ANALYSIS AND RESYNTHESIS OF QUASI-HARMONIC SOUNDS: AN ITERATIVE FILTERBANK APPROACH

Harvey D. Thornburg, Randal J. Leistikow

Stanford University, CCRMA
660 Lomita Drive
Stanford, CA 94305

ABSTRACT

We employ a hybrid state-space sinusoidal model for general use in analysis-synthesis based audio transformations. This model, which has appeared previously in altered forms (e.g. [5], [8], perhaps others) combines the advantages of a source-filter model with the flexible, time-frequency based transformations of the sinusoidal model. For this paper, we specialize the parameter identification task to a class of "quasi-harmonic" sounds. The latter represent a variety of acoustic sources in which multiple, closely spaced modes cluster about principal harmonics loosely following a harmonic structure (some inharmonicity is allowed.) To estimate the sinusoidal parameters, an iterative filterbank splits the signal into subbands, one per principal harmonic. Each filter is optimally designed by a linear programming approach to be concave in the passband, monotonic in transition regions, and to specifically null out sinusoids in other subband regions. Within each subband, the constant frequencies and exponential decay rates of each mode are estimated by a Steiglitz-McBride approach, then time-varying amplitudes and phases are tracked by a Kalman filter. The instantaneous phase estimate is used to derive an average instantaneous frequency estimate; the latter averaged over all modes in the subband region updates the filter's center frequency for the next iteration. In this way, the filterbank structure progressively adapts to the specific inharmonicity structure of the source recording. Analysis-synthesis applications are demonstrated with standard (time/pitch-scaling) transformation protocols, as well as some possibly novel effects facilitated by the "source-filter" aspect.

1. INTRODUCTION: ANALYSIS-SYNTHESIS GOALS

A worthwhile model-based transformation goal is to modify salient characteristics (e.g. pitch, time-scale evolution, formant structure, etc.) while preserving at least the essence of more subtly defined, textural characteristics. Organic sources, for instance, betray a rich variety of textures: the exact quality of "breathiness" of a vocal, the gestural squeak as the bow first meets the string, etc. all contribute greatly to the listener's perception of those sounds.

Since textural characteristics are difficult to model, it is desirable that any model for the salient characteristics account explicitly for the part of the signal which cannot be modeled. A fundamental condition is that of *perfect reconstruction*: if the model undergoes an identity transformation, the resynthesis is an exact copy of the original signal.

The signal decomposition for a perfect reconstruction system may be expressed:

$$y_t \quad = \quad m_t(\theta, y_{1:t-1}) + r_t, \forall t \in 1:N \tag{1}$$

Here y_t is the signal (either *input* or *resynthesis*) evaluated at time t, $y_{1:t-1}$ represents past signal values[1], θ is the *model parameter*, m_t is the *model resynthesis*, and r_t is the *residual*. In general, at least for a causal implementation, the model resynthesis can depend on past values of the signal.

2. OUTPUT RESIDUAL VS. SOURCE-FILTER MODELS

Two historically important classes of perfect reconstruction models are the *output residual* and *source-filter* models. Output residual models satisfy (1) directly, with resynthesis independent of past outputs.

For instance, SMS (Spectral Modeling Synthesis) [6] is an output residual model that enjoys widespread use. Here m_t becomes a time-varying sinusoidal model:

$$m_t(\theta) \quad = \quad \sum_{k=1}^{p} \left[\alpha_{k,t} \cos \left(\sum_{s=1}^{t} \omega_{k,s} + \phi_{k,t} \right) \right] \tag{2}$$

where $\theta = \{\alpha_{k,t}, \phi_{k,t}\}_{k=1:p,t=1:N} \cup \{\omega_k\}_{k=1:p}$. In practice, the amplitudes, frequencies and phases are estimated as piecewise constant on a framewise basis, then interpolated across frames [6]. The different interpolation choices for the frequency terms versus phase terms in (2) effectively resolve the inherent ambiguity between these terms.

Source-filter models, on the other hand, exploit the recursive dependence of m_t on $y_{1:t-1}$ in such a way that r_t appears at the *input*, i.e., as a "driving term" for the main recursion involving y_t. For instance, LPC (Linear Predictive Coding) [1] is one of the simplest source-filter models with audio processing applications (especially in the area of speech processing). In fact, LPC admits a linear transfer relation from residual to output. Here,

$$m_t(\theta, y_{1:t-1}) \quad = \quad \sum_{k=1}^{p} a_k y_{t-k} \tag{3}$$

where $\theta = a_{1:p}$ are the prediction coefficients.

Substituting (3) into (1) yields a linear recursion for y_t:

$$y_t = \sum_{k=1}^{p} a_k y_{t-k} + r_t \tag{4}$$

As (4) is recognized as a linear difference equation, one may derive

$$\frac{Y(z)}{R(z)} \quad = \quad \frac{1}{1 - A(z)} \tag{5}$$

[1] $A:B$ represents the sequence of integers $\{A, A+1, \ldots, B-1, B\}$.

where $A(z) = \sum_{k=1}^{p} a_k z^{-k}$. Hence, r_t represents a broadband "source" (usually interpreted as white noise and/or a periodic impulse train at the pitch period.)

Source-filter models possess several advantages over the output residual models. First, the input residual's role as an excitation for a dynamical system indicates that its energy will likely be concentrated in time. The time concentration property aids in transient detection. When a framewise parameter estimation is used, resynthesis artifacts can be minimized by aligning frame boundaries with the transient boundaries. Moreover, the excitation interpretation makes it more likely that the transient boundary corresponds to a musically meaningful event, such as a note onset.

Second, the input residual as "excitation" is likely to exhibit little cross-dependence on the model resynthesis. The latter makes viable *cross-synthesis*, where sounds may be hybridized by the interchange of models and residuals. Nevertheless the underlying model structures of the classical source-filter models (LPC, acoustic models, etc.) seem less amenable than the output-residual sinusoidal models to arbitrary time-frequency modifications.

3. HYBRID STATE-SPACE SINUSOIDAL MODEL

In this paper, we employ a general state-space resynthesis approach for extended sinusoidal models which has appeared previously in different forms; e.g., [8], [5], etc. Our method presents a hybrid approach, comprising *both* input and output residuals.

In the absence of any residual apart from the initial excitation, our model yields the following output:

$$y_t = \sum_{k=1}^{p} \left[\alpha_{k,t} e^{\sum_{u=1}^{t} \gamma_{k,u}} \cos\left(\sum_{u=1}^{t} \omega_{k,u} + \phi_{k,t} \right) \right] \quad (6)$$

The allowance for exponential decays is particularly useful in modeling the onset regions of many acoustic sounds.

Equation (6) admits a state-space resynthesis, as follows. Let $s_t \in \mathbb{R}^{2p}$ denote the *state* at time t. s_t encodes the information necessary to reconstruct the amplitudes and phases of all component sinusoids. Precisely, $s_t(2k-1)$ encodes the in-phase and $s_t(2k)$ the quadrature component of the k^{th} sinusoid. The amplitude and phase terms are retrieved accordingly:

$$\begin{aligned} \alpha_{k,t} &= \sqrt{s_t^2(2k-1) + s_t^2(2k)} \\ \phi_{k,t} &= \tan^{-1}\left[s_t(2k)/s_t(2k-1) \right] \end{aligned} \quad (7)$$

The state undergoes the recursion:

$$s_t(2k{-}1{:}2k) = F_t(2k{-}1{:}2k) s_{t-1}(2k{-}1{:}2k) + r_{i,t}(2k{-}1{:}2k) \quad (8)$$

where $r_{i,t} \in \mathbb{R}^{2p}$ is the *input residual*, and

$$F_t(2k-1:2k) = e^{-\gamma_{k,t}} \begin{bmatrix} \cos(\omega_{k,t}) & -\sin(\omega_{k,t}) \\ \sin(\omega_{k,t}) & \cos(\omega_{k,t}) \end{bmatrix} \quad (9)$$

The output y_t sums over the in-phase components plus an *output residual*, $r_{o,t} \in \mathbb{R}$:

$$y_t = H x_t + r_{o,t} \quad (10)$$

where $H \in \mathbb{R}^{1 \times 2p}$; $H(2k-1) = 1$; $H(2k) = 0 \; \forall k$.

The residuals are modeled stochastically, as white Gaussian processes:

$$\begin{aligned} r_{i,t} &\sim \mathcal{N}(0, rI) \\ r_{o,t} &\sim \mathcal{N}(0, q) \end{aligned} \quad (11)$$

Given the frequency/decay trajectories $\{\omega_{k,t}, \gamma_{k,t}\}_{k=1:p, t=1:N}$, and a noninformative initial state distribution, $p(s_0) \sim \mathcal{N}(0, \infty I)$, we have a complete Markov model for the observations $\{y_t\}_{t=1:N}$, i.e.:

$$\begin{aligned} s_t &\sim \mathcal{N}(F_t s_{t-1}, qI) \\ y_t &\sim \mathcal{N}(H x_t, r) \end{aligned} \quad (12)$$

Here $F_t = \text{blockdiag}_{k=1:p} F_{k,t}$.

The Kalman filter may be used to derive the posterior state distributions based on present/past outputs, i.e.

$$p(s_t | y_{1:t}) \sim \mathcal{N}\left(\hat{s}_{t|1:t}, P_{t|1:t} \right) \quad (13)$$

Input and output residuals are extracted using $\hat{s}_{t|1:t}$ in place of s_t in (8) and (10); i.e.:

$$\begin{aligned} \hat{r}_{i,t} &= \hat{s}_{t|1:t} - F_t \hat{s}_{t-1|1:t-1} \\ \hat{r}_{o,t} &= y_t - H \hat{s}_{t|1:t} \end{aligned} \quad (14)$$

To summarize, the entire analysis-synthesis algorithm works as shown in Fig. 1:

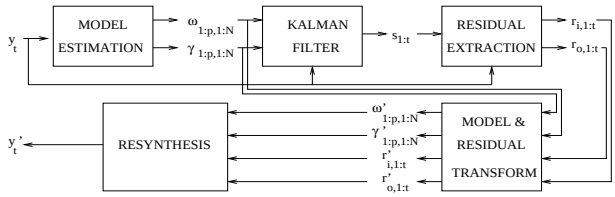

Figure 1: *Generic analysis-synthesis block diagram for the hybrid state-space sinusoidal model*

1. Estimate frequency and decay trajectories: $\{\omega_{k,t}, \gamma_{k,t}\}_{k=1:p, t=1:N}$ over time, by whatever means.
2. Derive Kalman filtered state estimates: $\left\{ \hat{s}_{t|1:t} \right\}_{t=1:N}$
3. Extract input/output residuals via (14).
4. Transform the model, input residuals, and/or output residual by user-specified means.
5. Reconstruct y_t via the state-space resynthesis (8,10).

The specification of q, r (only the ratio, $\rho = r/q$, matters) is vitally important towards obtaining a successful resynthesis. Please see Appx. A.

4. SPECIALIZATION TO QUASI-HARMONIC SOUNDS

A *quasi-harmonic* sound model may represent a variety of single, monophonic recordings of acoustic sounds, especially plucked or struck tones: piano, marimba, bells etc.

The defining criteria are as follows:

1. **QH1**: Frequencies and decay factors are modeled as time invariant for a single analysis frame. However, nonstationarities in the amplitudes and phases may proxy for small, local variations in frequencies and decay rates.
2. **QH2**: All frequencies cluster about principal harmonics. Frequency spacing may be arbitrarily close within a cluster.

3. **QH3**: Principal harmonics exist roughly about a harmonic series, though some inharmonicity is allowed. At minimum, a somewhat uniform separation between each harmonic must be guaranteed.

One may reorganize the generic sinusoidal model to reflect the hierarchy of principal harmonics $k = 1:p$ and clusters $l = 1:p_k$:

$$y_t \quad = \quad \sum_{k=1}^{p} \sum_{l=1}^{p_k} \alpha_{k,l,t} e^{-\gamma_{k,l} t} \cos(\omega_{k,l} t + \phi_{k,l,t}) \qquad (15)$$

Criterion **QH3** may be formalized w.r.t. (15), as follows: There exists $\omega_0 \in [0, 2\pi]$ and small $\epsilon > 0$ such that

$$\sup_{l_1=1:p_{k+1}, l_2=1:p_k} |\omega_{k+1,l_1} - \omega_{k,l_2} - \omega_0| \quad < \quad \epsilon \qquad (16)$$

Often, rich timbral dynamics arise from the coupling interaction of several physical vibrational modes. In piano tones, for instance, coupling between the transversal modes of several strings produces a characteristic "double decay" behavior in which an initial fast decay is followed by a lower amplitude sustained resonance [12], as well as beating and other effects characteristic of the timbral evolution. Though the coupling interaction does not result in the superposition of each vibrational mode in isolation, linearity of the overall system guarantees *equivalent modes* which do superpose [4]. The latter superposition, reflected in (15), can represent the double decay behavior with a stationary parameterization of each decay parameter. Though multiple sinusoids are required, the latter representation is nevertheless more parsimonious and less prone to overfitting risk than that of a single nonstationary sinusoid per harmonic, especially in the presence of beating caused by phase cancellations.

5. ITERATIVE FILTERBANK

The main idea behind our iterative filterbank approach is to perform a "STFT-like" preprocessing especially tailored to the quasi-harmonic model(15). The STFT dissects the input signal into uniformly spaced subbands, the spacing being a function of window length. Even if the window length is chosen to guarantee a spacing of ω_0 via the classic "pitch-synchronous" approach, there is no guarantee each group of modes surrounding a given harmonic will fall within a given subband, thanks to the flexibility in **QH3**. One may instead perform a long DFT of the entire analysis frame, in hopes to isolate each individual frequency within a cluster in their own subbands However, **QH2** precludes the guarantee of a minimum frequency separation.

A more flexible approach is to employ a custom, variable bank of bandpass filters to isolate all modes surrounding a given harmonic in exactly one subband. In this way, we exploit the quasi-harmonicity assumption, without explicitly forcing a rigid pitch-synchronous criterion.

The subband centers, defined as $\left\{\bar{\omega}_k^{(m)}\right\}_{k=1:p}$ for the m^{th} iteration, are initialized as harmonic: $\bar{\omega}_k^{(0)} = k\omega_0$. Here ω_0 is derived from an *ad hoc* DFT-based preprocessing. Through successive iterations, subband centers adapt to the inharmonicity present in the signal. For each subband, we extract the analytic signal, heterodyne such that the center frequency maps to DC, and maximally downsample. The foregoing preprocessing steps are shown in Fig.2.

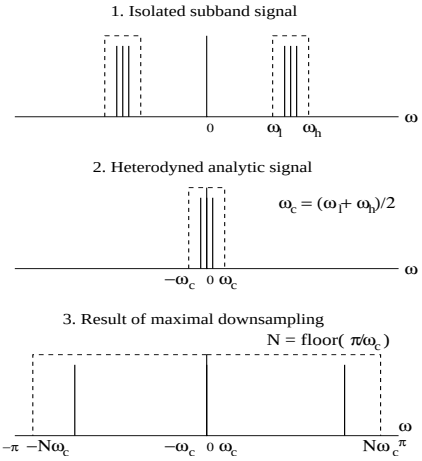

Figure 2: *Preprocessing steps.*

From the heterodyned and downsampled analytic signal $\bar{y}_{k,t}$ a modified Steiglitz-McBride approach [7] estimates the frequencies and decay rates of multiple (complex) sinusoidal components; see [9] for further details. Still in the heterodyned/downsampled domain, a Kalman/Rauch-Tung-Striebel smoother tracks instantaneous phases according to the model:

$$\begin{aligned} \hat{\bar{s}}_{k,t+1}(2l-1:2l) &\sim \mathcal{N}\left(\bar{F}_k(2l-1:2l)\hat{\bar{s}}_{k,t}(2l-1:2l),\ \bar{q}I\right) \\ \bar{y}_{k,t} &\sim \mathcal{N}\left(\bar{H}\hat{\bar{s}}_{k,t}(2l-1:2l), \bar{r}\right) \end{aligned} \qquad (17)$$

Here $\bar{F}_k$ is analogous to F_t in (9), except concerning the sinusoids only in the k^{th} cluster, and $H(2l-1) = 1, H(2l) = \sqrt{-1}$.

Define the (unwrapped) instantaneous phase and frequency estimates:

$$\begin{aligned} \hat{\bar{\phi}}_{k,l,t} &= \text{unwrap}\left(\tan^{-1}\left[\hat{\bar{s}}_{k,t}(2l)/\hat{\bar{s}}_{k,t}(2l-1)\right]\right) \\ \hat{\bar{\omega}}_{k,l} &= \left(\hat{\bar{\phi}}_{k,l,N_k} - \hat{\bar{\phi}}_{k,l,1}\right) / (N_k - 1) \end{aligned} \qquad (18)$$

where N_k, the number of samples in the downsampled subband, equals $N/N_{ds,k}$; $N_{ds,k}$ being the downsampling factor.

To the degree $\bar{\rho} = \bar{r}/\bar{q}$ is small, the instantaneous frequency may correct possible errors in the Steiglitz-McBride estimate. To the degree y_t is noisy, however, the former will become inaccurate unless $\bar{\rho}$ is increased. We find here that a "balanced" parameterization: $\bar{q} = 1, \bar{r} = 1$ where the state covariance is initialized as $P_0 = 10^6 I$ yields acceptable tracking performance

The next iteration's center frequency is updated by averaging instantaneous frequency estimates, then adjusting for the downsampling and heterodyning:

$$\omega_k^{(m+1)} \quad = \quad \omega_k^{(m)} + \frac{1}{N_{ds,k} \cdot p_k} \sum_{l=1}^{p_k} \hat{\bar{\omega}}_{k,l} \qquad (19)$$

5.1. Optimal filter design

In order to isolate the cluster of modes about a given partial, each passband filter is designed to null out the passband regions about all other partials. Each bandpass filter of odd length $2M + 1$, denoted $h = \{h_n\}_{n=-M:M}$, is designed as zero phase: $h_n = h_{-n}$. Various frequency response constraints are satisfied at a uniformly

sampled collection of frequency grid points: $\{\omega_{g,j}\}_{j=1:N_g}$, to produce the typical response shown in Fig. 3. Generally, we set N_g to 1.7 to 2 times the filter length.

Constraints are either *hard* (satisfied exactly) or *soft* (satisfied within tolerance $c^{(j)}\tau$). Here $1/c^{(j)}$ is the *importance weight* for the j^{th} constraint. The optimization proceeds as follows.

$$\begin{array}{lll} \text{Min.} & \tau & \\ \text{subj. to} & \left\{ \begin{array}{rcl} \left|\mathcal{L}^{(j)}_{\text{soft}}[h] - b^{(j)}_{\text{soft}}\right| & \leq & c^{(j)}\tau \\ \mathcal{L}^{(j)}_{\text{hard}}[h] & \leq & b^{(j)}_{\text{hard}} \end{array} \right. & (20) \end{array}$$

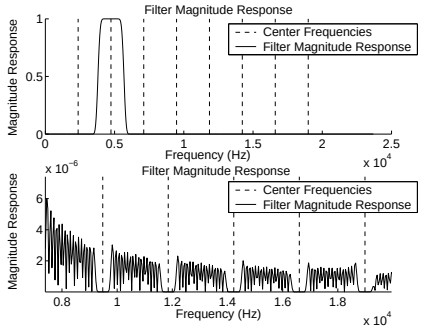

Figure 3: *Optimal bandpass filter response.*

If each $\mathcal{L}^{(j)}_{\text{soft}}[h]$ and $\mathcal{L}^{(j)}_{\text{hard}}[h]$ are linear in h, then clearly (20) admits a linear programming optimization.

The general form of $\mathcal{L}[h]$ is as a q^{th} derivative response constraint $\mathcal{L}[h] \propto \partial^q H(\omega_{g,j})/\partial\omega^q_{g,j}$. It is shown [9] that $\mathcal{L}[h]$ is linear in h:

$$\frac{\partial^q H(\omega)}{\partial\omega^q} \quad = \quad \mathbf{1}_{\{q=0\}} h_0 + 2\sum_{n=1}^{M} n^q T_q(\omega n) h_n \qquad (21)$$

where $T_q(\omega) = \partial^q/\partial\omega^q \cos(\omega)$.

To specify the constraints, we partition the desired frequency response into five regions: passband, low transition width, high transition width, nulling region (the union of all the nulled regions about the other partials), and general stopband. Hard constraints are summarized in Table 1, and soft constraints are summarized in Table 2.

Hard Constraints		
Type	*Constraint*	*Region*
Concave passband	$\partial^2 H(\omega_{g,j})/\partial\omega^2_{g,j} \leq 0$	Passband
Monotonic transition	$\partial H(\omega_{g,j})/\partial\omega_{g,j} \leq 0$	Low TW
	$-\partial H(\omega_{g,j})/\partial\omega_{g,j} \leq 0$	High TW

Table 1: *Hard optimization constraints, which must be met exactly.*

Values $c_s = 2.5, c_{null,0} = 0.05, c_{null,1} = 0.25$ produce the example stopband responses shown in Fig. 4. Here the fundamental frequency is 2375 Hz with a filter length of 499 samples. The combination of zeroth and first order derivative constraints enables attenuation in the nulling region on the order of 10^{-4} times the stopband attenuation, without causing the response artifacts due to setting $c_{null,0} = 10^{-4} c_{stop}$ directly.

Signals with low fundamental frequencies, such as an A0 piano tone, may contain hundreds of audible harmonics up to the Nyquist frequency. As such, a filter with hundreds of closely-spaced nulling regions must be designed. Though the optimization ensures as short as possible filter meeting the objectives, the filter length required for sounds such as the A0 piano tone (fundamental = 27.5 Hz) is on the order of thousands, of samples. With current processing capabilities (1.5 GHz Athlon), any filter over length 2000 seems prohibitive in repeated trials, taking several hours to design.

Soft Constraints		
Type	*Constraint*	*Region*
Passband	$\lvert H(\omega_{g,j}) - 1\rvert \leq \tau$	Passband
General Stopband	$\lvert H(\omega_{g,j})\rvert \leq c_s\tau$	General Stopband
Nulling	$\lvert H(\omega_{g,j})\rvert \leq c_{null,0}\tau$	Nulling Region
	$\lvert H(\omega_{g,j})\rvert \leq c_{null,1}\tau$	

Table 2: *Soft constraints, satisfied within tolerance $c^{(j)}\tau$.*

The reason for preferring a short filter at whatever the design cost is because a shorter filter will bias the sinusoids' decay profiles by a lesser amount, leading to an improved decay estimation. Computational cost is technically a secondary issue; nevertheless a compromise has been necessary in order to cut the design time to under a minute using length $400 - 500$ prototype filters.

Our compromise consists of a multistage approach in which a short secondary bandpass filter with constant 3 dB/octave decay is cascaded with a nulling filter which only nulls out the set of center frequencies within the passband region of the secondary filter. The nulling filter is first designed as a prototype in a stretched frequency space, then this prototype is upsampled via bandlimited interpolation and single sideband modulated. Fig. 4 illustrates the individual responses of the filters to be cascaded.

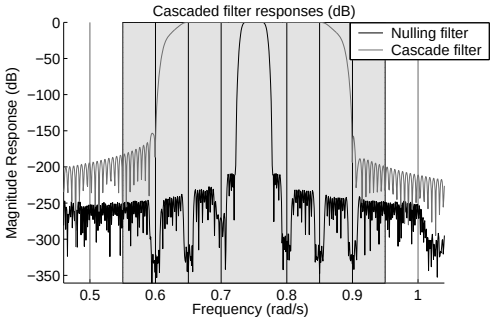

Figure 4: *Nulling filter and secondary bandpass filter.*

6. CONVERGENCE RESULTS

A typical convergence result is shown in Fig. 5 for the first eight partials of a piano tone over five filterbank iterations. The filterbank input consists of a recording of the piano note C2 (65 Hz) of duration 8.53 seconds at 44100 Hz. Each filterbank center frequency is initialized at 95% of the ideal harmonic profile estimated via [2], to simulate a gross frequency initialization error. True vs. initialized frequency profiles are shown in Table 3, as well as the filterbank center frequencies after four and five iterations.

Filter bandwidths are each initialized to 40% of the distance between neighboring center frequencies, and 33% of this region is devoted to transition width.

Partial No.	*"Ideal" Frequency*	*Freq. Init.*	*Freq. Iteration # 4*	*Freq. Iteration # 5*
1	65.41 Hz	62.14 Hz	65.47 Hz	65.48 Hz
2	130.83 Hz	124.29 Hz	131.07 Hz	131.08 Hz
3	196.29 Hz	186.48 Hz	196.81 Hz	196.81 Hz
4	261.82 Hz	248.71 Hz	262.57 Hz	262.57 Hz
5	327.38 Hz	311.00 Hz	328.39 Hz	328.40 Hz
6	393.03 Hz	373.38 Hz	394.29 Hz	394.29 Hz
7	458.80 Hz	435.85 Hz	460.26 Hz	460.26 Hz
8	524.66 Hz	498.43 Hz	526.40 Hz	526.40 Hz

Table 3: *Filterbank initialization and convergence.*

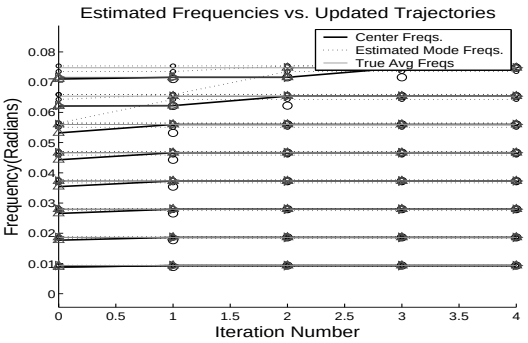

Figure 5: *Filterbank convergence.*

7. ANALYSIS-SYNTHESIS RESULTS

The ability to transform input residual, output residual, and model dynamics by independent means facilitates an immense variety of modifications. Aside of the standard pitch- and time-scaling effects, the separate modification of each input residual may generate some novel effects, such as applying a different regenerative delay to each input residual. Nevertheless, the standard pitch- and time-scaling protocols provide a sort of "benchmark" for comparison with other analysis-synthesis schemes, so we first detail how these are implemented within the present framework.

7.1. Pitch scale modifications

Each of the model's frequency trajectories are multiplied by the amount of the pitch scale. All residuals and decay trajectories are preserved. Clearly, one may generalize the pitch scaling model to allow separate modifications, and/or time-varying, audio rate pitch modulations.

7.2. Time scale modifications

Time scaling is rather more involved. First, one resamples the frequency and decay trajectories by a bandlimited interpolation (linear interpolation can be used if computations are scarce), and multiplies the decays by the inverse of the stretching factor. Next, each input/output residual must be individually timestretched. Since the residuals are by definition the parts which cannot be modeled, a nonparametric approach such as WSOLA [11] must be used. To the extent the output/input residuals are small, the WSOLA artifacts will be suppressed in the overall resynthesis. However, a further improvement occurs when the "excitation region" (defined as the sample window of length $2p$ maximizing the sum of the residual energies) is held out from the WSOLA algorithm then respliced with the WSOLA resynthesis. The resynthesized residuals are presented to the transformed model for the overall resynthesis (recall Fig. 1).

7.3. Cross synthesis

Many possibilities exist for the cross-synthesis or hybridization of multiple sounds. A traditional approach, adhering strictly to the "source-filter" interpretation, is that the input residuals for one model are resented as input to another model, while the output residual of the latter is preserved.

7.4. Results

The setting of $\rho = r/q$ is vital, for reasons detailed in Appx. A which will merely be summarized here. Our examples use $\rho = 20p$ with minimal artifacts. If ρ is too small, however, the input residuals and hence the individual state resyntheses become large, posisbly larger than y_t. However, since the output residual vanishes under small ρ, a precarious situation is established where large input resyntheses must undergo phase cancellation while summing to a model resynthesis closely approximating y_t. This cancellation may not survive model transformation. If so, the resynthesis amplitude envelope may become distorted; for instance, the "soft attack" phenomenon may occur under pitch-scaling when cancellation is maintained at the beginning of the signal yet dissipates over time as the component sinusoids drift out of their original phase relationships.

Such artifacts are easily recovered, however, by a simple envelope adjustment in postprocessing. By contrast, if one considers the loss of phase coherence or the stuttering behavior of canonical pitch-scaling algorithms, the latter prove far more difficult to correct in postproduction than the artifacts generated by our model.

A. TUNING OF THE RESIDUAL VARIANCES

Residual variance parameters q, r balance the distribution of energy among input residual and output residuals as well as govern the ability to track amplitude and phase nonstationarities. In fact, only the ratio $\rho = r/q$ matters. A high ρ means the filtered state estimate only loosely tracks the output, instead depending heavily on the prediction from the previous iteration's state estimate, yielding a large output residual but small state residuals. However, the lack of a swift response to y_t limits the state's ability to track amplitude/phase nonstationarities.

If ρ is small, however, large state residuals but a small output residual results. In the latter case, the magnitudes of the *individual* state resyntheses $\left\{\hat{s}_{t|1:t}(2k-1)\right\}_{k=1:p}$ may be large even with respect to y_t, but the individual resyntheses added together magically "cancel" to form a close approximation to y_t.

Indeed, underspecification of ρ can be problematic under *modifications* to F_t; e.g. pitch shift. For t large enough, the state resyntheses no longer maintain their original phase relationships, thus cancellation no longer occurs. Here the resynthesis' amplitude increases after the attack and the impression of a proportionately soft attack results. A closer view reveals the soft attack to be a classic symptom of the *overfitting phenomenon*: The Kalman filter tries too hard to fit the *specific* input y_t, failing to generalize

to the *class* of signals represented by y_t and/or signals undergoing transformation.

The foregoing observations concerning q, r are easily derived from the Kalman filter equations. The latter come by way of Bayes' Rule and the Markov independences $p(s_{t+1}|s_t, y_{1:t}) = p(s_{t+1}|s_t)$; $p(y_{t+1}|s_{t+1}, y_{1:t}) = p(y_{t+1}|s_{t+1})$, i.e.:

$$\begin{aligned} p(s_{t+1}|y_{1:t}) &= \int_{\mathbb{R}^{2p}} p(s_{t+1}|s_t)p(s_t|y_{1:t})ds_t \\ p(s_{t+1}|y_{1:t+1}) &= \frac{p(y_{t+1}|s_{t+1})p(s_{t+1}|y_{1:t})}{p(y_{t+1}|y_{1:t})} \end{aligned} \quad (22)$$

Straightforward application of the Gaussian potential rules in [3] transforms (22) as follows:

$$\begin{aligned} \hat{s}_{t+1|1:t} &= F_{t+1}\hat{s}_{t|1:t} \\ P_{t+1|1:t} &= F_{t+1}P_{t|1:t}F_{t+1}^T + qI \\ K_{f,t+1} &= P_{t+1|1:t}H^T(HP_{t+1|1:t}H^T + r)^{-1} \\ \hat{s}_{t+1|1:t+1} &= \hat{s}_{t+1|1:t} + K_{f,t+1}\left(y_{t+1} - H\hat{s}_{t+1|1:t}\right) \\ P_{t+1|1:t+1} &= \left(I - K_{f,t+1}H\right)P_{t+1|1:t} \end{aligned} \quad (23)$$

We now state and prove several facts concerning the setting of q, r:

- **S1** Only $\rho = r/q$ matters. That is, if for $c > 0$, $r' \triangleq cr$, $q' \triangleq cq$, and (23) are initialized by cP_0 in place of P_0, identical expressions for $\hat{s}_{t|1:t}$ and $P_{t|1:t}$ result.
- **S2** The relative contribution of $F_t\hat{s}_{t|1:t}$ vs y_t towards $\hat{s}_{t+1|1:t+1}$ increases with ρ. In the extreme case $\rho \uparrow \infty$, $\hat{s}_{t+1|1:t+1} = F\hat{s}_{t|1:t}$: the state residual vanishes.
- **S3** As $\rho \downarrow 0$, $y_{t+1} - H\hat{s}_{t+1|1:t+1} \rightarrow 0$, i.e. the output residual vanishes.

Proof S1 From (23) the identities follow:

$$P_{t+1|1:t+1} = \left[\left(F_{t+1}P_{t|1:t}F_{t+1}^T + qI\right)^{-1} + r^{-1}H^TH\right]^{-1} \quad (24)$$

$$\begin{aligned} K_{f,t+1} &= \left(F_{t+1}P_{t|1:t}F_{t+1}^T + qI\right)H^T \\ &\quad \times \left[H\left(F_{t+1}P_{t|1:t}F_{t+1}^T + qI\right)H^T + r\right]^{-1} \end{aligned} \quad (25)$$

$$\hat{s}_{t+1|1:t+1} = \left(I - K_{f,t+1}H\right)F\hat{s}_{t|1:t} + K_{f,t+1}y_{t+1} \quad (26)$$

Define $P^{(c)}_{t+1|1:t+1} = \left[\left(F_{t+1}P_{t|1:t}F_{t+1}^T + cqI\right)^{-1} + (cr)^{-1}H^TH\right]^{-1}$ The latter simplifies as follows:
$c^{-1}P^{(c)}_{t+1|1:t+1} = \left[\left(F_{t+1}(c^{-1}P_{t|1:t})F_{t+1}^T + qI\right)^{-1} + r^{-1}H^TH\right]^{-1}$
Hence, if q is replaced by cq and r by cr,
$P^{(c)}_{t+1|1:t+1} = cP_{t+1|1:t+1}\forall t$. The Kalman recursion for $P_{t|1:t}$ is unchanged except for the initialization: $P^{(c)}_0 = cP_0$.

Similarly, define $K^{(c)}_{f,t+1}$ via (25) replacing $P_{t|1:t}$ by $P^{(c)}_{t|1:t}$ and q by cq. It follows that $K^{(c)}_{f,t+1} = K_{f,t+1}\forall t$. Since $K^{(c)}_{f,t+1}$ is unaltered and no other term in (26) depends on c, $\hat{s}_{t|1:t}$ is unchanged for all t.

Proof S2 Fix $q = 1$, so $\rho = r$; by **S1** no loss of generality occurs. As r grows, $\left[H\left(F_{t+1}P_{t|1:t}F_{t+1}^T + qI\right)H^T + r\right]^{-1}$ decreases; thus $K_{f,t+1}$ decreases elementwise and the eigenvalues of $(I - K_{f,t+1}H)$ increase. Thus the relative contribution of $F_t\hat{s}_{t|1:t}$ towards $\hat{s}_{t+1|1:t+1}$ increases. When $r \uparrow \infty$, $K_{f,t+1} = 0$, so $\hat{s}_{t+1|1:t+1} = F\hat{s}_{t|1:t}$: all state residuals vanish.

Proof S3 Again fix $q = 1$. Multiplying both sides of (26) on the left by H obtains

$$H\hat{s}_{t+1|1:t+1} = \left(H - HK_{f,t+1}H^t\right)F_{t+1}\hat{s}_{t|1:t} + HK_{f,t+1}y_{t+1} \quad (27)$$

Since $HK_{f,t+1} = HP_{t|1:t}H^T\left(HP_{t|1:t}H^T + r\right)^{-1}$ converges to 1 as $r \downarrow 0$, (27) becomes simply $H\hat{s}_{t|1:t+1} = y_{t+1}$. The model resynthesis perfectly tracks the output, hence the output residual vanishes.

B. REFERENCES

[1] Atal, B.S., Hanauer, S.L., "Speech analysis and synthesis by linear prediction of the speech wave", J. Acoust. Soc. Am. 50: 637-655, 1971

[2] Bensa, J., Bilbao, S., et al. "From the physics of piano strings to digital waveguides", Proc. 2002 International Computer Music Conference, Göteborg, Sweden, 2002.

[3] Murphy, K. "Filtering, Smoothing, and the Junction Tree Algorithm", `http://www.ai.mit.edu/~murphyk/Papers/smooth.ps.gz`, 1999

[4] Nakamura, I. " Fundamental theory and computer simulation of the decay characteristics of piano sound", J. Acoustical Society of Japan 10(5), 289-297, 1989.

[5] Qi, Y., Minka, T.P., Picard, R.W., "Bayesian Spectrum Estimation of Unevenly Sampled Nonstationary Data", MIT Media Lab Technical Report Vismod-TR-556, 2002.

[6] Serra, X. and Smith, J.O. III "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System based on a Deterministic plus Stochastic Decomposition", Computer Music Journal 14(4): 12-24, 1990.

[7] Steiglitz, K., and McBride, L.E. "A Technique for the Identifi cation of Linear Systems", IEEE Trans. Automatic Control, AC-10: 461-464, 1965.

[8] Thornburg, H. and Gouyon, F. "A Flexible Analysis-Synthesis Method for Transients", Proc. Int'l Computer Music Conf. 2000, Berlin.

[9] Thornburg, H. and Leistikow, R. "An Iterative Filterbank Approach for Extracting Sinusoidal Parameters From Quasi-Harmonic Sounds", Proc. IEEE WASPAA 2003, New Paltz, NY.

[10] Vandenberghe, L. and Boyd, S. "Positive Defi nite Programming", SIAM Review, 38(1): 49-95, March 1996.

[11] Verhelst, W. and Roelands, M. "An overlap-add technique based on waveform similarity(WSOLA) for high-quality time-scale modifi cation of speech", Proc. IEEE ICASSP 1993, Minneapolis, pp.554-557

[12] Weinreich, G. "Coupled Piano Strings", J. Acoust. Soc. Amer., 62(6), pp. 1474–84, 1977.