Radiated Voice Pulse Modeling

QueueSevenM - August 3, 2026 | Updated: August 5, 2026

Justification

While [1] showed that Wide-Band Harmonic Sinusoidal Modeling (WBHSM) performed superior to prior methods (evaluated using the additive synthesis reconstruction signal-to-residual ratio method, and listening tests for transformations); they also demonstrated several issues. I will classify these issues into two separate categories. The first category is issues that arise from estimation errors in the priori information used by the algorithm. The second type is problems that are intrinsic to the processing technique itself.

For the first type of issue, [1] demonstrated that WBHSM is highly sensitive to estimation errors in the fundamental frequency. For the second type, [6] pointed out that WBHSM exhibits estimation errors when its input experiences a change in the MFPA [2] harmonic phase envelope. However, I will show that WBHSM also experiences estimation errors when there are changes in frequency as well as for changes in the harmonic amplitude envelope. Not only that, but I will also quantify mathematically exactly what errors these input irregularities produce in the model's estimation.

To understand the effect of input irregularities, we must go back to the duality of voice signal. That is, that it can be thought as either a train of overlapping pulses in the time-domain; or as harmonics which evolve in amplitude, frequency, and phase in the frequency-domain. To relate these two components, we must consider what each corresponds to in the other domain. Thinking of periodic signal, the harmonics in the frequency domain correspond to the spectral components at those harmonic frequencies. That is - the harmonics of period signal actually sample the spectrum of a single period [2, 3].

But what about a train of infinite (or just arbitrarily-sized) overlapped pulses? Well assuming all of the pulses are identical, the rate of repetition is constant, and the pulse train extends infinitely both backwards and forwards in time, then the harmonics of a single period is actually equivalent to those harmonics of a single pulse. If each pulse

$$p(t) = 0 \, \forall t < t_{start}$$

, then the pulse train only needs to extend to a finite amount of time, but still must extend backwards an infinite amount of time.

Lemma 1: The harmonics

$$R(hf_0) \, \forall h \in \mathbb{N}$$

of a single period

$$r(t) = \sum_{k=-\infty}^{\infty} s(t + kT_0) \, \forall t < T_0$$

taken from an infinite series of overlapping pulses repeated at a constant frequency $f_0 = 1/T_0$ are equivalent to values at those frequencies of the Fourier Transform of a single pulse.

Proof:

$$R(hf_0) = \int_0^{T_0} r(t)e^{-j2\pi hf_0 t}\,dt =$$
$$\int_0^{T_0} e^{-j2\pi hf_0 t} \sum_{k=-\infty}^{\infty} s(t+kT_0)\,dt =$$
$$\int_0^{T_0} \sum_{k=-\infty}^{\infty} s(t+kT_0)e^{-j2\pi hf_0 t}\,dt =$$
$$\int_0^{T_0} \sum_{k=-\infty}^{\infty} s(t+kT_0)e^{-j2\pi hf_0 t - j2\pi kh}\,dt =$$
$$\int_0^{T_0} \sum_{k=-\infty}^{\infty} s(t+kT_0)e^{-j2\pi hf_0 (t+kT_0)}\,dt =$$
$$\sum_{k=-\infty}^{\infty} \int_0^{T_0} s(t+kT_0)e^{-j2\pi hf_0 (t+kT_0)}\,dt =$$
$$\sum_{k=-\infty}^{\infty} \int_{kT_0}^{(k+1)T_0} s(t)e^{-j2\pi hf_0 t}\,dt =$$
$$\int_{-\infty}^{\infty} s(t)e^{-j2\pi hf_0 t}\,dt\,\square$$

Now let's consider the effect of irregularities in the input signal on the estimation. Let's consider the effect of a change to a voice pulse with index b to the estimation of a voice pulse with index a, that is, within its single period time range

$$(a-\frac{1}{2})T_0 \le t < (a+\frac{1}{2})T_0$$

. Specifically, we will look at the estimation error $\Delta A(hf_0)$ from the actual value $A(hf_0)$ from the original ideal pulse train. This error in a perfect processing method would ideally be zero if $a \neq b$.

First let's consider what contribution b has on $A(hf_0)$. From our definition of a single period of the overlapping pulses, it's

$$\Delta_b A = \int_{-T_0/2}^{+T_0/2} s((a-b)T_0 + t)e^{-j2\pi ft}\,dt.$$

The definite integral can be rewritten using the frequency-domain formula for the rectangular window (i.e. sinc) as

$$\Delta_b A = [e^{j2\pi fT_0(a-b)} \int_{-\infty}^{\infty} s(t)e^{-j2\pi ft}\,dt] \circledast \frac{sin(\pi fT_0)}{\pi f}.$$

Now let's consider the effect of moving voice pulse b in time on the estimation of voice pulse a. Applying a time-shift of Δt_b to b, we get

$$\Delta_{b,shifted} A = [e^{j2\pi fT_0(a-b)} \int_{-\infty}^{\infty} s(t-\Delta t_b)e^{-j2\pi ft}\,dt] \circledast \frac{sin(\pi fT_0)}{\pi f}$$

as the new contribution to $A(hf_0)$. Combining the previous two equations, we get

$$\Delta_{shift} A = [e^{j2\pi fT_0(a-b)} \int_{-\infty}^{\infty} s(t-\Delta t_b)e^{-j2\pi ft}\,dt - e^{j2\pi fT_0(a-b)} \int_{-\infty}^{\infty} s(t)e^{-j2\pi ft}\,dt] \circledast \frac{sin(\pi fT_0)}{\pi f}.$$

Using the property that adding two spectra is equivalent to the Fourier Transform of the sum of the inverse Fourier Transforms of those two spectra, and that the fact that convolution is distributive (which can be

seen by combining the fact that convolution in the frequency-domain is equivalent to multiplication in the time-domain with the property that addition is the same in both domains), this can be simplified to

$$\Delta_{shift}A = [e^{j2\pi fT_0(a-b)} \int_{-\infty}^{\infty} (s(t-\Delta t_b) - s(t))e^{-j2\pi ft}\,dt] \circledast \frac{sin(\pi fT_0)}{\pi f}.$$

Notice that this is actually a negated comb filter with a frequency equal to $\frac{1}{\Delta t_b}$; thus we can use the frequency response formula for a comb filter [4] to obtain

$$\Delta_{shift}A\Big|_{S(f)=\int_{-\infty}^{\infty} s(t)e^{-j2\pi ft}\,dt} =$$
$$[-2e^{-j\pi(f\Delta t_b-\frac{1}{2})}sin(\pi f\Delta t_b)e^{j2\pi fT_0(a-b)}S(f)] \circledast \frac{sin(\pi fT_0)}{\pi f} =$$
$$[2e^{-j\pi(f\Delta t_b-2fT_0(a-b)+\frac{1}{2})}sin(\pi f\Delta t_b)S(f)] \circledast \frac{sin(\pi fT_0)}{\pi f}.$$

Finally,

$$\Delta_{shift}A(hf_0) = \int_{-\infty}^{\infty} 2e^{-j\pi(f\Delta t_b-2fT_0(a-b)+\frac{1}{2})}\frac{sin(\pi(fT_0-h))}{\pi(f-hf_0)}sin(\pi f\Delta t_b)S(f)\,df.$$

Now let's look at the effect of a timbre change of voice pulse b:

$$\Delta_{timbre}A =$$
$$[e^{j2\pi fT_0(a-b)}S_{new}(f)] \circledast \frac{sin(\pi fT_0)}{\pi f} - [e^{j2\pi fT_0(a-b)}S(f)] \circledast \frac{sin(\pi fT_0)}{\pi f} =$$
$$[e^{j2\pi fT_0(a-b)}S_{new}(f) - e^{j2\pi fT_0(a-b)}S(f)] \circledast \frac{sin(\pi fT_0)}{\pi f} =$$
$$[e^{j2\pi fT_0(a-b)}\Delta_{timbre}S(f)] \circledast \frac{sin(\pi fT_0)}{\pi f}.$$

Then,

$$\Delta_{timbre}A(hf_0) = \int_{-\infty}^{\infty} e^{j2\pi fT_0(a-b)}\frac{sin(\pi(fT_0-h))}{\pi(f-hf_0)}\Delta_{timbre}S(f)\,df.$$

To compute the change that results from a change to multiple voice pulses, we can just sum up the individual changes from each one.

Proposed Algorithm

Based on the prior analysis, I propose a new processing technique I call Radiated Voice Pulse Modeling to minimize these estimation errors. As a prerequisite, this technique requires a set of harmonics that have been already estimated. In theory, any applicable processing technique could be obtain these; but another requirement of the proposed algorithm is a signal that just the harmonics approximate; so in practice, only WBHSM can be used to obtain the initial harmonic estimation.

After the harmonics have been estimated, the harmonic envelope is interpolated by any appropriate method (e.g. cubic spline, EpR [5]). This interpolation is used to obtain an interpolated spectrum with L bins (including the negative frequency, Nyquist, and DC bins). An inverse discrete Fourier Transform is then applied to this spectrum to obtain a voice pulse in the time-domain. Because of the interpolation, the values towards the edges of this window in the time-domain should be very close to zero. As an optimization, a fractional delay could have been applied in the frequency-domain so the start and end of the window can be at integer indexes in the time-domain. Since the bins near the edges are close to zero, the effects of this optimization should be negligible.

Next, we can add all of these voice pulses together in the time-domain to obtain a reconstruction of the original signal. Already we can see a benefit over WBHSM - no border interpolation is needed in the synthesis stage since the edges of each pulse are already approximately zero and thus the discontinuity is insignificant. Next, we obtain a residual signal by subtracting the reconstructed signal from the original signal. The idea then is to somehow improve our estimation by using this residual information. Importantly, this residual does not correspond to noise or transients, but actually corresponds to the difference between the original signal, with its irregularities, and our estimation of the spectrum of each voice pulse, with roughly the same irregularities.

One way of representing the reconstruction process is with mapping of each sample of the input to each sample of the time-domain overlapping estimated voice pulses that contribute to it. This can be done using matrix of size (N, M), where N is the number of samples in the signal and M is the maximum number of voice pulses each input signal sample can be mapped to. If L is constant, and there is a maximum fundamental frequency, then

$$M = \lceil \frac{L f_0}{f_s} \rceil$$

where f_0 is the maximum fundamental frequency and f_s is the sampling rate. If a row in a matrix corresponds to less than M voice pulses, then the remaining values are set to zero. By multiplying this matrix by an all-ones vector and setting it equal to a vector of the input samples, we obtain an expression that represents all possible model estimations that result in a perfect reconstruction.

$$\begin{bmatrix} v_{0,299} & v_{1,199} & v_{2,99} & 0 \\ v_{0,300} & v_{1,200} & v_{2,100} & v_{3,0} \\ v_{0,301} & v_{1,201} & v_{2,101} & v_{3,1} \\ v_{0,302} & v_{1,202} & v_{2,102} & v_{3,2} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} s_{299} \\ s_{300} \\ s_{301} \\ s_{302} \\ \vdots \end{bmatrix}$$

Since our aim is to update our existing estimation based upon the residual, we can a matrix of deltas that sums to the residual instead:

$$\begin{bmatrix} \Delta v_{0,299} & \Delta v_{1,199} & \Delta v_{2,99} & 0 \\ \Delta v_{0,300} & \Delta v_{1,200} & \Delta v_{2,100} & \Delta v_{3,0} \\ \Delta v_{0,301} & \Delta v_{1,201} & \Delta v_{2,101} & \Delta v_{3,1} \\ \Delta v_{0,302} & \Delta v_{1,202} & \Delta v_{2,102} & \Delta v_{3,2} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} r_{299} \\ r_{300} \\ r_{301} \\ r_{302} \\ \vdots \end{bmatrix}.$$

Since there are multiple overlapping voice pulses for each input sample, there are multiple possible combinations values for the values of the matrix. To determine the best ones, I propose to make use of several heuristics:

- Since each voice pulse is centered on an estimated glottal closure instant, the difference of the values at same sample index relative to each voice pulse between neighboring voice pulse should be minimal. Although this may be complicated by the previously proposed optimization involving a fractional delay to align the voice pulse windows to integer indexes in the input signal.
- The values close to the edges of the window should be close to zero.
- The samples within each voice pulse that neighbor each other should be close in value.
- In the frequency-domain, neighboring bins should be close in value.
- Also in the frequency-domain, the envelope, particularly the unwrapped phase envelope, should be similar between neighboring voice pulses.

If we are to consider frequency-domain heuristics, another interesting idea would be to consider the bins closer to the harmonics that were originally used to estimate the envelope as having more information than bins further away from those harmonics.

Once the deltas have been calculated, they are added to the original estimation of the voice pulses to obtain a refined estimation of each voice pulse. Further processing can then be done. For example, the voice pulses could be converted back into the frequency-domain. After transforming the voice pulses, the synthesis can be done in the same way the reconstruction was done for calculating the residual. Transforms on the voice pulses work the same way as in WBHSM, except that the frequency resolution is much higher and constant regardless of the fundamental frequency.

The main challenge of implementing this algorithm is how the optimization taking into account the heuristics should be performed. The simplest idea would be to optimize each row of the matrix individually using any appropriate technique (e.g. gradient descent, Simultaneous Perturbation Stochastic Approximation). Then repeat this many times. However this would presumably be very inefficient. Besides, we should ideally be optimizing globally or at least using regions larger than a single sample. Another issue is that optimization procedure may change the deltas to minimize the heuristics, instead of distributing the residual in the way that minimizes the heuristics' cost.

To solve the issue of the optimization process making baseline changes to minimize the heuristics, is to enforce a requirement $\frac{V}{S} \geq C$. Where V is the sum of a row, S is the sum of the absolute values of a row, and C is a constant. This requirement limits the amount of changes that exist with an opposing sign to that of the residual. Of course another issue is to prevent changes that go beyond magnitude of the residual, but this can be classed into the base heuristic that makes the deltas follow the residual. To ensure no baseline changes at all, C would be set to one. On the other end, to allow some amount of variation, i.e. allowing some small deltas with an opposing sign to the residual's sign, C could be set to be some value $0 < C \leq 1$.

In an optimization-based approach, these hard limits may be a problem. Thus I propose using the following differentiable cost function: $c_2 e^{-vc_1}$. Where v is a value that becomes negative when the condition is not met, and c_1 and c_2 are weights. For the previous proposed requirement, $v = \frac{V}{S} - C$, or $v = (\frac{V}{S} - C)S$ if we want it to be scaled also by the magnitude of the changes.

For the optimization algorithm itself, I propose using gradient descent. Of course, this will mean that it will be unlikely that the resulting algorithm. However, for some applications such as single voice synthesizers, this could be acceptable if the analysis is done during the database creation, since the gradient descent process doesn't have to be run for synthesis. Another interesting optimization could be downsampling the signal for the RVPM procedure, and then using the original values for the higher spectral components.

I have also thought of another approach, actually that doesn't use an optimizer at all. This approach presumably would not yield as good results as the optimizer-based approach, but it's more likely to able to run in real-time. The idea is to center a window function on each of the voice pulse onsets. Then, the RVPM residual is normalized according to the sum of all of those window functions. Finally, the residual is added to each voice pulse, after being multiplied by that voice pulse's corresponding window.

A parametric window like the Kaiser-Bessel is a good choice since the parameter can then be fine-tuned. The window values would be fixed if a constant window size is used. If the fractional delay optimization is employed, the voice pulse onset would not be exactly in the middle, so a fractional delay may need to be applied to the window. On the other hand, the difference from this effect may be insignificant enough that it may be able to ignored. Additionally, this window can be modified based on the fact that a real voice pulse should be zero before the start of the opening phase. We can determine a safe maximum opening phase duration, and then set the window to be zero before the earliest possible start.

References

1. Bonada Sanjaume, Jordi. "Wide-Band Harmonic Sinusoidal Modeling", 2008. Music Technology Group, Pompeu Fabra University.
2. Bonada Sanjaume, Jordi. "High-Quality Voice Transformations Modeling Radiated Voice Pulses in the Spectral Domain", 2004. Music Technology Group, Pompeu Fabra University.
3. Lyons, Richard. "Understanding Digital Signal Processing", Third Edition, 2011, pp. 120-123. Pearson Publishing.

4. Lyons, Richard. “Understanding Digital Signal Processing”, Third Edition, 2011, pp. 394-395. Pearson Publishing.
5. Bonada Sanjaume, Jordi; Celma, Òscar; Loscos, Àlex; Ortolà, Jaume; Serra, Xavier. “Singing Voice Synthesis Combining Excitation plus Resonance and Sinusoidal plus Residual Models”, 2001. International Computer Music Conference.
6. Bonada Sanjaume, Jordi. “Voice Processing and Synthesis by Performance Sampling and Spectral Models”, 2008. PhD Dissertation, Pompeu Fabra University.