Voice Processing

Marchioro Thomas

Project 2 - Sinusoidal model

1 Analysis-synthesis of speech based on the Sinusoidal model

1.1 Peak picking

I employed the same peak picking algorithm that I've designed in Project 0, which is based on the intuition that a peak must be higher than both the previous and the next sample. I ordered the peaks by magnitude and I kept the L largest peaks, then I ordered them again by increasing frequency. I also tried to compare my algorithm with the `findpeaks` Matlab function and they appear to find mostly the same peaks.

1.2 Frequency matching

Following the idea of birth and death processes from the original paper, I noticed that there are three main cases for frequency matching. Denoting with $f_1^{(i)}$ the i-th frequency of the previous frame and with $f_2^{(j)}$ the j-th frequency of the next frame, the three main cases are the following:

1. **Case** $f_2^{(j)} < f_1^{(i)} - \Delta$: in this case the current frequency on the next frame is below the current frequency of the previous frame. Since the comparison is done in increasing order, it means that there is no match for $f_2^{(j)}$, hence a birth must occur.
2. **Case** $f_2^{(j)} > f_1^{(i)} + \Delta$: in this case, instead, the current frequency on the next frame is above the current frequency of the previous frame. Hence the death of $f_1^{(i)}$ must occur.
3. **Case** $|f_2^{(j)} - f_1^{(i)}| < \Delta$: in this case the matching condition is satisfied. Nonetheless, there might be a better match, so the algorithm must enter into a loop that stops only once there is no improvement in the matching, i.e., $|f_2^{(j)} - f_1^{(i)}| < |f_2^{(j+1)} - f_1^{(i)}|$. All the sub-optimal matches become births.

After the frequency are matched, a matching table must be built and it must contain the values of magnitude, frequency and phase in the previous and next frames $(A_1, f_1, \phi_1, A_2, f_2, \phi_2)$. For the actual matches, the tables is just filled with the respective values, but for deaths and birds the counterpart in the other frame must be handcrafted. In the case of births, the amplitude in the previous frame must be set to $A_1^{(i)} = 0$, and – since the phase of a sinusoid $\sin(2\pi f_2^{(j)} nT/Fs + \phi_2^{(j)})$ is linear in $f_2^{(j)}$ – the phase in the previous frame is given by $\phi_1^{(i)} = \phi_2^{(j)} - 2\pi f_2^{(j)} S/Fs$, where S is the number of samples in a single frame. Conversely, for deaths $A_2^{(j)} = 0$ and $\phi_2^{(j)} = \phi_1^{(i)} + 2\pi f_2^{(j)} S/Fs$. The frequency is matched to itself in both cases.

1.3 Frame interpolation

In a given frame, the reconstructed sample at time nT is computed as

$$y[nT] = \sum_{k=1}^{L} A_k[nT] \cos(\omega_k[nT]) \tag{1}$$

where the instantaneous amplitude $A_k[nT]$ is obtained by linear interpolation and the instantaneous phase $\omega_k[nT]$ is obtained by cubic interpolation, following the procedure described in the original paper. The idea is to have a “smooth” transition from A_1 to A_2 and from ϕ_1 to ϕ_2.

2 Results

I employed the sinusoidal model to reconstruct the given audio track `arctic_bdl1_snd_norm.wav`, and also on a track of my recorded voice `marchiorot_dont_steal.wav`. In both cases the reconstruction is quite good and the spectrum obtained from the sinusoidal model resembles the original track's spectrum, as can be seen from Figures 1 and 2. I also tried to quantify how close the reconstruction is to the original by computing the average signal to noise ratio on a single frame. I repeated the computation for different values of L and plotted the results in Figure 3. Nonetheless, it appears that the curve saturates at 8dB, which is very strange since the perceptive quality is good. I hence decided to artificially apply additive Gaussian noise to visualize how 8dB of SNR look (and sound) like. Figure 4 shows that adding a distortion of 8dB leads alters visibly the Fourier transform of the signal. Hence I concluded that, due to some mistakes I have made while writing the code, the reconstructed frames ended up being slightly out of sync with respect to the original ones. Still, the final output is good from a perceptive point of view.

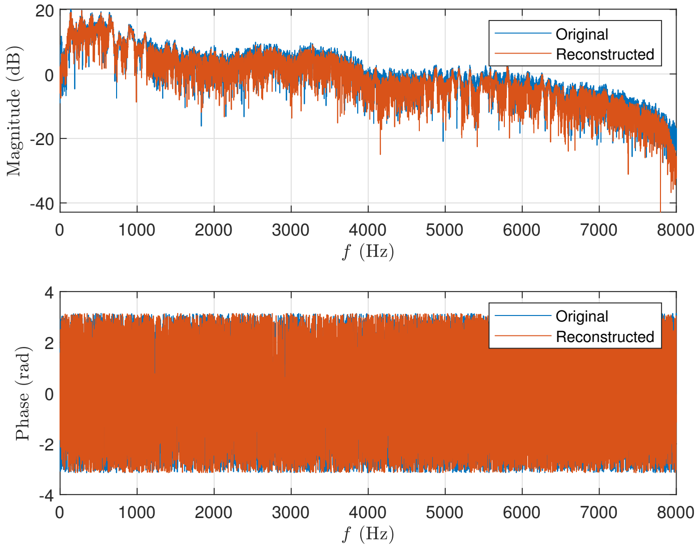

Figure 1: Reconstruction of the given audio track `arctic_bdl1_snd_norm.wav`.

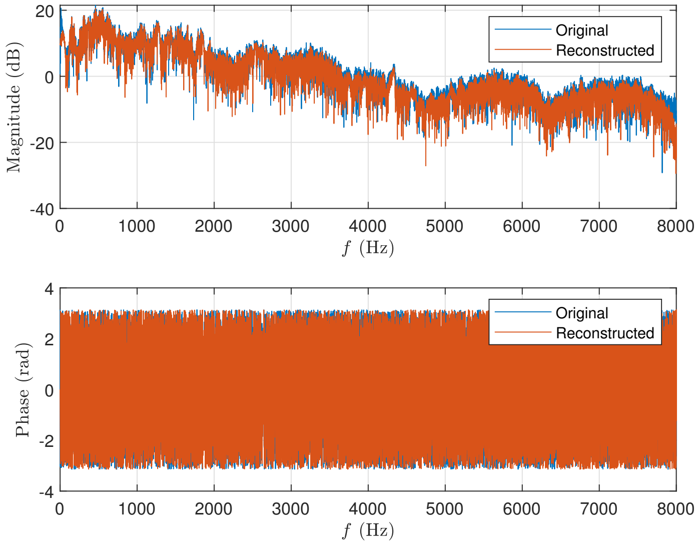

Figure 2: Reconstruction of the given audio track `marchiorot_dont_steal.wav`.

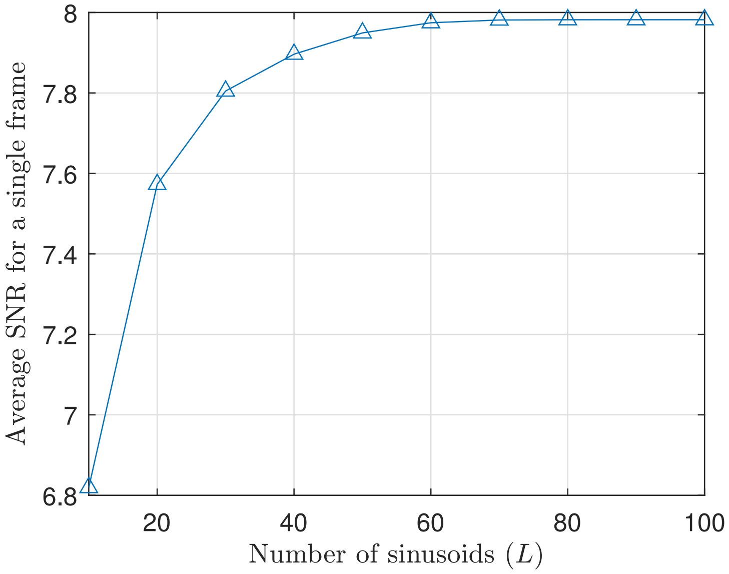

Figure 3: Computation of the average SNR per frame with respect to the number L of sinusoids employed.

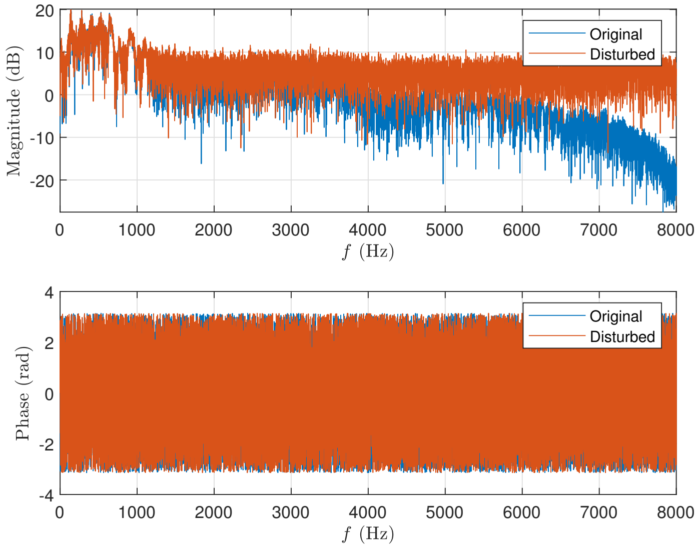

Figure 4: Distortion of 8dB obtained by applying additive Gaussian noise per each frame.