Blind Audio Source Separation Pipeline and Algorithm Evaluation

Wisam Reid, Kai-Chieh Huang & Doron Roberts-Kedes

Abstract—This report outlines the various methods and experiments employed by its authors in their search for algorithmic solutions for Blind Audio Source Separation. In the context of music, advancing BSS would lead to improvements in music information retrieval, computer music composition, spatial audio, and audio engineering. An understanding and evaluation on the advantage and disadvantage of different BSS algorithms will be beneficial for further usage of these BSS algorithm in different context. This report discusses three Blind Audio Source Separation Algorithms: GMM, NMF, and ICA, and evaluates their performance based on human perception of audio signals.

Index Terms— Audio Signal Processing, Blind Source Separation, Bark Coefficient Analysis, Non-negative Matrix Factorization, Gaussian Mixture Model, Critical Band Smoothing

1 Project Background

Blind Source Separation (BSS) is the separation of a set of source signals from a set of mixed signals, without the aid of information (or with very little information) about the source signals or the mixing process. One way of categorizing these algorithms is dividing them into the approach in time domain and frequency domain. An example of time domain approach is the Independent Component Analysis which works with input data that contains both positive and negative values. On the other hand, algorithms such as Non-negative Matrix Factorization works in the frequency domain where the input data is the magnitude of the spectrogram which can only be positive. Although BSS is an active research area where new techniques are continuously being developed, there is little literature studying the characteristics and the differences between different BSS algorithms and having an objective way of measuring the performance of the BSS algorithm in the context of human perception. Thus, it would be interesting to study the differences of BSS algorithms and compare them by evaluating their separation results. In this report, we studied three major BSS algorithms: Gaussian Mixture Model (GMM), Non-negative Matrix Factorization (NMF), and Independent Component Analysis (ICA) and examine their performance along with a perceptually relevant criteria for measuring the error, thus enabling regression on the parameters of our model.

2 Pipeline

With the goal of comparing the performance of each algorithm used in this paper in a convenient fashion, we designed a pipeline structure as shown in **Fig 1**. Performance of BSS supervised and unsupervised algorithms for both under-determined and determined systems, was compared using this process. Using this structure, we've implemented a system that automatically generates mixings and feeds the mixings into different algorithms to evaluate and compare their performance.

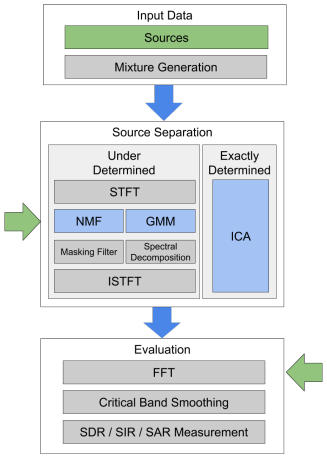

Fig. 1. Pipeline for Performing Algorithm Evaluation

3 NMF

Source separation can be viewed as a matrix factorization problem, where the source mixture is modeled as a matrix containing it's spectrogram representation. Spectrograms are commonly used to visualize the time varyingspectral density of audio, and other time domain signals [1].Audio signals can therefore be fully represented by a matrix with rows, columns, and element values corresponding to the horizontal axis t (representing time), the vertical axis f (representing frequency), and the intensity or color of each point in the image (indicating the amplitude of a particular frequency at a particular time) of a spectrogram respectively. The spectrogram of a signal x(t) can be estimated by computing the squared magnitude of the short-time fourier

transform (STFT) of the signal x(t), and likewise, x(t) can be recovered from the spectrogram through the inverse short-time fourier transform (ISTFT) after processing the signal in spectral domain [2].

3.1 Modeling Source Separation as an NMF

Adopting this view, it follows that source separation could be achieved by factorizing spectrogram data as a mixture of prototypical spectra. While there are many commonly practiced matrix factorization techniques such as Singular Value Decomposition (SVD), Eigenvalue Decomposition, QR Decomposition (QR), Lower Upper Decomposition (LU). NMF is a matrix factorization that assumes everything is non-negative, giving this technique an advantage when processing magnitude spectrograms. As an added advantage non-negativity avoids destructive interference, guaranteeing that estimated sources must cumulatively add during resynthesis. In general NMF decomposes a matrix as a product of two or more matrices as follows [3]

1) $V \in R_+^{F \times T}$ original non-negative data
2) $W \in R_+^{F \times K}$ matrix of basis vectors, dictionary elements
3) $H \in R_+^{K \times T}$ matrix of activations, weights, or gains

In the form:

$$[\quad \mathrm{V} \quad] \approx [\quad \mathrm{W} \quad] [\quad \mathrm{H} \quad]$$

Typically $K < F < T$ and K is chosen such that $FK + KT \ll FT$, hence reducing dimensionality [4]. In the context of source separation spectrogram data is modeled as V. The columns of V are approximated as a weighted sum (or mixture) of basis vectors W representing prototypical spectra and H representing time onsets or envelopes. NMF is underlaid by a well-defined statistical model of superimposed gaussian components and is equivalent to maximum likelihood estimation of variance parameters. NMF can accommodate regularization constraints on the factors through Bayesian priors. In particular, inverse-gamma and gamma Markov chain priors. Estimation can be carried out using a generalized expectation-maximization. This can also be solved as a minimization of D where D is a measure of divergence. In the literature, factorization is usually framed as an optimization problem $\min_{W,H \geq 0} D(V|WH)$ [4] Commonly solved using Euclidean or the generalized Kullback-Leibler (KL) divergence. Defined as

$$d_{KL}(x|y) = x\,log\frac{x}{y} - x + y \quad \text{giving}$$

$$\min_{W,H \geq 0} \sum_{f,t} V_{ft}\,log\frac{V_{ft}}{(WH)_{ft}} - V_{ft} + (WH)_{ft}$$

The former is convex in W and H separately, but is not convex in both simultaneously. NMF does not always give an intuitive decomposition, however explicitly controlling the sparseness and smoothness of the representation leads to representations that are parts-based and match the intuitive features of the data [5]. A deeper intuition is needed for how regularization techniques relate to the performance of these algorithms on audio data.

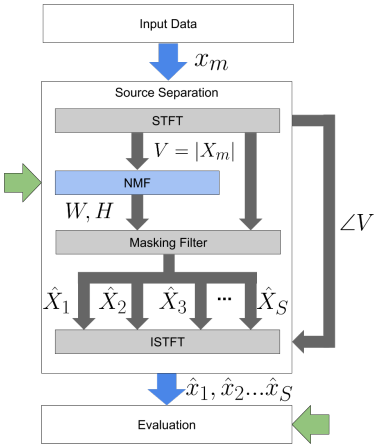

Fig. 2. A closer look at Non-Negative Matrix Factorization

Euclidean and KL divergence are both derived from a greater class of β-divergence algorithms, while it should be noted that the derivative of $d_\beta(x|y)$ with regard to y is continuous in β, KL divergence and the Euclidean distance are defined as $(\beta = 1)$ and $(\beta = 2)$ respectively. This is noteworthy since factorizations obtained with $\beta > 0$ will rely more heavily on the largest data values and less precision is to be expected in the estimation of the low-power components. This makes KL-NMF especially suitable for the decomposition of audio spectra, which typically exhibit exponential power decrease along frequency f and also usually comprise low-power transient components such as note attacks together with higher power components such as tonal parts of sustained notes [6]. Majorization-minimization (MM) can be performed using block coordinate descent, where H is optimize for a fixed W, then W is optimize for a fixed H, this is then repeated until convergence. Since solving a closed form solution is intractable, this is solved using Jensen's inequality, introducing the weights $\sum_k \phi_{ijk} = 1$ which gives $D(V|WH)$

$$\leq \sum_{f,t}(-V_{ft}\sum_k \phi_{ftk} log\frac{V_{ft}}{\phi_{ftk}} + \sum_k (W_{fk}H_{kt}))$$

Choosing ϕ_{ftk} to be $\frac{W_{fk}H_{kt}^{\ell}}{\sum_k W_{fk}(H)_{kt}^{\ell}}$ as suggested in [7] MM updates can be derived, where majorization is achieved by calculating ϕ_{ftk} and minimization is achieved by

$$\underset{W,H \geq 0}{\text{minimize}} \; -\sum_{f,t} V_{ft}\sum_k \phi_{ftk} log W_{fk}H_{kt} + \sum_k W_{fk}H_{kt}$$

As mentioned earlier, the MM estimation is equivalent to a generalized EM estimation with the added bene-

fit of accommodating regularization constraints on the factors through Markov chain priors. This is intuitive since the EM algorithm is a special case of MM, where the minorizing function is the expected conditional log likelihood. This approach stems from the fact that only V_{ft} is observed, but the full model involves unobserved variables k. EM is used to fit parameters that maximize the likelihood of the data, maximizing over $p(k|t)$ and $p(f|k)$ gives EM updates where the E-step involves calculating

$$P(k|f,t) = \frac{P(k|t)P(f|k)}{\sum_k P(k|t)P(f|k))}$$

and an M-Step maximizing

$$\sum_{f,t} V_{ft} \sum_k P(k|f,t) log P(k|t)P(f|k)$$

Unfortunately the number of parameters that need to be estimated is $FK + KT$, in such a high dimensional setting it is useful to impose additional structure. This can be done using priors and regularization. Priors can encode structural assumptions, like sparsity. Commonly the posterior distribution is calculated using the posterior mode (MAP). Another way is to impose structure is through regularization, by adding another term to the objective function

$$\underset{W,H \geq 0}{\text{minimize}} D(V||WH) + \lambda \Omega(H)$$

where Ω encodes the desired structure, and λ controls the strength [7]. As discussed earlier, sparsity and smoothness are good choices for $\Omega(H)$, these structures are useful when encoding the transient features of common audio sources. In addition, an interesting area of future work could be the inclusion of GMM derived structure through clustering. The beginnings of this research are discussed further in this report.

3.2 Seperating Sources

In the current research NMF is used to estimate W and H which are in turn used to derive masking filters $M_s$as seen in the signal flow in **Fig 2**. Here audio source mixtures are first represented as spectrograms by a STFT algorithm, then factorized into W and H in order to derive masking filters used to extract estimated sources $\hat{X}$ where $\hat{X} = M_s \circ V$ and $\circ$ is the Hadamard product (an element-wise multiplication of the matrices). These estimates are then synthesized into time domain signals via a ISTFT algorithm, the original phase components $\angle V$ are added back into the estimates, and passed down to the next stage of the pipeline for evaluation. Unsupervised, partially supervised, and fully supervised algorithms were evaluated, using this method NMF's separation and source estimation performance where compared to ICA and GMM algorithms against the performance criteria outline later in this report.

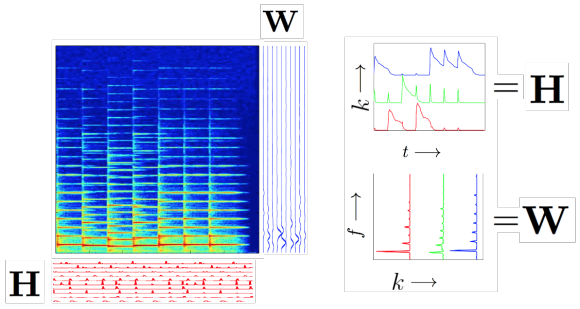

Fig. 3. Decomposing Spectrograms

4 GMM

N sources are separated from a mixed signal by fitting a Gaussian Mixture Model (GMM) with N components on the signal's magnitude spectrogram. Each bin in the original spectrogram is assigned to one of N GMM components. Spectrograms containing the bins assigned to each GMM component are inverted to produce estimations of the source signals. The spectrogram of the mixed signal was generated using a short-time Fourier transform (STFT). The STFT was computed with a 2048 sample kaiser window with a beta value of 18, a hop size of 64 samples, and zero-padding by a factor of 2. These parameters were chosen to minimize spectral leakage while providing extremely high resolution in both the time and frequency domain. The top most graph in **Fig 4** shows the spectrogram of a mixed signal, where the magnitude is in decibel scale. All data points with a magnitude less than -40db were discarded. The threshold value of -40db was chosen experimentally as the value that most effectively disambiguated between silence and acoustic events.

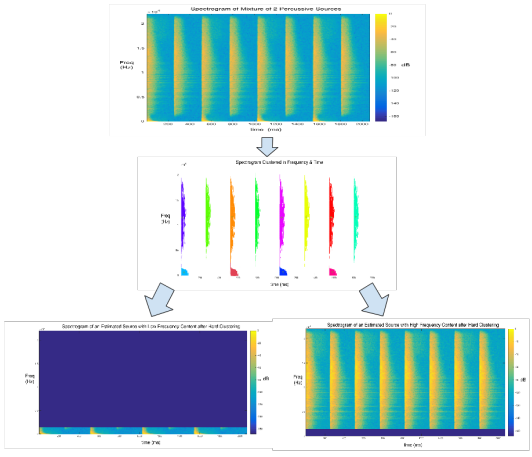

Fig. 4. GMM Source Separation Process

4.1 Clustering

After thresholding data based on magnitude, the magnitude feature was discarded from the data used to fit the GMM. Fitting a GMM on data without magnitude consistently outperformed the GMM fit on data with magnitude included. This is not surprising since the time and frequency of a bin are much more indicative of source signal membership than magnitude. The GMM was fit to the data with an equal number of components as source signals to be estimated. Each component was fit with a full covariance (non-diagonal) matrix independently of the other components. A full covariance matrix reflects the highly unpredictable nature of the covariance of time and frequency in audio sources. The independence of each components covariance matrix reflects tendency for some audio events to be highly concentrated in time or frequency, while other audio events are more dispersed in time and frequency. Initial values for the components were selected using the k-means algorithm.

4.2 Source Estimation

After fitting the GMM to the data, each bin in the positive frequency portion of the original mixed signal spectrogram underwent hard assignment to the component that maximized the posterior probability. The middle graph in **Fig 4** shows the result of this assignment after fitting a GMM using both frequency and time information. A new spectrogram was generated for each component consisting only of the bins in the original spectrogram that were assigned to the component. Finally, in the bottom of **Fig 4** shows the spectrograms of two estimated sources - the first having predominately low, narrow frequency content, and the second having predominately high, disperse frequency content. Each of these spectrograms was inverted using an inverse short-time Fourier transform to obtain estimates of the original source signals. The evaluation result of this algorithm is discussed in the end of this paper.

5 Performance Evaluation

In order to fully understand and compare the performance of each algorithm used in this paper, it is important to have an objective way of measuring the estimated source result against its true source. Several methods commonly used to evaluate audio quality such as PEAQ [8] are particular tailored to measure audio codec performance and are in consequence not ideal for evaluating audio source separation algorithms. One set of metrics used to measure the performance in the literature of BSS that particular suited for studying the characteristics of different BSS algorithms are Source to Distortion Ratio (SDR), Source to Interference Ratio (SIR), and Sources to Artifact Ratio (SAR) [9]. These ratio are derived based on decomposing the estimated source into true source part plus error term corresponding to the interference of other sources and the error term of artifact introduced by the algorithm, and then calculate the relative energy of these component in time domain. This evaluation technique hence has the advantage of providing information on how well a particular BSS algorithm suppresses the interference of other sources or the artifact introduced while performing the separation. However, the relationship between human perception on the quality of separated results to these metrics is not well established. Accordingly, we proposed an improved performance evaluation approach which relates human perception on audio signals to the metrics based on modifying the method proposed in [9].

5.1 Critical Band Smoothing

Since human hearing is only sensitive to spectral features wider in frequency than a critical bandwidth, we can model how human perceive audio signal by blurring spectral features smaller than a critical bandwidth in the spectrum using critical band smoothing procedure [10]. In this paper, the equivalent rectangular bandwidth (ERB) scale [11] is used for determining the critical band. The ERB and frequency in kHz are related by the equations $b_E = 21.4log_{10}(4.37f + 1.0)$ and $f = (10^{\frac{b_E}{21.4}} - 1.0)/4.37$. With this relationship in hand, we can smooth the spectral features of a certain audio signal by replacing the magnitude of each frequency bin with its average magnitude across one critical bandwidth using the calculation below:

$$P(\omega, \beta) = \sum_{\varphi = f(b(\omega) - \frac{\beta}{2})}^{f(b(\omega) + \frac{\beta}{2})} |H(\varphi))|^2 \quad (1)$$

where we are using a β = 0.5 in this paper. The effect of critical band smoothing can be understood through the graph in **Fig 5**.

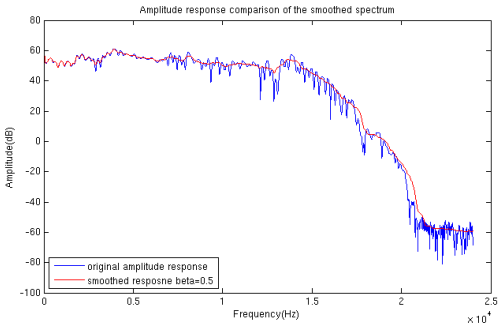

Fig. 5. Critical Band Smoothed Spectrum Example

5.2 Proposed Performance Criteria

In this article, a new performance criteria is designed to study the performance among different BSS algorithms in the context of human spectral perception. For the evaluation of the BSS algorithms in this paper,

the estimated source is decomposed into three parts as discussed previously in [9], but instead of measuring the energy ratio in time domain, we examine the energy ratio in critical band smoothed frequency domain to include the objectiveness of human hearing perception. For instance, the critical band smoothed spectrum of the estimated source is decomposed as $S_{estimate} = S_{true} + S_{interfere} + S_{artifact}$ where, $S_{estimate}$ is the critical band smoothed spectrum of the estimated source, S_{true} is the projection of the critical band smoothed spectrum of the estimated source onto the critical band smoothed spectrum of the targeted true source, $S_{interfer}$ is the summation of all the projections of the critical band smoothed spectrum of the estimated source onto all other true sources (excluding the targeted true source), and finally $S_{artifact}$ is calculated as $S_{estimate}$ - (S_{true} + $S_{interfer}$) which stands for additional error or artifact introduced by the BSS algorithm. The three metrics SDR, SIR, and SAR in this paper is defined as follows:

$$SDR = 10log_{10}(\frac{\|S_{true}\|^2}{\|S_{interfere} + S_{artifact})\|^2}) \quad (2)$$

$$SIR = 10log_{10}(\frac{\|S_{true}\|^2}{\|S_{interfere}\|^2}) \quad (3)$$

$$SAR = 10log_{10}(\frac{\|S_{true} + S_{interfere}\|^2}{\|S_{artifact}\|^2}) \quad (4)$$

where S_{true}, $S_{interfere}$, and $S_{artifact}$ are the critical band smoothed spectrums as stated previously. It is critical to note that SAR stands for Sources to Artifact Ratio and in the numerator: $\|S_{true} + S_{interfer}\|^2$ is the total energy of all the sources present in the estimated source. This arrangement makes SAR independent of SIR and make a robust and accurate evaluation on the artifact caused by the BSS algorithm. An evaluation test on an estimated source generated by a true source adding more and more white noise using the new established performance measurement as a demonstration is presented in **Fig 6**.

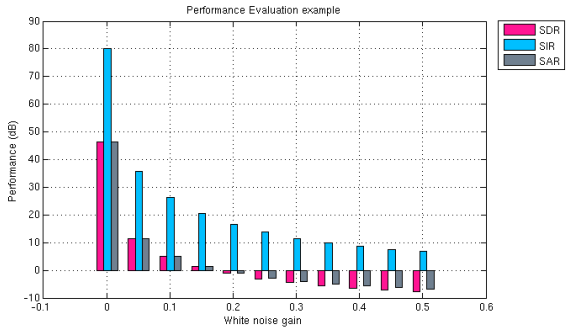

Fig. 6. Performance Evaluation Demonstration

6 Result & Conclusion

Through the algorithm evaluation pipeline, we have successfully generated the SDR, SIR, and SAR comparison

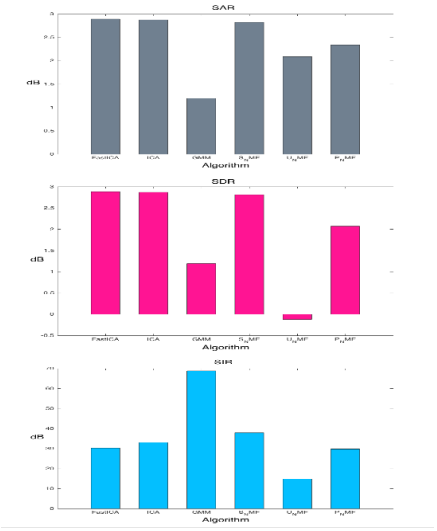

Fig. 7. Fast ICA, ICA, GMM, Supervised NMF, Unsupervised NMF, and Partially-Supervised NMF

among GMM, NMF, and ICA by testing on the same mixing generated using two sources: a Bass and a Drum audio clip. As mentioned in the performance evaluation section, SDR measures the general accuracy of how well the estimated source is compared to its targeted true source. SIR on the other hand, measures how well the algorithm prevent other sources from interfering the estimated source while separating. Finally, SAR evaluate how well the algorithm is at avoiding artifact. As is clear in **Fig 7**, fitting GMMs with spectrograms of mixed audio signals yielded the highest SIR measured, but yielded the lowest SDR except for unsupervised non-negative matrix factorization. These results are likely due to the sharp division in the frequency domain between source signal estimations created by hard assignment of spectral data points to GMM components. The results of GMM in **Fig 4** show a typical cutoff in the frequency domain. The abrupt cutoff in frequency prevents source signals in different frequency bands from interfering with the source being estimated. This suppression leads to GMMs high SIR value. However, a negative consequence of the sharp frequency cutoff is that signal content on the wrong side of the cutoff is excluded from the source estimation entirely. This suppression leads to a low SDR value. On the contrary, Fast ICA, ICA and NMF have similar SDR, SAR performance which are the highest among all the algorithm we've evaluated. There is however, a major difference among ICA and NMF, where ICA performs well in a determined system when there is sufficient mixing examples, while NMF performs well in an underdetermined system but requires a supervised learning process on the true sources examples. Furthermore, we also evaluate the performance on the unsupervised and partially supervised version of NMF. As we can see from the result in **Fig 7**, the performance of NMF drops according to the degree it is unsupervised.

REFERENCES

[1] Boualem Boashash. *Time Frequency Analysis*. Elsevier Science, 2003.

[2] E. Jacobsen and R. Lyons. The sliding dft. *Signal Processing Magazine, IEEE*, 20(2):74–80, Mar 2003.

[3] 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 (WASPAA)*, pages 177–180, October 2003.

[4] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural Comput.*, 21(3):793–830, March 2009.

[5] Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. *CoRR*, cs.LG/0408058, 2004.

[6] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *CoRR*, abs/1010.1763, 2010.

[7] Judith C. Brown Paris Smaragdis. Non-negative matrix factorization for polyphonic music transcription. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, October 19-22 2003.

[8] Treurniet W. Bitto R. Schmidmer C. Sporer T. Beerends J. Colomes C. Keyhl M. Stoll G. Brandenburg K. Feiten B. Thiede, T. Peaqthe itu standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, 48(1/2):3–29, January/February 2000.

[9] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1462–1469, July 2006.

[10] Jonathan S. Abel and David P. Berners. Signal processing techniques for digital audio effects. pages 273–280, Spring 2011.

[11] B. C. J. Moore and B. R. Glasberg. A revision of zwicker's loudness model. *Acta Acustica*, 82:335–345, Spring 1996.