Discovering Auditory Objects Through Non-Negativity Constraints

Paris Smaragdis

Mitsubishi Electric Research Laboratories
Cambridge MA, USA
paris@merl.com

Abstract

We present a novel method for discovering auditory objects from scenes in a self-organized manner. Our approach is using non-negativity constraints to find the building elements of a monaural auditory input. Surprisingly, although devoid of any statistical measures, this approach discovers independent elements in the scene similarly to previously reported methods employing ICA algorithms. The use of non-negativity constraints makes this work best suited for spectral magnitude analysis and provides a fairly robust method for discovery and extraction of auditory objects from scenes.

1. Introduction

Processing of magnitude spectra for the extraction of auditory objects has been a long standing practice. It is by now common knowledge that by visual inspection of time-frequency magnitude transforms it is possible to see individual objects and obtain an understanding of the structure of an auditory scene. Alas this visual inspection, although very successful using the researcher's eye, is rather challenging for a machine to perform automatically. This has resulted in the creation of a rich variety of algorithms that attempt to analyze audio using the structure in the time-frequency domain.

In this paper we will present a approach that gives accurate results through a surprisingly simple optimization process. This method is akin to the ICA approaches described in the past by Casey and Westner [2] and Smaragdis [8], it however lacks a statistical foundation and rather attempts to describe auditory scenes using a component-wise reconstruction approach. In this paper we will present two flavors of this technique, one dealing with static-spectrum objects, and an extension that can deal with time-varying objects.

2. Detection of objects using Non-Negative Matrix Factorization

In this section we will consider the discovery and extraction of static-spectrum objects from a scene. We use the term static-spectrum to imply an approximately constant, in time, spectral structure. This would be exemplified by the spectral characteristics of a constant frequency tone, or a piano note, or an impact sound. All of these examples might change slightly in time however a single spectrum can describe their structure concisely. To perform the object discovery we will employ the Non-Negative Matrix Factorization algorithm by Lee and Seung [5].

2.1. Non-Negative Matrix Factorization

Non-Negative Matrix Factorization (NMF), was introduced by Lee in Seung in a paper that described its application to face recognition and analysis [4]. Although its statistical underpinnings and reasoning behind its success is still an intense subject of research and speculation, it has been widely adopted as a very useful technique for linear decomposition and dimensionality reduction of non-negative data sets.

Its formulation is as follows. Given a non-negative $M \times N$ matrix $\mathbf{V} \in \mathbb{R}^{\geq 0, M \times N}$ the goal is to approximate it as a product of two non-negative matrices $\mathbf{W} \in \mathbb{R}^{\geq 0, M \times R}$ and $\mathbf{H} \in \mathbb{R}^{\geq 0, R \times N}$ where $R \leq M$, such that we minimize the error of reconstruction of $\mathbf{V}$ by $\mathbf{W} \cdot \mathbf{H}$. Lee and Seung [5] provided two cost functions by which we can measure the error or reconstruction. One is the norm of the difference between the input and the reconstruction, and the other is a mutated version of the Kullback-Leibler distance as applied to arbitrary functions. We will use the latter cost which is defined as:

$$D = \left\| \mathbf{V} \otimes \ln(\frac{\mathbf{V}}{\mathbf{W} \cdot \mathbf{H}}) - \mathbf{V} + \mathbf{W} \cdot \mathbf{H} \right\|_F \qquad (1)$$

where $\|\cdot\|_F$ is the Frobenius norm and $\otimes$ is the Hadamard product (an element-wise multiplication); the division is also element-wise. In order to optimize this cost function Lee and Seung [5] also provided a update algorithm which due to its multiplicative nature forgoes the need for non-negativity constraints during optimization (assuming the initial values of $\mathbf{W}$ and $\mathbf{H}$ are positive). The updates for the matrices $\mathbf{W}$ and $\mathbf{H}$ are defined as:

$$\mathbf{H} = \mathbf{H} \otimes \frac{\mathbf{W}^\top \cdot \frac{\mathbf{V}}{\mathbf{W} \cdot \mathbf{H}}}{\mathbf{W}^\top \cdot \mathbf{1}}, \quad \mathbf{W} = \mathbf{W} \otimes \frac{\frac{\mathbf{V}}{\mathbf{W} \cdot \mathbf{H}} \cdot \mathbf{H}^\top}{\mathbf{1} \cdot \mathbf{H}^\top} \qquad (2)$$

where $\mathbf{1}$ is a $M \times N$ matrix with all its elements set to unity, and the divisions are performed in an element-wise manner. These updates are applied iteratively until the two factors start converging to a constant solution. The variable R, which is the number of columns of $\mathbf{W}$ and the rows of

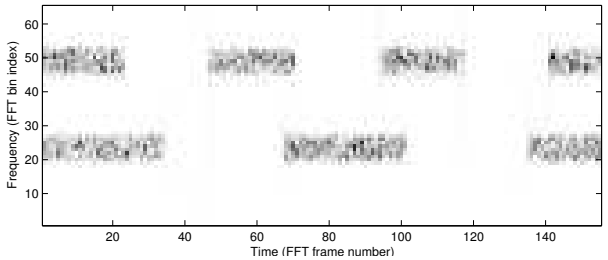

Figure 1: *Spectrogram of a scene composed of bandlimited noise bursts.*

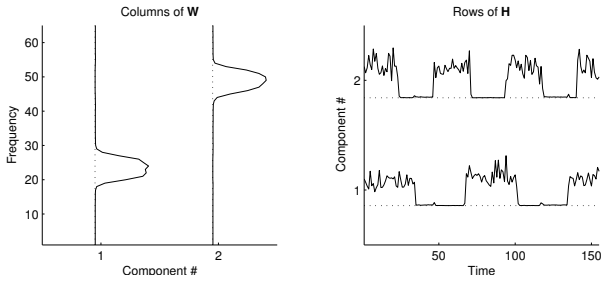

Figure 2: *The factors obtained by performing NMF on the spectrogram in figure 1.*

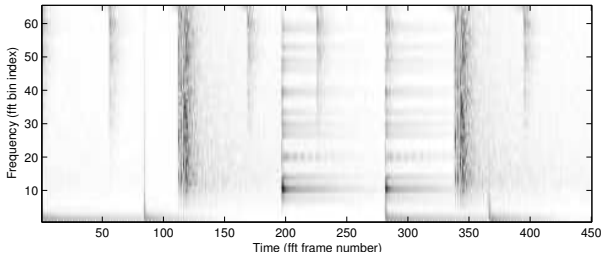

Figure 3: *The spectrogram of a drum loop. Four types of instruments can be visually identified.*

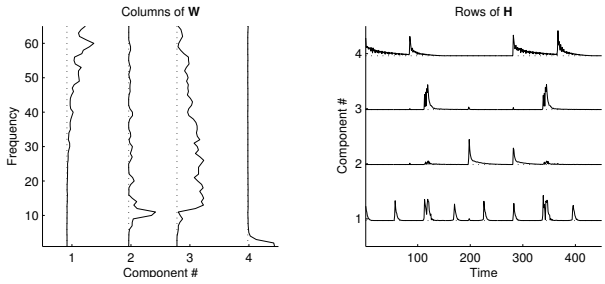

Figure 4: *NMF analysis of the spectrogram in figure 3. All four instruments are adequately described in both time and frequency by the columns of* $\mathbf{W}$ *and rows of* $\mathbf{H}$.

$\mathbf{H}$, determines the rank of the approximation. If $R = M$ we can obtain a complete reconstruction of the input, but as R is reduced we start obtaining low-rank approximations and notice that the elements of $\mathbf{W}$ and $\mathbf{H}$ start to reveal the structure of the input. The R columns of $\mathbf{W}$ tend to reveal the vertical structure of the input, and their corresponding R rows in $\mathbf{H}$ the horizontal structure. These pairs of columns and rows result into R linear models that will be describing the identified objects or components in the input (a more complete and intuitive description of these models and their ability to model objects follows in the next section).

Selecting R can be a complex procedure that requires estimating the dimensionality of the input matrix. Although various techniques for this estimation are available in the statistics literature, in this paper we will use prior knowledge of the structure of the input. If R is misestimated and is greater than the objects in the input scene, then some of these objects will be distributed between two or more NMF components. Interestingly enough this split is often intuitive, distributing sounds using distinct parts (e.g. a harmonic and an inharmonic component, or an impulse and a release part). If R is less than the optimal value, then objects are consolidated into NMF components and the results are not as significant.

2.2. Non-Negative Matrix Factorization applied on audio spectra

To illustrate the use of this approach in the audio domain consider the simple case of two distinct band-limited noise bursts in an auditory scene shown in figure 1.

In terms of the time-frequency composition of the scene we can say that the two components are a noise burst centered around frequency bin 50 which occurs four times, and a noise burst around frequency bin 23 which occurs three times. Using this magnitude spectrogram as the non-negative matrix input to NMF with $R = 2$ we obtain the results in figure 2.

In order to interpret the results let us examine the decomposition that NMF performs. The provided input spectrogram is a $M \times N$ matrix $\mathbf{V}$ representing the magnitude value for the M frequency bins at the N time frames. This is decomposed as a product of two matrices $\mathbf{V} \approx \mathbf{W} \cdot \mathbf{H}$. The matrix $\mathbf{W}$ is $M \times R$ and $\mathbf{H}$ is $R \times N$. Now consider the following interpretation of a matrix product, individually each of the columns of $\mathbf{W}$ is multipled with its corresponding row in $\mathbf{H}$ to produce a rank-1 approximation of $\mathbf{V}$. These approximations are then summed up to produce the final approximation of $\mathbf{V}$. Keeping this in mind we can see how the matrix $\mathbf{W}$ will encapsulate some of the vertical structure of $\mathbf{V}$ in its columns, whereas the matrix $\mathbf{H}$ will encapsulate the horizontal structure. Upon examination of the results in figure 2, we can see how this works with the noise burst example. The two columns of $\mathbf{W}$ describe the frequency structure of the two noise bursts (one centered around bin 50 and one around bin 23). Their corresponding rows in $\mathbf{H}$ describe the time evolution of each burst, one occurring four times and the other occurring thrice. Effectively this decomposition has revealed the structure of the input scene by describing its dominant elements in both frequency and time.

Now let us examine a more complex example where this type of analysis is applicable. Consider the spectrogram in

figure 3. This is a spectrogram from a drum loop consisting of four sounds; a bass drum (low frequency with four long instances), a snare drum (wideband with two instances), a cowbell (resonant harmonic structure, two instances), and a hi-hat (high-frequency wideband with eight instances). Although not entirely static, the spectra of these sounds are fairly constant in time, which means that we can apply the method developed so far. The results of this analysis with $R = 4$ are shown in figure 4. Note how the columns of $\mathbf{W}$ and rows of $\mathbf{H}$ describe the four instruments. The first column of $\mathbf{W}$ is a high-frequency wideband spectrum and the corresponding time envelope as described in the first row of $\mathbf{H}$ has eight peaks. These two describe the hi-hat instrument spectrally and temporally. Likewise, the second object has a resonant spectral structure and two time peaks (the cowbell), the third is wideband and has two instances in time (the snare drum) and the fourth is the bass drum, exhibiting a low frequency spectrum and four instances in time. Despite the fact that the four instruments were often overlapping in time and in frequency, NMF was successful in describing the scene in a semantically meaningful way.

This approach can be used with a variety of scenes, to extract information about its time and frequency composition. A notable case where it performs admirably under very complex conditions, is musical transcription of piano music as described by Smaragdis and Brown [9].

3. Detection of objects using Non-Negative Matrix Deconvolution

In this section we will introduce a new extension to NMF which will now allow us to deal with objects that have time-varying spectra. Unlike the previous approach where the auditory objects were described by a spectrum and its corresponding energy in time, this time we will consider a sequence of successive spectra and its corresponding energy across time. This will require a straightforward extension to the NMF update equations.

3.1. Non-Negative Matrix Deconvolution

NMF attempts to reconstruct a matrix $\mathbf{V}$ using a matrix product by $\mathbf{V} \approx \mathbf{W} \cdot \mathbf{H}$. In Non-Negative Matrix Deconvolution (NMD) we extend this expression to:

$$\mathbf{V} \approx \sum_{t=0}^{T-1} \mathbf{W}_t \cdot \overset{t\rightarrow}{\mathbf{H}} \tag{3}$$

where $\mathbf{V} \in \mathbb{R}^{\geq 0, M \times N}$ is the input we wish to decompose, and $\mathbf{W}_t \in \mathbb{R}^{\geq 0, M \times R}$ and $\mathbf{H} \in \mathbb{R}^{\geq 0, R \times N}$ are the two factors. The $\overset{i\rightarrow}{(\cdot)}$ operator is a shift operator that moves the columns of its argument by i spots to the right, such that:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}, \quad \overset{0\rightarrow}{\mathbf{A}} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix},$$
$$\overset{1\rightarrow}{\mathbf{A}} = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 5 & 6 & 7 \end{bmatrix}, \quad \overset{2\rightarrow}{\mathbf{A}} = \begin{bmatrix} 0 & 0 & 1 & 2 \\ 0 & 0 & 5 & 6 \end{bmatrix}, etc...$$

The leftmost columns of the resulting matrix are set to zero so as to maintain the original size of the input. Likewise we define the inverse operation $\overset{\leftarrow i}{(\cdot)}$, which shifts columns to the left (again appropriately zero-padding on the right).

In order to estimate the appropriate matrices $\mathbf{W}_t$ and $\mathbf{H}$ to estimate $\mathbf{V}$ we can use the already existing framework of NMF. We define our cost function as:

$$D = \left\| \mathbf{V} \otimes \ln(\frac{\mathbf{V}}{\mathbf{\Lambda}}) - \mathbf{V} + \mathbf{\Lambda} \right\|_F \tag{4}$$

Where $\mathbf{\Lambda}$ is our approximation to $\mathbf{V}$ defined as:

$$\mathbf{\Lambda} = \sum_{t=0}^{T-1} \mathbf{W}_t \cdot \overset{t\rightarrow}{\mathbf{H}} \tag{5}$$

Our new cost function can be seen as a set of NMF operations that are being summed to produce the final result. Keeping this observation in mind we can use the adaptation procedure for NMF, only this time instead of updating two matrices ($\mathbf{W}$ and $\mathbf{H}$), we will be updating $T+1$ matrices, all the $\mathbf{W}_t$ and $\mathbf{H}$. This results into the NMD update equations which are:

$$\mathbf{H} = \mathbf{H} \otimes \frac{\mathbf{W}_t^\top \cdot \overset{\leftarrow t}{[\frac{\mathbf{V}}{\mathbf{\Lambda}}]}}{\mathbf{W}_t^\top \cdot \mathbf{1}} \text{ and } \mathbf{W}_t = \mathbf{W}_t \otimes \frac{\frac{\mathbf{V}}{\mathbf{\Lambda}} \cdot \overset{t\rightarrow}{\mathbf{H}}^\top}{\mathbf{1} \cdot \overset{t\rightarrow}{\mathbf{H}}^\top} \tag{6}$$

In every updating iteration, for each t we update $\mathbf{H}$ and all $\mathbf{W}_t$. Through experience it has been found very practical to first update all $\mathbf{W}_t$ and then update $\mathbf{H}$ using the average result of its updates from all $\mathbf{W}_t$. Since $\mathbf{H}$ is dependent on all $\mathbf{W}_t$ serially updating it will result into an $\mathbf{H}$ heavily influenced by the last update using $\mathbf{W}_{T-1}$, and not equally by all $\mathbf{W}_t$. In terms of performance this technique largely depends on T. If $T = 1$ then it reduces to the NMF updates, otherwise it is burdened with extra updates.

3.2. Non-Negative Matrix Deconvolution applied on audio spectra

In this section we will consider the application of NMD on audio spectra and highlight its differences as compared to NMF. What NMD does is impose a temporal structure to the frequency description of each object. In NMF the spectra were constrained to be static and certain characteristics of their spectral evolution structure was lost. With the NMD representation the ith column of the $\mathbf{W}_t$ matrix describes the spectrum of the ith object t time steps after that object

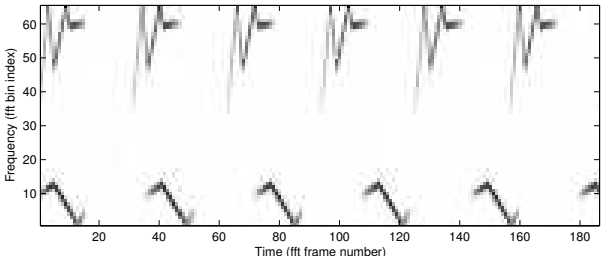

Figure 5: *Spectrogram of a scene composed of two repeating objects with distinct evolution in their spectral structure.*

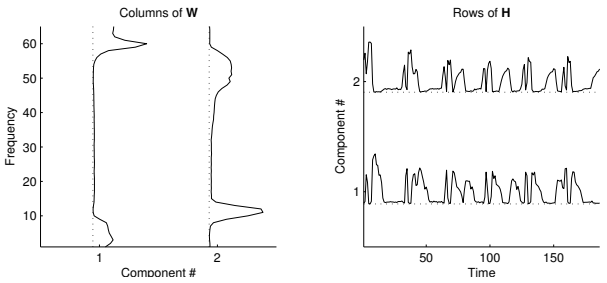

Figure 6: *NMF analysis of the data in figure 5. Note how the spectral evolution character of the two objects is lost.*

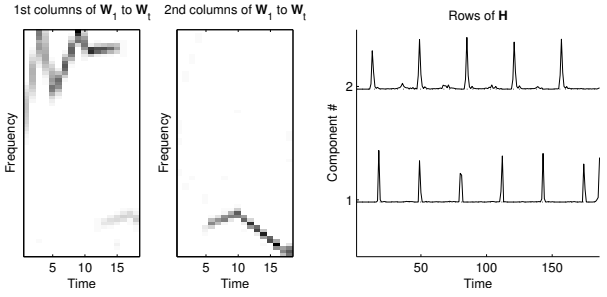

Figure 7: *NMD analysis of the data in figure 5. Note how the spectral evolution character of the two objects is now captured by the structure of the* $\mathbf{W}_t$ *matrices.*

it has begun. This provides an extra degree of freedom to describe how objects evolve spectrally once they have started. To demonstrate how this works as compared to NMF consider the synthetic spectrogram in figure 5. The sounds contained therein are two repeating sinusoidal patterns that are frequency modulated in a consistent manner.

In figure 6 we see the two object descriptions that NMF recovers. We can see that even though the region of frequency that each pattern dominates is encapsulated in the columns of $\mathbf{W}$, the results are not what we would hope for. The expressive power of this description is not enough to reveal the structure we are after.

Now let us consider the same input decomposed using NMD. This time in addition to defining how many component we want to discover we have to define the length T of spectral evolution that we are interested in. As with the value of R sophisticated statistical measures can be applied to find an optimal value for T, however this is beyond the scope of this paper and instead we'll select values heuristically. For this particular example we set $T = 18$, which is approximately the length of the two patterns. The results of the analysis are shown in figure 7. Note how this time the spectral evolution within each pattern is encapsulated in the columns of all $\mathbf{W}_t$. Likewise $\mathbf{H}$ now contains in its rows the temporal position of each pattern.

Now let us consider a more complex example with real sounds. We made two recordings of the same speaker continuously uttering the word "where" in one recording and the word "what" in the other. the pitch for the word "where" was going upwards, and for "where" downwards. The two recordings were mixed into a monophonic file and for most of the time the two words where overlapping. The spectrogram of the mixture is shown in figure 8. We applied NMD to this mixture with $R = 2$ and $T = 50$. The results are shown in figure 9. The columns of the resulting $\mathbf{W}_t$ have adapted to the spectrograms of the individual words "where" and "what", whereas the rows of $\mathbf{H}$ contain peaks where the corresponding words occur.

4. Extraction and reconstruction of objects

An additional advantage of these types of decompositions is the fact that we can reconstruct the input spectrogram using an arbitrary number of objects. Since the objects in the scene are often segregated in the columns of $\mathbf{W}$ (or $\mathbf{W}_t$) and the rows of $\mathbf{H}$, we can use only these to make a selective approximation. So in the case of NMF, if we wish to extract the nth object we do so by:

$$\mathbf{V}_n = \mathbf{W}^{(:,n)} \cdot \mathbf{H}^{(n,:)} \tag{7}$$

Where $\mathbf{W}^{(:,n)}$ is the nth column of $\mathbf{W}$ and $\mathbf{H}^{(n,:)}$ is the nth row of $\mathbf{H}$. Likewise for NMD we have:

$$\mathbf{V}_n = \sum_{t=0}^{T-1} \mathbf{W}_t^{(:,n)} \cdot \overset{t\rightarrow}{\mathbf{H}^{(n,:)}} \tag{8}$$

What $\mathbf{V}_n$ will be is the contribution of the nth object to the magnitude spectrogram, effectively recovering the time-frequency distribution of that object. This technique can be used for both extracting and suppressing individual components. We can only reconstruct one $\mathbf{V}_n$ to get the spectrogram of a single object, or reconstruct using a set of $\mathbf{V}_n$ in which case we get the contribution of selected objects. Finally we can always reconstruct and omit only one $\mathbf{V}_n$, which can be the component that models an unwanted noise source that we want to remove.

To illustrate the form of all the $\mathbf{V}_n$ consider the plots in figure 10, that reconstruct the spectrogram in figure 8, using one object at a time. The two plots display $\mathbf{V}_1$ and $\mathbf{V}_2$, which when summed will approximate the input spectrogram.

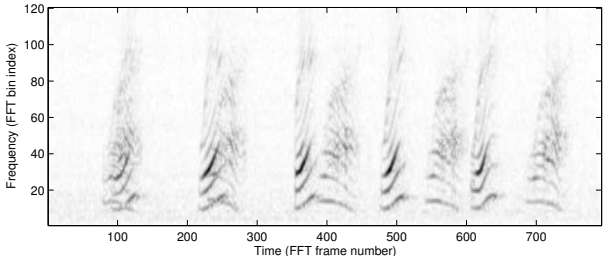

Figure 8: *Spectrogram of the overlapped repeating words "what" and "where". "Where" is discernible by the upwards moving formants, and "what" by the downward trend.*

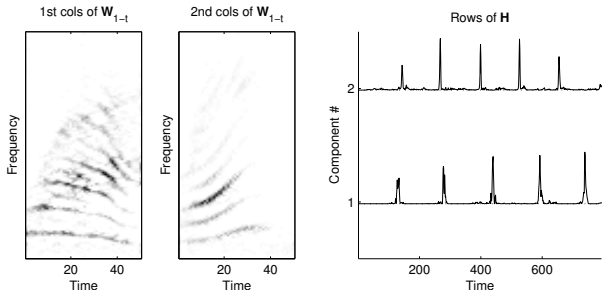

Figure 9: *NMD analysis of the data in figure 8. The two words have been discovered as the objects described in the columns of the $\mathbf{W}_t$ matrices. There respective location in time is denoted by spikes in the rows of $\mathbf{H}$.*

Since what we obtain with this process is a magnitude spectrogram if we wish to convert it back to the time domain we would need to deal with the missing phase information. Although multiple phase recovery techniques exist it is sufficient to just use the original phase of the input spectrogram and modulate it with the magnitude of each $\mathbf{V}_n$. Doing so results in fairly clean sounding extractions of each object, provided that the input data can be adequately modeled by these decompositions.

Finally it is also possible to modify the rows of $\mathbf{H}$ to alter the temporal composition of the scene (for example permuting the rows of $\mathbf{H}$ in the drum example would result into a "re-orchestration" of the drum loop), of tamper with the columns of $\mathbf{W}$ (or $\mathbf{W}_t$) to change the spectral character of the inputs.

5. Discussion

One of the interesting points that this paper brings up is the issue of what is an object. Given the lack of prior object knowledge in the adaptation method, the results we obtain are completely dependent on the input scene. The fact that this method gives us reasonable results has to do with how well the objects are exposed in the input scene. Considering the spoken words example, if both the spoken words were always perfectly synchronized and overlapping then the entire recording would constitute one object since the contained

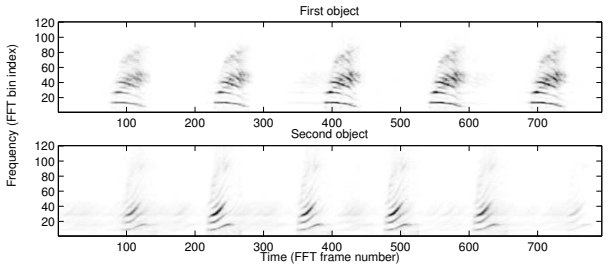

Figure 10: *Reconstruction of the spectrogram in figure 8 using only one component at a time.*

elements always appear together in a deterministic manner. The fact that the spoken words are not dependent in their temporal placing makes them stand out as individual objects. This means that the data provided to this algorithm needs to expose the individuality of each object by presenting it in a disassociated way with other objects.

The object description that we use raises the issue of how appropriate such a model is for general sounds. Although is seems inflexible to define objects as repeating spectra or spectral patterns, there is significant room to model real world sounds very well. This is certainly the case with musical signals where repeating patterns (either at the note scale, or the phrase scale) can be modeled very accurately. This has been extensively demonstrated by Casey [2], Smaragdis and Brown [9][1], Virtanen [10], and others. More complex signals such as speech recordings can also be analyzed this way although in this case we often see the extraction of individual phonemes as opposed to words or the voice of one speaker (Smaragdis [8]). Adding the temporal component in this model generalizes the nature of objects that are being sought and provides one more dimension of description. An obvious future work extension is to refine this model to deal with objects that scale in time length so that it can model temporal evolution in relative terms like Markov model or dynamic time warping.

Temporal and spectral overlaps are not an issue since the linear form of the model can deal with summed objects, we only need to make sure that objects do not repeat verbatim and are not summed the same way all the time. Once the approximation process is faced with the task of coming up with a concise description of the input it is forced to model each object because this is usually the most compact description that can perform a satisfactory approximation. Although this objective could be satisfied with an algorithm that explicitly attempts to find this description using a statistical foundation (such as PCA and ICA do), we have found NMF and NMD to be performing considerably better, despite the lack of statistical rigor and their simplistic cost function. One reason for this effect is that unlike other approaches there is a strict non-negativity constraint which fits naturally to finding components of magnitude spectra. Algorithms such as ICA which generally do not impose such constraints often extract mag-

nitude spectra that have both negative and positive elements (albeit one of the two in smaller quantities). The result of this effect is unwanted crosstalk between components. By imposing a non-negativity constraint the assumed model becomes more relevant to the data and thereby provides more suitable results for this particular problem. The has been solid work on enforcing non-negativity for ICA algorithms (Plumbley [6][7]) although the necessity of this constraint has been contested (Cichocki [3]). In general unconstrained ICA will produce similar results with non-negative ICA when confronted with non-negative data, however when it comes to sensitive domains such as audio, these slight differences make an audible difference. The complexity and domain constraints of non-negative ICA do not yet accommodate applications such as the ones presented in this paper, it is however a closely related area to NMF and can soon evolve to be a very useful audio analysis tool in this context.

Finally although this definition of an object certainly encompasses what we would perceive as an individual sound, it should not be thought so much as something that extracts what we perceive as an entire sound, but rather something that extracts building elements of auditory scenes. With analysis at a shorter time scale we can use this representation to find sets of temporal basis functions (analysis of speech results in bases being phonemes with various pitch inflections), whereas longer term and more coarse analysis can provide us with information over larger scale behavior (extraction of entire words and causal behaviors between sounds). The lack of semantics in this approach allows us to work at multiple levels and discover auditory objects at various resolutions.

6. Conclusions

In this paper we have shown how we can use simple non-negativity constraints to force low-rank approximations of magnitude spectra of audio scenes to reveal the existing structure. By attempting to model an input using only a handful of spectral and temporal information the models are forced to retain only the essential information and segment it in terms of the objects contained in the input. We have shown how we can use a simple NMF model to decompose scenes into static spectra and respective time profiles, and also extended the NMF model to a convolutive form that has more expressive power and is able to deal with more complex objects with a varying spectral character. Although these models are simple they perform admirably well on complex cases as we demonstrated with real world examples. The non-negativity constraint seems to be a natural fit for analyzing magnitude spectra and we anticipate it to be used extensively for more complex objects models in the future, not only because of its good performance, but also because of the simplicity of the mathematics involved.

7. References

[1] Brown, J.C. and P. Smaragdis. "Independent Component Analysis for Automatic Note Extraction from Musical Trills", In *Journal of the Acoustical Society of America*, Vol. 115, Issue 5, pp. 1851-2634, May 2004

[2] Casey, M.A. and A. Westner. "Separation of Mixed Audio Sources by Independent Subspace Analysis", in *Proceedings of the International Computer Music Conference*, Berlin, Germany, August, 2000.

[3] Cichocki, A. and P.G. Georgiev. "Blind source separation algorithms with matrix constraints," In *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol.E86-A, no.1, pp.522-531, Jan. 2003.

[4] Lee, D.D. and H.S. Seung. "Learning the parts of objects with nonnegative matrix factorization". In *Nature*, 401:788 791, 1999.

[5] Lee, D.D. and H.S. Seung. "Algorithms for Non-Negative Matrix Factorization". In *Neural Information Processing Systems* 2000, pp. 556-562, 2000.

[6] Plumbley, M. D. "Algorithms for non-negative independent component analysis". In *IEEE Transactions on Neural Networks*, 14(3), pp534- 543, May 2003.

[7] Plumbley, M. D. "Conditions for non-negative independent component analysis". In *IEEE Signal Processing Letters*, 9(6), pp177-180, June 2002.

[8] Smaragdis, P. "Redundancy Reduction for Computational Audition, a Unifying Approach", *Doctoral Dissertation*, MAS Dept. Massachusetts Institute of Technology, Cambridge MA, USA, 2001.

[9] Smaragdis, P. and J.C. Brown. "Non-Negative Matrix Factorization for Polyphonic Music Transcription", in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz, NY, October 2003.

[10] Virtanen, T. "Separation of Sound Sources by Convolutive Sparse Coding" In *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, October 2004.