Phase-Aware Non-Negative Spectrogram Factorization

R. Mitchell Parry and Irfan Essa

Georgia Institute of Technology
College of Computing / GVU Center
85 Fifth Street NW, Atlanta, GA USA
{parry,irfan}@cc.gatech.edu
http://www.cc.gatech.edu/

Abstract. Non-negative spectrogram factorization has been proposed for single-channel source separation tasks. These methods operate on the magnitude or power spectrogram of the input mixture and estimate the magnitude or power spectrogram of source components. The usual assumption is that the mixture spectrogram is well approximated by the sum of source components. However, this relationship additionally depends on the unknown phase of the sources. Using a probabilistic representation of phase, we derive a cost function that incorporates this uncertainty. We compare this cost function against four standard approaches for a variety of spectrogram sizes, numbers of components, and component distributions. This phase-aware cost function reduces the estimation error but is more affected by detection errors.

Key words: audio processing, source separation, sparse representations, time-frequency representations, unsupervised learning

1 Introduction

Non-negative spectrogram factorization (NSF) has been proposed for single-channel source separation [1–3], music transcription [4, 5], and speech recognition [6]. The input mixture is first transformed into a time-frequency representation such as the short-time Fourier transform (STFT). Because of phase-invariant aspects of human hearing the phase information in the STFT is removed yielding the absolute value or absolute square of the STFT (*i.e.,* magnitude or power spectrogram) [2]. The resulting spectrogram matrix is then factored into the sum of rank-one component spectrograms using independent component analysis (ICA) or non-negative matrix factorization (NMF). Each component comprises a static spectral shape and time-varying amplitude envelope. Ideally, each component contains information unique to a particular source for separation or a particular event for transcription. We focus on this basic approach although various other algorithms incorporate sparseness, convolution, or multiple channels [4, 7, 8].

NSF methods commonly assume that the mixture magnitude or power spectrogram is well approximated by the sum of source components. ICA forces this

relationship while maximizing the independence of the spectral components [1], whereas NMF minimizes a cost function between the mixture spectrogram and the sum of spectral components [9]. However, because of the nonlinearity of the absolute value function a mixture spectrogram is not the sum of the component spectrograms. Instead, the mixture spectrogram depends on the component spectrograms and their phases. We derive a cost function suitable for NSF by treating the phase as a uniform random variable and maximizing the likelihood of the mixture spectrogram. In previous work, we derived the explicit likelihood function for the case of two components [10]. In this paper, we extend this result to the case of more than two components and show that it is analogous to the multiplicative noise model employed by Abdallah and Plumbley [4]. Even though this cost function is specifically tailored to non-negative spectrograms, the Euclidean distance or generalized Kullback-Leibler divergence is more commonly used for NSF. We compare each cost function based on its ability to estimate the component spectrograms for a variety of spectrogram sizes, numbers of components, and component distributions.

2 Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) was first proposed for the decomposition of images [11]. Image data is inherently non-negative and a single image can be regarded as a linear combination of underlying image parts. NMF estimates these components by minimizing the distance between a set of mixture images contained in the columns of a matrix, A, and the sum of the component matrices, B. The two common distance functions are the Euclidean distance:

$$\|A - B\|^2 = \sum_{ij} (A_{ij} - B_{ij})^2 \tag{1}$$

and a generalized version of the Kullback-Leibler divergence:

$$D(A\|B) = \sum_{ij} \left(A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij} \right). \tag{2}$$

When applied to non-negative spectrograms, A represents the mixture spectrogram and B represents the sum of component spectrograms. Instead of decomposing multiple images, spectrogram factorization decomposes multiple spectral frames contained in the columns of A. Although magnitude or power spectrograms are non-negative they are *not* a linear combination of underlying component spectrograms because of the nonlinearity of the absolute value function used to generate them.

3 Non-negative Spectrograms

A popular way to transform an audio signal into a series of image-like representations is to extract its frequency spectrum at multiple time-points. We consider

the case of one mixture signal and model it as the sum of R source component signals:

$$x(t) = \sum_{r=1}^{R} s_r(t) \; . \tag{3}$$

The short-time Fourier transform (STFT) is a linear transformation into the frequency domain that preserves this relationship:

$$\mathcal{F}_x(k,t) = \sum_{r=1}^{R} \mathcal{F}_{s_r}(k,t) \; . \tag{4}$$

The magnitude spectrogram is the absolute value of the complex-valued STFT:

$$X_{kt} = |\mathcal{F}_x(k,t)| \qquad\qquad [S_r]_{kt} = |\mathcal{F}_{s_r}(k,t)| \; . \tag{5}$$

The original STFT contains additional phase information:

$$\mathcal{F}_x(k,t) = X_{kt}(\cos\Theta_{kt} + i\sin\Theta_{kt}) = \sum_r [S_r]_{kt}(\cos[\Theta_r]_{kt} + i\sin[\Theta_r]_{kt}) \; . \tag{6}$$

When applied to non-negative spectrograms, ICA and NMF estimate rank-one component spectrograms. The columns of a $K \times R$ matrix W specify the spectral shapes and the rows of an $R \times T$ matrix H represent the amplitude envelopes of all the component spectrograms:

$$[S_r]_{kt} = W_{kr}H_{rt} \; . \tag{7}$$

The various algorithms for NSF vary in the way that they estimate W and H.

4 Non-negative Spectrogram Factorization

The vast majority of NSF methods treat each column of a magnitude or power spectrogram matrix as though it were an image and use ICA or NMF to estimate the components. To our knowledge, there has been only one cost function specifically designed for non-negative spectrograms, namely that of Abdallah and Plumbley [4]. They derive a divergence function based on a multiplicative noise model for estimating the variance (*i.e.,* power) at each time-frequency bin. In this paper, we define the mixture magnitude spectrogram in terms of the component magnitude spectrograms and their phases. Using a probabilistic representation of the phase, we derive an analogous divergence function.

Both ICA- and NMF-based techniques implicitly assume that the mixture non-negative spectrogram, X, is well approximated by the sum of the spectral components, S_r. However, by incorporating the phase of the components, Θ_r, we make this relationship precise:

$$X_{kt} = \sqrt{\sum_{qr} [S_q]_{kt}[S_r]_{kt} \cos\left([\Theta_q]_{kt} - [\Theta_r]_{kt}\right)} \; . \tag{8}$$

The mixture magnitude spectrogram does not equal the sum of component magnitude spectrograms unless at most *one* component is active at a time or all active components have the *same* phase.

5 Probabilistic Representation of Phase

Given the mixture spectrogram's dependence on the phase in Equation 8, we represent the phase as a uniform random variable. We also make the simplifying assumption that the phase is independent at different time-frequency points. To some degree, this is true. However, the unwrapped phase of a steady state signal can be approximated from the previous two time-steps [12]. Although this violates the independence assumption, we have found that the resulting approach works well in practice.

We wish to maximize the likelihood of the mixture magnitude spectrogram as a function of the source component magnitude spectrograms. For the case of two components, Equation 8 is a function of one random variable (*i.e.,* $\Theta_d = \Theta_1 - \Theta_2$) and it is relatively straightforward to derive $p(X|S_1, S_2)$ directly [10]. However, for more components it becomes increasingly difficult to derive the precise likelihood function. Instead, we estimate the likelihood using the central limit theorem to capture the shape of the distribution for a large number of components.

The probability density function for a complex random variable with magnitude S_r and uniform random phase has a mean of zero and a variance of S_r^2. According to the Lindeberg-Feller central limit theorem [13], the sum of many such variables tends toward a complex Gaussian with zero mean and a variance of $\sum_r S_r^2$. This theorem is valid under the Lindeberg condition, which states that the component variances, S_r^2, are small relative to their sum [13]. Applied to magnitude spectrograms we have the following:

$$p(\mathcal{F}_x|S_1, \ldots, S_R) = \prod_{kt} \frac{1}{\pi \Lambda_{kt}} \exp\left(-\frac{X_{kt}^2}{\Lambda_{kt}}\right), \tag{9}$$

where $\Lambda_{kt} = \sum_r [S_r^2]_{kt}$. We find the likelihood of X by integrating with respect to phase, resulting in a Rayleigh distribution:

$$p(X|S_1, \ldots, S_R) = \prod_{kt} \frac{2X_{kt}}{\Lambda_{kt}} \exp\left(-\frac{X_{kt}^2}{\Lambda_{kt}}\right). \tag{10}$$

6 Maximum Likelihood

In order to estimate S_r, we propose minimizing the negative log likelihood of X:

$$-\log p(X|S_1, \ldots, S_R) = -\sum_{kt} \left[\log\left(\frac{2X_{kt}}{\Lambda_{kt}}\right) - \frac{X_{kt}^2}{\Lambda_{kt}}\right]. \tag{11}$$

For comparison, we frame our maximum likelihood approach in terms of a divergence function. The minimum of Equation 11 is $1 - \log(2/X_{kt})$ at $\Lambda_{kt} = X_{kt}^2$. By subtracting this value we find a divergence function that is non-negative reaching zero only when all $\Lambda_{kt} = X_{kt}^2$:

$$D_s = D(1\|X^2/\Lambda) = \sum_{kt} \frac{X_{kt}^2}{\Lambda_{kt}} - 1 + \log\left(\frac{\Lambda_{kt}}{X_{kt}^2}\right), \tag{12}$$

which is equivalent to Equation 8 in Abdallah and Plumbley [4]. We derive the gradient for D_s with respect to W_{kr}^2 and H_{rt}^2:

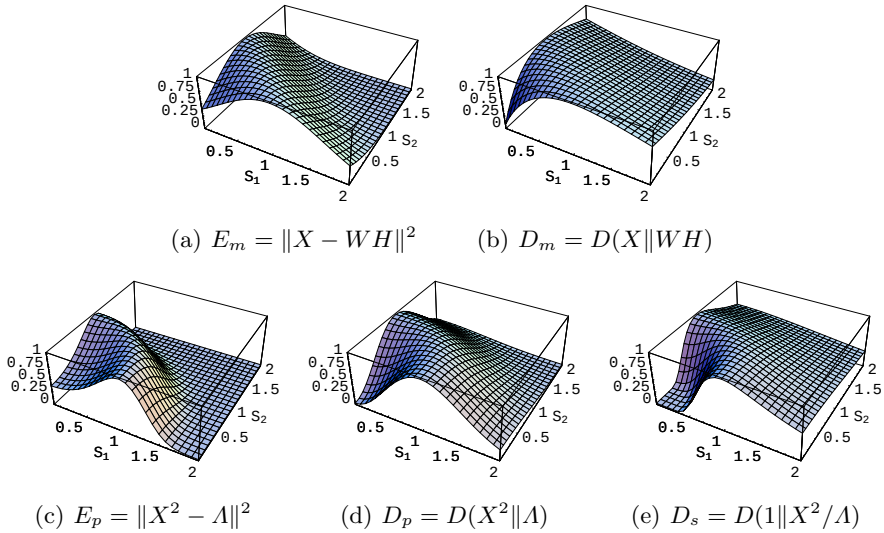
$$\frac{\partial D_s}{\partial (W_{kr}^2)} = \sum_t H_{rt}^2 \left(\frac{\Lambda_{kt} - X_{kt}^2}{\Lambda_{kt}^2} \right) \qquad \frac{\partial D_s}{\partial (H_{rt}^2)} = \sum_k W_{kr}^2 \left(\frac{\Lambda_{kt} - X_{kt}^2}{\Lambda_{kt}^2} \right), \quad (13)$$

where $\Lambda_{kt} = \sum_r W_{kr}^2 H_{rt}^2$. Although D_s is not convex with respect to W_{kr}^2 or H_{rt}^2, we find local minima using the following multiplicative update rules:

$$W_{kr}^2 \leftarrow W_{kr}^2 \frac{\sum_t H_{rt}^2 X_{kt}^2 / \Lambda_{kt}^2}{\sum_t H_{rt}^2 / \Lambda_{kt}} \qquad H_{rt}^2 \leftarrow H_{rt}^2 \frac{\sum_k W_{kr}^2 X_{kt}^2 / \Lambda_{kt}^2}{\sum_k W_{kr}^2 / \Lambda_{kt}} . \quad (14)$$

7 Results

We compare the phase-aware cost function, D_s, to four other cost functions based on Euclidean or Kullback-Leibler divergence for magnitude or power spectrograms. Figure 1 plots the shape of the likelihood functions for each of the cost functions with $X = 1$. Magnitude spectrogram methods (E_m and D_m) reach a maximum on the line $S_1 + S_2 = X$. Power spectrogram methods (E_p, D_p, and D_s) reach a maximum on the circle $S_1^2 + S_2^2 = X^2$. When $X = 1$, the sum of S_1 and S_2 must be greater than one. D_s encourages this result by penalizing solutions near the origin more than the other cost functions.



(a) $E_m = \|X - WH\|^2$ (b) $D_m = D(X\|WH)$

(c) $E_p = \|X^2 - \Lambda\|^2$ (d) $D_p = D(X^2\|\Lambda)$ (e) $D_s = D(1\|X^2/\Lambda)$

Fig. 1. The shape of the likelihood functions derived from the 5 labeled cost functions for the case of two components and $X = 1$.

In our experiment, we evaluate the performance of the cost functions for a variety of spectrogram sizes, numbers of components, and component distributions. Specifically, we construct square spectrograms and vary their size with $K = T \in [32, 64, 128, 256, 512, 1024]$, $R \in [1, \ldots, 30]$, and W and H drawn from the uniform, positive normal, or exponential distribution. After drawing W and H from the specified distribution, we construct X using Equations 5–7 with uniformly distributed random phase, Θ_r. We then estimate W and H using each cost function with multiplicative update rules derived in Section 6 or by Lee and Seung [9]. Because scaling W by α and H by $1/\alpha$ produces the same cost, we normalize the rows of H to unit L_2 norm after every update.

We evaluate each cost function according to the mean square error between the original and estimated $\{S_r\}$. Because the factorization technique is permutation invariant, we must determine the mapping between each estimated and original S_r. For this purpose, we use a greedy algorithm that matches the two most similar components (one original and one estimated) and then removes them from consideration. The process repeats until the mapping is complete.

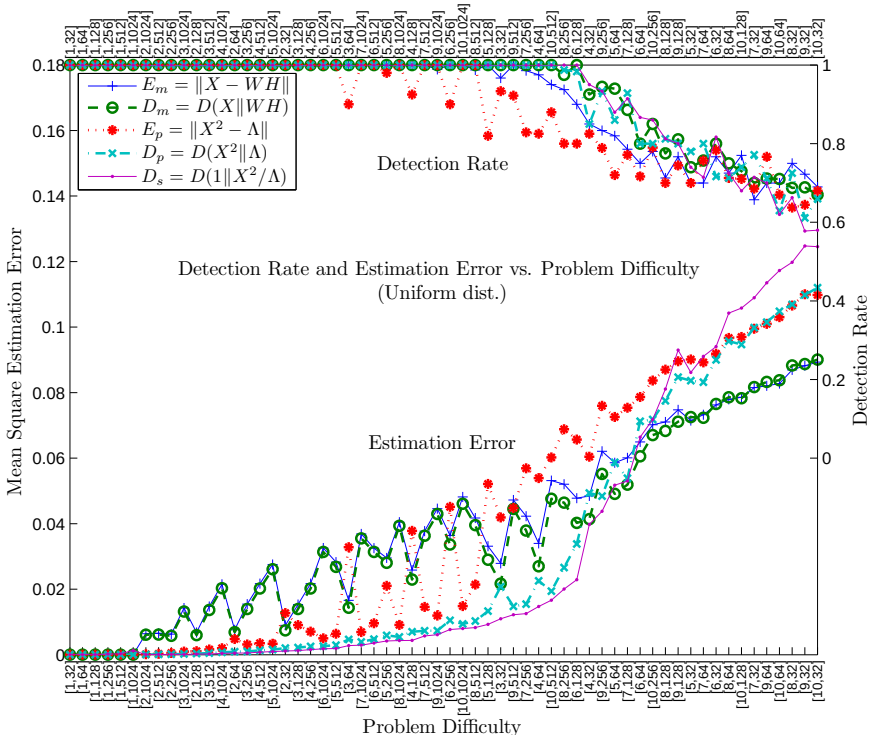


Fig. 2. Estimation error and detection rate for the five cost functions.

Figure 2 plots the average performance over ten trials for each configuration of parameters. For space considerations, we only show $R \in [1, \ldots, 10]$ and W and H drawn from the uniform distribution. Each of the 60 $[R, K]$ pairs are sorted along the x-axis in order of increasing minimum error among the five cost functions. Clearly, the problem becomes more difficult as R increases or as K decreases.

The bottom of Figure 2 plots the mean square estimation error. For simpler versions of the problem, D_s outperforms the rest. However, toward the right of the plot the performance becomes markedly worse and E_m and D_m perform better. This inversion of performance is linked to the detection rate.

The top of Figure 2 plots the detection rate. When each estimated component uniquely matches a real component, the detection rate is 100%. However, when none of the estimated components match one of the real components, that component is not detected. We compute the detection rate as the fraction of real components that are the closest match (in the mean square sense) for at least one estimated component. At $[R, K] = [4, 32]$, the detection rate for D_s drops below 100% for the first time and this corresponds to the first large increase in estimation error. After that, the estimation rate for D_s accelerates until it is the worst of the group. We speculate that if 100% detection could be maintained, D_s would continue to outperform the others.

The underlying distribution of W and H also affects estimation and detection. As presented, the cost functions implicity assume a uniform prior distribution on W and H in the maximum likelihood framework. Therefore, as the component distributions diverge from the uniform distribution (*e.g.,* become more sparse) the maximum likelihood approach becomes less realistic. The aggregated mean square error for the uniform, positive normal (more sparse), and exponential (most sparse) distribution is 0.036, 0.19, and 0.44, respectively. However, sparseness has the opposite effect on detection. All of the cost functions attain 100% detection for more problems as sparseness increases. Table 1 lists the number of problems that resulted in 100% detection and the number of times each algorithm provides the best estimation error for each of the distributions and R between 2 and 10.

Table 1. Summary of detection rate and lowest estimation error for $R = [2, 10]$.

Distribution:	Uniform		Positive Normal		Exponential	
Cost func.	100% det.	Best est.	100% det.	Best est.	100% det.	Best est.
E_m	27	9	37	3	44	0
D_m	34	8	43	6	47	6
E_p	23	0	29	0	30	0
D_p	33	0	38	4	41	3
D_s	35	37	40	41	42	45
Total	152	54	187	54	204	54

8 Conclusion

We present a new derivation of a divergence function, D_s, specifically tuned to non-negative spectrogram factorization. We compare its performance against four standard approaches for a variety of spectrogram sizes, numbers of components, and sparseness. We show that this divergence improves the estimation of the source components. However, it is more affected by detection error. Algorithms aimed at improving detection rates (*e.g.,* a prior distribution on W and H) are likely to improve D_s.

References

1. Casey, M., Westner, W.: Separation of mixed audio sources by independent subspace analysis. In: Proc. of the Int'l Computer Music Conf. (2000)
2. Smaragdis, P.: Redundancy Reduction for Computational Audition, a Unifying Approach. PhD thesis, MAS Dept., Massachusetts Institute of Technology (2001)
3. Wang, B., Plumbley, M.D.: Investigating single-channel audio source separation methods based on non-negative matrix factorization. In: ICA Research Network Int'l Workshop. (2006) 17–20
4. Abdallah, S.A., Plumbley, M.D.: Polyphonic transcription by non-negative sparse coding of power spectra. In: Proc. of the Int'l Conf. on Music Information Retrieval. (2004) 318–325
5. FitzGerald, D., Coyle, E., Laylor, B.: Sub-band independent subspace analysis for drum transcription. In: Proc. of Int'l Conf. on Digital Audio Effects. (2002) 65–69
6. Raj, B., Singh, R., Smaragdis, P.: Recognizing speech from simultaneous speakers. In: Eurospeech. (2005)
7. Virtanen, T.: Separation of sound sources by convolutive sparse coding. In: ISCA Tutorial & Research Wkshp on Statistical & Perceptual Audio Processing. (2004)
8. FitzGerald, D., Cranitch, M., Coyle, E.: Sound source separation using shifted non-negative tensor factorisation. In: Proc. of the IEEE ICASSP. (2006)
9. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in NIPS 13. MIT Press (2001) 556–562
10. Parry, R.M., Essa, I.: Incorporating phase information for source separation via spectrogram factorization. In: Proc. of IEEE ICASSP. (2007)
11. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401** (1999) 788–791
12. Bello, J.P., Sandler, M.B.: Phase-based note onset detection for music signals. In: Proc. of the IEEE ICASSP. Volume 5. (2003) 441–444
13. Feller, W.: An Introduction to Probability Theory and Its Applications. New York: Wiley (1971)