Chapter 4. Auditory Scene Analysis

4.1 Introduction

In the previous chapters we focused on low level perceptual elements. The operations that we performed were taking place in short timescales and they were modelling subconscious perceptual processes. In this chapter we will describe similar techniques that deal with larger time spans and result in seemingly more complex tasks. We will explore the applications of ICA on extracting features from auditory scenes and relate this to the perception of separate sounds. Our approach will draw inspiration from two very different fields with opposing viewpoints. One is the field of multichannel statistical methods, and the other the field of psychoacoustic-based monaural scene analysis.
Source separation is the holy grail of audio processing. Its goal is to extract auditory objects from a scene. It is an elusive process that has attracted a lot of attention and has commanded a lot of research. Due to this there has been development on many different approaches towards solving it. The two dominant methods that relate to this thesis, are the statistical and the psychoacoustic approaches.

4.2.1 Psychoacoustic Separation

Psychoacoustic separation methods have been investigated ever since the definition of the cocktail party problem. This problem was defined by Cherry (1953) and raised the issue on how it is possible that a human at a cocktail party, as subjected in many auditory sources, can understand and extract from the scene only the source that draws the listener’s attention. Through time this problem has been transformed to finding a method that can computationally isolate auditory sources of a scene from a monophonic recording. This work was termed as Computational Auditory Scene Analysis (CASA) and drew a wide array of approaches.

In their most basic form most of the CASA approaches are attempting to extract objects by some form of perceptual grouping in the frequency domain. The usual path has been to utilize a time-frequency transformation and then extracting individual elements based on a set psychoacoustic grouping procedures. This approach flourished in the early 90’s where any conceivable transform and set of grouping rules were used to this extend. Notable work in this vein has been by Weintraub (1986), Cooke (1991), Mellinger (1991), Ellis and Vercoe (1992), Brown (1992) and Ellis (1994, 1996), and has been partly described in the preceding chapter.

These various approaches introduced different methods that work best within their domain of sources, but they were not universal extractors. Most were bootstrapped for either music or vocal data, exploiting knowledge and procedures specific to these domains. In most of these cases separation extracted distorted sources, partly because of the loss of information due to mixing. In light of the realization that exact reconstruction is not always possible, recently the opinion of many researchers has changed to that accurate reconstruction should not be the goal of CASA. Auditory perception is more concerned with the detection and outlining of sound objects, rather than the exact reconstruction (which is not always feasible). This is a step closer to the initial cocktail party problem and presumably to human listening, which brought CASA back against a more reasonable challenge.

In a hostile viewpoint, most of CASA is still rather primitive and short-sighted. This is partly because of the fuzzy nature of its goals and unconditional faith to psychoacoustical literature. The most important problem however is that most of CASA approaches still don’t have a definition of what a source is. Most often a source is heuristically defined as something that fulfills the requirements of being a separate object according to perceptual grouping. Given that perceptual grouping itself is also poorly defined, it is no surprise that there is a high incidence of extremely complex systems that produce extremely basic results. There is also a significant disregard towards general perception.
which results in highly specific systems that cannot shed any light on the inner workings of the human mind. Finally the strong belief into psychological literature, prohibits the use of formal mathematical definitions. These problems are acknowledged by many in this field and are not personal musings. The desire to overcome these obstacles has been noted, and is an active subject of debate.

4.2.2 Multichannel Blind Source Separation

In contrast to CASA approaches, multichannel separation is a clean, well defined and understood operation. It deals with the case where we have a number of sound sources as recorded by a set of sensors. Each of these sensor recordings will contain a superposition of the sources, thereby concealing their individual form. By modelling the recording process and using knowledge of the setup, or some measurable statistics, we can cancel out interfering sources to extract the desired ones. This operation is easy to formulate in the context of linear algebra. Assuming a set of sources

$$s(t) = \begin{bmatrix} s_1(t) & \cdots & s_N(t) \end{bmatrix}^T$$

we can model a multisensor recording process as:

$$x(t) = A \cdot s(t)$$  \hspace{1cm} (1)

where $$x(t) = \begin{bmatrix} x_1(t) & \cdots & x_M(t) \end{bmatrix}^T$$ are the recorded series and $$A$$ is a real $$M$$ by $$N$$ matrix, known as the ‘mixing matrix’. The structure and contents of the mixing matrix are dependent on the sensor and source numbers and physical locations (Figure 1).

![Figure 1](image_url)  

**Figure 1**  
Example of a multichannel setup. The stars $$s_1$$ and $$s_2$$ are the sources and the dotted lines their paths to the microphones. The solid lines indicate their corresponding distances. Assuming a simple inverse distance attenuation rule for the propagation of sound, we can deduce the mixing matrix for this scene.
Auditory Scene Analysis

If $M \geq N$ and $\mathbf{A}$ is full rank it is possible to recover the exact original sources by:

$$\mathbf{s}(t) = \mathbf{A}^+ \cdot \mathbf{x}(t)$$ (2)

where $\mathbf{A}^+$ is the inverse (or pseudoinverse if $M \neq N$) of $\mathbf{A}$. In practice the invertibility constraints that $M \geq N$ and full rank of $\mathbf{A}$, mean that no two or more sensors or sources can be at the same position, and that we have at least as many sensors as sources (If these constraints do not apply, it is only possible to try to reinforce the sources as much as the situation allows, but in general it not possible to fully recover them. We will not consider these cases here). This scenario is referred to as the instantaneous mixing scenario, since it does not model propagation delays and other filtering issues, thereby assuming an instantaneous transmittance of sound.

More realistic mixing systems were subsequently formulated, modeling the mixing process as:

$$\mathbf{x}(t) = \mathbf{A} \cdot \mathbf{s}(t)$$ (3)

where $\mathbf{A}$ is an FIR matrix. That means that instead of scalar elements the matrix contains FIR filters and matrix multiplication of this object with a vector is defined as:

$$\mathbf{A} \cdot \mathbf{x}(t) = \begin{bmatrix} a_{(1,1)} & \cdots & a_{(1,N)} \\ \vdots & \ddots & \vdots \\ a_{(N,1)} & \cdots & a_{(N,N)} \end{bmatrix} \cdot \begin{bmatrix} x_1(t) \\ \vdots \\ x_N(t) \end{bmatrix} = \begin{bmatrix} \sum_i a_{(1,i)} \ast x_i(t) \\ \vdots \\ \sum_i a_{(N,i)} \ast x_i(t) \end{bmatrix}$$ (4)

where the $\ast$ operator denotes convolution and $a_{(i,j)}$ are FIR filters. This model accounts for the effect of reflections, propagation delay and reverberation in a recording environment.

In both of the above cases the objective of a source separation algorithm is to estimate an inverse of the mixing matrix ($\mathbf{A}$ or $\mathbf{A}$). In general this matrix is not available to us, and it is deduced from observation of the recordings, and/or knowledge of the sources and the geometry of the setup. Modern techniques, known as blind processing techniques, have no knowledge of the environment (hence the term blind), and only use the statistics of the recordings to make their estimation. The only assumptions made are that different sources are not dependent to each other, so by trying to find an unmixing matrix that produces maximally independent outputs it is possible to obtain the original sources. Although there are many ways to attack this problem when armed with this assumption, the most relevant to our work is the application of ICA methods. It is quite
obvious that the outputs of Equation (2) can be forced to be independent using an ICA algorithm. This has been exploited by many researchers (see Torkkola 1999 for exhaustive reference) and in the case of instantaneous mixing is regarded as a solved problem. Equation (3), commands more complex implementations of ICA, it has been however successfully solved for a reasonable number of sources and filter sizes (Smaragdis 1997, Lee et al 1997).

Overall the ICA approaches to source separation yielded impressive results, superior to their CASA counterparts. However, the use of multisensor methods for source separation has not been accepted by the CASA community since it deviates on a key point for perceptual studies. Auditory perception is based at most on a two channel input (our two ears), and the constraint that current ICA algorithms impose is that we have at least as many inputs as sources. Although there is this obstacle in bridging ICA and CASA, ICA can boast a relation to redundancy reduction and the relevant perceptual theories which by far supersede the CASA philosophies in both depth and breadth.

Scene analysis, as has been pursued in the auditory community, has been intimately linked to source separation. In the rest of this chapter we will not abide by this convention. We feel that it is important to realize that understanding a scene and extracting useful information from it, is a distinct task from source separation (it is however driven by the same principles). This is a realization that has come to mature only recently (Ellis 1997, Scheirer 2000), and source separation is nowadays seen as a probably impossible task, and not crucial to perceptual processing. We will adopt this point of view and continue with this chapter presenting methods that lie in the fine line between detection and separation. Our primary goal will be awareness of the structure of an auditory scene. Issues of separation will be examined as side effects. Our approach will start from a methodology introduced by Casey and Westner (2000), which combines the formal rigor of the multichannel approach, with the realistic constraints of the perceptual approach.

### 4.3 Decompositions of Magnitude Transforms

#### 4.3.1 Object Detection

In this section we will examine the problem of detecting individual sounds from an auditory scene. Our assumption, as in the previous chapters, is that by performing redundancy reduction we can yield interesting results. We will examine redundancy reduction as it applies to the time-frequency energy of signals. We will show that by extracting the energy of components of scenes we can deduce a lot of information about the events that take place. This is an approach that will try to use the philosophy of CASA approaches, but backed by the more rigorous theories relating to multichannel approaches.

Having a sound scene $s(t)$ we can analyze its energy content with respect to a set of bases using the magnitude of the transformation that these bases dictate. For example, the magnitude of a STFT transform will reveal the energy of all the sinusoidal compo-
Auditory Scene Analysis

\[ F = \mathbf{A} \cdot \begin{bmatrix} \uparrow & \uparrow & \ldots & \uparrow \\ s_1 & s_2 & \ldots & s_M \\ \downarrow & \downarrow & \ldots & \downarrow \end{bmatrix} \] (5)

where \( \mathbf{A} \) is an \( N \) by \( N \) matrix expressing the desired transformation, the \( \| \) operator performs element-wise magnitude extraction, and \( s_i = [s(i+1) \ldots s(i+N)]^T \).

The resulting matrix \( F \) can be interpreted in two directions. If we observe it by traversing its columns one after the other, we can think of them as instantaneous magnitude transforms of the scene, localized in successive times. Observing its rows, we can think each one as a time series denoting the amount of involvement of a basis. We will refer to the columns of this matrix as spectra and the rows as energy tracks.

Since our thesis is that redundancy reduction of data leads to perceptual-like outputs, we will apply the same techniques we’ve used in the previous chapters to this representation. In order to build some intuition we will start with a very simple two object scene and its magnitude STFT transform (so that the \( \mathbf{A} \) matrix in Equation (5) will be the Fourier matrix). The scene is one that displays interesting structure in both the time and the frequency axis. The two components of this scene are two amplitude modulated sinusoids:

\[ s(t) = \cos(1321t)g(t) + \cos(3211t)g(9.7t) \] (6)

where \( g(\cdot) \) is a periodic gate function defined as:

\[ g(t) = \begin{cases} 1, & \text{if } \sin(t) \geq 0 \\ 0, & \text{if } \sin(t) < 0 \end{cases} \] (7)

The odd nature of the constants defined in Equation (6) was such so as to avoid any gestalt grouping problems. The two gated sinusoids are intended to be separate and independent sources.

The magnitude of the STFT of this scene is depicted in Figure 2. To obtain it we used an FFT size of 128 and a hop size of 32 samples. This data will constitute our \( F \) matrix in Equation (5). Due to the symmetry properties of the FFT we only retained the first 65 samples from each transform (64 frequencies and the constant offset).
By observation of the data we can extract basis functions for the matrix $F$. We can do so in either of the two dimensions. We’ll first consider basis functions that describe the spectra (the columns of $F$). They would be 65 samples long and they should be able to reconstruct every column of $F$ (the vertical slices in Figure 2). We will try to estimate these basis functions using techniques for redundancy reduction presented in earlier chapters.

We will start by using the PCA transform on $F$. PCA will provide a matrix $W_p$ so that the operation:

$$C_p = W_p \cdot F$$  \hspace{1cm} (8)

will result in a matrix $C_p$ whose rows will be decorrelated. The rows of the $W_p$ matrix will be a set of bases comprising the transformation that results in $C_p$. If we wish to resynthesize the matrix $F$ using $W_p$ and $C_p$, we only need to solve Equation (8) with respect to $F$, from which we get:

$$C_p = W_p \cdot F \Rightarrow W_p^{-1} \cdot C_p = W_p^{-1} \cdot W_p \cdot F \Rightarrow F = W_p^{-1} \cdot C_p$$  \hspace{1cm} (9)
The columns of matrix $W_p^{-1}$ will contain a set of bases that are required to resynthesize $F$. In terms of transforms, $W_p^{-1}$ will be the inverse transform of $W_p$.

In our case we only require a small number of cases and we do not care as much for an accurate reconstruction. In order to perform this dimensionality reduction on our transform $W_p$ we only keep the first few rows, which correspond to the most significant bases. In order to then recover the inverse transform we need to take the pseudoinverse of $W_p$:

$$W_p^+ = (W_p^T W_p)^{-1} W_p^T$$

The columns of $W_p^+$ will be the bases of the inverse transform. They will effectively be the elements that when multiplied with $C_p$ will reconstruct the scene. We will think of them as the building blocks of our scene. We will now examine these bases and their resulting transformation.

In the case at hand we will obtain a set of resynthesis bases which will be 65 samples long and can be used to recreate all of the spectral instances of Figure 2 (the columns of matrix $F$). Since we only have two simple objects we will only keep two bases. The results we obtained are shown in Figure 3.

---

Figure 3 Results of the PCA analysis on the data in Figure 2. This figure displays the estimated spectral basis functions, which almost reveal the two sinusoidal spectra.

†. In the special case of PCA, the inverse transform matrix is just the transpose of the forward transform matrix (since $W$ is an orthogonal matrix), and most references are as such. However, we generalize and use $W^{-1}$ to accommodate a more abstract framework, which will help later on.

‡. In the case of PCA, ‘significant base’ implies a base that has a significant contribution in terms of variance contribution.
We note that these two basis functions roughly correspond to the two spectra that construct each of the two sinusoids. In order to increase the quality of the results, we will continue the analysis by a subsequent application of ICA on the resulting matrix $C_p$. By doing so we will make the operation:

$$C_I = W_I \cdot C_p = W_I \cdot W_P \cdot F = W \cdot F$$  \hspace{1cm} (11)

Where $C_I$ is the output of the ICA transform $W_I$, and $W$ the overall transform that we have applied on $F$ (the transform that includes both the PCA basis reduction and ICA steps). The inverse (or pseudo inverse if we reduce the number of bases) of $W$ will contain the new ICA resynthesis bases is its columns. Applying this to our scene we obtain the bases in Figure 4.

![ICA spectral bases](image)

**Figure 4** Spectral bases of the scene in Figure 2 as extracted using ICA. Each of the two bases corresponds to one of the sinusoids in the scene, effectively discovering its composition.

We now notice that the extracted bases are much more clean, and have a interesting relation to the spectral structure of the scene having highlighted the existence of the two different frequency peaks, each corresponding to one object. In effect we have isolated the frequency contribution of each object. It is also interesting to examine the resulting transformation $C_I$ (Figure 5).
Figure 5  The resulting transformation of the data in Figure 2. Note how the independent components are in fact the amplitude envelopes of the two objects.

The rows of $C_I$ contain the transformed energy tracks of $F$ so that they are now maximally independent. We notice that this transformation results into two energy tracks describing the amplitude of the two objects in our scene.

The operations that were performed on $F$ were that of reducing unnecessary data by the PCA dimensionality reduction, and then forcing the remaining elements to be statistically independent. As is easy to see from Figure 2 we had two kinds of significant energy tracks (the horizontal slices of $F$), the tracks that corresponded to the amplitudes of the two objects. These were the two tracks that were retained by the PCA transform, they were not however returned in a clearly separated form (and in general they will never be), they were returned as two mixtures. The ICA step was responsible for the fine tuning required for separating these two components into two independent tracks. In effect this last operation is equivalent to the work in multichannel separation, only we are performing it in a subsampled transformed domain. We have produced the extra channels by the magnitude STFT transform and that allowed us to run multichannel ICA on an originally monophonic sound. The motivation to do this operation stems from the fact that the individual objects in a scene are independent. That means that they feature a set of energy tracks that would also be independent. By attempting to decorrelate these energy tracks and then seeing which spectra correspond to them, we hope to reveal some of the structure in a scene. This is the same reasoning applied in the multichannel ICA work, only on a different domain.

We will now examine the same operation as applied on the other dimension of the data. We will extract time bases by striving for independence of the spectral bases. To do so we only need to repeat the experiment above by setting the input to our analysis being $F^T$ rather than $F$. Just like before we obtain a set of time bases and spectral components, which highlight the two objects in the scene. The results after the ICA step are shown in Figure 6.
Auditory Scene Analysis

Figure 6 PCA analysis of the data in Figure 2. These results, although similar to Figure 4 and Figure 5, were obtained by applying the same procedure, but this time on the other dimension. This operation has the effect of making the transformed spectra independent, as opposed to the previous approach that made the energy tracks independent.

It is interesting to note that we obtain qualitatively the same results as we have had using $F$ instead of $F^T$. This fact highlights an interesting point; independence between objects in a scene exists in both their frequency and time axes. Interestingly enough by looking at one domain we can extract information about the other. In the previous case we obtained our results by forcing the energy tracks to be independent and then seeing which spectral bases corresponded to them. In this case we performed the opposite, trying to get the spectral components to be independent and then extracting their corresponding energy tracks. Since clues of the objects’ identities existed in both the time and frequency axis, it became possible to detect both by forcing a constraint on only one of them.

Obviously the scene we used so far was quite a simple example where there was, on average, little overlap in the time domain and none in the frequency domain and good discrimination between the two objects. We’ll now use a more complex scene. It will be a three element scene comprised by the sum of the components displayed in Figure 7. These three components are drum instruments, a snare drum, a bass drum and a hi-hat. This is an example that displays a more complex mix of spectral and temporal variation. The solution is non-trivial since for some time instances we have overlap of multiple components and between all instruments we have spectral overlaps.
Auditory Scene Analysis

Figure 7 The left figure displays the individual components of an auditory scene in the time domain. The right figure displays their corresponding power spectra. The scene itself is composed by the summations of these components.

By noting the time and frequency characteristics of the instruments we can make some observations that will help us in deciphering our results later. The bass drum contains a lot of energy at the lower side of the spectrum and its only isolated instance is its second one. The first and third instances occur simultaneously with the hi-hat. The snare drum has a resonant character in the low-middle frequencies and is somewhat wideband. Of its two instances only the first is isolated, the other coincides with a hi-hat. The hi-hat has a wideband tone with some high resonant frequencies. In addition to its temporal overlaps we mentioned above there is also significant spectral overlap with the snare drum.

We analyze this scene using the magnitude STFT again and obtain the matrix $F$ shown in Figure 8. We used an FFT size of 512 and a hop size of 64 samples. As we had done before we truncated the frequency axis to keep only one half of the symmetric spectrum plus the constant offset.
Auditory Scene Analysis

Figure 8
The STFT of the drum loop scene. The individual components are easily identified by eye from their temporal and spectral characteristics.

We apply our previous procedure using PCA dimensionality reduction on $F$, down to four bases, and subsequently follow this step with application of ICA. This operation makes the temporal features maximally independent and provides the results in Figure 9.

Figure 9
Analysis of the data in Figure 8. The left figure displays the estimated resynthesis spectral bases, and the right figure their energy through time. In contrast to the original objects we have a fair decomposition of the scene to its original comprising elements.

Observation of the results unveils the structure of the scene. Easier to decipher is the resulting transform $C_p$ (right figure in Figure 9). We can easily see that the second and fourth energy tracks correspond to the bass drum. The fourth appears to deal with the snap in the attack portion of the bass drum, whereas the second one deals with its low

Decompositions of Magnitude Transforms
frequency resonance. By examining their spectral counterparts, we see that they are both exhibiting a low frequency content. The one corresponding to the attack portion has a slightly higher frequency content, whereas the other has most of the low end energy. Likewise we see that the snare drum energy track is well captured in the third component. As we might expect the corresponding spectral base has the same resonances as the snare drum. Finally the first component is the estimated energy of the hi-hat. Due to the fact that it temporally coincides with the bass and the snare drum for some instances, the time track is not perfect. The spectral base though is very accurate and encapsulates all of its characteristics.

In order to deal with the temporal problems in our estimation, we repeat the process using the other dimension of the spectrum by analysis of $F^T$. This will perform statistical separation of the spectra of the objects, rather than the time energies, which should alleviate the time tracking problems we have. The results are very similar, the spectra indeed correspond to the sounds we had in the scene, and the temporal functions are more complete now. We do note however that the energy track corresponding to the hi-hat does contain some extra events. Due to the spectral similarity of the snare drum and the hi-hat (they both feature a wideband spectrum), their separation was not perfect and we notice that the hi-hat track also includes the snare drum events.

![ICA analysis of the data in Figure 9. The two figures are the corresponding ICA outputs to Figure 9. In contrast we see that the temporal estimation of the hi-hat was more accurate, although there is confusion with the snare drum.](image)

In general, the issues that arise with these kinds of overlaps are not easy to solve. The process of monophonically mixing a set of sounds results in information loss that we cannot reconstruct without additional knowledge. For this reason it is hard to extract exactly what occurs at all points. In many cases this requires the use of a knowledge system to interpolate missing data. In our case the only form of knowledge is the knowledge discerned from the presented scene. Considering that the drum scene encapsulates all of the auditory experiences of our system, the results are not bad at all.

As hinted by the initial definition of this process, it is not imperative to use the magnitude STFT transform for this process. The magnitude STFT was used due to its desir-
Auditory Scene Analysis

Decompositions of Magnitude Transforms

able phase-invariance properties and since it is a fair approximation to the bases that we have discovered in the second chapter which does not complicate the implementation. It is quite possible to substitute this transform with a DCT transform, or even PCA or ICA derived bases. In fact, the transform can even be the unit matrix (or for that matter any linear transform). The performance of this algorithm however is strongly biased depending on the transform type. The main goal of the transform step is to obtain a sparse decomposition of the data which can consequently assist the PCA and ICA steps to have better convergence. By taking the magnitude of the transform we create an invariance to the sign of the data, and in the case of the STFT an invariance to the phase of the data. Perceptually these invariances exist in the human auditory system, and they help provide a good sparse coding of what we hear. Computationally these invariances venture in the domain of nonlinear ICA since they deal with a non-linear and non-invertible transformation, and complicate matters considerably. It is outside the scope of this thesis to delve into this territory (it is also a volatile and relatively unexplored subject), but we will however bring up the subject once more in the last chapter.

Similar results to the ones presented can be extracted from a variety of scenes, and the results are dependent mostly on the density and overlaps of the scene objects. The issue of invariance comes up again in the case of complex sources. So far we have dealt with auditory objects that have a fairly static spectral character. Once we attempt to extract sources with a more complicated structure, it is quite impossible to obtain results by forcing the spectra to be independent due to the absence of consistent spectral characters. For such complex scenes it is preferable to force independence of the energy tracks instead. This implies analysis of $F$ rather than $F^T$ and we will deal with that in later sections.

4.3.2 Applications in Music Processing

The estimation of independent spectral bases opens avenues of exploration for music processing. A traditionally hard operation on audio processing is music transcription. It involves the translation of a raw audio stream to musical semantics (most systems attempt only the extraction of musical notes, a daunting task by itself).

Fitting our framework, we present a brief example of how we could use the method developed in the previous section for note detection. We used the first two bars of Bach’s first invention in C major BWV 772 (Gould 1966). This being a solo piano piece, it offers a highly structured scene. There is a strong dependency to frequency templates which are the present piano notes. This is a structure we can easily extract by the aforementioned analysis method. We repeatedly performed the analysis using windows lasting one bar (total two windows to cover the entire scene). The first window of analysis contained a segment with one note sounding at any time, whereas the second one contained a polyphonic passage. We performed ICA on $F^T$, thereby striving for independence of frequency components, since we know that the structure we care for lies there. We kept five components from the first window analysis and twelve from the second. The resulting frequency bases from the first analysis are displayed in Figure 11.
Auditory Scene Analysis

As expected the frequency bases were the spectra of the notes in the scene, and their energy tracks provide their corresponding time location. Cross referencing this data with the score of the piece proves that we have correctly identified the position of every note. The second bar is a more challenging case, since it exhibits polyphony. To compensate for the added harmonic activity we extracted twelve bases and used a longer FFT length. The results were similar and are displayed in Figure 12. The first six bases correspond to the left hand notes, and we have only one undetected note and an octave ambiguity for the last G. The top six bases capture the trill between C and B, as stretched out and jagged energy tracks.

† An historically interesting point to make, is one that is reflected in our results. There is a reason why the third basis does not have a smooth energy track, and seems to accent the start and end of every note instance and is rather low and noisy during the middle. This is an effect due to the fact that the piano used in this recording was damaged during a transfer, and was recorded having a middle register that suffered from a mechanical tic that resulted in an audible hiccup. In the analyzed passage, the E note, represented by the third basis, is one of the affected notes, and the energy track has reflected that!
From an evolutionary standpoint this is a very interesting results, since it highlights the connection between statistically strong features and musical building blocks. The interesting point to make is that the notion of notes was implied by examination of the sound scene and it was not specifically pointed out. We have thus, created an analysis from which our system can deduce musical information from listening, rather than being explicitly told what to look for.

### 4.3.3 Auditory Object Extraction

Since we do have a way to detect and isolate objects in a transformed domain it is conceivable that we can perform an inverse operation to extract individual objects. Using the magnitude transform to obtain our data, we face the ambiguity of phase as we attempt to perform the inverse transform. Having kept only the magnitude data, we have no information about the phase. One, admittedly poor, way of doing this is to use the phase values from the initial scene and modulate them with the newfound amplitudes. Although this is not a precise and clean way to reconstruct the data is provides reasonable results (we should keep in mind that exact reconstruction of each auditory object is not always possible, and this pursuit is merely to satisfy our curiosity of how the components sound).

To explain this more formally, having obtained a set of bases $\mathbf{W}$ and their coefficients $\mathbf{C}$ from the original input $\mathbf{F}$, we can reconstruct $\mathbf{F}$ by:

$$\mathbf{F} = \mathbf{W}^{-1} \cdot \mathbf{C}$$  \hspace{1cm} (12)

As we mentioned we do not need to keep all of the bases in $\mathbf{W}$. If $\mathbf{W}_r$ is the matrix that contains the reduced set of bases, we can reconstruct the data-reduced $\mathbf{F}$ matrix (denoted as $\hat{\mathbf{F}}$) by first doing:

$$\mathbf{C}_r = \mathbf{W}_r \cdot \mathbf{F}$$  \hspace{1cm} (13)

to obtain the coefficients for the reduced bases, and then:

$$\hat{\mathbf{F}} = \mathbf{W}_r^+ \cdot \mathbf{C}_r$$  \hspace{1cm} (14)

to get the thinned-out $\hat{\mathbf{F}}$. For the sake of illustration had we wanted to obtain a reconstruction using only one component, we could use:

$$\hat{\mathbf{F}}_i = \mathbf{W}_{(i)}^+ \cdot \mathbf{C}_{(i)}$$  \hspace{1cm} (15)
where the subscript on the right hand matrices, selects the \(i\)th column of \(W^+\) and \(i\)th row of \(C\). This will give us the magnitude spectrum of only the \(i\)th component. In order to convert this to an invertible spectrum we can use the phase of the original input \(F\) and amplitude modulate it by our new \(\hat{F}_i\) as:

\[
S_i = \hat{F}_i (\cos(\angle F) + i \sin(\angle F))
\]  

(16)

\(S_i\) will contain the approximate time-frequency spectrum of one component, and we can use it as the input to an inverse STFT to obtain the component in the time domain. Using this reconstruction we can extract audible representations of the sinusoids, drums and piano notes presented in the previous examples.

Following is an example of a section from the pop song “Da da da” (Trio 1981). The particular section contains instances of the word “da”, a synthetic clave sound, a bass line, some guitar strumming and drums.

![Da da da scene spectrogram](image)

**Figure 13** Looking at the spectrogram we can locate the individual sounds. The utterances of the word da can be distinguished by the formant structure of speech. The clave is the high frequency ‘beep’ which we see halfway through the spectrum. The snare drum and the guitar strums coincide in time and are the large vertical columns. The bass and the bass drum are located at the bottom of the plot. For legibility purposes not all instances of each sound are marked.

Applying on this scene the aforementioned analysis and resynthesis method we can effectively, detect and pseudo-extract the instruments that are present. In this case we analyzed \(F\) (looking for independence of energy tracks) and used the STFT with an FFT size of 64 and a hop size of 8. We also applied a hanning window to improve clarity of reconstruction and avoid frame to frame clicking transitions. We used PCA to reduce
Auditory Scene Analysis

the number of energy components down to nine. Out of these nine components, at least six corresponded to the various instruments in the scene (Figure 14).

Figure 14

Clockwise from top left: extraction of the ‘da’ utterances (note the formant structure), the bass drum (low and short bursts), the clave (high pitched short ‘beeps’), the snare drum (wideband bursts, low-mid resonance), guitar strumming (somewhat contaminated by the snare drum, but exhibits a more harmonic structure than the snare drum track), and the bass part (with a clear harmonic structure and low notes).

The results have been quite satisfying. The algorithm successfully tracked the major components of the scene and provided a reasonable reconstruction. Extraction was not
comparable to multichannel techniques though. For example the guitar and snare drum, whose instances were synchronized, posed a hard, if not impossible, separation problem. The missing information due to their mixing cannot be reconstructed and the resulting components are not clearly separated (however, we can distinguish them quite easily as the guitar and snare drum tracks).

4.3.4 What is a Source? What is an Auditory Object?

An issue that arises at this point is the definition of a source, an auditory percept, or an auditory object. It might have been noticeable that in the preceding section we used the term “auditory object”, rather than “source”, to describe what it was we extracted from the scene. Preceding researchers have consistently defined their extracts as sources, a term we did not use. In this section we will attempt to make a distinction between a source and an object, as they relate to the extraction process.

The concept of a source is usually defined in terms of human perception. It is easy to make for most sounds; a passing car, a person speaking, or music playing from a speaker can be a source. These are percepts that provide some sense of continuity which makes us perceive them as sources. Even though the car can shift gears, the speaker stop speaking for a while, or the music be composed from a changing setting of instruments, we still perceive them as one auditory source, regardless of the drastic changes they undergo. This is because we have a prior knowledge of how these sources are produced and we are able to make higher level judgements on how to piece such varied sounds together to construct a source. For different reasons, both the CASA and the multichannel approaches perform extraction of sources. In the CASA case extracting sources is a hardwired function and a primary objective; in the case of multichannel systems the inputs are spatially constrained in such a way so that they imply a source structure.

Our approach, although heavily indebted to CASA and multichannel work, extracts a different kind of auditory entity. We will use the term auditory object. An auditory object is a simpler element which is easily defined as an independent set of data in the scene. The particular method we provide to extract this independent set works by defining spectral or temporal templates. In the case of spectral templates, an object is extracted as a frequency response which, in the analyzed scene, is amplitude modulated in a statistically independent manner. Likewise, extracting temporal templates results in statistically independent amplitude tracks, being harmonically modulated. It should be quite clear that this type of analysis is inadequate in capturing a source as defined in the previous literature. So it is important at this point to make a distinction between a statistically independent object in the scene and a source. In a strict mathematical sense a person speaking in between long pauses, although a source by itself, provides many auditory objects which can be either the syllables, the words, the sentences, or some other coherent mass of sound (a fact determined by our method and length of observation). Using our particular method, we extract objects that have some common frequency or time modulation. We cannot expect to extract speech, or music as one object because it is not. We can only pick out coherent pieces that compose them. If we bring ourselves to the ‘mindset’, of our algorithm it is easy to understand this. The submitted scene is the only stimulus the algorithm receives and all knowledge we use is extracted
Auditory Scene Analysis

from it. Had the scene had only spoked word, the algorithm not having the knowledge of what speech is, it would be quite natural to say that each different coherent section of the speech signal is a different object. We can not expect at such a low level to extract complex representations. We perceive this as one source, but this is because we also analyze the semantic content, and use the timbral continuation to deduce that it originates from a single speaker, thereby concluding it is one source. Our algorithm, does not have access to all this information and makes a simple statistical judgement. Now if the spoken word was such that it offered some kind of repetition to imply continuity, then the connection would be surely made (such is an example of the “da da da” extraction, where the repetition of the utterance da makes the algorithm extract all of the sung part as one object).

The following example highlights the difference between a source and an object. It is a segment from the jazz song “Blue moon” as sung by Billie Holiday (Holiday 1945). The section we analyzed contains Holiday singing “you saw me standing alone” over a bass line, a piano part and some faint drums (Figure 15). Our goal was to extract just the vocal part. Since this part is hard to represent with only one component, we tried to extract more components and selectively combine them to reconstruct all of the sung part.

![Blue moon spectrogram](image)

Figure 15 Extract from ‘Blue Moon’. The vocal parts, easily seen by their formant structure, cannot be extracted by a single frequency template as we have before. To overcome this problem we need to combine a set of components to reconstruct the entire sung line.

We extracted nine components, and some of them were corresponding to speech. Individual components have adapted to represent dominant and locally static parts of the singing line (the auditory objects). These were mostly the vowel parts which are sung most prominently (Figure 16) and display statistical coherence amongst them.
Once the vocal components are extracted we can reconstruct the entire vocal part by just combining them. Likewise, when we analyze complex scenes we need multiple components to extract complex sources such as speech, lines of a solo instrument etc.

Another way to deal with the issue of needing many components to represent one source is the use of analysis frames. As we have done in the previous example analyzing the Bach invention, we can subdivide the scene into multiple segments and analyze each one individually. Each frame, due to limited content, will require fewer components to represent all the sources, and by proper recombination of components we can reconstruct a source over the course of many analysis frames. Unfortunately, finding the successor of each component in the consecutive frames is a very hard problem and this method of analysis is rather hard in all but trivial cases. To remedy this, online algorithms can be used to provide a running estimate of the components on a sample by sample basis. Although this approach eliminates a lot of the work required to match components over time, it provides poorer estimates due to its short time frame of operation. In any case though, the results we have so far are promising and perform fairly well under reasonable conditions, the only major drawback being that this process requires human guidance. Most work in this area has been preliminary though, and we are confident that in time these problems can be addressed in a more mature manner.

4.4 Conclusions

In this chapter we presented a methodology for decomposing auditory scene to a set of independent features. The approach was a hybrid of the multichannel statistical approaches, and the monaural psychoacoustical models. This was inspired by the fact that although the psychoacoustic approaches were dealing with a perceptual problem (one input, many outputs), they were not using the elegant formulations of the multichannel approaches (which in turn were not dealing with a perceptually realistic problem). The formulations used for the multichannel approaches, aside from rigor and depth, had a strong link with perceptual theories and provided a, conceptually, very strong platform for scene analysis. Once combined with the scenario of the psychoa-
coustic approaches they resulted in, what we feel, is a much more plausible process for scene analysis.

This particular approach is however a starting point rather than a solution. It has many unclear points, and a fair amount of manual fine-tuning to obtain results. Its relative simplicity and conformance to perceptual theories make it a compelling approach which can potentially lead to more elegant auditory scene analysis. Hopefully it will mature in time and provide greater autonomy and robustness.