SIMULTANEOUS SEPARATION AND SEGMENTATION IN LAYERED MUSIC

Prem Seetharaman
Northwestern University
prem@u.northwestern.edu

Bryan Pardo
Northwestern University
pardo@northwestern.edu

ABSTRACT

In many pieces of music, the composer signals how individual sonic elements (samples, loops, the trumpet section) should be grouped by introducing sources or groups in a layered manner. We propose to discover and leverage the layering structure and use it for both structural segmentation and source separation. We use reconstruction error from non-negative matrix factorization (NMF) to guide structure discovery. Reconstruction error spikes at moments of significant sonic change. This guides segmentation and also lets us group basis sets for NMF. The number of sources, the types of sources, and when the sources are active are not known in advance. The only information is a specific type of layering structure. There is no separate training phase to learn a good basis set. No prior seeding of the NMF matrices is required. Unlike standard approaches to NMF there is no need for a post-processor to partition the learned basis functions by group. Source groups are learned automatically from the data. We evaluate our method on mixtures consisting of looping source groups. This separation approach outperforms a standard clustering NMF source separation approach on such mixtures. We find our segmentation approach is competitive with state-of-the-art segmentation methods on this dataset.

1. INTRODUCTION

Audio source separation, an open problem in signal processing, is the act of isolating sound producing sources (or groups of sources) in an audio scene. Examples include isolating a single person's voice from a crowd of speakers, the saxophone section from a recording of a jazz big band, or the drums from a musical recording [13].

A system that can understand and separate musical signals into meaningful constituent parts (e.g. melody, backing chords, percussion) would have many useful applications in music information retrieval and signal processing. These include melody transcription [18], audio remixing [28], karaoke [21], and instrument identification [8].

Many approaches have been taken to audio source separation, some of which take into account salient aspects of musical structure, such as musical scores, or pitch (see Section 2). Few algorithms have explicitly learned musical structure from the audio recording (using no prior learning and no musical score) and used it to guide source discovery and separation. Our approach is designed to leverage compositional structures that introduce important musical elements one by one in *layers*. In our approach, separation alternates with segmentation, simultaneously discovering the layering structure and the functional groupings of sounds.

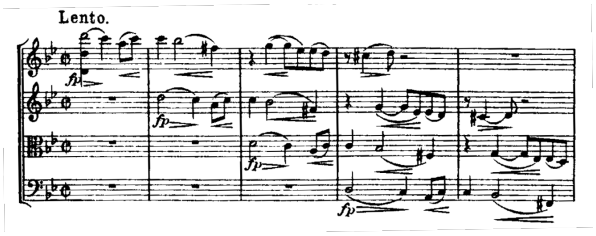

Figure 1. An exemplar layering structure in classical music - *String Quartet No. 1, Op. 27, Mvt IV, Measures 1-5.* by *Edvard Grieg*. The instruments enter one at a time in a layering structure, guiding the ear to both the content and the different sources.

In a layered composition, the composer signals how individual sound sources (clarinet, cello) or sonic elements (samples, loops, sets of instruments) should be grouped. For example, often a song will start by introducing sources individually (e.g. drums, then guitar, then vocals, etc) or in groups (the trumpet section). Similarly, in many songs, there will be a "breakdown", where most of the mixture is stripped away, and built back up one element at a time. In this way, the composer communicates to the listener the functional musical groups (where each group may consist of more than one source) in the mixture. This layering structure is widely found in modern music, especially in the pop and electronic genres (Figure 2), as well as classical works (Figure 1).

We propose a separation approach that engages with the composer's intent, as expressed in a layered musical structure, and separates the audio scene using discovered functional elements. This approach links the learning of the segmentation of music to source separation.

We identify the layering structure in an unsupervised manner. We use reconstruction error from non-negative matrix factorization (NMF) to guide structure discovery.

Reconstruction error spikes at moments of significant sonic change. This guides segmentation and also lets us know where to learn a new basis set.

Our approach assumes nothing beyond a layering structure. The number of sources, the types of sources, and when the sources are active are not known *a priori*. In parallel with discovering the musical elements, the algorithm temporally segments the original music mixture at moments of significant change. There is no separate training phase to learn a good basis set. No prior seeding of the NMF matrices is required. Unlike standard NMF there is no need for a post-processor that groups the learned basis functions by source or element [9] [25]. Groupings are learned automatically from the data by leveraging information the composer put there for a listener to find.

Our system produces two kinds of output: a temporal segmentation of the original audio at points of significant change, and a separation of the audio into the constituent sonic elements that were introduced at these points of change. These elements may be individual sources, or may be groups (eg. stems, orchestra sections).

We test our method on a dataset of music built from commercial musical loops, which are placed in a layering structure. We evaluate the algorithm based on separation quality, as well as segmentation accuracy. We compare our source separation method to standard NMF, paired with a post processer that clusters the learned basis set into groups in a standard way. We compare our segmentation method to the algorithms included in the Musical Structure Analysis Framework (MSAF) [16].

The structure of this paper is as follows. First, we describe related work in audio source separation and music segmentation. Then, we give an overview of our proposed separation/segmentation method, illustrated with a real-world example. We then evaluate our method on our dataset. Finally, we consider future work and conclude.

2. RELATED WORK

2.1 Music segmentation

A good music segmentation reports perceptually relevant structural temporal boundaries in a piece of music (e.g. verse, chorus, bridge, an instrument change, a new source entering the mixture).

A standard approach for music segmentation is to leverage the self-similarity matrix [7]. A novelty curve is extracted along the diagonal of the matrix using a checkerboard kernel. Peak picking on this novelty curve results in a music segmentation. The relevance of this segmentation is tied to the relevance of the similarity measure.

[12] describes a method of segmenting music where frames of audio are labeled as belonging to different states in a hidden Markov model, according to a hierarchical labeling of spectral features. [10] takes the self-similarity matrix and uses NMF to find repeating patterns/clusters within it. These patterns are then used to segment the audio. [17] expands on this work by adding a convex constraint to NMF.

[22] infers the structural properties of music based on structure features that capture both local and global properties of a time series, with similarity features.

The most similar work to ours is [27], which uses shift invariant probabilistic latent component analysis to extract musical riffs and repeated patterns from a piece of music. The activation of these recurring temporal patterns is used to then segment the audio. Our approach takes into account temporal groupings when finding these patterns, whereas their approach does not.

Our proposed method uses the reconstruction error of a source model over time in a musical signal in order to find structural boundaries. We explicitly connect the problem of music segmentation with the problem of audio source separation and provide an alternative to existing approaches to finding points of significant change from the audio.

2.2 Source separation

There are several source separation methods that leverage high-level musical information in order to perform audio source separation.

Separation from repeating patterns: REPET [21] separates the repeating background structure (e.g. bass, backing chords, rhythm from the guitar, and drums in a band) from a non-repeating foreground (e.g. lead singer with a melody) by detecting a periodic pattern (e.g. a sequence of chords that repeats every four bars). While REPET models the repeating background structure as a whole, our proposed method models the individual musical elements implicit in the composer's presentation of the material and does not require a fixed periodic pattern.

In [20], source models are built using a similarity matrix. This work looks for similar time frames anywhere in the signal, using a similarity measure (cosine similarity) that does not take musical grouping or temporal structure into account. Our method leverages the temporal groupings created by the composer's layering of sonic elements.

Informed separation: [4] incorporates outside information about the musical signal. A musical score gives information about the pitch and timing of events in the audio and is commonly used for informed musical separation. First, [4] finds an alignment between the low-level audio signal and the high-level musical score (in MIDI form). The pitch and timing of the event are then used to perform audio separation. These score-informed approaches are elaborated on in [6]. Our approach, does not require a score. Musical elements are discovered, modelled, and separated from the mixture using only the mixture, itself.

Non-negative matrix factorization (NMF). Our work uses NMF which was first proposed for audio source separation in [24]. Probabilistic latent component analysis (PLCA) can be seen as a probabilistic formulation of NMF, and is also used for source separation [23].

NMF finds a factorization of an input matrix $\mathbf{X}$ (the spectrogram) into two matrices, often referred to as spectral templates $\mathbf{W}$ and activations $\mathbf{H}$. Straightforward NMF has two weaknesses when used for source separation that will be elaborated on in Section 3: (1) there is no guar-

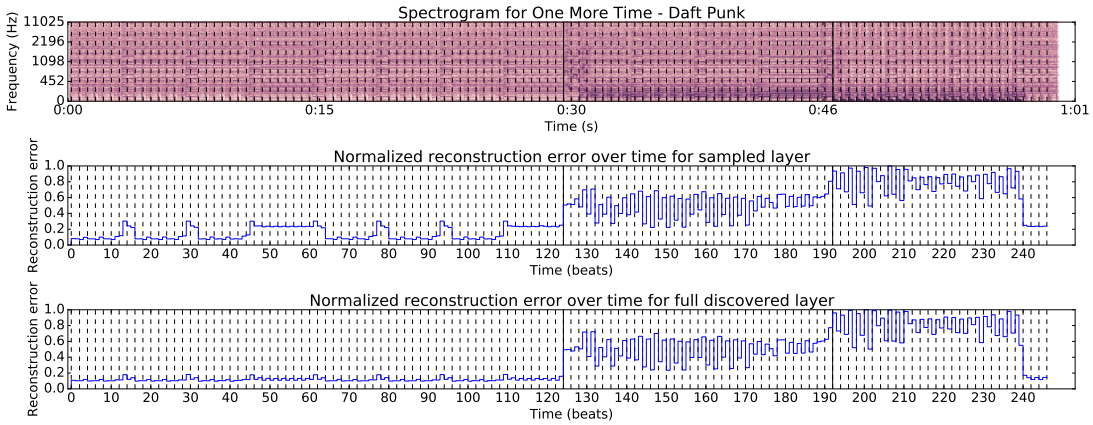

Figure 2. The top graph shows the spectrogram of 0:00 to 1:01 of *One More Time*, by *Daft Punk*. Ground truth segmentation is shown by the solid vertical black lines, where each line signals a new source starting. The middle graph shows the behavior of the reconstruction error of a sampled source layer over time ($\mathbf{e}$). When new layers begin, reconstruction error noticeably spikes and changes behavior. The bottom graph shows the reconstruction error over time for a full model of the first layer. Beats are shown by the vertical dashed black lines.

antee that an individual template (a column of $\mathbf{W}$) corresponds to only one source and (2) spectral templates are not grouped by source. Until one knows which templates correspond to a particular source or element of interest, one cannot separate out that element from the audio.

One may solve these problems by using prior training data to learn templates, or meta-data, such as musical scores [6], to seed matrices with approximately-correct templates and activations. User guidance to select the portions of the audio to learn from has also been used [2].

To group spectral templates by source without user guidance, researchers typically apply timbre-based clustering [9] [25]. This does not consider temporal grouping of sources. There are many cases where sound sources with dissimilar spectra (e.g. a high squeak and a tom drum, as in *Working in a Coal Mine* by DEVO) are temporally grouped as a single functional element by the composer. Such elements will not be grouped together with timbre-based clustering.

A non-negative Hidden Markov model (NHMM) [15] has been used to separate individual spoken voices from mixtures. Here, multiple sets of spectral templates are learned from prior training data and the system dynamically switches between template sets based on the estimated current state in the NHMM Markov model. A similar idea is exploited in [3], where a classification system is employed to determine whether a spectral frame is described by a learned dictionary for speech.

Our approach leverages temporal grouping created by composers in layered music. This lets us appropriately learn and group spectral templates without the need for prior training, user input or extra information from a musical score, or post-processing.

3. NON-NEGATIVE MATRIX FACTORIZATION

We now provide a brief overview of non-negative matrix factorization (NMF). NMF is a method to factorize a non-negative matrix $\mathbf{X}$ as the product of two matrices $\mathbf{W}$ and $\mathbf{H}$. In audio source separation, $\mathbf{X}$ is the power spectrogram of the audio signal, which is given as input. $\mathbf{W}$ is interpreted as a set of spectral templates (e.g. individual notes, the spectrum of a snare hit, etc.). $\mathbf{H}$ is interpreted as an activation matrix indicating when the spectral templates of $\mathbf{W}$ are active in the mixture. The goal is to learn this dictionary of spectral templates and activation functions. To find $\mathbf{W}$ and $\mathbf{H}$, some initial pair of $\mathbf{W}$ and $\mathbf{H}$ are created with (possibly random) initial values. Then, a gradient descent algorithm is employed [11] to update $\mathbf{W}$ and $\mathbf{H}$ at each step, using an objective function such as:

$$argmin_{\mathbf{W},\mathbf{H}}||\mathbf{X} - \mathbf{WH}||_F^2 \quad (1)$$

where $||\cdot||_F^2$ refers to the Frobenius norm. Once the difference between $\mathbf{WH}$ and $\mathbf{X}$ falls below an error tolerance, the factorization is complete.

There are typically many approximate solutions that fall below any given error bound. If one varies the initial $\mathbf{W}$ and $\mathbf{H}$ and restarts, a different decomposition is likely to occur. Many of these will not have the property that each spectral template (each column of $\mathbf{W}$) represents exactly one element of interest. For example, it is common for a single spectral template to contain audio from two or more elements of interest (e.g. a mixture of piano and voice in one template). Since these templates are the atomic units of separation with NMF, mixed templates preclude successful source separation. Therefore, something must be done to ensure that, after gradient descent is complete, each spectral template belongs to precisely one group or source of interest. An additional issue is that, to perform meaningful source separation, one must partition these spectral templates into groups of interest for separation. For example, if the goal is to separate piano from drums in a mixture of piano and drums, all the templates modeling the drums should be grouped together.

One can solve these problems by using prior training data, running the algorithm on audio containing an isolated element of interest to learn a restricted set of $\mathbf{W}$. One can repeat this for multiple elements of interest to separate audio from a mixture using these prior learned templates. This avoids the issues caused from learning the spectral templates directly from the mixture: one template having portions of two sources, and not knowing which templates belong to the same musical element. One may also use prior knowledge (e.g. a musical score) to seed the $\mathbf{W}$ and $\mathbf{H}$ matrices with values close to the desired final goal. We propose an alternative way of grouping spectral templates that does not require prior seeding of matrices [6] or user segmentation of audio to learn the basis set for each desired group [2], nor post-processing to cluster templates.

4. PROPOSED APPROACH

Our approach has four stages: estimation, segmentation, modeling, and separation. We cycle through these four stages in that order until all elements and structure have been found.

Estimation: We assume the composer is applying the compositional technique of layering. This means that estimating the source model from the first few audio frames will give us an initial model of the first layer present in the recording. Note that in our implementation, we beat track the audio [14] [5]. Beat tracking reduces the search space for a plausible segmentation, but is not integral to our approach.

We use the frames from the first four beats to learn the initial spectral dictionary. Consider two time segments in the audio, with i, j and k as temporal boundaries: $\mathbf{X} = [\mathbf{X}_{i:j-1}, \mathbf{X}_{j:k}]$. To build this model, we use NMF on a segment of $\mathbf{X_{i:j-1}}$ to find spectral templates $\mathbf{W_{est}}$.

Segmentation: Once an estimated dictionary $\mathbf{W_{est}}$ is found, we measure how well it models the mixture over time. Keeping $\mathbf{W_{est}}$ fixed, learn the activation matrix $\mathbf{H}$ for the second portion. The reconstruction error for this is:

$$error(\mathbf{W_{est}H}, \mathbf{X_{j:k}}) = ||\mathbf{X_{j:k}} - \mathbf{W_{est}H}||_F^2 \quad (2)$$

Equation 2 measures how well the templates in $\mathbf{W_{est}}$ model the input $\mathbf{X}$. For example, assume $\mathbf{W_{est}}$ was constructed a spectrogram of snare drum hits in segment $\mathbf{X_{i:j-1}}$. If a guitar and bass are added to the mixture somewhere in the range $j:k$,then the reconstruction error on $\mathbf{X_{j:k}}$ will be greater than the reconstruction error on $\mathbf{X_{i:j-1}}$. We use reconstruction error as a signal for segmentation. We slide the boundaries j, k over the the mixture, and calculate error for each of these time segments, as shown in Figure 2. This gives us $\mathbf{e}$, a vector of reconstruction errors for each time segment.

In the middle graph in Figure 2, we show reconstruction error over time, quantized at the beat level. Reconstruction error spikes on beats where the audio contains new sounds not modeled in $\mathbf{W_{est}}$. As layers are introduced by the artist, reconstruction error of the initial model rises considerably. Identifying sections of significant change in reconstruction error gives a segmentation on the music. In Figure 2, the segmentation is shown by solid vertical lines. At each solid line, a new layer is introduced into the mixture. Our method for identifying these sections uses a moving median absolute deviation, and is detailed in Algorithm 1.

Algorithm 1 Method for finding level changes in reconstruction error over time, where $\odot$ is the element-wise product. $\mathbf{e}$ is a vector where $\mathbf{e}(t)$ is the reconstruction error for $\mathbf{W_{est}}$ at time step t. lag is the size of a smoothing window for $\mathbf{e}$. p and q affect how sensitive the algorithm is when finding boundary points. We use $lag = 16$, $p = 5.5$ and $q = .25$ in our implementation. These values were found when training on a different dataset, containing 5 mixtures, than the one in Section 5.

```
lag, p, q ← initialize, tunable parameters
e ← reconstruction error over time for W_est
e ← (e ⊙ e)                          ▷ Element-wise product
d ← max(Δe/Δt)
for i from lag to length(e) do       ▷ Indexing e
    window ← e_{i−lag:i−1}
    m ← median(abs(window - median(window))
    if abs(e_i − median(window)) ≥ p ∗ m then
        if abs(e_i − e_{i−1}) ≥ q ∗ d then
            return i                 ▷ Boundary frame in X
        end if
    end if
end for
return length(e)                     ▷ Last frame in X
```

Modeling: once the segmentation is found, we learn a model using NMF on the entire first segment. This gives us $\mathbf{W_{full}}$, which is different from $\mathbf{W_{est}}$, which was learned from just first four beats of the signal. In Figure 2, the first segment is the first half of the audio. $\mathbf{W_{full}}$ is the final model used for separating the first layer from the mixture. As can be seen in the bottom graph of Figure 2, once the full model is learned, the reconstruction error of the first layer drops.

Separation: once the full model $\mathbf{W_{full}}$ is learned, we use it for separation. To perform separation, we construct a binary mask using NMF. $\mathbf{W_{full}}$ is kept fixed, and $\mathbf{H}$ is initialized randomly for the entire mixture. The objective function described in Section 3 is minimized only over $\mathbf{H}$. Once $\mathbf{H}$ is found, $\mathbf{W_{full}H}$ tells us when the elements of $\mathbf{W_{full}}$ are active. We use a binary mask for separation, obtained via:

$$\mathbf{M} = round(\mathbf{W_{full}H} \oslash max(\mathbf{W_{full}H}, abs(\mathbf{X})))$$

where $\oslash$ indicates element-wise division and $\odot$ is element-wise multiplication. We reconstruct the layer using:

$$\mathbf{X}_{layer} = \mathbf{M} \odot \mathbf{X} \quad (3)$$

$$\mathbf{X}_{residual} = (\mathbf{1} - \mathbf{M}) \odot \mathbf{X} \quad (4)$$

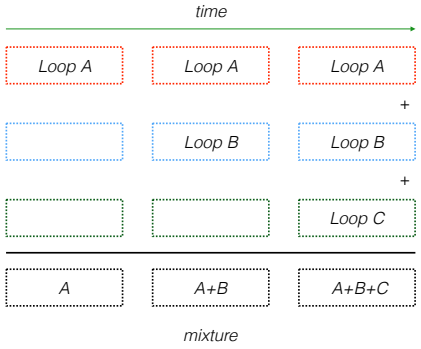

Figure 3. Construction of a single mixture using a layering structure in our dataset, from 3 randomly selected loops each from 3 sets A, B, and C.

where $\odot$ indicates element-wise product. $\mathbf{X}_{residual}$ is the mixture without the layer. We restart at the **estimation** stage above, this time using $\mathbf{X}_{residual}$ as the input, and setting the start point to the segmentation boundary found in the **segmentation** stage above. Taking the inverse Fourier transform of $\mathbf{X}_{layer}$ gives us the audio signal of the separated layer.

Termination: if $\mathbf{X}_{residual}$ is empty (no source groups remain in the mixture), we terminate.

5. EVALUATION

5.1 Dataset

We evaluate our approach in two ways: separation quality, and segmentation accuracy. To do this, we construct a dataset where ground truth is known for separation and segmentation. As our approach looks for a layering structure, we devise mixtures where this layering occurs. We obtain source audio from *Looperman* [1], an online resource for musicians and composers looking for loops and samples to use in their creative work. Each loop from *Looperman* is intended by its contributor to represent a single source. Each loop can consist of a single sound producing source (e.g. solo piano) or a complex group of sources working together (e.g. a highly varied drumkit).

From *Looperman*, we downloaded 15 of these loops, each 8 seconds long at 120 beats per minute. These loops are divided into three sets of 5 loops each. Set A contained 5 loops of rhythmic material (drum-kit based loops mixed with electronics), set B contained 5 loops of harmonic and rhythmic material performed on guitars, and set C contained 5 loops of piano. We arranged these loops to create mixtures that had layering structure, as seen in Figure 3.

We start with a random loop from set A, then add a random loop from B, then add a random loop from C, for a total length of 24 seconds. We produce 125 of these mixtures.

Ground truth segmentation boundaries are at 8 seconds (when the second loop comes in), and at 16 seconds (when the third loop comes in). In Figure 3, each row is ground truth for separation.

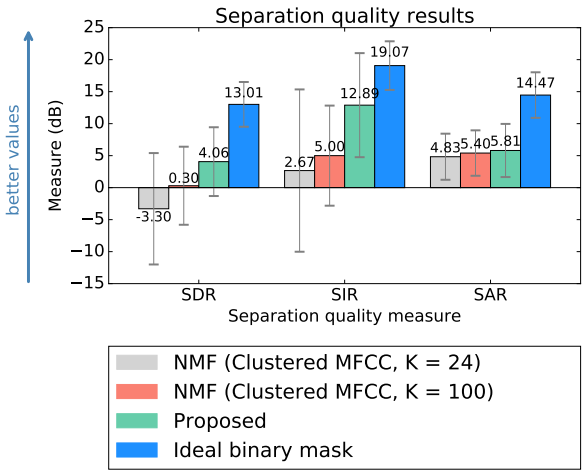

Figure 4. Separation performance of current and proposed algorithms, and an ideal binary mask. Higher numbers are better. The ideal binary mask is an upper bound on separation performance. The error bars indicate standard deviations above and below the mean.

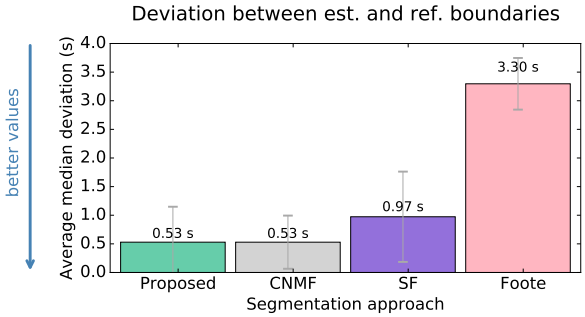

Figure 5. Segmentation performance of current and proposed algorithms. *Est. and ref.* refers to the median deviation in seconds between a ground truth boundary and a boundary estimated by the algorithm. Lower numbers are better. The error bars indicate standard deviations above and below the mean.

5.2 Methods for comparison

For separation, we compare our approach to a separation method in [25]. In this method, they use NMF on the entire mixture spectrogram, and then cluster the components into sources using MFCCs. Each cluster of components is then used to reconstruct a single source in the mixture. In our approach, the number of components (K) was fixed at $K = 8$, giving a total of $K = 24$ components for the entire mixture. For direct comparison, we give the method in [25] $K = 24$ components. We also look at the case where [25] is given $K = 100$ components.

For segmentation, we compare our approach with [22], [17], and [7].

Approach	Median deviation (s)	Avg # of segments
CNMF [17]	.53	6.152
SF [22]	.97	4.024
Foote [7]	3.3	3.048
Proposed	.53	3.216

Table 1. Segmentation results for various approaches. In the dataset, an accurate segmentation reports 3 segments. While CNMF reports similar average median deviation from estimated to reference boundaries to the proposed method, it finds almost twice the number of boundaries. Foote finds a number of segments closer to ground truth, but the boundaries are in the wrong place.

5.3 Results

5.3.1 Separation

To measure separation quality, we use the BSS Eval toolbox [26] as implemented in [19], which reports Source-to-Distortion (SDR), Source-to-Interference (SIR), and Source-to-Artifact (SAR) ratios. For all of these, we compare our proposed approach to an NMF clustering approach based on MFCCs in [25]. This clustering approach was given the number of sources to find in the mixture. This is in contrast to our algorithm, where the number of sources is unknown, and instead is discovered. We also compare to an ideal binary mask. Results are in Figure 4, which shows mean SDR, SIR, and SAR for different source separation methods.

As seen in Figure 4, our approach found sources that correlated with the target sources, giving SDR and SIR more comparable to the ideal binary mask. This is in contrast to the clustering approach, which found sources that poorly correlated with the actual target sources, resulting in low values for SDR and SIR, even when using more components than our approach ($K = 100$ vs. $K = 24$). The clustering mechanism in [25] leverages MFCCs, and finds sources that are related in terms of resonant characteristics (e.g. instrument types) but fails to model sources that have multiple distinct timbres working together.

Our results indicate that separation based on NMF reconstruction error is a useful signal to guide the grouping of spectral templates for NMF, and boost separation quality on layered mixtures.

5.3.2 Segmentation

To measure segmentation accuracy, we use the median absolute time difference from a reference boundary to its nearest estimated boundary, and vice versa. For both of these measures, we compare our proposed approach with [22], [17], and [7], implemented in MSAF [16], as shown in Figure 5.

We find that our approach is as accurate as existing state-of-the-art, as can be seen in Figure 5 and Table 1. Our results indicate that, when finding a segmentation of a mixture, in which segment boundaries are dictated by sources entering the mixture, current approaches are not sufficient. Our approach, because it uses reconstruction error of source models to drive the segmentation, finds more accurate segment boundaries.

6. CONCLUSIONS

We have presented a method for source separation and music segmentation which uses reconstruction error in non-negative matrix factorization to find and model groups of sources according to discovered layered structure. Our method does not require pre-processing of the mixture or post-processing of the basis sets. It requires no user input, or pre-trained external data. It bootstraps an understanding of both the segmentation and the separation from the mixture alone. It is a step towards a framework in which separation and segmentation algorithms can inform one another, for mutual benefit. It makes no assumptions on what a source actually is, but rather finds functional sources implied by a specific type of musical structure.

We showed that tracking reconstruction error of a source model over time in a mixture is a helpful approach to finding structural boundary points in the mixture. These structural boundary points can be used to guide NMF. This separation approach outperforms NMF that clusters spectral templates via heuristics. This work demonstrates a clear, novel, and useful relationship between the problems of separation and segmentation.

The principles behind this approach can be expanded to other source separation approaches. Since source separation algorithms rely on specific cues (e.g. repetition like in REPET, or a spectral model like in NMF), the temporal failure points of source separation algorithms (e.g. the repeating period has failed, or the model found by NMF has failed to reconstruct the mixture) may be a useful cue for music segmentation.

The approach presented here exploits the compositional technique of layering employed in many musical works. For future approaches, we would like to build separation techniques which leverage other compositional techniques and musical structures, perhaps integrating our work with existing work in segmentation.

7. ACKNOWLEDGEMENTS

This work is supported by National Science Foundation Grant 1420971.

8. REFERENCES

[1] Looperman. http://www.looperman.com.

[2] Nicholas J Bryan, Gautham J Mysore, and Ge Wang. Isse: An interactive source separation editor. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 257–266. ACM, 2014.

[3] Zhiyao Duan, Gautham J Mysore, and Paris Smaragdis. Online plca for real-time semi-supervised source separation. In *Latent Variable Analysis and Signal Separation*, pages 34–41. Springer, 2012.

[4] Zhiyao Duan and Bryan Pardo. Soundprism: An online system for score-informed source separation of music audio. *Selected Topics in Signal Processing, IEEE Journal of*, 5(6):1205–1215, 2011.

[5] Daniel PW Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, 36(1):51–60, 2007.

[6] Sebastian Ewert, Bryan Pardo, Mathias Muller, and Mark D Plumbley. Score-informed source separation for musical audio recordings: An overview. *Signal Processing Magazine, IEEE*, 31(3):116–124, 2014.

[7] Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 1, pages 452–455. IEEE, 2000.

[8] Toni Heittola, Anssi Klapuri, and Tuomas Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *ISMIR*, pages 327–332, 2009.

[9] Rajesh Jaiswal, Derry FitzGerald, Dan Barry, Eugene Coyle, and Scott Rickard. Clustering nmf basis functions using shifted nmf for monaural sound source separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 245–248. IEEE, 2011.

[10] Florian Kaiser and Thomas Sikora. Music structure discovery in popular music using non-negative matrix factorization. In *ISMIR*, pages 429–434, 2010.

[11] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[12] Mark Levy and Mark Sandler. Structural segmentation of musical audio by constrained clustering. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2):318–326, 2008.

[13] Josh H McDermott. The cocktail party problem. *Current Biology*, 19(22):R1024–R1027, 2009.

[14] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, 2015.

[15] Gautham J Mysore, Paris Smaragdis, and Bhiksha Raj. Non-negative hidden markov modeling of audio with application to source separation. In *Latent variable analysis and signal separation*, pages 140–148. Springer, 2010.

[16] O. Nieto and J. P. Bello. Msaf: Music structure analytis framework. In *The 16th International Society for Music Information Retrieval Conference*, 2015.

[17] Oriol Nieto and Tristan Jehan. Convex non-negative matrix factorization for automatic music structure identification. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 236–240. IEEE, 2013.

[18] Mark D Plumbley, Samer A Abdallah, Juan Pablo Bello, Mike E Davies, Giuliano Monti, and Mark B Sandler. Automatic music transcription and audio source separation. *Cybernetics &Systems*, 33(6):603–627, 2002.

[19] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. mir_eval: A transparent implementation of common mir metrics. *Proc. of the 15th International Society for Music Information Retrieval Conference*, 2014.

[20] Zafar Rafii and Bryan Pardo. Music/voice separation using the similarity matrix. In *ISMIR*, pages 583–588, 2012.

[21] Zafar Rafii and Bryan Pardo. Repeating pattern extraction technique (repet): A simple method for music/voice separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(1):73–84, 2013.

[22] Jean Serra, Mathias Muller, Peter Grosche, and Josep Ll Arcos. Unsupervised music structure annotation by time series structure features and segment similarity. *Multimedia, IEEE Transactions on*, 16(5):1229–1240, 2014.

[23] Madhusudana Shashanka, Bhiksha Raj, and Paris Smaragdis. Probabilistic latent variable models as nonnegative factorizations. *Computational intelligence and neuroscience*, 2008, 2008.

[24] Paris Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *Independent Component Analysis and Blind Signal Separation*, pages 494–499. Springer, 2004.

[25] Martin Spiertz and Volker Gnann. Source-filter based clustering for monaural blind source separation. In *Proceedings of International Conference on Digital Audio Effects DAFx09*, 2009.

[26] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1462–1469, 2006.

[27] Ron J Weiss and Juan P Bello. Unsupervised discovery of temporal structure in music. *Selected Topics in Signal Processing, IEEE Journal of*, 5(6):1240–1251, 2011.

[28] John F Woodruff, Bryan Pardo, and Roger B Dannenberg. Remixing stereo music with score-informed source separation. In *ISMIR*, pages 314–319, 2006.