Experiment 1: Partial grouping in one frame

Nicholas Esterer

May 2016

Contents

0.1 Introduction

To evaluate whether the grouping of partials with common 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).

0.2 Methodology

We synthesize theoretical sets of parameters as described above. 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 evalutate the results. A summary follows:

- Parameters are synthesized from a theoretical mixture of AM and FM sinusoids. Spurious data are added to these parameters.
- Principle components analysis is carried out on the parameters happening at one time instance.
- A histogram is made of the 1st 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 Gaussian mixture models are made by dividing the histogram into equal parts by area and choosing the centres of these parts.
- The EM algorithm for Gaussian mixture models is carried out to classify the sources.

0.3 Evaluation

The algorithm is run on a typical source separation problem to evaluate its plausibility.

0.3.1 Synthesis

Our model makes available the following parameters. Time values are in seconds, frequency values are in Hz and phase values are in radians.

- t time.
- f_s sampling frequency.
- N length of signal in samples.
- H duration between data point calculations in samples (i.e., the hop size).
- N_p 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.
- $t_{attack,p}$ 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.

To incorporate inharmonicity often observed in real string instruments where the strings exhibit some stiffness, we define the *stretched* harmonic numbers as follows

$$K_B(k) = k(1 + Bk^2)^{\frac{1}{2}} \tag{1}$$

Each source is synthesized using the following equation:

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

where

$$A_p(k,t) = \begin{cases} \exp(a_{60,p}t + a_{k,60,p}k)\cos^2(\frac{\pi}{2}(\frac{t}{t_{attack,p}} - 1)) & \text{if } t \leq t_{attack,p}, \\ \exp(a_{60,p}t + a_{k,60,p}k) & \text{if } t > t_{attack,p}, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

$$a_{60,p} = \frac{\log_{10}(10^{-3})}{k_{60,p}} \tag{4}$$

$$a_{k,60,p} = \frac{\log_{10}(10^{-3})}{t_{60,p}} \tag{5}$$

The piecewise amplitude function is based on the amplitude function of the FOF.

The estimation of these parameters is a separate problem addressed by the Derivative or Reassignment Methods (DM and RM). 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 the model of these methods.

The DM and RM seek signals $s_k \in \mathbb{C}$ of the following form:

$$s_k(n) = \exp(\log(A_k + \mu_k n + j(\phi_k + \omega_k n + \frac{1}{2}\psi_k n^2))) \tag{6}$$

Here n is the sample number. 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. The coefficents of the kth harmonic of the pth source from our synthetic model are given by

$$\mu_{k,p}(t) = \frac{a_{60,p}}{f_s} \tag{7}$$

$$A_{k,p}(t) = \frac{1}{2}\exp(a_{60,p}t + a_{k,60,p}k) \tag{8}$$

for the part of the signal after the attack portion.

For the attack portion, we estimate the parameters using least-squares on a (rectangular) windowed signal. Let

$$\hat{\mathbf{s}}(t) = \begin{pmatrix} \frac{1}{2}\exp(a_{60,p}(t - \frac{N_w}{2f_s}) + a_{k,60,p}k)\cos^2(\frac{\pi}{2}(\frac{t-\frac{N_w}{2f_s})}{t_{attack,p}} - 1)) \\ \vdots \\ \frac{1}{2}\exp(a_{60,p}(t + \frac{N_w}{2f_s}) + a_{k,60,p}k)\cos^2(\frac{\pi}{2}(\frac{t+\frac{N_w}{2f_s})}{t_{attack,p}} - 1)) \end{pmatrix} \tag{9}$$

then $\log(A_{k,p})$ and $\mu_{k,p}$ are found as the least-squares solution of

$$\begin{bmatrix} 1 & \frac{-N_w}{2} \\ \vdots & \vdots \\ 1 & \frac{N_w}{2} \end{bmatrix} \begin{pmatrix} \log(A_{k,p}(t)) \\ \mu_{k,p}(t) \end{pmatrix} = \log \hat{\mathbf{s}}(t) \tag{10}$$

for the argument parameters (those multiplied by j in (6))

$$\omega_{k,p}(t) = \frac{2\pi}{f_s}(f_{0,p} + A_{f,p}\sin(2\pi f_{f,p}t + \phi_{0,f,p}))K_{B_p}(k) \tag{11}$$

$$\psi_{k,p}(t) = (\frac{2\pi}{f_s})^2 A_{f,p} f_{f,p}(f_{0,p} + A_{f,p}\cos(2\pi f_{f,p}t + \phi_{0,f,p}))K_{B_p}(k) \tag{12}$$

$$\phi_k(t) = (2\pi f_{0,p}t - \frac{A_{f,p}}{f_{f,p}}\cos(2\pi f_{f,p}t + \phi_{0,f,p}))K_{B_p}(k) + \phi_{0,p} \tag{13}$$

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:

- $\tilde{\omega}_{k,p}(t) \sim \mathcal{N}(\omega_{k,p}(t), \omega_{no})$
- $\tilde{\psi}_{k,p}(t) \sim \mathcal{N}(\psi_{k,p}(t), \psi_{no})$
- $\tilde{\phi}_{k,p}(t) \sim \mathcal{N}(\phi_{k,p}(t), \phi_{no})$
- $\tilde{\mu}_{k,p}(t) \sim \mathcal{N}(\mu_{k,p}(t), \mu_{no})$
- $\tilde{A}_{k,p}(t) \sim \mathcal{N}(A_{k,p}(t), A_{no})$

The θ_{no} (where θ is replaced by ω etc.) specifies the variance of the particular parameter.

We also add spurious data points as a fraction r_s of the number of true data points. Their values are drawn from uniform distributions with boundaries $\theta_{s,min}$ and $\theta_{s,max}$, where θ is some parameter above, e.g., $\omega_{s,min}$ and $\omega_{s,max}$ for the ω parameter.

Data points are computed for the times $t = 0, \frac{H}{f_s}, \frac{2H}{f_s}, \ldots, \frac{\lfloor \frac{N}{H} \rfloor H}{f_s}$.

0.3.2 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 $\omega_l(t)$ (see thesis by Creager). The amplitude modulation $\mu_l(t)$ remains constant for all harmonics of the same source, only its initial value changes according to $k_{60,p}$. We compile the data points at one time into a set of observations

$$\mathbf{x}_l(t) = \begin{pmatrix} \frac{\psi_l(t)}{\omega_l(t)} \\ \mu_l \end{pmatrix} \tag{14}$$

$$\mathbf{X}(t) = \left[\mathbf{x}_1(t) \dots \mathbf{x}_L(t)\right] \tag{15}$$

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

Following the standard technique for producing principal components (see Jolliffe), 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

$$\mathbf{A}(t) = \mathbf{V}^T(t)\mathbf{X}(t) \tag{16}$$

We have found it sufficient to use only the first principal component and therefore only use the values in the first row of $\mathbf{A}(t)$

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, hence if we are observing a random variable drawn from multiple distributions with sufficiently separated means and small enough variances, this separation will be most observable in the principle components corresponding to the greatest eigenvalues.

0.3.3 Preparing data for clustering

The Expectation Maximization underlying the Gaussian mixture model parameter estimation will only converge to a local maximum (see Dempster et al), 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 τ_h other values are discarded. We find N_p 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 indeed follow a normal distribution). The initial guesses for the weights are simply $\frac{1}{N_p}$.

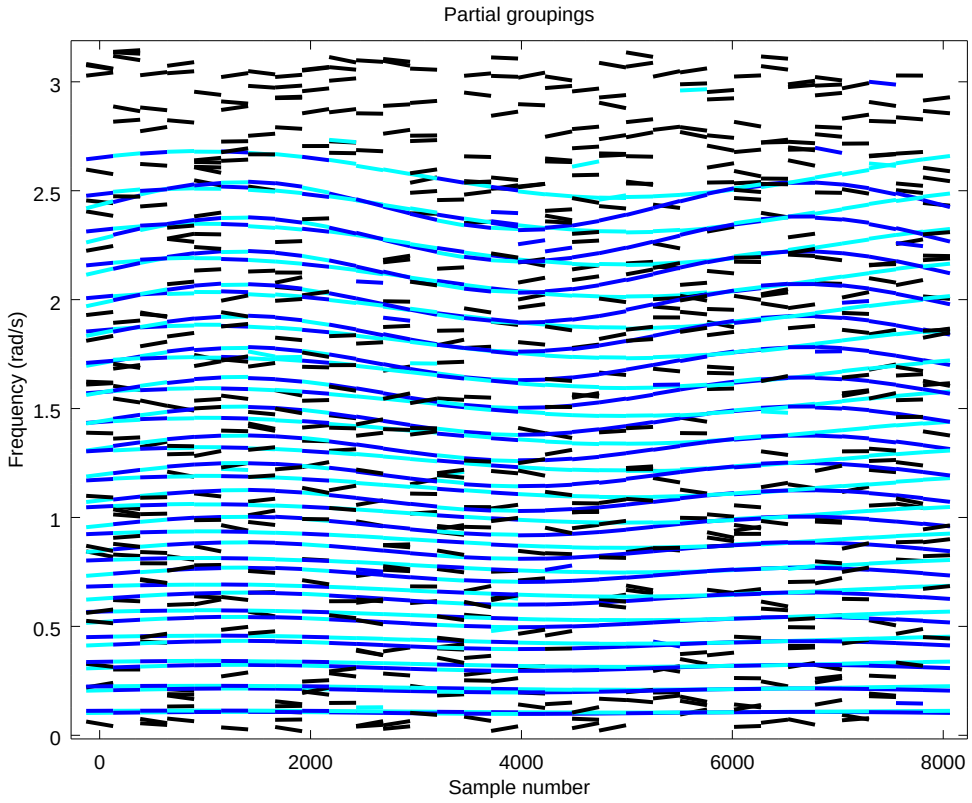
0.3.4 Clustering

Gaussian mixture model parameter estimation will be discussed in an appendix of this document. 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 occured. 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 following parameters: The length of the signal N is 8000 samples and the

Parameter	Source 1 value	Source 2 value
$f_{0,p}$	261.63	277.18
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
$t_{attack,p}$	0.1	0.1
$A_{f,p}$	11.486	11.486
$f_{f,p}$	3	2

anlaysis hop size H is 256 samples.

The data points are represented as line segments to show the frequency slope. Data points classified as coming from the same source are all dark blue or cyan. Data points classified as spurious are in black. Note that the classification is done on a frame by frame basis so it could be that data points adjacent in time that seem to come from the same source are in different colours. This is to be addressed in the sequel. (TODO: how do we choose levels of noise to test against? My current method is to multiply the value by a random number drawn from a normal distribution + 1. The amount of change is nicely


Partial groupings
Frequency (rad/s)
Sample number
0
0.5
1
1.5
2
2.5
3
0
2000
4000
6000
8000

interpretable as a percentage deviating from the original value, but it doesn't work when the value being multiplied is 0)