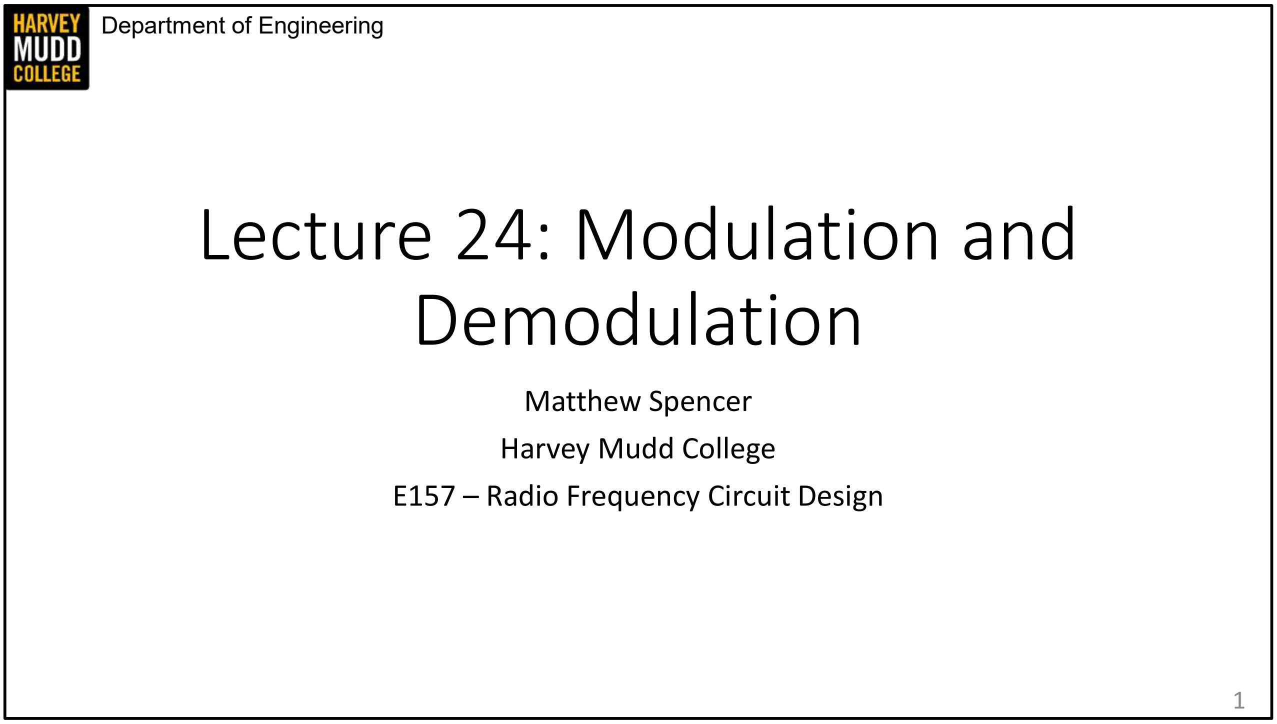
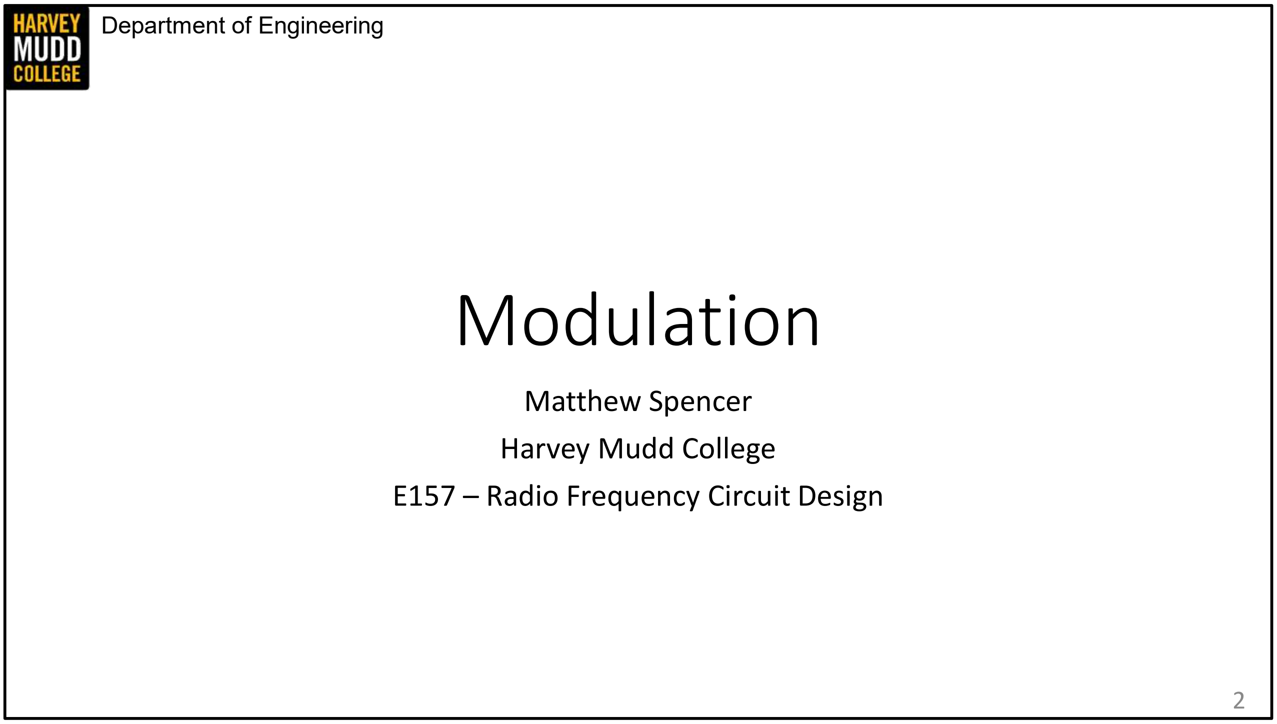


Lecture 24: Modulation and Demodulation

Matthew Spencer

Harvey Mudd College

E157 – Radio Frequency Circuit Design



Department of Engineering

Modulation

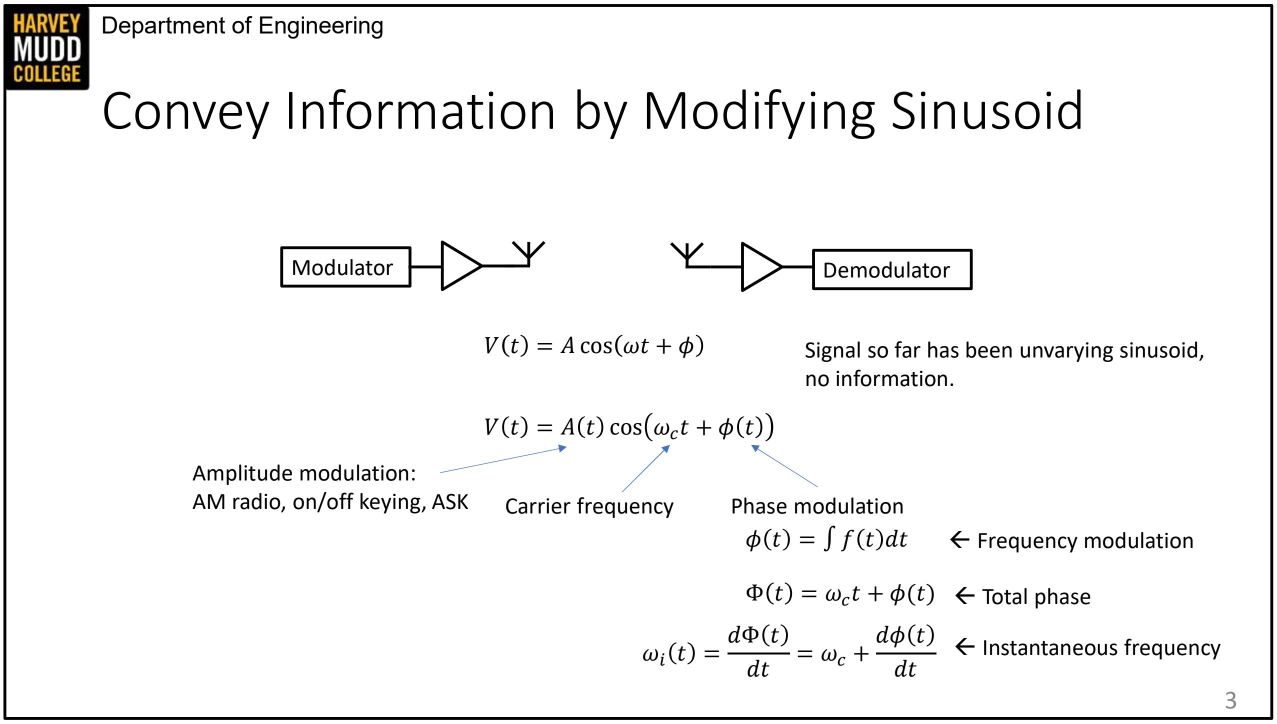
Matthew Spencer

Harvey Mudd College

E157 – Radio Frequency Circuit Design

2

In this video we are going to talk about how to encode information in sinusoidal signals using a process called modulation.



Department of Engineering

Convey Information by Modifying Sinusoid

$V(t) = A\cos(\omega t + \phi)$

Signal so far has been unvarying sinusoid, no information.

$V(t) = A(t)\cos(\omega_c t + \phi(t))$

Amplitude modulation:
AM radio, on/off keying, ASK

Carrier frequency

Phase modulation

$\phi(t) = \int f(t)dt$ ← Frequency modulation

$\Phi(t) = \omega_c t + \phi(t)$ ← Total phase

$\omega_i(t) = \frac{d\Phi(t)}{dt} = \omega_c + \frac{d\phi(t)}{dt}$ ← Instantaneous frequency

3

We need to do this because we've assumed that all of our signals up until this point have just been sinusoids. Sinusoids are handy because understanding how sinusoids move in our system and the Fourier Transform lets us analyze arbitrary signals. However, any individual sinusoid doesn't contain much information. If you know it's value at one time, you know how it's going to behave at all times, so a transmitter sending this over doesn't give us any information about what's happening on the transmitter end of this link.

CLICK We fix this by making small changes to the sinusoid that carry information. We have to build circuits to make these changes, which are called modulators, and circuits to detect them, which are called demodulators. This equation shows a few features of the sinusoid we can modulate: the amplitude and the phase.

CLICK Before we start talking about the modulation itself, it's worth emphasizing that modulation is generally small changes to an underlying sinusoid, which is called a carrier. This is an important fact because it means the signals we produce by modulating a sinusoid are still pretty close to the original sinusoid. As a result, many of the assumptions we've made about RF systems, particularly ones that required a narrow bandwidth, still hold reasonably well for modulated signals. For instance, antennas are often narrow band, but a modulated signal can still make it through antennas as long as the fractional bandwidth of the modulating signal isn't too large.

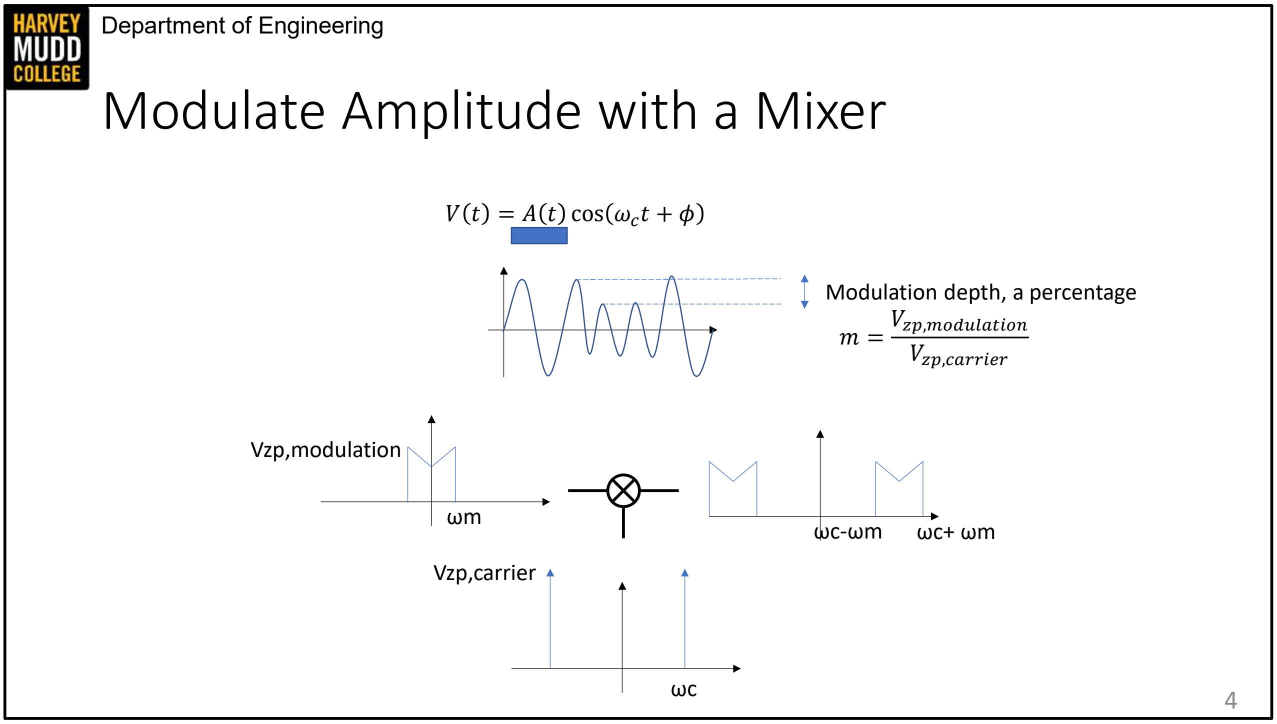
The carrier is often a fairly high frequency, which has the advantage of making the antennas that interface with it smaller. However, high frequencies make receiver deisgn more challenging, which might be a decent tagline for this whole class. We modulate a high frequency signal to make it easier to launch our signal off of an antenna, then we demodulate in order to build easier circuits for the relatively low bandwidth modulating signal.

CLICK Changing the coefficient in front of a sine wave is called amplitude modulation. This is the simplest form of modulation, and it is how AM radios work. You could argue that this type of modulation had an even earlier life in Morse code, which is a form of digital amplitude modulation – instead of sending a continuous amplitude modulated signal, you send bits that are indicated by the presence or absence of a high frequency signal. Every type of modulation comes in both analog and digital varieties, and digital versions of amplitude modulation include on/off-keying or OOK, and amplitude shift keying or ASK.

CLICK Changing the phase of a sinusoid is called phase modulation or angle modulation. It’s popular because phase modulation rejects amplitude noise gracefully and you spend more power on your information relative to your carrier than in amplitude modulation, but phase tends to be much more complex to modulate and demodulate than amplitude.

CLICK Frequency modulation is one extremely popular subset of phase modulation. In frequency modulation, we imagine the value of the phi(t) modulating our sinusoid is the integral of a frequency signal f(t).

CLICK One other useful paradigm for analyzing phase modulation and sinusoid behavior in general is to define the argument of the sinusoid here as the total phase. You can then define an instantaneous frequency as the derivative of your total phase. This is useful when your frequency is moving around as in FM or PM communication schemes.



Department of Engineering

Modulate Amplitude with a Mixer

$$V(t) = A(t)\cos(\omega_c t + \phi)$$

Modulation depth, a percentage

$$m = \frac{V_{zp,modulation}}{V_{zp,carrier}}$$

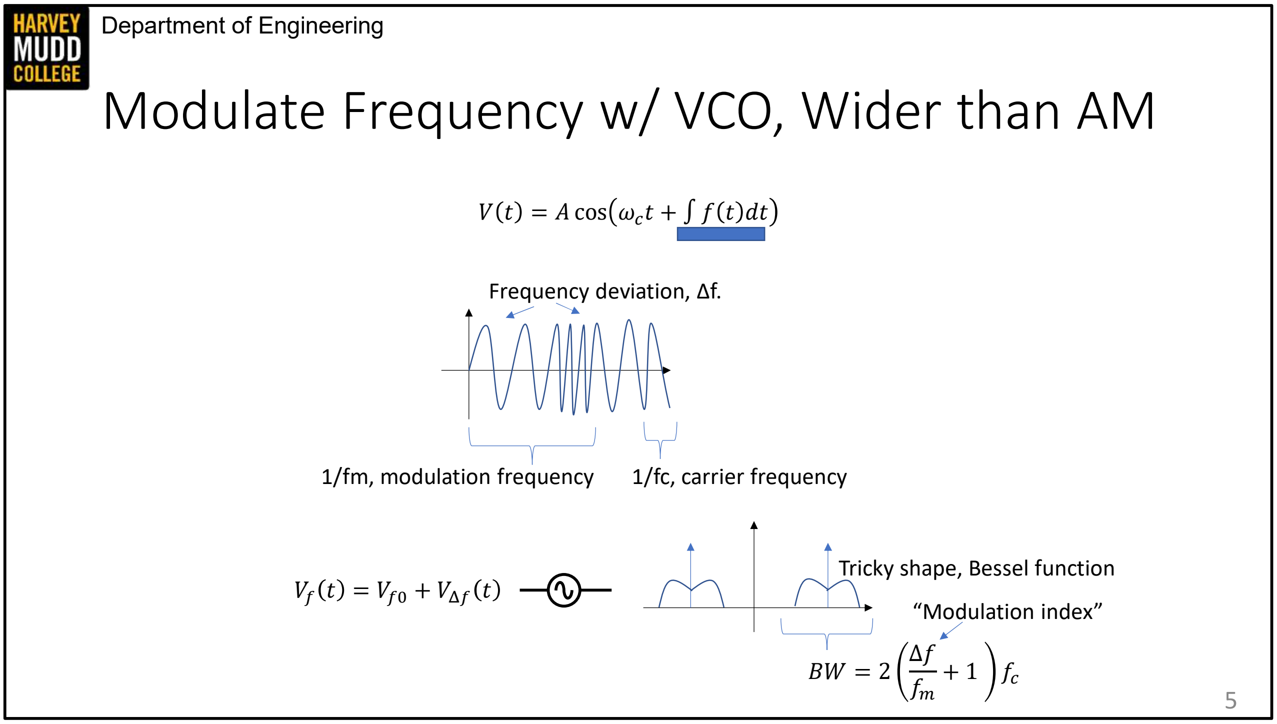
This slide contains a picture of an amplitude modulated wave and a circuit that implements amplitude modulation.

Amplitude modulated waves look about how you would expect: the size of the wave gets bigger and smaller while the frequency stays the same. This picture represents digital amplitude modulation sending the bit pattern of 101. This kind of amplitude modulation is often called amplitude shift keying, where zeros are indicated by one amplitude and ones by another. It's also possible to encode more bits by having more amplitude levels or to modulate amplitude continuously.

I've called out a feature of amplitude modulated waves called the modulation depth. This measures how dramatically the carrier signal is being changed by the modulation. If the amplitude of the carrier gets reduced to zero when it is being modulated, then you would say the modulation depth is 100%. As shown in this figure, it's about 50%. Modulation depth puts demands on the receive chain because small signals are susceptible to noise and large signals can cause distortion.

Amplitude modulation is relatively easy to implement with circuits we already know about. We need to multiply a carrier signal by some other signal, and mixers already explicitly implement multiplication. So here I've shown a modulating signal getting mixed with a carrier on the LO port of an amplifier. The modulating signal is also often referred to as the base band signal.

The output spectrum of the mixer is no surprise: multiplication in the time domain is convolution in the frequency domain, so we get two copies of the modulating signal, one at each of the peaks of the carrier. Amplitude modulation is often said to be bandwidth inefficient because the signal that appears at each of the carrier peaks has twice the baseband bandwidth: if you consider only positive spectrum, the low baseband signal has a bandwidth of omega_m, while the RF signal has a bandwidth of 2*omega_m. Single-side-band modulators can fix that issue.



Department of Engineering

Modulate Frequency w/ VCO, Wider than AM

$$V(t) = A\cos\left(\omega_c t + \int f(t)dt\right)$$

Frequency deviation, Δf.

1/fm, modulation frequency

1/fc, carrier frequency

$$V_f(t) = V_{f0} + V_{\Delta f}(t)$$

Tricky shape, Bessel function

“Modulation index”

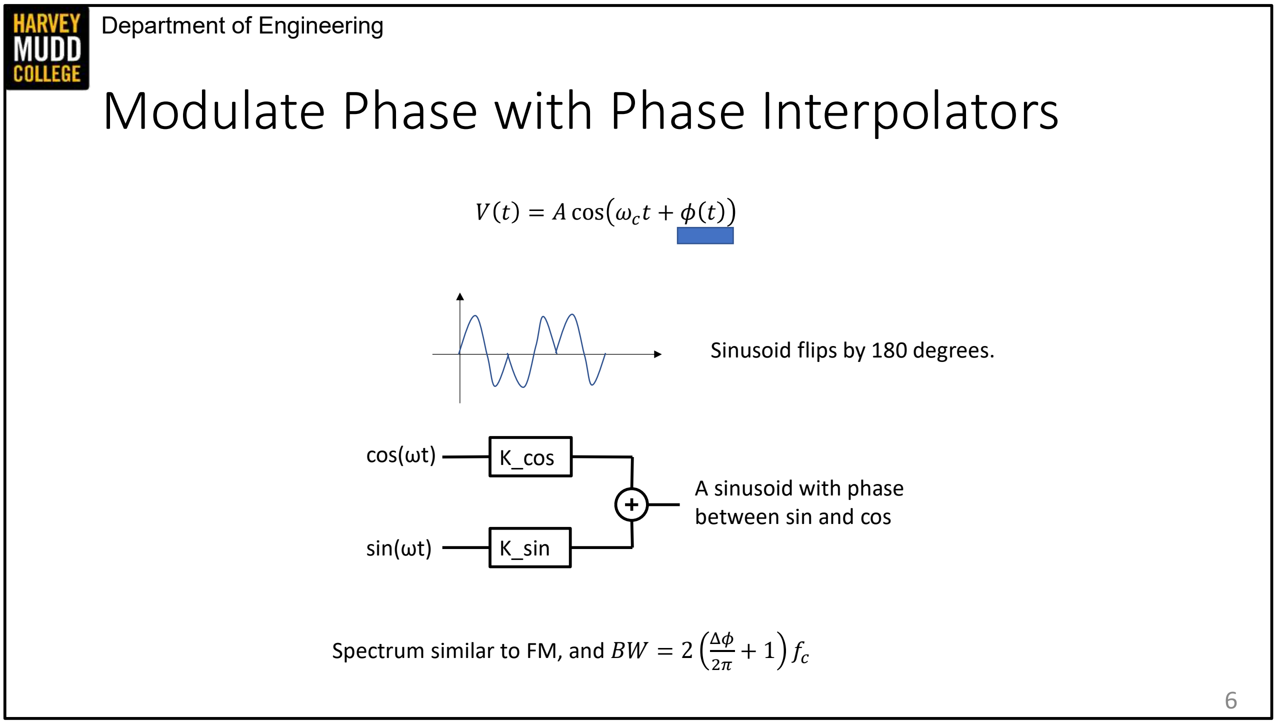
$$BW = 2\left(\frac{\Delta f}{f_m} + 1\right)f_c$$

This slide contains a picture of a frequency modulated wave and a circuit that implements frequency modulation.

Frequency modulated waves alternate between high and low frequencies depending on the modulating frequency, and I’ve tried to capture this in the graph I’ve drawn on this page. You need to be careful describing frequency modulation because there are a few different frequencies involved. The carrier frequency is one that we’re already acquainted with: it’s the frequency of an underlying sinusoidal signal that our modulator is messing with. The frequency deviation measures how much our frequency changes, so it’s the maximum possible frequency minus the carrier frequency (or the minimum signal minus the carrier if the frequency deviation is asymmetric).

As on the last slide, I’ve chosen to show a digital frequency modulation here called frequency shift keying or FSK, and I’m showing the pattern 101. In FSK, 1 bits are indicated by one frequency while zero bits are indicated by another. Continuous frequency modulation is also possible and common.

You can implement frequency modulation with voltage controlled oscillators. The DC voltage on the oscillator control terminal will set your carrier frequency, and variations from the DC voltage will set the modulation depth and frequency. The output spectrum produced by frequency modulation is not as straightforward as an amplitude modulated spectrum, and Bessel functions are frequently invoked when describing it. However, we mostly care about the total bandwidth of a signal when designing a transmitter or a receiver, and there is a simple expression for it shown below the output spectrum. The bandwidth increases as we modulate our signal more dramatically, which is captured by the dimensionless modulation index. The modulation index is usually in the range of 1-5 for FM signals.



Department of Engineering

Modulate Phase with Phase Interpolators

$$V(t) = A\cos\left(\omega_c t + \phi(t)\right)$$

Sinusoid flips by 180 degrees.

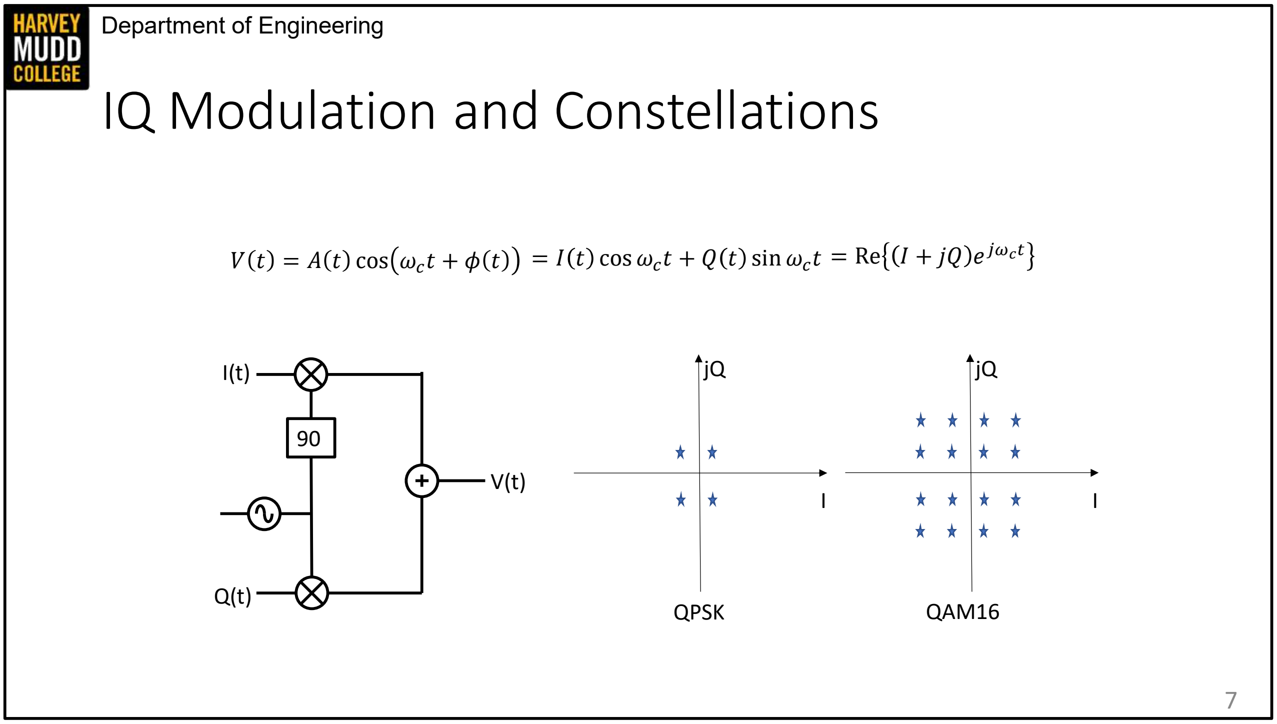
A sinusoid with phase between sin and cos

Spectrum similar to FM, and $BW = 2\left(\frac{\Delta\phi}{2\pi} + 1\right) f_c$

This slide contains a picture of a phase modulated wave and a circuit that implements phase modulation.

Phase modulated signals are often difficult to interpret, so I've included a simple modulation called binary phase shift keying or BPSK, which swaps between 0 degrees and 180 degress of phase in the modulating signal. This is a type of digital phase modulation, and this wave represents the code 101 by flipped from 0 phase in the first period to 180 degrees of phase in the second period and back to 0 degrees in the third period. The sharp transitions when phase flips by 180 degrees are particularly challenging for amplifiers and filters to render faithfully without smoothing out the signal.

You can make phase modulated signals in a few ways, but one that's within the scope of this class is called the phase interpolator. In a phase interpolator, cosine and sine signals are weighted by coefficients and added together, which results in a signal that has a phase somewhere between cosine and sine. The spectrum winds up looking similar to frequency modulated spectrums, and the total bandwidth varies with the amount of modulation, just like in the frequency modulated case.



Department of Engineering

IQ Modulation and Constellations

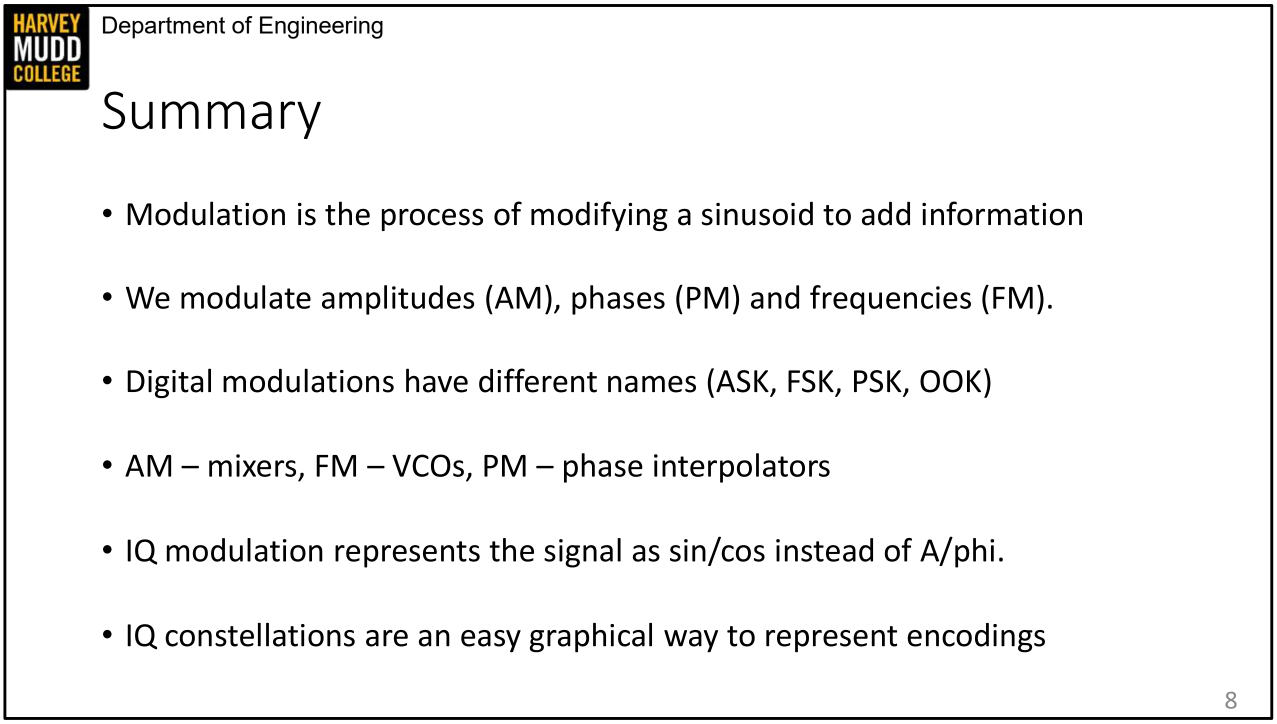
$$V(t) = A(t)\cos\big(\omega_c t + \phi(t)\big) = I(t)\cos\omega_c t + Q(t)\sin\omega_c t = \mathrm{Re}\{(I + jQ)e^{j\omega_c t}\}$$

7

One last type of modulation is worth mentioning. We tend to think of modulation in terms of changes to amplitude and phase, but it's possible express a modulated sinusoid as the weighted sum of a sine and a cosine wave, which are often called the in-phase and quadrature components of a signal. Accordingly, we can represent modulation using the coefficients of sine and cosine, which I've called I(t) and Q(t) instead of the amplitude and phase, A(t) and phi(t). We can take this further and make an analytical representation of I and Q which looks like a complex coefficient multiplied by a complex exponential.

This is a useful representation because it's easy to imagine how to build and IQ modulator, and I've shown how in the lower left of this slide. Note that the signal addition in this block diagram is probably accomplished using a power combiner, which is literally the power splitters we've discussed earlier in class plugged in backwards. Most modern cell phones are built using IQ modulators that look like more complex versions of this.

Further, plotting the I and Q values that a modulation scheme considers to be different codes provides a convenient way to talk about different types of modulation. This representation of a modulation technique is called an IQ constellation. I've included two common constellations on this slide. QPSK is quadrature phase shift encoding, and it's a type of modulation in which phase is varied between 45, 135 , 225 and 315 degrees. You can interpret those angles by looking at the angles on the IQ plane. Similarly, you can tell this modulation doesn’t use any amplitude modulation because all the codes are at the same distance from the origin. By way of contrast, QAM16, or 16 code quadrature amplitude modulation, uses both amplitude and phase modulation to densely pack possible signal values into IQ space.



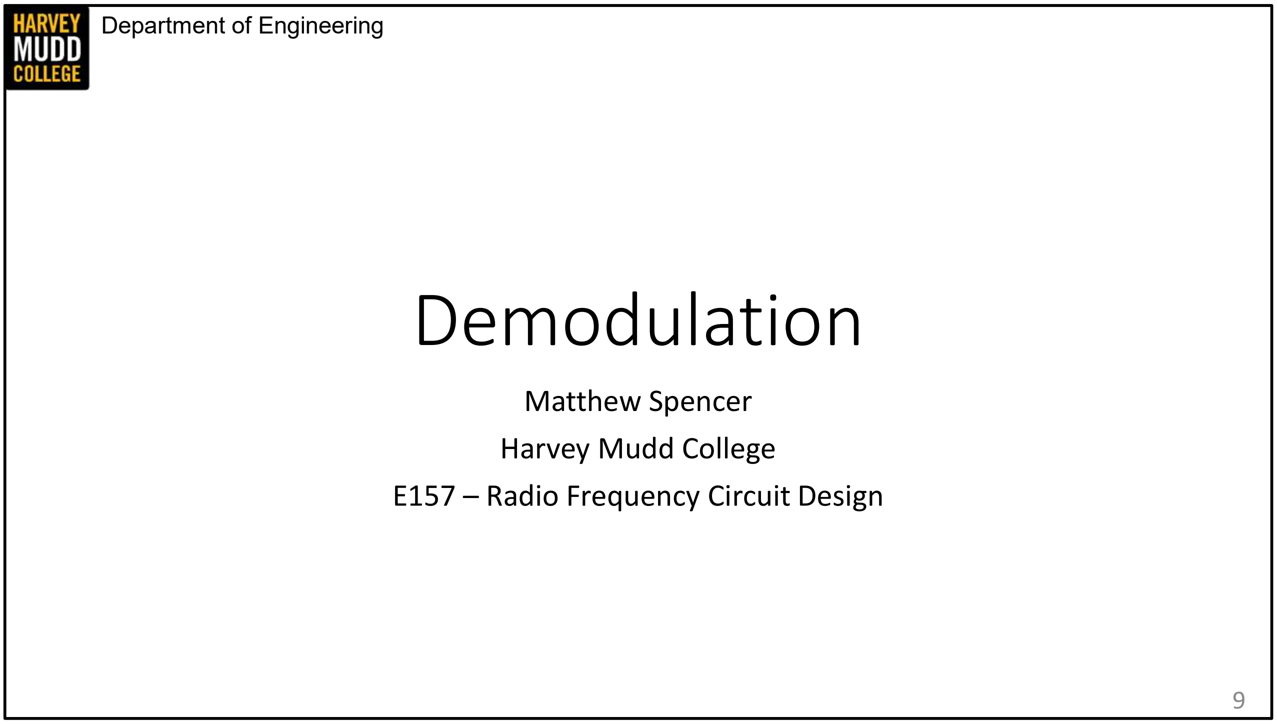
Department of Engineering

Summary

- Modulation is the process of modifying a sinusoid to add information
- We modulate amplitudes (AM), phases (PM) and frequencies (FM).
- Digital modulations have different names (ASK, FSK, PSK, OOK)
- AM – mixers, FM – VCOs, PM – phase interpolators
- IQ modulation represents the signal as sin/cos instead of A/phi.
- IQ constellations are an easy graphical way to represent encodings

8

Also ASK vs. AM, FSK vs. FM, PSK vs. PM



Department of Engineering

Demodulation

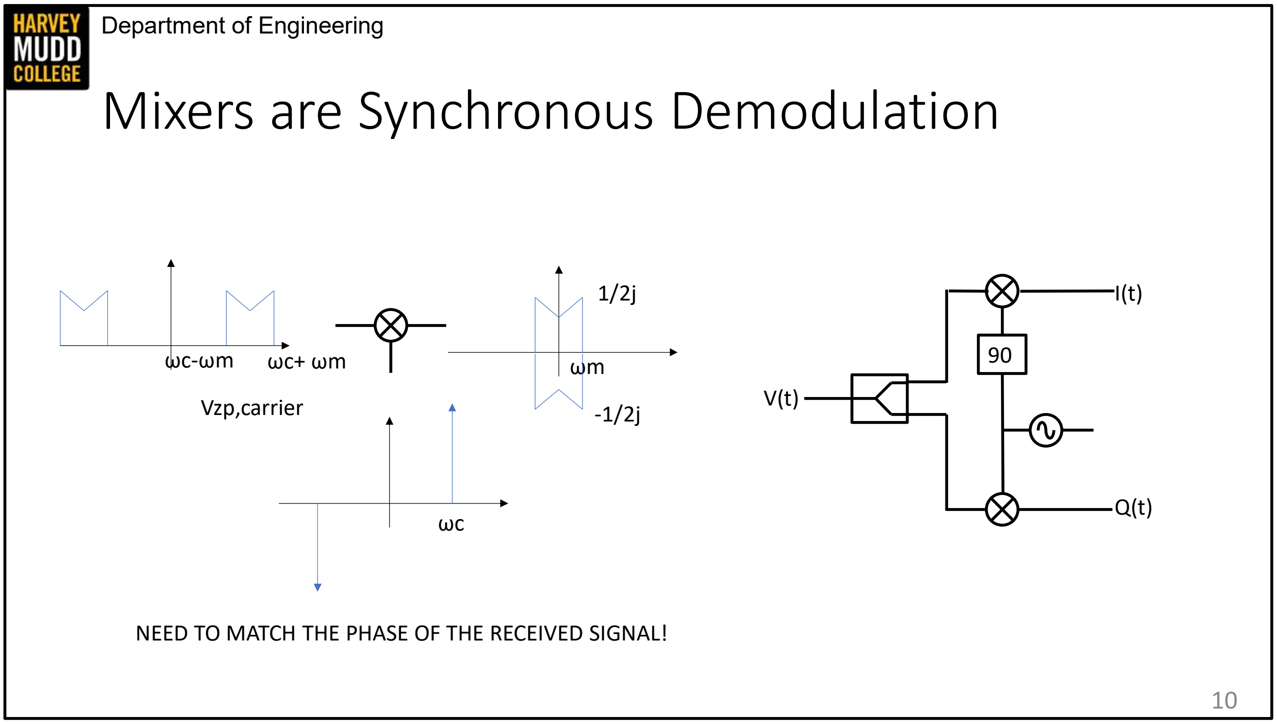
Matthew Spencer

Harvey Mudd College

E157 – Radio Frequency Circuit Design

9

In this video we’re going to discuss how to pick information off of varying sinusoids, a process called demodulation.



Department of Engineering

Mixers are Synchronous Demodulation

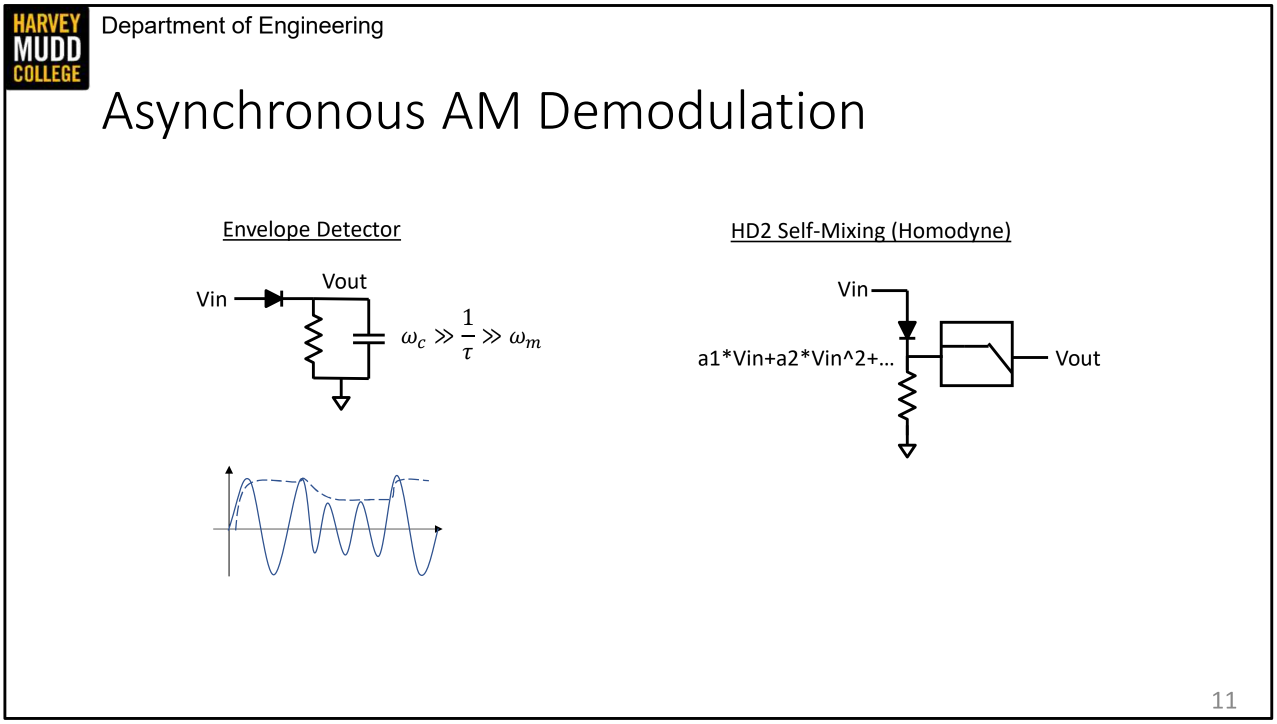
NEED TO MATCH THE PHASE OF THE RECEIVED SIGNAL!

The key issue that we’re trying to solve with demodulation is that our modulated sinusoid is at a high frequency, which makes amplifying and measuring it difficult. We already know that mixers allow us to translate the frequency of a signal, so can’t we just mix high frequency signals down to low frequencies and call this finished?

Unfortunately, no. That’s because we depend on the local oscillator being a in phase with our incident signal in order to mix the incident signal to baseband. For instance, in the picture I’ve drawin on the left, a signal was mixed up to RF using a cosine, and we’re using a sine wave to demodulate it. This results in a positive and a negative copy of the signal landing on top of eachother at baseband and cancelling completely. So we can only use a mixer to convert incoming signals to baseband if we know the phase of the incoming signal. We refer to this kind of demodulation as synchronous demodulation because of the need to be lined up with the incoming signal.

IQ demodulation, which is pictured on the right of this slide, is also synchronous demodulation. If the local oscillator is out of phase with the incoming signal, then some of the incident signals I value will wind up on the Q path, and some of the Q value will wind up on the I path. (As an aside: This results in the IQ constellation rotating, and the ability to analyze IQ imbalances by looking at the graphical behavior of constellations is a nice plus of IQ representations.) The block on the left side of the IQ demodulator is a power splitter.

Recovering the correct phase to use in your synchronous demodulation is an involved problem that requires a specialized frequency feedback loop called a phase-locked loop. Phase lock loops used to align phase like this are often called clock and data recovery circuits or CDRs.



Department of Engineering

Asynchronous AM Demodulation

Envelope Detector

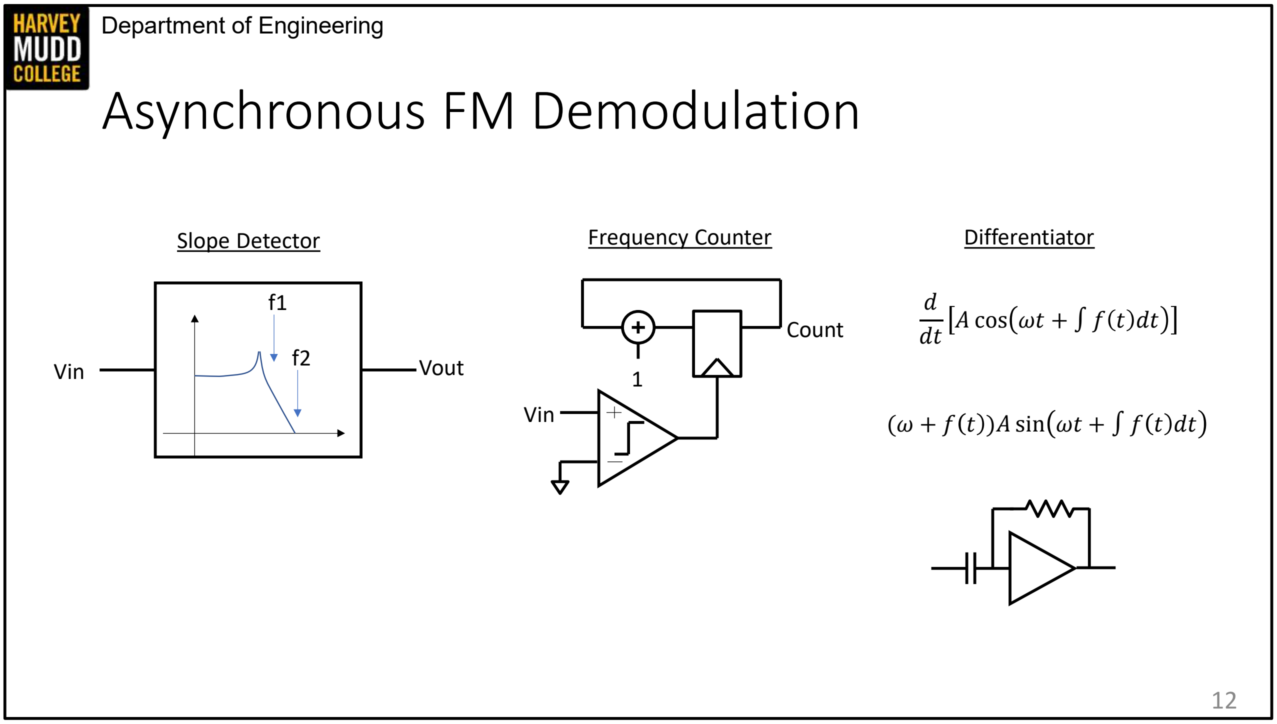
$\omega_c \gg \frac{1}{\tau} \gg \omega_m$

HD2 Self-Mixing (Homodyne)

Synchronous demodulation can provide very good performance, but it's also possible to demodulate signals asynchronously, which can reduce design complexity. Amplitude modulated signals are particularly amenable to asynchronous demodulation, which is one of the reasons they were so popular in early radios.

Envelope detectors explicitly pick the amplitude off of high frequency sine waves. The presence of a modulated signal helps us pick the time constant of the RC filter in the envelope detector: we need it to be slow enough to filter out the carrier frequency, but fast enough that it can respond to the modulation signal. An example of an envelope detector working on an amplitude modulated signal appears below the circuit. The dashed line is the output of the envelope detector.

It's also possible to use second order harmonic distortion to mix a signal back down to DC as pictured on the right of this slide. The second order harmonic distortion has a DC component, and that DC component will vary as the amplitude increases and decreases. This type of demodulation has a few issues – it relies on big signals and it's non-linear – but it's easy to implement. Any time that you mix a singal with itself it is called a homodyne measurement, so this type of demodulator can be called a diode homodyne.



Department of Engineering

Asynchronous FM Demodulation

$$\frac{d}{dt}\left[A\cos\left(\omega t+\int f(t)dt\right)\right]$$

$$(\omega+f(t))A\sin\left(\omega t+\int f(t)dt\right)$$

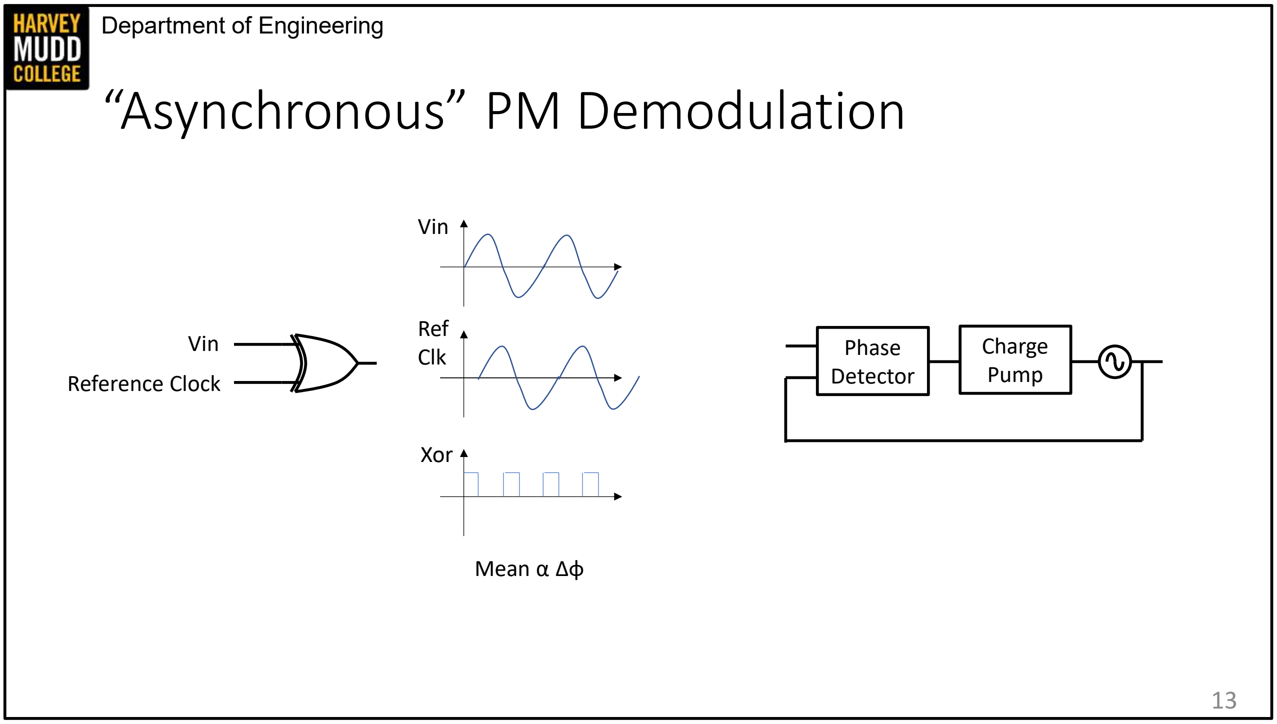
12

Asynchronously demodulating frequency modulation is harder than amplitude modulation, but it’s still doable.

One way to demodulate frequency modulated signal is to put them into a filter that has a sharp frequency response. The different frequencies in your input signal will have different amplitudes at the output because of the slope in the filter, which converts frequency modulation into easier to measure amplitude modulation. This type of measurement is called a slope detector. Various specific implementations of this exist, including one version that uses a center-tapped inductor and differential voltage signals called a Foster-Seeley detector.

Another option is using digital circuits to directly measure the frequency. The frequency counter circuit in the middle uses a comparator to convert the input signal into a digital signal that is then used to clock a flip-flop. The flip-flop is configured as an accumulator, so every clock stroke will increase count by one. By looking at the rate of change of clock, the digital system can extract the frequency of the input signal. This relies on very fast digital logic, which is a limitation of the technique.

Finally, one common piece of advice when demodulating FM is to just take a derivative of the signal. Doing so results in the frequency modulating signal appearing as a coefficient of the sinusoide. However, building a fast analog differentiator is quite difficult, so this demodulation technique is more often used in digital signal processing once the signal has been sampled.

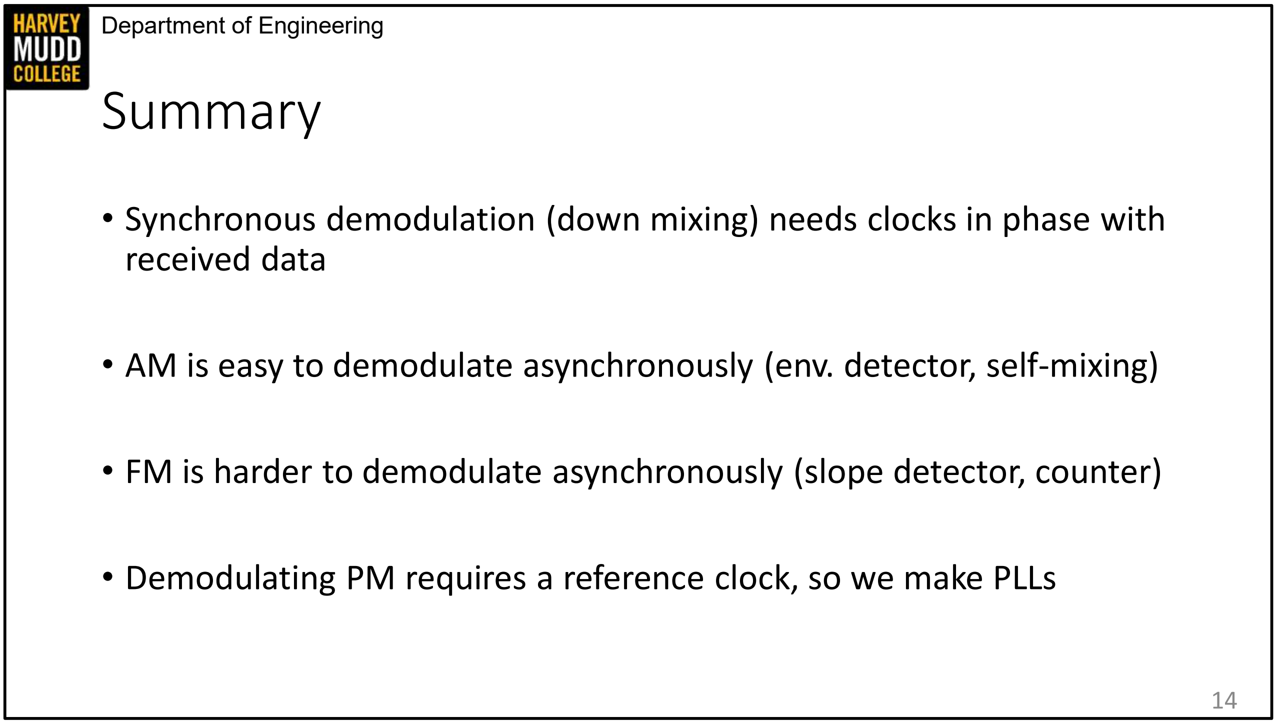


Department of Engineering

“Asynchronous” PM Demodulation

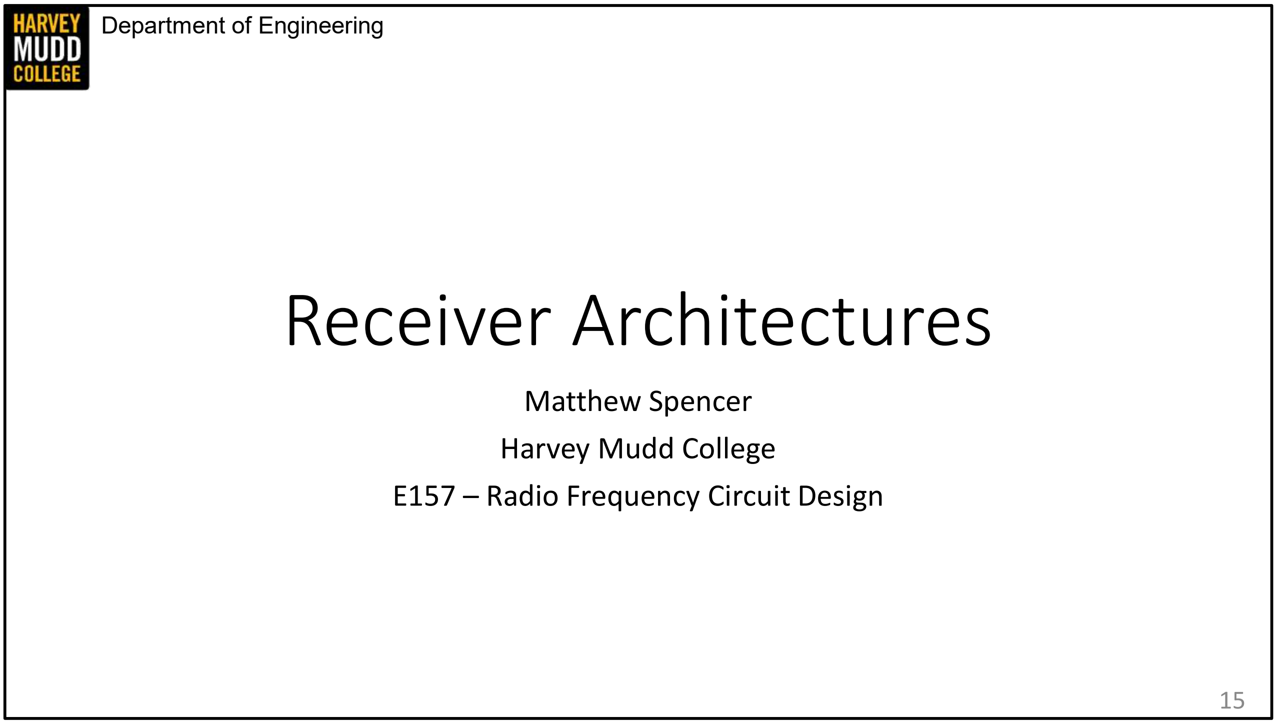
Demodulating phase asynchronously is impossible because phase is always measured relative to some kind of time reference, however it is possible to build circuits that detect the phase difference between a signal and a reference. A simple XOR does this job pretty well. When the sign of the input signal is different than the sign of the reference clock the XOR will be high, which means the average voltage at the output of the XOR will be proportional to the phase difference between Vin and the reference clock.

If you want to provide the reference clock yourself then you can use a phase detector and a phase to voltage converter and filter called a charge pump to wrap a feedback loop around a voltage controlled oscillator. This will align the VCO with the input clock and the control effort applied to the VCO can be used to extract changes in input phase. This type of feedback loop is called a phase lock loop, and there are lots of other things to say about them.



Summary

- Synchronous demodulation (down mixing) needs clocks in phase with received data
- AM is easy to demodulate asynchronously (env. detector, self-mixing)
- FM is harder to demodulate asynchronously (slope detector, counter)
- Demodulating PM requires a reference clock, so we make PLLs



Department of Engineering

Receiver Architectures

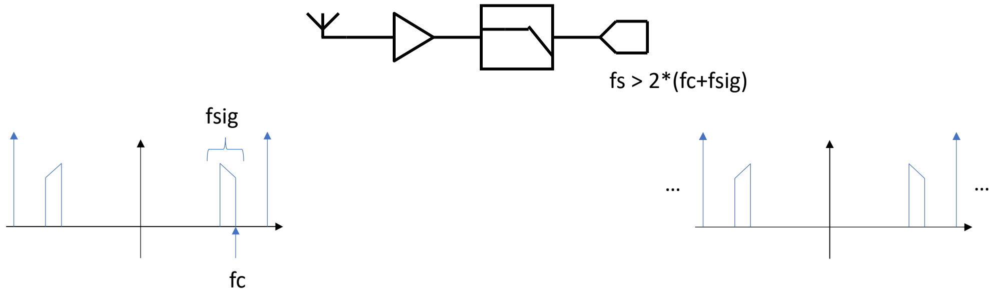
Matthew Spencer

Harvey Mudd College

E157 – Radio Frequency Circuit Design

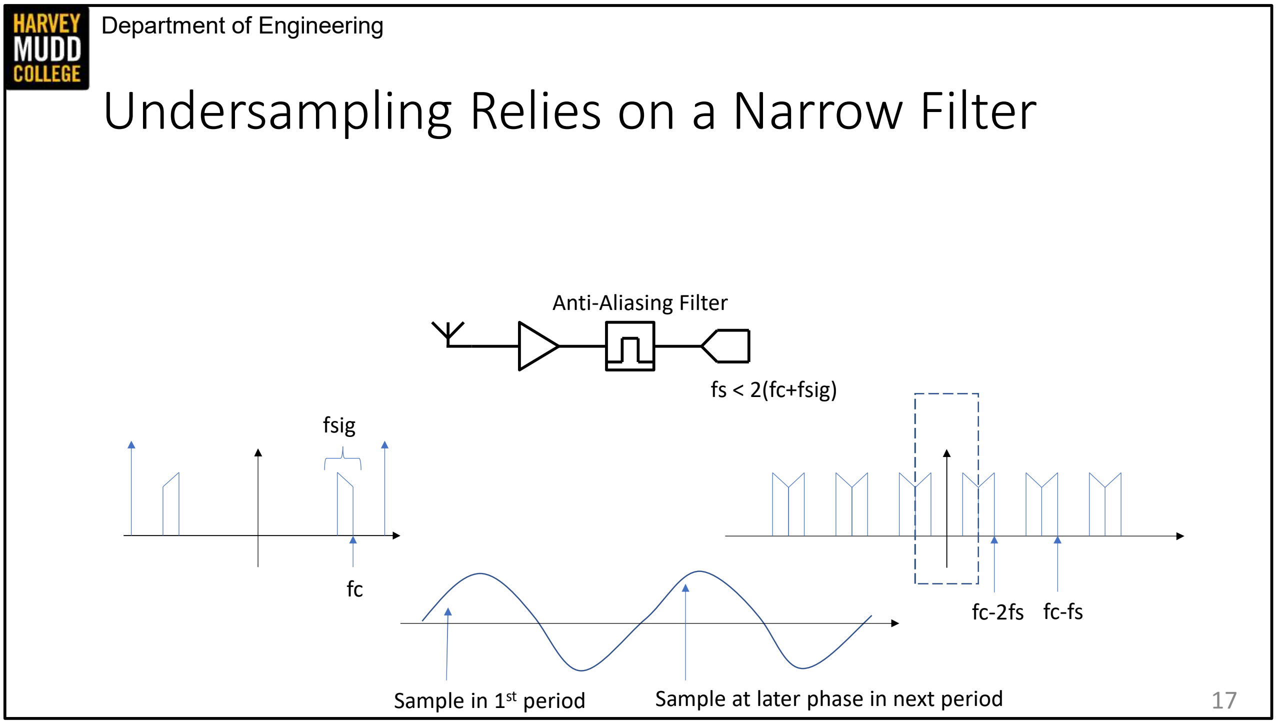
In this video we’re going to discuss several different ways to combine amplification, filtering, modulation and sampling to recover signals.

Direct Sampling Needs to Sample Very Fast



The simplest receiver you could imagine is attaching a sampler directly to an antenna. This type of receiver is called a direct sampling receiver. It requires extremely high sampling frequencies to accurately capture the input signal, especially if the input signal includes high frequency blockers, whiih we're not filtering in this architecture. This type of architecture is also wide-band, which will admit a lot of noise power into the final result. These drawbacks make direct sampling unpopular, but what this architecture lacks in noise performance and power efficiency (because of the expensive ADC) it makes up for in flexibility. You can use digital processing to measure signals anywhere in the sampling bandwidth. That means this architecture sees some use in software defined radios, particularly for lower frequencies.

I've drawn an example input spectrum, which consists of a single side-band and a blocker. Sampling fast enough means the output spectrum will be identical, though there will be higher frequency copies centered on the sampling frequency.



Department of Engineering

Undersampling Relies on a Narrow Filter

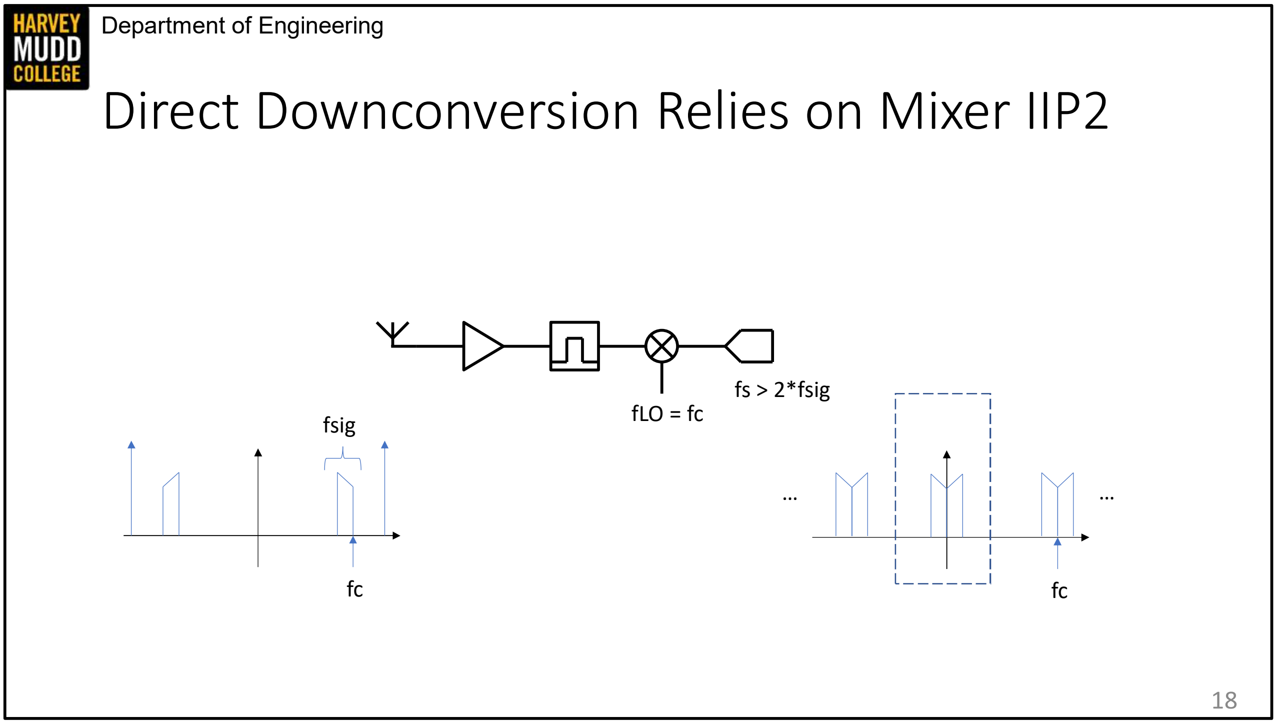
17

An undersampling receiver deliberately attempts to alias a high frequency signal to create a low frequency copy of it. Because we're trying to alias a signal, it's extremely important that the signal be band-limited. This architecture uses a very sharp anti-aliasing filter so that no blockers or other signals get aliased on top of the signal of interest. As a result, we add a narrow band-pass filter to this receiver in place of the low-pass that was used in a direct sampling receiver. This narrow band pass reduces the noise that's admitted to the undersampling receiver.

The output figure in this spectrum shows that the blocker was knocked out by the anti-aliasing filter and that we have many copies of our signal at high and low frequencies. Filtering out our signal from the nearby upper side-bands seems difficult in this example, but small changes in the sampling frequency could make this easier.

This seems kind of magical, but it has a natural time-domain interpretation. When you're undersampling, you're measuring the shape of a wave by measuring different phases in different periods of the wave. So if I capture a sample near the start of a wave I will wait for a whole period plus a little time to measure the next sample of the wave instead of trying to sample quickly and catch it right away.

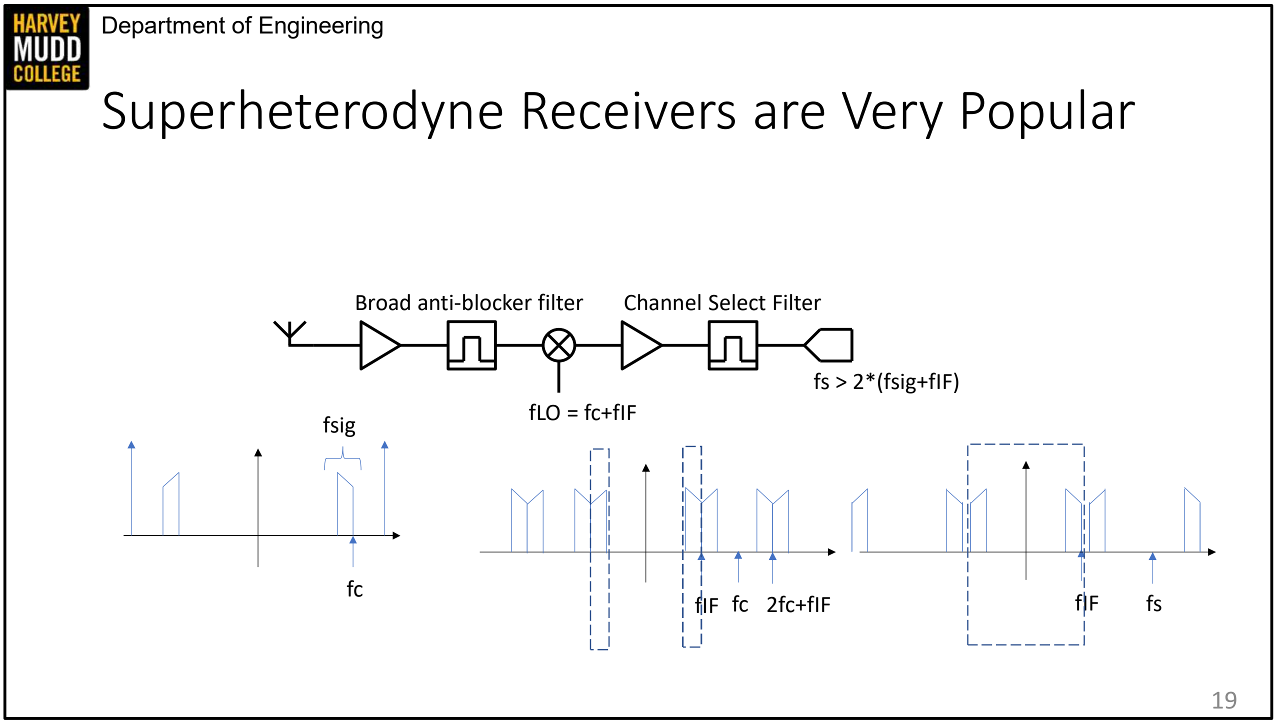
Changing the sampling frequency of this filter nominally allows it to sample other frequencies easily, but the anti-aliasing frequency filter is difficult to make programmable. As a result this receiver is usually tied into receiving one frequency.



Department of Engineering

Direct Downconversion Relies on Mixer IIP2

In a direct downconversion receiver we use a mixer to shift our frequency directly from RF to baseband. These mixers benefit from having both side-bands in the input signal because it doubles the received power. These receivers are very sensitive to IIP2, especially in the mixer, because low frequency distortion terms will make it directly into the output spectrum. These receivers also have limited programmability because the filtering occurs before the mixer. It’s relatively easy to change the mixer LO to look at a different frequency, but the fixed filter means only a small band of signals is available to the mixer.



Department of Engineering

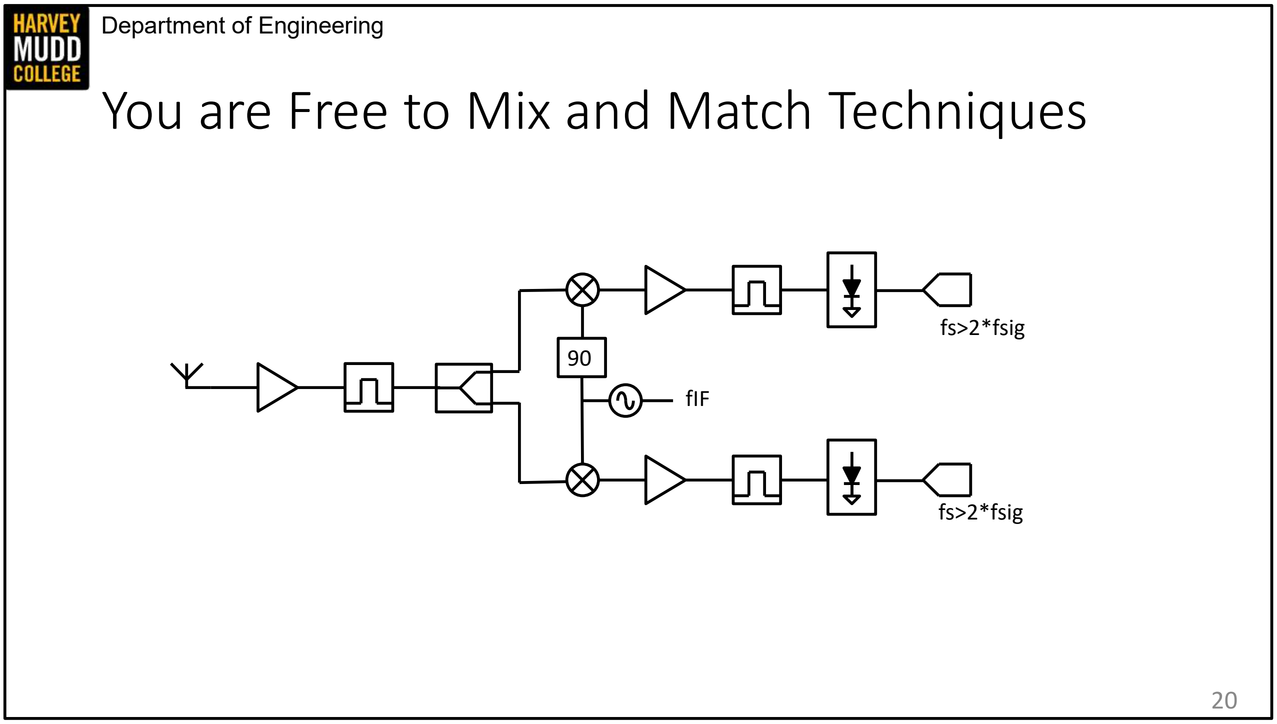
Superheterodyne Receivers are Very Popular

19

Superheterodyne receivers are interesting architectures that split up gain and filtering over multiple frequency bands, which has a number of advantages. The first mixer in system is driven by a frequency that shifts our input to an intermediate frequency or IF. We have one filter at our original RF frequency, but it can be a wide filter. It's main job is to limit high power blockers, but it often admits a whole band of interesting signals.

At the intermediate frequency we have more gain and a very narrow channel select filter. Splitting up gain over two frequencies can make it easier to avoid instability in amplifiers, which is a nice perk, and having a very narrow filter after a mixer gives this receiver some programmability. Changing the value of fLO can result in different signals falling into the channel filter.

I've put a sampler after the IF filter, and it may be possible to directly sample some low-IF systems. However, that sampler is just a short hand for demodulation, and you could use another mixer to shift the frequency to an even lower frequency or, depending on the modulation used, you could use asynchronous demodulation techniques to extract your modulating signal.

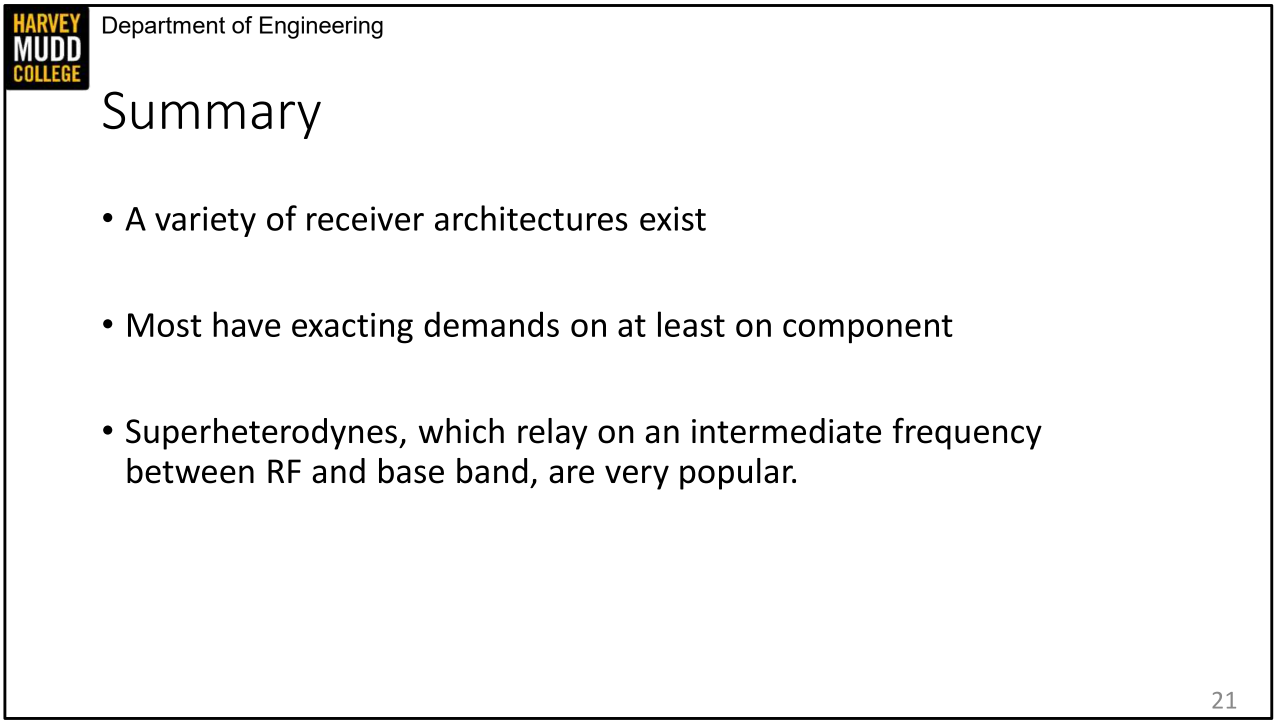


Department of Engineering

You are Free to Mix and Match Techniques

20

One of the joys of designing receivers is that there are an infinite number of ways to mix and match these receiver design techniques. Here I’m showing an IQ receiver that mixes the input signal to an intermediate frequency then uses power detectors to find the amplitude of the I and Q signals. You have lots of interesting design freedom in complex systems like a receiver.

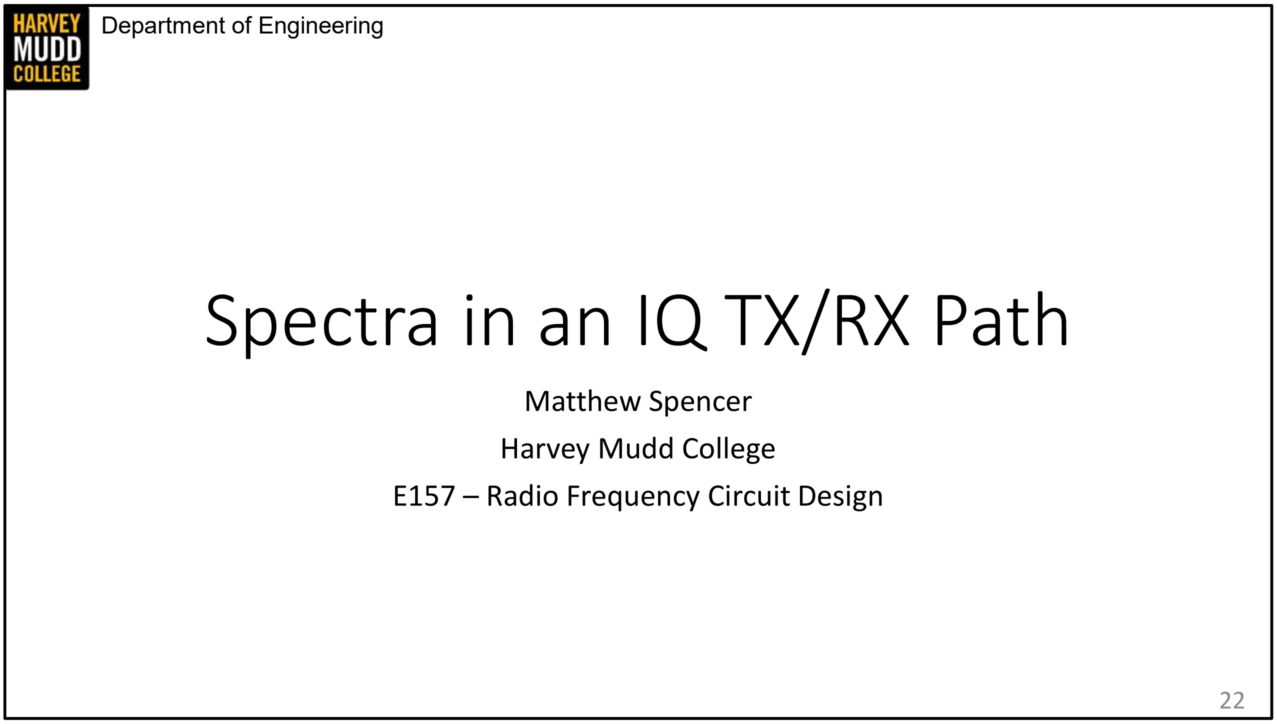


Summary

- A variety of receiver architectures exist

- Most have exacting demands on at least on component

- Superheterodynes, which relay on an intermediate frequency between RF and base band, are very popular.



Department of Engineering

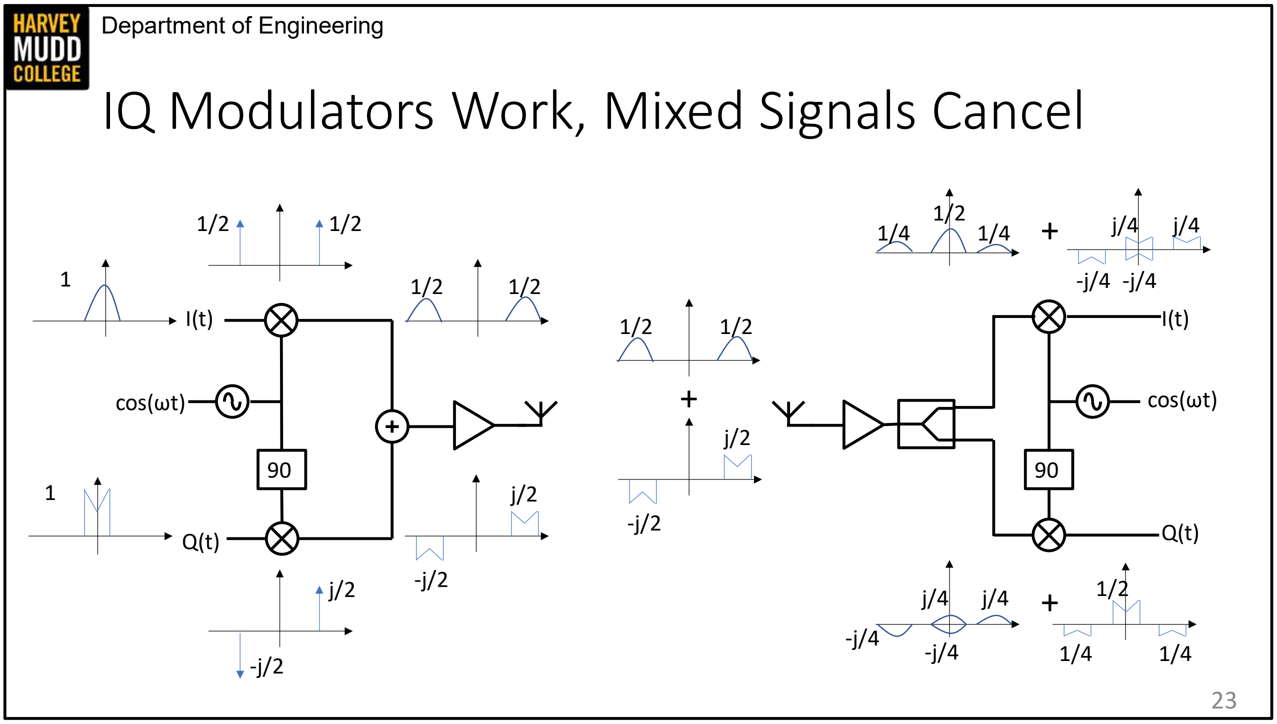
Spectra in an IQ TX/RX Path

Matthew Spencer

Harvey Mudd College

E157 – Radio Frequency Circuit Design

In this video we’re going to look at how signals move through an IQ transmitter and receiver to demystify the behavior of IQ transceivers.



HARVEY MUDD COLLEGE

Department of Engineering

IQ Modulators Work, Mixed Signals Cancel

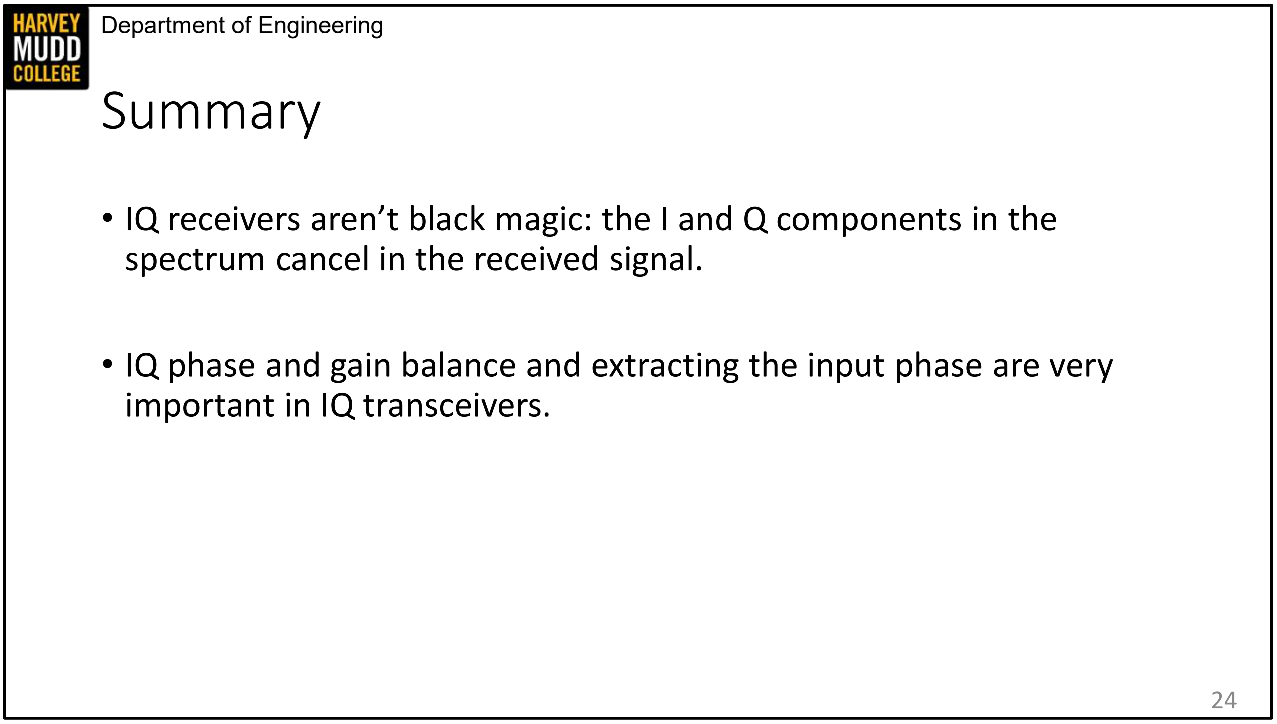
This slide shows an IQ transmitter and an IQ receiver. I've drawn some example I and Q signals on the left, and the LO that they'll be multiplied by appears by the mixers.

CLICK When we carry out the convolution we wind up with slightly different spectra from the I and Q paths. The spectrum on the I path is purely real and it's an even (and positive) spectrum because the I signal was convolved with a cosine. Both signals have been reduced by a factor of 2 relative to the input because the cosine splits its power into the upper and lower impulses. The Q path is purely imaginary and it's an odd spectrum because the input was convolved with a sine.

CLICK These signals are added together in the air, and because the I path is purely real while the Q path is purely imaginary, the paths don't affect one-another. It's often said that the I and the Q signals are orthogonal because of this property. The phase rotator is a very important part of the transmitter to preserve this orthogonality property; if the LO signals aren't exactly in quadrature then I will leak into Q and vice versa. Similarly, the phase match on the I and Q paths of the receiver matters a lot to preserve this orthogonality. When the I and Q paths don't have the same phase, it causes the received constellation to rotate, which naturally causes difficulties for the decoding path.

CLICK Finally, the transmitted signal is received and multiplied by a sine and a cosine again.

This figure is ignoring the delay introduced by the phase delay between the transmitter and the receiver, but you can imagine it’s there by multiplying every term on the receiver by e to the –2*pi*j*r/lambda. The result of multiplying the received signal by a cosine is a half-sized I signal with and image signal and a Q signal that interferes destructively at baseband plus some images. A low-pass filter will recover only the Q signal. The Q path has the opposite result, the I signal interferes destructively and the Q signal is reconstructed at baseband.



Summary

- IQ receivers aren’t black magic: the I and Q components in the spectrum cancel in the received signal.

- IQ phase and gain balance and extracting the input phase are very important in IQ transceivers.