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 \cite{1, 2, 3, 4} and its probabilistic counterparts applied to audio spectrogram data \cite{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 \cite{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.

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) \cite{8}. The method models normalized spectrogram data $\mathbf{X} \in \mathbb{R}^{N_F \times N_t}$ as a factorized two-dimensional proba-
Mixture Sound
Separated Sources
User Annotations
Separation Algorithm
User

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

Separability distribution over time and frequency

\[ P(f, t) = \sum_z P(z)P(f\mid z)P(t\mid z), \]  

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 \( P(f\mid z) \) represents the frequency basis vectors or characteristic frequency content of a given source. The marginal distribution \( P(t\mid z) \) represents the weights or activations of each basis vector, and the distribution \( P(z) \) represents 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 \( P(f\mid z), P(t\mid z), \) 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

\[ P(z\mid f, t) = \frac{P(z)P(f\mid z)P(t\mid z)}{\sum_{z'} P(z')P(f\mid z')P(t\mid z')} \]  

(2)

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

\[ P(f\mid z) = \frac{\sum_{f'} \sum_{t'} X_{(f,t)} Q(z\mid f, t)}{\sum_{f'} \sum_{t'} X_{(f', t')} Q(z\mid f', t')}, \]  

(3)

\[ P(t\mid z) = \frac{\sum_{f'} \sum_{t'} X_{(f,t')} Q(z\mid f', t')}{\sum_{f'} \sum_{t'} X_{(f', t')} Q(z\mid f', t')}, \]  

(4)

\[ P(z) = \frac{\sum_{f'} \sum_{t'} \sum_{z'} X_{(f,t')} Q(z\mid f', t', z')}{\sum_{z'} \sum_{f'} \sum_{t'} X_{(f,t')} Q(z\mid f', t', z')}, \]  

(5)

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

2.3. Supervised and Semi-Supervised Separation

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

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 \( P_s(f\mid z) \) of each source and discard the other model parameters.

2. Combine the frequency basis vectors of each source \( P_s(f\mid z) \) to form a larger collection of basis vectors by taking the union of each source basis \( P(f\mid z) = \bigcup_{s=1}^{N_s} P_s(f\mid z) \), 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 \( X \) by fixing \( P(f\mid z) \), and computing the unknown distributions \( P(t\mid z) \) and \( P(z) \).

4. Compute the posterior distribution \( P(z\mid f, t) \) (see (2) below) and \( P(s\mid f, t) = \sum_{z\in Z_s} P(z\mid f, t) \), 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 \( P(s\mid f, t) \) with the mixture spectrogram data \( X \) and mixture phase, resulting in \( \tilde{X}_s \).

6. Convert each \( \tilde{X}_s \) 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 \( H = \{ \sum_{z\in Z_s} \theta_z P(f\mid z) \mid \theta_z \geq 0, \sum_{z\in Z_s} \theta_z = 1 \} \), 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 Grac¸a 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 $\Lambda^s \in \mathbb{R}^{N_f \times N_s}$, $\forall s = 1, ..., N_s$, 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

$$\arg\min_{q} - q^T \ln p + q^T \ln q + q^T \lambda$$

subject to $q^T 1 = 1$, $q \geq 0$,\hspace{1cm} (6)

for each value of $f$ and $t$ in the posterior (in standard EM, the optimization does not include the penalty $q^T \lambda$ [13]). $p \in \mathbb{R}^{N_z}$ is the corresponding vector of posterior probability values for a given time-frequency point, $q \in \mathbb{R}^{N_z}$ is the corresponding vector of $Q(z|f,t)$, $T$ is a matrix transpose, $\geq$ is element-wise greater than or equal to, and $1$ is a column vector of ones. The weight vector $\lambda \in \mathbb{R}^{N_z}$ is then constructed as $\lambda_{(z)} = \Lambda_s^{f,t}$, $\forall s = 1, ..., N_s$, $\forall z \in Z_s$. When two sources are separated, for example, and one source is assigned two basis vectors and the other with three, $\lambda = [\alpha, \alpha, \beta, \beta, \beta]$, with $\alpha = \Lambda_s^{f,t}$ and $\beta = \Lambda_s^{f,t}$ 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,

$$Q(z|f,t) = \frac{P(z)P(f|z)P(t|z)}{\sum_{z'} P(z')P(f|z')P(t|z')} \Lambda_{(f,t,z)},$$  \hspace{1cm} (7)

where $\Lambda = \exp\{-\Lambda\}$, $\exp\{1\}$ is an element-wise exponential function, and $\Lambda \in \mathbb{R}^{N_f \times N_s} \times N_s$ represents the entire set of real-valued painting annotations. In this case, $\Lambda$ 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-
Table 1. SDR (in dB) for supervised and semi-supervised separation with/without refinement vs. ideal mask results.

<table>
<thead>
<tr>
<th>Example</th>
<th>Ideal</th>
<th>Supervised</th>
<th>Semi-Supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell</td>
<td>30.7</td>
<td>29.2 / 27.6</td>
<td>28.4 / 06.5</td>
</tr>
<tr>
<td>Drum</td>
<td>14.8</td>
<td>09.7 / 08.5</td>
<td>07.7 / 03.9</td>
</tr>
<tr>
<td>Cough</td>
<td>15.8</td>
<td>14.0 / 12.5</td>
<td>12.0 / 10.5</td>
</tr>
<tr>
<td>Piano</td>
<td>26.1</td>
<td>26.0 / 21.6</td>
<td>14.9 / 08.4</td>
</tr>
<tr>
<td>Siren</td>
<td>27.8</td>
<td>23.8 / 18.9</td>
<td>21.0 / 19.9</td>
</tr>
</tbody>
</table>

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. (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


