Instantaneous and Frequency-Warped Techniques for Source Separation and Signal Parametrization

Avery L. Wang

Chromatic Research, Inc.
800A East Middlefield Rd.
Mountain View, CA 94043-4030
avery@chromatic.com

ABSTRACT

This paper summarizes several contributions to the study of audio signal parameterization and source separation, more fully described in [1]. The philosophy of *Frequency-Warped Signal Processing* provides powerful means for separating the AM and FM contributions to the MS bandwidth of a complex-valued, frequency-varying sinusoid $p[n]$, transforming it into a signal with slowly-varying parameters. The use of frequency- and harmonic-locked loops to track the phase variation of analytic signals provides the necessary accurate instantaneous frequency information to drive frequency-warped filtering.

1. Bandwidth

The mean-square bandwidth of a signal $z(t)$ may be expressed as [2]

$$\sigma_{\text{BW}}^2 \triangleq \frac{\int_{-\infty}^{\infty}(f-\langle f\rangle)^2|Z(f)|^2\,df}{\int_{-\infty}^{\infty}|Z(f)|^2\,df} \tag{1}$$

$$= \frac{\int_{-\infty}^{\infty}\left[\left(f(t)-\langle f\rangle\right)^2+\left[\frac{a'(t)}{2\pi a(t)}\right]^2\right]|z(t)|^2\,dt}{\int_{-\infty}^{\infty}|z(t)|^2\,dt} \tag{2}$$

$$= \sigma_{\text{FM}}^2+\sigma_{\text{AM}}^2, \tag{3}$$

where $a(t)=|z(t)|$,

$$\langle f\rangle \triangleq \frac{\int_{-\infty}^{\infty}f|Z(f)|^2\,df}{\int_{-\infty}^{\infty}|Z(f)|^2\,df}, \tag{4}$$

and σ_{FM}^2 and σ_{AM}^2 are respectively the FM and AM contributions to the total mean-square bandwidths.

Conventional signal processing techniques often make use of the Fourier transform, which is suited for the spectral analysis of stationary signals, i.e., signals whose statistics are time-independent [3, 4]. However, the Fourier transform is ill-suited for analyzing signals with rapidly varying parameters. By frequency warping a signal, i.e., selectively frequency modulating it, it may be possible to transform it so that the resulting signal is stationary and has a simplified spectrum which is more amenable to analysis. Frequency warping may be thought of as attempting to "straighten out" nonstationarities due to continuous variations in the instantaneous frequency of a signal. The FM component σ_{FM}^2 of the mean-square bandwidth may be minimized, leaving only the AM component.

2. Instantaneous Frequency

Instantaneous Frequency has been discussed by several authors previously, for example [2, 5]. It may be defined for an analytic signal $p(t)$ as

$$f(t) \triangleq \frac{1}{2\pi}\frac{d\arg\{p(t)\}}{dt}. \tag{5}$$

This definition is fraught with perils, as it is possible to get arbitrarily large values with well-behaved signals [1, 2]. It makes sense to talk about instantaneous frequency if we restrict the dialog to signals of the form

$$p(t) = a(t)\exp\left(j2\pi\int_0^t f(\tau)\,d\tau\right), \tag{6}$$

where $a(t) > 0$. We see here that in this context $f(t)$ has reasonably intuitive meaning.

3. Frequency-Warping

Let $\xi(t)$ be a continuous, real-valued function specifying the *ξ-frequency warp factor* defined by

$$\Xi(t) \triangleq \exp\left(j2\pi\int_0^t \xi(\tau)\,d\tau\right). \tag{7}$$

Intuitively, the frequency-warping factor $\Xi^*(t)$ demodulates a signal

$$p(t) = a(t)\exp\left(j2\pi\int_0^t f(\tau)\,d\tau + j\phi_0\right) \tag{8}$$

by displacing its instantaneous frequency $f(t)$ by $-\xi(t)$. This is easy to see because

$$\Xi^*(t)p(t) = a(t)\exp\left(j2\pi\int_0^t\{f(\tau)-\xi(\tau)\}\,d\tau + j\phi_0\right), \tag{9}$$

and, by Eqn. (5), the instantaneous frequency is

$$\frac{1}{2\pi}\frac{d\arg\{\Xi^*(t)p(t)\}}{dt} = f(t)-\xi(t). \tag{10}$$

Hence, we see the motivation for the name "frequency-warping". We introduce the notation

$$\langle p\,|\,\xi\rangle(t) \triangleq \Xi_\xi^*(t)p(t) \tag{11}$$

and call this the *ξ-frequency-warped transform* of the signal $p(t)$, where $\xi(t)$ is a continuous function.

We note that if we set $\xi(t)$ to be the instantaneous frequency $f(t)$ of a signal, the signal $p(t)$ is then demodulated by $\Xi_f^*(t)$ to a constant-phase signal

$$\langle p\,|\,f\rangle(t) = \Xi_f^*(t)p(t) \tag{12}$$

$$= a(t)\exp\left(j2\pi\int_0^t\{f(\tau)-f(\tau)\}\,d\tau + j\phi_0\right) \tag{13}$$

$$= a(t)\exp(j\phi_0). \tag{14}$$

In this case the demodulating function $\Xi_f(t)$ is *frequency-matched* to $p(t)$. We see that frequency warping a signal $p(t)$ is nothing more than multiplying it by a unit-amplitude phase factor $\Xi^*(t)$. To invert the frequency warping we simply multiply the result by $\Xi(t)$. Notationally, the inversion can be stated as

$$p(t) = \langle \langle p \wr \xi \rangle \wr -\xi \rangle(t) \tag{15}$$

Another obvious fact is the linearity in the first argument of frequency warping:

$$\langle w + z \wr \xi \rangle(t) = \langle w \wr \xi \rangle(t) + \langle z \wr \xi \rangle(t). \tag{16}$$

4. Isolating single time-varying partials from mixtures

If we have a low-pass filter **LPF** with an impulse response $h(t)$ and a passband equal in width to the bandwidth of the amplitude envelope $a(t)$ of $p(t)$ then we may isolate $p(t)$ from an additive mixture $z(t) = p(t) + \nu(t)$, where $\nu(t)$ is some unknown signal. We simply calculate

$$\hat{p}(t) = \langle h * \langle z \wr f \rangle \wr -f \rangle(t), \tag{17}$$

where we have neglected the group delay of the low-pass filter.

To the extent that the frequency-warped interfering signal $\langle \nu \wr f \rangle(t)$ does not intersect the bandwidth of the filter **LPF** we may isolate $p(t)$ relatively cleanly, due to its decreased bandwidth after frequency-matched frequency warping.

We may, similarly, *high-pass* filter the signal $\langle z \wr f \rangle(t)$ with a filter **HPF** whose stop band coincides with the pass band of **LPF**. The resulting signal after unwarping (remodulating) the high-passed signal should be a signal without $p(t)$.

5. Frequency-Locked Loop

In order to apply the frequency-warped technique practically to isolating partials from mixtures it is necessary to obtain a good estimate of the instantaneous frequency $f_k(t)$ for each partial $p_k(t)$ we wish to isolate from a mixture. To implement frequency tracking, a *Frequency-Locked Loop* algorithm is introduced which uses the complex winding error to update its frequency estimate. The input signal is dynamically demodulated and filtered to extract the envelope. This envelope may then be remodulated to reconstruct the target partial, which may be subtracted from the original signal mixture to yield a, quickly-adapting form of notch filtering. Similar work has been done by Costas and Kumaresan [6, 7].

Let the target signal $p[n]$ be a complex-valued discrete-time signal defined for $n \geq 0$ with sampling frequency f_s

$$p[n] = a[n] \exp\left(\frac{j2\pi}{f_s} \sum_{k=1}^{n} f[k] + j\phi_0 \right), \tag{18}$$

where $a[n]$ is the instantaneous amplitude envelope, $f[n] > 0$ is the instantaneous frequency, and ϕ_0 is the phase offset at time $n = 0$.

We assume that $f[n]$ and $a[n]$ are slowly varying with respect to the loop time constant of the tracking system, which will be defined later.

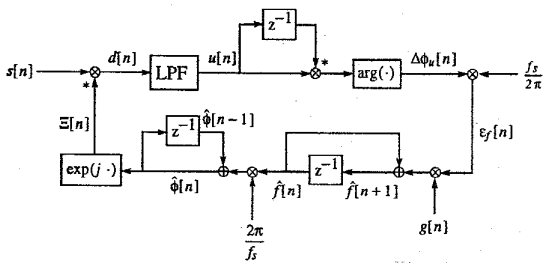

Figure 1: Signal flow for a basic FLL tracker.

Define the *discrete-time slew rate* of the instantaneous frequency as

$$\dot{f}[n] \triangleq \left(f[n] - f[n-1] \right) f_s \tag{19}$$

The input signal $z[n]$ will be assumed to be a mixture of $p[n]$ and some unknown disturbance signal $\nu[n]$, i.e.

$$z[n] = p[n] + \nu[n]. \tag{20}$$

We start the FLL algorithm by demodulating the input signal $z[n]$ by multiplying it by the complex conjugate of the frequency-warping signal

$$\Xi[n] \triangleq \exp\left(j\hat{\phi}[n] \right), \tag{21}$$

where

$$\hat{\phi}[n] \triangleq \begin{cases} \hat{\phi}[n-1] + 2\pi \hat{f}[n]/f_s, & \text{for } n \geq 1 \\ 0, & n = 0 \end{cases} \tag{22}$$

$$= \frac{2\pi}{f_s} \sum_{k=1}^{n} \hat{f}[k], \tag{23}$$

and where $\hat{f}[k] \in (0, f_s/2)$, is the estimate of $f[k]$ at time step k. We assume that $\hat{f}[1]$ is reasonably close to $f[1]$ by the criterion suggested by Eqn. (29).

The complex demodulated signal is

$$d[n] \triangleq \Xi^*[n] z[n] \tag{24}$$

$$= a[n] \exp\left(\frac{j2\pi}{f_s} \sum_{k=1}^{n} (f[k] - \hat{f}[k]) + j\,\phi_0 \right) + \Xi^*[n]\nu[n]. \tag{25}$$

This demodulated signal is subjected to a low-pass filter **LPF** which has a cut-off frequency of f_c, unity gain at DC, and is assumed, for now, to have zero group delay at all frequencies, resulting in a signal $u[n]$ such that

$$u[n] = h_{\text{LPF}} * d[n] \tag{26}$$

$$= h_{\text{LPF}} * \{\Xi^*[n] z[n]\} \tag{27}$$

$$\approx a[n] \exp\left(\frac{j2\pi}{f_s} \sum_{k=1}^{n} (f[k] - \hat{f}[k]) + j\,\phi_0 \right) \tag{28}$$

with the assumptions that the slew rate $\dot{f}[n]$ is small and that

$$\left| f[k] - \hat{f}[n] \right| < f_c. \tag{29}$$

In Eqn. (28) we have assumed that the filtered disturbance term $h_{\text{LPF}} * \{\Xi^*[n]\nu[n]\}$ is negligible. More precisely, if the *winding criterion*

$$h_{LPF} * \{\Xi^*[n]\nu[n]\} < h_{LPF} * \{\Xi^*[n]p[n]\} \tag{30}$$

holds, the *Winding Theorem* guarantees that the phase winding is dominated by the frequency of the partial $p[n]$ [1].

We now consider the phase $\phi_u[n]$ of the signal $u[n]$. By using the *change* in phase we can avoid the phase unwrapping problem:

$$\begin{aligned}
\Delta\phi_u[n] &\triangleq \phi_u[n] - \phi_u[n-1] \qquad \text{(formally)} && (31)\\
&= \arg\left(u[n]u^*[n-1]\right) && (32)\\
&= \frac{2\pi}{f_s}\left(f[n] - \hat{f}[n]\right) + \zeta[n] - \zeta[n-1] && (33)\\
&\approx \frac{2\pi}{f_s}\left(f[n] - \hat{f}[n]\right), && (34)
\end{aligned}$$

where we assume that the phase disturbance $\zeta[n]$ is small. Because the signal $s[n]$ is over-sampled, we are guaranteed that $|\Delta\phi_u[n]| \leq \pi$. Consequently, $\Delta\phi_u[n]$ has a well-defined value which we may calculate by using a standard complex arg$(\cdot)$ function.

Define the frequency tracking error at time n as

$$\begin{aligned}
\varepsilon_f[n] &\triangleq \frac{f_s}{2\pi}\Delta\phi_u[n] && (35)\\
&\approx f[n] - \hat{f}[n]. && (36)
\end{aligned}$$

We see immediately that we may close the loop to form an estimate $\hat{f}[n+1]$ of $f[n+1]$ by computing

$$\begin{aligned}
\hat{f}[n+1] &= \hat{f}[n] + g[n]\varepsilon_f[n] && (37a)\\
&\approx \hat{f}[n] + g[n]\left(f[n] - \hat{f}[n]\right) && (37b)\\
&= (1 - g[n])\hat{f}[n] + g[n]f[n], && (37c)
\end{aligned}$$

where $g[n]$ is the tracking gain of the system at time step n. This system is especially nice because the frequency tracking system simply reduces to a first-order difference equation in $\hat{f}[n]$ with closed-loop gain

$$g_\ell[n] \triangleq 1 - g[n]. \tag{38}$$

The tracking time constant of the system at time step n, in terms of time steps, is given by

$$\begin{aligned}
\tau_\ell[n] &\triangleq \frac{-1}{\log\left(g_\ell[n]\right)} && (39)\\
&\approx \frac{1}{g[n]}, && (40)
\end{aligned}$$

for small, slowly-varying $g[n]$. This system will converge if

$$\forall k, \quad |g_\ell[k]| < 1 \tag{41}$$

6. Harmonic-Locked Loop

The frequency-locked loop (FLL) algorithm of the previous section performs fast and accurate tracking of the instantaneous frequency of a single target partial in isolation. However, if the signal-to-noise ratio is too large, tracking may break down. Acoustical signals are often composed of complex mixtures of signals which bring the signal-to-noise ratio for target partials down below the level needed for tracking by the FLL method. Hence, for analyzing natural signals, the FLL algorithm is an interesting, but fragile, tool for signal analysis. In this section, we find that we may take advantage of the harmonic structure of many natural acoustical signals to increase the robustness of tracking significantly.

A *harmonic signal* $\Gamma[n]$ is the sum

$$\Gamma[n] \triangleq \sum_{k=1}^{M} p_k[n] \tag{42}$$

of the members $p_k[n]$ of a *harmonic set* Γ, which is defined to be a set of M harmonic partials $\Gamma = \{p_k[n]\}_{k=1}^{M}$, where each $p_k[n]$ is of the form in Eqn. (18) and also has

$$f_k[n] \triangleq k f_0[n]. \tag{43}$$

$f_0[n]$ is the *instantaneous fundamental frequency*, or *pitch*, of $\Gamma[n]$.

The problem, then, is to estimate $f_0[n]$ and the amplitudes $a_k[n]$ for some specified range of n from the signal

$$z[n] = \Gamma[n] + \nu[n], \tag{44}$$

where $\nu[n]$ is some unknown signal satisfying certain conditions to be given later.

We take advantage of the information from each tracker by combining each tracker's instantaneous frequency correction term $\varepsilon_{f,k}[n]$ to form a weighted average of the ensemble correction as

$$\hat{\varepsilon}_{f,0}[n] = \sum_{k=1}^{M} w_k[n]\frac{\varepsilon_{f,k}[n]}{k}, \tag{45}$$

where the $\varepsilon_{f,k}[n]$ are defined as in Eqn. (35) and

$$\sum_{k=1}^{M} w_k[n] = 1. \tag{46}$$

There are many possible weighting schemes. A useful scheme is to weight each update estimate by the reciprocal of its ε_f-variance. If we assume that

$$\varepsilon_{f,k}[n] = f_k[n] - \hat{f}_k[n] + \eta_k[n], \tag{47}$$

where $\eta_k[n]$ is assumed to have a zero-mean Gaussian distribution of variance $\sigma^2_{\eta,k}[n]$ and is independent across k, but not necessarily across n, then the k-th estimate of the fundamental update is

$$\begin{aligned}
\hat{\varepsilon}^{(k)}_{f,0}[n] &\triangleq \varepsilon_{f,k}[n]/k && (48)\\
&= f_0[n] - \hat{f}_0[n] + \eta_k[n]/k && (49)
\end{aligned}$$

and the maximum-likelihood fundamental frequency update, under the stated assumptions, is

$$\hat{\varepsilon}^{\dagger}_{f,0}[n] = \frac{\sum_{k=1}^{M}\left(k^2/\sigma^2_{\eta,k}[n]\right)\hat{\varepsilon}^{(k)}_{f,0}[n]}{\sum_{k=1}^{M}\left(k^2/\sigma^2_{\eta,k}[n]\right)}. \tag{50}$$

Eqn. (50) is derived by maximizing the probability

$$p\left(\varepsilon_{f,1}[n],\ldots,\varepsilon_{f,M}[n]\middle|\varepsilon_{f,0}[n]\right) = \frac{1}{(2\pi)^{M/2}\prod_{k=1}^{M}\sigma_{\eta,k}[n]}\exp\left(-\sum_{k=1}^{M}\frac{(\varepsilon_{f,k}[n]-k\varepsilon_{f,0}[n])^2}{2\sigma^2_{\eta,k}[n]}\right). \quad (51)$$

We may form the refined updates of the $f_k[n]$ by setting

$$\hat{\varepsilon}^{\dagger}_{f,k}[n] \triangleq k\hat{\varepsilon}^{\dagger}_{f,0}[n]. \quad (52)$$

From Eqn. (50) and the assumption that the $\hat{\varepsilon}^{(k)}_{f,0}[n]$ are independent across k, we find that the variance of this result is

$$\sigma^2_{\hat{\varepsilon}^{\dagger},0}[n] = \frac{\sum_{k=1}^{M}\left(k^2/\sigma^2_{\eta,k}[n]\right)^2\left(\sigma^2_{\eta,k}[n]/k^2\right)}{\left\{\sum_{k=1}^{M}\left(k^2/\sigma^2_{\eta,k}[n]\right)\right\}^2} \quad (53)$$

$$= \left\{\sum_{k=1}^{M}\left(k^2/\sigma^2_{\eta,k}[n]\right)\right\}^{-1} \quad (54)$$

$$< \min_{k}\left\{\sigma^2_{\eta,k}[n]/k^2\right\}. \quad (55)$$

The variance of $\hat{\varepsilon}^{\dagger}_{f,k}[n]$ is thus

$$\sigma^2_{\hat{\varepsilon}^{\dagger},k}[n] = k^2\sigma^2_{\hat{\varepsilon}^{\dagger},0}[n] \quad (56)$$

$$< k^2\min_{m}\left\{\sigma^2_{\eta,m}[n]/m^2\right\} \quad (57)$$

$$\leq \sigma^2_{\eta,k}[n], \quad (58)$$

where the last inequality may be quite substantial.

Eqns. (50) and (52) may be used to update the individual frequency trackers, using Eqn. (37a), constraining the harmonic frequency estimates to being exact integer multiples of the fundamental. The variance reduction in Eqn. (58) yields a powerful algorithm which can track a harmonic signal through interfering noise. All of the harmonic estimators pool their estimates together, sharing their mutual information so that the ensemble $f_0[n]$ update estimate becomes at least as good as the best single-harmonic estimator, and usually much better.

7. Summary

The novel technique of *Harmonic-Locked Loop* tracking, using N harmonically constrained FLL trackers, results in fast and accurate estimation of the fundamental frequency of harmonic signals, such as voices and certain musical instruments. The estimated fundamental frequency is computed from a maximum-likelihood weighting of the N tracking estimates, making it highly robust. The result is that harmonic signals, such as voices, can be isolated from complex mixtures in the presence of other spectrally overlapping signals. Additionally, since phase information is preserved, the targeted harmonic signals may be resynthesized and removed from the original mixture with relatively little damage to the residual signal.

Applications include music restoration, source separation, speech enhancement, speech compression, parametric MIDI control, and time-frequency signal analysis.

This work was supported by a National Science Foundation Graduate Fellowship, as well as a CCRMA Affiliates Scholarship.

References

1. A. L. Wang, "Instantaneous and frequency-warped techniques for auditory source separation," Tech. Rep. STAN-M-86, Center for Computer Research in Music and Acoustics (CCRMA), Department of Music, Stanford University, Stanford, CA 94305-8180, Aug. 1994.
2. L. Mandel, "Interpretation of instantaneous frequency," *Amer. J. Phys.*, vol. 42, pp. 840–846, 1974.
3. R. N. Bracewell, *The Fourier Transform and its Applications*. New York: McGraw-Hill, second, revised ed., 1986.
4. S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1993.
5. B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal–part 1: Fundamentals," *Proc. IEEE*, vol. 80, pp. 519–538, Apr. 1992.
6. J. P. Costas, "Residual signal analysis–a search and destroy approach to spectral analysis," in *Proc. of the first ASSP Workshop on Spectral Estimation*, (Hamilton, Canada), pp. 6.5.1–6.5.8, Aug. 1981.
7. R. Kumaresan, C. S. Ramalingam, and A. Rao, "RISC: an improved Costas estimator-predictor filter bank for decomposing multicomponent signals," in *Proc. Seventh SSAP Workshop*, (Quebec City), June 1994.

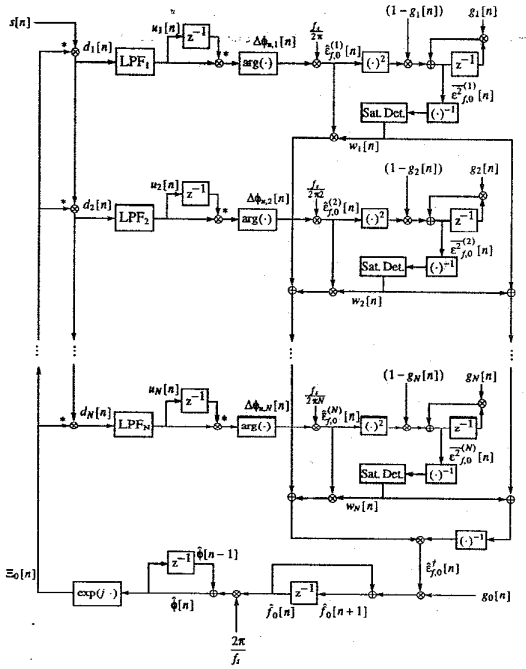

Figure 2: Flow diagram for harmonic-locked loop.