A METHOD FOR SPECTRUM SEPARATION AND ENVELOPE ESTIMATION OF THE RESIDUAL IN SPECTRUM MODELING OF MUSICAL SOUND

Nicola Laurenti and Giovanni De Poli

Dipartimento di Elettronica ed Informatica, Università di Padova
via Gradenigo 6/A, 35131 Padova, Italy — E-mail: {nil,depoli}@dei.unipd.it

ABSTRACT

We propose an original technique for separating the spectrum of the noisy residual component from that of the harmonic, quasi-deterministic one, and to estimate the envelope of the residual, for the spectrum modeling of musical sounds. The algorithm for spectrum separation relies on nonlinear transformations of the amplitude spectrum of the sampled signal (obtained via FFT), which allow to eliminate the dominant partials without the need for precisely tuned notch filters. The envelope estimation is performed by calculating the energy of the signal in the frequency domain, over a sliding time window. Eventually the residual can be obtained by combining its spectrum and envelope, so that separate processing can be performed on the two.

1. INTRODUCTION

In the digital synthesis of musical sound from time-frequency representation, the harmonic, quasi deterministic component, and the noisy, stochastic one are treated separately, due to their quite different features. In particular, in spectrum modeling techniques[1], the deterministic component is usually modeled as a sum of stable (with slow amplitude and frequency variations) sinusoids, whereas the noisy component is modeled through time-varying filtering of stationary white noise. Other different approaches for modeling the noisy component were proposed in [2, 3].

Therefore, in the analysis of digitally sampled sound waveforms, the two components must be carefully separated in order to extract the two sets of parameters that are needed for their synthesis and processing, respectively. On the other hand, being the energy of the sinusoidal component larger than that of the noisy one by several orders of magnitude, the separation process is not an easy task.

Methods for separation can follow two distinct approaches. Time-domain methods rely on a precise estimate of the sinusoidal component, then subtract it from samples of the original sound. The main drawbacks of such a direct method are the sensitivity to finite arithmetics effects and estimate errors depending on window and hop sizes. This gives rise to an undesired and unstable “harmonic residual”, the energy of which can be larger than that of the non harmonic one. Such effects can result perceptually annoying and prevent use of the results for further processing.

More complicated, frequency-domain methods perform a filtering of the original sound, with a filter that exhibits deep notches corresponding to the frequencies of its partials. They provide a more realistic noise residual, especially since it preserves the amplitude envelope of the original noise, but if the notches are not very selective, the resulting spectrum turns out more “anti-harmonic” rather than stochastic. The complexity of the problem calls for nonlinear and/or stochastic analysis methods.

2. OUTLINE OF THE PROPOSED METHOD

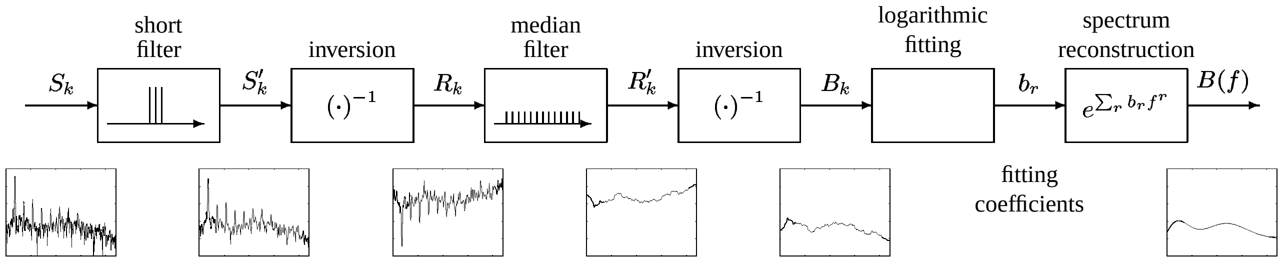
We propose an original nonlinear technique to estimate the spectrum of the noisy residual. Given the original sampled sound s_n, we divide it into short time windows (and calculate the amplitude spectrum S_k of the DFT for each windowed portion of the signal. Due to the highly energetic sinusoidal components, we can not obtain a smooth and reliable estimate of the noise spectrum through linear operations. Thus, after eliminating the (rare and randomly distributed) zeros in the DFT by averaging the amplitude of each DFT bin with the adjacent ones, we calculate its inverse $R_k = 1/S'_k$ so that the spectral lines of S_k become zeros of R_k. The values of R_k are then averaged within a sliding frequency window of N_f bins, and by inverting the result we obtain an estimate of the stochastic component spectrum, which is eventually, furtherly smoothed through a logarithmic fitting operation to yield the desired amplitude spectrum.

After performing spectral separation of the noisy component we face the problem of estimating its amplitude envelope. In order to do that, we make use of the stochastic spectrum samples obtained as above. In a dual fashion with respect to the spectrum smoothing operation previously described, we first perform a local estimate of the time-varying noise power by averaging its spectral energy within a sliding window. Then we take its square root as an estimate e_n of the noise envelope and smooth it by logarithmic fitting.

The advantage of our method is to provide a spectrum model in which the partials are removed without suppressing them below the level of noise. Moreover, the model can accurately track slow time evolution of the partial frequencies. The results do not exhibit artifacts and can be useful for further processing such as time stretching, pitch shifting and separate processing of sinusoidal and noisy components.

As an example, the so obtained estimates of the spectrum and envelope of the noisy component for a piano sound, have been combined for its synthesis. Beside a possible common time stretching, the sinusoidal part is pitch shifted according to the desired interval, while the noisy part (representing hitting of the keys, and percussive noise) is reproduced unshifted.

By generating random, independent and uniform phases φ_k for each frequency bin of the signal DFT and pairing them with the corresponding amplitudes A_k we obtain, via FFT, a stationary colored noise, which is then modulated by the envelope estimate sequence v_n The synthesized sound results plausible and doesn’t have any harmonic residual present.



$s(t)$, modeled as the sum of partials and a wideband stochastic residual $g(t)$ with a much lower energy, we are faced with the task of finding a smooth function $B(f)$, possibly updatable as we move our analysis window along the sound sample, that approximates the time-varying spectrum of the stochastic residual, by representing, in a sense, an average spectrum of its realizations.

It is evident that, by performing a median filtering in frequency of the signal spectrum, we would only obtain a spreading of the narrowband partials, that are by orders of magnitude ampler than the underlying stochastic spectrum. Also, if we removed the partials with a comb filter before the median filtering, the effect of the latter would be to spread the rather wide comb notches, unless the comb filter is very precisely tuned and capable of tracking the partial frequencies. As we said above, this would give rise to a "complementary harmonic" spectrum.

On the other hand, if we consider the inverse amplitude spectrum $R(f) = 1/|S(f)|$, then in place of the highly energetic partials in $S(f)$ we will find deep and selective notches in $R(f)$, which can be eliminated through the median filter, whereas the reciprocal of stochastic spectrum will play a prominent role in the averaging performed by the filter. Once obtained the filtered reciprocal spectrum, we must in turn take its reciprocal, to have the required function $B(f)$ approximating the stochastic residual spectrum.

We note that when taking the reciprocal $R(f)$ we might end up with some very high accidental peaks, that correspond to zeros of $S(f)$ due to the particular realization we have chosen. Such peaks would corrupt the result of filtering $R(f)$, in much the same way that the partials would corrupt the result of filtering $|S(f)|$. However, since the zeros are much more isolated and randomly distributed than the partials, they can be canceled, without altering much the spectrum shape, by passing $|S(f)|$ through a very short median filter with an arbitrarily small length Δf, prior to taking its reciprocal.

When performed numerically on the portion of $s(t)$ that belongs to an analysis window of N_t samples $\{s_n\}$, $n = 0, N_t - 1$ the above technique requires the following steps, shown in Figure 1:

1. calculate the amplitude spectrum of $\{s_n\}$ by taking the absolute value of its N_t-points DFT with a suitable window function w_n (e.g. Hann, Hamming or Blackman)

$$S_k = \left|\sum_{n=0}^{N-1} s_n\, w_n\, e^{-j2\pi nk}\right| \quad , \quad k = 0, \dots, N_t - 1 \tag{1}$$

2. remove accidental zeros in $\{S_k\}$, by replacing it with

$$S'_k = (S_{k-1} + S_k + S_{k+1})/3 \tag{2}$$

3. calculate the reciprocal spectrum

$$R_k = \frac{1}{S'_k} \tag{3}$$

4. smooth R_k by cyclic convolution with the N_f-points median filter impulse response

$$\overline{R}_k = \frac{1}{N_f} \sum_{h=-\lfloor N_f/2 \rfloor}^{\lceil N_f/2 \rceil - 1} R_{(k-h) \bmod N_t} \tag{4}$$

5. calculate the reciprocal of $\overline{R}_k$ that gives the samples required approximation for the residual spectrum

$$B_k = \frac{1}{\overline{R}_k} \tag{5}$$

Observe that, as described, the algorithm has two adjustable parameters of analysis: the length N_t of the analysis window in the time domain, and the length N_f of the median filter in the frequency domain, and both must be set according to the time and frequency variability of the stochastic residual spectrum. Typically they should be set to rather low values in order to be able to track fast variations in time and frequency shaping.

We have thus obtained the function $B(f)$ as N_t frequency samples. However, for further processing of the residuale spectrum, it is desirable to have a closed form expression of $B(f)$, depending on few parameters, so that it is possible to calculate it for different frequencies. To this aim we can perform a *logarithmic fitting* of the samples, so that $B(f)$ when expressed in dB is a polynomial of order p,

$$B(f) = \exp \sum_{r=0}^{p} b_r f^r \tag{6}$$

and the coefficients $b_0, \dots, b_p$ are chosen to minimize the error measure

$$\sum_{k=0}^{N_t - 1} \left[\log B(f_k) - \log B_k\right]^2 \quad , \quad f_k = kF_c/N_t \tag{7}$$

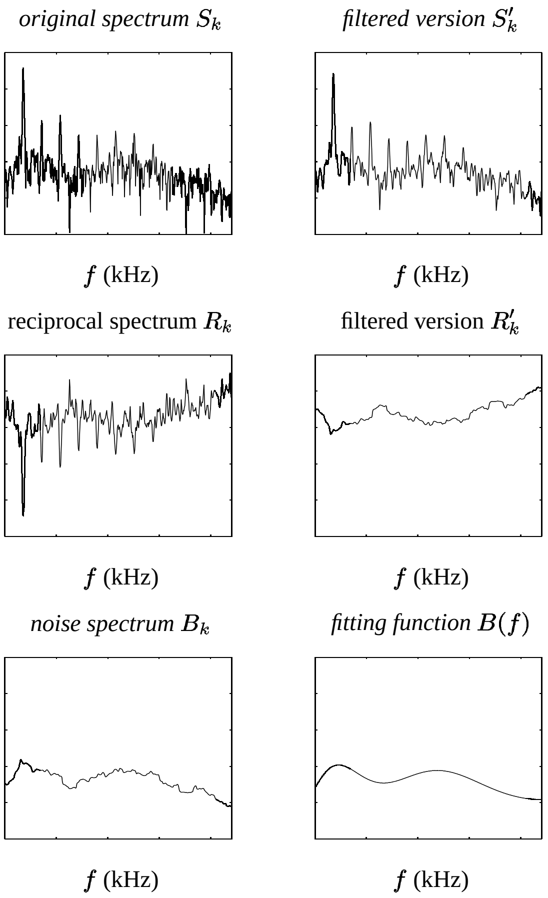


Figure 2: *Spectra resulting from the consecutive steps in the separation procedure*

In this way, we have added a third degree of freedom, that is the number of spectral parameters $p+1$. The spectral shapes obtained from each step of the procedure are plotted in Figure 2, for a flute sound with pitch at 1780 Hz (A6).

4. ENVELOPE ESTIMATION

Consider the spectrum separation procedure performed on the portion of the signal samples within the analysis window $[n_0, (n_0 + N_t - 1)]$. From the noise spectral samples $\{B_k\}$ obtained in step 5 we can derive a measure of the energy of the noise process g_n within the window. In fact, since windowing and DFT are linear operations, we can expect B_k to be a good approximation to the amplitude spectrum of $g_n w_n$. Therefore we must have

$$E_B = \sum B_k^2 \simeq \sum (g_n w_n)^2 \simeq E_g E_w \qquad (8)$$

Indeed, if we expect the noise g_n to be nearly stationary within the analysis window, the average noise envelope

$$\bar{r}_{n_0} = \sqrt{\frac{E_g^{[n_0,n_0+N_t]}}{N}}_t \qquad (9)$$

can be assumed as the value of the envelope at the window midpoint $t = n_0 + N_t/2$.

By progressively shifting the analysis window along the signal sequence by small amounts, we can obtain a rather dense grid of envelope values, which have to be interpolated to yield the required envelope $r(t)$. Again, the interpolation is performed by piecewise polynomial fitting of the logarithm of the amplitude. Piecewise linear interpolation is sufficient for most cases, providing that the length of single pieces matches the signal dynamics. Better results are obtained with higher order polynomials or splines. Observe that, if a single polynomial is used (e.g. for short sounds), it is appropriate to fit an even order polynomial with a negative leading coefficient, since the envelope must vanish at both ends of the signal sequence.

The steps of this procedure are implemented as follows

1. for the ℓ-th time window, starting at $n_0 = \ell\Delta$ with length N_t samples, perform spectral separation of the noise component and evaluate its energy as

$$E_g^{[n_0,n_0+N_t]} = \sum B_k^2 / E_w \qquad (10)$$

2. calculate the envelope at the midpoint instant

$$r(\ell\Delta + N_t/2) = \sqrt{\frac{E_g^{[n_0,n_0+N_t]}}{N}}_t \qquad (11)$$

3. fit a piecewise polynomial curve $q(t)$ to the logarithm of $r(\ell\Delta + N_t/2)$,

4. obtain the required envelope as

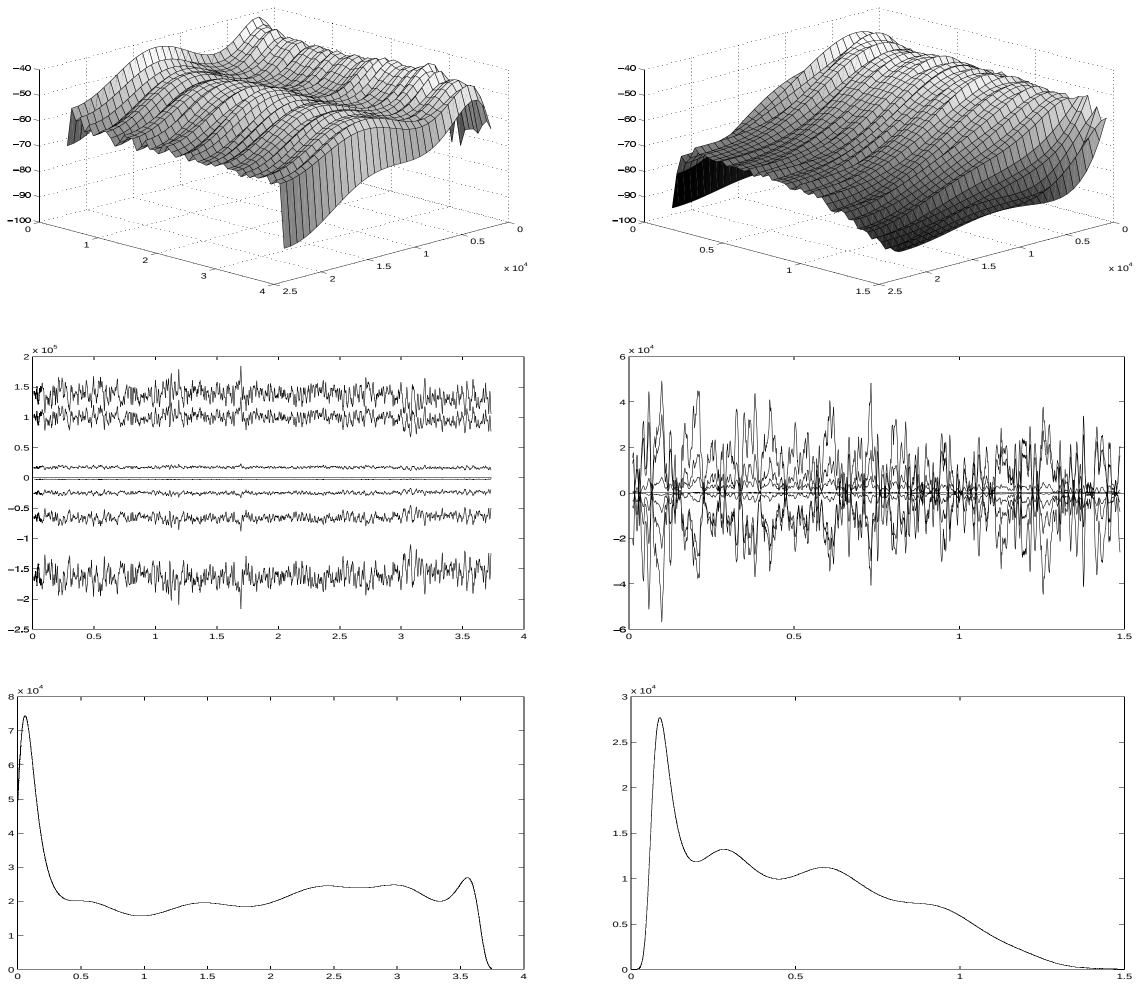
$$r(t) = \exp q(t) \qquad (12)$$

5. ANALYSIS EXAMPLES

We have applied the above procedure to some sound sample sequences, given by X. Serra in [5]. The results are shown in Figure 3, with reference to a flute sound with pitch 1780 Hz (A6) and an oboe sound with pitch 522 Hz (C5), both sampled at 44.1 kHz. The parameter values in both cases were:

- window size: $N_t = 1024 \simeq 23ms$;
- window shift: $\Delta = 32 \simeq 0.73$ ms;
- window type: Hann $w_n = \cos^2(\pi n/N)t)$;
- median filter length: $N_f = 25 \simeq 1077$ Hz;
- spectrum fitting order: $p = 8$;
- number of divisions for piecewise envelope interpolation: $N_e = 1$
- envelope fitting order: $q = 12$:

The three plots indicate the time-varying amplitude spectrum of the noisy component, the trajectories of the polynomial coefficients in time and the envelope. It can be observed that the noise spectrum is nearly stationary throughout the sound duration, but, while in the case of the flute sound so are the coefficients b_r, this is not the case with the oboe sound, showing that, probably b_r are not very significant parameters.


−40
−50
−60
−70
−80
−90
−100
0
1
2
3
4
2.5
2
1.5
1
0.5
0
x 10⁴
0.5
1
1.5
x 10⁵
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
3.5
x 10⁴
6
4
2
0
−2
−4
−6
x 10⁴
8
7
6
5
4
3
2
1
0
x 10⁴
3
2.5
2
1.5
1
0.5
0