft} p_t\left(z \mid f\right)}{\lambda_t + 1/\alpha_z}. \tag{7}$$

Computation of the Lagrange multiplier λ_t is not trivial in (7) because the denominator is not the same for different latent components z as it was in (5). However, since λ_t is a scalar variable we can use a simple iterative algorithm, e.g., Newton's method, to find its optimal value.

In contrast to (5), (7) leads to nonnegative estimates for the activations for any value of the hyperparameters $\boldsymbol{\alpha}$. Now consider a simple problem where we are interested to force a certain basis vector to be inactive. To do this using the Dirichlet priors, we have to boost the activity of all the other basis vectors since to ensure nonnegativity we should avoid choosing $\alpha_z < 1$. Using the exponential priors, we only need to use a small hyperparameter in the prior distribution corresponding to the given basis vector. In this case α_z can approach to 0 without making any theoretical problem.

2.3. Example: Separation of Sources with One Common Basis Vector

We consider a source separation problem to illustrate how the exponential distribution can be used to avoid certain type of activations. In this toy example, we generate 3-d nonnegative data for two sources. Let $\mathbf{e}_i$ denote a 3-d indicator vector whose i-th element is 1 and the rest of its elements are zero. We considered two basis vectors per each source from which one is shared between the sources: $\mathbf{B}^{(1)} = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 \end{bmatrix}$ and $\mathbf{B}^{(2)} = \begin{bmatrix} \mathbf{e}_2 & \mathbf{e}_3 \end{bmatrix}$. We also added small nonnegative random noise to the basis matrices. Data was generated by multiplying the bases by an activation vector with elements sampled from a uniform distribution in the interval $[0, 1]$. For our example, this procedure yielded to:

$$\mathbf{x}^{(1)} = \begin{bmatrix} 0.30 \\ 0.63 \\ 0.07 \end{bmatrix}, \mathbf{x}^{(2)} = \begin{bmatrix} 0.03 \\ 0.32 \\ 0.65 \end{bmatrix}.$$

These vectors together with the mixture

$$\mathbf{x} = \mathbf{x}^{(1)} + \mathbf{x}^{(2)} = \begin{bmatrix} 0.33 \\ 0.95 \\ 0.72 \end{bmatrix},$$

are shown in Figure 1. Note that to be on the 2-d simplex, all of these vectors are normalized to sum to one. Applying our proposed method with $\boldsymbol{\alpha} = [1\ \ 1\ \ 0.5\ \ 1]$[1] leads to the estimates which are shown in the figure. Numerically, we got:

$$\hat{\mathbf{x}}^{(1)} = \begin{bmatrix} 0.3 \\ 0.9 \\ 0.08 \end{bmatrix}, \hat{\mathbf{x}}^{(2)} = \begin{bmatrix} 0.03 \\ 0.05 \\ 0.64 \end{bmatrix}.$$

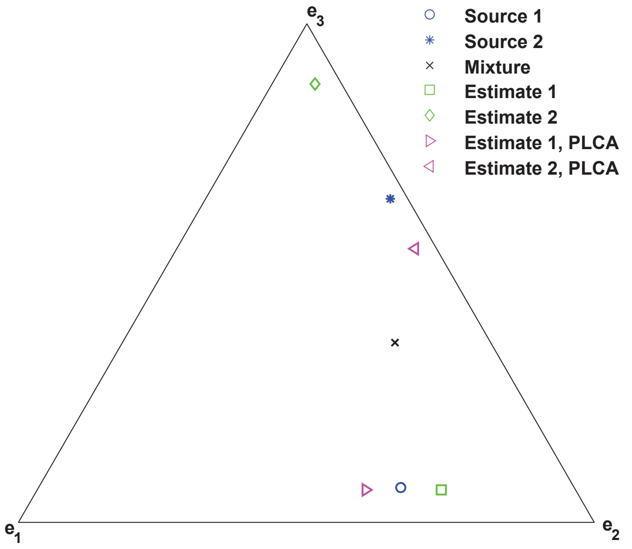

Figure 1: This example shows the original sources, the mixture and the estimated sources on a 2-d simplex. To be on the simplex, all the vectors are normalized to sum to one. $\mathbf{e}_i$ is a 3-d indicator vector whose i-th element is 1. Each source has two basis vectors from which $\mathbf{e}_2$ is shared between them. Since the shared basis vector will introduce artifacts in the estimate of the desired source (estimate 1), a prior is constructed such that a big portion of the mixture's second element is taken as the corresponding component in "Estimate 1".

As we can see in Figure 1 also, first source has taken over and the second dimension of its estimate (0.93) is very close to the corresponding element in the mixture (0.95). If we use PLCA here (with the same initial value for the activation of the similar bases), the second dimension will be divided almost equally between two estimates (see Figure 1), which means that we lose some part of the desired source.

2.4. Identifying Common Bases

We need to know which basis vectors are shared between sources to use the algorithm given in Section 2.2. In the following, we describe a maximum a-posteriori (MAP) approach to get this information. Let us assume that we have trained I_1 and I_2 basis vectors for the desired source (source 1) and the interfering source (source 2), respectively, using some training data. Our goal in this section is to develop an approach to identify a subset of basis vectors that belong to the interfering source and can also explain the desired source with a given accuracy. This subset can be actually seen as the common set of bases between two sources. By having this information, we can construct the vector $\boldsymbol{\alpha}$ to automatically prevent the activity of this subset. This will reduce artifacts in the estimate of the desired source.

We start by concatenating small development sets of both of the sources (clean signals) as: $\mathbf{X} = [\mathbf{X}^{(1)}\ \mathbf{X}^{(2)}]$, where we have T_1 and T_2 observations (columns) in $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, respectively. Also, we concatenate the basis vectors of the two sources to obtain a larger basis matrix to explain both of the sources. We now apply PLCA to the concatenated signal $\mathbf{X}$ with the basis matrix being fixed. The probability of choosing a basis vector, given the source, can be written as:

$$\begin{aligned} p\left(z \mid s^{(j)}\right) &= \sum_{t=1}^{T_1+T_2} p\left(z, t \mid s^{(j)}\right) \\ &= \sum_{t=1}^{T_1+T_2} p\left(z \mid t, s^{(j)}\right) \frac{p\left(s^{(j)} \mid t\right) p\left(t\right)}{p\left(s^{(j)}\right)} \\ &= \frac{1}{T_j} \sum_{t \in T_{s^{(j)}}} p_t\left(z\right), z \in \left\{1 \dots I_1 + I_2\right\}, \end{aligned}$$

where $T_{s^{(j)}}$ includes the indices of the observations from $s^{(j)}$ and for these observations we have: $p(s^{(j)} \mid t) = 1$. To get the last line we have used a uniform distribution for t as $p(t) = 1/(T_1 + T_2)$, and also we have $p(s^{(j)}) = T_j/(T_1 + T_2)$.

To get a MAP classifier, we can use the Bayes' theorem (with a flat prior over the sources) to obtain the probability of each source given a basis vector. For $j = 1$, this results to:

$$p\left(s^{(1)} \mid z\right) = \frac{\sum_{t \in T_{s^{(1)}}} p_t\left(z\right)/T_1}{\sum_{t \in T_{s^{(1)}}} p_t\left(z\right)/T_1 + \sum_{t \in T_{s^{(2)}}} p_t\left(z\right)/T_2}.$$

To identify the basis vectors from the dictionary of $s^{(2)}$ that can also explain $s^{(1)}$, we now compare $p(s^{(1)} \mid z)$, $z \in \{I_1 + 1 \dots I_1 + I_2\}$ with a given threshold $0 \leq \gamma \leq 1$. If $p(s^{(1)} \mid z)$ was larger than γ, it means that this basis vector can also explain $s^{(1)}$ good enough. We should avoid the activity of this basis vector in a given mixture so that its similar basis vector that belongs to $s^{(1)}$ takes over and explains the mixture. $\gamma = 1$ recovers the basic PLCA, and $\gamma = 0$ corresponds to outputting the mixture signal as the estimate of the desired source. Thus, $\gamma = 0$ will neither suppress the interfering signal nor will introduce any artifacts in the final estimate.

3. EXPERIMENTS USING SPEECH DATA

We consider two problems to demonstrate the application of the proposed algorithm. In our experiments with the source separation and noise reduction, we used speech signals from the TIMIT and babble noise from the NOISEX-92 databases. Here, we considered instantaneous mixtures of sources where the mixed signal is obtained by adding the speech and noise waveforms, or by adding the speech signals of two speakers. All the signals were down-sampled to 16 kHz. The discrete Fourier transform (DFT) with a frame length of 64 ms, 50% overlap, and a Hann window was used in our simulations.

3.1. Source Separation

We learned 30 basis vectors using a set of training sentences for each speaker. The results presented here are averaged over 40 pairs of randomly-selected speakers. We aim to separate the speech signal from the first source with low artifacts. Hence, we first find a subset of the basis vectors belonging to the second source that can also explain the first source, and then we set the corresponding elements in $\boldsymbol{\alpha}$ to 0.5. All the other elements are given a value of 1.

We study the performance of the algorithm for the same gender (male-male or female-female) and different gender scenarios. The

[1] Each element in $\boldsymbol{\alpha}$ reflects our preference of having this basis vector active. If we set an element to a value smaller than the average of $\boldsymbol{\alpha}$, that basis vector is encouraged to be inactive. The choice of 0.5 was arbitrary in this example.

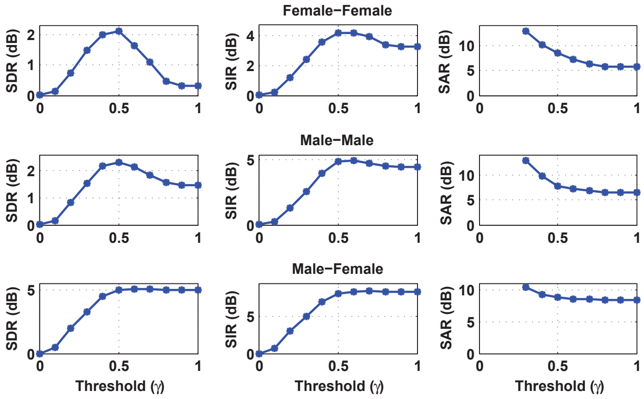

Figure 2: Results of source separation for female-female (top panel), male-male (middle panel), male-female (bottom panel) speakers. $\gamma = 1$ corresponds to the basic PLCA. Results show that by reducing the threshold a lower-artifact estimate is obtained for the desired source.

performance of the separation is measured using the source to distortion ratio (SDR), source to interference ratio (SIR) and source to artifact ratio (SAR) [8]. The results are shown in Figure 2. A high value of SAR corresponds to a low-artifact estimate.

Our simulations show that by reducing the threshold (γ) we get a higher SAR in all scenarios. In fact, the lower we set the threshold, the more number of basis vectors (from the interfering source) are recognized as the common bases and we put a prior that motivates these basis vectors to remain inactive. As a result, the corresponding parts of the mixture signal are taken in the favor of the desired source and we lose less and less of the desired signal.

Figure 2 shows that the performance is maximized in terms of SDR and SIR at a threshold equal to 0.5. For SIR, by increasing threshold we first get a higher suppression of the interfering signal. But when we set a very high value for the threshold, we get lower suppression. This might be explained by noting that with a proper value of γ, some shared basis vectors (of the interference signal) are inactive while the other basis vectors get higher activations, which results in a stronger suppression. Considering SDR, we again see that we get the best quality for $\gamma = 0.5$.

Another interesting result that can be seen in Figure 2 is that for the Male-Female configuration we do not get any improvement in SDR by using our algorithm. However, we can recover the desired source with lower artifacts. This is intuitive since we do not expect many common basis vectors in this case.

3.2. Reducing Babble Noise

As our second experiment, we consider a noise reduction problem where a speech signal is degraded by a babble noise. As discussed earlier, we expect the two sources to share some basis vectors. So we apply our method to reconstruct the speech signal with low artifacts and better quality. Here, we learned 30 and 50 basis vectors for the speech and babble signals, respectively. The results are averaged over 40 speech signals from different speakers and are shown in Figure 3.

The experimental results are similar to the ones in Figure 2. Again, we see that the SAR value is reducing as a monotonic function of the threshold γ while SIR and SDR exhibit a maximum around $\gamma = 0.3 \sim 0.5$. Our informal listening tests were consistent with these results.

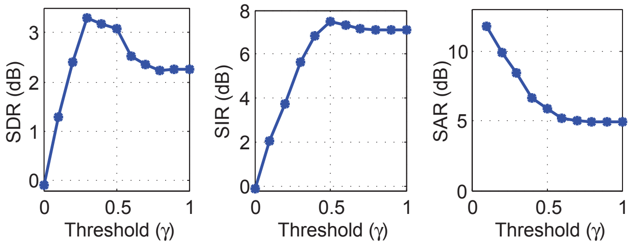

Figure 3: Speech denoising results for the babble noise. Similar to Figure 2, the lower the threshold is, the lower the artifacts are. A trade-off between noise suppression and artifact absence is obtained for $\gamma = 0.3$ where the SDR is maximum. Input SNR is 0 dB.

4. CONCLUSIONS

In this paper, we discussed a source separation problem in which the sources share some basis vectors. We proposed a PLCA-based approach to extract a desired source with an arbitrarily low artifacts. This was achieved by keeping the common basis vectors from the interfering source's dictionary inactive. We developed a MAP approach to automatically detect the similar basis vectors. We considered applications of the proposed method in speech source separation and noise reduction. Our simulations show that when the underlying speakers have the same gender or the speech is contaminated with babble noise (for which we expect to see a sufficient number of common basis vectors) the proposed method can be used to reduce the artifacts and increase the quality in the estimate of the desired source.

5. REFERENCES

[1] M. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete latent variable decomposition of counts data," in *Advances in Neural Information Process. Systems (NIPS)*. MIT Press, 2007, pp. 1313–1320.

[2] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 1066–1074, 2007.

[3] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using NMF," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 10, pp. 2140–2151, oct. 2013.

[4] A. Cichocki, R. Zdunek, and S. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 5, may 2006.

[5] N. Mohammadiha and A. Leijon, "Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 998–1011, may 2013.

[6] K. Yagi, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Music signal separation by orthogonality and maximum-distance constrained nonnegative matrix factorization with target signal information," in *Proc. Audio Engineering Society Int. Conf.*, mar. 2012.

[7] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Applications of Signal Process. Audio Acoust. (WASPAA)*, oct. 2003, pp. 177–180.

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