A7: Sinusoidal plus Residual Model

Audio Signal Processing for Music Applications

Introduction

This is a peer assessed assignment where you will analyze and synthesize sounds using the Harmonic plus Stochastic (HPS) model. There are two questions in this assignment. In the first one you will analyze a speech sound that we give and in the second one you will analyze a sound of your choice, in both cases using the HPS model. For each question, you will first describe some of the sound characteristics by listening to the sounds and visualizing their spectrogram, characteristics that should be of relevance for the analysis/synthesis with the HPS model. Then from the described characteristics you will set the appropriate values for the different analysis parameters, explaining the choices you make, and analyze and synthesize the sounds with the HPS model software of the sms-tools package. You should write the answers on the text box provided for each question. On each text box you can write text and hyperlinks. You can upload wav audio files using the "Upload file" button. When asked to upload several files, please compress them into one zip file which you upload.

Guidelines

For this assignment, you can use the models GUI, `models_GUI.py`, or directly the HPS model function, `hpsModel_function.py` (both available in the directory `software/models_interface`). Feel free to modify the code if needed. You should save the synthesized output sounds and upload them here when asked. If you use the models GUI, the output sounds get stored in the folder `sms-tools/software/models_interface/output_sounds` .

To help you with the assignment, we first give a brief description of the analysis parameters used by the HPS model function:

- **Window type (window) and Window size (M):** The choice of window size and window type has a time-frequency trade-off. Choosing a longer window helps resolve sinusoidal components that are close in frequency, but gives a poorer temporal resolution. Shorter windows track transients better, maintaining sharp onsets, but may not resolve frequency components so well. For monophonic harmonic sounds, the window size is best chosen based on the lowest value of f0 and the fastest change in pitch.
- **FFT size (N):** The FFT size is chosen as a power of 2 larger than the window size M. A large FFT size N, compared with M, results on an interpolated DFT spectrum and leads to better estimation of spectral peak values. However, given that the software also uses parabolic interpolation we can achieve good peak estimates with not too big FFT sizes, for example just the next power of 2 larger than M.
- **Threshold in negative dB (t):** The peak picking threshold is the lowest amplitude peak that will be identified. Setting a very low threshold (< -120dB) will take most peaks, but the threshold should be set as high as possible to minimize the presence of peaks that do not correspond to sinusoidal peaks (the window main-lobe).
- **Maximum number of harmonics (nH):** The maximum number of harmonics that can be detected in a harmonic sound is influenced by the brightness of the sound, but also by the sampling rate and by how low is the f0. The recording quality can also have an impact. For a compact representation, we should only capture the relevant harmonics, the ones that affect the perceptual quality of the reconstruction.

- **Minimum f0 frequency in Hz (minf0) and Maximum f0 frequency in Hz (maxf0):** The `minf0` and `maxf0` are the parameters used by the fundamental frequency detection algorithm to obtain possible `f0` candidates to be passed to the TWM algorithm. Choosing a correct range of `f0`, but the smallest possible, greatly improves the `f0` estimation by TWM algorithm, specially minimizing octave errors, which are very common in `f0` detection algorithms. You should select the values by first looking at the spectrogram of the sound and identifying the lowest and highest fundamental frequencies present.

- **Error threshold in the f0 detection (f0et):** This is the maximum error allowed in the TWM algorithm. If the value is too large, the algorithm might detect fundamental frequencies that might not be actually so. Instead, if is it too small, good fundamental frequencies might not be detected, returning value 0 at that frame. The smaller the value the more restrictive the algorithm behaves. A normal strategy is to start with a big value (> 10) and then making it smaller until we only keep what we consider to be the relevant f0 components, discarding the `f0` values in the parts of the sound that do not have a clear harmonic structure.

- **Slope of harmonic deviation (harmDevSlope):** Slope of the harmonic deviation allowed in the estimated harmonic frequencies, compared to a perfect harmonic frequencies. If the value is 0 it means that we allow the same deviation for all harmonics, which is hard coded to `f0`/3. A value bigger than 0 means that higher harmonics will be allowed to deviate more than the lower harmonics from perfect harmonicity (which is a common behaviour). It normally works better to have a value slightly bigger than 0, for example around 0.01.

- **Minimum length of harmonics (minSineDur):** Any harmonic track shorter, in seconds, than minSineDur will be removed. This is a good parameter for discarding harmonic tracks that are too short and thus that do not correspond to stable harmonics of the sound. Typically we put a value bigger that 0.02 seconds.

- **Decimation factor of magnitude spectrum for stochastic analysis (stocf):** The stochastic approximation of the residual is a decimated version of the magnitude spectrum of the residual. This leads to a compact and smooth function that approximates the magnitude spectrum of the residual at each frame. The smaller the stocf, higher the decimation will be and thus will result in a more compact representation. A value of 1 means no decimation, leaving the residual magnitude spectrum as it is. A value of 0.2 (a good starting value) will decimate the original residual magnitude spectrum by a factor of 1/5.

The most compact and useful representation of a sound, least number of analysis data values while maintaining the sound quality in the synthesis, will be obtained by using a high t, a small `nH`, a small decimation factor for stochastic analysis, stocf, and by succeeding in detecting only the harmonics they are perceptually relevant. The values of `nH`, `minf0` and `maxf0` should be chosen by first visualizing the spectrogram of the sound with a large enough window size. There is usually a range of all parameter values for which we get a good reconstruction. Also the analysis parameters are not independent of each other and hence they need to be considered together. For testing if the detection of the harmonics have been done correctly (and for improving it) it is very useful to perform the Harmonic plus Residual analysis/synthesis (HPR model) using the same parameters and listen to the residual component for possible artifacts resulting from the harmonic analysis.

Question 1. Obtain a good harmonic+stochastic analysis of a speech sound

Analyze and synthesize the sound speech-female.wav, available from the sounds directory in sms-tools, using the Harmonic Plus Stochastic model. The goal is to obtain the best possible reconstruction using the most compact representation. Perform the following two tasks:

- **Part 1.1:** Analyze the sound with the STFT using the models GUI, or with any other analysis tool you might wish, and describe the characteristics of the sound that might be relevant to perform the HPS analysis. Specially important characteristics for the analysis include pitch range and maximum number of harmonics. Write no more than a paragraph for this description.

- **Part 1.2:** Select the analysis parameters that give a good reconstruction and at the same time result in the most compact representation possible, specially related to the number of harmonics and the number of stochastic coefficients. We recommend that you first perform the harmonic plus residual analysis and by listening to the residual make some decisions on the best parameters to use. You can listen to the output sounds (harmonic, residual, stochastic components) and fine tune the parameters. Save the output sounds. Explain the choices for the following parameters: `window type`, `window size`, `FFT size`, `minimum f0`, `maximum f0`, `error threshold in f0 detection`, `number of harmonics`, and `stochastic decimation factor`. In your descriptions do not use more than one sentence per parameter.

After the parameter explanation, upload the three synthesized output sounds. You should compress them into one zip file which you upload. Also, name the sounds so as to be consistent among all students: 1. `speech-harmonic.wav`: harmonic part of the analysed speech sound; 2. `speech-stochastic.wav`: stochastic part of the analysed speech sound; 3. `speech-reconstructed.wav`: resynthesized output sound using the HPS model.

An Example

An example of an analysis/synthesis for a male speech sound can be found here:

- Input sound: `http://freesound.org/people/xserra/sounds/317744/`
- Harmonic component: `http://freesound.org/people/xserra/sounds/327139/`
- Residual component: `http://freesound.org/people/xserra/sounds/327141/`
- Stochastic component: `http://freesound.org/people/xserra/sounds/327137/`
- Harmonic+stochastic resynthesis: `http://freesound.org/people/xserra/sounds/327140/`

Question 2. Obtain a good harmonic+stochastic analysis of a monophonic musical phrase

Analyze and synthesize a harmonic sound of your choice from Freesound using the harmonic plus stochastic model. The goal is to obtain the best possible reconstruction using the most compact representation possible. Return an explanation of what you have done and why, together with the re-synthesized sounds.

The sound from freesound to use could be in any format, but to use the sms-tools software you will have to first convert it to be a monophonic file (one channel), sampling rate of 44100, and 16bits samples. You might also have to select a fragment of the sound.

- **Part 2.1:** Choose a sound from freesound to be analyzed. It should be a short monophonic musical fragment of a harmonic sound, not longer than 5 seconds. Put the freesound link of the sound selected and write a brief explanation of why you chose this sound. You can even use a specific sound of your own for this question. Just upload it to freesound and provide a link.
- **Part 2.2:** Analyze the chosen sound with the STFT, or with any other analysis or tool you might wish, and describe the characteristics of the sound that will be relevant to perform the harmonic plus stochastic analysis. Important characteristics for the analysis include the pitch range and the maximum number of harmonics. Write no more than a paragraph.

- **Part 2.3:** Select the analysis parameters that give a good reconstruction and at the same time result in to the most compact representation, specially related to the number of harmonic and the number of stochastic coefficients. We recommend that you first perform the harmonic plus residual analysis and by listening to the residual make some decisions on the best parameters to use. You can listen to the output sounds (harmonic, residual, stochastic components) and fine tune the parameters. Save the output sounds. Explain the choices for the following parameters: `window type`, `window size`, `FFT size`, `minimum f0`, `maximum f0`, `error threshold in f0 detection`, `number of harmonics`, and `stochastic decimation factor`. In your descriptions do not use more than one sentence per parameter.

After the parameter explanation, upload the three synthesized output sounds. You should compress them into one zip file which you upload. Also, name the sounds so as to be consistent among all students: 1. `a7p2-harmonic.wav`: harmonic part of the analysed sound; 2. `a7p2-stochastic.wav`: stochastic part of the analysed sound; 3. `a7p2-reconstructed.wav`: resynthesized output sound using the HPS model.