An Overview of Radar Waveform Diversity

Shannon D. Blunt, IEEE Fellow, Eric L. Mokole, IEEE Fellow

S.D. Blunt is with the Electrical Engineering & Computer Science Dept., Radar Systems & Remote Sensing Lab, University of Kansas, Lawrence, KS. He recently chaired the NATO SET-179 research task group on Dynamic Waveform Diversity & Design.

E.L. Mokole, recently retired from the Radar Division of the US Naval Research Laboratory in Washington DC, is an independent consultant. He was the US Navy service lead for the DoD Waveform Diversity working group and the Panel mentor for NATO SET-179.

Outline

I. Introduction
   • History & Definition of WD
   • A Biological Perspective
II. Radar Pulse Compression
   II.A. Pulsed vs. CW
   II.B. Waveform Classes
      • Linear FM
      • Nonlinear FM
      • Phase codes
      • Frequency codes
      • Noise radar
      • Ultrawideband
II.C. Matched Filtering
II.D. Delay-Doppler Ambiguity Function
II.E. Coherent Processing of Multiple Pulses
II.F. Waveform Metrics
      • Peak sidelobe level
      • Integrated sidelobe level
      • Power spectral density
II.G. Practical Considerations
      • Pulse eclipsing
      • Transmitter distortion
      • Range Straddling
II.H. Mismatched Filtering
      • Amplitude tapering
      • Mismatch loss
      • Optimized mismatched filtering
III. Waveform Diversity
   III.A. Practical Waveform Optimization
      • Transforming codes into waveforms
      • Incorporating transmitter distortion
      • Diversifying design metrics
   III.B. Environment-Specific Waveforms
      • Cognitive spectral avoidance
      • Matched illumination
      • Nonlinear harmonic radar
      • Radar-embedded communication
   III.C. Colocated MIMO Radar
      • Far-field signal vs. spatial angle
      • Impact of mutual coupling
      • Reactive power in the invisible space
      • Frequency diverse array
   III.D. Distributed Aperture Radar
      • Aspect angle RCS dependence
      • Coherently transmit
      • Bistatic ambiguity function
      • Illumination of opportunity
   III.E. Waveform Agility
      • Complementary codes
      • Stepped frequency
      • Frequency agility
      • Twin inverted pulses
      • Range sidelobe modulation
   III.F. Polarization Diversity
      • Poincaré sphere
      • Co-pol/cross-pol separation
IV. Conclusions
I. INTRODUCTION

The term Waveform Diversity (WD) [1-3] was coined by Dr. Michael C. Wicks of the Air Force Research Laboratory (AFRL) Sensors Directorate in 2002 [4]. He raised the notion of jointly pursuing a long-term roadmap for research, development, and manufacturing in the broad area of WD with representatives from the U.S. Army (Dr. Robert W. McMillan) and the U.S. Navy (Dr. Eric L. Mokole). Because this group felt that technology was sufficiently mature for extending and implementing the waveform research of the preceding 60 years across all pertinent scientific and engineering disciplines, they began a concerted effort to foster programs in this area. The main goals of the ensuing research have been to address a) the ever-increasing competition for radar spectrum and encroachment into what have historically been radar bands [5] and b) to leverage the rapid advances that are being made in digital signal generation (e.g. see [6-8]) and adaptive signal processing. The purpose of this tutorial is to provide the reader with the context in which WD has arisen, a sense of the tremendous breadth of the subject, and a sufficient starting point from which to explore WD further.

A good point of reference for a survey of waveform diversity (WD) is to state the IEEE Standard 686-2008 definition [9], which reads as follows.

“Waveform Diversity: Optimization (possibly in a dynamically adaptive manner) of the radar waveform to maximize performance according to particular scenarios and tasks. May also jointly exploit other domains, including the antenna radiation pattern (both on transmit and receive), time domain, frequency domain, coding domain and polarization domain.”

The pending update to this definition adds the following:

“Also used to denote adaptive receive processing that is applicable to such waveforms. See also: waveform; pulse compression; ambiguity function.”
This rather broad definition provides the general sense that WD can involve the modulation and/or exploitation of any aspect of the radar signal structure, both on transmit and receive. Further, the design, processing, and evaluation of the waveform clearly play a key role.

Fundamentally, a waveform comprises the modulation of an emitted signal such that, via appropriate filtering of the subsequent echoes at the receiver, desired aspects of the illuminated environment can be accurately measured [10]. Most often the radar itself generates this waveform, though there has been considerable work on passive radar that exploits the waveforms emitted by other spectrum users (e.g. FM radio) [11-14]. Generally speaking, research in WD can be categorized according to the areas delineated in Table I-1, though it is not uncommon for two or more categories to be considered jointly.

<table>
<thead>
<tr>
<th>Waveform design/optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interference rejection/avoidance</td>
</tr>
<tr>
<td>Multi-dimensional waveforms and processing</td>
</tr>
<tr>
<td>Bio-mimetic/bio-inspired operation</td>
</tr>
<tr>
<td>Multi-function operation</td>
</tr>
<tr>
<td>RF spectrum utilization</td>
</tr>
</tbody>
</table>

Diverse waveforms have been present in nature well before the information and technological explosions in the latter part of the 20th century. For example, echo-locating mammals (bats, dolphins, whales) have been exploiting WD for over 50 million years [15,16]. Recent technology has permitted replication and characterization of naturally occurring waveforms from such creatures [17], and this newly acquired understanding is being used to improve methods in man-made sensing systems like radar and sonar, which have been in existence for roughly a century. For example, bats emit dynamically adaptive acoustic waveforms that enable autonomous orientation, detection, localization, classification, discrimination, pursuit, and capture of prey. Researchers have shown that bats use a set of different waveforms (constant frequency, linear frequency modulation, hyperbolic frequency modulation, multiple harmonics, and possibly other nonlinear frequency modulation) to meet biological imperatives [18-20]. These
investigations suggest that a combination of flight profile, WD, and multi-algorithmic processing are crucially important factors in a bat’s success. Likewise, it has been found that dolphins vary the nature of sequential waveforms to enable discrimination in bubble-rich environments [21-23] and can generally perform sensing tasks far better than one might expect given their “mediocre equipment” [24]. In addition, researchers are trying to characterize how humpback whales use very loud acoustic emissions to trap prey within cylindrical walls of bubbles (bubble nets) of their creation [25,26].

What is currently called WD can be thought to have originated conceptually in 1933, when Edwin Armstrong invented frequency-modulated (FM) radio to improve audio signals conveyed via radio by controlling the noise static generated by electrical equipment and the earth's atmosphere. This invention led to the development of theoretical and experimental techniques for FM radar applications [27]. Prior to the 1990s, WD activities occurred as parts of other investigations such as the high-power microwave efforts of the 1950s and 1960s, waveform design for clutter rejection, electromagnetic compatibility, and spread spectrum techniques for communication and radar systems.

For traditional high-power radar, Klauder, et al [28] published a seminal paper on linear-frequency-modulation (LFM) radar, which was followed by investigations on optimum transmit waveforms (e.g. Barker, polyphase, and complementary codes) to tailor range sidelobes without suffering mismatch loss or mainlobe broadening [29-42]. In the 1970s and 1980s, new theoretical waveform designs were developed to improve the detection performance of radar [43]. Specifically, sub-complementary sequences and a new class of polyphase codes were found to improve pulse compression [44,45], and studies using Costas codes as detection waveforms yielded nearly ideal properties of the range-Doppler ambiguity function [46].

Since these seminal contributions, research in WD has truly experienced what Dr. Joe Guerci, in his 2014 IEEE Radar Conference keynote address, articulated as a “Cambrian explosion” [47]. Given that, a disclaimer is in order: this survey is intended to provide context
(something of a “phylum/genus/species” cataloguing if you will) for how the myriad forms of WD may be applicable to radar. Of course, despite our best efforts it is quite possible that we have missed or insufficiently detailed some salient features. Hopefully, however, this tutorial serves as an adequate starting point for further, deeper investigations into WD.

The next section serves an introduction to the fundamentals of radar waveforms and filtering, along with discussion of pertinent practical considerations. Section III then provides an overview of the different areas of research in radar waveform diversity, including waveform optimization and environmentally adaptive waveforms, MIMO and distributed aperture radar, waveform agility, and polarization diversity.

II. RADAR PULSE COMPRESSION

Before delving into the various topics with WD, it is first instructive to consider the essentials of radar pulse compression, which plays a pivotal role in WD. The concept of pulse compression was developed independently in Germany, Britain, and the US during World War II to address the problem of how to attain high range resolution, such as provided by a short pulse, while ensuring sufficient “energy on target” to provide detectable signal-to-noise ratio (SNR) in the receiver despite the peak power limitation on transmit [28]. The pulse compression solution entails the modulation of a much longer pulse that, after application of appropriate receive filtering to the reflected version of the waveform (for now assume a point scatterer), yields a response (Fig. II-1) having a mainlobe with resolution commensurate with what the short pulse would have provided, along with the addition of range sidelobes. Specifically, the range resolution (mainlobe width) is inversely proportional to the waveform bandwidth.
Much of the work in this area [10] involves the design of the waveform (the modulation scheme) and the subsequent receive filtering to suppress the range sidelobes, while minimizing degradation to the mainlobe in term of range resolution (widening of the mainlobe) and SNR loss at the mainlobe peak. Sufficient radial motion also induces a Doppler shift that must likewise be considered.

To illustrate why it is desirable to minimize the sidelobes, consider the result when two point targets have a range separation that is less than the pulsewidth $T$ and whose received powers are considerably different. Figure II-2 depicts the matched filter response to an LFM waveform used to illuminate these two point targets. Given that standard radar detection methods [48] rely on how much a prospective target stands out relative to the immediate surroundings, it is likely that the smaller target would not be detectable due to the sidelobes induced by the pulse compression response to the larger target.
This section explains the general principles of radar pulse compression, the classes of waveforms that are used, how receive filtering is performed, and describes the metrics that are employed to evaluate the goodness of a waveform/filter pair. Practical considerations for pulse compression are also explained. These general principles establish the operational sensing framework from which the larger study of waveform diversity has emerged.

A) Pulsed vs CW

It is quite common for a waveform to be modulated repeatedly onto multiple segments of the transmitted signal so that the echoes from these segments can be combined on receive (usually coherently) for enhanced gain and as a means to enable discrimination in the Doppler domain. If the transmitted segments are interleaved with the receive intervals, this scheme is referred to as a pulsed mode, where the pulsewidth $T$ is less than the pulse repetition interval (PRI) denoted $T_{PRI}$ and $T/T_{PRI}$ is called the duty cycle. Due to the wide variety of radar applications and implementations, the duty cycle could be as low as 0.1% or as high as 25% or more [5]. For a pulsed mode, $1/T_{PRI}$ is the pulse repetition frequency (PRF).
In contrast to pulsed operation, a continuous wave (CW) radar is one in which the emission of waveform-modulated segments does not alternate with the receive operation, but instead performs transmission and reception simultaneously. Frequency modulated CW (FMCW) is the most common form of CW in use, where the frequency is swept as a function of time. FMCW is primarily used by high frequency (HF) over-the-horizon (OTH) radars [49], as well as for short-range applications such as automotive radar [50] and is being explored as a means of on-board sense-and-avoid for small unmanned air systems (UAS) that require low size, weight, and power (SWaP) [51].

The CW mode can be thought of as a special case of pulsed operation in which the duty cycle is 100%. As such, the different types of waveforms are generally applicable to either mode. Pulsed operation is far more prevalent than CW in modern radar systems. Therefore, in the following we shall use terminology appropriate to pulsed operation, with the understanding that the same is generally applicable to each repeated interval of a CW mode.

B) Waveform Classes

There exist myriad varieties of waveforms, though they can generally be sorted into just a few categories. These categories are frequency modulated (FM) waveforms, phase codes, frequency codes, and random noise waveforms. While arguably not a separate class of waveform, one may also consider various forms of modulation across a set of pulses (to be coherently combined on receive) as a waveform attribute as well (further discussed in Sect. II-E). Of these categories, the most commonly used in operational radar systems are FM waveforms and the subset of phase codes denoted as binary codes.

Of the frequency modulated (FM) waveforms, the most prevalent is linear FM (LFM) [10, Chap. 4] because it is easy to implement in hardware, has attractive Doppler properties (see Sect. II-D), and for wideband operation is amenable to computationally efficient stretch processing
[52] on receive. The complex baseband representation of an arbitrary FM waveform of pulsewidth $T$ (normalized to unit energy) is

$$s_{FM}(t) = \frac{1}{\sqrt{T}} \exp(j\theta(t)), \quad (II-B1)$$

where $\theta(t)$ is the instantaneous phase, and its scaled derivative

$$\frac{1}{2\pi} \frac{d\theta(t)}{dt} = f(t) \quad (II-B2)$$

is the instantaneous frequency.

The LFM waveform $s_{LFM}(t)$, commonly referred to as a *chirp*, thus has the phase function

$$\theta_{LFM}(t) = \pm \pi B t^2 / T \quad \text{for} \quad 0 \leq t \leq T, \quad (II-B3)$$

where $B$ closely approximates the 3 dB power bandwidth for practical waveforms, and the $\pm$ indicates either an *up-chirp* (increasing frequency) or *down-chirp* (decreasing frequency) [10, Chap. 4]. The product $BT$ is referred to as the *time-bandwidth product*, which is also the coherent processing gain (or compression ratio) of the waveform when applying the matched filter (see Sect. II-C). Upon substituting (II-B3) into (II-B2), the instantaneous frequency for LFM is found to be $f_{LFM}(t) = \pm B t / T$, which is clearly a linear function of frequency. Taking another derivative yields the rate at which the frequency changes with time, otherwise known as the *chirp rate*, and is the constant $\pm B / T$ for LFM. In other words, LFM linearly sweeps over the bandwidth $B$ during the pulsewidth $T$.

The primary limitation of LFM is the high range sidelobes it produces (depicted in Fig. II-1), the largest of which is generally only about 13 dB below the mainlobe. One way in which these sidelobes can be reduced is by applying an amplitude taper as

$$s_{Tapered-LFM}(t) = a(t) s_{LFM}(t), \quad (II-B4)$$

where $0 \leq a(t) \leq 1$ for $0 \leq t \leq T$ is commonly one of the window functions otherwise used for digital filter design or antenna beampattern design (e.g. Taylor, Hamming, etc.) [10, Chap. 4].
The trade-off for sidelobe reduction in this manner is a broadening of the mainlobe (degraded range resolution) and SNR loss relative to the absence of tapering. The mainlobe is broadened because the amplitude weighting of an LFM serves to produce a “rounded off” spectral content, as opposed to the relatively flat LFM spectrum, which in turn reduces the 3 dB bandwidth. Since the LFM waveform defined in (II-B1) and (II-B3) has unit energy, the loss due to amplitude tapering on transmit can be determined as

\[
\text{SNR Loss}_{\text{transmit taper}} = -10 \log_{10} \left[ \int_{0}^{T} a^2(t) \, dt \right],
\]

noting that \( a(t) \) is a real function and bounded between 0 and 1.

For example, Fig. II-B1 shows the pulse compression response for a square-root Hamming-weighted LFM, where the largest sidelobe has been reduced to just below –40 dB, with an SNR loss of 2.7 dB and a resolution degradation (increase) by a factor of 1.5 (measured 3 dB below the peak of the mainlobe). Such amplitude control also tends to limit operation to lower-power radar applications since a high-power transmitter generally operates in saturation, thus necessitating the emission of a constant amplitude waveform [53, Chap. 10]. An alternative that alleviates the transmitter limitation is to taper only at the receive filter, which in this case becomes a form of mismatched filter (see Sect. II-H), though resolution degradation and SNR loss still occur to some degree.
Fig. II-3. Pulse compression response for LFM (black) and square-root-Hamming weighted LFM (red)

Besides LFM, many varieties of nonlinear FM (NLFM) have also been developed [53-58]. The general premise behind the development of NLFM waveforms is that the inherent spectral shaping performed by amplitude tapering in (II-B4) can also be achieved via determination of a time-varying chirp rate function, which avoids the need for amplitude control via tapering (see Fig. II-B2). Put another way, the instantaneous frequency is a nonlinear function of time (hence the name) or, comparing with (II-B3), the instantaneous phase is no longer a quadratic function of time. As such, NLFM can reap the sidelobe reduction benefit of tapered LFM without the associated transmitter limitations and SNR loss discussed above, though NLFM does still incur resolution degradation from the spectral shaping (see Figs. II-B3 and II-B4).
Fig. II-B2. Time-frequency relationship for an LFM waveform (left) and a generic NLFM waveform (right)

Fig. II-B3. Power spectral densities (PSDs) of LFM and NLFM waveforms with the same 99% power bandwidth (but different 3 dB power bandwidths)
Amplitude tapered LFM waveforms and NLFM waveforms are generally designed such that their overall time-frequency response achieves a desired power spectral density (PSD), due to the Fourier relationship between PSD and the waveform autocorrelation (see Sect. II-C). For NLFM design the stationary phase principle was proposed [55], which says the spectral density at a given frequency is inversely related to the rate of frequency change (chirp rate) at that frequency. Thus one can determine the phase function in (II-B1) that provides a PSD whose corresponding autocorrelation has low sidelobes (known to occur when the spectrum tapers towards the band edges [59]). For example, the Fourier transform of a Gaussian shape is another Gaussian. As such, if the PSD is designed to be Gaussian, the associated autocorrelation would, in theory, exhibit a mainlobe that rolls-off with no sidelobes at all. Also, just like with the tapered LFM in (II-B4), the NLFM can be amplitude tapered as well, which is referred to as hybrid FM [60-62].

Another class of waveforms that has received significant attention is that of phase-modulated codes (or simply phase codes), in which the pulsewidth $T$ is temporally subdivided into a set of constant-amplitude sub-pulses (or chips) of duration $T_c = T/N$, with each chip being modulated by a fixed phase value $\theta$ drawn from a discrete set referred to as the phase
constellation. Phase codes are generally classified as binary (or biphase) codes, where the phase constellation is composed of only $\theta = 0^\circ$ and $\theta = 180^\circ$, or as polyphase codes in which the set of phase values may only be limited by numerical precision (or $360^\circ/2^{(# of \text{bits})}$). Figure II-B5 illustrates the structure for a phase-coded (PC) waveform, which can be expressed mathematically as

$$s_{PC}(t) = \frac{1}{\sqrt{T}} \sum_{n=1}^{N} \exp(j\theta_n) \text{rect} \left[ \frac{t-(n-1)T_c}{T_c} \right] \text{ for } 0 \leq t \leq T,$$

where the $n^{th}$ of $N$ chips is modulated by the phase $\theta_n$ that is drawn from a constellation of $P$ possible values. Like the LFM waveform, the energy of the PC waveform is normalized to unity.

![Figure II-B5: Phase-coded waveform structure](image)

Myriad different phase codes have been developed because the determination of “good codes” can be achieved through various optimization approaches that permit searching over the set of $P^N$ possible codes. Well-known examples of binary codes include Barker codes, minimum peak sidelobe (MPS) codes, and maximal length sequences [63-66]. Likewise, well-known polyphase codes include Frank codes, P codes, and polyphase Barker codes [38,45,67,68]. See [10, Chap. 6] and [53, Chap. 20] for further details.

Clearly the number of possible codes becomes tremendously large as the number of chips $N$ increases (which approximates the time-bandwidth product, based on 3 dB power bandwidth). Because the construction of a phase-coded waveform inherently involves an abrupt phase transition every $T_c$ seconds (yielding poor spectral containment due to the resulting $\sin(x)/x$
spectral roll-off), it is also important to consider how to implement codes in a manner that is physically amenable to the radar transmitter (discussed in Sect. III-A).

In contrast to phase coding, a frequency-coded (FC) waveform structure can be defined as

\[
s_{\text{FC}}(t) = \frac{1}{\sqrt{NT}} \sum_{n=1}^{N} \exp(j \theta_n) \exp \left[ j2\pi \left( n - \frac{N+1}{2} \right) \frac{t}{T} \right] \quad \text{for} \quad 0 \leq t \leq T, \tag{II-B7}
\]

which amounts to a phase weighting (via the first exponential term of each summand) of a set of complex sinusoids with frequency separation \( \Delta f = 1/T \) (thus the sinusoids are orthogonal). This FC waveform possesses the same 3 dB bandwidth (and therefore the same \( BT \approx N \)) as the PC waveform structure from (II-B6) and is likewise normalized to unity energy. The combination of these complex sinusoids (also referred to as the multiple carrier frequencies), introduces a time-varying amplitude that inhibits high-power operation through the use of a saturated power amplifier, with the peak-to-average power ratio (PAPR) [69] providing a measure of how much the power must be backed off from the peak power to avoid distortion due to amplitude clipping. Thus PAPR also directly implies the loss in SNR one could expect for a frequency-coded waveform relative to a constant-amplitude waveform. This formulation is actually a single symbol interval of orthogonal frequency division multiplexing (OFDM), which is widely used in cellular communications [69]. Further, because this form is already non-constant amplitude, one may relax the restriction of using a constant-amplitude code (the first exponential term in (II-B7)) by replacing it with a constellation that employs both amplitude and phase (e.g. quadrature amplitude modulation (QAM)) to provide greater design freedom.

A more general scheme that incorporates both phase coding and frequency coding is the multicarrier phase-coded (MCPC) structure [70], which can also be viewed as multiple OFDM symbol intervals. For the same time-bandwidth product \( N \) as for the PC and FC structures in (II-B6) and (II-B7), respectively, in this case \( N_{\text{MC}} \) sub-codes having \( N_{\text{PC}} \) chips each are modulated
onto $N_{MC}$ carriers, such that the total dimensionality remains as $N_{MC}N_{PC} = N$. The $N_{PC}$ chips in each sub-code have a duration of $T_{MC} = T / N_{PC}$. The MCPC structure is thus [10, Chap. 11]

$$s_{MCPC}(t) = \frac{1}{\sqrt{N_{MC}T}} \sum_{n=1}^{N_{PC}} \sum_{m=1}^{N_{MC}} \exp(j \theta_{n,m}) \exp \left[ j 2\pi \left( m - \frac{N_{MC} + 1}{2} \right) \frac{t}{T_{MC}} \right] \text{rect} \left[ \frac{t - (n-1)T_{MC}}{T_{MC}} \right],$$

(II-B8)

where the frequency difference between each pair of adjacent carriers is now $\Delta f = 1/T_{MC}$ to maintain orthogonality of the carriers. Since (II-B8) is a phase-coded generalization of (II-B7), the MCPC structure can likewise be viewed as multiple symbol intervals of an OFDM signal. In addition to the PAPR issue and potential for a more general amplitude/phase coding as discussed for the FC scheme, the MCPC structure also has the same issue with spectral containment as mentioned for the PC scheme from the abrupt changes in the code values across all $M$ carriers every $T_{MC}$ seconds.

The last general waveform class that is based on pulse-compression processing is noise radar [71-75], in which the waveform is made to appear as noise, for which a bandpass representation can be expressed as [72]

$$s_{\text{noise}}(t) = a(t) \cos[\omega_0 t + \theta(t)]$$

(II-B9)

for the Rayleigh distributed amplitude $a(t)$ and a uniformly distributed phase $\theta(t)$. The random, and otherwise unstructured, nature of this emission scheme makes it inherently low probability of intercept (LPI) [76]. Noise radar is generally implemented as a form of non-repeating CW and generally is used as an ultrawideband (UWB) waveform to facilitate high range resolution. This emission scheme can also be filtered such that the resulting spectral shape yields a PSD having an associated autocorrelation with low range sidelobes [75]. Due to significant amplitude modulation (AM), combined with the fact that the amplitude is predominantly near zero (PAPR could exceed 10 dB), noise radar tends to be limited to lower-power, short-range applications.
Finally, it is worth mentioning the distinctly different class of UWB waveforms, which have very short temporal durations and very broad instantaneous bandwidths. A primary objective of radar is to achieve a large enough SNR to detect targets of interest, with sufficiently high range resolution to separate the different targets. To achieve very high range resolution, the notion of impulse radar has been pursued, subsequently leading to new UWB waveforms of much shorter duration than standard radar signals. The major demonstrated benefit that UWB radar provides is ultra-high resolution, which can be used for object characterization and identification. In particular, such waveforms and their associated radar systems were initially developed for forestry applications [77,78], for characterizing sea scatter [79], and for detecting underground utilities, land mines, and unexploded ordnance (ground penetrating radar) [80-86]. More recent systems have been designed for through-structure detection in urban areas [87], for imaging in search and rescue operations, and for obstacle avoidance in automobiles and micro air vehicles [88]. As the technology for UWB radar has developed and improved over the last fifty years, the sophistication and performance of these radars have increased. Nonetheless, several crucial issues remain that are problematic to UWB radar: spectral availability, hardware limitations in the transmission chain, electromagnetic interference and compatibility, difficulties with waveform control/shaping, and the unreliability of high-power sources for sustained use above 2 GHz. Consequently, UWB radar will probably be limited to short-range, low-power, directive, niche applications. To overcome these deficiencies, recent systems have taken advantage of increased memory, throughput, and computational speed to build stepped-frequency UWB radars for sensing through walls [89]. Since a significant body of literature exists on UWB and this venue has insufficient space to cover it adequately, this interesting field will not be discussed further. The authors suggest that interested readers avail themselves of the voluminous literature on UWB waveforms and systems.

To summarize the discussion of different waveform classes, Table II-1 lists the different waveforms and provides a brief synopsis of their attributes.
Table II-1: Waveform classes and attributes

<table>
<thead>
<tr>
<th>Waveform class</th>
<th>attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear FM (LFM)</td>
<td>easy to generate/process wideband, high peak sidelobes</td>
</tr>
<tr>
<td>Nonlinear FM (NLFM)</td>
<td>trade LFM resolution for lower sidelobes</td>
</tr>
<tr>
<td>Phase codes</td>
<td>easy to optimize, binary or polyphase sub-classes</td>
</tr>
<tr>
<td>Frequency codes</td>
<td>modulate onto different sub-carriers, high AM effects</td>
</tr>
<tr>
<td>Noise radar</td>
<td>non-repeating form of CW, high AM effects, LPI</td>
</tr>
<tr>
<td>Ultrawideband</td>
<td>very short pulse, very wide band, ultra-high resolution</td>
</tr>
</tbody>
</table>

C) Matched Filtering

Denote $s(t)$ as the complex baseband representation of an arbitrary waveform with temporal extent $T$ (pulsewidth for a pulsed waveform). For this waveform there exists a matched filter $h_{MF}(t)$ such that the SNR after filtering is maximized. This filter, originally derived by North [90], has the form $h_{MF}(t) = C s^*(T - t)$, where $(\bullet)^*$ denotes complex conjugation and $C$ is an arbitrary constant. For the purpose of comparison among different waveform classes and with the mismatched filters in Sect. II-H, it is convenient to define $C$ such that $\left(\int_{0}^{T} |h_{MF}(t)|^2 \, dt\right)^{1/2} = 1$, thus yielding a normalized matched filter that produces a unity noise-power gain regardless of the waveform.

The matched filter response to the waveform (without appreciable Doppler shift during the pulsewidth and for the waveform energy assumed to be unity) is the convolution

$$h_{MF}(t) * s(t) = \int_{t=0}^{T} s(t) \, s^*(t + \tau) \, d\tau,$$  \hspace{1cm} (II-C1)

which is also the autocorrelation of the waveform. If the energies of both $s(t)$ and $h_{MF}(t)$ are normalized to unity, (II-C1) likewise produces a unity magnitude at $\tau = 0$ (peak of the mainlobe).

As mentioned in Section II-B, the sidelobes of the autocorrelation can be minimized through proper design of the waveform PSD.
For $x(t)$ the scattering response in an illuminated environment, which consists of an unknown number of targets and ubiquitous clutter, the received signal at the radar can be expressed as

$$y(t) = s(t) * x(t) + v(t),$$  \hspace{1cm} (II- C2)

where $v(t)$ is additive noise and the influence of Doppler during the pulsewidth is neglected. The matched filter response for this received signal is therefore

$$\hat{x}_{\text{MF}}(t) = h_{\text{MF}}(t) * y(t),$$  \hspace{1cm} (II- C3)

in which the $\hat{x}_{\text{MF}}(t)$ term is the matched filter (MF) estimate of the true scattering $x(t)$. Note that it has become increasingly more common to perform this filtering operation in the digital domain, which is discussed in Sect. II-G.

For most radar systems, the matched filter response of (II-C3) is collected over a set of pulses for subsequent filtering over this set of responses via Doppler processing (possibly including clutter cancellation) or by synthetic aperture radar (SAR) processing (possibly including additional image focusing). The time interval for this set of pulses is referred to as the \textit{coherent processing interval} (CPI). These “next stage” processes after pulse compression rely on the \textit{slow time} Doppler shift that occurs between successive pulses, as opposed to the \textit{fast time} Doppler shift during a pulsewidth.

\textit{D) Delay-Doppler Ambiguity Function}

Thus far we have only considered the response to scattering that exhibits no Doppler shift during the pulsewidth. When there is radial motion between the radar platform and a given scatterer, a Doppler frequency shift is induced and may not be negligible. With regard to pulse compression, this Doppler shift imparts a time-varying phase change to the reflected waveform, thereby changing the response of the matched filter. Specifically, the matched filter response of
(II-C1) for a pulsed waveform of pulsewidth $T$ can be generalized to arbitrary delay $\tau$ and Doppler frequency $f_D$ as

$$\chi(\tau, f_D) = \int_{\tau=0}^{T} e^{i2\pi f_D t} s(t) s^*(t + \tau) \, dt,$$  

(II-D1)

which is known as the delay-Doppler ambiguity function as formulated by Woodward [91]. Note that one plots (II-D1) as $20\log_{10}(|\chi(\tau, f_D)|)$ versus $\tau$ and $f_D$.

For example, Fig. II-D1 depicts the ambiguity function for the well-known LFM waveform. Observe that the mainlobe is actually part of a delay-Doppler ridge that exhibits a gradual roll-off from the peak at $(\tau = 0, f_D = 0)$. The existence of this ridge is why LFM is also referred to as a \textit{Doppler tolerant} waveform, since an appreciable Doppler shift induces little SNR loss relative to the peak. The sidelobes of the autocorrelation of the LFM waveform in Fig. II-1 are part of a larger Doppler-dependent pattern known as Fresnel lobes (the lobing pattern surrounding the large delay-Doppler ridge).

![Fig. II-D1. Delay-Doppler ambiguity function for an LFM waveform (brightness in dB)](image)

An important property of the ambiguity function is that the maximum value occurs at $(\tau = 0, f_D = 0)$ [10, Chap.3]. It can likewise be shown [10, Chap.3] that
\[
\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |\chi(\tau, f_D)|^2 d\tau df_D = 1
\]  \quad (II-D2)

for arbitrary waveform structure, assuming the waveform energy is normalized to unity. In other words, there exists a conservation of ambiguity such that, if a waveform has lowered sidelobes in one location (say on the \( f_D = 0 \) axis), then a commensurate increase must occur elsewhere in delay-Doppler space. Finally, the 3 dB resolution in range (the distance between the peak and first null) for useful waveforms is approximately the reciprocal of the 3 dB bandwidth \( B \) of the waveform, while the Doppler resolution (peak to first null) is \( 1/T \) for pulsewidth \( T \).

\( E \)  \quad Coherent Processing of Multiple Pulses

For a single pulse, the Doppler resolution is \( 1/T \). However, for a coherent processing interval (CPI) of \( M \) identical pulses with constant pulse repetition interval (PRI) \( T_{PRI} \), the Doppler resolution is greatly improved to \( 1/(MT_{PRI}) \). Further, the ambiguity function for a single waveform-modulated pulse via (II-D1), for a CPI of identical pulses [10, Chap. 7], is now scaled as

\[
|\chi_{CPI}(\tau, f_D)| = |\chi(\tau, f_D)| \frac{\sin(M\pi f_D T_{PRI})}{M \sin(\pi f_D T_{PRI})}
\]

for \( |\tau| \leq T \).  \quad (II-E1)

Recalling from Section II-A that an FMCW waveform can be viewed as a set of pulses for which the duty cycle is 100\% (\( T_{PRI} = T \)), it is thereby evident that (II-E1) is likewise applicable to the processing of \( M \) segments of FMCW.

The coherent processing of multiple pulses also provides a mechanism for the inclusion of additional degrees of freedom for design of the overall delay-Doppler response. For example, greater control of the Doppler sidelobes can be achieved via interpulse (and also intrapulse) weighting over the CPI (see [10, Chap. 7]). Likewise, changing the PRI during the CPI (known as \textit{PRI/PRF staggering} or \textit{jittering}) can address the problem of blind speeds for moving target
indication (MTI) radar [92,93] (also see [10, Chap. 8]). One can even employ completely
different waveforms over the set of pulses in a CPI, which is generically referred to as pulse
agility or waveform agility and is discussed in greater detail in Section III-C.

F) Waveform Metrics

The determination of “goodness” of a waveform is generally dependent upon the evaluation
of attributes of the delay-Doppler ambiguity function from Section II-D, along with consideration
of practical characteristics such as time-bandwidth product (waveform dimensionality), spectral
containment, and amenability to the transmitter (see Sect. II-H). Here we focus on the attributes
that relate to the ambiguity function defined in (II-D1) for a single pulse.

Perhaps the most common metric for waveform range sidelobes is the peak sidelobe level
(PSL), or peak sidelobe ratio (PSR) [53, Chap. 20], here denoted $\Phi_{PSL}$ and typically defined
using the delay-Doppler ambiguity function of (II-D1) as

$$
\Phi_{PSL}[\chi(\tau,f_d)]_{f_d=0} = \max_{\tau} \frac{\chi(\tau,0)}{\chi(0,0)} \text{ for } \tau \in [\tau_{main},T]. \tag{II-F1}
$$

Only the zero-Doppler cut ($f_d = 0$) is considered, which amounts to neglecting the effect of
radial motion during the pulsewidth. Accounting for symmetry of $\chi(\tau,0)$ about $\tau = 0$, the
interval $[0,\tau_{main}]$ corresponds to the time (range) mainlobe, such that the interval $[\tau_{main},T]$
contains sidelobes. The PSL metric indicates the largest degree of interference that one point
scatterer can cause to another at a different delay offset (for both having zero Doppler).

For the class of waveforms known as linear period modulation (LPM) [94], which is also
referred to as hyperbolic FM (HFM) and is employed in sonar and by some species of bats, the
value [61]

$$
PSL_{LPM\text{ bound}} = [-20\log_{10}(BT) - 3] \text{ dB} \tag{II-F2}
$$
serves as a lower bound on PSL for the waveform time-bandwidth product $BT$ as defined in Section II-B. Although this PSL bound only holds for LPM (HFM) waveforms, it is nonetheless a useful benchmark with which to compare the performance of untapered FM waveforms (i.e. a fixed “measuring stick” for optimization purposes).

Another important metric is the integrated sidelobe level (ISL) [53, Chap. 20], which for the zero-Doppler cut ($f_D = 0$) of the ambiguity function can be defined as

$$\Phi_{\text{ISL}}(\chi(\tau, f_D)|_{f_D=0}) = \frac{\int_{\tau} |\chi(\tau,0)|^2 d\tau}{\int_{0}^{\tau_n} |\chi(\tau,0)|^2 d\tau}. \quad \text{(II-F3)}$$

The ISL metric is particularly useful for establishing the susceptibility to distributed scattering such as clutter. Conceptual depictions of PSL and ISL are shown in Fig. II-F1. Consideration of a Doppler interval could be included in the PSL and ISL metrics by generalizing the mainlobe and sidelobe regions to correspond to the interior and exterior, respectively, of a delay-Doppler ellipse.

![Fig. II-F1. Conceptual definition of PSL and ISL measured on the zero-Doppler ambiguity function response of a waveform](image-url)
Finally, given the Fourier relationship between a waveform’s autocorrelation (zero-Doppler cut of the ambiguity function) and PSD, it is useful to define a PSD-based metric as well. In this case, it is necessary first to establish a desired PSD $|W(f)|^2$ that corresponds to some desired autocorrelation response (with a sufficiently narrow mainlobe and sufficiently low sidelobes). A good example is a Gaussian PSD that is scaled to have the same energy as a constant-amplitude pulse of duration $T$ (Fig. II-F2). A metric in this context would then permit measuring “how close” (in some sense) the actual frequency response of the waveform is to this desired PSD.

![Fig. II-F2. Gaussian PSD in dB](image)

For instance, the frequency template error (FTE) metric [95] is defined as

$$
\Phi_{\text{FTE}}[S(f),W(f)] = \left( \frac{1}{f_H - f_L} \right) \int_{f_L}^{f_H} \left| \frac{|S(f)|^p - |W(f)|^p}{W(f)} \right|^q df ,
$$

(II-F4)

where $f_L$ and $f_H$ demarcate the edges of the frequency interval of interest (which should include enough of the spectral roll-off region to provide sufficient fidelity). The positive real values $p$ and $q$ define the degree of emphasis placed on different frequencies. For $p = 1$ and $q = 2$, the FTE metric defines a form of frequency-domain mean-square error (MSE). Alternatively, $p > 1$ overly
emphasizes frequencies with higher in-band power, and $p < 1$ overly emphasizes frequencies with lower out-of-band power. Note that only the spectral envelope (magnitude) is used in (II-F4) so that the phase response remains free for design.

\[ \text{Practical Considerations} \]

There are several practical aspects one must consider when designing/selecting a pulse compression waveform. From an operational perspective, the bandwidth and pulsewidth (and thus the time-bandwidth product) are selected to be suitable to the application (MTI, SAR, etc.) and, combined with the selection of PRF, to achieve acceptable maximum ambiguous range and Doppler values. For a pulsed mode, in which the receiver and transmitter operation are interleaved, the notion of \textit{pulse eclipsing} also arises.

Pulse eclipsing [96,97] (Fig. II-G1) occurs when the receiver turns on or off during the reception of a waveform-induced echo. Reflected echoes experience reduced SNR since a portion of the reflected pulse is not captured at the receiver. For frequency-swept waveforms such as LFM and many useful forms of NLFM, pulse eclipsing also translates into degraded range resolution because part of the received bandwidth is lost.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{image.png}
\caption{Fig. II-G1. Echoes (a) and (c) are eclipsed because they arrive at the receiver when the radar is transmitting. More echoes will be eclipsed if the duty cycle $T/T_{PRI}$ is increased.}
\end{figure}
For example, Fig. II-G2 illustrates a comparison between the matched filter response of a complete LFM echo and the matched filter response for an LFM echo that is eclipsed by 50%. Compared to the former, the latter suffers a 6 dB SNR loss and a factor-of-2 degradation in range resolution. Thus, while a longer pulse provides greater energy on target, and therefore higher received SNR in the absence of eclipsing, one must decide if the subsequent increase in the occurrence of eclipsing is worth the trade.

Another attribute that must be considered is the impact of the radar transmitter upon the waveform one wishes to emit. The purpose of the transmitter is to generate and amplify the waveform to a degree that the reflected echoes of much lower power can be adequately captured by the receiver relative to the noise and interference. The most common ways to generate a waveform are:

(i) a frequency swept local oscillator (LO) which is often used to produce an LFM chirp,
(ii) a surface acoustic wave (SAW) device which can be used to generate either LFM or NLFM waveforms, and
(iii) a digital arbitrary waveform generator (AWG) which is becoming increasingly popular due to its tremendous flexibility.
Following waveform generation, the transmitter high-power amplifier (HPA) then serves to produce the necessary emitted power, typical values for which could be ~100 W up to several MW depending on the application [5]. While vacuum tube HPAs such as klystrons, traveling wave tubes (TWTs), and crossed field amplifiers (CFAs) are still in widespread use due to their high power efficiency, achievable transmit power, and reliability [98], solid-state HPAs continue to make advances and have begun to be used in radars that employ active electronically scanned array (AESA) antennas [99]. Of course, advances in tube technology [100,101] likewise continue.

The overall transmit chain introduces two forms of distortion on the intended waveform, linear and nonlinear. Linear distortion is a direct result of the spectral shaping that arises from the finite bandwidths of the individual transmitter components, and its impact is that the waveform may experience amplitude ripple and phase distortion (dispersion). Nonlinear distortion is primarily caused by the typical HPA operation in saturation (particularly for tube-based HPAs), thereby creating the formation of intermodulation products (harmonics) from the pairwise multiplication of different frequency components in the waveform [102]. These intermodulation products introduce leakage into the surrounding spectrum, an effect generally known as spectral regrowth (see Fig. II-G3), which should be avoided as spectral congestion continues to increase [5].

![spectral regrowth](image)

**Fig. II-G3. Spectral regrowth can create interference for adjacent spectral users**

The presence of transmitter distortion is arguably the main reason why the use of certain forms of coded waveforms (namely polyphase codes) has thus far been limited. Specifically, the abrupt transitions between adjacent chips in a code correspond to out-of-band spectral content
that cannot pass the bandlimited transmitter. Further, this linear distortion introduces AM that is subsequently further distorted if the HPA is operated in saturation.

For example, Fig. II-G4 shows the spectral content of a $N = 64$ chip P4 code [103], which represents a complex baseband sampled version of an LFM waveform, and the spectral content of an LFM waveform with the same $BT$ for comparison. Using the form described in (II-B5) for the coded waveform, both the P4 and LFM have been implemented on an AWG and driven into an S-band radar testbed that includes a mixer, pre-amplifier, bandpass filter, and a class AB solid-state Gallium Arsenide (GaN) HPA. The resulting “emissions” were captured by a receiver in a loopback configuration (i.e. not emitted into free space) where each is down-converted to baseband, analog lowpass filtered, and then sampled at rate of 150 samples / chip interval (same rate as the version of each waveform that was loaded onto the AWG).

![Fig. II-G4. Spectral content of (left) P4 code before/after transmitter distortion and (right) LFM waveform before/after transmitter distortion](image)

For each waveform, Fig. II-G4 compares the spectrum of the original “AWG waveform” and the resulting captured “loopback emission”, thereby representing a before/after perspective on the impact of transmitter distortion. The inherent spectral shaping of the transmitter clearly exhibits significant distortion for the coded waveform, but far less so for LFM. This result is to be expected because, unlike the coded waveform, LFM contains no abrupt phase changes. Note that
for these examples spectral regrowth is for the most part not observed due to the use of a Class AB solid-state HPA, as compared to what one could expect from a tube-based HPA that can produce much greater output power (and distortion).

Examination of the pulse shape of each waveform after transmitter distortion (Fig. II-G5) also reveals that the abrupt chip transitions in the code translate into amplitude nulls commensurate with the amount of phase change. Since a P4 code exhibits larger phase changes near the beginning/end of the code, deeper nulls are present near the ends of the transmitter-distorted pulse shape. In contrast, the loopback-measured LFM shows only a small amount of amplitude ripple that is expected from any real system. As an aside, note that both pulse shapes in Fig. II-G5 exhibit rapid pulse rise/fall times, which are additional contributors to broader radar spectral content.

![Fig. II-G5. Pulse shape after transmitter distortion for (left) P4 code and (right) LFM waveform](image)

To help explain why these amplitude nulls occur, Fig. II-G6 illustrates the unit circle on which the phase constellation of the code is defined (here with just 8 equally spaced phase values). Where an FM waveform moves around this circle continuously according to the instantaneous frequency (i.e. instantaneous rate of phase change), a phase-coded waveform makes abrupt jumps from one phase value to another (here from $\theta_n$ to $\theta_{n+1}$). While we would wish these phase transitions to move around the unit circle, the high spectral content required to do so is
suppressed by the inherent bandlimiting of the transmitter, so that the transition instead moves through the interior of the unit circle, thus producing an amplitude null.

Fig. II-G6. Desired and actual phase transitions for a phase code due to transmitter effects

Distortion of the waveform is problematic because the considerable time and effort put into designing a code that is optimal in some sense (e.g. minimal PSL) may have been wasted if the distorted version of the coded waveform deviates from this optimality condition (which is rather likely). An additional problem that arises from the amplitude nulls observed above is that the saturated HPA is still generating power even though a low amplitude value is occurring at the output. As such this generated power is effectively reflected back into the system as a time-varying spike in voltage standing wave ratio (VSWR) that serves to produce intermittent increases in the operating temperature of the system. At best, such temperature increases translate into higher phase noise; while at worst these temperature fluctuations could potentially damage the system.

It should be noted that the limitations discussed above do not imply that phase codes are necessarily a poor option for waveform modulation. For lower power systems where AM effects can be well controlled or for operating modes/environments where distortion and spectral regrowth effects are acceptable (particularly if an accurate replica of the actual emitted waveform can be captured after the HPA for use in matched filtering), the design freedom provided by phase codes may still be an attractive option. Further, as discussed in Sect. III-A, there are
existing methods to convert binary codes into waveforms that are amenable to an HPA (thereby making use of binary codes quite common even for high-power systems), and recent work has shown that a well-known scheme from communications can likewise convert arbitrary polyphase codes into new kinds of HPA-ready NLFM waveforms.

There are practical aspects to be considered for the frequency-coded waveforms from (II-B7) and (II-B8) as well since they involve the weighted combination of multiple carriers, which subsequently induces significant AM effects [104]. While such waveforms are attractive from a design freedom perspective, the AM effects require significant power back-off to avoid distortion (discussed in Sect. II-B), thereby leading to a substantial SNR loss as determined by (II-B5) for the AM envelope $a(t)$. To demonstrate why the power back-off is so important, Fig. II-G7 illustrates the autocorrelation of an $N = 64$ optimized FC waveform from [105] using (II-B7) along with the autocorrelation of the same waveform after undergoing distortion by a saturated power amplifier. In terms of PSL as defined in (II-F1), the distorted waveform experiences a PSL degradation of 14.7 dB relative to the ideal case. The amplitude envelopes for these before/after distortion cases are also shown in Fig. II-G8. Because linear amplification is a necessity to avoid this distortion, such waveforms are thus restricted for use in lower-power radar applications.

![Fig. II-G7. Autocorrelation of an optimized frequency-coded (FC) waveform before and after distortion by a saturated power amplifier](image-url)
Fig. II-G8. Amplitude envelope of an optimized frequency-coded (FC) waveform before and after distortion by a saturated power amplifier

A final practical consideration for all waveforms occurs when performing pulse compression (receiver matched filtering) digitally. In this case, after anti-aliasing filtering and A/D conversion, the continuous baseband received signal from (II-C2) can be expressed in discrete notation as

\[ y(n) = x^T(n)s + v(n), \]  \hspace{1cm} (II-G1)

where the length-\(N\) vector \(s = [s_1, s_2, \ldots, s_{N-1}]^T\) is the discretized version of the waveform for \(N \approx BT\), the vector \(x(n) = [x(n), x(n-1), \ldots, x(n-N+1)]^T\) is the collection of \(N\) contiguous samples of the unknown illuminated scattering, \(v(n)\) is a sample of additive noise, \((\cdot)^T\) is the transpose operation, and the influence of Doppler shift during the pulsewidth is neglected. Collecting \(N\) contiguous samples of \(y(n)\) from (II-G1) to form the vector \(y(n)\), the matched filter response from (II-C3) thus becomes

\[ \hat{s}_{MF}(n) = h^H_{MF} y(n) = C^H s^H y(n), \]  \hspace{1cm} (II-G2)
in which $(\cdot)^{H}$ is the complex-conjugate transpose (Hermitian) operation and the scalar $C$ is again selected to provide unity noise power gain ($\|h_{\text{MF}}\|=1$).

The implication of discretizing the waveform into $N$ samples (for $N \approx BT$ with $B$ the 3 dB bandwidth) is that the receiver sampling rate is likewise defined according to the waveform 3 dB bandwidth. While the 3 dB bandwidth is intrinsically related to the range resolution of the matched filter, which is itself generally measured by the autocorrelation 3 dB mainlobe width, the actual spectral content of the waveform and subsequent echo are considerably greater. If one performs (complex) receiver sampling at a rate corresponding to the 3 dB bandwidth, which is likewise approximately the chip rate for phase codes, then there is a possibility that the relative delay of a reflected echo may be offset by an amount that introduces a mismatch loss when applying the matched filter. In other words, for sampling period $T_s$ the received reflected echoes may arrive with a delay offset of as much as $\pm 0.5T_s$ relative to the sampled structured of the matched filter. Generally referred to as range straddling or scalloping, this effect occurs because the received signal and the associated matched filter (obtained from the waveform) are under-sampled according to Nyquist. Thus there exists a continuum of possible delay-shifted versions of the waveform, some of which differing enough from $h_{\text{MF}} = Cs$ that an appreciable loss occurs (as much as a couple dB) [106]. Further, for a transmitter-distorted phase code, the presence of abrupt phase changes (even after transmitter bandlimiting) means there are certain delay-shifted versions coinciding with these transitions that may be considerably different from the nominal versions that are well-matched to the code (since phase is constant during a chip for a phase code).

On one hand, there is a rather simple way to minimize the mismatch loss due to range straddling: use a higher receiver sampling rate. In doing so, (II-G1) does not change aside from the discretized version of the waveform $s = [s_1, s_2, \ldots, s_{NK-1}]^T$ and the over-sampled collection of
scatterers $x(n) = [x(n) x(n-1) \cdots x(n-NK+1)]^T$, which are both now length $NK$. Likewise, $NK$ samples of the over-sampled received signal $y(n)$ are now collected to form $y(n)$ for subsequent application of the (now) length-$NK$ matched filter $h_{MF}$, which is still normalized such that $\|h_{MF}\| = 1$.

The trade-off for operating at a higher rate is an increase in the computational cost to perform pulse compression, which already tends to be a bottleneck due to the need to process a large amount of data rapidly (usually in real-time). This computational burden may be alleviated somewhat by performing pulse compression filtering in the frequency domain [107, Chap. 7], which is already commonly done. Continued improvements in computing speed are also helping to ease this bottleneck.

\( H) \quad \text{Mismatched Filtering} \)

As discussed in Sect. II-B, an amplitude-tapered version of LFM can significantly reduce range sidelobes at the cost of degraded range resolution and SNR loss. Further, the need for amplitude control prevents the transmitter from operating in saturation, thereby inducing additional SNR loss relative to what could be achieved if the HPA were operated in saturation. Besides NLFM, another alternative is to transmit an untapered LFM with a receive filter that is different from the matched filter, that is, a mismatched filter (MMF).

The simplest MMF involves tapering of the matched filter as

$$h_{MMF}(t) = a(t) h_{MF}(t). \quad \text{(II-H1)}$$

For example, where the response shown in Fig. II-B1 involves a square-root Hamming-weighted LFM and associated matched filter (which thus also contains a square-root Hamming taper), Fig. II-H1 depicts the response of a Hamming-weighted MMF via (II-H1) to an untapered LFM waveform. Where the former distributes the weighting equally over the waveform and filter via the square-root, the latter employs the entire weighting only at the receive filter. The MMF yields
the same resolution degradation factor of 1.5 relative to the use of untapered LFM with a matched filter. The largest sidelobe is now about $-36$ dB while the SNR loss is about 1.4 dB.

![Fig. II-H1. LFM pulse compression response for the matched filter (black) and Hamming-weighted mismatched filter (red)](image)

While the tapered LFM with matched filtering and untapered LFM with mismatched filtering exhibit SNR losses of 2.7 dB and 1.4 dB, respectively, these losses are for different reasons. The amplitude-tapered LFM clearly exhibits SNR loss due to the deviation from a constant-amplitude waveform (see (II-B5)), yet the subsequent (normalized) matched filter still maximizes the received SNR (with unity noise power gain). In contrast, untapered LFM maximizes the transmit power while the subsequent receive MMF accepts a mismatch loss as a trade-off for lower sidelobes. By likewise normalizing the MMF to produce unity noise power gain via scaling such that $\left(\int_0^T |h_{\text{MMF}}(t)|^2 \, dt\right)^{1/2} = 1$ (or $\|h_{\text{MMF}}\| = 1$ from a digital perspective), one can surmise that the mismatch loss, as the name suggests, is a result of the filter not being exactly matched to the waveform so that the received echo signals do not experience the maximum coherent processing gain provided by the waveform time-bandwidth product.

Generally speaking, the loss in SNR due to mismatched filtering is
\[
\text{SNR Loss}_{\text{mismatch}} = -10 \log_{10} \left( \frac{\text{SNR}_{\text{MMF}}}{\text{SNR}_{\text{MF}}} \right), \quad (\text{II-H1})
\]

where \( \text{SNR}_{\text{MMF}} \) and \( \text{SNR}_{\text{MF}} \) are defined at the mainlobe peaks of the respective filter responses and, if the filters are normalized to produce unity noise power gain as discussed above and in Section II-C, the noise power terms cancel out such that the ratio is between achievable signal powers after filtering. For the amplitude-weighted MMF, assuming the weighting is a real function, the mismatch loss can be expressed as

\[
\text{SNR Loss}_{\text{mismatch (weighted)}} = -10 \log_{10} \left( \frac{\int_0^T a(t) \, dt}{\int_0^T a^2(t) \, dt} \right), \quad (\text{II-H2})
\]

Likewise, the discretized representation of (II-H2) is

\[
\text{SNR Loss}_{\text{mismatch (weighted)}} = -10 \log_{10} \left( \frac{\sum_{n=1}^{NK} a(n)^2}{NK \sum_{n=1}^{NK} a^2(n)} \right), \quad (\text{II-H3})
\]

for \( N \approx BT \) and the receive over-sampling factor \( K \) relative to 3 dB bandwidth. While (II-H2) and (II-H3) provide a way to compute the mismatch loss for the specific form of MMF based on amplitude tapering of the matched filter, (II-H1) is the more general formulation that is useful for all manner of mismatched filtering.

Another prominent MMF instantiation arises from least squares estimation [108]. For arbitrary receive over-sampling \( K \) and filter-length increase-factor \( b \) (typically on the order of 2 to 4), the length \( bNK \) least squares (LS) MMF formulation is posed as

\[
\mathbf{A}_b \mathbf{h}_{\text{LS-MMF}} = \mathbf{e}_m, \quad (\text{II-H4})
\]

where \( \mathbf{e}_m \) is the length \((b+1)NK - 1\) elementary vector with a 1 in the \( m \)th element and zero elsewhere and
\[
\mathbf{A} = \begin{bmatrix}
    s_1 & 0 & \cdots & 0 \\
    \vdots & s_1 & \ddots & \vdots \\
    s_{NK} & \vdots & \ddots & 0 \\
    0 & s_{NK} & \cdots & s_1 \\
    \vdots & \ddots & \ddots & \vdots \\
    0 & \cdots & 0 & s_{NK}
\end{bmatrix}
\]  

(II-H5)

is a \((b+1)NK-1) \times bNK\) Toeplitz matrix. If \(K=1\) (not over-sampled), the optimal MMF in the
LS sense is thus

\[
\mathbf{h}_{\text{LS-MMF}} = \left(\mathbf{A}^H \mathbf{A} + \epsilon \mathbf{I}\right)^{-1} \mathbf{A}^H \mathbf{e}_m, 
\]

where the diagonal loading term \(\epsilon \mathbf{I}\), for real and positive \(\epsilon\) and identity matrix \(\mathbf{I}\), has been added to the LS solution to provide further control over MMF performance. Once determined, the filter is subsequently scaled such that \(\|\mathbf{h}_{\text{LS-MMF}}\| = 1\).

If \(\epsilon = 0\) the true LS MMF is obtained, though the resulting mismatch loss determined via (II-H1), which is waveform dependent, may be unacceptable. In contrast, if \(\delta\) is made large, the LS MMF in (II-H6) effectively becomes a scaled version of the matched filter with surrounding zeros. While the matched filter may not provide acceptable sidelobe performance, it yields no mismatch loss, straddling effects notwithstanding. Thus the \(\epsilon \mathbf{I}\) term enables determination of an acceptable trade-off between sidelobe reduction and mismatch loss.

If the received signal is over-sampled \((K > 1)\) to combat mismatch loss from range straddling, the LS MMF in (II-H6) produces a super-resolution condition that, while yielding a narrower mainlobe for the pulse compression filter response, also suffers from considerable mismatch loss (several dB) and increased sidelobes [109]. This effect can be remediated by replacing \(\mathbf{A}\) with \(\tilde{\mathbf{A}}\), for which some number of rows above and below the \(m\)th row are replaced with zeros to provide a “beam-spoiling” effect. The precise number of zeroed rows to achieve the nominal resolution (same as the matched filter) depends on the waveform and the value of \(K\) [110].
Besides mismatch loss, an additional effect arises when range straddling occurs for the LS MMF. Because this filtering scheme is constructed from the waveform via (II-H5) and (II-H6) to suppress sidelobe to the greatest degree possible (for a given acceptable mismatch loss and resolution), the LS MMF is particularly sensitive to model mismatch effects such as occurs in a range straddling condition. When the sampled version of the received waveform differs from the version used to construct the LS MMF, the degree of sidelobe suppression is hindered. Thus there is a need to continue exploring MMF robustness measures such as the filter averaging approach considered in [110].

In addition to the LS MMF above, which is based on minimization of the $L_2$ norm, many different MMF formulations have also been developed. These approaches include the use of different $L_p$ norms with convex optimization [111-114], iterative reweighting of least squares [115,116], inverse filtering [117], the two-sample sliding window adder [118,119], linear programming [120], minimax optimization [121,122], and even alternative signal representations such as the Laurent decomposition of the waveform [123].

Adaptive forms of MMF have also been developed, in which the pulse compression response from the initial matched/mismatched filtering is used as prior knowledge to enable further sidelobe suppression. The earliest of these approaches [124], which eventually became commonly known as the CLEAN algorithm [125,126], sequentially subtracts the estimated sidelobe responses generated by large scatterers. A more recent approach, Adaptive Pulse Compression (APC) [127], performs adaptive nulling in the range domain by using the current estimate of the measured pulse compression response to generate an updated adaptive filter specific to each particular range cell. Subsequent variants of APC address fast-time Doppler [128-130], pulse eclipsing [97,131], post matched filter processing [132,133], the application to FM waveforms [110], and multistatic [134,135] and dual-polarized [136] operation. More computationally efficient versions have likewise been developed [137,138].
\section*{1) Bandwidth Considerations}

One means for characterizing a signal, device, or system in a meaningful way is to use some measure of the bandwidth $B$ of the spectral density of its transfer function to define a categorization scheme. Consequently, defining a signal, device, or system unambiguously is a two-step process: (1) clearly specify the notion of bandwidth, and (2) categorize the signal, device, or system in terms of its bandwidth. This process begs the questions of what is bandwidth and what is an appropriate categorization scheme? The answers to these questions are neither obvious nor unambiguous [139,140], as numerous definitions of bandwidth exist in the literature and various standards [9,141,142], and these definitions are influenced by differing needs and viewpoints of communities of interest (radar, communication, directed energy, electromagnetic interaction, high-power EM, etc.). Even though many of these communities are related, no codified definition across them exists. So when using the term bandwidth, the user should clearly define what is meant and how the bounding frequencies are selected. In addition, because well-known standard definitions for narrowband signals and hardware either do not apply or are not easily extendable to ultrawideband (UWB) signals and hardware, it is imperative that the meaning of bandwidth be clearly stated and well formulated. In fact, issues with understanding and classifying UWB short-pulse and signals and devices for radar and communication applications led to the categorization schemes by the US Office of the Secretary of Defense / Defense Advanced Research Projects Agency (OSD/DARPA) Panel [143], the International Electrotechnical Commission (IEC) [144], and the US Federal Communications Commission (FCC) [145].

Generally speaking, there are three common ways in which to measure bandwidth (RMS, power-level, and energy-level), albeit with many different variations thereof. For the time-domain signal $s(t)$ with finite energy and it’s frequency-domain representation $S(f)$ determined by the Fourier transform pair
these bandwidth measures are as follows.

In radar signal theory, the RMS bandwidth $B_{\text{RMS}}$ is often used. The most recent IEEE Standard 686-2008 [9] now defines $B_{\text{RMS}}$ according to [146, Chap. 2] as the 2nd moment of the square magnitude of $S(f)$ about a designated frequency. Specifically,

$$
B_{\text{RMS}} = \sqrt{\frac{\int_{0}^{\infty} [2\pi(f - f_{mp})]^2 S(f)^2 df}{\int_{0}^{\infty} S(f)^2 df}}.
$$

where the denominator is half of the signal energy and the mean frequency $f_{mp}$ over positive frequencies is given by

$$
f_{mp} = \frac{\int_{0}^{\infty} f S(f)^2 df}{\int_{0}^{\infty} S(f)^2 df}.
$$

Relative to $f_{mp}$, the high end of the band is $(f_{mp} + 0.5B_{\text{RMS}})$, while the low end of the band is set as $\max\{0, (f_{mp} - 0.5B_{\text{RMS}})\}$ because $(f_{mp} - 0.5B_{\text{RMS}})$ could be negative. For well-behaved spectra, $f_{mp}$ is usually very near to the frequency associated with the maximum value of the energy density, which is usually the carrier frequency for radiated narrowband signals.

The $X$ dB power-level bandwidth $B_{\text{X,db}}$ is [139]

$$
B_{\text{X,db}} = f_{\text{high}} - f_{\text{low}},
$$

where the lowest $f_{\text{low}}$ and highest $f_{\text{high}}$ frequencies are solutions of

$$
20\log_{10}|S(f)| = 20\log_{10}|S(f_{\text{max}})| - X,
$$
for positive $X$, where $f_{\text{max}}$ is the frequency at which the power spectral density $|S(f)|^2$ achieves its maximum value. For example, $B_{3\text{dB}}$ corresponds to values of $f$ at which $|S(f)|^2$ is half its maximum value. Power-level bandwidths are used in a wide variety of applications. Filter design and control theory traditionally use $B_{3\text{dB}}$, the FCC employs $B_{10\text{dB}}$ to define UWB signals, and the spectrum-management community uses $B_{20\text{dB}}$ and $B_{40\text{dB}}$.

Finally, for each value $X$ in $(0, 1]$, let $A_X$ be the collection of nonnegative pairs $\{f_{\text{low}}, f_{\text{high}}\}$ of real numbers that satisfy

$$
\int_{f_{\text{low}}}^{f_{\text{high}}} S(f)^2 \, df = X \int_0^\infty S(f)^2 \, df.
$$

The $X$ fractional energy bandwidth is thus $[139,140]$

$$
B_{X\text{EB}} = \inf \left\{ (f_{\text{high}} - f_{\text{low}}) ; f_{\text{low}}, f_{\text{high}} \text{ in } A_X \right\}.
$$

Although $A_X$ may contain more than a single pair of frequencies, $B_{X\text{EB}}$ is unique. For example, if the spectral magnitude is a rectangular function, the $X$ fractional bandwidth is a single value, even though $A_X$ contains an infinite number of distinct pairs. The fractional energy bandwidth provides good information on how the signal energy is distributed in the frequency domain. This quality makes $B_{X\text{EB}}$ a useful measure for characterizing signals in terms of their spectral occupancy (spectrum management) and their electromagnetic interference on other sources (directed-energy systems and electromagnetic hardening).

Where the above measures provide different definitions of bandwidth, it is likewise useful to classify the nature of a signal/system as narrowband, wideband, or ultrawideband according to its fractional bandwidth, which is defined as

$$
B_f = \frac{(f_{\text{pass,high}} - f_{\text{pass,low}})}{(f_{\text{pass,high}} + f_{\text{pass,low}})/2} \times 100\%.
$$
where \( f_{\text{pass,high}} \) and \( f_{\text{pass,low}} \) denote the upper and lower edges of the passband, respectively. As such, a signal/component/system is categorized in terms of its fractional bandwidth as [140,147]:

- narrowband if \( 0\% \leq B_F \leq 1\% \),
- wideband if \( 1\% \leq B_F \leq 25\% \),
- ultrawideband if \( 25\% \leq B_F \leq 200\% \).

Note that the 1\% demarcation between narrowband and wideband is not used in the IEEE Radar Standard [9] and should be taken as one possible summary of the literature. Some references suggest that a 10-20\% fractional bandwidth could be considered as being effectively narrowband. For example, Engler [148] states that “a typical narrowband signal will have 10\% bandwidth or less” and Urkowitz, et al [149] denotes a signal as narrowband if \( B_F < 20\% \) because in such case the “range and range rate (have) no dependence upon the bandwidth of the transmitted signal”. Likewise, according to Richards [150] “Few radars achieve 10\% bandwidth. Thus most radar waveforms can be considered narrowband, bandpass functions.”

One could also consider the point at which group delay dispersion become noticeable or when VSWR exceeds a specified value. Clearly, there are various different definitions of bandwidth and means of categorizing bandwidth. The take away here is that one must be careful to specify which definition is being used and to remember that the notion of spectral content is more complicated than the statement of a single number.

III. WAVEFORM DIVERSITY

Due to the combination of increasing RF spectrum pressure, an increasingly complex interference environment, and the continued desired for improved radar sensitivity/discrimination capability, research in waveform diversity (WD) has flourished. For this very reason, it is really not feasible to survey all the myriad developments. Instead, we shall take a general view of the different types of WD, with a focus on the practical problems and attributes. Specifically, while
one could argue that WD research is largely arising from a signal processing / waveform design perspective, the RF system and electromagnetics effects play crucially important roles in what is physically achievable. These effects become particularly important when considering the impact of coupling between the various dimensions of fast-time (range), slow-time (Doppler), space, polarization, and coding (modulation). When operating frequency is included, this set has been referred to as the *transmission hypercube* or *transmission hyperspace* [151].

A) *Practical Waveform Optimization*

The properties of a waveform that are the most conducive to its emission from a radar are 1) constant amplitude and 2) sufficient spectral containment. The former helps to avoid some of the nonlinear distortion that would otherwise be imparted to AM waveforms by the high power amplifier (HPA) and facilitates maximization of power-added efficiency (PAE) and subsequent “energy on target” for detection sensitivity. The latter property helps to minimize the spectral shaping imposed by the transmitter that can produce additional AM effects leading into the HPA, subsequently compounding distortion and potentially creating additional problems (see Section II-G).

As discussed in Section II-B, FM waveforms are attractive because they are constant amplitude and inherently well-contained spectrally, thus making them amenable to a physical radar transmitter, particularly the distortion induced by the HPA. However, binary codes have also been widely used, due in large part to the existence of implementation schemes through which the code structure can be converted into a transmitter-appropriate physical waveform. The two most common implementation schemes are derivative phase shift keying (DPSK) [152] and biphase-to-quadriphase (BTQ) transformation [153], the latter being a form of minimum shift keying (MSK). For example, specifying $s_{\text{BC}}(t)$ as the binary coded version of (III-B6) with $\theta = 0^\circ$ or $180^\circ$, the resulting DPSK-implemented waveform can be expressed as [152]
\[ s_{\text{DPSK}}(t) = s_{\text{BC}}(t - T_c/2)[\cos(\pi t / T_c) - js_{\text{BC}}(t)\sin(\pi t / T_c)], \]  

(III-A1)

thereby ensuring that the phase is continuous by avoiding the abrupt chip transitions (see DPSK implementation of a length-5 Barker code in Fig. III-A1). Likewise, the BTQ transformation \[153\] causes any transition from 0° to 180°, or vice-versa, first to transition to ±90°, thus forcing the phase to traverse the unit circle instead of going through its center (such as we observed in Figs. II-G5 and II-G6). While widely used, the main limitation for binary-coded waveforms is a lack of design freedom due to the \( P = 2 \) phase constellation.

![Phase trajectory](image)

**Fig. III-A1. Phase trajectory of a binary code (ideal) and its DPSK implementation**

Just as DPSK or MSK can be used to implement binary codes, it has recently been shown that arbitrary polyphase codes can likewise be implemented using a modified form \[103,154\] of continuous phase modulation (CPM) \[155\] that is otherwise commonly employed in aeronautical telemetry \[156\], deep space communications \[157\], and the Bluetooth™ wireless standard \[158\]. The resulting waveform is actually a form of FM and thus is denoted as polyphase-coded FM (PCFM). For this formulation, a train of \( N \) impulses is formed that have time separation \( T_p \) and thus a total time support of \( T = NT_p \). The \( n \)th impulse is weighted by \( -\pi \leq \alpha_n \leq \pi \), which is the phase change occurring over a \( T_p \) interval, and thus can be viewed as a discretized representation of the instantaneous frequency in (II-B2)). From a design standpoint, it is possible either to determine the \( \alpha_n \) values directly or to obtain them from a standard length \( N + 1 \) polyphase code via
\[
\alpha_n = \begin{cases} 
\tilde{\alpha}_n & \text{if } |\tilde{\alpha}_n| \leq \pi \\
\tilde{\alpha}_n - 2\pi \text{sgn}(\tilde{\alpha}_n) & \text{if } |\tilde{\alpha}_n| > \pi
\end{cases}
\] (III-A2)

where

\[
\tilde{\alpha}_n = \theta_n - \theta_{n-1} \quad \text{for} \quad n = 1, \ldots, N
\] (III-A3)

sgn(*) is the sign operation, and \(\theta_n\) is the phase value of the \(n^{th}\) chip in the length \(N+1\) polyphase code.

Given the phase-change code \(x = [\alpha_1 \alpha_2 \cdots \alpha_N]^T\) and arbitrary starting phase \(\theta_0\), the resulting PCFM waveform is generated as [103]

\[
s_{\text{PCFM}}(t; x) = \exp\left\{ j \int_0^t g(\tau) \ast \left[ \sum_{n=1}^N \alpha_n \delta(\tau - (n-1)T_p) \right] d\tau + \theta_0 \right\},
\] (III-A4)

where the shaping filter \(g(t)\) must integrate to unity over the real line and have time support on \([0, T_p]\) and \(\ast\) denotes convolution. For example, a rectangular filter meets these requirements and, upon inclusion in (III-A4), serves as a linear interpolation of phase that can be viewed as a first-order hold representation of the phase function. By comparison, the standard phase-code structure of (II-B6) can be viewed as a zero-order hold representation, since the phase is constant between the abrupt transitions. Figure III-A2 depicts an optimized PCFM waveform from [95] that has a time-bandwidth product of 64. Compared to the LPM bound from (II-F2), for which the PSL value can be computed to be \(-39.1\) dB, this optimized FM waveform realizes an improved PSL of \(-40.2\) dB.
Fig. III-A2. Autocorrelation of an optimized PCFM waveform with $BT = 64$

Noting that the phase component of the first-order representation of (III-A4) can be written as

$$\theta_{1st}(t; x_1) = \int \sum_{n=1}^{N} \alpha_n g_1(\tau - (n-1)T_p) \, d\tau + \theta_0,$$

(III-A5)

where the notation $g_1(t)$ and $x_1 = [\alpha_1 \, \alpha_2 \, \cdots \, \alpha_N]^T$ are used to explicitly denote this shaping filter and phase-change code as corresponding to first-order, higher-order phase functions can also be defined [159]. For example, a second-order coded representation can be defined as

$$\theta_{2nd}(t; x_2) = \int \int \sum_{n=1}^{N} b_n g_2(\tau' - (n-1)T_p) \, d\tau' d\tau + \int \theta_0 \, d\tau + \theta_0,$$

(III-A6)

and likewise a third-order coded representation as

$$\theta_{3rd}(t; x_3) = \int \int \int \sum_{n=1}^{N} c_n g_3(\tau'' - (n-1)T_p) \, d\tau'' d\tau' d\tau + \int \beta_0 \, d\tau' d\tau + \int \omega_0 \, d\tau + \theta_0,$$

(III-A7)

and so on for higher orders, where $x_2 = [b_1 \, b_2 \, \cdots \, b_N]^T$ and $x_3 = [c_1 \, c_2 \, \cdots \, c_N]^T$ are therefore frequency-change (chirp rate) and chirp-rate-change (“chirp acceleration”) codes, respectively, with associated shaping filters $g_2(t)$ and $g_3(t)$. Also, $\theta_0$ is the starting phase, and $\omega_0$ and $\beta_0$
are the starting frequency and chirp rate, respectively. These coding structures in (III-A5), (III-A6), and/or (III-A7) may even be combined [159] to permit multi-order coding for even greater freedom in FM waveform design. This increased freedom means more ways in which to represent the continuum of possible phase trajectories, thus enabling the potential to obtain waveforms whose pulse compression response yields even lower sidelobes for the zero (or at least small) Doppler regime of the ambiguity function. For example, again using a time-bandwidth product of 64, Fig. III-A3 depicts the autocorrelations of waveforms obtained via joint optimization of the first- and second-order components as well as the first-, second-, and third-order components. These waveforms realize PSL values of $-48.4 \text{ dB}$ and $-48.7 \text{ dB}$, respectively.

![Fig. III-A3. Autocorrelation of optimized higher-order PCFM waveforms for $BT = 64$](image)

Besides higher-order phase functions, the polyphase-coded FM implementation of (III-A4) can also be expanded to accommodate what has been referred to as over-coding [160]. In the over-coded formulation, 1) the phase-change intervals of $T_p$ are subdivided into smaller intervals, and 2) the amount of phase change over the interval of $T_p$, which in (III-A2) was limited to $|\alpha_n| \leq \pi$ due to extraction from traditional polyphase coding, is now allowed to exceed this limit.
as long as the aggregate spectral containment is maintained. Thus even more different continuous phase functions may be realized, thereby enabling waveforms such as the one demonstrated in Fig. III-A4, in which a PSL value of $-52.0$ dB is attained, again for a time-bandwidth product of 64.

![Graph of Autocorrelation](image)

**Fig. III-A4. Autocorrelation of an optimized over-coded PCFM waveform for $BT = 64$**

One may also consider how structures such as these higher-order and over-coding formulations could be combined, potentially to yield even greater sidelobe reduction. Further, because the continuum of possible continuous phase functions supports a theoretically infinite number of possibilities, there are certainly other coding implementation structures that could be developed. Other examples include the recent design of FM waveforms based on the use of Bézier curves [161], polynomial function design [58] (which inspired the higher-order form above), the Zak transform [57], and various forms of piecewise NLFM [62]. Also, recent work on hybrid FM (amplitude-tapered NLFM) discussed in Section II-B has experimentally demonstrated a PSL better than $-83$ dB ($-108$ dB in simulation) with only a quarter dB of SNR loss [162].
Clearly there is significant room for improvement in operational systems with regard to sidelobe-limited sensitivity.

These various forms of FM waveforms, along with DPSK/MSK implemented binary codes, provide different ways to parameterize a continuous, constant-amplitude waveform that is amenable to a high-power radar transmitter. However, one should not infer that such waveforms experience no distortion at all. If the goal is to achieve very low sidelobes that are many 10s of dB below the mainlobe peak, then even a small degree of distortion can become the limiting factor on performance. Consequently, it becomes necessary to consider the impact of the transmitter on the generation of the emitted waveform. In so doing, we introduce a stratified nomenclature in which the code (if one exists) comprises a discrete set of parameters that, via some subsequent implementation scheme (such as those discussed above), then realizes the waveform that is subsequently injected into the transmitter for amplification, thereby ultimately producing the physical emission that is launched into the environment.

With the HPA generally representing the most significant source of transmitter distortion due to its inherent nonlinearity, many different linearizing transmit architectures have been developed, including the Kahn technique, envelope tracking, various outphasing methods, the Doherty technique, etc. (see [163] for a review of such methods). Likewise, predistortion techniques (see [164] for an overview) rely upon a variety of models such as a look-up-table (LUT), Volterra model, polynomial model, Wiener model, Hammerstein model, and variations thereof to parameterize and subsequently estimate the nonlinear nature of the HPA so as to undo such effects upon the waveform. Collectively, all these approaches seek to avoid the top scenario in Fig. III-A5 in favor of the bottom scenario in which the actual radar emission is a close approximation to the intended waveform.
An alternative perspective was recently proposed in [95]. By denoting $s(t;\mathbf{x}) = T_{C2W}\{\mathbf{x}\}$ as some arbitrary code-to-waveform implementation operation followed by the operation $u(t;\mathbf{x}) = T_{Tx}[s(t;\mathbf{x})]$ that represents the distortion imposed by the transmitter, a holistic waveform design formulation can be posed as shown in Fig. III-A6, in which $\Phi[u(t;\mathbf{x})]$ corresponds to the application of some metric such as those described in Sect. II-F. As opposed to the linearization approaches above that seek to compensate for transmitter distortion, this “transmitter-in-the-loop” paradigm instead seeks to optimize the final emission inclusive of the transmitter distortion effects. These distortion effects could leverage known mathematical models for the transmitter via a Model-in-the-Loop (MiLo) framework or by directly using the actual radar system via a Hardware-in-the-Loop (HiLo) framework. The trade-off between these is much faster convergence for the former and greater accuracy for the latter. Some form of hybridization of MiLo and HiLo would likewise yield the best speed vs accuracy trade-off in practice.
An interesting feature of the transmitter-in-the-loop design paradigm is that the linearization methods discussed previously can still be incorporated into the transmitter architecture as a means of facilitating greater design freedom. This notion of joint transmitter/waveform optimization, which was inspired by observations of sensing performance by dolphins despite their “mediocre equipment” [24], has been suggested [165] as a promising direction to explore in order to address the expected continued erosion of radar spectrum [5] combined with increased “network densification” of interferers expected from future wireless systems [166]. Leveraging previous “spectrally clean” emission schemes [152], such joint design approaches have already begun to emerge [167,168]. Specifically, [168] proposes the Smith Tube concept as an extension to the well-known Smith Chart used in RF systems engineering, whereby the vertical component (making it a tube) can be some other optimizable parameter such as waveform bandwidth. Moreover, Fellows et al. envision further extension to a veritable Smith Hyper-Tube comprised of multiple optimizable waveform parameters while maintaining transmitter power-added efficiency.

Finally, given the high-dimensional solution space for waveform design, further complicated if one considers the emission induced by transmitter distortion like that in Fig. III-A6, one can surmise that numerous local minima exist. Though it is not necessary to determine the global minimum as long as a predetermined performance specification is met (multiple “good enough” solutions could actually be beneficial from an operational flexibility perspective), the
determination of the sufficiently good waveform(s) may still be a challenge. Thus, while myriad different search strategies exist that one could take [169], some general observations about waveform design are useful to consider: 1) per (II-D2) the delay-Doppler ambiguity function integrates to a constant, thereby establishing a *conservation of ambiguity*; 2) as depicted in Fig. II-D1 for LFM, the ambiguity function for a chirp-like waveform exhibits a delay-Doppler ridge so that, by using the previous observation, a significant portion of the total ambiguity is already “absorbed”; and 3) metrics such as PSL (II-F1) and ISL (II-F3), and even PSD-based metrics such as FTE (II-F4), are complementary measures of the same delay-Doppler ambiguity function. Based on these observations, the recently emerged *performance diversity* paradigm [95] uses an LFM signal as an initialization to start with a well-consolidated ambiguity ridge and then alternates between different metrics during a greedy search to help avoid local minima, since each complementary metric still exhibits a different performance surface. One could even consider various combinations of these metrics to provide even more different performance surfaces upon which to search. The reader is referred to [95,103] as a starting point for further reading on the optimization of physical waveforms.

**B) Environment-Specific Waveforms**

On the one hand, the RF environment in which radar operates continues to become more congested, which competes with the radar community’s sustained need for ever-better detection, discrimination, and tracking. Thus, where Sect. IV-A discussed the optimization of physically realizable waveforms in a general context, this section considers the impact of the radar environment on waveform design.

In Section II-F it was discussed how the power spectral density (PSD), due to its Fourier relationship with the autocorrelation, is useful for waveform design. For example, one could determine a desired autocorrelation, determine the associated PSD, and then optimize a waveform to match that PSD (e.g. the FTE metric defined in (II-F4)). It is known for NLFM waveform
design that to achieve low range sidelobes the signal spectrum should decrease towards the band edges [59]. However, growing spectral congestion driven by the demands for commercial cellular [5] has motivated research into how radar and communications could share spectrum [170-173]. For example, one could insert notches into the radar spectrum to avoid other in-band/near-band spectrum users, both as a means to facilitate more efficient use of the spectrum via sharing [174] and to remediate the associated degradation to radar performance [175]. Doing so in a manner that involves listening to the spectral environment [176] and modifying one’s emissions accordingly [177] is considered a form of cognitive sensing [178,179]. It is important to note that because they collectively operate in the congested HF, VHF, and UHF bands, the modalities of over-the-horizon (OTH) radar [148], foliage penetration (FOPEN) radar [180], ground penetrating radar (GPR) [181], and urban sensing [89] have been already been contending with this problem for quite some time. A survey of the challenges of radar spectrum engineering can be found in [5,165].

An early approach to avoid other in-band spectrum users incorporates notches into swept-frequency waveforms [182]. The same could be achieved for a stepped-frequency waveform, for which the center frequency of the \( m \)th pulse in the CPI is incremented by \( m\Delta f \) (described in Sect. III-E), either by skipping the pulses for which the associated frequencies are to be avoided [180, Chap. 5] or using an additional within-pulse phase coding to “thin the spectrum” [183]. Many subsequent phase-code and NLFM “sparse frequency” waveform design approaches have been developed [184-193].

An important practical aspect involved with designing waveforms that possess spectral notches is whether the waveform remains constant amplitude for injection into an HPA, or at least to what degree the AM effects are minimized if linear amplification is feasible for the sensing application. Further, because ambiguity is conserved, the presence of in-band spectral notches tends to translate into a broadening of the out-of-band spectrum if constant amplitude is preserved. As spectral congestion continues to grow, one must also be cognizant of the increase
in range sidelobes that is incurred as the penalty for spectral notching [194], as well as the prospect of “power struggles” [195] as cognitive systems attempt to outmaneuver one another.

Another form of environment-specific waveform arises when one considers how to design a waveform to emphasize known (or at least hypothesized) attributes of a desired target. In [196], Bell applied information theory to formulate waveform designs that rely on presumed knowledge of a target’s impulse response either to maximize probability of detection or to maximize the amount of information gleaned from the target response. Such signals are referred to as *matched illumination* as coined by Gjessing [197,198]. Since then, considerable work has appeared (e.g. [199-210]) exploring the ways in which the radar could perform the alternating processing of 1) observing the environment with a given waveform and then 2) reformulating a new waveform to capture/enhance some additional salient feature of the environment. For example, successive refinement of the waveform may permit better discrimination between different classes of targets or between targets and the ambient clutter. Because it relies on this “query and revise” strategy, the concept of *time reversal* has also been investigated for this problem (e.g. [211-213]), albeit with a cautionary note on the electromagnetics provided in [214]. Regardless of the specific approach, this notion of *adaptive waveform design* can be viewed as a form of cognitive sensing [178,179]. Of course, such waveforms must still adhere to the physical requirements imposed by the radar transmitter, antenna included, as discussed in Sections II-G and III-A. See [178,179] as a starting point for further reading on cognitive sensing and adaptive waveform design.

The waveforms used for *nonlinear harmonic radar* [215-220] generally require particular consideration of their spectral containment to enable adequate discrimination between nonlinear and (typically far stronger) linear scattering. For example, because electronics typically contain diodes and transistors that can produce such a nonlinear response, this form of radar is considered a means to detect, and perhaps even to discriminate, electronics in the illuminated environment. In principle, if one can generate a pure sinusoid at frequency $f_0$, then a nonlinear response would occur at $2f_0$ and higher integer multiples. The difficulty is that this harmonic response tends to be
orders-of-magnitude smaller than the linear response [221], and could be masked by harmonics generated by the transmitter if the emission does not possess sufficient spectral purity [218]. This form of radar has yielded some rather interesting applications, such as a nonlinear junction detector to enable counter-surveillance by sweeping for listening devices [220] and the tracking of insects using RF tags comprised of Schottky diodes [216].

The related concept of stimulated emissions relies on the fact that RF receivers, such as those in cell phones, produce an identifiable signal when illuminated by an appropriate stimulation signal [222]. In fact, one can even induce an intermodulation effect within the nonlinear device by using multiple signals to produce a desired emitted waveform [217,223]. For example, higher-order intermodulation products are produced when a nonlinear device is simultaneously interrogated with tones having frequencies of $f_1$ and $f_2$. These and higher-order mixing products facilitate the “fingerprinting” of different commercial RF devices [217]. From an operational sensing standpoint, the detection/identification of unknown electronic devices could therefore follow the sequential interrogation paradigm discussed above for adaptive waveform design, albeit with the inclusion of the nonlinear response [224]. A survey of recent work in this area can be found in [217].

Finally, the clutter response generated by the radar may provide the spectral environment in which other signals could reside. This notion of radar-embedded communication involves the generation of either inter-pulse [225-228] or intra-pulse [229-232] signals that are designed to be embedded within ambient radar scattering by RF transponders/tags as a means to self-identify friendly targets (“blue force tracking”), to enable environmental monitoring, or to provide a covert communication link. Generally speaking, the inter-pulse form [225-228] encodes information into the radar backscatter via modulation on a pulse-to-pulse basis so that the communication signal resides in the slow-time (Doppler) domain. This form is well tested, with Sandia’s Athena tag being a notable example [233]. The data rate, however, is on the order of
bits-per-CPI, which is quite low for any useful communication mode aside from self-
identification.

In contrast, the intra-pulse form [229-232] involves the determination of a set of $K = 2^{\#\text{bits}}$ communication symbols $\{c_1(t), c_2(t), \cdots, c_K(t)\}$ that have minimal mutual cross-correlation, yet have commensurate correlation with the ambient scattering produced by the radar waveform $s(t)$. The former requirement maximizes the separability of the symbols on receive, while the latter mitigates receiver bias that could be generated by the radar clutter. This form of radar-embedded communication, while still at the theoretical stage, enables a data rate of bits-per-pulse, thus scaling with PRF such that data rates commensurate with speech may be possible. Starting points for further reading on inter-pulse and intra-pulse radar-embedded communication are [226] and [229], respectively.

C) Colocated MIMO Radar

The category of waveform diversity denoted as multiple-input multiple-output (MIMO) has arguably received more attention than all the rest combined based on the sheer volume of publications. Inspired by the capacity and performance gains enabled by spatial diversity for MIMO communications [234,235], the early notions of MIMO radar [236-238] sought to generalize the prior concept of ubiquitous radar [239], in which the transmitter illuminates a wide spatial beam combined with multiple narrow receive beams, to facilitate spatial diversity on transmit. The terminology was also separately used to refer to joint operation of multiple, widely separated transmitters and receivers [240]. Many of these ideas built on even earlier work such as multiple simultaneous transmit beams for phased arrays [241], that was realized on the AMRFC test bed [242], and spatio-temporal coding for radar array processing [243], which were themselves predated by the French RIAS system [244-246]. The latter was in fact experimentally
performing “transmit beamforming on receive” well before the term MIMO was even used in the radar context.

A collection of much of the early theoretical work on MIMO radar was compiled in [247]. Today, MIMO radar is generally referred to as belonging to one of two types: colocated MIMO in which phase coherence can be assumed; and distributed (or statistical) MIMO in which phase coherence typically cannot be assumed, though there are exceptions [248,249]. The latter can be treated under the umbrella of the more traditional nomenclature of multistatic radar and thus will be discussed in Section III-D, while we shall focus here on the notion of colocated MIMO.

Generally speaking, colocated MIMO involves the generation of different waveforms from different antenna elements (or sub-arrays of elements [250-252]). For example, given a linear array of \( L \) elements indexed by \( \ell = 1, 2, \ldots, L \), the set of distinct waveforms \( \{ s_{\ell}(t) \} \) can be defined according to the waveform structures from Section II-B. Likewise for a planar array, the MIMO waveforms could be indexed according to the horizontal and vertical antenna elements, thereby providing a spatially diverse emission structure in both azimuth and elevation angles.

For the ubiquitous MIMO emission of a wide transmit beam, these waveforms should be designed to possess a low cross correlation – the term “orthogonal” has been widely used, but it is really a misnomer for radar if the waveforms have overlapping spectral support, per Parseval’s theorem, since no assumption of synchronicity of radar echoes can be made. This ubiquitous mode could be a means to enable multi-function operation [238,239] if the trade-off between the loss in transmit spatial gain and the increased dwell time is feasible for the given operating parameters [253,254]. Note that there are other means to achieve a multi-functional capability. For example, the Advanced Multifunction RF concept (AMRFC) [242] assigns different functions to different array sub-apertures. Further, enhanced spatial resolution and reduced spatial sidelobes have been demonstrated for MIMO relative to a non-MIMO mode [255]. Of course, sensing modes such as synthetic aperture radar (SAR), in which the transmit beam is intentionally
broad so as to capture the long synthetic aperture, could be inherently well-suited for the wide beamwidth provided by such a MIMO emission [256].

The notion of partially correlated waveforms has also been proposed [251], in which the set of MIMO waveforms lie somewhere between the extremes of fully coherent (identical aside from a subsumed phase shift for beam steering) and completely independent waveforms that facilitate ubiquitous operation. As such, greater design freedom is available to shape the spatio-temporal structure of the radar emission, thereby providing greater control over the trade between mainbeam gain and spatial diversity.

MIMO radar has also been a source of controversy as some radar systems engineers question the validity of some of the theoretical claims of MIMO [253,254,257], going so far as to suggest that it could be “snake oil” [253]. However, there are some clear practical applications of MIMO radar. A case in point is over-the-horizon (OTH) radar [258] in which the transmitted and received signals are reflected off various layers of Earth’s ionosphere in transit each way. Experimental results in [259] have shown that the “transmit beamforming on receive” spatial diversity of MIMO radar, which Frazer et al. refer to as “non-causal transmit beamforming”, provides greater separability of OTH radar echoes due to the inherent range/angle coupling that exists for the skywave propagation channel.

Another way of looking at the MIMO emission structure is via the far-field fast-time signal as a function of spatial angle, which is intuitively attractive because it is this signal that is physically incident upon a scatterer. Consider a uniform linear array with inter-element spacing $d$ and wavelength $\lambda$, in which the $\ell$th antenna element emits the narrowband pulsed waveform $s_{\ell}(t)$ for $0 \leq t \leq T$. The far-field emission (baseband representation) at time $t$ for transmit spatial angle $-90^\circ \leq \phi_T \leq 90^\circ$ can thus be expressed as

$$g(t,\phi_T) = \frac{1}{L} \sum_{\ell=1}^{L} s_{\ell}(t) e^{j(t-\ell)\tilde{\phi}_T},$$

(III-C1)
where \( \phi_T = 2\pi d\sin(\phi_T)/\lambda \) is the transmit electrical angle for \( \phi_T = 0^\circ \) at array boresight.

Integrating (III-C1) over the pulsewidth and dividing by \( T \) therefore yields the aggregate beampattern

\[
B(\phi_T) = \frac{1}{T} \int_0^T g(t, \phi_T) g^*(t, \phi_T) \, dt
\]  

(III-C2)

as a function of transmit spatial angle \( \phi_T \). For the simplified case in which \( s_i(t) = s(t)e^{-j\phi_{\text{look}}} \) and the electrical angle \( \phi_{\text{look}} \) corresponds to an arbitrary spatial “look” direction, (III-C2) is just the standard array factor beampattern [260].

Where the delay-Doppler ambiguity function developed by Woodward (II-D1) describes the matched filter response to different Doppler-shifted versions of a waveform, it is likewise useful to define a delay-angle ambiguity function for MIMO operation. For receive spatial angle \( \phi_R \) and the similarly specified receive electrical angle \( \phi_{\text{look}} \), the unity-gain normalized matched filter can be defined as

\[
h(t, \phi_R) = \frac{g(t, \phi_R)}{\sqrt{\int_0^T g(t, \phi_R) g^*(t, \phi_R) \, dt}}.
\]  

(III-C3)

Thus, the combination of receive beamforming by the \( L \) antenna elements in the direction \( \phi_R \) and associated pulse-compression matched filtering as a function of receive angle realizes the delay-angle ambiguity function

\[
\chi_{\text{delay-angle}}(\tau, \phi_T, \phi_R) = \sqrt{L \int_0^T \left[ \sum_{i=1}^L g(t, \phi_T) e^{j(\tau-i)\phi_{\text{look}}} g^*\left(t - \tau, \phi_R\right) e^{-j(\tau-i)\phi_{\text{look}}} \right] \, dt}
\]  

(III-C4)

that describes the response that a given set of emitted waveforms would produce when reflected by a point scatterer in the environment as a function of angle and relative delay. Using (II-D1),
the delay-angle ambiguity function of (III-C4) readily generalizes to a delay-Doppler-angle ambiguity function given by

\[
\chi_{\text{delay-Doppler-angle}}(\tau, f_D, \phi_T, \phi_R) = \int \sum_{L=1}^{T} e^{i2\pi f_D t} g(t, \phi_T) e^{i(\ell)\hat{\phi}_T} \left( t - \tau, \phi_R \right) e^{-i(\ell)\hat{\phi}_R} \, dt
\]

and likewise incorporates the impact of a coherent pulse train in the same manner as (II-E1). Different forms of the MIMO ambiguity function can be found in [247,261-263], along with more detailed discussion of the associated properties.

The true discrimination capability of the much-increased dimensionality provided by MIMO may also necessitate adaptive receive processing that leverages this high dimensionality, which comes with an additional trade-off that could involve a significant increase in computation cost. However, just as space-time adaptive processing (STAP) [264] enables an interference suppression capability that could not be achieved using adaptive beamforming or Doppler processing alone, MIMO-oriented adaptive processing may likewise enable such new capabilities. For example, it was recently demonstrated experimentally [265,266] that a MIMO formulation could be used to generate a joint space-frequency null on transmit to avoid interfering with other nearby spectrum users.

It is important to note that MIMO waveforms possess the same physical requirements and undergo the same transmitter effects as non-MIMO waveforms per Section II-G, therefore necessitating consideration of practical waveform design as discussed in Section III-A. Further, because the physical MIMO emission inherently depends on the interaction between the set of \( L \) waveforms and the distributed antenna elements in the array, the electromagnetic effects of the array must likewise be considered. For example, mutual coupling among antenna elements, which involves neighboring antenna elements receiving and reradiating the waveform from a given
element, produces a distortion of the far-field delay-angle emission structure relative to the ideal case of no mutual coupling (see Fig. III-C1) [267,268].

![Fig. III-C1. Delay-angle ambiguity function for 16 waveforms generated via DPSK implementation of length-50 random binary codes where (left) no mutual coupling is present and (right) \(-10\) dB nearest neighbor mutual coupling is present but not accounted for on receive. The result is degraded resolution and 1.1 dB mismatch loss.](image)

Another practical impact of MIMO arises when attempting to emit wideband signals over a wide beamwidth, such as desired for SAR [269]. For narrowband operation, the wavelength \(\lambda\) of the center frequency is an adequate approximation over the entire bandwidth. Thus inter-element spacing of \(d = \lambda/2\) is the maximum value that avoids grating lobes (though if the MIMO array takes advantage of the “virtual array” concept to more widely separate the elements for enhanced spatial resolution, phase discontinuity effects must still be considered [270]). In contrast, when the bandwidth becomes sufficiently large that a single wavelength is not a good approximation, one could set \(d = \lambda_{\text{min}}/2\), for the shortest wavelength \(\lambda_{\text{min}}\) corresponding to the highest in-band frequency to avoid grating lobes for all corresponding frequencies in the passband [271]. However, this choice has the undesired effect of yielding inter-element spacing for the longer wavelengths (lower frequencies) such that \(d/\lambda \ll 0.5\), which can result in “emission” of power into the imaginary space (or invisible space) [258] that exists beyond the endfire spatial directions at \(\phi_T = \pm90^\circ\). In fact, this power is not actually emitted as it becomes energy that is stored in the reactive near-field of the array and can lead to large amounts of power being
reflected back into the transmitter, potentially damaging the radar [253]. The implication for MIMO is that waveform design must account for the relative instantaneous phase difference between waveforms on adjacent antenna elements so as to avoid exceeding the boundaries of real space [272].

Fig. III-C2. Wideband frequency content vs. electrical angle for MIMO emission. Element spacing is half-wavelength for the center frequency $f_{\text{cent}}$ (so $d=0.5\lambda_{\text{cent}}$) and bandwidth is 30% relative to the center frequency. The red triangles identify the invisible space. Alternatively, setting $d=0.5\lambda_{\text{min}}$ would raise the location of the red triangles in the figure, so that more of the emission would be into the invisible space.

A notable subset of MIMO is the frequency diverse array (FDA) [273-277], in which all antenna elements emit the same waveform, aside from a small frequency shift that is incremented across the array. In other words, the carrier frequency for the waveform generated by the $\ell$th antenna element is $f_c + (\ell - 1)\Delta f$, where $f_c$ is a nominal carrier frequency and $\Delta f$ is the small frequency increment. As a result, the beamforming look direction varies in fast time and sweeps across space at a rate depending on the value of $\Delta f$. Because it maintains a coherent mainbeam (the location of which changes with time), the FDA can be expected to experience less degradation than arbitrary MIMO as a result of mutual coupling effects. The very relevant concept of circulating codes [278,279] provides an alternative perspective to forming such spatially swept beams through the use of small time shifts $t + (\ell - 1)\Delta t$ across the array.
Inspired by fixational eye movement [280, 281], the FDA framework has also recently been subsumed by the notion of spatial modulation [282], which provides the freedom to change the rate and direction of fast-time spatial steering via an extension of the PCFM coding in (III-A4). For example, relative to standard beamforming Fig. III-C3 illustrates the aggregate beampattern from (III-C2) when using spatial modulation with a planar array to traverse a circle during the pulsewidth [283].

![Fig. III-C3. 2D aggregate beampattern via a planar array for (left) standard beamforming and (right) circular spatial modulation](image)

As is known for space-time adaptive processing (STAP) [264], the increase in useful degrees-of-freedom obtained by coupling the antenna’s spatial channels with the slow-time (Doppler) channels of the pulses in the CPI is most notably useful when employed adaptively. The same holds true for the increased degrees of freedom afforded by MIMO emissions. Due to the myriad ways in which MIMO can be realized, juxtaposed against the numerous different types of radars, it is not surprising that many different adaptive MIMO receive processing algorithms have emerged (e.g. [284-289]). These approaches all seek to exploit the much-increased dimensionality to enhance sensing performance, that is, to improve resolution/discrimination, sidelobe suppression, and interference rejection. Of course, as with STAP, MIMO incurs the “curse of dimensionality” in terms of higher computational cost and possibly higher sample data requirements (depending on the nature of the algorithm). Further, any such adaptive algorithm must rely on models of physically realizable waveforms that can be
transmitted by a radar, if possible even using captured replicas of the actual emissions to ensure maximum fidelity, and appropriately must consider the straddling effect that arises from discretization [110,263,288].

Finally, it is worth noting some of the concepts that have emerged from the ubiquitous notion of MIMO radar. As a countermeasure to combat passive bistatic exploitation of the emission of an active radar, the idea of bistatic denial was developed [290], in which a secondary waveform is emitted to mask the spatial sidelobes of the primary waveform. More recently, a counter-countermeasure was proposed [291] that seeks to estimate the primary waveform in the presence of the secondary waveform interference (from the perspective of the bistatic radar). The MIMO-enabled increased degrees of freedom have also been exploited to facilitate the embedding of communications within the radar emission [292-294]. Other forms of radar-embedded communication are discussed in the next two sections. A good subset of references to begin further reading on colocated MIMO is [238,247,253,256,258,276].

D) Distributed Aperture Radar

While the colocated MIMO formulation in the previous section employs waveform diversity to realize delay-angle coupled radar emissions with greater design degrees-of-freedom, the notion of distributed aperture radar considers multiple antenna apertures with considerable spatial separation. This arrangement has been known by many names including “netted radar” [295,296], “statistical MIMO” [240,297-299], and more classically as “multistatic radar” or “multisite radar” [300-305]. In this formulation, practical realization of phase coherency may be difficult (though progress continues [306,307]), thus often necessitating non-coherent combining to perform target detection using the distributed apertures. Further, this arrangement also includes the advantageous exploitation of other emitters in the environment, for which one has no control over the structure of the associated waveforms.
A prospective benefit of this distributed arrangement is a way to address the aspect angle dependence of a target’s radar cross section (RCS), which may vary by 10s of dB as the target moves with respect to the radar. The well-known Swerling models [308] can be used to represent how such RCS fluctuations behave statistically as a function of the target decorrelation time (scan-to-scan or pulse-to-pulse) [53, Chap. 7]. In fact, frequency diversity has long been used to realize this same fluctuating RCS effect to improve target detection [309-312].

Of particular note are the distributed coherent aperture X-band radar experiments undertaken by MIT Lincoln Lab [248, 249] that demonstrated the use of low cross-correlation waveforms to perform time and phase synchronization of the incident waveforms upon a selected target (in this case the application was ballistic missile defense). Once sufficient synchronization was achieved, the same waveform could be used across the distributed aperture to realize a cohere-on-transmit (or effectively “cohere-on-target”) mode yielding a 9 dB SNR gain.

It is important to note that the general notion of a distributed aperture dates back to the very earliest work in radar, where the waveforms employed were typically CW and the transmit and receive antennas were separated in a bistatic configuration for various operational and technical reasons (see Chaps. 1 and 2 of [11] for a fascinating historical review of the subject). For the bistatic arrangement, it has been shown [313, 314] that the delay-Doppler ambiguity function, referenced to the receiver, can be expressed as (modifying the nomenclature for consistency here)

\[
X_{\text{bistatic}} \left( R_R^{(h)}, R_R^{(a)}, v^{(h)}, v^{(a)}, \phi_R, L_B \right) = \int_{-\infty}^{\infty} s(t - \tau^{(a)}(R_R^{(a)}, \phi_R, L_B)) s^*(t - \tau^{(h)}(R_R^{(h)}, \phi_R, L_B)) \times \exp \left\{ -j 2\pi \left[ f_{D}^{(h)}(R_R^{(h)}, v^{(h)}, \phi_R, L_B) - f_{D}^{(a)}(R_R^{(a)}, v^{(a)}, \phi_R, L_B) \right] \right\} .
\]

The superscripts (a) and (h) denote the actual and hypothesized values of delay \( \tau \) and Doppler frequency \( f_D \), which are themselves nonlinear functions of the target velocity vector \( v \) and the bistatic geometry depicted in Fig. III-D1. The bistatic angle is implicitly defined via the
transmitter-to-target, receiver-to-target, and baseline transmitter-to-receiver distances $R_T$, $R_R$, and $L_B$, respectively.

**Fig. III-D1. Bistatic geometry**

In recent years there has been an explosion of new research on many aspects of bistatic radar, particularly the exploitation of *illuminators of opportunity* such as FM radio, broadcast television, commercial cellular, etc. (see surveys in [13,315]). This general research area has been referred to as “passive radar”, “hitchhiking”, “passive coherent location”, and more recently as “commensal radar”. While one has no control over the waveforms used by the illuminators of opportunity, and thus it is fair to assume that such waveforms are generally not optimal for a sensing application, the delay-Doppler ambiguity functions for these signals can be inferred based on the type of signal and its purpose (see [316,317] and [13, Chap. 6]). For example, analog FM radio has been found to realize an ambiguity function that is highly dependent on the nature of signal content (for example, speech vs music, as well as the tempo of the latter as shown in Fig. III-D2) and the signal structure (the chrominance subcarrier of analog television contains a repeating structure that produces strong ambiguities in range and Doppler per Fig. III-D3). In contrast, more recent digital modulation schemes produce delay-Doppler ambiguity functions that tend towards a thumbtack response (Fig. III-D4).
Fig. III-D2. Delay-Doppler ambiguity function of fast tempo jazz on FM radio (courtesy of Prof. Hugh Griffiths, University College London)

Fig. III-D3. Delay-Doppler ambiguity function of the chrominance subcarrier of analog TV (courtesy of Prof. Hugh Griffiths, University College London)
The bistatic ambiguity function of (III-D1) has also been generalized in different ways to account for multiple emitters in a multistatic configuration [318-320]. It then becomes necessary to consider how best to select from among the myriad emitters (and their subsequent echo responses) that may be present [321,322] and then how to combine the responses in the most advantageous way based on the observed/known signal structures and the multistatic geometry [323-326]. When concerned with image formation, this approach has also been referred to as RF/microwave tomography [327-330]. Surveys of bistatic and multistatic radar techniques can be found in [11-14,302-304].

E) Waveform Agility

Where the previous two sections considered the impact of waveform diversity via the spatial dimension, here we consider the implications of changing the waveform on a pulse-to-pulse basis during a coherent processing interval (CPI). Known as pulse agility, pulse diversity, or waveform agility, this arrangement provides increased degrees-of-freedom that may be used to suppress range sidelobes, to extend maximum unambiguous range, to enable radar-embedded
communication, and potentially to improve robustness to structured interference (e.g. commercial communications).

Complementary codes, originally proposed by Golay in the 1960s [39], are a form of waveform agility in which the pulse compression responses from two or more codes sum to produce an overall response whereby the range sidelobes completely cancel, leaving only the mainlobe [44,331-335]. Using (III-C1), the combined responses from a complementary set comprised of $M$ waveforms can be expressed as

$$\sum_{m=1}^{M} h_{MF,m}(t) \ast s_m(t) \approx \delta(t),$$

(III-E1)

where $\delta(t)$ is a Dirac delta and the approximation denotes the fact that this ideal response would really comprise the pulse compression mainlobe of finite bandwidth waveforms. It was shown in [336] that a similar result could be obtained by modulating a given waveform (e.g. LFM) with a set of orthonormal codes. While the complementary arrangement would seem to solve the range-sidelobe problem, the sidelobe suppression performance for complementary codes is known to degrade when the signal structure deviates from the ideal, such as when generated by a physical transmitter (due to bandlimiting and distortion) or when Doppler is present. That said, continued work is exploring ways in which to improve the Doppler tolerance limitation [335,337,338] and practical transmitter effects could be incorporated such as discussed in Section III-A.

Another well-known form of waveform agility is called stepped frequency, synthetic wideband, or frequency jumped burst and essentially involves a center frequency offset of $\Delta f$ between adjacent pulses [339, Chap. 5]. The benefit of stepped frequency waveforms is that wideband, hence high range resolution, sensing can be achieved while avoiding the complexity and cost of wideband hardware. Instead, a “burst” of narrowband pulses with offset center frequencies is generated, with the resulting received echoes pulse compressed as usual according to each individual waveform and then coherently combined across the set of pulses (e.g. via inverse FFT). Assuming the same waveform repeated over the set of frequency steps, it was
shown in [10] that the delay-Doppler ambiguity function for the overall stepped-frequency emission can be generalized from (II-E1) as

$$|\mathcal{X}_{SF}(\tau, f_D)| = |\mathcal{X}(\tau, f_D)| \frac{\sin(M \pi (f_D + k_s \tau) T_{PRI})}{M \sin(\pi (f_D + k_s \tau) T_{PRI})}$$

for $|\tau| \leq T$, \hspace{1cm} (III-E2)

where $|\mathcal{X}(\tau, f_D)|$ is the ambiguity function of the single waveform and

$$k_s = \pm \frac{A_f}{T_{PRI}}$$

is the linear FM ramp applied across the CPI.

A problem with using a constant pulse-to-pulse frequency offset across the CPI is the appearance of grating lobes. Consequently, much of the work on this type of emission scheme has focused on how to design the frequency offsets and the selection of the individual pulsed waveforms so as to reduce these grating lobes [183,340-345]. As discussed in Section IV-B and [180, Chap. 5], the stepped-frequency emission structure also provides a convenient way in which to avoid other in-band spectrum users, albeit with an expected degradation in sidelobe performance. The related concepts of frequency agility [309-312] and frequency diversity [346, Chap. 12], which provide robustness to frequency-dependent target RCS fluctuations and radar countermeasures, respectively, likewise involve a pulse-to-pulse change of the center frequency, even though generally not in as clearly structured a manner as with stepped frequency. It is also interesting to note that the stepped-frequency scheme bears some similarity in mathematical construction, if not the domain in which it is applied, to the more recent MIMO concept of the frequency-diverse array of Section III-C. See [10,53,346] as a primer for further reading on complementary codes, stepped-frequency, and frequency-diverse operation.

Besides the complementary coding and stepped-frequency concepts above, waveform agility may also provide an alternative to using multiple PRFs to extend the unambiguous range ([347,348], [53, Chap. 17], and see Fig. III-E1); it may facilitate the embedding of communication or navigation information in radar emissions [349] and may even enable new
forms of discrimination. As an example of the latter, a biomimetic form of waveform agility, called “twin inverted pulse” (TWIP), was recently developed for sonar [23] and radar [350]. TWIP mimics a waveform scheme employed by dolphins to discriminate between linear and nonlinear scattering in bubble-rich underwater environments [351,352], such as those from the wakes of passing ships. In this formulation, pulse pairs are emitted, with the two pulses having opposite polarity \( s_2(t) = -s_1(t) \). Computing the ratio between the difference and sum of the resulting matched filter responses to these waveforms then provides a way to emphasize the nonlinear scattering relative to the linear scattering, as demonstrated experimentally in [350].

![Diagram](image)

**Fig. III-E1.** The same pulse is repeated throughout the CPI (top) so a distant target is range-ambiguous while the target could be range-unambiguous if different waveforms were used within the CPI (bottom).

An earlier idea that arose to address range ambiguities was to incorporate a pulse-to-pulse phase coding [353-357]. Waveform agility can be viewed as a generalization of this idea – instead of phase coding the same waveform across the set of pulses, each pulse could be a different waveform. However, when changing the waveform on a pulse-to-pulse basis, clutter cancellation may be degraded for radar modes on which it is performed. Figure III-E2 illustrates the matched filter response to each of four DPSK-implemented length-100 randomly generated binary codes, in which we can observe that the sidelobe structure is different for each. Clutter illuminated by
waveforms such as these would experience a range sidelobe modulation effect over the CPI that in turn would limit the efficacy of clutter cancellation. Thus if one is to consider waveform agility in the context of clutter cancellation, it is necessary 1) to compensate for the modulation effect; 2) to expand the dimensionality of the receive processing so as to perform fast-time (range) and slow-time (Doppler) processing in a joint manner; or 3) to expand the overall dimensionality of the radar emission such that these sidelobes are simply driven into the noise since they do not combine coherently.

![Fig. III-E2. Matched filter responses for four length-100 random binary codes implemented with DPSK. Note the sidelobes are quite different.](image)

In the case that only two different waveforms are used and under the assumption that fast-time Doppler effects are negligible, one can satisfy the sidelobe similarity constraint [349]

$$h_{MF,1}(t) \ast s_1(t) = h_{MF,2}(t) \ast s_2(t)$$

by setting $s_2(t) = s_1^*(T - t)$, such that $h_{MF,1}(t) = s_2(t)$ and $h_{MF,2}(t) = s_1(t)$. However, this constraint cannot be met for more than two waveforms because, in general, the frequency response for the $m$th filter would be
which is an infinite impulse response (IIR) filter due to the term in the denominator. Thus the similarity constraint can only be met approximately for more than two waveforms by using mismatched filters of sufficient length. It was shown in [349], and subsequently extended in [358,359], that a joint formulation of the least-squares based mismatched filter described in Section II-H provides this capability, with the caveat that the number of waveforms still be relatively small (say 4 or 5) due to the associated trade-off of a general increase in sidelobes across the (now similar) mismatched filter responses. In [360] this filter design problem was also considered from a convex optimization perspective.

One may also consider performing pulse compression and (slow-time) Doppler processing jointly. An adaptive processing formulation for this joint perspective on waveform agility was developed in [361], though the associated computational cost is rather high. A non-adaptive approach was likewise conceived in [362] and subsequently was extended in [363]. While these joint processing schemes are clearly more complex and generally have a higher computational cost than performing pulse compression and Doppler processing separately, the multiplicative increase in adaptive degrees of freedom (not unlike that obtained via STAP [264]) provides the means to address the range-Doppler coupling that arises from range-sidelobe modulation of clutter. Further, these schemes can be extended to incorporate multiple-time-around clutter, also known as range-ambiguous clutter or folded clutter that becomes more prevalent at higher PRF [364,365], [366, Chap. 9.5].

Finally, an arguably more straightforward approach to addressing range-sidelobe modulation is simply to expand the dimensionality of the radar emission to such a degree that the sidelobes are driven into the noise since they do not coherently combine. While not necessarily designed for this reason, one could contend that this very effect occurs for noise radar [71-75] and the similar concept of chaotic radar [367-369]. For example, for the recently developed FMCW noise radar
concept in [370], Fig. III-E3 illustrates the RMS average sidelobe response over $10^4$ waveform segments (left) that is reduced by roughly 40 dB when coherently integrating over these segments (right) such as occurs with Doppler processing. Recently, this concept was also examined for a pulsed mode [371] in which it is observed that a high PRF (100 kHz in this case) provides performance very similar to the FMCW mode while a lower PRF (1 kHz) necessitates additional receive processing since the dimensionality is no longer high enough to drive the sidelobes into the noise. See [349,361,371] as a starting point for further reading on waveform agility.

![Fig. III-E3. Range sidelobes for FM noise radar for (left) the RMS average over $10^4$ waveform segments and (right) the coherent integration over these segments](image)

\( F) \quad \text{Polarization Diversity} \)

Polarization provides another dimension to utilize for radar detection, tracking, and imaging. In 1986 Giuli [372] provided an excellent survey on polarization diversity. Simply put, a dual-polarized antenna has twice the degrees of freedom as the similar antenna with only one polarization channel, though with the caveat of increased complexity and implementation cost. Assuming each antenna element (if part of an array) has orthogonally polarized channels, such as the simple crossed dipoles depicted in Fig. III-F1, then the additional freedom on both transmit and receive may provide several advantages. These advantages include enhanced classification/identification/discrimination between targets and clutter [373-380], measurement of scattering depolarization for remote sensing applications [381-385], dual-polarized generalization
of the matched illumination concept discuss in Section III-B [386,387], dual-polarized SAR [388,389], and polarization coding on transmit [390,391].

![Crossed dipoles to enable dual-polarized operation](image1)

**Fig. III-F1.** Crossed dipoles to enable dual-polarized operation

![Poincaré sphere representing the different polarization states](image2)

**Fig. III-F2.** Poincaré sphere representing the different polarization states

A useful and well-known way to visualize the polarization state is the Poincaré sphere (Fig. III-F2), which includes the basic states (horizontal, vertical, left/right-hand circular) as well as all the variations in between. Any two antipodal states on the Poincaré sphere are orthogonal. Thus, it is common to express the received radar scattering at a given instant in time in terms of the linear horizontal and vertical components via the scattering matrix

\[
X = \begin{bmatrix}
X_{HH} & X_{HV} \\
X_{VH} & X_{VV}
\end{bmatrix},
\]

(III-F1)
where the subscripts denote the receive and transmit channel for each component. Therefore the horizontal (vertical) antenna captures a \textit{co-polarized} (\textit{cross-polarized}) signal as well as a \textit{cross-polarized} (\textit{co-polarized}) signal. It should be noted that many different decompositions of the scattering matrix have been developed to achieve greater understanding of the inherent scattering properties of an object and to develop recognizable features for various scattering environments and target structures [392-395].

Because the co-polarized and cross-polarized components of (IV-F1) are superimposed at each orthogonal antenna element, accurate estimation of the scattering matrix terms as a function of range is often performed by alternating between which of the orthogonal antenna elements transmits so as to achieve isolation [396,397]. However, this isolation comes at the cost of increased ambiguities since the measurement time is doubled relative to simultaneous dual-polarized operation. It was suggested in [398,399] that simultaneous operation could be performed if the waveforms emitted by the orthogonal antenna elements are sufficiently separable. According to Parseval’s theorem, if these two waveforms have the same spectral support (which is generally desired in this context to ensure phase coherence between to various channels), then there is a limit on how low the cross-correlation between the waveforms can be (which is dependent on the time-bandwidth product). Because their cross-correlation is rather flat, one could use two LFM waveforms, one an up-chirp and the other a down-chirp, on the respective orthogonal antenna elements [398,399]. Various other coding approaches have been examined, including bias removal of coherent cross-channel coupling [400], pulse-to-pulse phase coding [401], and a frequency-interleaved OFDM structure that minimizes the spectral overlap (the spectral maxima of one waveform coincides with the spectral minima of the other) while still maintaining the same general spectral support [402]. It has also recently been experimentally demonstrated [391] that adaptive range-domain processing using knowledge of the two waveforms can separate the co-pol/cross-pol components just as well as the time alternating approach (with a higher computational cost, of course).
IV. CONCLUSIONS

Waveform Diversity is an exciting technology that has sparked intense interest from the research community in recent years due to advances in high-fidelity electronic components and high-performance computing. WD is expected to have a profound impact on radar spectrum management, particularly in light of increasing competition for spectrum usage, as well as to facilitate enhanced radar sensitivity/discrimination and perhaps even to enable new sensing modes. As mentioned at the beginning of this tutorial, the definition of WD is clearly rather broad, but such is to be expected for a topic that continues to evolve.

ACKNOWLEDGEMENTS

The authors would like to thank several individuals for their help in preparing this lengthy manuscript. Mr. John Jakabosky of the US Naval Research Laboratory (previously of the University of Kansas) and Mr. Patrick McCormick of the University of Kansas provided valuable figures used throughout. Prof. Hugh Griffiths of University College London generously allowed use of figures on ambiguity functions of illuminators of opportunity. Prof. Chris Allen of the University of Kansas provided some useful references on current waveform generation technology.

The individuals participating in the NATO SET-179 and SET-182 research task groups on Dynamic Waveform Diversity & Design and Radar Spectrum Engineering & Management, respectively, provided invaluable different perspectives and research contributions these last few years. Likewise, the material surveyed in this tutorial would not exist if not for the continued advancements by a multitude of persons in the radar research community.

Dr. Dale Blair and Prof. Maria Greco provided the opportunity to prepare a Systems Magazine tutorial summarizing and cataloguing the vast array of contributions in Waveform Diversity. We have certainly learned some new things in the process.
Finally, a special thank you to Dr. Michael C. Wicks for guiding the radar community further down this path. To borrow more from Joe Guerci’s “Cambrian explosion” analogy, it will be very interesting to see what emerges next from this “primordial radar soup”.

REFERENCES


[27] E. Hüttman, German Patent No. 768,068, Mar. 22, 1940.


