RESIDUAL MODELING IN MUSIC ANALYSIS-SYNTHESIS

Michael Goodwin
Center for New Music and Audio Technologies (CNMAT) &
Department of Electrical Engineering and Computer Science
University of California at Berkeley
e-mail: michaelg@eecs.berkeley.edu

ABSTRACT

In analysis-synthesis of musical sounds based on a sinusoidal model, the difference between the original signal and the synthesized signal, termed the *residual*, is typically a broadband noise process. It contains such musical phenomena as flute breath noise or violin bow noise. Synthesis without such "noise" tends to sound artificial; it is desirable to improve the synthesis realism by modeling the residual in such a way that it can be reinjected in the synthesized signal. This paper deals with a model of noise perception based on the *equivalent rectangular bands* (ERBs) of the auditory system. Since a broadband noise is perceptually well-represented by the time-varying energy in each of these frequency bands, the residual is parametrized in terms of these energies in the proposed model. An application of the model to music synthesis based on the inverse fast Fourier transform (FFT) is described in detail.

1. SINUSOIDAL MODELING AND ADDITIVE SYNTHESIS

In analysis-synthesis, the difference between the original signal and the synthesized signal is called the *residual*. The properties of the residual are determined by the underlying signal model and the analysis-synthesis method. A common model for musical signals is a deterministic plus stochastic decomposition where the deterministic component $d(t)$ is a sum of slowly-evolving sinusoids called *partials* and the stochastic component $s(t)$ is a broadband noise process [1]:

$$d(t) = \sum_{q=1}^{Q} A_q(t) \cos \Theta_q(t) \qquad (1)$$

Here, Q is the number of partials; $A_q(t)$ describes the time-varying amplitude of the q-th partial and the total phase $\Theta_q(t)$ describes its frequency evolution and phase offset. This model is useful for musical applications since the parameters allow for desirable modifications such as pitch-shifting, time-scaling, and a wide variety of spectral transformations such as cross-synthesis.

For this signal model, the analysis typically consists of a short-time Fourier transform (STFT) followed by a spectral peak-picking algorithm that tracks the peaks from one analysis frame to the next; such an analysis typically finds the amplitude, frequency, and phase of each partial in each frame [1, 2]. If the partial parameters vary slowly with respect to the analysis frame rate, the frame rate samples can be reliably interpolated in the ensuing synthesis using low-order evolution models such as linear amplitude. Any error in the synthesis interpolation appears in the residual, as do any partials missed by the analysis; such partials can be removed by re-analyzing the residual. More notably, the residual also contains any part of the input signal that the analysis is not designed to extract, namely the stochastic component of the signal model, which is a broadband noise since narrowband stochastic components correspond to sinusoids with slowly varying amplitude and phase that are accordingly extracted by the analysis as partials. This broadband component accounts for such musical phenomena as breath noise in a flute or saxophone; electronic music synthesis without such "noise" tends to sound artificial. To improve the realism of the synthesized signal, the residual can be parametrized and reinjected before synthesis.

In a time-domain additive synthesizer, the interpolated parameter tracks are used as control inputs to a bank of oscillators whose respective outputs are accumulated to form the deterministic component; this is a direct realization of the sum-of-partials model. The stochastic component is typically generated by filtering white noise. The two components are then added to yield the final synthesis output. Equivalently, additive synthesis can be done in the frequency domain by using the parameter tracks from the analysis to derive a short-time spectrum for each synthesis frame and then performing an inverse discrete Fourier transform (IDFT) followed by overlap-add (OLA) to create the time-domain output [3]. The short-time spectrum of the deterministic component is built by accumulating a weighted spectral lobe for each partial, where the lobe corresponds to the main lobe of the DFT of an appropriately chosen time-domain window. Since it is computationally advantageous to use the IDFT to generate the deterministic and stochastic components simultaneously rather than to generate the components independently and then combine them in the time domain, it is desirable to find a frequency-domain parametrization of the residual that can be incorporated into the short-time spectrum before the IDFT.

2. NOISE PERCEPTION MODEL

Auditory models commonly include a set of overlapping bandpass filters whose bandwidths increase roughly in proportion to their center frequencies. The classical critical

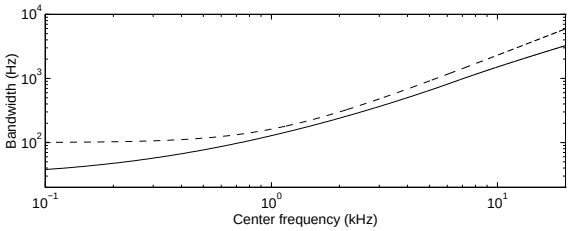

Figure 1. *Bandwidth vs. center frequency for critical bands (dashed) and ERBs (solid).*

bandwidths, derived in experiments on noise masking and perception of complex sounds, are often considered as related to the bandwidths of these auditory filters at given center frequencies [4]. Early estimates of the critical bandwidth as a function of center frequency indicate a roughly constant value below 500 *Hz* and a linear increase for higher frequencies, resulting in the common interpretation of the auditory system as a constant-Q filter bank. More recent experiments by Moore and Glasberg suggest that the low-frequency critical bandwidths are quadratically related to the center frequency [5]. They provide an expression for the *equivalent rectangular bandwidths* of the auditory filters that differs somewhat from classical critical band theory; this is shown in Figure 1.

A simple model of noise perception can be arrived at by dividing the spectrum into a set of equivalent rectangular bands (ERBs) based on the Moore-Glasberg formulation. In perceiving a broadband noise, the auditory system is primarily sensitive to the total short-time energy in each of these bands, and not to the specific distribution of energy within the bands. An analysis based on this model filters a broadband noise $x[n]$ through an ERB filter bank $\{h_1[n], h_2[n], \ldots, h_B[n]\}$ to derive the ERB signals $\{x_1[n], x_2[n], \ldots, x_B[n]\}$. These signals are then parameterized on a frame-rate basis in terms of their energies; for the i-th frame, the energy of the b-th ERB signal is given by

$$E_i(b) \;=\; \sum_{n=0}^{N-1} x_b[n - iL]^2 \qquad (2)$$

where N is the frame size and L is the analysis stride. Synthesis according to this model is achieved by filtering white noise through the ERB filter bank with a time-varying gain $g_i(b)$ on each channel given by the ERB energy parameter from the analysis:

$$g_i(b) \;=\; \sqrt{\frac{E_i(b)}{N\sigma^2 \sum_m h_b[m]^2}} \qquad (3)$$

where σ^2 is the variance of the excitation white noise. The following section discusses an FFT-based implementation of this model for use in a frequency-domain music synthesizer.

3. APPLICATION OF THE MODEL

3.1. Frequency-Domain Additive Synthesis

Frequency-domain additive synthesis was introduced by Rodet and Depalle as an alternative to time-domain approaches [3]. The first stage in the synthesis is the construction of the short-time spectrum of a given frame. For each partial, a spectral *motif* $B(\omega)$ is added into the spectrum centered at frequency ω_q and weighted by $A_q e^{j\phi_q}$, where the parameters $\{A_q, \omega_q, \phi_q\}$ are derived by the analysis. Since time-domain multiplication is equivalent to frequency-domain convolution, the accumulated spectrum corresponds to a time-windowed sum of sinusoids:

$$\sum_{q=1}^{Q} A_q e^{j\phi_q} B(\omega - \omega_q) \;=\; B(\omega) * \sum_{q=1}^{Q} A_q e^{j\phi_q} \delta(\omega - \omega_q) \quad (4)$$

$$\stackrel{\mathcal{F}^{-1}}{\Longrightarrow} \quad b[n] \sum_{q=1}^{Q} A_q \cos(\omega_q n + \phi_q) \qquad (5)$$

where $*$ denotes convolution, $\mathcal{F}^{-1}$ denotes an inverse discrete Fourier transform (implemented as an IFFT), and where the time window $b[n]$ is the IDFT of the motif $B(\omega)$, which is designed to be localized in frequency so that the spectral accumulation of partials is computationally efficient. The conjugate symmetric components of the spectrum have been omitted from the formulas for simplicity.

After the IDFT, $b[n]$ is divided out, which leaves a frame containing a sum of partials with constant amplitudes A_q and frequencies ω_q. The final output is synthesized using OLA with a triangular window, which provides linear amplitude interpolation across synthesis frames. The partial frequencies are interpolated nonlinearly based on phase-matching constraints at the frame boundaries [3, 6].

3.2. Residual Analysis

The residual analysis is based on three requirements:

- It should derive a small set of perceptually meaningful parameters. The perceptual relevance is desirable since it enables transformations of the parameters to create a specific perceptual effect.
- The residual parameters should be in such a form that they can be combined economically with the partial parameters. In the IFFT synthesizer, they should yield a spectral representation of the residual that can be combined with the sinusoidal spectrum before the IFFT.
- The analysis-synthesis of the residual itself should be *perceptually lossless*, meaning that the input and output should sound the same; there is some leeway here since imperfections in the synthesized residual tend to be masked when it is combined with the partials.

These requirements are met by the following FFT-based implementation of the ERB noise analysis-synthesis model.

The analysis uses a sliding window $w_i[n] = w[n - iL_a]$ of length N_a to extract frames of the residual $s[n]$ at times spaced by the analysis hop size L_a. The frame signal $x_i[n] = w_i[n]s[n]$ is then transformed by an FFT of size M_a, where $M_a > N_a$. Then, the spectrum is divided into bands according to the ERB model; for the sake of data reduction, the number of bands can be decreased somewhat by scaling each ERB width by the same factor. After the band allocation is established, the energy in each of the bands is computed from the FFT magnitudes; the negative

frequency components are not included since the spectrum is conjugate symmetric:

$$E_i(b) \;=\; \frac{1}{M_a} \sum_{k \in \text{band } b} |X_i[k]|^2 \qquad (6)$$

These ERB energies serve as the residual parameters for the i-th frame; changes in the characteristics of the residual are reflected in frame-to-frame variations of the ERB energies. Recall that the psychoacoustic model is that the perceptual qualities of broadband noise are determined by the total energy in each band, and not by the specific distribution of energy within the bands. In addition, the irrelevance of the FFT phase to the ERB energy calculation is justified since the auditory system is primarily sensitive to the magnitude of the short-time spectrum. Note that the sum of the energies of all the bands across the spectrum is precisely the signal energy of Parseval's theorem:

$$\sum_b E_i(b) \;=\; \frac{1}{M_a} \sum_{k=0}^{M_a-1} |X_i[k]|^2 \;=\; \sum_{n=0}^{N_a-1} x_i[n]^2 \qquad (7)$$

3.3. Residual Synthesis

The noise is synthesized as follows. First, the ERB energies are converted into a piecewise constant spectrum wherein the energy of each constant piece is determined by the corresponding ERB analysis parameter. This is illustrated in Figure 2, which shows the magnitude spectrum of an analysis frame and the corresponding piecewise constant spectral estimate for synthesis based on twelve ERBs. Synthesis using piecewise linear spectral estimates, sloped in each ERB to fit the analysis spectrum, gives perceptually the same output as the piecewise constant approach, verifying that the ear is insensitive to the specific spectral distribution within each ERB.

The following equations demonstrate the preservation of ERB energies in the analysis-synthesis pathway; $X_i[k]$ denotes the analysis FFT, $\hat{X}_i[j]$ denotes the spectral estimate derived in the synthesis, Δ_b is the number of bins in the b-th ERB at the synthesis stage, and M_s is the size of the synthesis IFFT. Note that the analysis FFT and synthesis IFFT do not have to be the same size.

$$E_i(b) \;=\; \frac{1}{M_s} \sum_{j \in \text{band } b} |\hat{X}_i[j]|^2 \;=\; \frac{1}{M_a} \sum_{k \in \text{band } b} |X_i[k]|^2 \qquad (8)$$

$$E_i(b) \;=\; \frac{\Delta_b}{M_s} |\hat{X}_i[j]|^2 \;\Longrightarrow\; |\hat{X}_i[j]| \;=\; \sqrt{\frac{M_s}{\Delta_b} E_i(b)} \qquad (9)$$

Energy preservation will be considered further in the section on normalization.

After the magnitude spectrum is constructed, a uniform random phase is applied on a bin-by-bin basis. Frame-to-frame phase correlations can be introduced to control the texture of the synthesized residual; for instance, varying the smoothness of the residual may be musically desirable. After the phase is incorporated, the stochastic spectrum is added to the partial spectrum (in rectangular coordinates) and transformed into a time-domain signal by the IFFT and OLA. This approach has proven perceptually viable for broadband residuals such as saxophone breath noise.

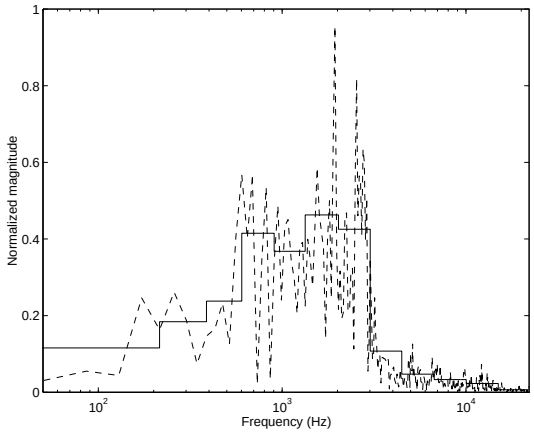

Figure 2. *Piecewise constant ERB estimate (solid) of the residual magnitude spectrum (dotted) for a frame of a breathy saxophone note.*

3.4. Normalization

The ERB noise analysis-synthesis, which is depicted in Figure 3, clearly satisfies the first two requirements proposed in section 3.2. Namely, the ERB energies comprise a small set of perceptually meaningful parameters that can be readily combined with the partials before the IFFT. The final criterion of perceptual losslessness necessitates a consideration of signal scaling; due to the multiple windowing steps and the possibility of different analysis and synthesis frame sizes and sampling rates, the synthesized residual may not have the same loudness as the original residual. In the following treatment, the subscripts a and s refer to analysis and synthesis parameters, respectively.

The proper scaling can be derived by considering the energy in the time-domain signal. For an input segment of length τ_a corresponding to N_a samples at the rate $f_a = \frac{1}{T_a}$, the energy in the continuous-time signal is

$$E'_a \;=\; \int_{\tau_0}^{\tau_0+\tau_a} s(t)^2 dt \;\approx\; \sum_{n=0}^{N_a-1} s(\tau_0 + nT_a)^2 T_a \qquad (10)$$

where the $\approx$ refers to the approximation of the integral by the sum of the areas of rectangles of width T_a. In discrete time, the energy of this analysis frame of length N_a is

$$E_a \;=\; \sum_{n=0}^{N_a-1} s[n]^2 \;=\; \frac{E'_a}{T_a} \qquad (11)$$

This frame energy is now traced through the system.

The output $w[n]s[n]$ of the analysis window has energy

$$E_w \;=\; \sum_{n=0}^{N_a-1} w[n]^2 s[n]^2. \qquad (12)$$

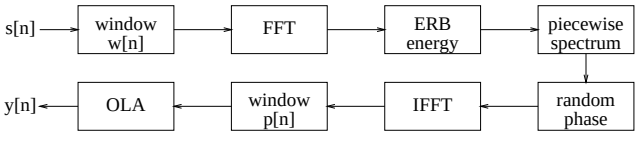

Figure 3. *Residual analysis-synthesis block diagram.*

Replacing $s[n]^2$ by its expected value in equations 11 and 12 gives

$$E_w = E\{s[n]^2\} \sum_{n=0}^{N_a-1} w[n]^2 = \frac{E_a}{N_a} \sum_{n=0}^{N_a-1} w[n]^2 \quad (13)$$

which indicates how the windowing process affects the signal energy. By Parseval's theorem, the M_a-point FFT preserves this energy measure, as does the ERB energy (by construction); the M_s-point IFFT likewise preserves the energy. Using a similar argument as for the analysis window, the effect of OLA with the length-M_s window $p[n]$ can be shown to be

$$E_s = \frac{2E_w}{M_s} \sum_{i=0}^{M_s-1} p[n]\,(p[n] + p_1[n] + p_2[n]) \quad (14)$$

where $p_1[n]$ and $p_2[n]$ correspond to the second half of the window from the previous frame and the first half of the window from the subsequent frame (a 50% overlap factor is assumed in the derivation).

$$p_1[n] = \begin{cases} p\left[n + \frac{M_s}{2}\right] & 0 \le n < \frac{M_s}{2} \\ 0 & \frac{M_s}{2} \le n < M_s \end{cases} \quad (15)$$

$$p_2[n] = \begin{cases} 0 & 0 \le n < \frac{M_s}{2} \\ p\left[n - \frac{M_s}{2}\right] & \frac{M_s}{2} \le n < M_s \end{cases} \quad (16)$$

As a check on the accuracy of this formulation, for a window $p[n]$ that overlap-adds to one, the post-windowing and OLA do not affect the energy. In this system, since the ERB spectrum is added to the partial spectrum before the IDFT and the subsequent division by $b[n]$, the effective OLA window $p[n]$ for the residual is a triangular window divided by the motif window. This hybrid window does not overlap-add to one, so the OLA scale factor must be included.

The energy E_s given by equation 14 is the discrete-time energy for a synthesis frame of length M_s. The energy of the continuous time output signal $y(t)$ is

$$E'_s = \int_{\tau_0}^{\tau_0+\tau_s} y(t)^2 dt \quad (17)$$

$$\approx \sum_{n=0}^{M_s-1} y(\tau_0 + nT_s)^2 T_s = E_s T_s \quad (18)$$

where T_s is the synthesis sampling period and τ_s is the duration of the M_s-sample output frame, $\tau_s = M_s T_s$. However, since the input energy corresponds to an input segment of duration τ_a, what is required is an equalization of the energy for an output segment of that same duration τ_a. Let N_s denote the number of output samples (at rate $\frac{1}{T_s}$) in a segment of duration τ_a; the energy in this segment is

$$E''_s = \sum_{n=0}^{N_s-1} y[n]^2 T_s = \frac{N_s}{M_s} E'_s \quad (19)$$

Noting that

$$\frac{N_s}{M_s} = \frac{\tau_a}{\tau_s} \implies \frac{N_s}{M_s}\frac{T_s}{T_a} = \frac{N_a}{M_s} \quad (20)$$

the entire transformation of the continuous time energies can be expressed as

$$E''_s = G_s G_a E'_a \quad (21)$$

where

$$G_a = \sum_{n=0}^{N_a-1} w[n]^2 \quad (22)$$

is the energy scaling incurred in the analysis, and

$$G_s = \frac{2}{(M_s)^2} \sum_{m=0}^{M_s-1} p[m]\,(p[m] + p_1[m] + p_2[m]) \quad (23)$$

is the effect of synthesis. In the analysis, then, the signal should be multiplied by the scale factor $1/\sqrt{G_a}$ before the ERB energies are calculated; at the synthesis stage, the output should be multiplied by $1/\sqrt{G_s}$ to equalize the energies. Listening tests have verified that the signal energy of Parseval's theorem is an accurate measure of the loudness of broadband noise, and that the outlined approach provides input-output equalization in the ERB analysis-synthesis.

4. CONCLUSION

The ERB residual model has proven useful for improving the realism of frequency-domain additive synthesis. It is primarily effective for representing the breath noise that appears in the sinusoidal analysis-synthesis residuals of saxophones, trumpets, flutes, and similar instruments. For instruments such as the marimba, the sharp transients of the note attacks are problematic for the sinusoidal analysis-synthesis and thus appear in the residual. Because the phase is not carefully treated, the ERB model muddies these residual attacks somewhat. Since both the sinusoidal model and the ERB model have difficulty with such highly time-localized events as mallet strikes, signal models more flexible than sinusoids plus noise should be considered.

REFERENCES

[1] X. Serra and J. Smith. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4), Winter 1990.

[2] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4), August 1986.

[3] X. Rodet and P. Depalle. Spectral envelopes and inverse FFT synthesis. *Proceedings of the 93rd Audio Engineering Society Convention*, October 1992.

[4] E. Zwicker. Subdivision of the audible frequency range into critical bands. *Journal of the Acoustical Society of America*, 33(2), February 1961.

[5] B. Moore and B. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74(3):750–753, September 1983.

[6] M. Goodwin and X. Rodet. Efficent Fourier synthesis of nonstationary sinusoids. In *Proceedings of the International Computer Music Conference*, 1994.