INTERACTIVE REFINEMENT OF SUPERVISED AND SEMI-SUPERVISED SOUND SOURCE SEPARATION ESTIMATES

*Nicholas J. Bryan**

Center for Computer Research in Music and Acoustics, Stanford University

Gautham J. Mysore

Adobe Research

ABSTRACT

We propose an interactive refinement method for supervised and semi-supervised single-channel source separation. The refinement method allows end-users to provide feedback to the separation process by painting on spectrogram displays of intermediate output results. The time-frequency annotations are then used to update the separation estimates and iteratively refine the results. The initial separation is performed using probabilistic latent component analysis and is then extended to incorporate the painting annotations using linear grouping expectation constraints via the framework of posterior regularization. Using a prototype user-interface, we show that the method is able to perform high-quality separation with minimal user-interaction.

Index Terms— source separation, probabilistic latent component analysis, user-interaction

1. INTRODUCTION

One of most promising source separation approaches is based on supervised and semi-supervised non-negative matrix factorization (NMF) methods [1, 2, 3, 4] and its probabilistic counterparts applied to audio spectrogram data [5, 6, 7, 8, 9]. While these methods can achieve high-quality separation in some cases, the results are often far from perfect. Sound artifacts such as musical noise are also a problem.

To overcome these issues, we propose an interactive refinement method that enables end-users to provide feedback into the separation process. We do this by allowing users to paint on spectrogram displays of intermediate output results. The annotations are then used to update the separation estimates and the entire process is repeated as shown in Fig. 1. The initial separation is performed via supervised or semi-supervised probabilistic latent component analysis [8], and then refined in an interactive fashion by constraining the probabilistic model via the framework of posterior regularization. We show how this can greatly refine the separation estimates, and we provide intuition as to why this happens.

*This work was performed while interning at Adobe Research.

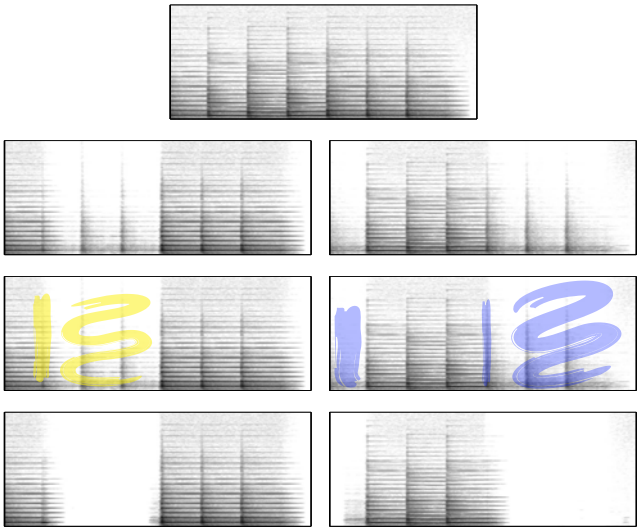

Fig. 1. (First Row) Mixture spectrogram of *Mary Had A Little Lamb*. (Second Row) Initial separated E note (left) and remaining notes (right) after supervised separation using PLCA. (Third Row) Annotations overlaid on the incorrectly separated regions. (Bottom) Refined separated E note and remaining notes after incorporating the annotations.

2. PROPOSED APPROACH

The proposed refinement method is an iterative procedure that separates a mixture signal into its respective sources, and then allows a user to correct for errors in the output estimates and update the results, creating a feedback-loop as shown in Fig. 2. The initial separation method is discussed in Section 2.1, Section 2.2, and Section 2.3, and the method of user annotations is discussed in Section 2.4.

2.1. Probabilistic Model

To initially separate a mixture recording into its respective components, we use probabilistic latent component analysis (PLCA) [8]. The method models normalized spectrogram data $\mathbf{X} \in \mathbf{R}_{+}^{N_f \times N_t}$ as a factorized two-dimensional proba-

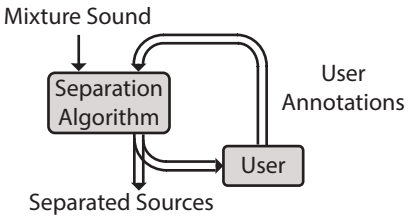

Fig. 2. Block diagram of the proposed refinement method.

bility distribution over time and frequency

$$P(f,t) = \sum_z P(z)P(f|z)P(t|z), \quad (1)$$

where f, t, and z are outcome values of the random variables F, T, and Z, and N_f, N_t, and N_z are the possible number of outcomes respectively. Note, Z is a latent variable. N_z is typically chosen by a user and N_f and N_t are a function of the recording length and STFT parameters. The marginal distribution $P(f|z)$ represents the frequency basis vectors or characteristic frequency content of a given source. The marginal distribution $P(t|z)$ represents the weights or activations of each basis vector, and the distribution $P(z)$ represent the weighting of the latent variable outcomes.

2.2. Parameter Estimation

To fit our probabilistic model to data, we use an expectation-maximization (EM) algorithm to find maximum-likelihood estimates of the model parameters of $P(f|z)$, $P(t|z)$, and $P(z)$. The resulting EM algorithm consists of an iterative two-stage optimization procedure. In the E step, the posterior distribution is computed using the model parameters

$$P(z|f,t) = \frac{P(z)P(f|z)P(t|z)}{\sum_{z'} P(z')P(f|z')P(t|z')}. \quad (2)$$

In the M step, this result is used to update the maximum-likelihood estimates of the model parameters,

$$P(f|z) = \frac{\sum_t \mathbf{X}_{(f,t)}\, Q(z|f,t)}{\sum_{f'}\sum_{t'} \mathbf{X}_{(f',t')}\, Q(z|f',t')}, \quad (3)$$

$$P(t|z) = \frac{\sum_f \mathbf{X}_{(f,t)}\, Q(z|f,t)}{\sum_{f'}\sum_{t'} \mathbf{X}_{(f',t')}\, Q(z|f',t')}, \quad (4)$$

$$P(z) = \frac{\sum_f \sum_t \mathbf{X}_{(f,t)}\, Q(z|f,t)}{\sum_{z'}\sum_{f'}\sum_{t'} \mathbf{X}_{(f',t')}\, Q(z'|f',t')}, \quad (5)$$

where we define the distribution $Q(z|f,t) = P(z|f,t)$, which, in this case, is the posterior distribution. $\mathbf{X}_{(f,t)}$ refers to time-frequency bin f,t of the spectrogram data. Both steps are repeated in secession until convergence.

2.3. Supervised and Semi-Supervised Separation

We use this model to perform supervised source separation [8] as follows:

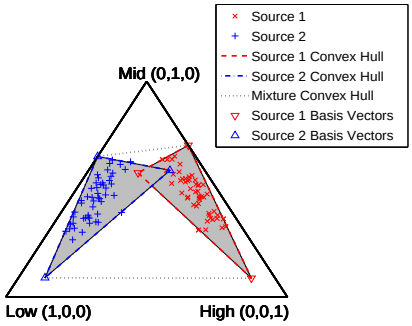

Fig. 3. A 2-simplex of normalized three-frequency spectra of two sources. The convex hulls and basis vectors of each source are shown, along with the mixture convex hull. The bottom-left, top-middle, and bottom-right corners of the simplex represent low, middle, and high frequencies respectively.

1. For each source $s \in 1, ..., N_s$ in a mixture, obtain a spectrogram of isolated training data of that source, and learn the PLCA model parameters as shown above. Retain the learned basis vectors $P_s(f|z)$ of each source and discard the other model parameters.

2. Combine the frequency basis vectors of each source $P_s(f|z)$ to form a larger collection of basis vectors by taking the union of each source basis $P(f|z) = \{\bigcup_{s=1}^{N_s} P_s(f|z)\}$, assuming each value of z is unique and N_f for each source is the same.

3. Estimate the contribution of each source in the mixture spectrogram data $\mathbf{X}$ by fixing $P(f|z)$, and computing the unknown distributions $P(t|z)$ and $P(z)$.

4. Compute the posterior distribution $P(z|f,t)$ (see (2) below) and $P(s|f,t) = \sum_{z \in Z_s} P(z|f,t)$, which serves as a soft mask that corresponds to the estimated proportion of each source at each time-frequency bin.

5. For each s, element-wise multiply the soft mask $P(s|f,t)$ with the mixture spectrogram data $\mathbf{X}$ and mixture phase, resulting in $\hat{\mathbf{X}}_s$.

6. Convert each $\hat{\mathbf{X}}_s$ to a time domain signal via the inverse STFT, completing the separation process.

This process is illustrated in Fig. 3, where we display 3-dimensional normalized spectrogram data (i.e. spectra with low, middle, and high frequencies) of two sound sources using a standard 2-simplex diagram. We additionally show three learned basis vectors and the convex hulls for each source. The convex hull of a source represents the general geometrical region within the simplex of the given source's normalized spectra and is defined by the set $H = \{\sum_{z \in Z_s} \theta_z P(f|z) \mid \theta_z \geq 0, \sum_{z \in Z_s} \theta_z = 1\}$, where Z_s is the set of values of z for source s.

When one or more sources lack isolated training data, semi-supervised separation must be used in place of supervised separation. The process is largely similar, except that

the basis vectors (or elements of $P(f|z)$) of sources with no training data are estimated from the mixture signal along with $P(t|z)$ and $P(z)$.

2.4. Posterior Regularization

Posterior regularization (PR) [10, 11, 12], initially proposed by Graça et al. is a flexible method of incorporating rich, typically data-dependent, constraints into latent variable models using EM algorithms. Prior regularization (maximum a posteriori estimation using a prior distribution), constrains the space of model parameters in the maximization step of the EM algorithm. In contrast, posterior regularization directly constrains the posterior distribution in the expectation step. This is done by first computing the posterior distribution $P(z|f,t)$ via (2), and then finding a distribution $Q(z|f,t)$ that best fits the posterior, given constraints. In doing so, the method facilitates constraints that would otherwise be difficult to encode using standard priors.

We use PR to constrain the posterior distribution $P(z|f,t)$ as a function of the time-frequency painting annotations. We do this by considering the real-valued annotations, stored as a set of matrices $\mathbf{\Lambda}^s \in \mathbf{R}^{N_f \times N_t},\ \forall\ s = 1, ..., N_s$, as penalty weights that discourage one source or another from explaining the observed spectrogram data for each time-frequency point. The annotation color and opacity indicate the source and strength of the penalty weights (initialized to all zeros). In this way, a user can independently control how each source contributes to a mixture recording at each time-frequency point, which is not possible using standard priors.

The penalties are incorporated via linear grouping constraints into the E step. This leads us to minimize the Kullback-Leibler divergence between the posterior and a free distribution Q [13], given the linear constraints, resulting in

$$\begin{aligned} \arg\min_{\mathbf{q}} \quad & -\mathbf{q}^{\mathrm{T}} \ln \mathbf{p} + \mathbf{q}^{\mathrm{T}} \ln \mathbf{q} + \mathbf{q}^{\mathrm{T}} \boldsymbol{\lambda} \\ \text{subject to} \quad & \mathbf{q}^{\mathrm{T}} \mathbf{1} = 1,\ \mathbf{q} \succeq 0, \end{aligned} \tag{6}$$

for each value of f and t in the posterior (in standard EM, the optimization does not include the penalty $\mathbf{q}^{\mathrm{T}} \boldsymbol{\lambda}$ [13]). $\mathbf{p} \in \mathbf{R}^{N_z}$ is the corresponding vector of posterior probability values for a given time-frequency point, $\mathbf{q} \in \mathbf{R}^{N_z}$ is the corresponding vector of $Q(z|f,t)$, $^{\mathrm{T}}$ is a matrix transpose, $\succeq$ is element-wise greater than or equal to, and $\mathbf{1}$ is a column vector of ones. The weight vector $\boldsymbol{\lambda} \in \mathbf{R}^{N_z}$ is then constructed as $\boldsymbol{\lambda}_{(z)} = \mathbf{\Lambda}^s_{(f,t)},\ \forall s = 1, ..., N_s, \forall z \in Z_s$. When two sources are separated, for example, and one source is assigned two basis vectors and the other with three, $\boldsymbol{\lambda} = [\alpha, \alpha, \beta, \beta, \beta]$, with $\alpha = \mathbf{\Lambda}^{s_1}_{(f,t)}$ and $\beta = \mathbf{\Lambda}^{s_2}_{(f,t)}$ for a particular time-frequency point. When the annotations are all zero, the constraints have no effect, and the process reduces to standard PLCA described above.

When we solve (6) for each time-frequency point and rearrange terms, we arrive at a new E step update,

$$Q(z|f,t) = \frac{P(z)P(f|z)P(t|z)\tilde{\mathbf{\Lambda}}_{(f,t,z)}}{\sum_{z'} P(z')P(f|z')P(t|z')\tilde{\mathbf{\Lambda}}_{(f,t,z')}}, \tag{7}$$

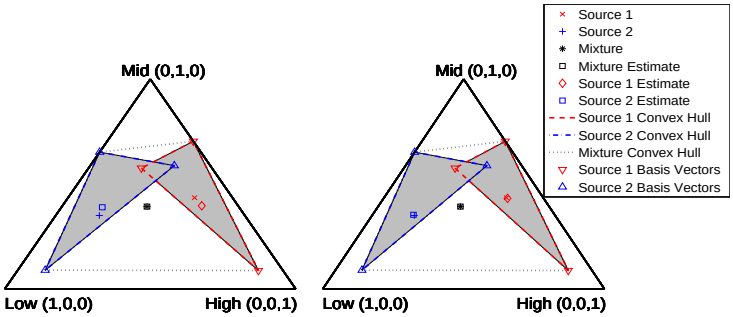

Fig. 4. A 2 simplex diagram illustrating the reconstructions of the mixture and individual sources when using supervised separation. In both cases, the mixture is well reconstructed. However, the reconstruction error of the individual sources are noticeably higher using standard PLCA (left) compared to the proposed refinement method (right).

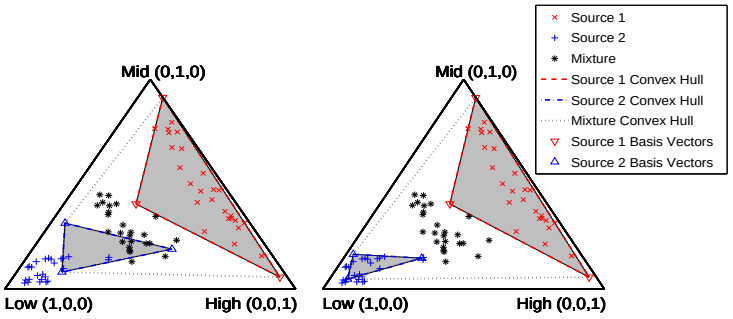

Fig. 5. A 2-simplex diagram illustrating semi-supervised separation. The basis vectors of the first source are learned in a supervised manner, while the basis vectors of the second source are learned via semi-supervised learning. The proposed method (right) results in better estimates of the second source basis vectors compared to standard PLCA (left).

where $\tilde{\mathbf{\Lambda}} = \exp\{-\mathbf{\Lambda}\}$, $\exp\{\}$ is an element-wise exponential function, and $\mathbf{\Lambda} \in \mathbf{R}^{N_f \times N_t \times N_z}$ represents the entire set of real-valued painting annotations. In this case, $\mathbf{\Lambda}$ is indexed by f, t, and z (instead of f, t, and s) for convenience, knowing which values of z correspond to each source s. Equation (2) and the definition of $Q(z|f,t)$ from before are then replaced with equation (7), resulting in an updated posterior-regularized EM algorithm.

We show the benefit of PR in Fig. 4 for a simulated supervised separation experiment. In this case, we learn basis vectors for two sources and then separate an unknown mixture of the two. On the left, we show a single mixture point and the original source points used to create the mixture, along with the separation estimates of the individual sources using standard PLCA. On the right, we apply posterior regularization to refine the procedure, resulting in better separation estimates of the original unmixed sources.

We further depict the benefit of PR in Fig. 5 for the case of semi-supervised separation. In this case, we learn the ba-

EXAMPLE	IDEAL	SUPERVISED	SEMI-SUPERVISED
CELL	30.7	29.2 / 27.6	28.4 / 06.5
DRUM	14.8	09.7 / 08.5	07.7 / 03.9
COUGH	15.8	14.0 / 12.5	12.0 / 10.5
PIANO	26.1	26.0 / 21.6	14.9 / 08.4
SIREN	27.8	23.8 / 18.9	21.0 / 19.9

Table 1. SDR (in dB) for supervised and semi-supervised separation with/without refinement vs. ideal mask results.

sis vectors from one source using supervised learning and the basis vectors of another source using semi-supervised learning. When we apply standard PLCA, the basis vectors from the second source do not accurately represent its true distribution as shown on the left. When we apply posterior regularized PLCA, the basis vectors for the second source more accurately outline its true distribution as shown on the right, resulting in better separation estimates.

3. EVALUATION

To test the proposed method, we developed a prototype user-interface and used it to separate five example mixture sounds using both supervised and semi-supervised separation over the course of five minutes per task. The mixture sounds were artificially mixed at 0dB levels and include: ambulance siren + speech, cell phone + speech, drum + bass loop, orchestra + coughing, and piano chords + incorrect note.

The separation estimates are compared to standard PLCA via the source-to-distortion ratio (SDR) metric [14]. The SDR metric considers both the suppression of the unwanted sources and added artifacts introduced by the separation algorithm. We also compare both results to an ideal masking filter, generated by using the original unmixed source recordings to compute the source probabilities $P(s|f,t)$.

Table 1 shows the SDR results for supervised separation, semi-supervised separation, and ideal mask results using 100 basis vectors per source. For supervised separation, the original unmixed tracks were used for training. For semi-supervised separation, we used a portion of the mixture data in which only one source was present as the training data for that source. In all cases, the refinement procedure increased the SDR for both supervised and semi-supervised separation and, in certain cases, approached the quality of the ideal mask. These results were insensitive to varying the number of basis vectors unlike standard PLCA [8].

The cell phone example, using semi-supervised separation, is illustrated in Fig. 6. In this case, the proposed method significantly outperforms standard PLCA with minimal user interaction (only a single harmonic of a single ring is annotated). For sound examples and demonstration videos, please see `https://ccrma.stanford.edu/~njb/research/iss/`.

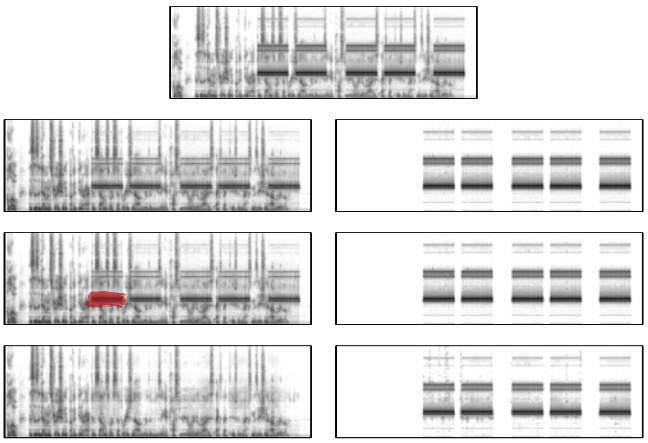

Fig. 6. (First Row) A mixture spectrogram of speech + cell phone. (Second Row) Initial separated speech (left) and cell phone (right) after semi-supervised PLCA. (Third Row) Painting annotation overlaid on the incorrectly separated regions. Note: only a single harmonic of a single ring is roughly annotated. (Bottom) Refined separated speech and cell phone.

4. RELATED WORK

There are a few related works that leverage user-guided information to improve the source separation process. In [15], a user is tasked to sing or hum a desired source signal to guide what source should be removed from the mixture. The guide signal is then used to inform PLCA using prior probabilities. In [16], a user is tasked to select the fundamental frequency of a source. The annotations are then used to inform a source-filter NMF model. In [17], binary time-frequency annotations are used to select patches of one source or another to inform NMF with the intention of training a completely automatic user-free system. While similar, these works use both different user-input and different algorithms. More specifically, they do not allow real-valued, time-frequency constraints to guide the separation process (enabled by PR) or provide a feedback mechanism to iteratively improve results, which we believe are the fundamental advantages of this work.

5. CONCLUSIONS

We have presented an interactive refinement method for supervised and semi-supervised single-channel source separation. The method allows end-users to provide feedback into the separation process by painting on spectrogram displays of intermediate output results to refine the separation estimates. To do so, we extend supervised and semi-supervised probabilistic latent component analysis via posterior regularization. Initial evaluation shows promising results with minimal user-interaction and little additional computation cost. For future work, we hope to extend the technique for the case of separation without training data and to develop techniques to reduce computational complexity.

6. REFERENCES

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*. 2001, pp. 556–562, MIT Press.

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

[3] T. 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, Mar. 2007.

[4] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 550–563, 2010.

[5] B. Raj and P. Smaragdis, "Latent variable decomposition of spectrograms for single channel speaker separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2005, pp. 17 – 20.

[6] P. Smaragdis, B. Raj, and M. Shashanka, "A Probabilistic Latent Variable Model for Acoustic Modeling," in *Workshop on Advances in Modeling for Acoustic Processing, Advances in Neural Information Processing Systems*, 2006.

[7] P. Smaragdis and B. Raj, "Shift-Invariant Probabilistic Latent Component Analysis," *MERL Tech Report*, 2007.

[8] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *International Conference on Independent Component Analysis and Signal Separation*, Berlin, Heidelberg, 2007, pp. 414–421, Springer-Verlag.

[9] M. Shashanka, *Latent variable framework for modeling and separating single-channel acoustic sources*, Ph.D. thesis, Boston University, Boston, MA, USA, 2008.

[10] J. Graça, K. Ganchev, and B. Taskar, "Expectation maximization and posterior constraints," in *Advances in Neural Information Processing Systems*, 2007.

[11] J. Graça, K. Ganchev, B. Taskar, and F. C. N. Pereira, "Posterior vs parameter sparsity in latent variable models," in *Advances in Neural Information Processing Systems*, 2009, pp. 664–672.

[12] K. Ganchev, J. Graça, J. Gillenwater, and B. Taskar, "Posterior Regularization for Structured Latent Variable Models," *Journal of Machine Learning Research*, vol. 11, pp. 2001–2049, Aug. 2010.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[14] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462 –1469, July 2006.

[15] P. Smaragdis and G. J. Mysore, "Separation by "humming": User-guided sound extraction from monophonic mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 69–72.

[16] J.-L. Durrieu and J.-P. Thiran, "Musical audio source separation based on user-selected f0 track," in *International Conference on Latent Variable Analysis and Signal Separation*, 2012, pp. 438–445.

[17] A. Lefevre, F. Bach, and C. Févotte, "Semi-supervised nmf with time-frequency annotations for single-channel source separation," in *International Society for Music Information Retrieval Conference*, 2012.