Lecture notes on Instantaneous Frequency and the Phase Vocoder.
S. Dubnov, Music, UCSD

Music is a signal that changes in time. If we think of it in terms of a musical notation, the representation of music consists of a sequence of notes (pitches) played at different times, with additional parameters like intensities, type of the instrument, articulation instructions and etc. Once the score is actually interpreted by the musicians, the acoustic signal preserves this information, with additional information about the interpretation, room acoustics, and etc.

It is only natural, when doing signal analysis of musical recordings, to seek an analysis that would yield a representation close to that of notation. The problem of musical transcription is formidable, and will not be solved in these notes. Our goal, for the moment, is to try and outline the basic concepts related to representing acoustical signals (recording) in time – frequency space, which is at least remotely, a variant of a score since it captures information both about the signal frequency, but it does it in various times, so we can tell what frequencies appear when.

The first step towards time-frequency analysis is actually to define what frequency is. In mechanics, the frequency of vibratory motion is defined as the number of oscillations per unit time, where vibratory motion is any to-and-fro motion and an oscillation is a complete to-and-fro motion. The time it takes to the motion to complete a cycle is called a period. Generating a sine tone on a computer requires calculating a function that would have a desired period. This equivalently creates a signal whose frequency is 1/time of one period. In order to do so we may calculate a set of values of a periodic function, such as a sine function, so, the frequency of a sinusoidal signal is a well defined quantity. However, often in practice, signals are not truly sinusoidal, they turn on and off and their period varies over time. Moreover, signals such as tones of musical instrument that have a well defined pitch, are actually aggregates of sinusoidal components (and so of course chord or signals resulting from simultaneously playing multiple notes). Signals that vary in time are called in engineering jargon “nonstationary signals” and they do not lend themselves to decomposition into pure sinusoidal components. For such signals, the notion of frequency loses its effectiveness, and one needs to use a parameter which accounts for the time-varying nature of the process. This need has given rise to the idea of instantaneous frequency.

Before proceeding to define instantaneous frequency, let us review the definition of frequency for a stationary, infinite duration signal. If the waveform of the signal is not a pure sine wave but is exactly periodic (i.e. the shape of the signal in every period is an exact replica of itself), then one can show that such a signal can be decomposed into sinusoidal components, with all frequencies of the components being integer multiples of the lowest frequency which corresponds to the signal period, called sometimes the “fundamental frequency”. Any stationary signal, not only a purely periodic one, but even signals of no evident periodic shape or of a finite duration, can be represented as the weighted sum of sine and cosine waves with particular frequencies, amplitudes and phases (note that for a particular frequency f, amplitude and phase of sine are constant).

In this situation the concept of frequency is unambiguous, and can be considered in some sense as a mathematical trick, i.e. just a different representation of the same signal,
instead of ‘in-time’, it is a signal in the frequency domain. Going back and forth between 
the two is possible by the Fourier and inverse Fourier transform. Let us imagine a sinusoidal 
motion as a rotation around a circle (actually, the sinusoidal 
shape is drawn as a projection on the y axis of the point that moves around the circle). 
The frequency of this motion and its related sinusoidal wave is the rate with which the 
age of the point on the circle increases. So, if at a moment in time T the angle of the 
point is Theta, then it arrived there by moving with speed of omega radiance per second, 
accomplishing Theta = omega * T radians at time T. In order to have a clear terminology, 
we shall define Theta as the phase. This should not be confuse with initial phase, which is 
a constant factor that tells us at what initial angle the motion started and distinguishes 
between sine and cosine or any other sinusoidal type of signal with initial phase offset. 
Now let us assume that the rotation around the circle is with varying speed. IF is a 
generalization of the definition of constant frequency, i.e., it is the rate of change of phase 
age at time t, which becomes an instantaneous property since the “speed” of the 
rotation can change in time. The problem is that if the changes in speed are very irregular, 
the whole process losses the meaning of periodicity and frequency. In 1946 Van der Pol 
approached the problem of instantaneous frequency by analyzing an expression with time 
varying amplitude and phase.

Then the frequency becomes the derivative of the angle, or writing it the other way 
around, the angle Theta becomes and integral of IF as function of time, plus possibly an 
initial phase factor. The question now becomes how can we extract or measure IF? 
In the case of a stationary periodic signal, the standard Fourier analysis “does the job”. 
Actually if the signal was a complex exponential (a “phasor”) instead of a sine or cosine, 
we could directly measure the frequency as derivative of the argument, i.e. if 
$z(t) = a(t)e^{-j\omega(t)}$ then $f(t) = \frac{1}{2\pi} \frac{d}{dt} \arg[z(t)]$.

In 1947 Gabor actually proposed a method for generating a unique complex signal from a 
real one by doing a Fourier analysis of the real signal and then to “suppress the 
amplitudes belonging to negative frequencies and multiply the amplitudes of positive 
frequencies by two.” This transformation is known today as adding a Hilbert transform 
($H[s(t)]$) to the real signal $s(t)$, giving $z(t) = s(t) + jH[s(t)]$. The complex signal is known 
as “analytic signal”, which is constructed from the real signal $s(t)$. 
So, we come to the conclusion that in order to know an IF of a complex phasor there is 
actually no need for Fourier analysis, except for using it as a step in Hilbert transform?! 
Not quite so, and the reason for that is the difficulty in applying the notion of IF to multi-
component signals or signals that have a fast varying amplitude (such as the case with tremolo or fast attacks and dumping of notes).

To see that IF is poorly defined for multi-component signals one may look at the phasor plot above. When more the one sinusoidal signal sum together to create the signal that we are trying to analyze, the resulting analytic signal (after applying Hilbert transform to the sum of sine waves and adding this as a conjugate to the original sum signal) can still be mathematically represented as a time varying amplitude and time varying phase 

\[ z(t) = a(t)e^{-i\omega(t)} \]

Let us assume that the amplitude is constant. Even in this case, we might get a phase that has a very complex time behavior – it could accelerate and decelerate in time and even stop or reverse its direction, depending on the way the phases of its two constituent sinusoids (or actually their analytic counterpart, i.e. phasors) behave. For instance, considering a signal comprising of two close sinusoids close in frequency

\[ z(t) = a_1e^{-j\phi_1(t)} + a_2e^{-j\phi_2(t)} \]

with \( \phi_1 = (\omega_0 - \delta)t \) and \( \phi_2 = (\omega_0 + \delta)t \), it can be shown that the instantaneous frequency of the signal equals

\[ f_i(t) = f_0 + \delta \frac{a_1^2 - a_2^2}{a_1^2 + a_2^2 + 2a_1a_2 \cos(\delta t)} \],

which is either ascending or descending, depending on the relative amplitudes of the two components.

So, in order to perform meaningful analysis of a multi-component signal, we need to find a way to separate it into single component sinusoids. Let us look at a typical scheme of phase vocoder (pvoc) analysis:

![Phase Vocoder Diagram]

This analysis in fact performs a combination of Hilbert and bandpass analysis around frequency \( f \). This view is the so-called Filter Bank (FB) view of phase vocoder: the signal is analyzed by passing it through a bank of filters, each having a small aperture around some frequency \( f \), i.e. using a bandpass filter around \( f \). Instead of making multiple filters, the FB in the picture above uses a heterodyning method: by multiplying a signal by sine and cosine functions (or complex exponential) of frequency \( f \), the spectrum of the signal is shifted down by \( f \), i.e. now the components around frequency \( f \) turns to be around frequency 0. Then the bandpass is realized by a lowpass filter around frequency 0. In order to complete our analysis and show how FB performs IF analysis, we will introduce the notion of short time Fourier transform (STFT).
STFT of a signal is defined as

$$S(\tau, \omega) = \int s(t) w(t - \tau) e^{-i\omega t} dt,$$

where $w(t)$ is a windowing function that “localizes” the signal $s(t)$ in time around time instance $\tau$ (the “localization” actually means that the signal after multiplication by the window is still of infinite duration, but all signal values outside some finite duration window around instance $\tau$ are zero). The STFT integral can be viewed as an expression for convolution between the signal $s(t)$ multiplied by cosine and sine components (real and imaginary) $e^{i\omega t}$ and a filter impulse response $w(-t)$, which is exactly the pvoc scheme describe above. One might notice that the windowing function is written here in a time-inverted manner. This should be considered as a time inverted notation, which actually makes no difference if the window is symmetrical in time, a case common in most analysis methods.

A mathematically precise version of the FB is actually slightly different then the pvoc or STFT scheme. As the name implies, this representation describes a multi-component signal as a bank of filter responses. This is different from the pvoc scheme, where the heterodyning “trick” brings all bands to be around zero frequency. In order to bring the bands back to their original frequency location, we need to perform an inverse heterodyning trick, defining FB as $FB(\tau, \omega) = e^{i\omega \tau} S(\tau, \omega)$.

This allows expressing $s(t)$ as a linear combination of basis functions $g(t, \omega) = w(t) e^{i\omega t}$:

$$s(t) = \frac{1}{2\pi} \frac{1}{E_w} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} S(\tau, \omega) g(\tau-t, \omega) d\tau d\omega,$$

with normalization (energy) constant $E_w = \int w^2(t) dt$. IF at the point $(\tau, \omega)$ is defined as

$$\omega(\tau, \omega) = \frac{\partial}{\partial \tau} \arg[F(\tau, \omega)].$$

Writing it explicitly in term of complex components $FB(\tau, \omega) = X(\tau, \omega) + jY(\tau, \omega)$ gives an expression for IF as

$$\omega(\tau, \omega) = \frac{X \dot{Y} - \dot{X} Y}{X^2 + Y^2}.$$

Where the $\dot{X}, \dot{Y}$ denote time derivatives of $X$ and $Y$, respectively.

Discussion:

Some confusion arises at this point due to the fact that the new, FB-based, IF concept depends on both time $\tau$ and frequency $\omega$. Equivalently, we might notice that in our previous discussion IF was obtain from one Fourier and one inverse Fourier transform over the complete duration of a signal, while here there are multiple (actually a continuum of) Fourier Transforms done at times $\tau$, with no inverse Fourier transform.

Well, indeed these notions ARE different, and understanding the correspondence between them is crucial to understanding the assumptions and limitations of pvoc or FB approaches.

In the first part of these notes we introduced a notion of instantaneous frequency as an instantaneous angular speed of a phasor with a time-changing phase. In order to “read out” the phase from the real signal, we had to convert it into an analytical signal, which is a complex signal in time that has only positive frequencies. In the new representation, we
consider a set of signals that result from filtering an input signal with a bank of complex filters (filters that have a complex impulse response). The output signals \( FB(\tau, \omega) \) (and also \( S(\tau, \omega) \)) are complex, and each can be represented now as a complex phasor with instantaneous amplitude and phase. In this representation we actually consider \( FB(\tau, \omega) \) as functions of time for different parameters \( \omega \). Using a slightly different notation \( FB(\tau, \omega) \rightarrow FB_\omega(\tau) \) might help express this “complex signal in time with analysis parameter \( \omega \)” view. So, strictly speaking, we are not measuring the frequencies of the input signal, but the frequencies of the filter bank outputs. They are already analytic, so no Hilbert transform is needed and all we do is simply apply the notion of IF to these filtered signals, directly. In order to relate this to the original signal, let us consider how FB analysis performs in case of two specific cases of periodic and quasi periodic signals.

Examples:
Let us consider a periodic signal \( s(t) = A \cos(\omega_0 t + \phi) \). Rewriting this as a combination of complex phasors gives \( s(t) = \frac{A}{2} [e^{-j(\omega_0 t + \phi)} + e^{+j(\omega_0 t + \phi)}] \). Let us assume moreover that there is a filter of the FB at frequency \( \omega \) close enough to \( \omega_0 \) so that it falls in the pass band of the lowpass \( w(t) \). Let us denote by \( W(\omega) = \Im[w(t)] \) the Fourier transform of \( w(t) \).

\[
S(\tau, \omega) = \int \left[ \frac{A}{2} [e^{-j(\omega_0 t + \phi)} + e^{+j(\omega_0 t + \phi)}] w(t - \tau)e^{-j\omega t} dt \right.
\]
\[
= \frac{A}{2} \left[ e^{-j(\omega_0 + \omega)\tau} + e^{j(\omega_0 - \omega)\tau} \right] W(\omega_0 + \omega) + e^{j\phi} e^{-j(\omega_0 - \omega)\tau} W(\omega_0 - \omega) \]
\]
\[
\approx \frac{A}{2} e^{j\phi} e^{j(\omega_0 - \omega)\tau} W(\omega_0 - \omega)
\]

The left component disappears since the filter is a lowpass, so that it does not let a frequency as high as \( \omega_0 + \omega \) to go through. (More precise expression for \( \omega_0 \) being close to \( \omega \) is that \( \omega_0 \) falls within the bandwidth \( \text{BW} \) of \( w(t) \), \( |\omega_0 - \omega| \approx \text{BW} \).

So, a pure sine passes through the FB with amplitude distorted by factor \( \frac{1}{2} |W(\omega_0 - \omega)| \), and phase equal to \( \varphi(\tau) = (\omega_0 - \omega)\tau + \phi \). Thus the instantaneous frequency in this representation (analysis) is \( \omega(\tau) = (\omega_0 - \omega) \), i.e. a constant equal to the deviation of the sine wave frequency from the center band of the analysis filter. We should note also that the output of the FB is the analytical signal corresponding to the original sine wave (up to the lowpass filter amplitude correction), which is the STFT after demodulation (i.e. adding \( \omega \) in frequency)

\[
FB_\omega(\tau) = e^{j\omega \tau} S(\omega, \tau) \approx \frac{A}{2} e^{j(\omega_0 \tau + \phi)} W(\omega_0 - \omega)
\]
In the quasi-periodic case we assume that the frequency $\omega_h(t)$ is time varying but still “close enough” to $\omega$ for all times. Written in “IF way”, the signal becomes

$$s(t) = A \cos\left(\int_0^t \omega_h(\tau)d\tau + \phi\right).$$

Repeating the above analysis yields time varying instantaneous frequency $\omega_i(\tau) = \omega_h(\tau) - \omega$, and the FB output at this channel is

$$FB_\omega(\tau) \approx \frac{A}{2} e^{i\int \omega_h(\tau)d\tau + \phi} W(\omega_h - \omega),$$

corresponding exactly to the analytical signal with notion of IF that we were trying to develop from the beginning.

Then, it is important to note, that if the signal is either

- NOT bandlimited (i.e. it moves outside the analysis filter)
- More then ONE component fall inside one filter, or
- The amplitude envelope $A(t)$ is FAST varying

the FB and STFT analysis “loose” their precise meaning, although in principle, analysis and resynthesis are still be possible. We shall discuss that in the notes about the so-called “overlap add” view of the pvoc.