INCORPORATING PHASE INFORMATION FOR SOURCE SEPARATION VIA SPECTROGRAM FACTORIZATION

R. Mitchell Parry and Irfan Essa

Georgia Institute of Technology
College of Computing / GVU Center
Atlanta, Georgia 30332-0760, USA

ABSTRACT

Spectrogram factorization methods have been proposed for single channel source separation and audio analysis. Typically, the mixture signal is first converted into a time-frequency representation such as the short-time Fourier transform (STFT). The phase information is thrown away and this spectrogram matrix is then factored into the sum of rank-one source spectrograms. This approach incorrectly assumes the mixture spectrogram is the sum of the source spectrograms. In fact, the mixture spectrogram depends on the phase of the source STFTs. We investigate the consequences of this common assumption and introduce an approach that leverages a probabilistic representation of phase to improve the separation results.

Index Terms— non-negative matrix factorization, source separation, audio

1. INTRODUCTION

Spectrogram factorization methods have been proposed for single channel source separation [1–4] and audio analysis [5–8]. These methods estimate the magnitude of each source spectrogram using the magnitude of the mixture spectrogram. The incorrect implicit assumption common to all these methods is that the magnitude of the mixture spectrogram is the sum of the magnitudes of the source spectrograms. In fact, the mixture spectrogram depends on the magnitude *and* phase of the source spectrograms. This paper investigates the role of phase in determining the mixture spectrogram and incorporates a probabilistic representation of phase into a novel method for source spectrogram estimation.

When multiple mixture signals are available, independent component analysis [9] (ICA) is a statistical technique that separates as many independent source signals as mixture signals. When there is only one mixture signal, the signal may be transformed into a time-frequency representation such as the magnitude of the short-time Fourier transform (*i.e.,* magnitude spectrogram). Casey and Westner [1] originated the idea of spectrogram factorization by applying ICA to the magnitude of the mixture spectrogram, treating each frequency channel as a separate mixture signal. Using this approach, ICA separates as many sources as frequency channels. However, the expressiveness of each source is necessarily diminished. Each source magnitude spectrogram is a rank-one matrix formed by the product of a column vector containing the spectral shape and a row vector containing the time-varying gain. The actual source spectra are deemed to be a combination of multiple rank-one source spectrograms.

The problem with ICA for spectrogram factorization is that it extracts components that have negative elements, whereas magnitude spectrogram data is always non-negative. Therefore, non-negative matrix factorization [11] (NMF) has been proposed for source spectrogram estimation. NMF does not require independence but maintains non-negative elements.

An underlying assumption of ICA- and NMF-based approaches is that the mixture magnitude spectrogram is the sum of the source magnitude spectrograms. This assumption is valid only in the unlikely event that all sources have the *same* phase at *every* time-frequency point or in the trivial case when only one source is active. In all other cases, the mixture spectrogram necessarily depends on the phase information in the short-time Fourier transform (STFT) of the sources. We present a method to incorporate the unknown source phase information into the estimation of the source magnitude spectrograms using a probabilistic representation of phase.

2. RELATED WORK

Source separation via spectrogram factorization suffers from several problems. Complex sources must be represented as the combination of several rank-one source spectrograms. Deciding how to cluster the rank-one spectrograms to form complex source spectra is a difficult problem. Casey and Westner [1] propose clustering via spectral similarity. Once a suitable spectrogram for each source has been formed, it is necessary to determine the phase of each time-frequency point in order to invert the resulting STFT. The most common method is to reuse the phase from the mixture STFT. However, statistical approaches have been proposed [10]. This paper re-examines the method for estimating rank-one source spectro-

grams from a mixture spectrogram. Prior work disregards the unknown phase of the sources in this analysis. We incorporate a probabilistic representation of phase and show an improvement in the estimation of the unknown rank-one source spectrograms.

3. REPRESENTATION

We consider the case of a single mixture signal that is the sum of multiple source signals:

$$x(t) = \sum_r s_r(t) \tag{1}$$

We transform this signal into the time-frequency domain using the short-time Fourier transform (STFT):

$$X(k,t) = \sum_n h(n-t)x(n)e^{-jkn} \tag{2}$$

where h is a localization window. In matrix form, $\mathbf{X}_{kt} = X(k,t)$, and the mixing equation becomes:

$$\mathbf{X} = \sum_r \mathbf{S}_r \tag{3}$$

Each element of these matrices is a complex number which we represent as a phasor:

$$\mathbf{X} = |\mathbf{X}|e^{j\mathbf{\Theta}}, \qquad \mathbf{S}_r = |\mathbf{S}_r|e^{j\mathbf{\Theta}_r} \tag{4}$$

where all operations are element-wise and $\mathbf{\Theta}$ is a phase matrix. However, in contrast to the assumption made by the algorithms described above, the mixture spectrogram is not generally the sum of source spectrograms:

$$|\mathbf{X}| \neq \sum_r |\mathbf{S}_r| \tag{5}$$

Instead, the mixture spectrogram is a function of the source spectrograms and the phase difference between them:

$$|\mathbf{X}|^2 = \sum_{qr} |\mathbf{S}_q||\mathbf{S}_r| \cos \mathbf{\Theta}_{qr} \tag{6}$$

where $\mathbf{\Theta}_{qr}$ is the phase difference, $\mathbf{\Theta}_q - \mathbf{\Theta}_r$, between $\mathbf{S}_q$ and $\mathbf{S}_r$. Notice that if only one source, $\mathbf{S}_r$, is active, $|\mathbf{X}|^2 = |\mathbf{S}_r|^2$, and if all sources have the same phase,

$$|\mathbf{X}|^2 = \sum_{qr} |\mathbf{S}_q||\mathbf{S}_r| = \left(\sum_r |\mathbf{S}_r|\right)^2. \tag{7}$$

4. NON-NEGATIVE MATRIX FACTORIZATION

Non-negative matrix factorization is a technique for estimating a non-negative $K \times T$ matrix $\mathbf{V}$ as the product of a non-negative $K \times R$ matrix $\mathbf{W}$ and a non-negative $T \times R$ matrix $\mathbf{H}$:

$$\mathbf{V} \approx \mathbf{W}\mathbf{H}^T \tag{8}$$

If $\mathbf{V}$ is a spectrogram, it has K frequency bins and T samples. The R source components are estimated by minimizing a distance metric between $\mathbf{V}$ and $\mathbf{W}\mathbf{H}^T$ such as the square of the Euclidian distance:

$$\|\mathbf{V} - \mathbf{W}\mathbf{H}^T\|^2 = \sum_{kt} \left(\mathbf{V}_{kt} - \left[\mathbf{W}\mathbf{H}^T\right]_{kt}\right)^2 \tag{9}$$

Minimizing this function using gradient descent under the constraint of non-negativity leads to a solution for $\mathbf{W}$ and $\mathbf{H}$ [11].

When applied to a mixture spectrogram, the columns of $\mathbf{W}$, $\mathbf{w}_r$, represent spectral shapes and the columns of $\mathbf{H}$, $\mathbf{h}_r$, represent amplitude envelopes. The source and mixture spectrograms are given by $|\mathbf{S}_r| = \mathbf{w}_r\mathbf{h}_r^T$ and $|\mathbf{X}| = \mathbf{V}$, respectively. This approach optimizes one possible configuration of phase. By incorporating the true distribution of phase, we improve the estimates of $\mathbf{W}$ and $\mathbf{H}$.

5. PROBABILISTIC REPRESENTATION OF PHASE

We consider the phase at each time-frequency point of each source to be a uniformly distributed random variable. We simplify the notation for the case of two components so that $v = |\mathbf{X}_{kt}|$, $a = |[\mathbf{S}_1]_{kt}|$, $b = |[\mathbf{S}_2]_{kt}|$, and $\theta = [\mathbf{\Theta}_{12}]_{k,t}$ for a particular value of k and t. The magnitude of the sum of two complex numbers is a function of the magnitude of each number and the phase difference between them:

$$v = \sqrt{a^2 + b^2 + 2ab\cos\theta} \tag{10}$$

Because of the circularity of phase, the difference in two uniformly distributed random phases is also a uniformly distributed random variable, $\theta = U(-\pi, \pi)$. Because v is a function of θ, v is also a random variable with the following probability density function given a and b:

$$p(v|a,b) = \frac{2v}{\pi\sqrt{-(v+a+b)(v+a-b)(v-a+b)(v-a-b)}} \tag{11}$$

The roots of the polynomial inside the square root are $v = \pm a \pm b$. The function is defined in the interval $(|a-b|, a+b)$ and approaches infinity as v approaches $|a-b|$ and $a+b$. Figure 1 plots $p(v|a,b)$ with $a = 2$ and $b = 1$.

In our problem, the mixture spectrogram is known, and the source spectrograms need to be estimated. Therefore, we maximize the likelihood in Equation 11 as a function of a and b. Figure 2 shows the surface of $p(v|a,b)$ with $v = 1$. The dark lines on the ab-plane are intersections of the asymptotic planes.

For values of a and b inside the region defined by the asymptotes in Figure 2, maximizing $p(v|a,b)$ for every time-frequency point is equivalent to minimizing the following:

$$\underset{a,b}{\operatorname{argmin}} \left|(v+a+b)(v+a-b)(v-a+b)(v-a-b)/v^2\right| \tag{12}$$

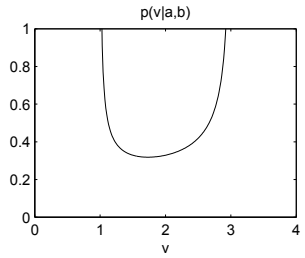

Fig. 1. Likelihood function for v when $a = 2$ and $b = 1$.

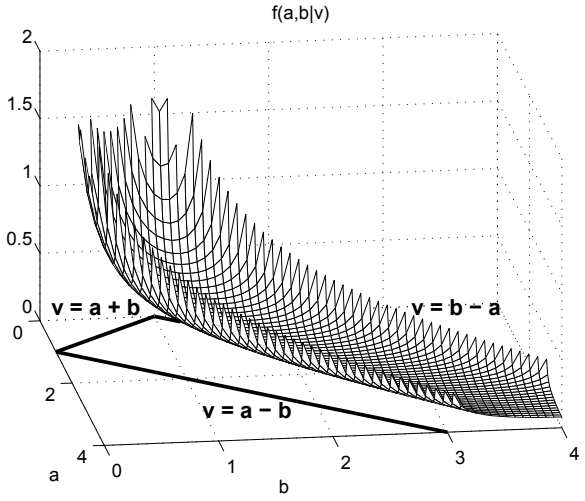

Fig. 2. Energy function for a and b when $v = 1$.

The absolute value avoids imaginary values for points outside this region. Notice that each of the roots defines an asymptote in Figure 2 and the function reaches a minimum of zero when (a, b) falls on an asymptote. (One asymptote is not visible because it does not intersect the positive a or b axis, namely $v = -a - b$.) In order to take advantage of the efficiency and convergence properties of the squared Euclidian distance function, we propose estimating a and b by minimizing the following function:

$$\underset{a,b}{\operatorname{argmin}} \left((v + a - b)(v - a + b)(v - a - b)/v^2\right)^2 \qquad (13)$$

which reaches a minimum of zero for the same points, (a, b), as Equation 12, specifically at the asymptotes in the positive ab-plane.

6. UPDATE RULES

We minimize the function D, which is the sum of Equation 12 across all time-frequency points:

$$D = \frac{1}{2} \sum_{kt} \mathbf{P}_{kt}^2 \mathbf{Q}_{kt}^2 \mathbf{R}_{kt}^2 / \mathbf{V}_{kt}^4 \qquad (14)$$

where $\mathbf{V} = |\mathbf{X}|$, $\mathbf{A} = |\mathbf{S}_1|$, $\mathbf{B} = |\mathbf{S}_2|$, $\mathbf{P} = \mathbf{V} + \mathbf{A} - \mathbf{B}$, $\mathbf{Q} = \mathbf{V} - \mathbf{A} + \mathbf{B}$, $\mathbf{R} = \mathbf{V} - \mathbf{A} - \mathbf{B}$, and all the operations are element-wise. Taking the derivative of D with respect to each of the columns of $\mathbf{W}$ and $\mathbf{H}$ yields:

$$\frac{\partial D}{\partial \mathbf{w}_1} = \left(\frac{\mathbf{P}\mathbf{Q}^2\mathbf{R}^2 - \mathbf{P}^2\mathbf{Q}\mathbf{R}^2 - \mathbf{P}^2\mathbf{Q}^2\mathbf{R}}{\mathbf{V}^4} \right) \mathbf{h}_1 \qquad (15)$$

$$\frac{\partial D}{\partial \mathbf{h}_1} = \mathbf{w}_1^T \left(\frac{\mathbf{P}\mathbf{Q}^2\mathbf{R}^2 - \mathbf{P}^2\mathbf{Q}\mathbf{R}^2 - \mathbf{P}^2\mathbf{Q}^2\mathbf{R}}{\mathbf{V}^4} \right) \qquad (16)$$

$$\frac{\partial D}{\partial \mathbf{w}_2} = \left(\frac{-\mathbf{P}\mathbf{Q}^2\mathbf{R}^2 + \mathbf{P}^2\mathbf{Q}\mathbf{R}^2 - \mathbf{P}^2\mathbf{Q}^2\mathbf{R}}{\mathbf{V}^4} \right) \mathbf{h}_2 \qquad (17)$$

$$\frac{\partial D}{\partial \mathbf{h}_2} = \mathbf{w}_2^T \left(\frac{-\mathbf{P}\mathbf{Q}^2\mathbf{R}^2 + \mathbf{P}^2\mathbf{Q}\mathbf{R}^2 - \mathbf{P}^2\mathbf{Q}^2\mathbf{R}}{\mathbf{V}^4} \right) \qquad (18)$$

where the operations inside the parentheses are element-wise with a matrix-vector product on the outside. We randomly initialize $\mathbf{W}$ and $\mathbf{H}$, and minimize D using gradient descent.

7. RESULTS

In order to compare our probabilistic phase algorithm against standard non-negative matrix factorization we construct source and mixture spectrograms as follows:

$$\begin{array}{rclrcl} \mathbf{W}_{kr} & = & |N(0,1)| & \mathbf{H}_{tr} & = & |N(0,1)| \\ [\mathbf{\Theta}_1]_{kt} & = & U(-\pi, \pi) & [\mathbf{\Theta}_2]_{kt} & = & U(-\pi, \pi) \\ \mathbf{S}_1 & = & \left(\mathbf{w}_1 \mathbf{h}_1^T\right) e^{j\mathbf{\Theta}_1} & \mathbf{S}_2 & = & \left(\mathbf{w}_2 \mathbf{h}_2^T\right) e^{j\mathbf{\Theta}_2} \\ \mathbf{X} & = & \mathbf{S}_1 + \mathbf{S}_2 & \mathbf{V} & = & |\mathbf{X}| \\ \hat{\mathbf{W}}_{kr}^{\text{init}} & = & |N(0,1)| & \hat{\mathbf{H}}_{tr}^{\text{init}} & = & |N(0,1)| \end{array}$$

We chose $K = T = 100$, $R = 2$, and ran both algorithms for 1000 trials, each time drawing new source spectrograms and initializing NMF with random matrices $\hat{\mathbf{W}}^{\text{init}}$ and $\hat{\mathbf{H}}^{\text{init}}$. We initialize our approach with the NMF solution. The scatter plot of time-frequency bins from one representative trial is shown in Figure 3a-c. Each point represents one time-frequency point of the source spectrograms. The position of each point is normalized to the $v = 1$ scale. That is, the position of each time-frequency bin is at $(a/v, b/v)$. Notice that our approach (Figure 3c) more closely resembles the actual scatter plot (Figure 3a) than traditional NMF (Figure 3b). Figure 3d-f shows the combined histograms for all trials. Notice that our approach (Figure 3f) has visible tails along $v = a - b$ and $v = b - a$ similar to the correct histogram, whereas NMF (Figure 3e) does not. We compute the mean square error between the estimated and true $\mathbf{W}$ and $\mathbf{H}$ after normalizing the columns of each to unit L_2 norm as follows:

$$MSE = \frac{1}{KR} \sum_{kr} \left(\hat{\mathbf{W}}_{kr} - \mathbf{W}_{kr} \right)^2 + \frac{1}{TR} \sum_{tr} \left(\hat{\mathbf{H}}_{tr} - \mathbf{H}_{tr} \right)^2 \qquad (19)$$

Over the 1000 trials, the mean square error for NMF was 3.37e-4, whereas our approach attained a mean square error of 2.43e-4 for an improvement of 28%.

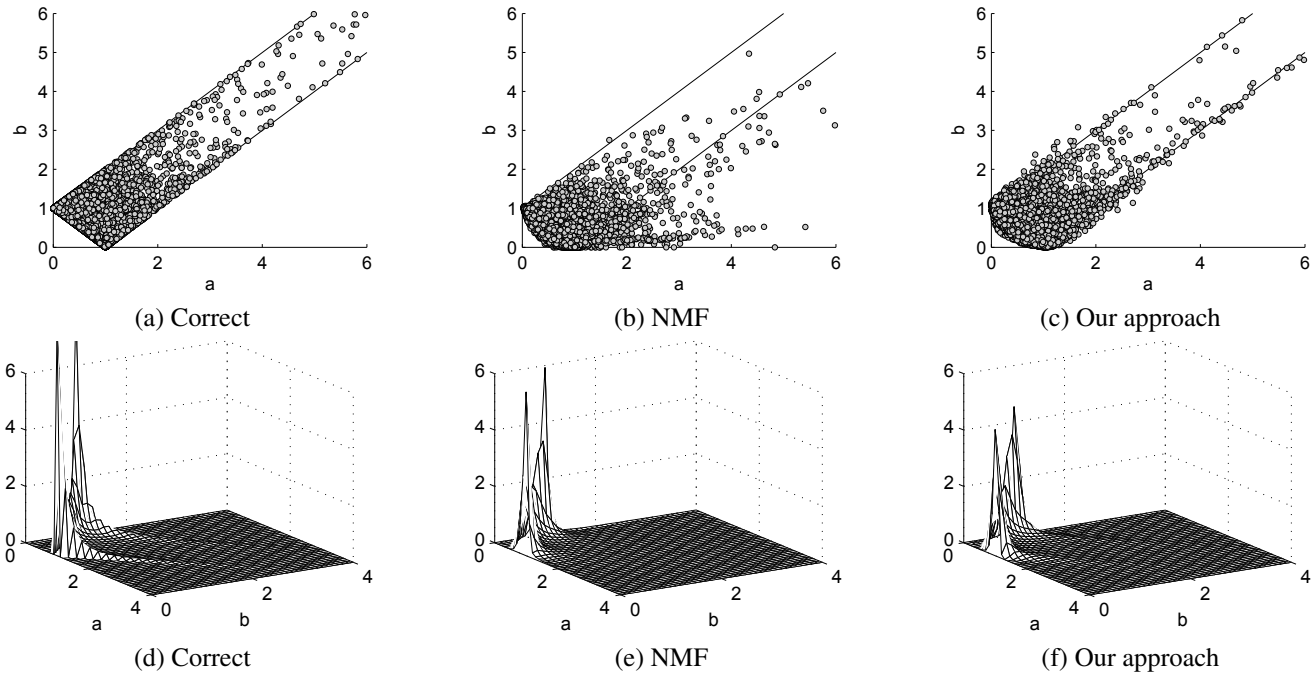

Fig. 3. Scatter plot of bins for one trial (a-c) and histogram for all trials in units of 10^5 (d-f).

8. CONCLUSION AND FUTURE WORK

We have shown that phase plays an important role in the determination of the mixture spectrogram from a number of source spectrograms. By incorporating a probabilistic representation of phase, we propose an improvement on NMF that more closely follows the true distribution of mixture spectrogram points given the source spectrograms. Our integrated approach provides a substantial improvement on traditional NMF, warranting further work in this area. Future work includes leveraging a probabilistic representation of phase for mixtures of more than two source spectrograms.

9. REFERENCES

[1] M. Casey and W. Westner, "Separation of mixed audio sources by independent subspace analysis," in *Proc. of the Int'l Computer Music Conference*, Berlin, 2000.

[2] P. Smaragdis, *Redundancy Reduction for Computational Audition, a Unifying Approach*, Ph.D. thesis, MAS Department, Massachusetts Institute of Technology, 2001.

[3] T. Virtanen, "Separation of sound sources by convolutive sparse coding," in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, 2004.

[4] B. Wang and M. D. Plumbley, "Investigating single-channel audio source separation methods based on non-negative matrix factorization," in *ICA Research Network Int'l Workshop*, 2006, pp. 17–20.

[5] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *Proc. of the Int'l Conference on Music Information Retrieval*, Barcelona, Spain, 2004, pp. 318–325.

[6] D. FitzGerald, E. Coyle, and B. Laylor, "Sub-band independent subspace analysis for drum transcription," in *Proc. of Int'l Conference on Digital Audio Effects*, Hamburg, Germany, 2002, pp. 65–69.

[7] P. D. O'Grady and B. A. Pearlmutter, "Convolutive non-negative matrix factorisation with sparseness constraint," in *Proc. of the IEEE Int'l Workshop on Machine Learning for Signal Processing*, 2006.

[8] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2003, pp. 177–180.

[9] A. Hyvärinen, *Independent Component Analysis*, New York: Wiley, 2001.

[10] K. Achan, S. T. Roweis, and B. J. Frey, "Probabilistic inference of speech signals from phaseless spectrograms," in *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

[11] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in NIPS 13*, pp. 556–562. MIT Press, 2001.