Multiple Frequency Harmonics Analysis and Synthesis of Audio Signals

A. Jbira, A. Kondoz
Centre for Communication Systems Research
University of Surrey
Guildford, Surrey GU2 5XH, UK

ABSTRACT

In this paper we propose a multiple frequency harmonics model for analysis and synthesis of audio signals. The novelty of this model is that a composite sound can be represented by a few number (two or three) of frequency harmonics, each with time-varying fundamental frequency. We present the model and some key issues in tuning it. Results demonstrate that the model is valid and promising. It is convenient for time- and frequency- scale modifications and is of interest for low bit rate audio coding.

1. INTRODUCTION

Speech and audio signals modelling has received a significant attention in a number of areas such as coding, synthesis and audio editing. Traditional models that represent audio signals are characterised by their attempt to minimise a mean square or weighted mean square distance between the input and the output signals. These models offer a natural sound, in which the reconstructed spectrum matches well the input spectrum at least in the perceptual sense. For example, in frequency models, the spectrum is split into individual components using a Short Term Fourier Transform (STFT), and each of the component is reconstructed accurately according to its perceptual content. One particularity of these models is their response to a pure sinusoid as they attempt to maintain its phase and amplitude. However, subjective listening show that, whatever phase is selected, the sinusoid will sound in the same way. On the applications side, these models are of limited use for time- and frequency- scale modifications because of artefacts caused by phase uncertainty and discontinuities when trying to modify the spectrum. Moreover, these models do not allow low bit rate audio coding because several parameters need to be encoded in order to synthesise the sequence, e.g. phases, and amplitudes. In contrast, models based on pitch tracking analysis methods, such as those used in vocoders, have been used in the past to solve some of these problems, but their use is restricted to the class of strongly harmonic signals.

In this paper, we show how a complex spectrum can be closely matched by multiple frequency harmonics, and how a synthetic spectrum can be constructed using a set of fundamental frequencies and a spectral amplitude vector.

2. OVERVIEW OF THE MODEL

Frequency harmonics in a speech or audio sequence can be localised in time and frequency domain using different techniques such as time domain waveform similarity method based on the normalised auto-correlation function. Figure 1(a) shows an example of the first 200 correlation coefficients calculated on a sequence of 40 ms of voiced speech. The fundamental frequency of the sequence corresponds to the position of the maximum value of the correlation function (here $f_0 = 8000 / 36 \approx 222 Hz$). One method of generating the synthetic spectrum is to use the Fourier transform of a suitable window (e.g. sinusoidal window) centred on each harmonic frequency and normalised to the corresponding harmonic magnitude. Figure 1(b) shows the result of this operation. Unfortunately, most of audio signals can not be modelled accurately using this approach. Figure 2 shows a case where one frequency harmonic is not sufficient to represent all the peaks in the spectrum.

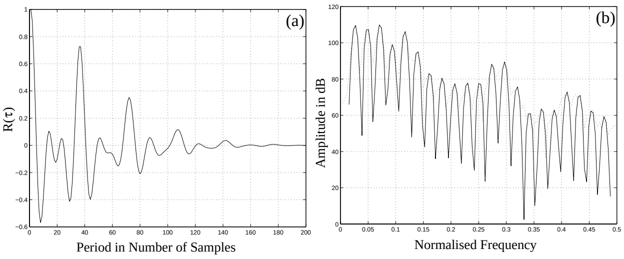

Figure 1. Normalised auto-correlation function (a) and synthetic magnitude spectrum (b). The original magnitude is represented in doted points.

The idea suggested in this paper is to use a second set of frequency harmonics. This is achieved by selecting the position of the second maximum of the correlation function, which does not correspond to a multiple of the first period. This position corresponds in Figure 2(a) to $\tau = 65$. This value, as we will explain in the following section, is an over-estimation of the second period. The correct value is in fact 32. Now that the second period has been found, the secondary spectrum can be synthesised using the same process as before. The result can be viewed in Figure 3(a). The resulting overall spectrum (spectrum-1 + spectrum-2) is shown in Figure 3(b). As we can see that, now the input spectrum is better matched.

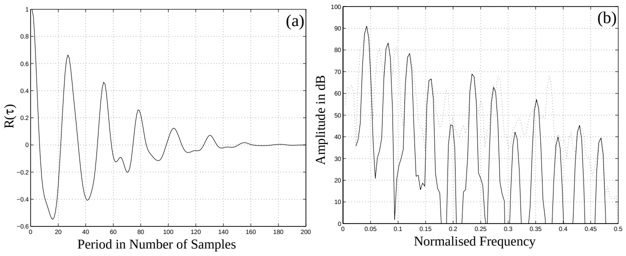

Figure 2. Normalised auto-correlation function (a) and synthetic magnitude spectrum using one fundamental frequency (b) for a sequence of music.

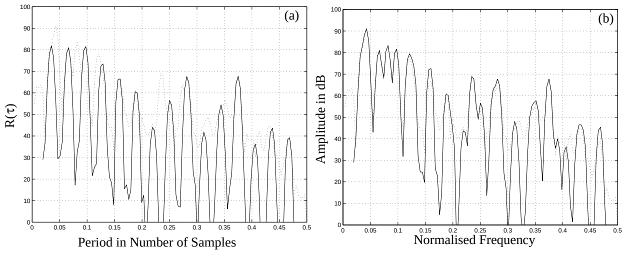

Figure 3. Synthetic magnitude spectrum using the second fundamental frequency (a) and overall magnitude using both the fundamental frequencies (b).

The choice of a model is important to achieve good performance, but tuning as usual, is also important because significant differences in quality may be expected from two systems that are based on the same model. The following sections expose some issues that should be taken into account in tuning the model, and in the mean time explain one possible implementation of the model.

3. AUDIO ANALYSIS

The parameters of the model are analysed on a frame by frame basis to determine the time-varying parameters. The length of the frame (N) can be set to a fixed value (e.g. 160, 20ms), but can be shorter in non-stationary regions of the input signal and larger in stationary regions.

Several methods have been developed to detect accurately the period in a speech signal. These methods may be classified as time domain waveform similarity methods or frequency domain spectral similarity methods. For instance, let consider the method based on the normalised auto-correlation function $R(\tau)$. This method can be extended to estimate two (or more) periods by selecting the secondary maximum of $R(\tau)$, that does not correspond to a multiple of the first (main) period. In cases where the first period is dominant, the detection of the second period may be wrong, an example is shown in Figure 2(a). The secondary period may even be completely masked by the first period, the example is given in Figure 4(a), where $R(\tau)$ for an other frame of the same music signal is shown. As we can note, the maximisation process would provide only one period around 26. In other cases, the position of the primary period may not quite correspond to the period, due to the presence of a secondary period.

A better idea is to use the multistep "estimate-cancel-estimate" algorithm [1]. This is a process in which the period sounds are cancelled in turn: a first period estimate is used to suppress correlates of the dominant sound, and a second period is then estimated from the remainder. The process is repeated to estimate further periods, or else to recursively refine the initial estimates. We propose to use the following algorithm to estimate the periods. The first period estimate T_0 is selected by considering the first maximum of the function $R(\tau)$. Then its contribution in the composite sound is suppressed by applying a comb filter in time domain. A comb filter of lag parameter $\tau = T_0$ defined by the impulse response: $h(t) = \delta(t) - \delta(t - T_0)$ would normally reduce this contribution, because this filter has zeros at the frequency $f_0 = 1/T_0$ and all its multiples, and its response to a periodic signal of period T_0 is zero everywhere. However, the first period estimate is usually not accurate. A better idea is to use a comb filter that suppresses the contribution of the periods T_0, $T_0 - 1$, and $T_0 + 1$. The impulse response of this filter is:

$$\left[\delta(t) - \delta(t - T_0 - 1)\right] * \left[\delta(t) - \delta(t - T_0 + 1)\right] * \left[\delta(t) - \delta(t - T_0)\right]$$
$$=$$
$$\left[\delta(t) - \delta(t - T_0 + 1) - \delta(t - T_0) - \delta(t - T_0 - 1)\right] +$$
$$\left[\delta(t - 2T_0 + 1) + \delta(t - 2T_0) + \delta(t - 2T_0 - 1) - \delta(t - 3T_0\right]$$

This filter has been tested for real signals, but found not reliable, because it uses a large window of the signal. Better results have been found when using the filter of impulse response:

$$h(t) = \delta(t) - \delta(t - T_0 + 1) - \delta(t - T_0) - \delta(t - T_0 - 1)$$

This linear filter is applied to the composite harmonics to suppress the first set, and then the secondary function $R(\tau)$ is calculated from the remainder. Then, the maximum and its position of this function are found. If the maximum is smaller than a threshold (here set to 0.25), then the second period is declared absent and the process of period estimation finishes at this point. Otherwise the second period T_1 is declared present and its value is the position of the maximum of $R(\tau)$. The next step is to operate the comb filter that corresponds to this second period to refine the first period estimate, and later refine the second period estimate. The same process may be applied to detect more than two periods.

In order to determine the magnitudes of the harmonics, the input signal is Hamming windowed and Fourier transformed to get a spectrum. The length of the window (N_a) should be large enough to allow a sufficient frequency resolution to detect the harmonic magnitudes. It is important to ensure that the Fourier transform of the chosen window is a peak (around frequency zero) that dies down to a very small value at $\pm[f_0]_{\min}$. In practice, the length should be around twice the maximum period. This value gives an attenuation of -45 dB at $\pm[f_0]_{\min}$, -7dB at $\pm 0.5\times[f_0]_{\min}$, and -1.6dB at $\pm 0.25\times[f_0]_{\min}$. As we can note, a wrong harmonic position would attenuate considerably its magnitude. In order to estimate how accurate the estimated period needs to be, let consider the worst case of a perturbed harmonic $f = k/(T_0+\varepsilon)$. The perturbation $\xi = |f - k/T_0|$ is maximum when $k/T_0 \approx 0.5$, and $T_0 = T_{\min}$. To ensure that the worst attenuation (in absolute value) is less than 1.6dB, we need to have $\xi < 0.25/T_{\max}$, hence $\varepsilon < T_{\min}/(2T_{\max})$. Hence if the method used to estimate the period is not accurate enough (this is in particular the case of the normalised auto-correlation method), it should be refined in frequency domain. This combination of non-accurate detection of the period and refinement reduces the computational complexity incurred in frequency domain search [2].

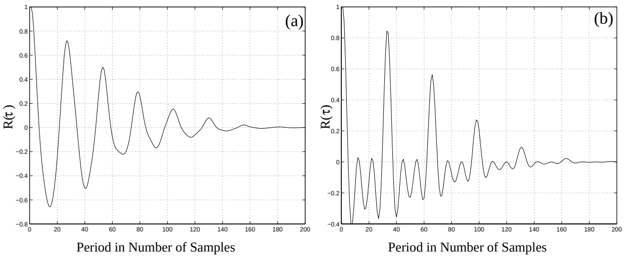

Figure 4. Multistep estimation of periods in a music signal before (a) and after (b) cancelling the first period.

Once the period has been accurately estimated, then the Fourier transform of the Hamming window is centred to each of the harmonics. To ensure that the peak is correctly adjusted to the harmonic, the Fourier transform of the Hamming window is calculated on a large number of points (up to 64 times the size N_a of the analysis window). The magnitudes are the weighting coefficients of the peaks that maximise the similarity between the synthetic spectrum and the input spectrum in the bands $\pm f_0$ around the harmonics. The bands where the resulting weighted peak has a smaller energy in respect to the spectrum are called non-harmonic bands. In these bands, the spectrum is represented by randomly distributed peaks, normalised to the energy of the input spectrum.

A further issue that should be discussed is the size of the analysis window. In order to estimate accurately the periods and the magnitudes it is important to use large-enough analysis windows. The periods range typically between 16 and 160 ($f_0 = 50$ to $500 Hz$). Hence for long periods the analysis window should be of around $2\times 260 = 320$ long. Naturally, we may decide to set the length of the analysis window to this value, but our experiments show a degradation for speech signals at the onsets and offsets. The reason is that such large windows spread on the previous and next frames and creates artificial peaks in the current frame. A solution is to use two analysis windows : a shorter will be used for small periods and a longer for large periods. This method alleviates the suggested artefacts.

A methodology to follow for tuning the analysis process is to test it on synthetic signals (sinusoids + noise), then on a mixture of instruments and complex sounds. The objective is to minimise the cases of failure and increase the output quality.

4. AUDIO SYNTHESIS

It is important to take some precautions when considering the amplitudes of very close harmonics that are represented in the short term spectrum by around the same index, otherwise one single peak would be represented twice in the synthetic spectrum. We can notice this case in Figure 3(b), where the amplitude of some harmonics are 6dB greater than their original values. As there is a problem of uncertainty in estimating the original amplitudes of very close peaks, a strategy is to force the lowest amplitude to zero. Every two harmonics are declared very close if their difference is smaller than the minimum allowed fundamental frequency ($\Delta f < 1/P_{\max}$). The remaining frequency positions are then sorted from the smallest value to the biggest.

An oscillator which is defined by its amplitude and phase is assigned to each harmonic. Windowing technique is used to reduce the effect of spreading of the harmonics one on the other. The synthetic spectrum is generated according to the following equation

$$\hat{s}(n) = w_s(n)\times\sum_f A_f(n)\cos\left[\theta_f(n)\right] \text{ , } 0\le n\le N_s-1$$

where $\{w_s(n)\}$ is the synthesis window of size N_s (e.g. sinusoidal window), $\{A_f(n)\}$ are the estimated magnitude spectrum at the harmonics f , $\{\theta_f(n)\}$ the corresponding phase. Overlap-add is used to compensate the effect of windowing.

Non harmonic bands, i.e. bands around the harmonics that does not correspond to peaks in the input spectrum, are synthesised using randomly distributed peaks around the frequency harmonics normalised to the harmonic magnitudes. The phases of these peaks are random uniformly distributed between 0 and 2π . Harmonic bands are synthesised by sinusoids centred on each harmonic frequency and normalised to the corresponding harmonic magnitude. The phases of these sinusoids can be presented as input to the synthesis module, but alternatively, and this is one advantage of the model, they can be synthesised linearly using the previous frame’s frequency harmonics and phases and the current frame’s frequency harmonics to ensure a smooth continuity between successive adjacent oscillators. However, due to the fact that the sound is not periodic over several analysis frames, the variations in the estimated parameters at adjacent frames can cause discontinuities at the edges of frames, resulting in significant degradation of quality. To solve this problem, during synthesis, both the current and previous frames’ parameters are checked to make sure a smooth transition at the frame boundaries take place. For each frequency harmonic f_k of the current frame, we find all the frequency harmonics of the previous frame that belong to the interval

$$J_k = \left[f_k - (f_{k-1} + f_k)/2,\ \ldots,\ f_k + (f_{k-1} + f_k)/2 \right]$$

Only the frequency harmonics of J_k that are very close to f_k, e.g. $|f_k - f| < 0.1 \times f_k$, need to be interpolated to the frequency harmonic f_k. The amplitudes are linearly interpolated and the phases $\{\theta_f(n)\}$ are quadratic interpolated. As there may be more than one harmonic of J_k to be interpolated, if each of the harmonics are interpolated, the energy of the harmonic f_k would be greater than what it should be. Hence a precaution should be taken. One strategy to solve this problem is to interpolate each frequency harmonic f of J_k with an attenuated version of the frequency harmonic f_k. The attenuation is the contribution of the energy of the harmonic f in the overall energy of the harmonics to be interpolated.

5. SIMULATIONS AND RESULTS

The proposed analysis and synthesis model has been implemented with the following parameters: $N = 160$, $N_a = 201$ for small periods, and $N_a = 321$ for large periods, $N_s = 221$. Subjective listening have been carried out to test the quality of the proposed model. Different sequences have been used including: speech signals (English male and female), musical instruments (piano, organ and violin), and a music orchestra. First, the number of periods has been set to one. The output quality for speech signals suffers from a degradation. This is due to the period estimation. In fact, the reliability of period estimation algorithm not only depends on its estimation function but also on its tracking method. Hence, a forward tracking period has been used to keep the period estimate change smoothly between succeeding frames. The function that is optimised (e.g. normalised auto-correlation function) takes into consideration the values of the period in the next one or two frames. After this modification, the speech quality was very clean and comparable to the original one. And as expected the single period works well for music instruments, except for some regions in violin, but fails for the music orchestra. But when the number of periods has been set to two, the quality has been considerably improved for all the sequences. Figure 2 explains how a multiple frequency harmonics model represents a short term spectrum calculated on a window of 512 samples of a music sequence. Figure 2(a) represents the input spectrum (doted) and the synthesised spectrum using one single period. As we can see, some peaks of the input spectrum have not been covered by the first harmonic spectrum. Figure 2(b) represents the result when using two periods. This time, as we can see that, more peaks have been covered.

6. CONCLUSION

The tests show that our representation is valid and promising. The proposed representation of audio signals has the same advantages of any type of harmonic model, e.g. time- and frequency- scale modification. Moreover, it will allow low bit rate audio coding because it uses few parameters to synthesise the signal. Finally, the design will benefit from several advances in vocoders, e.g. how to encode efficiently the time-varying parameters, and improvements are expected to be brought to the model.

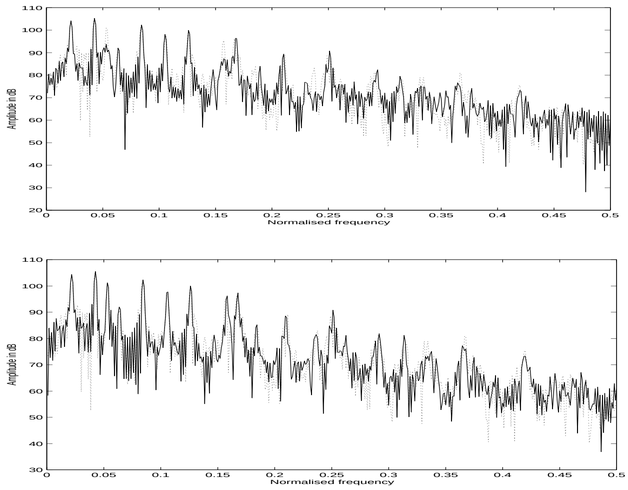

Figure 5. Above is a synthetic spectrum using one period and below is a synthetic spectrum using two periods.

7. REFERENCES

[1] A. de Cheveigné and H. Kawahara. "Multiple period estimation and pitch perception model," Speech Communication 27, pp.175—185, 1999.

[2] DVSI, “Inmarsat-M Voice Codec, Version 2,” Inmarsat-M Specification, Inmarsat, Feb. 1991.