ECE 4880: RF Systems

Fall 2016

Chapter 8: Frequency Strategy

Reading Assignments:

1. T. H. Lee, *The Design of CMOS Radio Frequency Integrated Circuits*, *2nd Ed*, Cambridge, 2004. Sec. 13.1 – 13.2; 19.1 – 19.4.
2. W. F. Eagan, *Practical RF System Design*, Wiley, 2003, Chap. 7. (No detailed understanding of the “Spur Plot” is needed)

Game Plan for Chap. 8: Frequency strategy

1. Mixer fundamentals
2. Image signals and noises, and the mitigation
3. Signals with high nonlinearity: *m:n* spurs
4. Frequency strategy for superheterodyne, homodyne and up conversion
5. Case studies into the mixers of transmitters and receivers

8.1 Mixer fundamentals

We have been dealing mostly with *linear* and *time-invariant* modules (which by definition can be characterized by a single small-signal frequency response, such as those represented by the S, T, Y, Z and ABCD matrices). Nonlinearity is treated with Taylor expansion in the hope that the higher-order terms can be eventually truncated. Nonlinearity has been seen as things to stay away, such as in the definition of spur free dynamic range (SFDR). However, there are several fundamental issues that the frequency in the RF system has to be converted, such as antenna size (needs to be around quarter wavelength to be reasonably efficient), channel selection (e.x., 1MHz bandwidth to be directly selected from a 2.4GHz carrier will be too difficult to be implemented by tunable filters), and last but not least, FCC regulation.

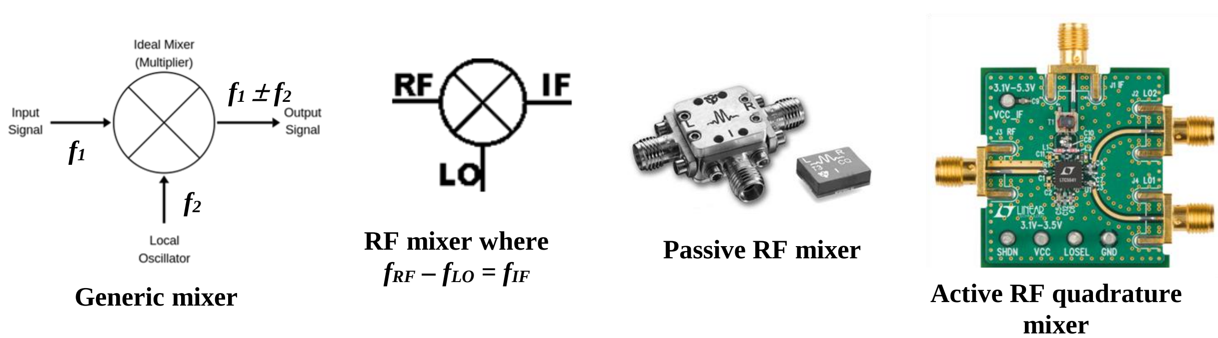

Fig. 8.1. Mixer schematics and examples.

Mixers are thus the critical and unique module in its frequency conversion role. Schematics and examples of RF mixers are shown in Fig. 8.1. Nevertheless, the analysis techniques on *gain, noise* and *nonlinearity* we have learned up to now can all be applied to the mixer when one of the terminals (most often the local oscillator LO) is held constant. Actually we finally have enough techniques in our toolbox to do a proper analysis. For sure, LO will change its frequency for channel selection, as well as change its magnitude to either improve the RF gain or decrease the spurs (will be clear later).

As frequency conversion is involved, the mixer needs to be either nonlinear (such as a diode, where we will use the 2^{nd}-order Taylor term in the nonlinear IV relationship to perform the mixing) or time variant (by injecting another frequency through the multiplication function). When the noise figure is considered, the passive nonlinear implementation may not be inferior, especially when the RF gain does not need to be large. However, we will restrict our focus to the more popular mixer topology: the active multiplier, where usually a cross-coupled Gilbert cell is used (you can learn that in ECE 5790).

The term "heterodyne" means "different frequency" or "different power", where frequency conversion happens. The heterodyne concept for easier radio implementation was first developed by Reginald Fessenden in 1901 as a direct-conversion heterodyne receiver, where the "baseband" audible beat was transferred to a higher RF frequency for smaller antennas (kHz RF will need a very large antenna to cause far-field radiation). The "direct conversion heterodyne" used the frequency conversion once, and has several names: homodyne, synchrodyne and zero-IF, to differentiate it from the much more popular way of "superheterodyne" later invented by Edwin Armstrong in 1918 which converts the frequency twice. The naming of homodyne and synchrodyne is actually a bit confusing, but please be patient. We will first understand superheterodyne and then we will use an asymptotic analysis to see this design variation.

The most important (and historical) frequency conversion technique is the superheterodyne, meaning converting the frequency two times. We will look at the receiver first, and then generalize to the transmitter later. A general superheterodyne receiver, which you have taken a peek in the dB link budget calculation, is shown in Fig. 8.2, where the RF signal around the RF frequency f_{RF} is first converted to an intermediate frequency f_{IF} by mixing with the LO frequency f_{LO} and then converted again to the baseband to be sampled and possibly digitized. Early radios had the baseband at the audible bands (20kHz – 40kHz bandwidth) to drive a speaker directly (audible detector). Two mixers perform the two frequency conversions, and when designed correctly, will greatly relieve the channel selection task from the filters.

The signal multiplication in the mixer can be understood from Eq. (8.1):

$$\begin{aligned} p_{out,mixer} &= g_{multi} \cdot A_{RF} \cos(\omega_{RF} t) \cdot A_{LO} \cos(\omega_{LO} t) \\ &= g_{multi} \cdot \frac{A_{RF} A_{LO}}{2} \left[\cos(\omega_{RF} - \omega_{LO})t + \cos(\omega_{RF} + \omega_{LO})t\right] \end{aligned} \tag{8.1}$$

We notice that in addition to the frequency conversion, the output at f_{IF} has a gain, in reference to A_{RF}, $g_{multi}A_{LO}$ (g_{multi} will be in V^{-1} and A_{LO} in V). In active multipliers with topology like a differential amplifier, g_{multi} can be significant, and the mixer gain is then proportion to A_{LO}! A tunable RF gain is readily available. However, the choice of the LO power level has many side effects (in particular, the spurs by nonlinearity), and we need to treat them more carefully later. Passive mixer often has a significant attenuation due to the magnitude of a_2 (the coefficient for the 2^{nd}-order nonlinearity). However, it is not to say that active mixer is always better than passive mixer. Besides the apparent power consumption, passive mixer often has lower noise figure, and therefore, a more detailed analysis will justify the choice. The LO oscillator signal can be an ideal monotone, or a delta function in the power density plot. The desired signal and its associated baseband bandwidth will keep its shape (and hence modulation) after the convolution with the delta function, which is just a pure shift in frequency. For sure, there can be noises in LO, which will be translated directly into the baseband. Again, we will treat the phase noise in LO individually in Chap. 9.

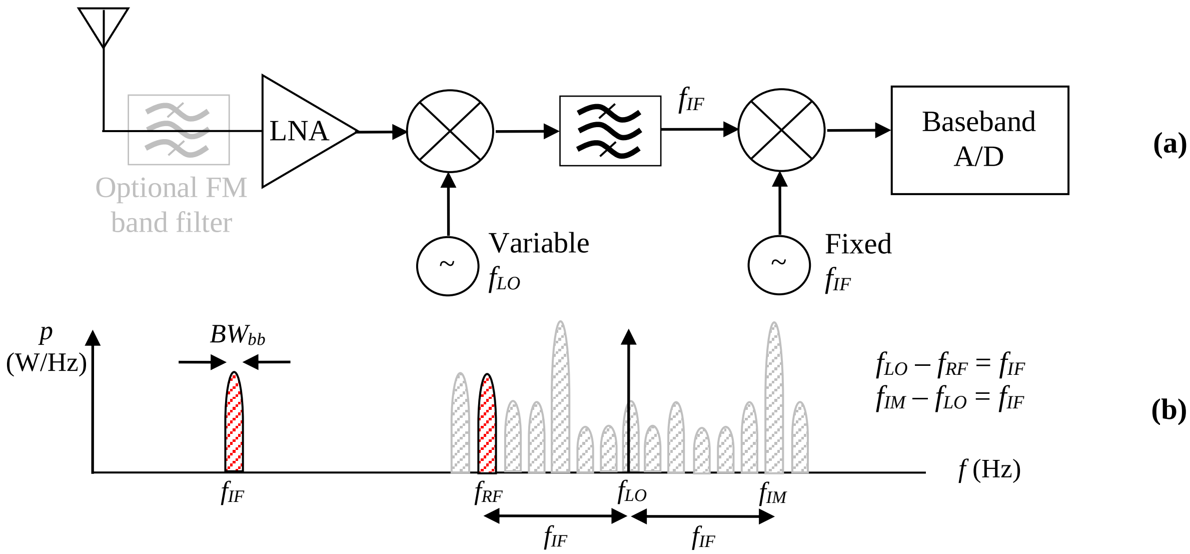

Fig. 8.2. (a) A conventional FM superheterodyne receiver, with optional components shown. (b) An example of f_{RF}, f_{LO} and f_{IF} in a conventional FM radio. f_{RF} can be between 88 – 108MHz, f_{IF} is set at 10.7MHz, $f_{LO} = f_{RF} + f_{IF}$, and baseband BW_{bb} = 200kHz. The frequency strategy is single-side band (f_{RF} on only one side of f_{LO}) with high side injection ($f_{LO} > f_{RF}$) for down conversion ($f_{IF} << f_{RF}$).

8.1.1 Advantages of the superheterodyne receivers

In most superheterodyne receivers, the major advantages include that:

1. The different f_{RF} frequencies and the associated baseband bandwidth BW_{bb}, in different "channels", such as the FM radio stations, are ALL converted to the same f_{IF}, which greatly simplifies the baseband processing designs. That is, the "station selection" within the FM band is performed entirely by the mixer by tuning f_{LO}, not by filters.

2. The cost of bandpass filters is often proportional to its quality factor $Q = f/\Delta f$. If we have to choose a narrow Δf at a high f, the required Q can be too high and hence the filter too expensive. For the FM example in Fig. 8.1, if channel selection is performed around the original f_{RF}, the required filter Q will be about f_{RF}/BW_{bb} = 100MHz/200kHz = 500. After down conversion to around f_{IF}, to purify the intended channel will only need a filter Q of 10.7Mz/200kHz = 54. The filter cost is heavily influence by its Q factor, order (how sharp the transition is) and precision (frequency and gain). This economical feature however significantly changes when filtering can be achieved by "software definition", where Q is related to the sampling precision, which greatly relax the cost structure to construct high-Q filtering.

3. To optimize the usage of the shared wireless spectrum, inaccuracy in the frequency between two radio transceivers will cut in the allocated spectral budget (actually you need to consider the worst case of any two transceivers that can communicate). The precision in the LO frequency f_{LO} can be more "economically" achieved than passive filters, which has been true since 1920's (the availability of stable crystal oscillators) up to around 2000s. Still remember that this economical feature however significantly changes when filtering can be achieved by "software definition".

4. The phase noise in LO (which we will treat in the last Chapter) is generally limited by the Flicker noise, and has a $1/f^{\alpha}$ characteristics around f_{LO}. When f_{RF} is about f_{IF} away, we have some design rooms to deal with the LO phase noise.

5. The intermediate frequency f_{IF} is an internally defined frequency, and can be much more accurate and predictable than f_{RF}, which has gone through distance with many interference and nonlinearity. Filters around f_{IF} can perform better even with the same Q factor!

8.1.2 Choices of single side, high side and down conversion

In Fig. 8.2, we have implied several frequency strategies, which are matters of choices. We will use the FM radio receiver as the example still. By FCC, f_{RF} in the air can be between 88 – 108MHz, f_{IF} is set at 10.7MHz, $f_{LO} = f_{RF} + f_{IF}$, and baseband BW_{bb} = 200kHz (sufficiently large for several audible bandwidths of 40kHz[1]). We have chosen f_{RF} to be on only one side of f_{LO}, which is called the *single-side band* (SSB). SSB is more efficient than *double-side band* (DSB) to arrange channels, although DSB is more useful in RF imaging. SSB will also require less absolute accuracy of f_{LO}, as the sampled bandwidth is f_{IF} away. We also choose here $f_{LO} > f_{RF}$, which is called the *high-side injection*. As f_{RF} is f_{IF} away from f_{LO}, $f_{LO} > f_{RF}$ will give a lower f_{LO} ratio (variable f_{LO} for channel selection) than the case for $f_{LO} < f_{RF}$.

For FM of f_{RF} between 88 – 108MHz with f_{IF} at 10.7MHz, high side injection requires f_{LO} changes between 98.7MHz and 118.7MHz, where the ratio f_{LOmax}/f_{LOmin} = 1.2, an easily achievable ratio by the frequency synthesizer. For low side injection, f_{LO} needs to change between 77.3MHz and 97.3MHz, where the ratio f_{LOmax}/f_{LOmin} = 1.3. This is not much different, so we actually can do either high side or low side injection for FM.

For AM radio, it is however another story. AM bands are between 530kHz – 1610kHz with a choice of f_{IF} = 455 kHz (poor choice historically, but probably will not change any longer). BW_{bb} is 20kHz[2]. High side injection requires f_{LO} changes between 985kHz and 2065kHz, where the ratio f_{LOmax}/f_{LOmin} = 2.1, an easily achievable ratio by the frequency synthesizer. For low side injection, f_{LO} needs to change between 75kHz and 1155kHz, where the ratio f_{LOmax}/f_{LOmin} = 15.4, which is a tuning range too large for most good quality oscillators before the rational frequency synthesizer is available (depending on a good quality of phase-lock loops).

Observed from Eq. (8.1), the frequency conversion has generated two frequencies at the output of the multiplier: $f_{LO} - f_{RF} = f_{IF}$ and $f_{LO} + f_{RF}$. When we choose $f_{LO} - f_{RF}$ and $f_{IF} << f_{RF}$, this is called down conversion. It is actually possible to use $f_{LO} + f_{RF}$ as well, although the first two advantages above will no longer be valid. However, up conversion has literally NO image frequency issues, and is worth considering for new applications indeed when digitization happened early in the signal chain.

8.2 Image signals and noises

The superheterodyne scheme (down conversion) has created two problems related to the frequency conversion: the image and the spurs. We will first deal with the image frequency issues, which you have seen briefly in the discussion of noises in Chap. 4. In Fig. 8.2, in addition to the presently selected

---

[1] FM radio usually uses subcarriers to give more than one "sound channels". Stereo means L-R speakers are driven independently, and there are quadraphonic, Dolby, etc. for various sound effects. BW_{bb} of 200kHz is large enough to further encompass captions and reading services.

[2] Due to BW_{bb} = 20kHz, AM has poorer sound quality and only monotone possible. Many of the AM designs inherited from, yes, telephones, where good-quality music sound is not expected. Even if the radio gives full audible information, the speaker/headset of old telephones will not give good sound any way.

channel, the other channels may also have signals, maybe even stronger signals than the present one that can cause severe interference. The mixing during down conversion, whether for high side or low side injection, will also map the image frequency f_{IM} into the base band, where $f_{IM} – f_{LO} = f_{IF}$ here, or f_{IM} is $2f_{IF}$ away from the desired signal. If around f_{IM} there is just thermal noise, the mixer will have just two times more thermal noise (3dB more) by the image frequency. However, there can be VERY strong interference signals around f_{IM}, as shown in Fig. 8.1, which can entirely swamp the intended signal!

Image frequency rejection has plagued superheterodyne from its very perception, and there are many possible solutions, all with some system cost or design tradeoffs.

8.2.1 Image rejection by f_{IF} choices

From the filter Q factor argument, the superheterodyne relieves the channel selection filtering from f_{RF}/BW_{bb} to f_{IF}/BW_{bb}. The first mitigation of image rejection is to put an "overall-band filter" before the LNA with a properly chosen f_{IF}. For the FM example, if we only allow any channel of FM to come in the signal chain, the Q factor of the "overall-band filter" will be just 88MHz/(108MHz – 88MHz) = 4.4, which can be readily achieved. Actually, most of your antennas have some bandwidth already, which can add up to the image-rejection purpose. If f_{IF} is chosen as: $f_{IF} > BW_{RF}/2$, we will find that all possible image frequencies will fall outside BW_{RF}, and can be attenuated by the "overall-band filter", which here for its purpose is the "image-rejection filter". This scheme will work for both high-side and low-side injection of f_{LO} in the receiver mixer.

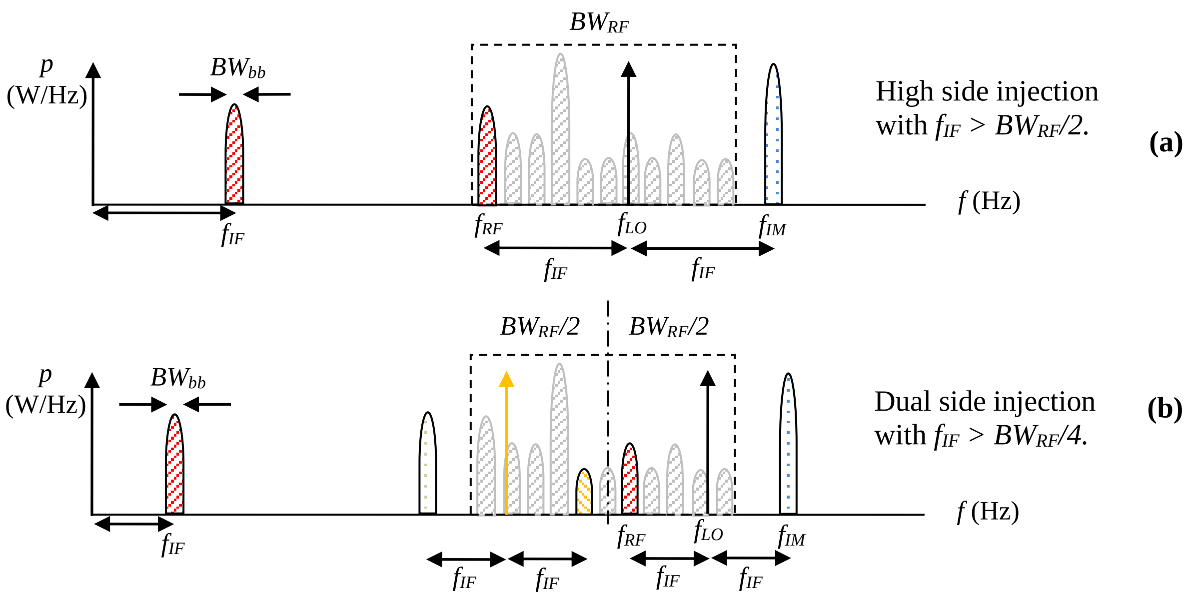

Fig. 8.3. (a) Image-rejection filtering for BW_{RF} band pass with high-side LO injection. As long as $f_{IF} > BW_{RF}/2$, the image frequency will fall outside the allowable BW_{RF}. (b) Image-rejection filtering for BW_{RF} band pass with dual-side LO injection, i.e., low-side LO injection for the lower half of $BW_{RF}/2$, and high-side LO injection for the higher half of $BW_{RF}/2$. As long as $f_{IF} > BW_{RF}/4$, the image frequency will fall outside the allowable BW_{RF}.

We can actually cut f_{IF} in half by employing dual-side LO injection. As we know the channel to be selected will be in the lower or higher half of BW_{RF}, if we emply low-side LO injection for the lower half of $BW_{RF}/2$, and high-side LO injection for the higher half of $BW_{RF}/2$, then with the previous overall band image-rejection filter, all image frequency will fall outside BW_{RF}. The penalty is the larger LO tuning range. For the FM example, we will set f_{IF} = 5.2MHz > (108MH – 88MHz)/4 for dual-side LO injection. The required tunable LO will have f_{LOmax} = 108MHz + 5.2MHz = 113.2MHz and f_{LOmin} = 88MHz – 5.2MHz = 82.8MHz, and the tuning range is: (113.2MHz/82.8MHz) = 1.4, which is still easily achievable.

8.2.2 Image rejection by up conversion and dual conversion

From Eq. (8.1), it is actually possible to use information around $f_{LO} + f_{RF}$ instead of $f_{LO} - f_{RF}$. This is called up conversion. An image interference is still possible through that mixer, which will be at $2f_{LO} + f_{RF}$. With any reasonable choice of f_{LO}, this frequency is easily rejected at the antenna end. For up conversion, the second advantage of superheterodyne (easier handing around the lower f_{IF}) disappears. Therefore, in all-analog processing where filter cost is significant, up conversion is not popular. If we have sufficiently high A/D sampling after the up-conversion mixer, then this becomes a possibly competitive choice. Notice that the tuning range of LO will be relaxed as well. Most up conversion schemes are employed for lower frequency band indeed (< 50MHz). This does not only include the classical AM radio, but very importantly the ground radar, the near-field communication (NFC)[3] and the long-range amateur radio[4].

(Exercise) For the AM radio above, if we choose the up conversion scheme and set f_{IF} = 5.1MHz, work out the frequency strategy of the varialbe f_{LO} for channel selection. What is the benefit of doing this? Observe why this will be more difficult for the FM radio?

In the superheterodyne scheme, we notice that when f_{IF} is low, we may need a better filter around f_{RF} to help channel selection and image rejection, as well as a more precise LO with larger tuning range (on the first order, a low f_{IF} will make f_{LO} track f_{RF} closely, so the tuning range is across the allowable channels approximately). On the other hand, a high f_{IF} will make image rejection easier but increase the burden for the further processing around f_{IF}. It is actually possible to use two f_{IF}, or even more, to combine the advantages of small and large f_{IF}. This is called "dual conversion" (not combining down and up conversion, but use of two f_{IF}!). In the extreme, we can use a homodyne receiver assisted by a superheterodyne with a relatively high f_{IF}. Such complexity is out of the scope of this class, and its implementation is mostly in the software-defined radio.

8.2.3 Image rejection by parrallel cancellation

There is a "third" way to reject image interference (followed the original title by D. K. Weaver in 1956). We notice that the signals at $f_{IM} - f_{LO}$ and $f_{LO} - f_{RF}$ after the mixing will have a phase difference of 180°! This cannot be differentiated by a single $\cos(\omega_{LO}t)$ multiplication because:

$$\cos(\omega_{LO} - \omega_{RF})t = \cos(\omega_{LO} - \omega_{IM})t = \cos(\pm\omega_{IF})t \qquad (8.2)$$

However, if we also mix with the quadrature, we can make the intended signal in phase and the image signal out of phase as shown in Eq. (8.3) for upside injection where $f_{IM} - f_{LO} = f_{LO} - f_{RF} = f_{IF}$

---

[3] NFC is used mostly for security purposes at frequencies around 13.56MHz (smart card), or even as low as 8kHz (entrance article control).

[4] Ionspheres for 6 – 10km above the earth surface (gas ionized by sun UV radiation) will reflect most of the radio waves in the HF band (1.6 – 30MHz), although with much additional noises. Radios that use this bouncing back and forth between ionspheres and earth surface are nicknamed as the "short-wave" radio. Before we have satellites (or during time when satellite communication is compromised by attackers), this is the only wireless way to transmit signals very far without putting a very high antenna, as the earth is spherical and waves need some line of sight or bouncing back and forth to be detected on far-way distance on earth surface.

$$p_{out,Q} = g_{multi} \cdot A_{RF}\left[\cos(\omega_{RF}t) + \cos(\omega_{IM}t)\right] \cdot A_{LO}\sin(\omega_{LO}t)$$
$$= g_{multi} \cdot \frac{A_{RF}A_{LO}}{2}\left[\sin(\omega_{LO} - \omega_{RF})t - \sin(\omega_{IM} - \omega_{LO})t + \sin(\omega_{RF} + \omega_{LO})t + \sin(\omega_{IM} + \omega_{LO})t\right] \quad (8.3)$$

We can then use the two possible architectures in Fig. 8.4 for image rejection. The first architecture, invented by Hartley, in Fig. 8.4a relies the precise shift of I/Q and the quadrature hybrid, which is difficult when percentage bandwidth is large. The next architecture in Fig. 8.4b, invented by Weaver, relies only on identical I_1/I_2 and Q_1/Q_2 and their phase precision, although 180° shift can be more easily implemented for broad band, such as a quasi-static inverter. This is easier now, as we are in the domain of the relatively low f_{IF}, not f_{RF}.

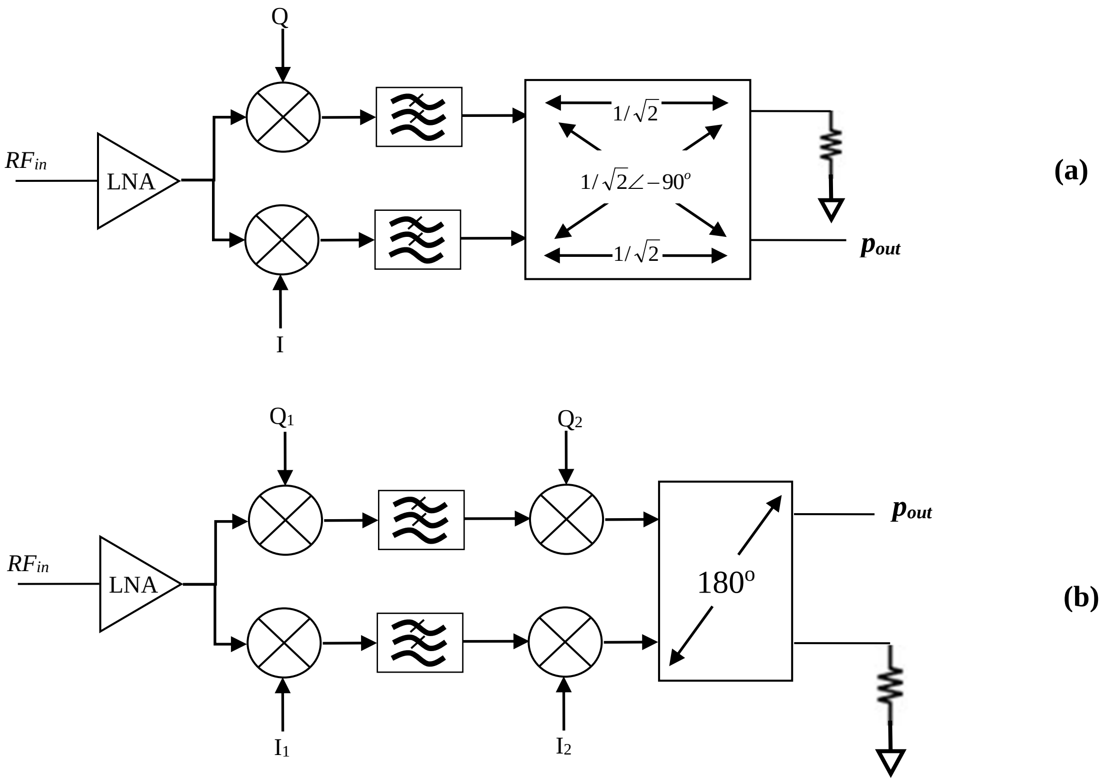

Fig. 8.4. (a) Image rejection mixer with a Hartley architecture. Signal: in-phase; image: 180° out of phase. (b) Weaver architecture for image rejection.

Although the functional derivation can be rigorous and clear to some students, it is probably more intuitive to understand the quadrature scheme for image rejection by the spectral representation. Recall that we can represent a real periodic signal in the Fourier domain as shown in Fig. 8.5. Notice that the phase offset in each sinusoidal function has been absorbed into t. However, with the same t, cosine and sine are 90° apart. We can use a low-side injection example to show the image frequency problem. Assume that both the desirable signal around ω_{RF} and the unwanted interfering signal around ω_{IM} are represented by their respective cosine functions. Surely here the desirable function can have its own phase modulation, and the interfering image signal can have a different offset from the signal. For the low-side injection example here, we have $\omega_{RF} - \omega_{IM} = 2\omega_{IF}$, as shown in Fig. 8.6.

$$\cos(\omega t) = \frac{e^{j\omega t} + e^{-j\omega t}}{2}; \qquad \sin(\omega t) = \frac{e^{j\omega t} - e^{-j\omega t}}{2}$$

Fig. 8.5. Spectral representation of real signals in cosine and sine functions. Notice that the phase offset has been absorbed into *t*. However, with the same *t*, cosine and sine are 90° apart.

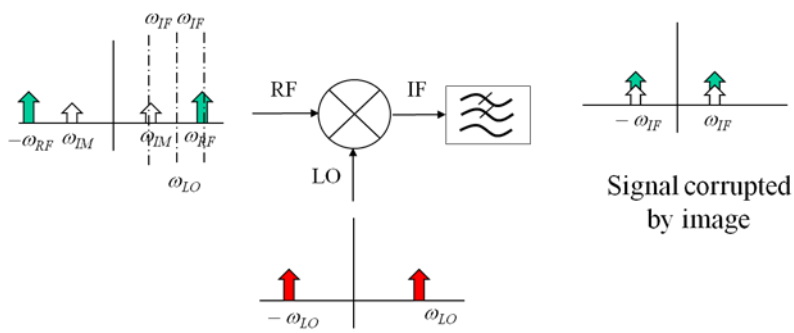

Fig. 8.6. Low-side injection example of the desirable signal at ω_{RF} to be corrupted by the interfering signal around ω_{IM}.

The Hartley image rejection can then be represented in the spectral domain, as shown in Fig. 8.7, whereas the Weaver image rejection is shown in Fig. 8.8.

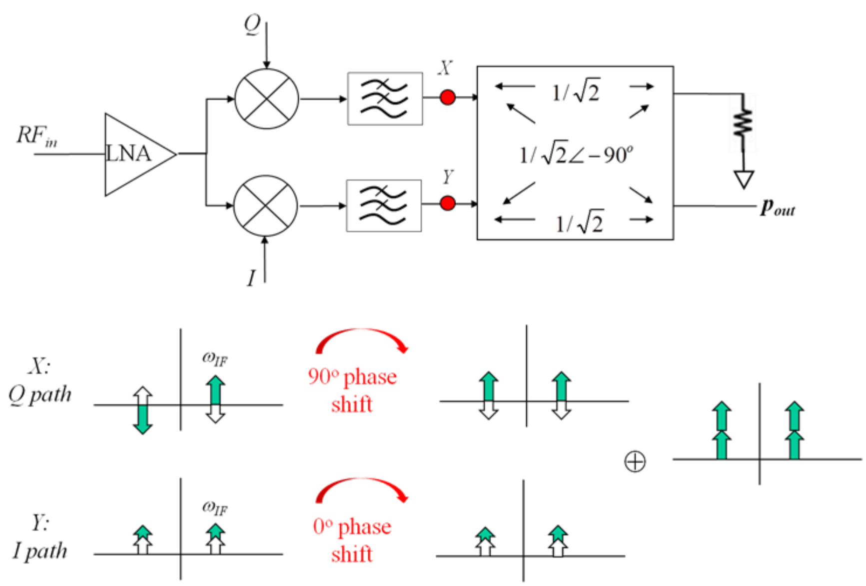

Fig. 8.7. Hartley receiver architecture in the spectral representation for image rejection.

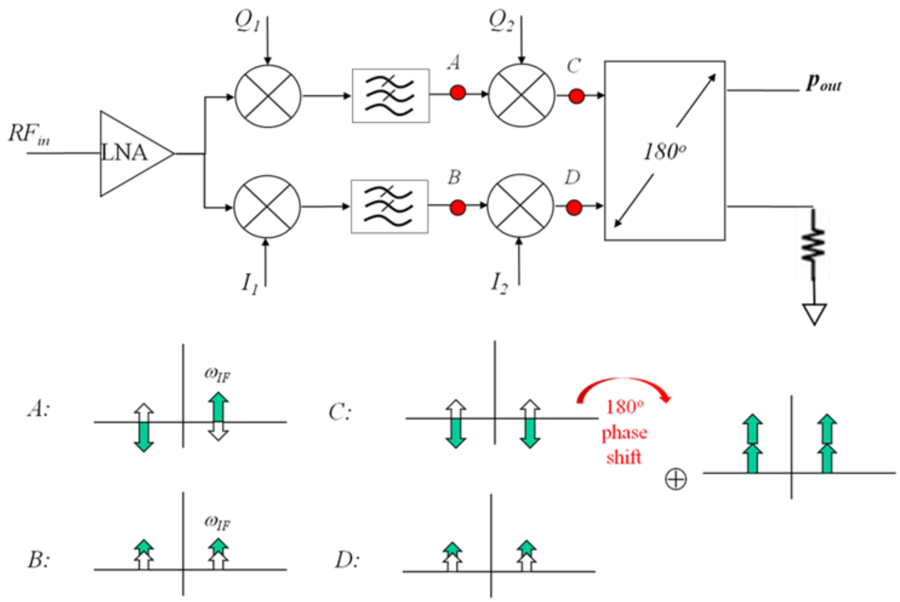

Fig. 8.8. Weaver receiver architecture in the spectral representation for image rejection.

Exercise: From the arguments of advantages and disadvantages in large and small f_{IF}, evaluate the homodyne receivers (f_{IF} = 0) in Fig. 8.9 against superheterodyne receivers, especially for hardware and software-defined implementations. Take special notice when f_{LO} cannot be exactly "synchronous" with f_{RF} and the potential image interference issues.

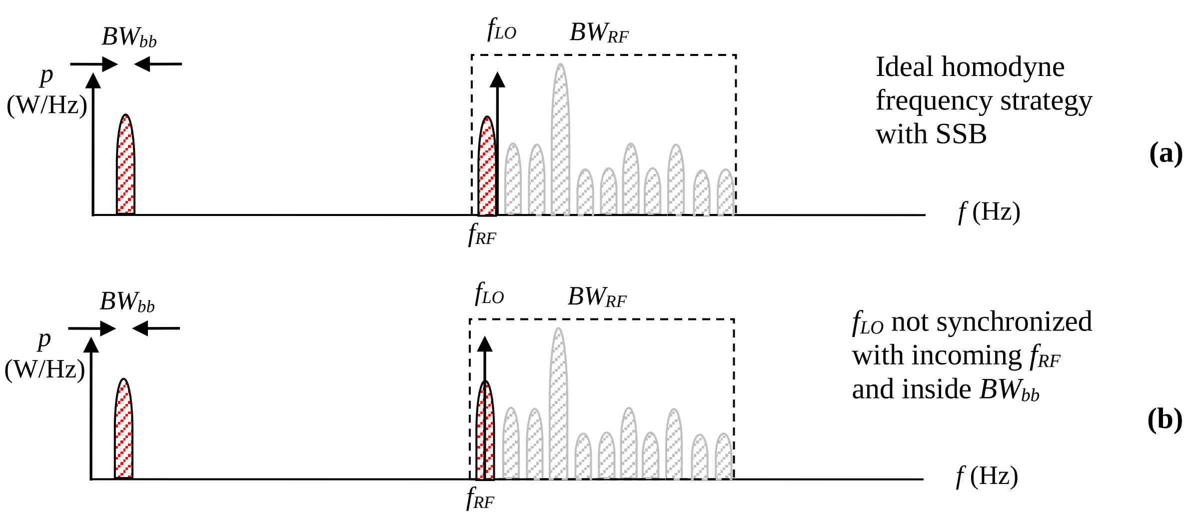

Fig. 8.9. (a) Homodyne frequency strategy with ideal single-side band; (b) f_{LO} not synchronized with f_{RF} and sitting inside the data bandwidth.

8.3 Signals with high nonlinearity: *m:n* spurs

LO is generated by some oscillators and frequency synthesizers. If A_{LO} is very small to minimize any possible nonlinearity during LO generation, the phase noise will cause a relatively small SNR for the LO signal. Larger A_{LO} will relieve this issue somewhat (but the phase noise of LO is multiplicative, so it will still be reflected to f_{IF}), but now we may have some harmonics in LO. The interaction of LO harmonics to the in-band RF signal (or whatever is not filtered out before hitting the mixer) is called "spurs". In the old days of radio, larger A_{LO} is purposely chosen to get some gain in the mixer as well as to boost SNR, but the RF engineers need to "catch" those spurs when they cause too much signal interference, and then iteratively re-design the filters or decrease the nonlinearity in LO generation (for sure, this will make the system more expensive). Spurs are often named by *m:n*, where *m* is the multiplication factor for f_{RF} and *n* is the multiplication factor for f_{LO}. Both m and n can be any integer including negative. With a reasonable A_{LO}, for sure, the higher harmonics are probably too small to cause issues. However, even lower ones have a large combination of possible *m:n*. Table 8.1 shows a possible "spur" table with $|m, n| < 4$ in conventional FM radios. With the improved LO generation and A/D data conversion, spurs are no longer done by playing cat-and-mouse in the spur plots and tables to optimize cost, but mostly become part of the CAD.

Table 1. A spur table example for FM radio for RF at –20dBm and LO at 10dBm.

m	*n*	$f_{RF,low}$ (MHz)	$f_{RF,high}$ (MHz)	Spur power level below signal (dB)
-3	3	73.8	93.9	50
-2	2	72.0	92.1	74
1	**-1**	**88.0**	**108.2**	**0**
2	-2	82.7	102.8	73
3	-3	80.9	101.0	49

8.4 Implications to the frequency strategy with software-defined ratio

If your Simulink is fast enough to catch the sampled waveform to be processed, then you have a software radio, and you do not need to fake in the noises to emulate the hardware. After digitization, most digital processing modules (gain, mixing, etc.) have noise temperature close to 0^oK (no additional noises due to signal regeneration, and if necessary, error correction code.)

8.5 Case studies into the mixers of transmitters and receivers