Chapter 6

Experiment: Partial grouping by amplitude- and frequency-modulation

6.1 Introduction

To evaluate whether the grouping of partials with common amplitude- and frequency-modulation (AM and FM) parameters is plausible, we synthesize a set of parameters and test by corrupting the parameters with noise and adding spurious sets of parameters that should not belong to any sources.

6.2 Methodology

We assume parameters have been estimated already so we start from theoretical values for the amplitude, frequency, frequency-modulation and amplitude-modulation. On each frame of analysis data, i.e., for parameters belonging to the same time instant, we consider each data-point as a multi-dimensional random variable. With these random variables, we compute principal components in order to produce a variable with maximum variance. This variable is classified using a clustering algorithm and we evaluate the results. A summary follows:

- Parameters are synthesized from a theoretical mixture of AM and FM sinusoids. Spurious data are added to these parameters.

- Principal components analysis is carried out on the parameters happening at one time instance.

- A histogram is made of the first principal components. Values sharing a bin with too few other values are discarded to remove spurious data points.

- Initial means and standard deviations for the GMMs (see Appendix B) are made by dividing the histogram into equal parts by area and choosing the centres of these parts.

- The EM algorithm (see Appendix B) for GMMs is carried out to classify the sources.

6.3 Evaluation

The algorithm is run on a typical source separation problem to evaluate its plausibility. Two sources are synthesized, each exhibiting both frequency- and amplitude-modulation. The amplitude and frequency of the frequency-modulation are chosen to be realistic with respect to musical instrument sounds — ± 12.5 cents surrounding the fundamental at around 6 Hz (see Table 6.2), similar to the measurements obtained in [38] for violin vibrato. Because musical sounds exhibit a wide variety of amplitude envelopes, one is chosen that is realistic, but that is not based on any particular instrument or recording. For this process all that is important to carry out source separation is the relative amplitude-modulation of the two sources — a reasonable assumption for recordings of a mixture of two different instruments or two performers of the same instrument.

To have control over the frequency- and amplitude-modulation separately, we compute the parameters of a function describing the amplitude envelope, and one describing oscillatory part. The parameters are combined when carrying out the classification.

6.4 Synthesis

Our model makes available the parameters summarized in Table 6.1. To incorporate inharmonicity often observed in real string instruments where the strings exhibit some stiffness,

H	Duration between data-point calculations in samples (i.e., the hop size).
N	Number of sources.
p	Which source.
$f_{0,p}$	Fundamental frequency.
K_p	Number of harmonics.
$k_{60,p}$	Harmonic number 60 dB lower than the first.
B_p	The inharmonicity coefficient.
$\phi_{0,p}$	Initial phase.
$\phi_{0,f,p}$	Initial FM phase.
$t_{60,p}$	Time until amplitude of partial has dropped 60 dB.
$t_{\text{attack},p}$	Time duration of attack portion.
$A_{f,p}$	Amplitude of FM.
$f_{f,p}$	Frequency of FM.
s_p	The signal representing the pth source.
$a_{60,p}$	The slope of the line in the argument of the exponential describing the amplitude variation.
$a_{k,60,p}$	The coefficient of the harmonic number in the argument of the exponential describing the initial amplitude of a harmonic as a function of its harmonic number.

Table 6.1. Synthesis parameters. Time values are in seconds, frequency values are in Hz and phase values are in radians.

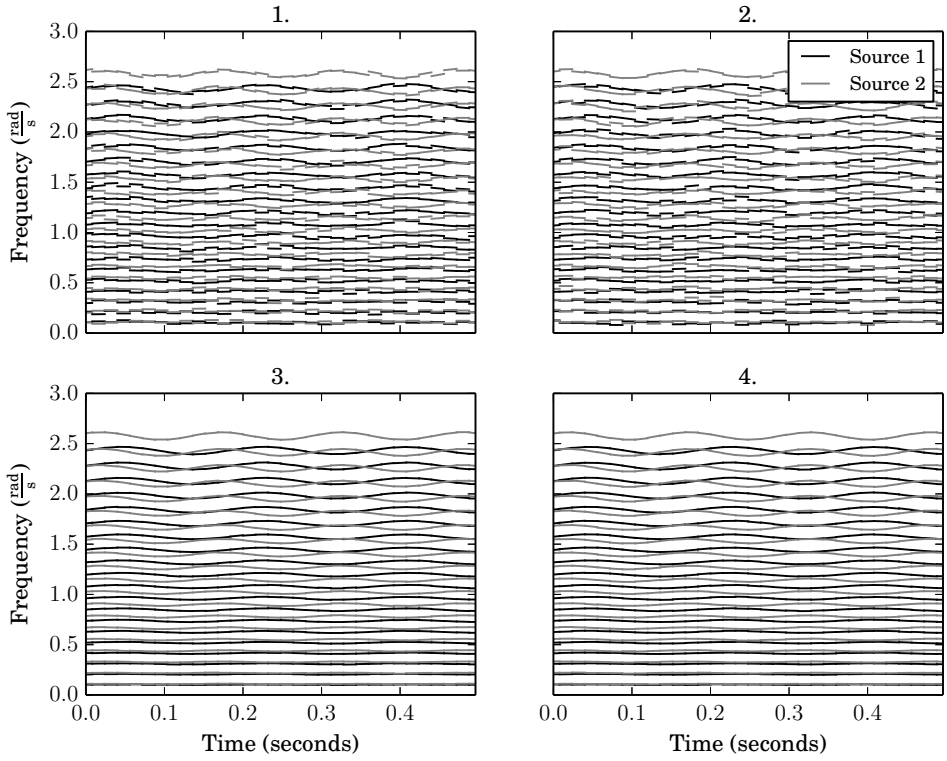

Fig. 6.1: **Original data-points.** Line-segments describing the frequency and frequency-modulation of both sources with no added spurious parameters. The amount of noise added and the corresponding plot title number are summarized in Table 6.3.

we define the *stretched* harmonic numbers as follows [56][1]

$$\mathcal{K}_B(k) = k(1 + Bk^2)^{\frac{1}{2}}$$

Each source is synthesized using the following equation:

$$s_p(t) = \sum_{k=1}^{K_p} A_p(k,t) \exp(j((2\pi f_{0,p} t - \frac{A_{f,p}}{f_{f,p}} \cos(2\pi f_{f,p} t + \phi_{0,f,p}))\mathcal{K}_{B_p}(k) + \phi_{0,p}))$$

[1] `http://ccrma.stanford.edu/~jos/pasp/Dispersion_Filter_Design_I.html`

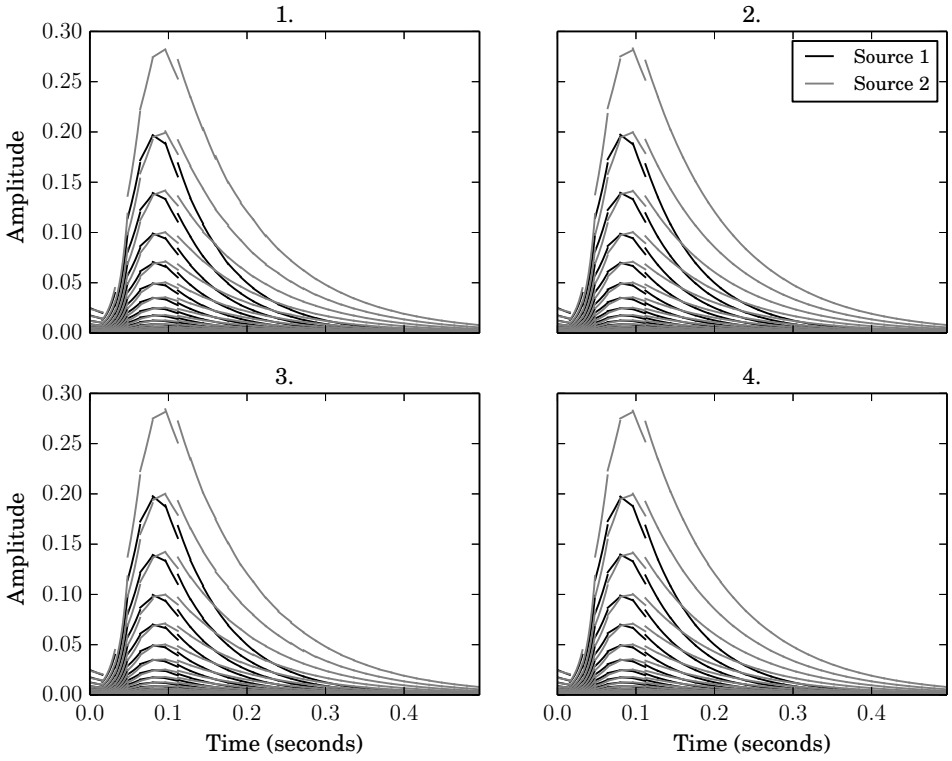

Fig. 6.2: **Amplitude function for each source (true).** Line-segments representing the instantaneous amplitude and amplitude slope at analysis time points. The amount of noise added and the corresponding plot title number are summarized in Table 6.3.

where

$$A_p(k,t) = \begin{cases} \exp(a_{60,p}t + a_{k,60,p}k)\cos^2(\frac{\pi}{2}(\frac{t}{t_{\text{attack},p}} - 1)) & \text{if } t \leq t_{\text{attack},p}, \\ \exp(a_{60,p}t + a_{k,60,p}k) & \text{if } t > t_{\text{attack},p}, \\ 0 & \text{otherwise.} \end{cases}$$

$$a_{60,p} = \frac{\log(10^{-3})}{t_{60,p}}$$

$$a_{k,60,p} = \frac{\log(10^{-3})}{k_{60,p}}$$

The piecewise amplitude function is based on the amplitude function of the *Formant Wave Function (FOF)*[2] described in [48, p. 19].

[2]FOF stands for *Forme d'Onde Formantique*.

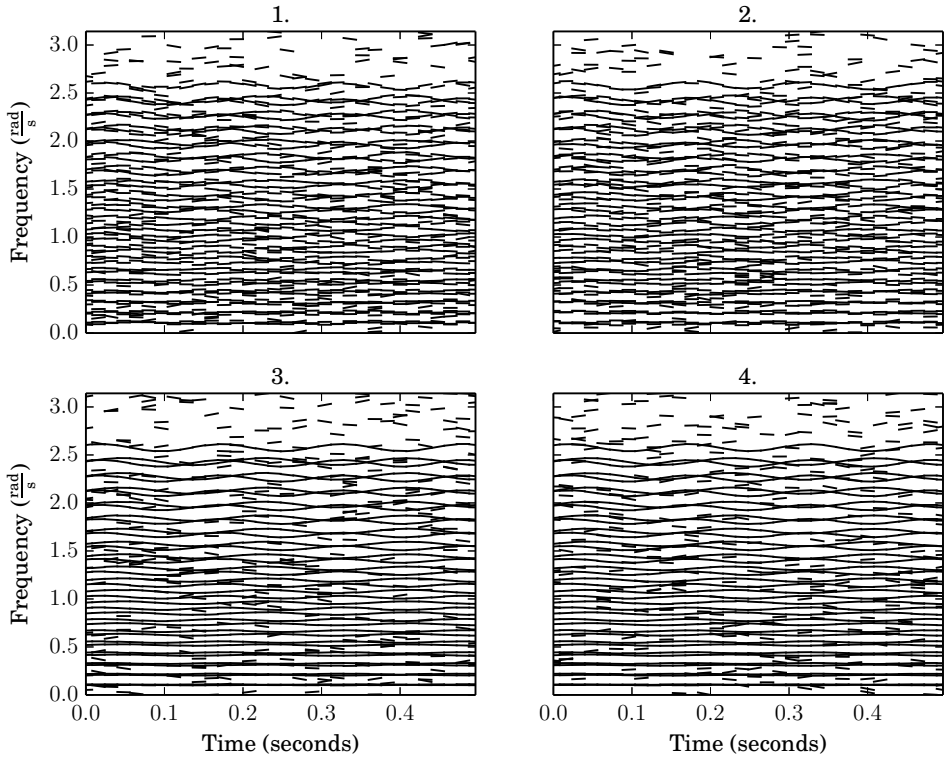

Fig. 6.3: **Original and spurious data-points.** Line-segments describing the frequency and frequency-modulation of the original data and the spurious data. The amount of noise added and the corresponding plot title number are summarized in Table 6.3.

6.5 Analysis

The estimation of these parameters is a separate problem addressed by the DDM (see Section 3.4). We use theoretical values calculated directly from the model signals. For interpretation, and to make it possible to simply replace the theoretical values with those obtained from an analysis, we compute parameters that correspond to a model whose parameters could be estimated through a technique such as the DDM.

[3]These are the fundamental frequencies of a C_4 and $C_4^\sharp$ respectively.

[4]These values are found by computing $f_{0,p}2^{1/48} - f_{0,p}$ giving ± 12.5 cents of frequency-modulation centred around the fundamental frequency.

Parameter	Source 1 value	Source 2 value
$f_{0,p}$	261.63	277.18 [3]
K_p	20	20
$k_{60,p}$	20	20
B_p	0.001	0.001
$\phi_{0,p}$	0	0
$\phi_{0,f,p}$	0	0.8
$t_{60,p}$	0.5	0.75
$t_{\text{attack},p}$	0.1	0.1
$A_{f,p}$	3.805	4.032 [4]
$f_{f,p}$	6.5	5.5

Table 6.2. Synthesis parameters for source separation by frequency- and amplitude-modulation.

Title number	ψ_{no}	ω_{no}	α_{no}	A_{no}
1	1.0×10^{-2}	1.0×10^{-2}	1.0×10^{-2}	3.2×10^{-5}
2	1.0×10^{-2}	1.0×10^{-2}	1.0×10^{-2}	1.0×10^{-5}
3	1.0×10^{-3}	1.0×10^{-3}	1.0×10^{-3}	3.2×10^{-5}
4	1.0×10^{-3}	1.0×10^{-3}	1.0×10^{-3}	1.0×10^{-5}

Table 6.3. The plot title numbers and the amount of noise added to the synthesis parameters for that realisation.

For this experiment we seek signals $s_k \in \mathbb{C}$ of the following form:

$$s_k(n) = \exp(\log(A_k) + \alpha_k n + j(\phi_k + \omega_k n + \frac{1}{2}\psi_k n^2)) \tag{6.1}$$

Here n is the sample number. This is the model of a sinusoid with linear amplitude-modulation and quadratic phase-modulation. We compute from the synthesis model what these parameters would be and add noise to simulate measurement error.

Typically when performing a short-time analysis, the time corresponding to $n = 0$ is made to be the centre of the window, therefore, t is the time at the centre of the window and N_w, in samples, is the length of the middle (usually non-zero) portion of the window.

6.5.1 The amplitude signal

The coefficients describing the amplitude-slope of the kth harmonic of the pth source from our synthetic model are given by

$$\alpha_{k,p}(t) = \frac{a_{60,p}}{f_s}$$

$$A_{k,p}(t) = \exp(a_{60,p}t + a_{k,60,p}k)$$

for the part of the signal after the attack portion. Note that the amplitude-slope is not time-varying.

For the attack portion, we estimate the amplitude parameters of Equation 6.1 using least-squares on a rectangular-windowed signal[5]. Let

$$\hat{\mathbf{s}}_{k,p}(t_n) = \begin{pmatrix} Q(t_n - \frac{N_w}{2f_s}) \\ Q(t_n - \frac{N_w}{2f_s} + 1) \\ \vdots \\ Q(t_n + \frac{N_w}{2f_s} - 1) \\ Q(t_n + \frac{N_w}{2f_s}) \end{pmatrix}$$

with

$$Q(t) = \exp\left(a_{60,p}\left(t\right) + a_{k,60,p}k\right)\cos^2\left(\frac{\pi}{2}\left(\frac{t}{t_{\text{attack},p}} - 1\right)\right)$$

the function describing the amplitude during the attack portion of the signal. Then $\log(A_{k,p})$ and $\alpha_{k,p}$ are found as the least-squares solution of

$$\begin{bmatrix} 1 & \frac{-N_w}{2} \\ \vdots & \vdots \\ 1 & \frac{N_w}{2} \end{bmatrix} \begin{pmatrix} \log(A_{k,p}(t_n)) \\ \alpha_{k,p}(t_n) \end{pmatrix} = \log \hat{\mathbf{s}}_{k,p}(t_n)$$

where $t_n = \frac{n}{f_s}$ is the time in seconds at sample n. For the argument parameters (those multiplied by j in Equation (6.1))

$$\phi_k\left(t\right) = \left(2\pi f_{0,p}t - \frac{A_{f,p}}{f_{f,p}}\cos\left(2\pi f_{f,p}t + \phi_{0,f,p}\right)\right)\mathcal{K}_{B_p}\left(k\right) + \phi_{0,p}$$

[5]We cannot simply compute the modulation parameters of Equation 6.1 using the Taylor-series expansion of $Q(t)$ because the attack portion is an exponential function modulated by a raised cosine, which does not match the analysis model.

$$\omega_{k,p}(t) = \frac{2\pi}{f_s}\left(f_{0,p} + A_{f,p}\sin\left(2\pi f_{f,p}t + \phi_{0,f,p}\right)\right)\mathcal{K}_{B_p}(k)$$

$$\psi_{k,p}(t) = \left(\frac{2\pi}{f_s}\right)^2 A_{f,p}f_{f,p}\cos\left(2\pi f_{f,p}t + \phi_{0,f,p}\right)\mathcal{K}_{B_p}(k)$$

To simulate the noise that would be present in an estimation of the signal parameters from an arbitrary signal, we create noise corrupted values by substituting the random variables:

- $\tilde{\psi}_{k,p}(t) \sim \psi_{k,p}(t) + \mathcal{N}(0, \psi_{no})$
- $\tilde{\omega}_{k,p}(t) \sim \omega_{k,p}(t) + \mathcal{N}(0, \omega_{no})$
- $\tilde{\alpha}_{k,p}(t) \sim \alpha_{k,p}(t) + \mathcal{N}(0, \alpha_{no})$
- $\tilde{A}_{k,p}(t) \sim A_{k,p}(t) + \mathcal{N}(0, A_{no})$

The θ_{no} (where θ is replaced by ψ etc.) specifies the variance of the particular parameter. Most likely in practice these random variables would be correlated but not knowing the estimation method, we cannot at this point say anything about this correlation. Therefore the noisy parameters are uncorrelated random variables for this experiment.

We also add spurious data-points as a fraction r of the number of true data-points. Their values are drawn from uniform distributions with boundaries $\theta_{\min}$ and $\theta_{\max}$, where θ is some parameter above, e.g., $\omega_{\min}$ and $\omega_{\max}$ for the ω parameter. For this experiment $r = 0.25$, which is quite a large number of spurious points. This value is chosen to show that, given an acceptable accuracy of estimation of the true parameters, good source separation results can be achieved, even with such a high proportion of spurious points. The parameters of the uniformly distributed random variables are given in Table 6.4. Data-points are computed for the times $t_n = 0, \frac{H}{f_s}, \frac{2H}{f_s}, \ldots, \frac{\lfloor \frac{N}{H} \rfloor H}{f_s}$.

Table 6.4. Distribution parameters of uniformly distributed random variables

Parameter	$\theta_{\min}$	$\theta_{\max}$
ω	0	π
ψ	-1×10^{-4}	1×10^{-4}
α	-1×10^{-3}	1×10^{-3}

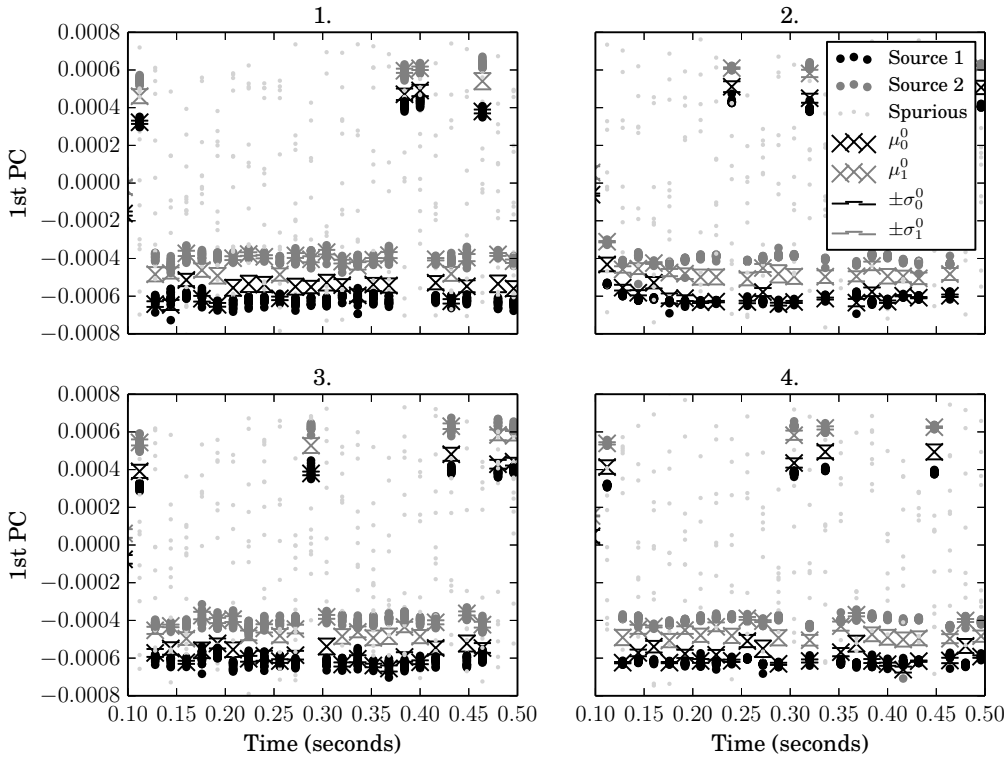

Fig. 6.4: **Principal components and their classification.** The PCs for each theoretical analysis time-point. μ_i^0 and σ_i^0 are respectively the initial mean and standard deviation guesses for the EM algorithm fitting the GMM parameters to the ith source. These values are also visible on the plot. The spurious points rejected using the process described in Section 6.7 are included for comparison. The amount of noise added and the corresponding plot title number are summarized in Table 6.3.

6.6 Computation of principal components

At each time t we have L data-points. As the source of each data-point is now unknown, we replace the k and p indices with index l. We only consider the amplitude- and frequency-modulation. According to our model, the frequency-modulation is greater for harmonics of greater centre frequency. To take this into consideration, we divide the frequency-modulation estimate $\psi_l(t)$ by the constant frequency estimate $\omega_l(t)$. This is similar to the approach taken in [6]. The amplitude-modulation $\alpha_l(t)$ remains constant for all harmonics of the same source[6], only its initial value changes according to $k_{60,p}$. We compile the data-points at one time into a set of observations.

$$\mathbf{x}_l(t) = \begin{pmatrix} \frac{\psi_l(t)}{\omega_l(t)} \\ \alpha_l \end{pmatrix}$$

$$\mathbf{X}(t) = \Big[\mathbf{x}_1(t) \dots \mathbf{x}_L(t)\Big]$$

From these L observations the correlation matrix $\mathbf{S}$ is computed[7]. We use the correlation matrix because the values in each row of $\mathbf{x}_l(t)$ do not have the same units, see [24, p. 22] for a discussion about this.

Following the standard technique for producing principal components (see Appendix A and also [24, p. 11]), we obtain a matrix $\mathbf{V}(t)$ of eigenvectors sorted so that the eigenvector corresponding to the largest eigenvalue is in the first column, etc. The principal components $\mathbf{A}(t)$ are then computed as

$$\mathbf{A}(t) = \mathbf{V}^T(t)\mathbf{X}(t)$$

[6] We acknowledge that this might not be realistic for all sounds. If the amplitude-modulation is a function of (normalized) frequency as well as time $\beta_l(t) = \mathcal{A}(\omega)$ we need only perform the transformation $\alpha_l(t) = \mathcal{A}^{-1}(\beta_l(t))$ to obtain the same data points as classified here.

[7] If we have N samples of random variables X_i and X_j, the entry in the ith row and jth column of correlation matrix $\boldsymbol{S}$ is their estimated correlation, i.e.,

$$S_{i,j} = \frac{\sum_{n=1}^{N}(x_{i,n} - \overline{x}_i)(x_{j,n} - \overline{x}_j)}{\sqrt{\sum_{n=1}^{N}(x_{i,n} - \overline{x}_i)^2 \sum_{n=1}^{N}(x_{j,n} - \overline{x}_j)^2}}$$

where $\overline{x}_i = \frac{1}{N}\sum_{n=1}^{N} x_{i,n}$, the sample mean [21, p. 66].

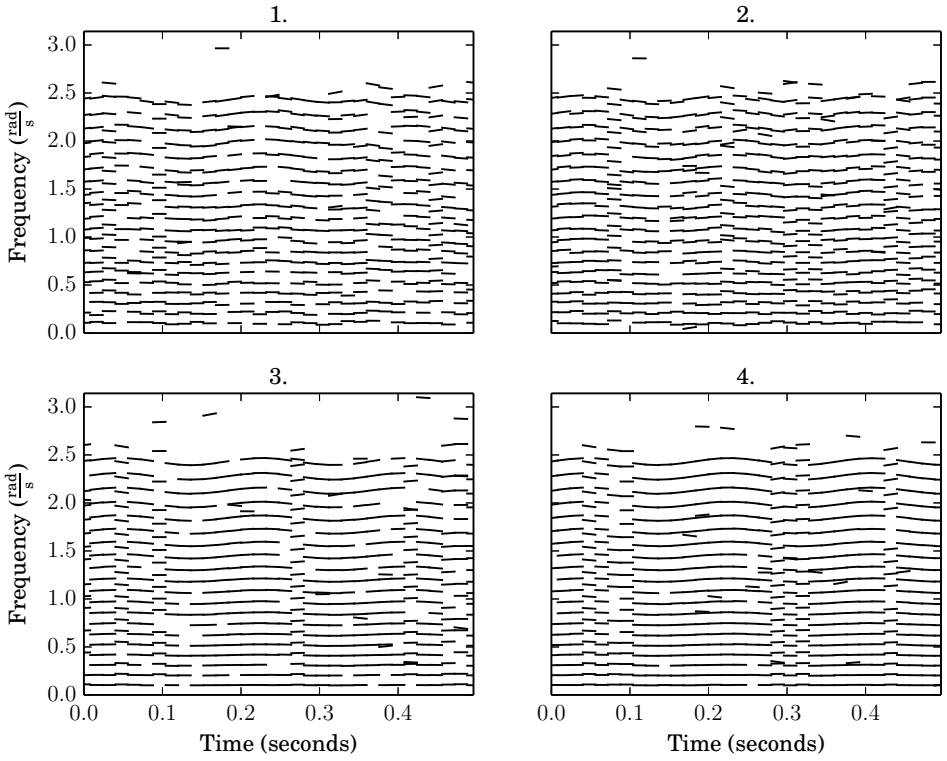

Fig. 6.5: **Source 1 (estimated).** Line-segments classified as belonging to Source 1. The classification is done on each frame so classifications in consecutive frames may not belong to the same true source. This is because the ordering of the clusters in each frame in Figure 6.4 is not predictable. The amount of noise added and the corresponding plot title number are summarized in Table 6.3.

We have found it sufficient to use only the first principal component and therefore only use the values in the first row of $\mathbf{A}(t)$. The ith principal component of sample l at time t is written $a_{i,l}(t)$.

If we see the $\mathbf{x}_l(t)$ as realizations of a random variable, the above computation of principal components has the effect of projecting realizations of $\mathbf{x}_l(t)$ to points $a_{1,l}(t)$ on a 1-dimensional subspace. It is a fundamental theory of principal components that the transformation above maximizes the expected Euclidean distance between the points $a_{1,l}(t)$. This is desirable for the current problem because it will always produce a variable emphasizing the parameter with the most variance. More specifically, if a scatter plot of the frequency-modulation measurements shows multiple distinct clusters whereas the

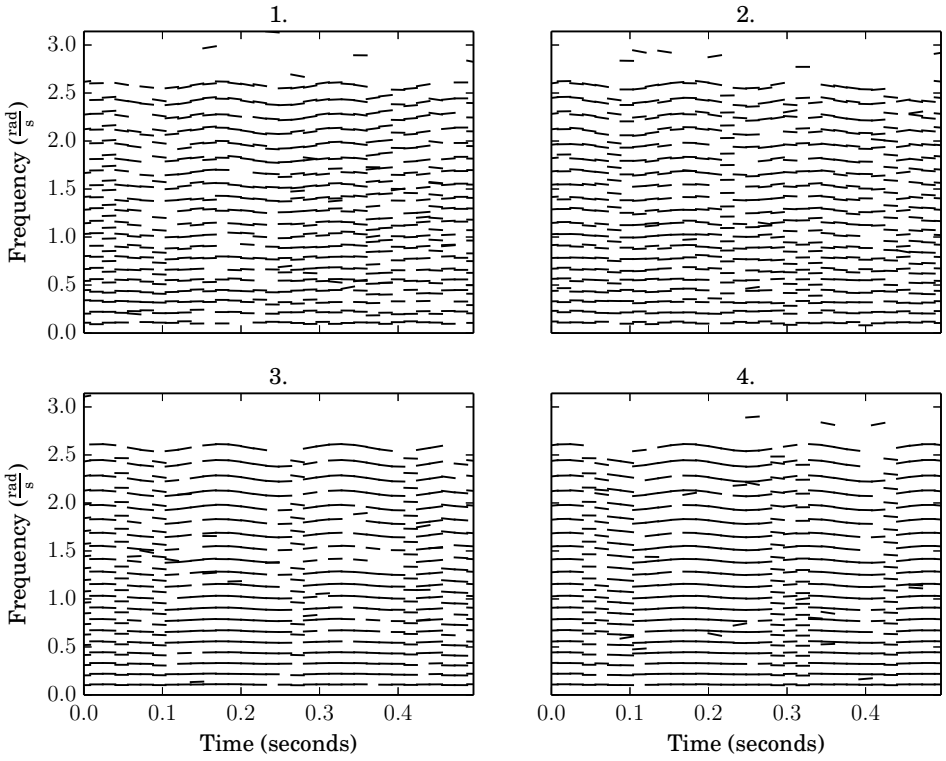

Fig. 6.6: **Source 2 (estimated).** Line-segments classified as belonging to Source 2. See Figure 6.5 for more information. The amount of noise added and the corresponding plot title number are summarized in Table 6.3.

amplitude-modulation measurements are all close and show only one cluster in a scatter plot, the first PC will emphasize the frequency-modulation measurements, which we desire for ease of clustering. The drawback of this approach is that if one parameter is very noisy and the other is not, the noisy parameter will be emphasized but forming informative clusters will be difficult. In that case it would be better to reject this parameter or use more PCs on which to perform clustering.

6.7 Preparing data for clustering

The EM underlying the Gaussian mixture model parameter estimation can converge to a local maximum [8], therefore, for the best results, we compute a good initial guess and remove obvious outliers before carrying out the clustering algorithm.

The $a_{1,l}(t)$ are compiled into a histogram of N_b bins. The minimum and maximum bin boundaries are computed from the maximum and minimum values of $a_{1,l}(t)$ respectively. Values in a bin with less than λ_h other values are discarded. We find N contiguous sections of equal area in the new histogram omitting the discarded values. We use the centres of these sections as the initial mean guesses and half their width as the distance 3 standard deviations from the mean (roughly 99.7 percent of values drawn from one distribution will lie within this interval if they follow a normal distribution). The initial guesses for the weights are simply $\frac{1}{N}$.

6.8 Clustering

GMM parameter estimation is discussed in Section B. After convergence we have an estimated probability p($a_{1,l}(t)$ from distribution p). We choose the distribution p for each $a_{1,l}(t)$ that gives the highest probability of it having occurred. The values $\mathbf{x}_t(t)$ corresponding to the $a_{1,l}(t)$ have this same classification. Those sharing the same classification can be interpreted as coming from the same source. The figure shows the results of the above steps carried out on a mixture of two sources synthesized with the parameters summarized in Table 6.2. The length of the signal N is 8000 samples and the analysis hop size H is 256 samples.

6.9 Results

Here we show source separation results for the synthesized signals with noise added to the synthesis parameters ψ, ω, α and A. Each plot shows four realisations with varying amounts of noise. The amount of noise added and the corresponding plot title number are summarized in Table 6.3.

The Figures 6.1, 6.3, 6.4, 6.5, 6.6 and 6.9 summarize the results of the source separation experiment before the smoothing step. Figure 6.1 shows a time-frequency representation of the original data, Figure 6.3 shows the original data with spurious data added, Figure 6.4 shows the principal components at each time frame and the initial guesses for the EM algorithm, Figure 6.5 shows the initial classifications for source 1 and Figure 6.6 the initial classifications for source 2 before the smoothing step.

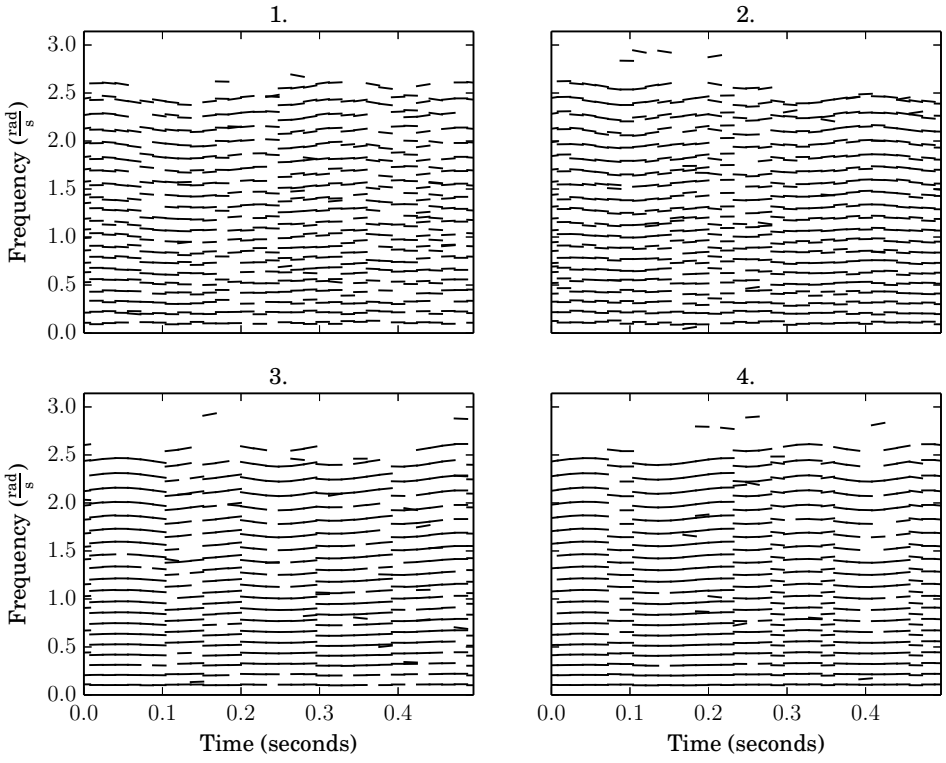

Fig. 6.7: **Source 1 (estimated) after smooth amplitude path search.** Line-segments classified as belonging to source 1 after smoothing in amplitude using $\mathcal{D}_a$. The amount of noise added and the corresponding plot title number are summarized in Table 6.3.

As seen in Figure 6.5 and 6.6, while classified well in individual frames, the overall classification does not always correspond to a single source. We must find a collection of frames with high plausibility of belonging to one source. We consider the collections of classified data-points corresponding to each source as a node in a lattice. Each frame of the lattice contains two nodes, one for each source. A best path through the lattice should connect together those nodes belonging to a single source. We use the results of Section 4.2 to find the two best paths through this lattice. We compare two distance metrics for the cost function.

The first prefers smoothness in frequency between two frames. For frame h with initial classification $\tilde{p}$ we have frequency measurements $\omega^h_{k,\tilde{p}}$ and frequency slope measurements $\psi^h_{k,\tilde{p}}$. The set of parameters at time h from initially classified source $\tilde{p}$ we will denote $\boldsymbol{\theta}^h_{\tilde{p}}$. Between frame h and frame $h+1$ we use Algorithm 1 on the pairs $\left\{\boldsymbol{\theta}^h_{\tilde{m}}, \boldsymbol{\theta}^{h+1}_{\tilde{n}}\right\}$ with

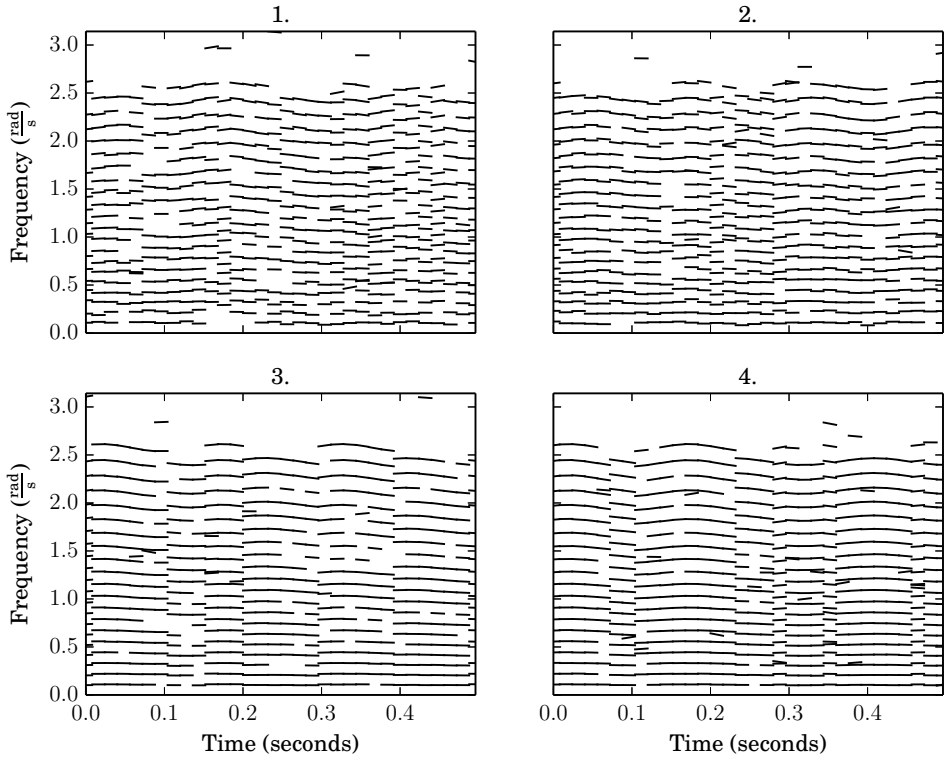

Fig. 6.8: **Source 2 (estimated) after smooth amplitude path search.** Line-segments classified as belonging to source 2 after smoothing in amplitude. The amount of noise added and the corresponding plot title number are summarized in Table 6.3.

$(m, n) \in \{0, 1\} \times \{0, 1\}$. For each pair, L is set to $\min(\#\boldsymbol{\theta}^h_{\tilde{m}}, \#\boldsymbol{\theta}^{h+1}_{\tilde{n}})$.[8] The cost function is the absolute error in predicting the frequency in the next frame from parameters in the current frame, i.e.,

$$\mathcal{D}_f\left(\theta^h_{i,\tilde{m}}, \theta^{h+1}_{j,\tilde{n}}\right) = |\omega^h_{i,\tilde{m}} + \psi^h_{i,\tilde{m}} H - \omega^{h+1}_{j,\tilde{n}}|$$

where H is the hop-size in samples between the two frames. The second distance metric measures the smoothness in amplitude between two frames by predicting the next frame's amplitude parameters using the amplitude and amplitude-modulation parameters of the current frame. It is given as

$$\mathcal{D}_a\left(\theta^h_{i,\tilde{m}}, \theta^{h+1}_{j,\tilde{n}}\right) = |\log(A^h_{i,\tilde{m}}) + \alpha^h_{i,\tilde{m}} H - \log(A^{h+1}_{j,\tilde{n}})|$$

[8]Here, the threshold parameter $\Delta = \infty$, i.e., a connection of any cost is possible.

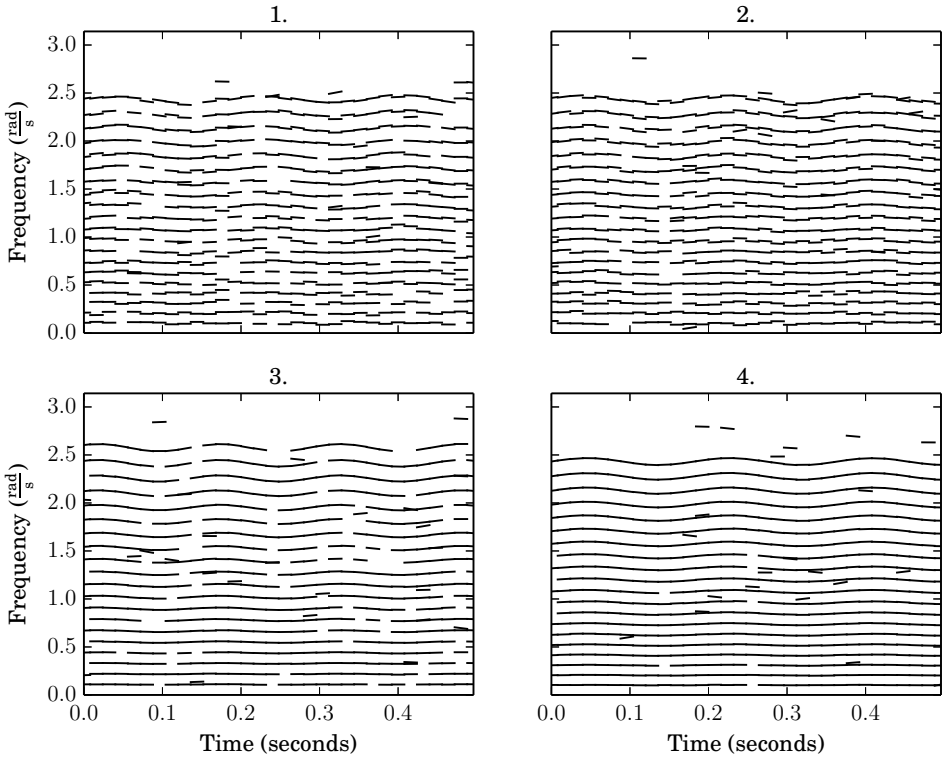

Fig. 6.9: **Source 1 (estimated) after smooth frequency path search.** Line-segments classified as belonging to source 1 after smoothing in frequency. The amount of noise added and the corresponding plot title number are summarized in Table 6.3.

We have found the absolute error to give better results than the squared error.

The costs of these connections are summed over the index pairs Γ_h to give the entries of the cost vector $\boldsymbol{c}$ in the LP[9]

$$c_{4h+2m+n} = \sum_{i^*,j^* \in \Gamma_h} \mathcal{D}\left(\theta^h_{i^*,\tilde{m}}, \theta^{h+1}_{j^*,\tilde{n}}\right)$$

The specification of the constraint matrices is done according to the topology of the lattice and the requirement that we find 2 non-overlapping paths (see Section 4.2). An example of

[9]The indexing of $\boldsymbol{c}$ is explained as follows. There are 4 possible classification connections between frame h and $h+1$. Each source m at time h has a cost of being associated with a source n at time $h+1$, which is stored in the index $4h+2m+n$. This is merely how the indices are laid out in the array representing the cost vector $\boldsymbol{c}$. See Section 4.2 for more about $\boldsymbol{c}$.

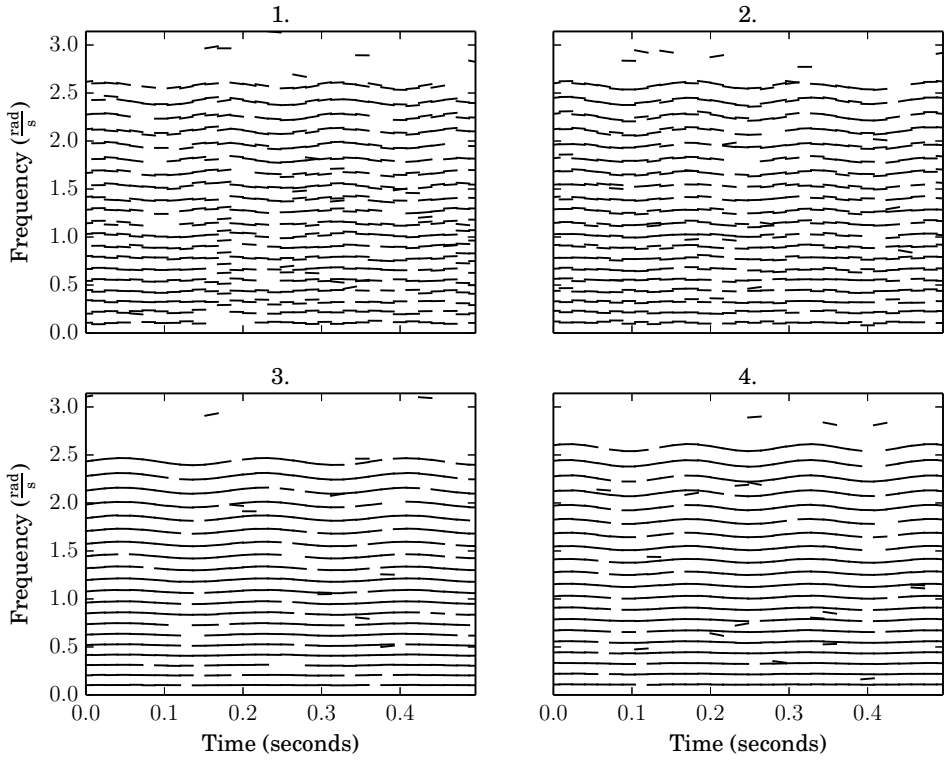

Fig. 6.10: **Source 2 (estimated) after smooth frequency path search.** Line-segments classified as belonging to source 2 after smoothing in frequency using $\mathcal{D}_f$. The amount of noise added and the corresponding plot title number are summarized in Table 6.3.

discovered paths is given in Figure 6.11. The estimated sources after smoothing in frequency using $\mathcal{D}_f$ are shown in Figures 6.9 and 6.10. The estimated sources after smoothing in amplitude using $\mathcal{D}_a$ are shown in Figures 6.7 and 6.8. We see that when smoothed in frequency, the results are acceptable. However, when both sets of parameters are close and give close costs, the spurious data-points can influence the cost function causing a false classification. This difficulty is not surprising, looking at Figure 6.1 we see that there are some segments where the frequency slopes are close.

When smoothed in amplitude, the results are less convincing. This is not surprising as smoothness in amplitude is not the best criterion at all time points. In Figure 6.2 we see that the amplitudes of both sources are similar at many points, e.g., at around 0.05 and 0.15 seconds.

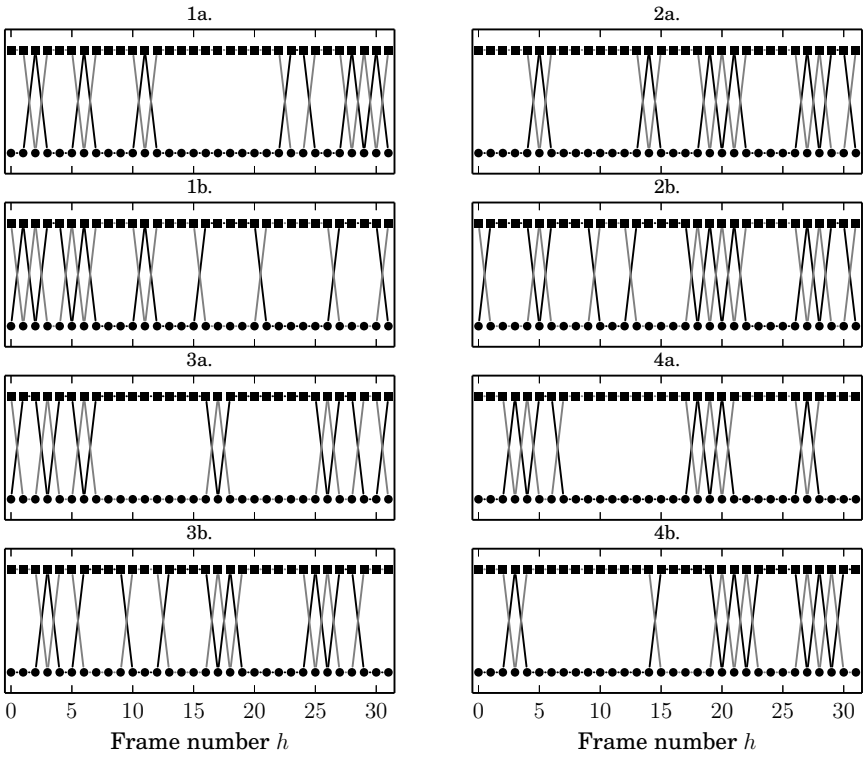

Fig. 6.11: **Smoothed paths.** The points originally classified as source 1 are marked with circles and those originally classified as source 2 squares. The paths in black or grey connect the points for source 1 or source 2 respectively with optimal smoothness. Plots indicated with an a are the paths with frequency smoothness as the criterion and those indicated with b are with amplitude smoothness as the criterion. The amount of noise added and the corresponding plot title number are summarized in Table 6.3.

6.10 Conclusion

In this chapter we evaluated the plausibility of separating two mixed sources based on their theoretical frequency- and amplitude-modulation. We obtained acceptable results for signals with small measurement errors. The method is also robust in the presence of spurious data points. A shortcoming of the method is the requirement that the frequency and frequency-modulation of the signals be known. Although for this experiment synthetic data were used, if the signals are sufficiently separated in frequency and have small bandwidth, as shown in Section 3.4, the DDM can be used to estimate these parameters. There are also techniques for estimating amplitude- and frequency-modulation that were

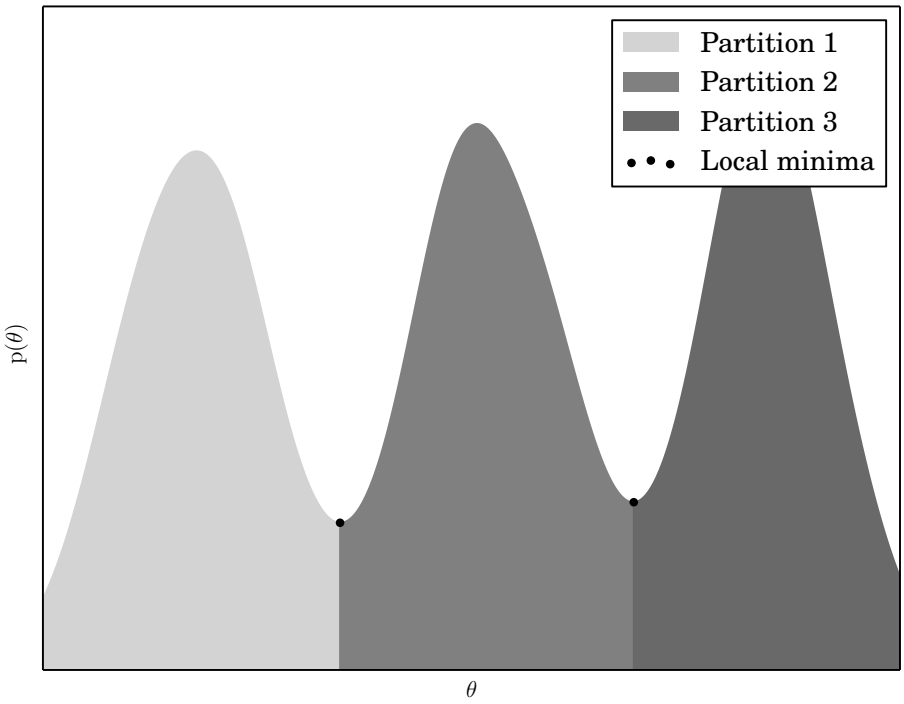

Fig. 6.12: **Paritioning example.** The data points are convolved with some smoothing kernels giving a function with a small number of extrema. The minima, indicated by circles, are used as boundaries between the partitions which are illustrated with different shades of grey.

not explored in this thesis. If signals are close in frequency, but the number of partials is known, and these exhibit slow modulations, signal subspace methods could be used [50] where the estimations at different time points are connected as in Chapter 4 and the modulation parameters postulated via interpolation similarly to Section 3.3.1. If it is possible to make uncorrupted measurements of the two signals and identify where the signals are not easily identified (e.g., the locations that their partials cross in the time-frequency plane) a strategy might be to use two measurements of one source and extrapolate the parameters of the signal in the part corrupted by the other source. A similar approach is explored in [27] where the Hough transform is used to identify crossing partials. Another shortcoming of the technique presented here is the use of the costly EM algorithm to classify data points using GMM (see Appendix B). A more ad hoc approach could be taken to save on

these computations, perhaps partitioning the data sets using local minima as illustrated in Figure 6.12. In any case, the source separation technique presented here, being iterative, is of a complexity similar to NMF or PLCA but can also resolve the phases of the sinusoids which are discarded in most NMF or PLCA implementations[10].

[10]See [4] for an approach that does take into consideration the phase information in the spectrogram.

Chapter 7

Experiment: Separation of two sources using partial decay rate

7.1 Introduction

In this section we demonstrate how the techniques described above can be used to perform audio source separation on signals obtained from recordings of acoustic instruments. Specifically, we show that in the absence of frequency-modulation, amplitude modulation — the decay rate — can be used to classify partials in a mixture of two sources into two groups, each group representing an underlying source.

7.2 Description of problem

We start with a recording of an acoustic guitar playing A_3 and a xylophone playing $F^{\sharp}_4$. The recordings are from [43] and have been mixed down to one channel (by simply adding the two signals together) and resampled at 16 kHz, coded simply as a stream of 64-bit floating-point numbers. Spectrograms of the original signals are shown in Figure 7.1 and Figure 7.2. The spectrograms were produced with a Hann window, DFT size of 4096 samples and a hop size of 512 samples. We see that neither source exhibits much frequency-modulation. The spectrogram of the mixture can be seen in Figure 7.3 and the partial paths in Figure 7.4.

The mixture of the two signals was analysed using the DDM for finding the coefficients of a cubic complex phase polynomial. Local maxima in each frame were found using the technique described in [52, p. 42]. For each of these local maxima, the polynomial

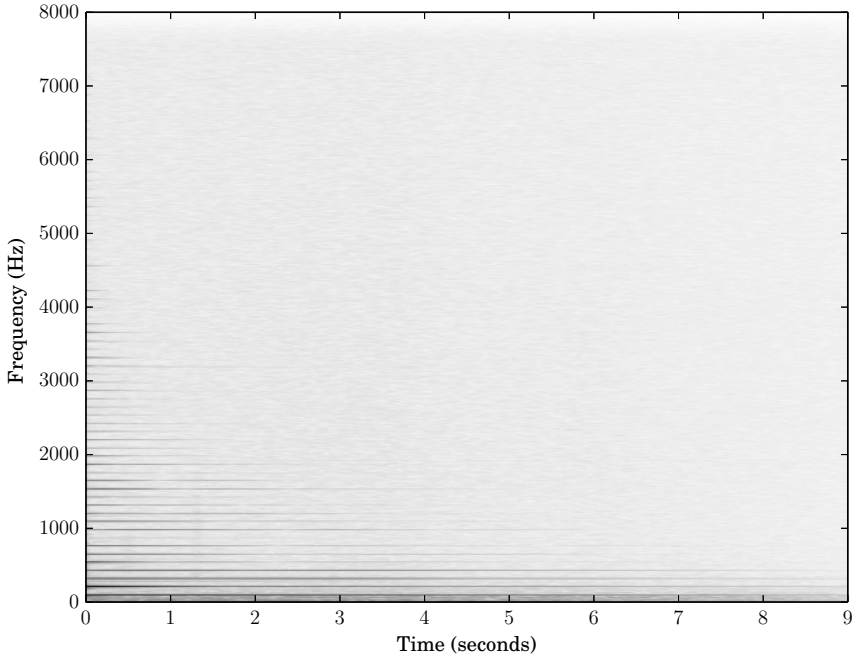

Fig. 7.1: **Spectrogram of acoustic guitar.** The fundamental is A_3 or 220 Hz.

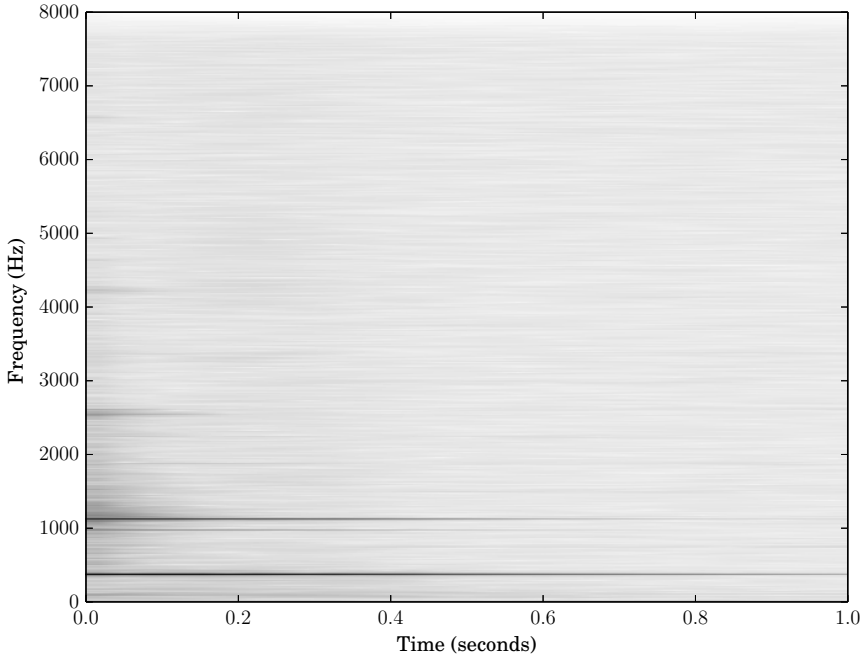

Fig. 7.2: **Spectrogram of xylophone.** The fundamental is $F_4^\sharp$ or approximately 370 Hz.

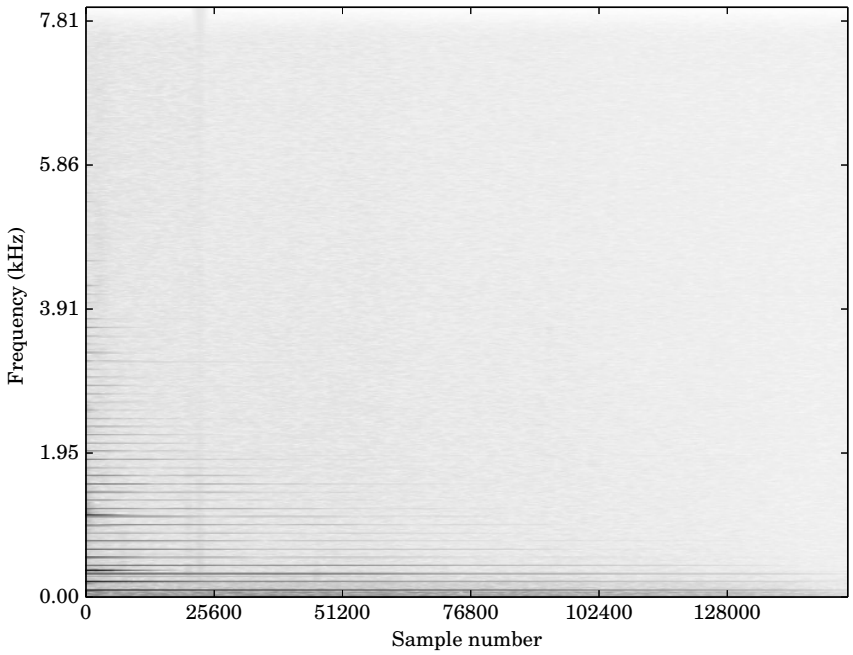

Fig. 7.3: **Spectrogram of mixture.** This spectrogram is the sum of the spectrograms in Figure 7.1 and 7.2.

coefficients were estimated. The analysis used the $\mathcal{C}^1$ 4-Term Blackman-Harris window that was designed in Chapter 3. To obtain partials it was then necessary to connect the local maxima. As the partials of these two sound sources are quite stable in frequency it sufficed to use the Viterbi algorithm [13] and the cost metric $\mathcal{D}_{\text{pr.}}$ from Section 4.3 to connect local maxima in sub-bands of the spectrum. The cost function is simply the Euclidean distance between the frequencies of two local maxima. Partial starting points are considered in sub-bands of width 15 Hz and these sub-bands overlap by 7.5 Hz. A partial path starts on the first local maximum in the band exceeding -100 dB and ends at the last maximum exceeding -100 dB. The path search algorithm will also look ahead to further frames if no maximum is present in the next frame. Because of this, sometimes unrealistic paths are discovered that jump between spurious maxima. These are filtered out by discarding paths whose cost-length ratio is excessive. See Figure 7.6 to see a plot of these values and the thresholding function.

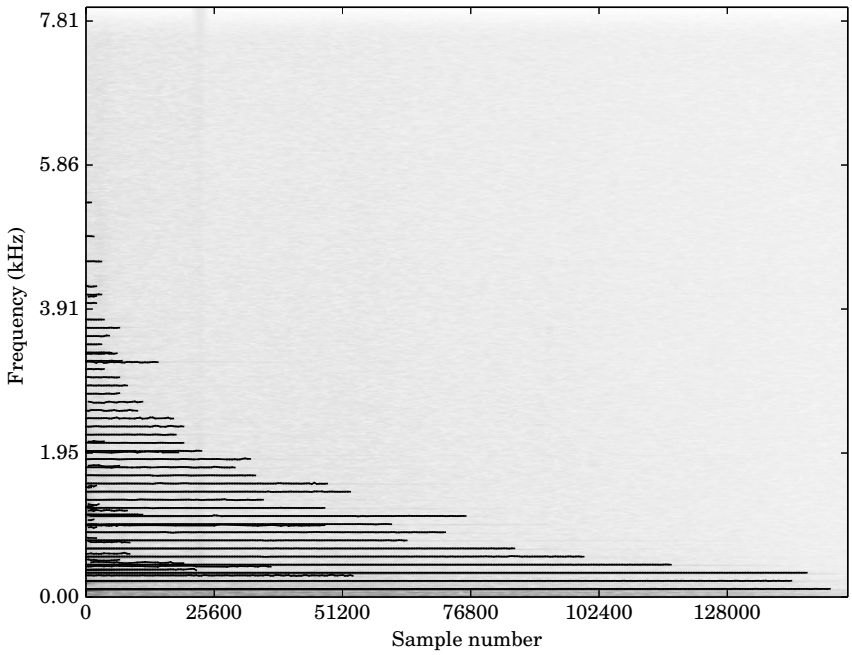

Fig. 7.4: **Spectrogram of mixture and partial trajectories**

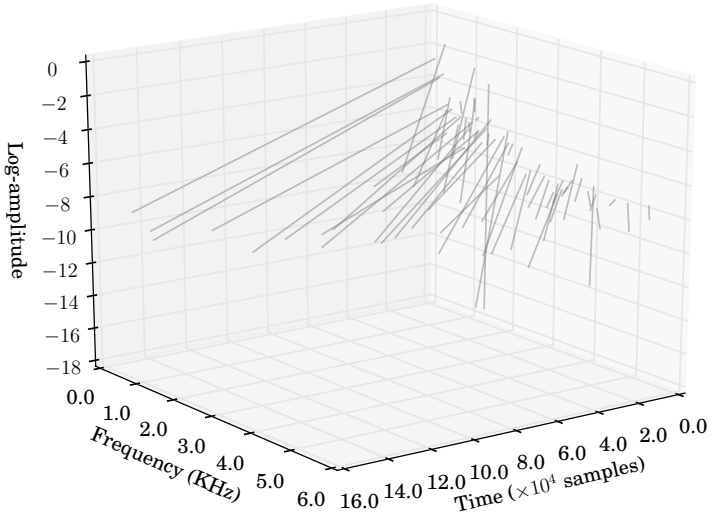

Fig. 7.5: **Partial trajectories.** Line functions are fit to the partial trajectory data of each partial to examine their general amplitude slopes.

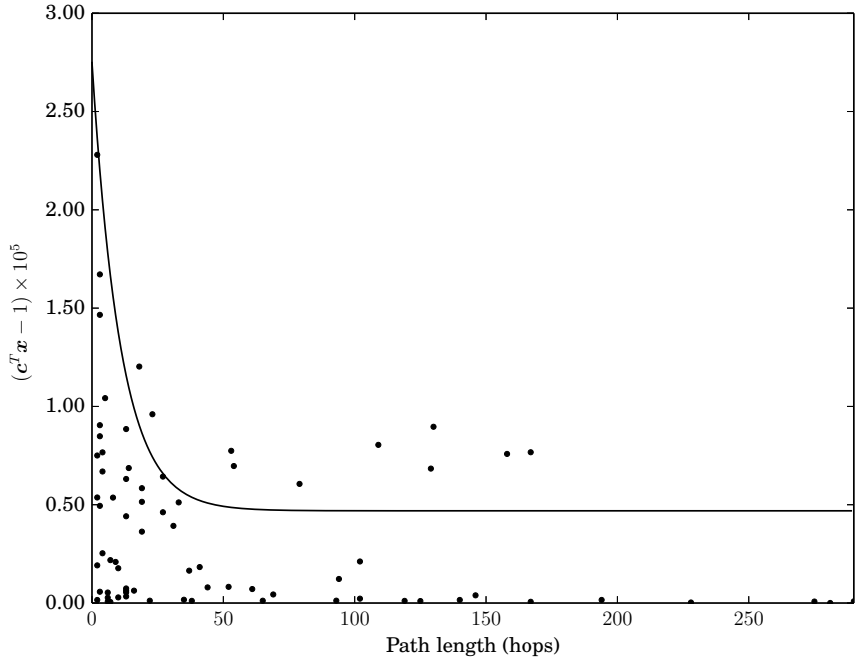

Fig. 7.6: **Path cost vs. length and thresholding boundary.** Paths, represented as circles, are considered only if their path cost is smaller than the threshold function, the black curve. The threshold is higher for shorter partials as to not reject those that represent the transient region of the sound. The partials during this time typically have rapidly changing frequency- and amplitude-modulation, so their path costs could be disproportionately high.

7.3 Motivation

Line functions (functions of time) are fit to the amplitude and frequency data on each partial trajectory via least-squares, as shown in Figure 7.5. Here roughly two kinds of partial slope with respect to amplitude are observed — those that are steep and brief and others that are longer and more gradual. Our goal is to classify based on the amplitude modulation of each partial, or to an approximation, the slope of these line functions. We found that examining the log-length of the partials gives better results than examining the slope directly. This is perhaps because the log-length encodes both the starting amplitude and the slope. Recall that the partials start on the first local maximum exceeding an amplitude threshold — those with lower starting amplitude and steeper amplitude slope will be shorter, while those with a higher starting amplitude and shallower slope will be

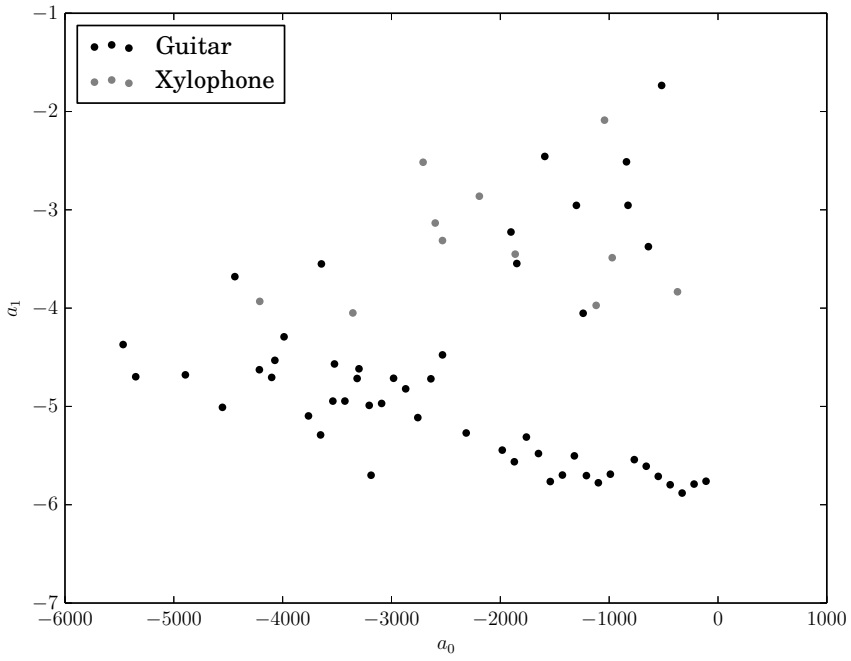

Fig. 7.7: **Log-partial-length vs. frequency: principal components.** This shows the distribution of partials when plotting their two principal components derived from their mean frequency and log-length. In this case, the source memberships of the partials are known. We see that there is generally a separation of the partials into two clusters corresponding to the two sources.

longer. We see in Figure 7.7 that using both the amplitude-slope and the initial amplitude of the partial gives clear separation in a plot of the log-length vs. the average frequency of the partials. These partials are from separate overlaid analyses of the guitar and xylophone signals. The experiment uses an analysis of a signal consisting of a mixture of the sources, of course.

The data-points have the form

$$\boldsymbol{a}_i = \begin{pmatrix} a_{i,0} \\ a_{i,1} \end{pmatrix}$$

where $a_{i,0}$ is the first principal components and $a_{i,1}$ the second and are computed via a linear transformation of

$$\boldsymbol{x}_i = \begin{pmatrix} \overline{f}_i \\ \ell_i \end{pmatrix}$$

where $\overline{f}_i$ is the mean frequency of the ith partial and ℓ_i its log-length (see Appendix A for the computation of principal components). The set of principal components will be denoted $\{\boldsymbol{a}\}$. We see that, for the most part, the partials belonging to the two sources are separated appropriately into two clusters. The partials from the xylophone present in the guitar cluster belong to higher partials, whose omission in the final rendering of the xylophone source would not be detrimental to its perceptual quality. Similarly, partials belonging to the guitar present in the xylophone cluster are short and most likely belong to briefly excited modes of the guitar body.

7.4 Classification

Our intention is now to use GMM (see Appendix B) on a set of unclassified partials to yield a plausible source separation. GMM fitting is sensitive to its initial guess of the parameters as the algorithm can converge to a local maximum of the likelihood function [26, p. 187]. To find an initial guess we convolve the scatter plot with kernel $\mathscr{K}$, giving a continuous function. $\mathscr{K}$ is defined[1]

$$\mathscr{K}(\boldsymbol{x}, \boldsymbol{\beta}) = \exp(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\beta}^{-1}\boldsymbol{x})$$

Here $\boldsymbol{x} \in \mathbb{R}^2$ and $\boldsymbol{\beta} \in \mathbb{R}^{2\times 2}$ controls the extent of the kernel, i.e., how much it smooths in each dimension.

We use the two local maxima of this function as the initial means for the two sought classifying Gaussian distributions. The convolution function evaluated at $\hat{\boldsymbol{a}}$ is

$$f(\hat{\boldsymbol{a}}) = \sum_{\boldsymbol{a}_i \in \{\boldsymbol{a}\}} \mathscr{K}\left(\hat{\boldsymbol{a}} - \boldsymbol{a}_i, \boldsymbol{\beta}_{\boldsymbol{a}}\right)$$

To make the variance proportional to the extent of each dimension, $\boldsymbol{\beta}_{\boldsymbol{a}}$ is defined as

$$\boldsymbol{\beta}_{\boldsymbol{a}} = \begin{pmatrix} \frac{\Delta_{\boldsymbol{a}_1}}{\Delta_{\boldsymbol{a}_0+\Delta_{\boldsymbol{a}_1}}}\theta_{\boldsymbol{\beta}_{\boldsymbol{a}}} & 0 \\ 0 & \frac{\Delta_{\boldsymbol{a}_0}}{\Delta_{\boldsymbol{a}_0+\Delta_{\boldsymbol{a}_1}}}\theta_{\boldsymbol{\beta}_{\boldsymbol{a}}} \end{pmatrix}$$

where

$$\Delta_{\boldsymbol{a}_0} = \max\left(\boldsymbol{a}_0\right) - \min\left(\boldsymbol{a}_0\right)$$

[1] Note its similarity to the normal distribution, defined in Appendix C.

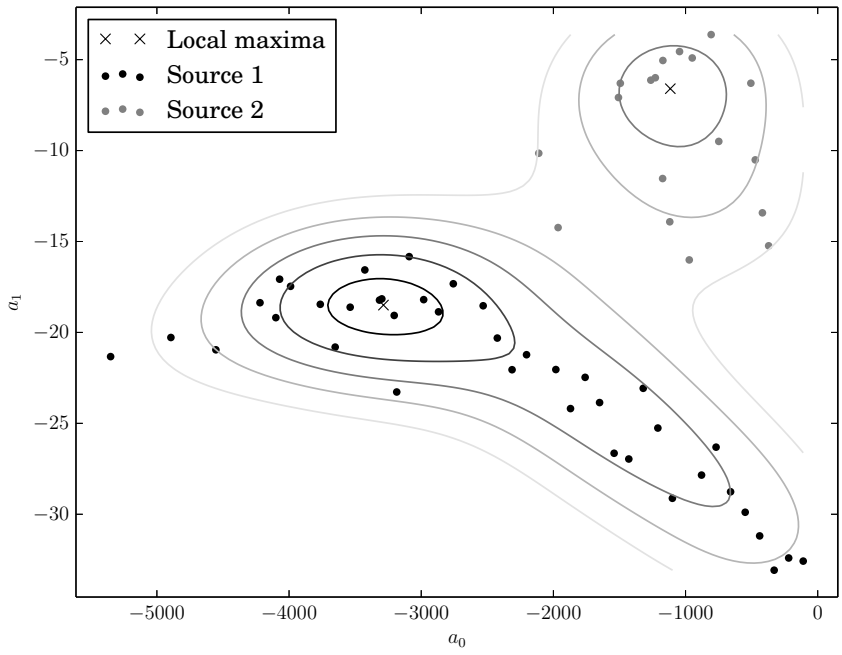

Fig. 7.8: **Estimated memberships.** The contours of the function resulting from convolving the data-points with kernels are represented by the lines. It is from the local maxima of these functions that the EM algorithm begins its search for the mixture of Gaussians (not shown) that give the classifications. A marker's classification is indicated by its colour.

$$\Delta_{\boldsymbol{a}_1} = \max(\boldsymbol{a}_1) - \min(\boldsymbol{a}_1)$$

and $\theta_{\boldsymbol{\beta}_{\boldsymbol{a}}}$ is a parameter to control the smoothness of the resulting function, here $\theta_{\boldsymbol{\beta}_{\boldsymbol{a}}} = 1.2$. A contour plot of the resulting function $f(\hat{\boldsymbol{a}})$ is shown in Figure 7.8.

To initialize GMM the initial means $\boldsymbol{\mu}^0$ are chosen to be the points corresponding to the local maxima of the smoothed scatter plot[2]. To determine initial weights $\boldsymbol{w}^0$ we first determine the value of the function at the two local maxima, $f(\boldsymbol{a}_0^*)$ and $f(\boldsymbol{a}_1^*)$. To weight relative to these two values, we compute

$$w_i^0 = \frac{\Theta_w\{f(\boldsymbol{a}_i^*)\}}{\sum_{p=0}^{R-1}\Theta_w\left\{f(\boldsymbol{a}_p^*\right\}}$$

[2]Recall that the superscript here refers to the iteration number of the algorithm.

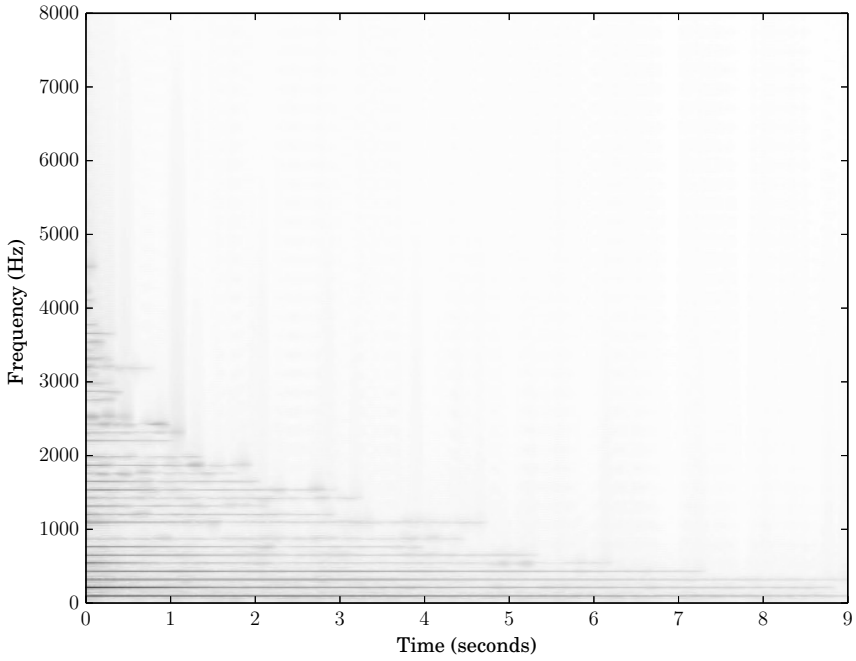

Fig. 7.9: **Spectrogram of source separated acoustic guitar**

Ichiro asked to add the fundamental

where Θ_w is some kind of weighting operator to have parametric control over the influence of each function value and R is the number of maxima. Here

$$\Theta_w \left\{ f(\boldsymbol{a}_i^*) \right\} = \begin{cases} f(\boldsymbol{a}_i^*)\theta_w & i = 0 \\ f(\boldsymbol{a}_i^*) & \text{otherwise} \end{cases}$$

and $R = 2$, i.e., only the first maximum is weighted. For this experiment the parameter set as $\theta_w = 1.1$ gave the best results. The covariance matrix $\boldsymbol{\Sigma}^0$ is computed as

$$\boldsymbol{\Sigma}^0 = \boldsymbol{S}(\{\boldsymbol{a}\}) + \epsilon \boldsymbol{I}$$

where $\boldsymbol{S}$ computes the sample covariance and $\boldsymbol{\epsilon I}$ is a matrix whose only non-zero entries are on the main diagonal and are equal to a small constant to avoid a singular initial covariance matrix. 100 iterations of the EM algorithm (see B) are performed to compute classifications. Each point is assigned to its most likely cluster using the final estimated

See appendix B

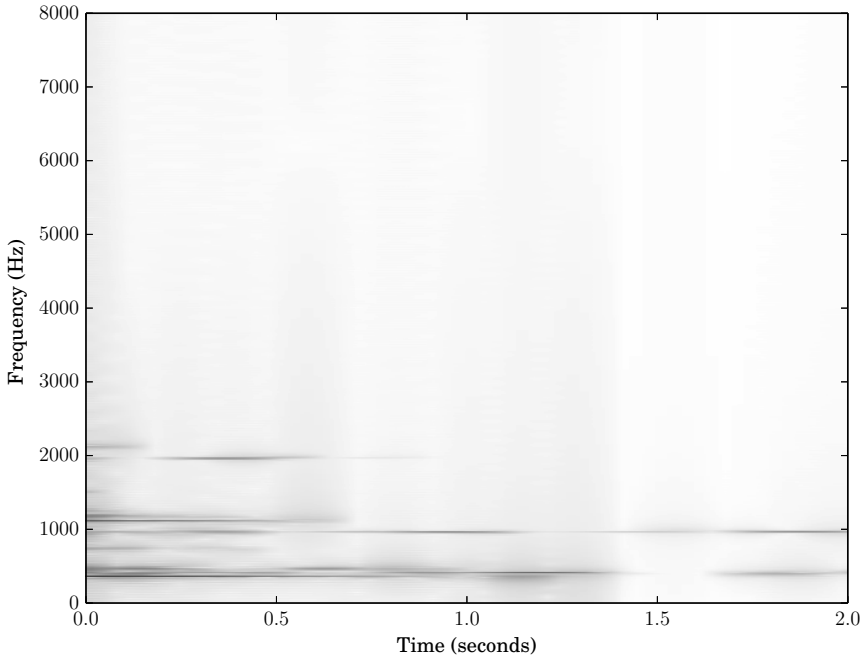

Fig. 7.10: **Spectrogram of source separated xylophone**

Ichiro asked to add the fundamental

Gaussian distributions. The final classifications for this classification task can be seen in Figure 7.8.

7.5 Synthesis

After the classifications have been made, synthesizing the separated sources simply involves only synthesizing the partials classified as belonging to the same source. For the synthesis, we use the technique described in Section 5.4.2. Spectrograms of the source separated signals are shown in Figure 7.9 and Figure 7.10.

7.6 Conclusion

After an informal listening, the source separation is perceptually convincing.[3] At least one partial from the guitar can be heard in the xylophone recording, however — it is difficult

[3]Soundfiles can be downloaded from
`https://drive.google.com/file/d/0B8B4c04j8tBwZDFraEZ1dFZHRFU/view?usp=sharing`

to separate partials that do not have sufficient spatial separation in Figure 7.8. Another drawback of the current technique is that it requires some tuning of the parameters θ_w and $\theta_{\boldsymbol{\beta_a}}$. From the spectrograms of the resynthesized sources, we see that some of the partials from both sounds were lost in the analysis. Although a shortcoming of the analysis rather than the classification, if partials are not sufficiently separated in time or frequency, they cannot be separated as their analysis will yield simply one partial when there are in fact many. In any case, it is important to see that source separation can be carried out by only considering the amplitude modulation (in this case, the decay rate) in relation to the partial frequency. Apart from the combination of instruments presented here many plausible situations can be imagined where this technique could be carried out: e.g., with a mixture of sustained instruments such as violin, voice or horns and pitched percussive instruments such as the piano or guitar.