Radar Signal Processing

Jonelle Loper

First Edition, 2012

ISBN 978-81-323-4017-1

Published by:
White Word Publications
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter-1

Digital Signal Processing

Digital signal processing (**DSP**) is concerned with the representation of signals by a sequence of numbers or symbols and the processing of these signals. Digital signal processing and analog signal processing are subfields of signal processing. DSP includes subfields like: audio and speech signal processing, sonar and radar signal processing, sensor array processing, spectral estimation, statistical signal processing, digital image processing, signal processing for communications, control of systems, biomedical signal processing, seismic data processing, etc.

The goal of DSP is usually to measure, filter and/or compress continuous real-world analog signals. The first step is usually to convert the signal from an analog to a digital form, by *sampling* it using an analog-to-digital converter (ADC), which turns the analog signal into a stream of numbers. However, often, the required output signal is another analog output signal, which requires a digital-to-analog converter (DAC). Even if this process is more complex than analog processing and has a discrete value range, the application of computational power to digital signal processing allows for many advantages over analog processing in many applications, such as error detection and correction in transmission as well as data compression.

DSP algorithms have long been run on standard computers, on specialized processors called digital signal processors (DSPs), or on purpose-built hardware such as application-specific integrated circuit (ASICs). Today there are additional technologies used for digital signal processing including more powerful general purpose microprocessors, field-programmable gate arrays (FPGAs), digital signal controllers (mostly for industrial apps such as motor control), and stream processors, among others.

Signal sampling

With the increasing use of computers the usage of and need for digital signal processing has increased. In order to use an analog signal on a computer it must be digitized with an analog-to-digital converter. Sampling is usually carried out in two stages, discretization and quantization. In the discretization stage, the space of signals is partitioned into equivalence classes and quantization is carried out by replacing the signal with representative signal of the corresponding equivalence class. In the quantization stage the representative signal values are approximated by values from a finite set.

The Nyquist–Shannon sampling theorem states that a signal can be exactly reconstructed from its samples if the sampling frequency is greater than twice the highest frequency of the signal; but requires an infinite number of samples . In practice, the sampling frequency is often significantly more than twice that required by the signal's limited bandwidth.

A digital-to-analog converter is used to convert the digital signal back to analog. The use of a digital computer is a key ingredient in digital control systems.

DSP domains

In DSP, engineers usually study digital signals in one of the following domains: time domain (one-dimensional signals), spatial domain (multidimensional signals), frequency domain, autocorrelation domain, and wavelet domains. They choose the domain in which to process a signal by making an informed guess (or by trying different possibilities) as to which domain best represents the essential characteristics of the signal. A sequence of samples from a measuring device produces a time or spatial domain representation, whereas a discrete Fourier transform produces the frequency domain information, that is the frequency spectrum. Autocorrelation is defined as the cross-correlation of the signal with itself over varying intervals of time or space.

Time and space domains

The most common processing approach in the time or space domain is enhancement of the input signal through a method called filtering. Digital filtering generally consists of some linear transformation of a number of surrounding samples around the current sample of the input or output signal. There are various ways to characterize filters; for example:

- A "linear" filter is a linear transformation of input samples; other filters are "non-linear". Linear filters satisfy the superposition condition, i.e. if an input is a weighted linear combination of different signals, the output is an equally weighted linear combination of the corresponding output signals.

- A "causal" filter uses only previous samples of the input or output signals; while a "non-causal" filter uses future input samples. A non-causal filter can usually be changed into a causal filter by adding a delay to it.

- A "time-invariant" filter has constant properties over time; other filters such as adaptive filters change in time.

- Some filters are "stable", others are "unstable". A stable filter produces an output that converges to a constant value with time, or remains bounded within a finite interval. An unstable filter can produce an output that grows without bounds, with bounded or even zero input.

- A "finite impulse response" (FIR) filter uses only the input signals, while an "infinite impulse response" filter (IIR) uses both the input signal and previous samples of the output signal. FIR filters are always stable, while IIR filters may be unstable.

Filters can be represented by block diagrams which can then be used to derive a sample processing algorithm to implement the filter using hardware instructions. A filter may also be described as a difference equation, a collection of zeroes and poles or, if it is an FIR filter, an impulse response or step response.

The output of a digital filter to any given input may be calculated by convolving the input signal with the impulse response.

Frequency domain

Signals are converted from time or space domain to the frequency domain usually through the Fourier transform. The Fourier transform converts the signal information to a magnitude and phase component of each frequency. Often the Fourier transform is converted to the power spectrum, which is the magnitude of each frequency component squared.

The most common purpose for analysis of signals in the frequency domain is analysis of signal properties. The engineer can study the spectrum to determine which frequencies are present in the input signal and which are missing.

In addition to frequency information, phase information is often needed. This can be obtained from the Fourier transform. With some applications, how the phase varies with frequency can be a significant consideration.

Filtering, particularly in non-realtime work can also be achieved by converting to the frequency domain, applying the filter and then converting back to the time domain. This is a fast, O(n log n) operation, and can give essentially any filter shape including excellent approximations to brickwall filters.

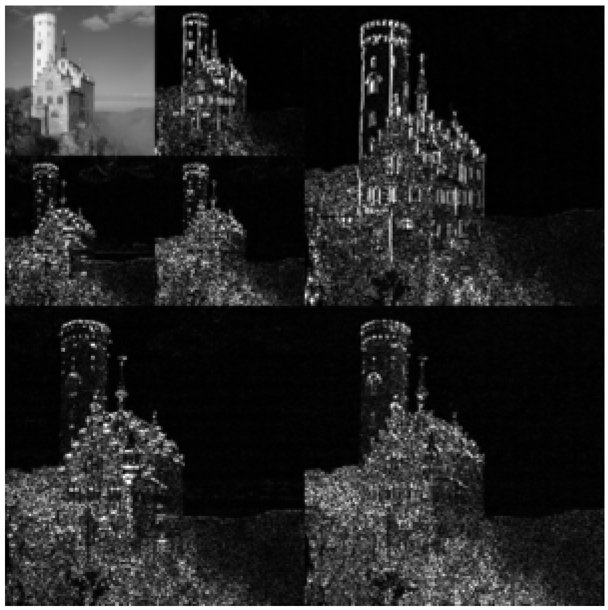
There are some commonly used frequency domain transformations. For example, the cepstrum converts a signal to the frequency domain through Fourier transform, takes the logarithm, then applies another Fourier transform. This emphasizes the frequency components with smaller magnitude while retaining the order of magnitudes of frequency components.

Frequency domain analysis is also called *spectrum-* or *spectral analysis*.

Z-plane analysis

Whereas analog filters are usually analysed in terms of transfer functions in the s plane using Laplace transforms, digital filters are analysed in the z plane in terms of Z-transforms. A digital filter may be described in the z plane by its characteristic collection of zeroes and poles.

Wavelet



An example of the 2D discrete wavelet transform that is used in JPEG2000. The original image is high-pass filtered, yielding the three large images, each describing local changes in brightness (details) in the original image. It is then low-pass filtered and downscaled, yielding an approximation image; this image is high-pass filtered to produce the three smaller detail images, and low-pass filtered to produce the final approximation image in the upper-left.

In numerical analysis and functional analysis, a **discrete wavelet transform** (DWT) is any wavelet transform for which the wavelets are discretely sampled. As with other wavelet transforms, a key advantage it has over Fourier transforms is temporal resolution: it captures both frequency *and* location information (location in time).

Applications

The main applications of DSP are audio signal processing, audio compression, digital image processing, video compression, speech processing, speech recognition, digital communications, RADAR, SONAR, seismology and biomedicine. Specific examples are speech compression and transmission in digital mobile phones, room correction of sound in hi-fi and sound reinforcement applications, weather forecasting, economic forecasting, seismic data processing, analysis and control of industrial processes, medical imaging such as CAT scans and MRI, MP3 compression, computer graphics, image manipulation, hi-fi loudspeaker crossovers and equalization, and audio effects for use with electric guitar amplifiers.

Implementation

Digital signal processing is often implemented using specialised microprocessors such as the DSP56000, the TMS320, or the SHARC. These often process data using fixed-point arithmetic, although some versions are available which use floating point arithmetic and are more powerful. For faster applications FPGAs might be used. Beginning in 2007, multicore implementations of DSPs have started to emerge from companies including Freescale and Stream Processors, Inc. For faster applications with vast usage, ASICs might be designed specifically. For slow applications, a traditional slower processor such as a microcontroller may be adequate. Also a growing number of DSP applications are now being implemented on Embedded Systems using powerful PCs with a Multi-core processor.

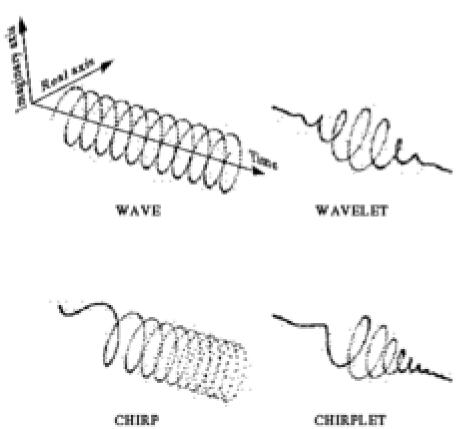
Techniques

- Bilinear transform
- Discrete Fourier transform
- Discrete-time Fourier transform
- Filter design
- LTI system theory
- Minimum phase
- Transfer function
- Z-transform
- Goertzel algorithm
- s-plane

Chapter-2

Chirplet Transform and Electronic Counter-countermeasures

Chirplet transform



Comparison of wave, wavelet, chirp, and chirplet

In signal processing, the **chirplet transform** is an inner product of an input signal with a family of analysis primitives called **chirplets**.

Similarity to other transforms

Much as in the wavelet transform, the chirplets are usually generated from (or can be expressed as being from) a single *mother chirplet* (analogous to the so-called *mother wavelet* of wavelet theory).

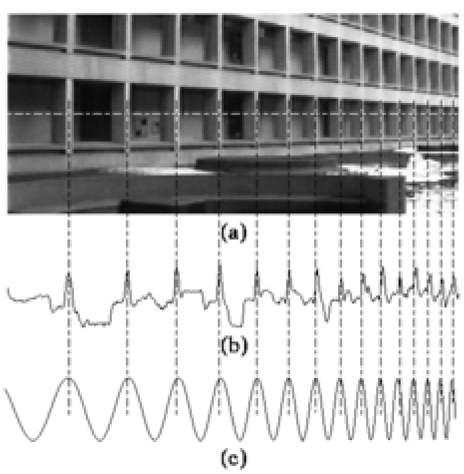
What is the chirplet and chirplet transform?

The term *chirplet transform* was coined by Steve Mann, as the title of the first published paper on chirplets. The term *chirplet* itself (apart from chirplet transform) was also used by Steve Mann, Domingo Mihovilovic, and Ronald Bracewell to describe a windowed portion of a chirp function. In Mann's words:

A wavelet is a piece of a wave, and a chirplet, similarly, is a piece of a chirp. More precisely, a chirplet is a windowed portion of a chirp function, where the window provides some time localization property. In terms of time–frequency space, chirplets exist as rotated, sheared, or other structures that move from the traditional parallelism with the time and frequency axes that are typical for waves (Fourier and short-time Fourier transforms) or wavelets.

The chirplet transform thus represents a rotated, sheared, or otherwise transformed tiling of the time–frequency plane. Although chirp signals have been known for many years in radar, pulse compression, and the like, the first published reference to the *chirplet transform* described specific signal representations based on families of functions related to one another by time–varying frequency modulation or frequency varying time modulation, in addition to time and frequency shifting, and scale changes. In that paper, the Gaussian chirplet transform was presented as one such example, together with a successful application to ice fragment detection in radar (improving target detection results over previous approaches). The term *chirplet* (but not the term *chirplet transform*) was also proposed for a similar transform, apparently independently, by Mihovilovic and Bracewell later that same year.

Applications



(a) In image processing, periodicity is often subject to a linear scaling. (b) In this image, repeating structures like the alternating dark space inside the windows, and light space of the white concrete, *chirp* (increase in frequency) towards the right. (c) The chirplet transform is able to represent this modulated variation compactly.

The chirplet transform is a useful signal analysis and representation framework that is widely used in

- Radar;

- Biomedical, the most commonly and widely used chirplet applications in biomedical being:
    - Heart sound processing;
    - EEG processing.
- Signal processing;
- Image processing;
- SETI@home uses chirp functions to compensate for Doppler drift;
- Chirplet Time Domain Reflectometry.

Taxonomy of chirplet transforms

There are two broad categories of chirplet transform:

- Fixed
- Adaptive

These categories may be further subdivided by:

- choice of chirp
- choice of window

In either the fixed or adaptive case, the chirplets may be:

- q-chirplets (quadratic chirplets) of the form $\exp[i2\pi(at^2 + bt + c)]$ or, in general, some kind of quadratically varying exponent, linear swept wave packet, or the like. These are sometimes called *linear FM chirplets* (linear frequency-modulated chirplets, since quadratic phase is linear frequency). Commonly used families of q-chirplets are metaplectomorphisms of one another (i.e., the energy distribution of any member of the family of q-chirplets can be generated from any other member by shear-in-time, shear-in-frequency, dilation, translation-in-time, and translation-in-frequency).
- w-chirplets, also known as *warblets*. A family of warblets are like the sound made by birds called warblers. Unwindowed warblets have a sinusoidally varying time–frequency distribution, or similar cyclostationary or periodically varying time–frequency plot. The sound of a police siren is an example, in which the pitch goes up and down periodically. Of course, the warblet is a "piece of" a warble (i.e., a windowed section of something that has a time–frequency periodicity).
- d-chirplets, also known as *Doppler chirplets*. These are analysis functions that mimic the Doppler shift of a passing tone (e.g., the sound you hear from a train whistle as it moves past).
- p-chirplets, in which the scale varies projectively. Whereas the wavelet transform is based on wavelets of the form $g(ax + b)$, the p-type chirplet transform is based on chirplets of the form $g((ax + b)/(cx + 1))$, where a is the scale, b is the translation, and c is the *chirpiness* (chirp-rate, as defined by the degree of perspective, or projection).

The choice of window is also another matter of decision. A Gaussian window is one possible choice, leading to a four parameter chirplet transform (for which time–shear and frequency–shear only give one degree of freedom that may thus be encapsulated as rotation angle—Radon transform of the Wigner distribution may, for example, be used, as may the fractional Fourier transform).

Another possible choice is the rectangular window, and discrete prolate spheroidal sequences (also called Multitaper#The_Slepian_sequences) may be used, by way of the *method of multiple mother chirplets*. This method gives a total chirplet transform as the sum of energies in various contributory chirplet transforms made from multiple windows, akin to the way in which DPSSs are used to get a perfect rectangular tiling of the time–frequency plane. Thus it is now possible to get perfect parallelogram tiling of the time–frequency plane, using the method of multiple mother chirplets.

Related work

The chirplet transform is a generalized representation that includes as special cases:

- The Fourier transform
- The short-time Fourier transform (STFT), also known as the spectrogram
- The Wigner-Ville distribution
- The wavelet transform
- Canonical conjugate variables
- Segal–Shale–Weil distribution

Josef Segman proposed the idea of incorporating scale into the Heisenberg group (position, momentum, phase, or equivalently any canonical conjugate variables taken together with phase, such as, for example, time, frequency, and phase). This gave rise to a four parameter space of time, frequency, phase, and scale. Segman introduced this idea of *phase scale*. (Personal communication with Mann, from Josef Segman, at Harvard University and at Massachusetts Institute of Technology). Further personal communication between Irving Segal (the principal behind the Segal, Shale Weil representation, known also as the metaplectic representation—a double covering of the symplectic group) and Mann led to additional insight into the chirplet transform, in particular, to the variation of the chirplet transform that is based on q-chirplets.

Electronic counter-countermeasures

Electronic counter-countermeasures (ECCM) is a part of electronic warfare which includes a variety of practices which attempt to reduce or eliminate the effect of electronic countermeasures (ECM) on electronic sensors aboard vehicles, ships and aircraft and weapons such as missiles. ECCM is also known as **electronic protective measures** (EPM), chiefly in Europe. In practice, EPM often means resistance to jamming.

History

Ever since electronics have been used in battle in an attempt to gain superiority over the enemy, effort has been spent on techniques to reduce the effectiveness of those electronics. More recently, sensors and weapons are being modified to deal with this threat. One of the most common types of ECM is radar jamming or spoofing. This originated with the Royal Air Force use of what they codenamed *window* during World War II, which is now often referred to as *chaff*. Jamming also may have originated with the British during World War II, when they began jamming German radio communications.

In perhaps the first example of ECCM, the Germans increased their radio transmitter power in an attempt to 'burn through' or override the British jamming, which by necessity of the jammer being airborne or further away produced weaker signals. This is still one of the primary methods of ECCM today. For example, modern airborne jammers are able to identify incoming radar signals from other aircraft and send them back with random delays and other modifications in an attempt to confuse the opponent's radar set, making the 'blip' jump around wildly and be impossible to range. More powerful airborne radars means that it is possible to 'burn through' the jamming at much greater ranges by overpowering the jamming energy with the actual radar returns. The Germans were not really able to overcome the chaff spoofing very successfully and had to work around it (by guiding the aircraft to the target area and then having them visually acquire the targets).

Today, more powerful electronics with smarter software for operation of the radar might be able to better discriminate between a moving target like an aircraft and an almost stationary target like a chaff bundle.

With the technology going into modern sensors and seekers, it is inevitable that all successful systems have to have ECCM designed into them, lest they become useless on the battlefield. In fact, the 'electronic battlefield' is often used to refer to ECM, ECCM and ELINT activities, indicating that this has become a secondary battle in itself.

Specific ECCM techniques

The following are some examples of EPM (other than simply increasing the fidelity of sensors through techniques such as increasing power or improving discrimination):

ECM detection

Sensor logic may be programmed to be able to recognize attempts at spoofing (e.g., aircraft dropping chaff during terminal homing phase) and ignore them. Even more sophisticated applications of ECCM might be to recognize the type of ECM being used, and be able to cancel out the signal.

Pulse compression by "chirping", or linear frequency modulation

One of the effects of the pulse compression technique, is boosting the apparent signal strength as perceived by the radar receiver. The outgoing radar pulses are chirped, that is, the frequency of the carrier is varied within the pulse, much like the sound of a cricket chirping. When the pulse reflects off a target and returns to the receiver, the signal is processed to add a delay as a function of the frequency. This has the effect of 'stacking' the pulse so it seems stronger, but shorter in duration, to further processors. The effect can increase the received signal strength to above that of noise jamming. Similarly, jamming pulses (used in deception jamming) will not typically have the same chirp, so will not benefit from the increase in signal strength.

Frequency hopping

Frequency agility ('frequency hopping') may be used to rapidly switch the frequency of the transmitted energy, and receiving only that frequency during the receiving time window. This foils jammers which cannot detect this frequency switch quickly enough, and switch their own jamming frequency accordingly during the receiving time window.

This method is also useful against barrage jamming, in that it forces the jammer to spread its jamming power across multiple frequencies in the jammed system's frequency range, reducing its power in the actual frequency used by the radar at any one time. The use of similar spread-spectrum techniques allow signals to be spread over a wide enough spectrum to make jamming of such a wideband signal difficult.

Sidelobe blanking

Radar jamming can be effective from directions other than the direction the radar antenna is currently aimed. When jamming is strong enough, the radar receiver can detect it from a relatively low gain sidelobe. The radar, however, will process signals as if they were received in the main lobe. Therefore, jamming can be seen in directions other than where the jammer is located. To combat this, an omnidirectional antenna is used for a comparison signal. By comparing the signal strength as received by both the omnidirectional and the (directional) main antenna, signals can be identified that are not from the direction of interest. These signals are then ignored.

Polarization

Polarization can be used to filter out unwanted signals, such as jamming. If a jammer and receiver do not have the same polarity, the jamming signal will incur a loss that reduces its effectiveness. The four basic polarities are horizontal, vertical, right-hand circular, and left-hand circular. The signal loss inherent in a cross polarized (transmitter different from receiver) pair is 3 decibels for dissimilar types, and 17 dB for opposites.

Aside from power loss to the jammer, radar receivers can also benefit from using two or more antennas of differing polarity and comparing the signals received on each. This effect can effectively eliminate all jamming of the wrong polarity, although enough jamming may still obscure the actual signal.

Radiation homing

The other main aspect of ECCM, is to program sensors or seekers to detect attempts at ECM and possible even to take advantage of it. For example, some modern fire-and-forget missiles like the Vympel R-77 and the AMRAAM are able to home in directly on sources of radar jamming if the jamming is too powerful to allow them to find and track the target normally. This mode, called 'home-on-jam', actually makes the missile's job easier. Some missile seekers actually target the enemy's radiation sources, and are therefore called "anti-radiation missiles" (ARM). The jamming in this case effectively becomes a beacon announcing the presence and location of the transmitter. This makes the use of such ECM a difficult decision; it may serve to obscure an exact location from a non-ARM missile, but in doing so it must emit signals which can be exploited by an ARM type missile.

Chapter-3

Matched Filter

In telecommunications, a **matched filter** (originally known as a **North filter**) is obtained by correlating a known signal, or template, with an unknown signal to detect the presence of the template in the unknown signal. This is equivalent to convolving the unknown signal with a conjugated time-reversed version of the template. The matched filter is the optimal linear filter for maximizing the signal to noise ratio (SNR) in the presence of additive stochastic noise. Matched filters are commonly used in radar, in which a known signal is sent out, and the reflected signal is examined for common elements of the out-going signal. Pulse compression is an example of matched filtering. Two-dimensional matched filters are commonly used in image processing, e.g., to improve SNR for X-ray pictures.

Derivation of the matched filter

The following section derives the matched filter for a discrete-time system. The derivation for a continuous-time system is similar, with summations replaced with integrals.

The matched filter is the linear filter, *h*, that maximizes the output signal-to-noise ratio.

$$y[n] = \sum_{k=-\infty}^{\infty} h[n-k]x[k].$$

Though we most often express filters as the impulse response of convolution systems, as above, it is easiest to think of the matched filter in the context of the inner product, which we will see shortly.

We can derive the linear filter that maximizes output signal-to-noise ratio by invoking a geometric argument. The intuition behind the matched filter relies on correlating the received signal (a vector) with a filter (another vector) that is parallel with the signal, maximizing the inner product. This enhances the signal. When we consider the additive stochastic noise, we have the additional challenge of minimizing the output due to noise by choosing a filter that is orthogonal to the noise.

Let us formally define the problem. We seek a filter, *h*, such that we maximize the output signal-to-noise ratio, where the output is the inner product of the filter and the observed signal *x*.

Our observed signal consists of the desirable signal *s* and additive noise *v*:

$$x = s + v.$$

Let us define the covariance matrix of the noise, reminding ourselves that this matrix has Hermitian symmetry, a property that will become useful in the derivation:

$$R_v = E\{vv^H\}$$

where $.^H$ denotes Hermitian (conjugate) transpose, and *E* denotes expectation. Let us call our output, *y*, the inner product of our filter and the observed signal such that

$$y = \sum_{k=-\infty}^{\infty} h^*[k]x[k] = h^H x = h^H s + h^H v = y_s + y_v.$$

We now define the signal-to-noise ratio, which is our objective function, to be the ratio of the power of the output due to the desired signal to the power of the output due to the noise:

$$SNR = \frac{|y_s|^2}{E\{|y_v|^2\}}.$$

We rewrite the above:

$$SNR = \frac{|h^H s|^2}{E\{|h^H v|^2\}}.$$

We wish to maximize this quantity by choosing *h*. Expanding the denominator of our objective function, we have

$$E\{|h^H v|^2\} = E\{(h^H v)(h^H v)^H\} = h^H E\{vv^H\}h = h^H R_v h.$$

Now, our *SNR* becomes

$$SNR = \frac{|h^H s|^2}{h^H R_v h}.$$

We will rewrite this expression with some matrix manipulation. The reason for this seemingly counterproductive measure will become evident shortly. Exploiting the Hermitian symmetry of the covariance matrix R_v, we can write

$$SNR = \frac{|(R_v^{1/2}h)^H (R_v^{-1/2}s)|^2}{(R_v^{1/2}h)^H (R_v^{1/2}h)},$$

We would like to find an upper bound on this expression. To do so, we first recognize a form of the Cauchy-Schwarz inequality:

$$|a^H b|^2 \leq (a^H a)(b^H b),$$

which is to say that the square of the inner product of two vectors can only be as large as the product of the individual inner products of the vectors. This concept returns to the intuition behind the matched filter: this upper bound is achieved when the two vectors a and b are parallel. We resume our derivation by expressing the upper bound on our *SNR* in light of the geometric inequality above:

$$SNR = \frac{|(R_v^{1/2}h)^H (R_v^{-1/2}s)|^2}{(R_v^{1/2}h)^H (R_v^{1/2}h)} \leq \frac{\left[(R_v^{1/2}h)^H (R_v^{1/2}h)\right] \left[(R_v^{-1/2}s)^H (R_v^{-1/2}s)\right]}{(R_v^{1/2}h)^H (R_v^{1/2}h)}.$$

Our valiant matrix manipulation has now paid off. We see that the expression for our upper bound can be greatly simplified:

$$SNR = \frac{|(R_v^{1/2}h)^H (R_v^{-1/2}s)|^2}{(R_v^{1/2}h)^H (R_v^{1/2}h)} \leq s^H R_v^{-1} s.$$

We can achieve this upper bound if we choose,

$$R_v^{1/2}h = \alpha R_v^{-1/2}s$$

where α is an arbitrary real number. To verify this, we plug into our expression for the output *SNR*:

$$SNR = \frac{|(R_v^{1/2}h)^H (R_v^{-1/2}s)|^2}{(R_v^{1/2}h)^H (R_v^{1/2}h)} = \frac{\alpha^2 |(R_v^{-1/2}s)^H (R_v^{-1/2}s)|^2}{\alpha^2 (R_v^{-1/2}s)^H (R_v^{-1/2}s)} = \frac{|s^H R_v^{-1} s|^2}{s^H R_v^{-1} s} = s^H R_v^{-1} s.$$

Thus, our optimal matched filter is

$$h = \alpha R_v^{-1} s.$$

We often choose to normalize the expected value of the power of the filter output due to the noise to unity. That is, we constrain

$$E\{|y_v|^2\} = 1.$$

This constraint implies a value of α, for which we can solve:

$$E\{|y_v|^2\} = \alpha^2 s^H R_v^{-1} s = 1,$$

yielding

$$\alpha = \frac{1}{\sqrt{s^H R_v^{-1} s}},$$

giving us our normalized filter,

$$h = \frac{1}{\sqrt{s^H R_v^{-1} s}} R_v^{-1} s.$$

If we care to write the impulse response of the filter for the convolution system, it is simply the complex conjugate time reversal of h.

Though we have derived the matched filter in discrete time, we can extend the concept to continuous-time systems if we replace R_v with the continuous-time autocorrelation function of the noise, assuming a continuous signal $s(t)$, continuous noise $v(t)$, and a continuous filter $h(t)$.

Alternative derivation of the matched filter

Alternatively, we may solve for the matched filter by solving our maximization problem with a Lagrangian. Again, the matched filter endeavors to maximize the output signal-to-noise ratio (*SNR*) of a filtered deterministic signal in stochastic additive noise. The observed sequence, again, is

$$x = s + v,$$

with the noise covariance matrix,

$$R_v = E\{vv^H\}.$$

The signal-to-noise ratio is

$$SNR = \frac{|y_s|^2}{E\{|y_v|^2\}}.$$

Evaluating the expression in the numerator, we have

$$|y_s|^2 = {y_s}^H y_s = h^H s s^H h.$$

and in the denominator,

$$E\{|y_v|^2\} = E\{{y_v}^H y_v\} = E\{h^H v v^H h\} = h^H R_v h.$$

The signal-to-noise ratio becomes

$$SNR = \frac{h^H s s^H h}{h^H R_v h}.$$

If we now constrain the denominator to be 1, the problem of maximizing *SNR* is reduced to maximizing the numerator. We can then formulate the problem using a Lagrange multiplier:

$$\begin{aligned}
&h^H R_v h = 1\\
&\mathcal{L} = h^H s s^H h + \lambda(1 - h^H R_v h)\\
&\nabla_{h^*}\mathcal{L} = s s^H h - \lambda R_v h = 0\\
&(s s^H) h = \lambda R_v h
\end{aligned}$$

which we recognize as an eigenvalue problem

$$h^H (s s^H) h = \lambda h^H R_v h = \lambda.$$

Since ss^H is of unit rank, it has only one nonzero eigenvalue. It can be shown that this eigenvalue equals

$$\lambda_{\max} = s^H R_v^{-1} s,$$

yielding the following optimal matched filter

$$h = \frac{1}{\sqrt{s^H R_v^{-1} s}} R_v^{-1} s.$$

This is the same result found in the previous section.

Frequency-domain interpretation

When viewed in the frequency domain, it is evident that the matched filter applies the greatest weighting to spectral components that have the greatest signal-to-noise ratio. Although in general this requires a non-flat frequency response, the associated distortion is not significant in situations such as radar and digital communications, where the original waveform is known and the objective is to detect the presence of this signal against the background noise.

Example of matched filter in radar and sonar

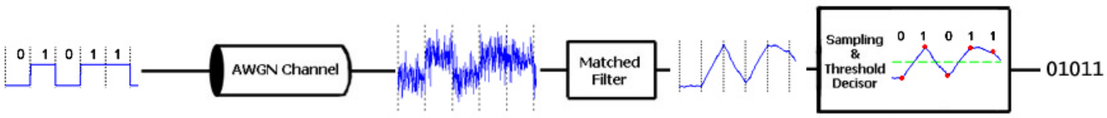
Matched filters are often used in signal detection. As an example, suppose that we wish to judge the distance of an object by reflecting a signal off it. We may choose to transmit a pure-tone sinusoid at 1 Hz. We assume that our received signal is an attenuated and phase-shifted form of the transmitted signal with added noise.

To judge the distance of the object, we correlate the received signal with a matched filter, which, in the case of white (uncorrelated) noise, is another pure-tone 1-Hz sinusoid. When the output of the matched filter system exceeds a certain threshold, we conclude with high probability that the received signal has been reflected off the object. Using the speed of propagation and the time that we first observe the reflected signal, we can estimate the distance of the object. If we change the shape of the pulse in a specially-designed way, the signal-to-noise ratio and the distance resolution can be even improved after matched filtering: this is a technique known as pulse compression.

Additionally, matched filters can be used in parameter estimation problems. To return to our previous example, we may desire to estimate the speed of the object, in addition to its position. To exploit the Doppler effect, we would like to estimate the frequency of the received signal. To do so, we may correlate the received signal with several matched filters of sinusoids at varying frequencies. The matched filter with the highest output will reveal, with high probability, the frequency of the reflected signal and help us determine the speed of the object. This method is, in fact, a simple version of the discrete Fourier transform (DFT). The DFT takes an N-valued complex input and correlates it with N matched filters, corresponding to complex exponentials at N different frequencies, to yield N complex-valued numbers corresponding to the relative amplitudes and phases of the sinusoidal components.

Example of matched filter in digital communications

The matched filter is also used in communications. In the context of a communication system that sends binary messages from the transmitter to the receiver across a noisy channel, a matched filter can be used to detect the transmitted pulses in the noisy received signal.



Imagine we want to send the sequence "0101100100" coded in non polar Non-return-to-zero (NRZ) through a certain channel.

Mathematically, a sequence in NRZ code can be described as a sequence of unit pulses or shifted rect functions, each pulse being weighted by +1 if the bit is "1" and by 0 if the bit is "0". Formally, the scaling factor for the k^{th} bit is,

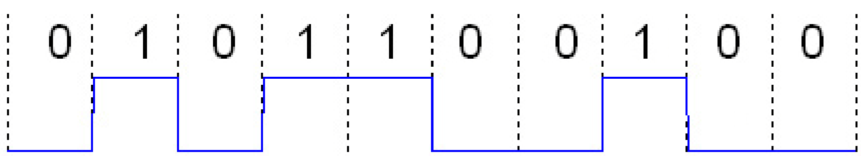
$$a_k = \begin{cases} 1, & \text{if bit } k \text{ is } 1, \\ 0, & \text{if bit } k \text{ is } 0. \end{cases}$$

We can represent our message, $M(t)$, as the sum of shifted unit pulses:

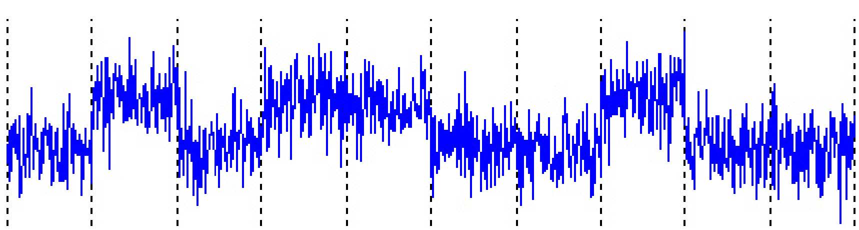
$$M(t) = \sum_{k=-\infty}^{\infty} a_k \times \Pi\left(\frac{t-kT}{T}\right).$$

where T is the time length of one bit.

Thus, the signal to be sent by the transmitter is



If we model our noisy channel as an AWGN channel, white Gaussian noise is added to the signal. At the receiver end, for a Signal-to-noise ratio of 3dB, this may look like:



A first glance will not reveal the original transmitted sequence. There is a high power of noise relative to the power of the desired signal (i.e., there is a low signal-to-noise ratio). If the receiver were to sample this signal at the correct moments, the resulting binary message would possibly belie the original transmitted one.

To increase our signal-to-noise ratio, we pass the received signal through a matched filter. In this case, the filter should be matched to an NRZ pulse (equivalent to a "1" coded in NRZ code). Precisely, the impulse response of the ideal matched filter, assuming white (uncorrelated) noise should be a time-reversed complex-conjugated scaled version of the signal that we are seeking. We choose

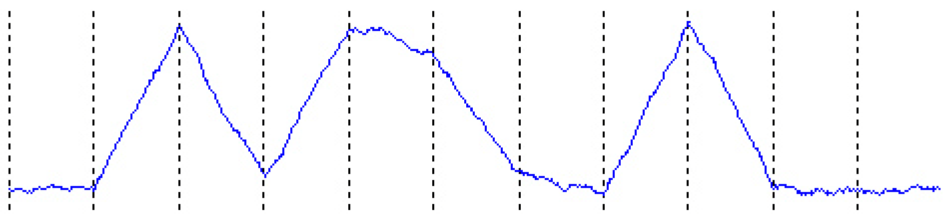
$$h(t) = \Pi\left(\frac{t}{T}\right).$$

In this case, due to symmetry, the time-reversed complex conjugate of $h(t)$ is in fact $h(t)$, allowing us to call $h(t)$ the impulse response of our matched filter convolution system.

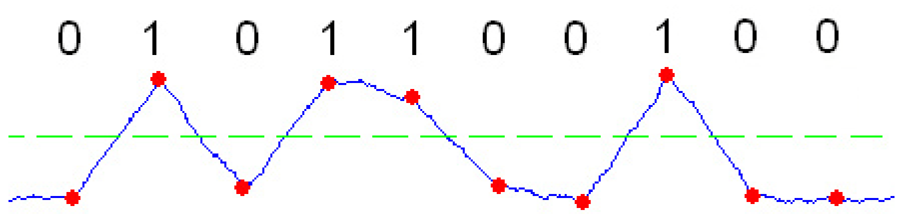
After convolving with the correct matched filter, the resulting signal, $M_{\text{filtered}}(t)$ is,

$$M_{\text{filtered}}(t) = M(t) * h(t)$$

where * denotes convolution.



Which can now be safely sampled by the receiver at the correct sampling instants, and compared to an appropriate threshold, resulting in a correct interpretation of the binary message.



Chapter-4

Pulse Compression

Pulse compression is a signal processing technique mainly used in radar, sonar and echography to augment the range resolution as well as the signal to noise ratio. This is achieved by modulating the transmitted pulse and then correlating the received signal with the transmitted pulse.

Simple pulse

Signal description

The simplest signal a pulse radar can transmit is a sinusoidal pulse of amplitude, $A$and carrier frequency, f_0, truncated by a rectangular function of width, T. The pulse is transmitted periodically, but that is not the main topic here; we will consider only a single pulse, s. If we assume the pulse to start at time $t=0$, the signal can be written the following way, using the complex notation:

$$s(t) = \begin{cases} Ae^{2i\pi f_0 t} & \text{if } 0 \leq t < T \\ 0 & \text{otherwise} \end{cases}$$

Range resolution

Let us determine the range resolution which can be obtained with such a signal. The return signal, written $r(t)$, is an attenuated and time-shifted copy of the original transmitted signal (in reality, Doppler effect can play a role too, but this is not important here.) There is also noise in the incoming signal, both on the imaginary and the real channel, which we will assume to be white and Gaussian (this generally holds in reality); we write $B(t)$to denote that noise. To detect the incoming signal, matched filtering is commonly used. This method is optimal when a known signal is to be detected among an additive white Gaussian noise.

In other words, the cross-correlation of the received signal with the transmitted signal is computed. This is achieved by convolving the incoming signal with a conjugated and time-reversed version of the transmitted signal. This operation can be done either in software or with hardware. We write $<s,r>(t)$ for this cross-correlation. We have:

$$< s, r > (t) = \int_{t'=0}^{+\infty} s^\star(t') r(t+t') dt'$$

If the reflected signal comes back to the receiver at time t_r and is attenuated by factor K, this yields:

$$r(t) = \begin{cases} KAe^{2i\pi f_0 (t - t_r)} + B(t) & \text{if } t_r \leq t < t_r + T \\ B(t) & \text{otherwise} \end{cases}$$

Since we know the transmitted signal, we obtain:

$$< s, r > (t) = KA^2 \Lambda \left(\frac{t - t_r}{T} \right) e^{2i\pi f_0 (t - t_r)} + B'(t)$$

where $B'(t)$, the result of the intercorrelation between the noise and the transmitted signal, remains a white noise of same characteristics as $B(t)$ since it is not correlated to the transmitted signal. Function Λ is the triangle function, its value is 0 on $[-\infty, -\frac{1}{2}] \cup [\frac{1}{2}, +\infty]$, it augments linearly on $[-\frac{1}{2}, 0]$ where it reaches its maximum 1, and it decreases linearly on $[0, \frac{1}{2}]$ until it reaches 0 again. Figures at the end of this paragraph show the shape of the intercorrelation for a sample signal (in red), in this case a real truncated sine, of duration $T = 1$ seconds, of unit amplitude, and frequency $f_0 = 10$ hertz. Two echoes (in blue) come back with a delay of 3 and 5 seconds, respectively, and have an amplitude equal to 0.5 and 0.3; those are just random values for the sake of the example. Since the signal is real, the intercorrelation is weighted by an additional ½ factor.

If two pulses come back (nearly) at the same time, the intercorrelation is equal to the sum of the intercorrelations of the two elementary signals. To distinguish one "triangular" envelope from that of the other pulse, it is clearly visible that the times of arrival of the two pulses must be separated by at least T so that the maxima of both pulses can be separated. If this condition is false, both triangles will be mixed together and impossible to separate.

Since the distance travelled by a wave during T is cT (where *c* is the speed of the wave in the medium), and since this distance corresponds to a round-trip time, we get:

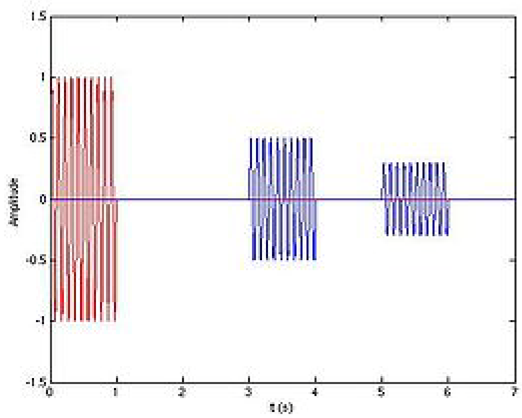
Result 1

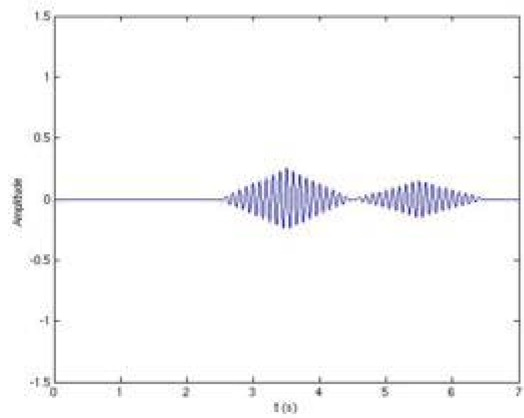
The range resolution with a sinusoidal pulse is $\frac{1}{2}cT$ where T is the pulse Duration and, c, the speed of the wave.

Conclusion: to augment the resolution, the pulse length must be reduced.

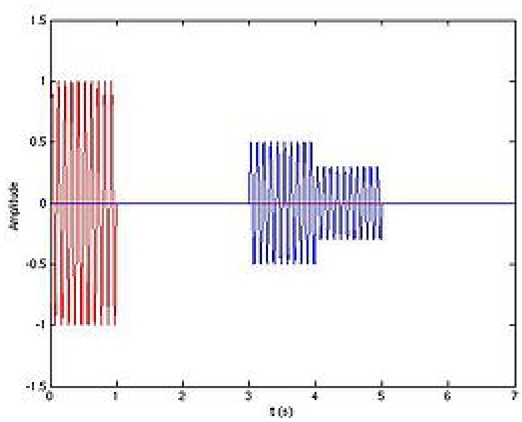
Example (simple impulsion): transmitted signal in red (carrier 10 hertz, amplitude 1, duration 1 second) and two echoes (in blue).

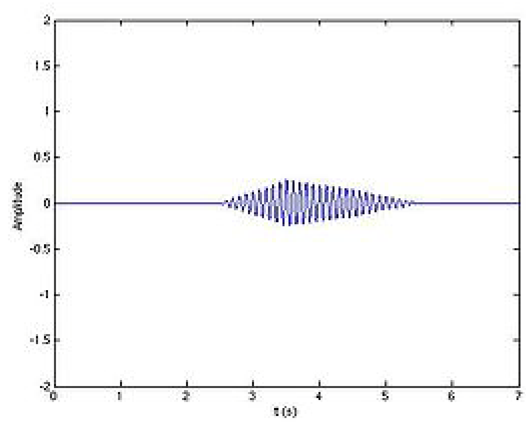
Before matched filtering **After matched filtering**





If the targets are separated enough... ...echoes can be distinguished.





If the targets are too close... ...the echoes are mixed together.

Required energy to transmit that signal

The instantaneous power of the transmitted pulse is $P(t) = |s|^2(t)$. The energy put into that signal is:

$$E = \int_0^T P(t)dt = A^2T$$

Similarly, the energy in the received pulse is $E_r = K^2A^2T$. If σ is the standard deviation of the noise, the signal-to-noise ratio (SNR) at the receiver is:

$$SNR = \frac{E_r}{\sigma} = \frac{K^2 A^2 T}{\sigma}$$

The SNR augments with the pulse duration, if other parameters are frozen. This goes against the resolution requirements, since generally one wants a large resolution.

Pulse compression by linear frequency modulation (or chirping*)*

Basic principles

How can one have a large enough pulse (to still have a good SNR at the receiver) without poor resolution? This is where pulse compression enters the picture. The basic principle is the following:

- a signal is transmitted, with a long enough length so that the energy budget is correct
- this signal is designed so that after matched filtering, the width of the intercorrelated signals is smaller than the width obtained by the standard sinusoidal pulse, as explained above (hence the name of the technique: pulse compression).

In radar or sonar applications, linear chirps are the most typically used signals to achieve pulse compression. The pulse being of finite length, the amplitude is a rectangle function. If the transmitted signal has a duration T, begins at $t=0$and linearly sweeps the frequency band Δfcentered on carrier f_0, it can be written:

$$s_c(t) = \begin{cases} A e^{2i\pi\left(f_0 + \frac{\Delta f}{2T} t - \frac{\Delta f}{2}\right)t} & \text{if } 0 \le t < T \\ 0 & \text{otherwise} \end{cases}$$

N.b., the chirp is written that way so the phase of the chirped signal (that is, the argument of the complex exponential), is:

$$\phi(t) = 2\pi \left(f_0 + \frac{\Delta f}{2T} t - \frac{\Delta f}{2}\right) t$$

thus the instantaneous frequency is (by definition):

$$f(t) = \frac{1}{2\pi} \left[\frac{d\phi}{dt}\right]_t = f_0 - \frac{\Delta f}{2} + \frac{\Delta f}{T} t$$

which is the intended linear ramp going from $f_0 - \frac{\Delta f}{2}$at $t=0$to $f_0 + \frac{\Delta f}{2}$at $t=T$.

Cross-correlation between the transmitted and the received signal

As for the "simple" pulse, let us compute the cross-correlation between the transmitted and the received signal. To simplify things, we shall consider that the chirp is not written as it is given above, but in this alternate form (the final result will be the same):

$$s_{c'}(t) = \begin{cases} Ae^{2i\pi\left(f_0 + \frac{\Delta f}{2T}t\right)t} & \text{if } -\frac{T}{2} \le t < \frac{T}{2} \\ 0 & \text{else} \end{cases}$$

Since this cross-correlation is equal (save for the *K* attenuation factor), to the autocorrelation function of $s_{c'}$, this is what we consider:

$$< s_{c'}, s_{c'} > (t) = \int_{-\infty}^{+\infty} s_{c'}^{\star}(-t') s_{c'}(t - t') dt'$$

It can be shown that the autocorrelation function of $s_{c'}$ is:

$$< s_{c'}, s_{c'} > (t) = T\Lambda\left(\frac{t}{T}\right) sinc\left[\pi\Delta f t\Lambda\left(\frac{t}{T}\right)\right] e^{2i\pi f_0 t}$$

The maximum of the autocorrelation function of $s_{c'}$ is reached at 0. Around 0, this function behaves as the sinc term. The −3 dB temporal width of that cardinal sine is more or less equal to $T' = \frac{1}{\Delta f}$. Everything happens as if, after matched filtering, we had the resolution that would have been reached with a simple pulse of duration T'. For the common values of Δf, T' is smaller than T, hence the *pulse compression* name.

Since the cardinal sine can have annoying sidelobes, a common practice is to filter the result by a window (Hamming, Hann, etc.). In practice, this can be done at the same time as the adapted filtering by multiplying the reference chirp with the filter. The result will be a signal with a slightly lower maximum amplitude, but the sidelobes will be filtered out, which is more important.

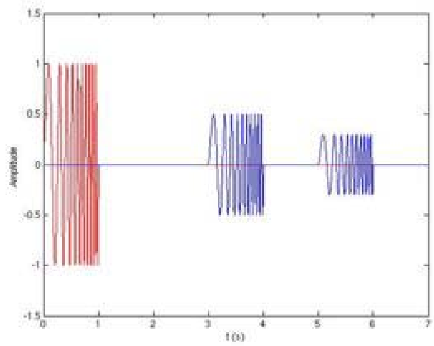
Result 2

The distance resolution reachable with a linear frequency modulation of a pulse on a bandwidth Δf is: $\frac{c}{2\Delta f}$ where c is the speed of the wave.

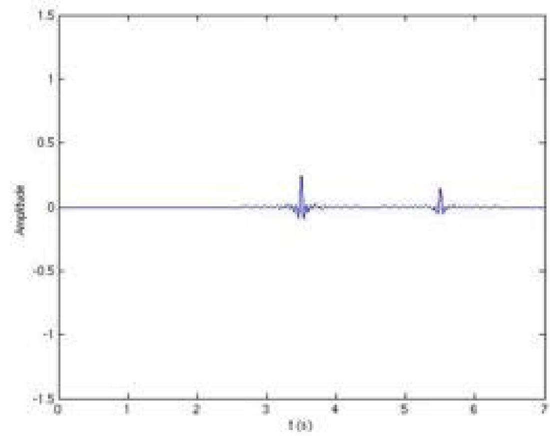
Definition

Ratio $\frac{T}{T'} = T\Delta f$ is the pulse compression ratio. It is generally greater than 1 (usually, its value is 20 to 30).

Example (chirped pulse): transmitted signal in red (carrier 10 hertz, modulation on 16 hertz, amplitude 1, duration 1 second) and two echoes (in blue).



Before matched filtering



After matched filtering: the echoes are shorter in time.

SNR augmentation through pulse compression

The energy of the signal does not vary during pulse compression. However, it is now located in the main lobe of the cardinal sine, whose width is approximately $T' \approx \frac{1}{\Delta f}$. If $P$is the power of the signal before compression, and $P'$the power of the signal after compression, we have:

$$P \times T = P' \times T'$$

which yields:

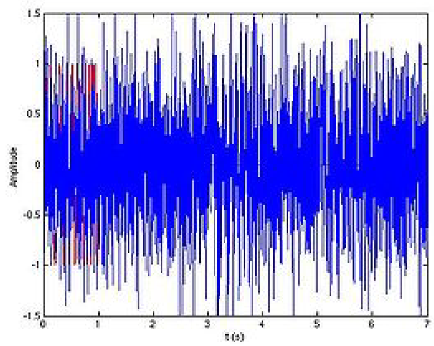
$$P' = P \times \frac{T}{T'}$$

Besides, the power of the noise does not change through intercorrelation since it is not correlated to the transmitted pulse (it is totally random). As a consequence:

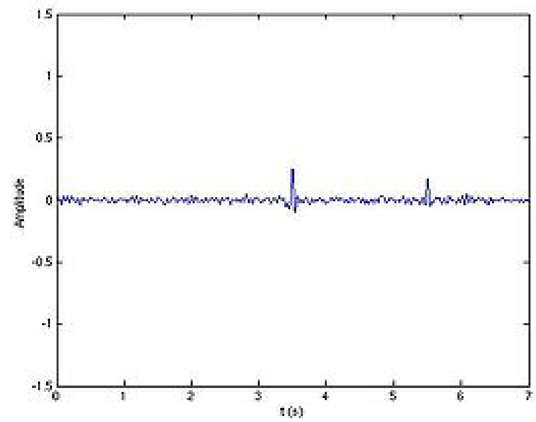
Result 3

After pulse compression, the power of the received signal can be considered as being amplified by $T\Delta f$. This additional gain can be injected in the radar equation.

Example: same signals as above, plus an additive Gaussian white noise ($\sigma = 0.5$)



Before matched filtering: the signal is hidden in noise



After matched filtering: echoes become visible.

Pulse compression by phase coding

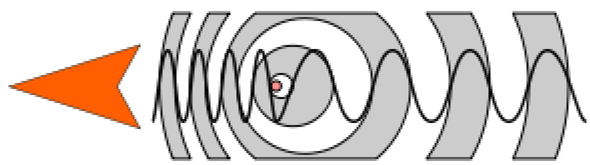
There are other means to modulate the signal. Phase modulation is a commonly used technique; in this case, the pulse is divided in N time slots of duration $\frac{T}{N}$ for which the phase at the origin is chosen according to a pre-established convention. For instance, it is possible not to change the phase for some time slots (which comes down to just leave the signal as it is, in those slots) and de-phase the signal in the other slots by π (which is equivalent of changing the sign of the signal). The precise way of choosing the sequence of $\{0, \pi\}$ phases is done according to a technique known as Barker codes. It is possible to code the sequence on more than two phases (polyphase coding). As with a linear chirp, pulse compression is achieved through intercorrelation.

The advantages of the Barker codes are their simplicity (as indicated above, a π de-phasing is a simple sign change), but the pulse compression ratio is lower than in the chirp case and the compression is very sensitive to frequency changes due to the Doppler effect if that change is larger than $\frac{1}{T}$.

Chapter-5

Doppler Radar and Moving Target Indication

Doppler radar



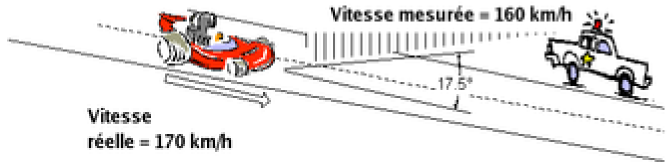
Doppler effect

A **Doppler radar** is a specialized radar that makes use of the doppler effect to produce velocity data about objects at a distance. It does this by beaming a microwave signal towards a desired target and listening for its reflection, then analyzing how the frequency of the returned signal has been altered by the object's motion. This variation gives direct and highly accurate measurements of the radial component of a target's velocity relative to the radar. Doppler radars are used in aviation, sounding satellites, meteorology, police speed guns, and radiology.

The specific term "*Doppler Radar*", due in part to its extremely common use by television meteorologists in on-air weather reporting, has erroneously become popularly synonymous with the type of radar used in meteorology. Most modern weather radars use the pulse-doppler technique to examine the motion of precipitation, but it is only a part of the processing of their data. So, while these radars use a highly specialized form of *doppler radar*, the term is much broader in its meaning and its applications.

Concept

Doppler Effect



The emitted signal toward the car is reflected back with a variation of frequency that depend on the speed away/toward the radar (160 km/h). This is only a component of the real speed (170 km/h).

The Doppler effect (or Doppler shift), named after Austrian physicist Christian Doppler who proposed it in 1842, is the change in frequency of a wave for an observer moving relative to the source of the waves. It is commonly heard when a vehicle sounding a siren approaches, passes and recedes from an observer. The received frequency is increased (compared to the emitted frequency) during the approach, it is identical at the instant of passing by, and it is decreased during the recession. This variation of frequency also depends on the direction the wave source is moving with respect to the observer; it is maximum when the source is moving directly toward or away from the observer, and diminishes with increasing angle between the direction of motion and the direction of the waves, until when the source is moving at right angles to the observer, there is no shift.

An analogy would be pitcher throwing one ball every second in a person's direction (a frequency of 1 ball per second). Assuming that the balls travel at a constant velocity, if the pitcher is stationary, the man will catch one ball every second. However, if the pitcher is jogging towards the man, he will catch balls more frequently because the balls will be less spaced out (the frequency increases). The inverse is true if the pitcher is moving away from the man; he will catch balls less frequently due to the pitcher's backward motion (the frequency decreases). If the pitcher were to move at an angle but with the same speed, the variation of the frequency at which the receiver would catch the ball would be less as the distance between the two would change more slowly.

Note that, from the point of view of the pitcher, the frequency remains constant (whether he's throwing balls or transmitting microwaves). Since with electromagnetic radiation like microwaves frequency is inversely proportional to wavelength, the wavelength of the waves is also affected. Thus, the relative difference in velocity between a source and an observer is what gives rise to the doppler effect.

Frequency variation

The exact formula for radar doppler shift is the same as that for reflection of light by a moving mirror. There is no need to invoke Einsteins' theory of special relativity, because all observations are made in the same frame of reference. The exact result derived with c as the speed of light and v as the target velocity gives the shifted frequency (F_r) as a function of the original frequency (F_t) :

$$F_r = F_t(\frac{1+v/c}{1-v/c})$$

The exact "beat frequency", aka Doppler Frequency (F_d), is thus:

$$F_d = F_r - F_t = 2v\frac{F_t}{(c-v)}$$

Since for most all practical applications of radar, $v \ll c$, so $(c-v) \to c$. We can then write:

$$F_d \approx 2v\frac{F_t}{c}$$

Technology

U.S. Army soldier using a radar gun, an application of Doppler radar, to catch speeding violators.

There are three ways of producing the Doppler effect. Radars may be Coherent pulsed (CP), Continuous wave (CW), or Frequency modulated (FM). CW doppler radar only provides a velocity output as the received signal from the target is compared in frequency with the original signal. Early doppler radars were CW, but these quickly led to the development of frequency modulated continuous wave(FM-CW) radar, which sweeps the transmitter frequency to encode and determine range.

The CW and FM-CW radars can normally only process one target, which limits their use. With the advent of digital techniques, Pulse-Doppler radars (PD) were introduced, and doppler processors for coherent pulse radars were developed at the same time. The advantage of combining doppler processing with pulse radars is to provide accurate velocity information. This velocity is called Range-Rate. It describes the rate that a target moves towards or away from the radar. A target with no range-rate reflects a frequency near the transmitter frequency, and cannot be detected. The classic zero doppler target is one which is on a heading that is tangential to the radar antenna beam. Basically, any target that is heading 90 degrees in relation to the antenna beam cannot be detected by its velocity (only by its conventional reflectivity).

History

FM radar was highly developed during World War II for the use by US Navy aircraft. Most used the UHF spectrum, and had a transmit yagi antenna on the port wing, and a receiver yagi antenna on the starboard wing. This allowed bombers to fly an optimum speed when approaching ship targets. Later when magnetrons and microwaves became available, the use of FM radar fell into disuse.

When the digital Fast Fourier transform became available, it was immediately connected to Coherent Pulsed radars, where velocity information was extracted. This quickly proved useful in both weather and air traffic control radars. The velocity information provided another input to the software tracker, and improved computer tracking. Due to the low *pulse repetition frequency* (PRF) of most coherent pulsed radars, which maximizes the coverage in range, the amount of doppler processing is limited. The doppler processor can only process velocities up to ±1/2 the PRF of the radar. This was not a problem for weather radars.

Specialized radars quickly were mechanized when digital techniques became affordable. Pulse-Doppler radars combine all the benefits of long range, and high velocity capability. Pulse-Doppler radars use a medium to high PRF (on the order of 30 kHz). This high PRF allows for the detection of either high speed targets, or high resolution velocity measurements. Normally it is one or the other, that is, a radar designed for detecting targets from zero to Mach 2, does not have a high resolution in speed, while a radar designed for high resolution velocity measurements does not have a wide range of speeds. Weather radars are high resolution velocity radars, while air defense radars have a large range of velocity detection, but the accuracy in velocity is in the 10's of knots.

Antenna designs for the CW and FM-CW started out as separate transmit and receive antennas before the advent of affordable microwave designs. In the late 1960s traffic radars began being produced which used a single antenna. This was made possible by the use of circular polarization, and a multi-port waveguide section operating at X band. By the late 1970s this changed to linear polarization and the use of ferrite circulators at both X and K bands. PD radars operate at too high a PRF to use a Transmit-Receive gas filled switch, and most use solid-state devices to protect the receiver Low Noise Amplifier when the transmitter is fired.

Moving target indication

Moving target indication (**MTI**) is a mode of operation of a radar to discriminate a target against clutter.

In contrast to another mode, stationary target indication, it takes an advantage of the fact that the target moves with respect to stationary clutter. The most common approach is taking an advantage of the Doppler effect. For a sequence of radar pulses the moving target will be at different distance from the radar and the phase of the radar return from the target will be different for successive pulses, while the returns from stationary clutter will arrive at the same phase shift.

Radar MTI may be specialized in terms of the type of clutter and environment: airborne MTI (**AMTI**), ground MTI (**GMTI**), etc., or may be combined mode: stationary and moving target indication (**SMTI**).

Design Parameters

Basic Geometry

A platform is going at velocity v_p with a maximum range R_{max} with elevation angle EL and azimuth AZ to the target.

Probability of Detection (Pd)

The probability of detecting a given target at a given range any time the radar beam scans across it, Pd is determined by factors that include the size of the antenna and the amount of power it radiates. A large antenna radiating at high power provides the best performance. For high quality information on moving targets the Pd must be very high.

Target Location Accuracy

Location accuracy is a function of platform self-location performance, radar-pointing accuracy, azimuth resolution, and range resolution. A long antenna or very short wave length can provide fine azimuth resolution. Short antennas tend to have a larger azimuth error, an error that increases with range to the target because signal-to-noise ratio varies inversely with range. Location accuracy is vital to tracking performance because it prevents track corruption when there are multiple targets and makes it possible to determine which road a vehicle is on if it is moving in an area with many roads.

The target location accuracy is proportional to the slant range, frequency and aperture length.

Target Range Resolution (High Range Resolution or HRR)

Target range resolution determines whether two or more targets moving in close proximity will be detected as individual targets. With higher performance radars, target range resolution—known as High Range Resolution (HRR)—can be so precise that it may be possible to recognize a specific target (i.e., one that has been seen before) and to place it in a specific class (e.g., “a T-80 tank”). This would allow more reliable tracking of specific vehicles or groups of vehicles, even when they are moving in dense traffic or disappear for a period due to screening.

Minimum Detectable Velocity MDV

The MDV comes from the frequency spread of the mainlobe clutter. MDV determines whether the majority of military traffic, which often moves very slowly, especially when traveling off-road, will be detected. A GMTI radar must distinguish a moving target from ground clutter by using the target’s Doppler signature to detect the radial component of the target’s velocity vector (i.e., by measuring the component of the target’s movement directly along the radar-target line). To capture most of this traffic, even when it is moving almost tangentially to the radar (i.e., perpendicular to the radar-target line), a system must have the ability to detect very slow radial velocities. As the radial component of a target’s velocity approaches zero, the target will fall into the clutter or *blind zone*. This is calculated as:

$$MDV = \frac{\lambda}{2}\left(\frac{4v_p}{B}\sqrt{(sin(AZ)sin(EL))^2 + (cos(AZ)cos(EL))^2}\right)$$

Any target with a velocity less than this minimum (MDV) cannot be detected because there is not sufficient Doppler shift in its echo to separate it from the mainlobe clutter return.

Area Search Rate

The area coverage rate (measured in area per unit time) is proportional to system power and aperture size. Other factors which may be relevant include grid spacing, size of the power amp, module quantization, the number of beams processed and system losses.

Stand-off Distance

Stand-off distance is the distance separating a radar system from the area it is covering.

Coverage Area Size (breadth and depth)

Coverage area size is the area that the system can keep under continuous surveillance from a specific orbit. Well known design principles cause a radar’s maximum detection range to depend on the size of its antenna (radar aperture), the amount of power radiated from the antenna, and the effectiveness of its clutter cancellation mechanism. The earth’s curvature and screening from terrain, foliage, and buildings cause system altitude to be another key factor determining depth of coverage. The ability to cover an area the size of an army corps commander's area of interest from a safe stand-off distance is the hallmark of an effective, advanced GMTI system.

Coverage Area Revisit Rate

This equates to the frequency with which the radar beam passes over a given area. Frequent revisits are very important to the radar's ability to achieve track continuity and contribute to an increased probability of target detection by lessening the chance of obscuration from screening by trees, buildings, or other objects. A fast revisit rate becomes critical to providing an uncorrupted track when a target moves in dense traffic or is temporarily obscured, if only by trees along a road.

Chapter-6

Pulse-Doppler Radar

Pulse-Doppler is a radar system capable of not only detecting target location (bearing, range, and altitude), but also measuring its radial velocity (range-rate). It uses the Doppler effect to determine the relative velocity of objects; pulses of RF energy returning from the target are processed to measure the phase shift between carrier cycles successive pulses at the transmitted frequency. To achieve this, the transmitter frequency source must have very good phase stability and the system is said to be coherent.

They are used in different fields. In meteorological radars, pulse-Doppler measures instantaneous speed at discrete range intervals to detect wind field in weather as the beam is slewed across the sky. Doppler On Wheels, NEXRAD, Terminal Doppler Weather Radar, and ARMOR Doppler Weather Radar are examples. In search and track radars, the purpose is to measure the speed of all targets and eliminate environmental reflections from weather, the surface of the earth, and biological objects like birds, which hides target signals, by their characteristic velocity.

Pulse repetition frequency

Pulse-Doppler typically uses medium Pulse repetition frequency (PRF) from about 3 kHz to 30 kHz. Systems using PRF below 3 kHz are considered low PRF because direct range can be measured to a range of 50 km, and Doppler processing becomes an increasing challenge due to coherency limitations as PRF falls below 3 kHz. Systems using PRF above 30 kHz function better as interrupted CW radar because direct velocity can be measured up to 4.5 km/s at L band, and range measurement becomes more of a challenge as PRF increases above 30 kHz. Range and velocity cannot be measured directly using medium PRF, and a technique called ambiguity resolutions is used to identify range.

Underlying principle

Pulse-Doppler radar is based on the Doppler effect, where movement in range produces frequency shift on the signal reflected from the target.

$$DopplerFrequency = 2F_o \left(\frac{RangeVelocity}{C} \right)$$

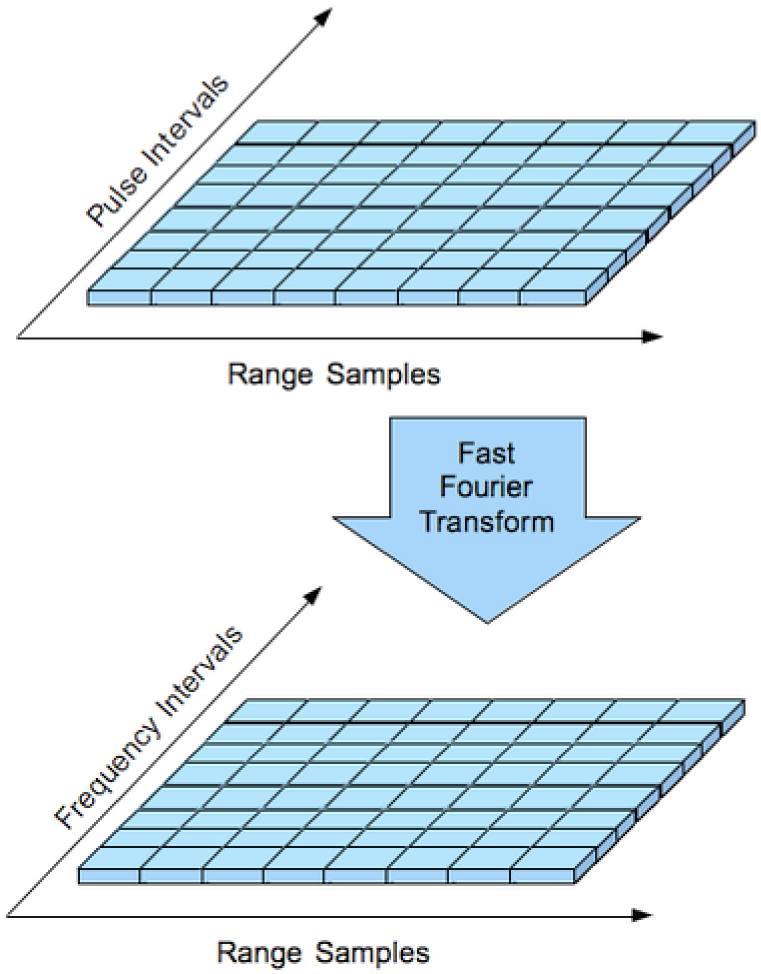
In some systems, the velocity measurement is compared to the change in range to determine if the detected signal is an airborne vehicle or an electronic artifact. Angular measurement is made using monopulse scanning and sidelobe blanking as the radar beam is scanned across the sky.

Multiple receive samples are taken between transmit pulses. Each sample includes an in-phase (I) value and a quadrature (Q) value. I an Q are the real and imaginary component of a complex number.

A spinning wheel, mirror and strobe-light can be used to visualize I and Q. The mirror is placed at a 45 degree angle above the wheel so that you can see the front and top of the wheel at the same time. The strobe-light is attached to the wheel so that you can see the wheel spin when the room lights are turned off. You sit directly in front of the wheel while a friend spins the wheel. The view of the front of the wheel (I) and the top of the wheel (Q) tell you whether your friend has spun the wheel clockwise or counterclockwise. Clockwise is like inbound Doppler. Counterclockwise is like outbound Doppler.

Pairs of I and Q values form a complex number explained as real and imaginary parts. I is the real components. Q is the imaginary component.

Signal Processing



Pulse-Doppler signal processing. The *Range Sample* axis represents individual samples taken in between each transmit pulse. The *Range Interval* axis represents each successive transmit pulse interval during which samples are taken. The Fast Fourier Transform process converts time-domain samples into frequency domain spectra.

I(t) and Q(t) are both required during the sampling process so that radar signal processing can include information about closing (approaching) versus opening (leaving) Doppler velocities. This is crucial for proper operation of pulse-Doppler systems.

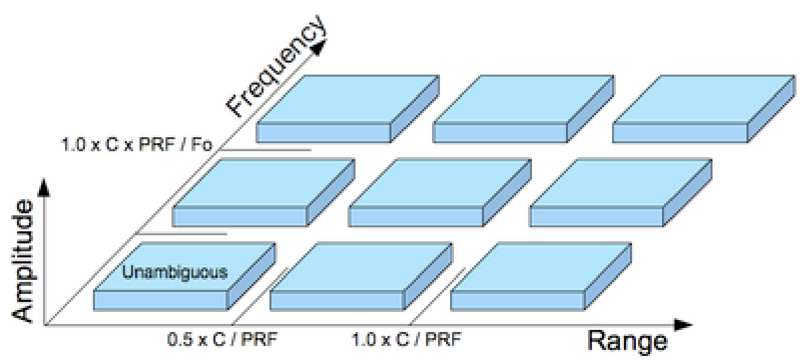
Pulse-Doppler relies on taking up to a few dozen individual I and Q samples in between each transmit pulse. Each different range sample has its own separate processing, and these time samples are converted to frequency domain using a digital filter, usually involving a fast Fourier transform (FFT). Side-lobes are produced during signal processing and a side-lobe suppression strategy, such as Dolph-Chebyshev Windowing, is commonly used to reduce false alarms .

The digital filter produces as many frequency outputs as there are time inputs. Production of one FFT requires as many samples as the number of points in the FFT (a 256 point FFT requires 256 transmit pulses). Each separate spectrum undergoes separate detection processing.

Pulse-Doppler typically uses a form of constant false alarm rate detection processing to identify target signals. Detection processing for pulse-Doppler produces an ambiguous range and ambiguous velocity corresponding to one of the FFT outputs from one of the range samples. The reflection from clutter falls into filters corresponding to a different frequency than the filter corresponding to most aircraft. Sub-clutter visibility in a pulse-Doppler radar is a measure of the power difference between the smallest signals that can be detected in the presence of a large signal at a different frequency. This Measure of Performance is called Sub-clutter Visibility . Sub-clutter visibility for pulse-Doppler is typically over 50dB. Conventional non-coherent radar using Moving target indication provides up to about 25dB sub-clutter visibility. Pulse-Doppler is used in many military applications for this reason.

Pulse-Doppler signals are audible, so a helicopter sounds like a helicopter and a jet sounds like a jet. The actual size of the target can be calculated using these audible signals, which is impossible with pulse compression and low PRF systems. Audible Doppler and target size support passive vehicle type classification when identification friend or foe is not available. This is another reason pulse-Doppler has military applications.

Ambiguity Resolution



Pulse-Doppler ambiguity zones. Each blue zone with no label represents a velocity/range combination that will be folded into the unambiguous zone. Areas outside the blue zones are blind ranges and blind velocities. Blind velocities correspond with the clutter notch.

Pulse Doppler relies on medium pulse repetition frequency (PRF) from about 3 kHz to 30 kHz. Each transmit pulse is separated by between 5 km and 50 km of distance. Operation in an environment larger than 5 km to 50 km means reflections arrive at the antenna from multiple ranges simultaneously. The true range and speed of the target is effectively folded by a modulo operation produced by the transmit and receive process.

Range Ambiguity Resolution

True range is found using two different PRF. Each pair of PRF is a separate detection scheme with unique mathematical properties that are configured into the radar. Target detection requires samples from both PRF, which are transmitted alternating back and forth rapidly during the search process.

The time between transmit pulses is split into several range samples. Each range sample contains the receive signal from a specific range interval. This is called sampling.

This is explained best using the following example, where PRF A produces a transmit pulse every 6 km and PRF B produces a transmit pulse every 5 km.

Transmit Sample 1 Sample 2 Sample 3 Sample 4 Sample 5

Target PRF A

Target PRF B

The apparent range for PRF A falls in the 2 km sample, and the apparent range for PRF B falls in the 4 km sample. This combination places the true target distance at 14 km (2x6+2 or 2x5+4). This can be seen graphically when range intervals are stacked end-to-end as shown below.

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 29 29

A A A A A

B B B B B B

"A" represents target range possibilities for PRF A, and "B" represents target range possibilities for PRF B.

This process is usually automated with a look-up table, and the size of this table limits the maximum range. A convolution algorithm may be used instead of a table.

Each individual PRF has blind ranges, where the transmitter pulse occurs at the same time as the target reflection signal arrives back at the radar. A third and fourth PRF are required to complete a detection by repeating the process described above. PRF alternates rapidly while scanning to fill in any blind ranges.

Multiple target reflections can corrupt the range ambiguity resolution process. Frequency ambiguity resolution improves performance for multiple target scenarios by separating target signals using speed differences.

Frequency Ambiguity Resolution

Each range samples is converted from time domain I/Q samples into frequency domain. Older systems use individual filters for this. Newer systems use digital sampling and a Fast Fourier transform or Discrete Fourier transform instead of physical filters. Each filter converts time samples into a frequency spectrum. Each spectrum frequency corresponds with a different speed. The ambiguous velocity is as follows.

$$AmbiguousVelocity = -0.5\left(\frac{DopplerFrequency * C}{TransmitFrequency}\right)$$

Frequency is folded for high speed targets where radial velocity produces a frequency shift above the Nyquist frequency. The true speed of the target may be folded by a modulo operation produced by the sampling process.

$$TrueVelocity = AmbiguousVelocity + 0.5N\left(\frac{PRF * C}{TransmitFrequency}\right)$$

$$N = IntegerBetween \pm \left(\frac{0.5Bandwidth}{PRF}\right)$$

The Nyquist frequency will also change when the PRF is changed. The target frequency will shift if the speed is greater than the Nyquist frequency. This shift can be used to establish the frequency interval, much the same way as described above for range ambiguity resolution.

A blind velocity occurs when Doppler frequency falls close to the PRF. This folds the return signal into the same filter as stationary clutter reflections. Rapidly alternating different PRF while scanning eliminates blind frequencies.

Special Consideration

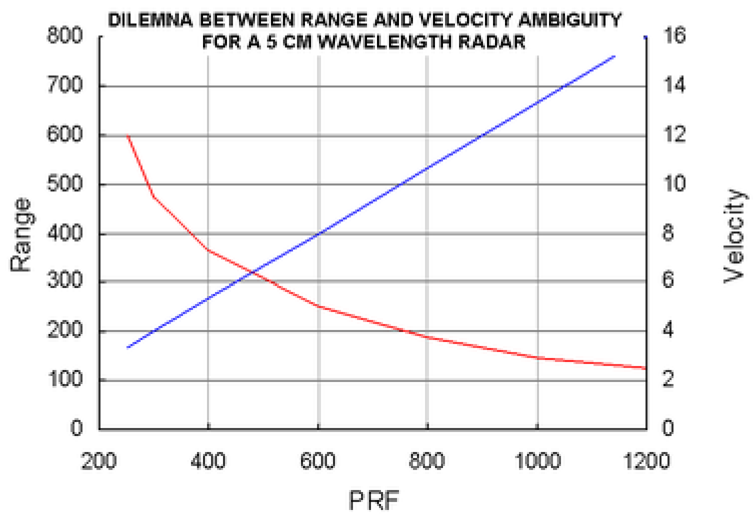
Pulse-Doppler radar has special requirements that must be satisfied to achieve acceptable performance.

Coherency

Pulse-Doppler radar requires an oscillator with very little noise. The amount of tolerable FM or phase noise depends upon the size of the filters used for signal processing. If range samples are split into 100 different frequencies using a PRF of 10 kHz, then the tolerable limit for FM transmit noise requires signal no greater than the sub-clutter visibility reduction factor 100 Hz away from the carrier frequency, and this must be maintained over a time span of no less than 10ms.

FM noise above this threshold paints stationary objects with false Doppler signal, which reduces sub-clutter visibility performance. Noise reduction includes shaping the transmit pulse to reduce ringing artifacts. Pulse Doppler requires coherent amplifiers, like solid state, klystron or traveling wave tube. Extra FM noise originating from transmitter amplification will induce apparent motion on stationary objects which will reduce performance, so magnetron and crossed-field amplifier are not appropriate because noise introduced by these devices interfere with detection performance.

In order for Pulse-Doppler radar to be effective, it is essential that the received echoes are coherent with the carrier signal, at least during the sampling period. To achieve this, a number of techniques are employed, the most common being that the transmitter signal is derived from a highly stable oscillator (the COHO) and the received signal is demodulated using an equally stable local oscillator (the STALO), which is phase locked to it. Doppler shift may then be accurately resolved by comparing the frequency components of the returned echo with the frequency components of the transmitted signal.



Maximum range from reflectivity (red) and unambiguous Doppler velocity range (blue) with a fix pulse repetition rate.

Scalloping

Scalloping is a phenomenon unique to two-PRF detection schemes. This creates **detection gaps** with a pattern of discrete ranges, each of which has a blind velocity.

Objects moving at the radial velocity that produces a harmonic of the PRF produce zero Doppler at the receiver. This is the blind velocity for a PRF. The target will become invisible in a two PRF detection scheme when the target reflection arrives during the transmit pulse on the other PRF.

Assume two pulse frequencies PRF-A and PRF-B. There will be blind zones at every range for PRF-A where there is a transmit pulse at the same time as when a signal is arriving. PRF-B will produce detections in those blind ranges, except that PRF-B will have blind velocities different from PRF-A.

This leaves intermittent detection gaps at discrete velocities in each blind range. Aircraft with a Doppler frequency that is a harmonic of PRF-B will not be detectable at blind ranges for PRF-A. Aircraft with a Doppler frequency that is a harmonic of PRF-A will not be detectable at blind ranges for PRF-B.

These detection gaps are filled in by using three or more alternating PRF in the detection scheme. The other alternative is to increase the pulse rate high enough to eliminate frequency ambiguity.

Velocity becomes ambiguous at a 3 kHz PRF for an L band radar at 300 m/s (near the speed of sound). Velocity becomes ambiguous at a 3 kHz PRF for an X band radar at 30 m/s (speed of a fast car).

Increasing PRF increases the Nyqist frequency. Velocity becomes ambiguous at a 30 kHz PRF for an L band radar at 3000 m/s (about half the velocity of low earth orbit). Velocity becomes ambiguous at a 30 kHz PRF for an X band radar at 300 m/s (speed of sound).

Alternating PRF requires a minimum level of complexity that depends upon maximum radial velocity of the object. The maximum radial velocity is determined by the environment. Objects moving inside earth's atmosphere are limited to about 1 km/s due to thermal heating caused by shock wave physics. Space begins above about 20miles, and objects in low earth orbit travel about 8 km/s. There is no velocity limit in space.

Modeling and simulation is required used to evaluate performance for the chosen PRF combination to ensure no detection gaps for all possible range and velocity combination.

Range Performance

The theoretical range performance is as follows.

$$R = \left(\frac{P_t G_t A_r \sigma F^4 D}{16\pi^2 K_b T B N} \right)^{-4}$$

where

- R = distance to the target
- P_t = transmitter power
- G_t = gain of the transmitting antenna
- A_r = effective aperture (area) of the receiving antenna
- σ = radar cross section, or scattering coefficient, of the target

- F = pattern propagation factor
- D = number of discrete Doppler filters for each range sample
- R = distance to the target
- K_b = Boltzmann's constant
- T = Temperature (kelvin)
- B = Receiver Bandwidth (band pass filter)
- N = Noise Factor

This is derived by combining the Radar equation with the Noise equation and accounting for in-band noise distribution across multiple detection filters. The value *D* is added to the standard radar range equation to account for noise power that arrives at the receiver being distributed equally among each of the Doppler filters.

Angular movement of the antenna must be slow enough so each volume of space remains within the main antenna lobe no less than the duration of three full PRF samples to achieve the level of performance described by this equation. Increasing the value of *D* will increase detection range at the expense of requiring more time to complete a full scan. It is impractical to increase *D* to a value that that requires phase noise lower than the transmitter specification.

Detection range is increased by the number of filters. For example, use of 100 filters will reduce noise contribution in each filter 20dB below the level of noise (electronics) begin sampled in the receiver. Each filter holds a small amount of the total noise arriving at the receiver but all of the signal arriving from the target goes into no more than two filters. This reduces the bandwidth associated with the detection process. A 20dB improvement corresponds with the ability to detect a target 100 times smaller than with conventional pulse-amplitude radar, all else being equal.

Active RF phase shifters commonly used with electronically steered phased array antenna normally have a phase settling time that exceeds the coherency limit needed for pulse-Doppler. This limits pulse-Doppler antennas to a mechanically steered configuration.

Application considerations

Type of Radar

The maximum velocity that can be unambiguously measured is inherently limited by the PRF, as discussed above. The PRF-value must therefore be chosen carefully, based on a tradeoff between maximum velocity resolution and the reduction of velocity aliasing and range ambiguity problems. This tradeoff is highly application dependent, as e.g. weather radars measure velocities at a totally different scale as compared to radars designed to detect supersonic missiles and aircraft.

Moving targets

Stationary targets such as earth ground clutter (land, buildings, etc.) will be dominant in the low doppler frequencies, while moving targets will produce much higher doppler shifts. The radar processor can be designed to mask out clutter by the use of doppler filters (digital or analogue) around the main spectral line (called the clutter-notch), which will result in the display of moving targets only (in relation to the radar). If the radar itself is moving, such as on a fighter aircraft, or a surveillance aircraft, then much more processing will be required, as the clutter in the filters will be based on platform speed, terrain under the radar, antenna depression angle, and antenna rotation/steered angle.

Monopulse radar

Doppler signal processing improves performance of monopulse radar systems by reducing or eliminating signals from slow object, like the earth surface, trees, buildings, and weather.

Chapter-7

Radar Tracker

A **radar tracker** is a component of a radar system, or an associated command and control (C2) system, that associates consecutive radar observations of the same target into tracks. It is particularly useful when the radar system is reporting data from several different targets or when it is necessary to combine the data from several different radars or other sensors.

Role of the radar tracker

A classical rotating air surveillance radar system detects target echoes against a background of noise. It reports these detections (known as "plots") in polar coordinates representing the range and bearing of the target. In addition, noise in the radar receiver will occasionally exceed the detection threshold of the radar's Constant false alarm rate detector and be incorrectly reported as targets (known as false alarms). The role of the radar tracker is to monitor consecutive updates from the radar system (which typically occur once every few seconds, as the antenna rotates) and to determine those sequences of plots belonging to the same target, whilst rejecting any plots believed to be false alarms. In addition, the radar tracker is able to use the sequence of plots to estimate the current speed and heading of the target. When several targets are present, the radar tracker aims to provide one track for each target, with the track history often being used to indicate where the target has come from.

When multiple radar systems are connected to a single reporting post, a **multiradar tracker** is often used to monitor the updates from all of the radars and form tracks from the combination of detections. In this configuration, the tracks are often more accurate than those formed from single radars, as a greater number of detections can be used to estimate the tracks. In addition to associating plots, rejecting false alarms and estimating heading and speed, the radar tracker also acts as a filter, in which errors in the individual radar measurements are smoothed out. In essence, the radar tracker fits a smooth curve to the reported plots and, if done correctly, can increase the overall accuracy of the radar system. A **multisensor tracker** extends the concept of the multiradar tracker to allow the combination of reports from different types of sensor - typically radars, secondary surveillance radars, identification friend or foe (IFF) systems and electronic support measures (ESM) data. A radar track will typically contain the following information

- Position (in two or three dimensions)
- Heading
- Speed
- Unique track number

In addition, and depending on the application or tracker sophistication, the track will also include:

- Civilian SSR Modes A, C, S information
- Military IFF Modes 1, 2, 3, 4 and 5 information
- Call sign information
- ADS-B information
- Track reliability or uncertainty information

General approach

There are many different mathematical algorithms used for implementing a radar tracker, of varying levels of sophistication. However, they all perform steps similar to the following every time the radar updates:

- Associate a radar plot with an existing track (plot to track association)
- Update the track with this latest plot (track smoothing)
- Spawn new tracks with any plots that are not associated with existing tracks (track initiation)
- Delete any tracks that have not been updated, or predict their new location based on the previous heading and speed (track maintenance)

Perhaps the most important step is the updating of tracks with new plots. All trackers will implicitly or explicitly take account of a number of factors during this stage, including:

- a model for how the radar measurements are related to the target coordinates
- the errors on the radar measurements

- a model of the target movement
- errors in the model of the target movement

Using these information, the radar tracker attempts to update the track by forming a weighted average of the current reported position from the radar (which has unknown errors) and the last predicted position of the target from the tracker (which also has unknown errors). The tracking problem is made particularly difficult for targets with unpredictable movements (i.e. unknown target movement models), non-Gaussian measurement or model errors, non-linear relationships between the measured quantities and the desired target coordinates, detection in the presence of non-uniformly distributed clutter, missed detections or false alarms. In the real world, a radar tracker typically faces a combination of all of these effects; this has led to the development of an increasingly sophisticated set of algorithms to resolve the problem. Due to the need to form radar tracks in real time, usually for several hundred targets at once, the deployment of radar tracking algorithms has typically been limited by the available computational power.

Plot to track association

In this step of the processing, the radar tracker seeks to determine which plots should be used to update which tracks. In many approaches, a given plot can only be used to update one track. However, in other approaches a plot can be used to update several tracks, recognising the uncertainty in knowing to which track the plot belongs. Either way, the first step in the process is to update all of the existing tracks to the current time by predicting their new position based on the most recent state estimate (e.g. position, heading, speed, acceleration, etc.) and the assumed target motion model (e.g. constant velocity, constant acceleration, etc.). Having updated the estimates, it is possible to try to associate the plots to tracks.

This can be done in a number of ways:

- By defining an "acceptance gate" around the current track location and then selecting:
    - the closest plot in the gate to the predicted position, or
    - the strongest plot in the gate
- By a statistical approach, such as the Probabilistic Data Association Filter (PDAF) or the Joint Probabilistic Data Association Filter (JPDAF) that choose the most probable location of plot through a statistical combination of all the likely plots. This approach has been shown to be good in situations of high radar clutter.

Once a track has been associated with a plot, it moves to the **track smoothing** stage, where the track prediction and associated plot are combined to provide a new, smoothed estimate of the target location.

Having completed this process, a number of plots will remain unassociated to existing tracks and a number of tracks will remain without updates. This leads to the steps of **track initiation** and **track maintenance**.

Track initiation

Track initiation is the process of creating a new radar track from an unassociated radar plot. Obviously, when the tracker is first switched on, all of the initial radar plots are used to create new tracks, but once the tracker is running, only those plots that couldn't be used to update an existing track are used to spawn new tracks. Typically a new track is given the status of **tentative** until plots from subsequent radar updates have been successfully associated with the new track. Tentative tracks are not shown to the operator and so they provide a means of preventing false tracks from appearing on the screen - at the expense of some delay in the first reporting of a track. Once several updates have been received, the track is **confirmed** and displayed to the operator. The most common criterion for promoting a tentative track to a confirmed track is the "M-of-N rule", which states that during the last N radar updates, at least M plots must have been associated with the tentative track - with M=3 and N=5 being typical values. More sophisticated approaches may use a statistical approach in which a track becomes confirmed when, for instance, its covariance matrix falls to a given size.

Track maintenance

Track maintenance is the process in which a decision is made about whether to end the life of a track. If a track was not associated with a plot during the plot to track association phase, then there is a chance that the target may no longer exist (for instance, an aircraft may have landed or flown out of radar cover). Alternatively, however, there is a chance that the radar may have just failed to see the target at that update, but will find it again on the next update. Common approaches to deciding on whether to terminate a track include:

- If the target was not seen for the past M consecutive update opportunities (typically M=3 or so)
- If the target was not seen for the past M out of N most recent update opportunities
- If the target's track uncertainty (covariance matrix) has grown beyond a certain threshold

Track smoothing

In this important step, the latest track prediction is combined with the associated plot to provide a new, improved estimate of the target state as well as a revised estimate of the errors in this prediction. There are a wide variety of algorithms, of differing complexity and computational load, that can be used for this process.

Alpha-beta tracker

An early tracking approach that assumed fixed Gaussian errors and a constant-speed, non-maneuvering target model to update tracks.

Kalman filter

The role of the Kalman Filter is to take the current known state (i.e. position, heading, speed and possibly acceleration) of the target and predict the new state of the target at the time of the most recent radar measurement. In making this prediction, it also updates its estimate of its own uncertainty (i.e. errors) in this prediction. It then forms a weighted average of this prediction of state and the latest measurement of state, taking account of the known measurement errors of the radar and its own uncertainty in the target motion models. Finally, it updates its estimate of its uncertainty of the state estimate. A key assumption in the mathematics of the Kalman filter is that measurement equations (i.e. the relationship between the radar measurements and the target state) and the state equations (i.e. the equations for predicting a future state based on the current state) are linear - i.e. can be expressed in the form $y = A.x$ (where A is a constant), rather than $y = f(x)$.

The Kalman filter assumes that the measurement errors of the radar, and the errors in its target motion model, and the errors in its state estimate are all zero-mean Gaussian distributed. This means that all of these sources of errors can be represented by a covariance matrix. The mathematics of the Kalman filter is therefore concerned with propagating these covariance matrices and using them to form the weighted sum of prediction and measurement.

In situations where the target motion conforms well to the underlying model, there is a tendency of the Kalman filter to become "over confident" of its own predictions and to start to ignore the radar measurements. If the target then manoeuvres, the filter will fail to follow the manoeuvre. It is therefore common practice when implementing the filter to arbitrarily increase the magnitude of the state estimate covariance matrix slightly at each update to prevent this.

Multiple hypothesis tracker (MHT)

The MHT allows a track to be updated by more than one plot at each update, spawning multiple possible tracks. As each radar update is received every possible track can be potentially updated with every new update. Over time, the track branches into many possible directions. The MHT calculates the probability of each potential track and typically only reports the most probable of all the tracks. For reasons of finite computer memory and computational power, the MHT typically includes some approach for deleting the most unlikely potential track updates. The MHT is designed for situations in which the target motion model is very unpredictable, as all potential track updates are considered. For this reason, it is popular for problems of ground target tracking in Airborne Ground Surveillance (AGS) systems.

Interacting multiple model (IMM)

The IMM is an estimator which can either be used by MHT or JPDAF. IMM uses two or more Kalman filters which run in parallel, each using a different model for target motion or errors. The IMM forms an optimal weighted sum of the output of all the filters and is able to rapidly adjust to target maneuvers. While MHT or JPDAF handles the association and track maintenance, an IMM helps MHT or JPDAF in obtaining a filtered estimate of the target position.

Nonlinear tracking algorithms

Non-linear tracking algorithms use a Non-linear filter to cope with the situation where the measurements have a non-linear relationship to the final track coordinates, where the errors are non-Gaussian, or where the motion update model is non-linear. The most common non-linear filters are:

- the Extended Kalman filter
- the Unscented Kalman filter
- the Particle filter

Extended Kalman filter (EKF)

The EKF is an extension of the Kalman filter to cope with cases where the relationship between the radar measurements and the track coordinates, or the track coordinates and the motion model, is non-linear. In this case, the relationship between the measurements and the state is of the form h = f(x) (where h is the vector of measurements, x is the target state and f(.) is the function relating the two). Similarly, the relationship between the future state and the current state is of the form x(t+1) = g(x(t)) (where x(t) is the state at time t and g(.) is the function that predicts the future state). To handle these non-linearities, the EKF linearises the two non-linear equations using the first term of the Taylor series and then treats the problem as the standard linear Kalman filter problem. Although conceptually simple, the filter can easily diverge (i.e. gradually perform more and more badly) if the state estimate about which the equations are linearised is poor.

The unscented Kalman filter and particle filters are attempts to overcome the problem of linearising the equations.

Unscented Kalman filter (UKF)

The UKF attempts to improve on the EKF by removing the need to linearise the measurement and state equations. Although it retains the assumption that the filter errors are Gaussian distributed, rather than model these as covariance matrices, it instead chooses an explicit sample of those Gaussian errors by choosing a small number (typically 5 or so) of different state estimates that have the required mean and variance.

These points are then propagated directly through the non-linear equations, and the resulting five updated samples are then used to calculate a new mean and variance. This approach then suffers none of the problems of divergence due to poor linearisation and yet retains the overall computational simplicity of the EKF.

Particle filter

The Particle Filter could be considered as a generalisation of the UKF. It makes no assumptions about the distributions of the errors in the filter and neither does it require the equations to be linear. Instead it generates a large number of random potential states ("particles") and then propagates this "cloud of particles" through the equations, resulting in a different distribution of particles at the output. The resulting distribution of particles can then be used to calculate a mean or variance, or whatever other statistical measure is required. The resulting statistics are used to generate the random sample of particles for the next iteration. The particle filter is notable in its ability to handle multi-modal distributions (i.e. distributions where the PDF has more than one peak). However, it is computationally very intensive and is currently unsuitable for most real-world, real-time applications.

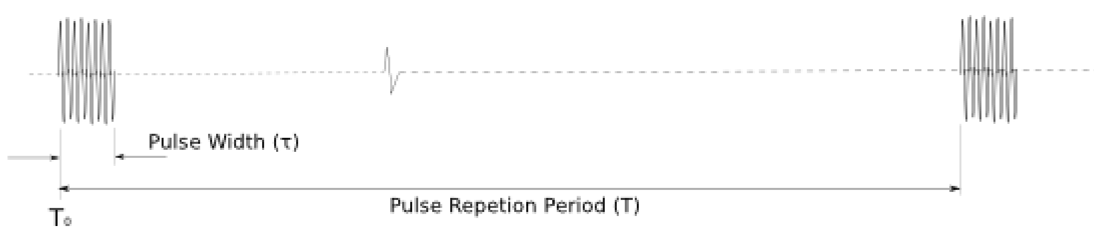
Chapter-8

Radar Signal Characteristics

A radar system uses a radio frequency electromagnetic signal reflected from a target to determine information about that target. In any radar system, the signal transmitted and received will exhibit many of the characteristics described below.

The radar signal in the time domain

The diagram below shows the characteristics of the transmitted signal in the time domain. Note that in this and in all the diagrams within here, the x axis is exaggerated to make the explanation clearer.



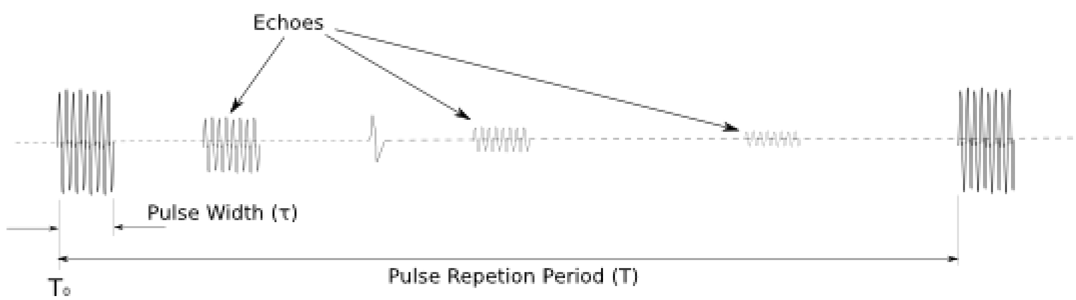
Carrier

The carrier is an RF signal, typically of microwave frequencies, which is usually (but not always) modulated to allow the system to capture the required data. In simple ranging radars, the carrier will be pulse modulated and in continuous wave systems, such as Doppler radar, modulation may not be required. Most systems use pulse modulation, with or without other supplementary modulating signals. Note that with pulse modulation, the carrier is simply switched on and off in sync with the pulses; the modulating waveform does not actually exist in the transmitted signal and the envelope of the pulse waveform is extracted from the demodulated carrier in the receiver. Although obvious when described, this point is often missed when pulse transmissions are first studied, leading to misunderstandings about the nature of the signal.

Pulse width

The pulse width (τ) (or pulse duration) of the transmitted signal is to ensure that the radar emits sufficient energy to allow that the reflected pulse is detectable by its receiver. The amount of energy that can be delivered to a distant target is the product of two things; the output power of the transmitter, and the duration of the transmission. Therefore pulse width constrains the maximum detection range of a target.

It also determines the range discrimination, that is the capacity of the radar to distinguish between two targets fairly close together. At *any* range, with similar azimuth and elevation angles and as viewed by a radar with an unmodulated pulse, the range discrimination is approximately equal in distance to half of the pulse duration.



Pulse width also determines the dead zone at close ranges. While the radar transmitter is active, the receiver input is blanked to avoid the amplifiers being swamped (saturated) or, (more likely), damaged. A simple calculation reveals that a radar echo will take approximately 10.8 μs to return from a target 1 standard mile away (counting from the leading edge of the transmitter pulse (T_0), (sometimes known as transmitter main bang)). For convenience, these figures may also be expressed as 1 nautical mile in 12.4 μs or 1 kilometre in 6.7 μs. (For simplicity, all further discussion will use metric figures.) If the radar pulse width is 1 μs, then there can be no detection of targets closer than about 150 m, because the receiver is blanked.

All this means that the designer cannot simply increase the pulse width to get greater range without having an impact on other performance factors. As with everything else in a radar system, compromises have to be made to a radar system's design to provide the optimal performance for it's role.

Pulse repetition frequency (PRF)

In order to build up a discernible echo, most radar systems emit pulses continuously and the repetition rate of these pulses is determined by the role of the system. An echo from a target will therefore be 'painted' on the display or integrated within the signal processor every time a new pulse is transmitted, reinforcing the return and making detection easier. The higher the PRF that is used, then the more the target is painted. However with the higher PRF the range that the radar can "see" is reduced. Radar designers try to use the highest PRF possible commensurate with the other factors that constrain it, as described below.

There are two other facets related to PRF that the designer must weigh very carefully; the beamwidth characteristics of the antenna, and the required periodicity with which the radar must sweep the field of view. A radar with a 1° horizontal beamwidth that sweeps the entire 360° horizon every 2 seconds with a PRF of 1080 Hz will radiate 3 pulses over each 1-degree arc. If the receiver needs at least 6 reflected pulses of similar amplitudes to achieve an acceptable probability of detection, then there are three choices for the designer: double the PRF, halve the sweep speed, or double the beamwidth. In reality, all three choices are used, to varying extents; radar design is all about compromises between conflicting pressures.

Staggered PRF

Staggered PRF is a transmission process where the time between interrogations from radar changes slightly, *in a patterned and readily-discernible repeating manner.* The change of repetition frequency allows the radar, on a pulse-to-pulse basis, to differentiate between returns from its own transmissions and returns from other radar systems with the same PRF and a similar radio frequency. Consider a radar with a constant interval between pulses; target reflections appear at a relatively constant range related to the flight-time of the pulse. In today's very crowded radio spectrum, there may be many other pulses detected by the receiver, either directly from the transmitter or as reflections from elsewhere. Because their apparent "distance" is defined by measuring their time relative to the last pulse transmitted by "our" radar, these "jamming" pulses could appear at any apparent distance. When the PRF of the "jamming" radar is very similar to "our" radar, those apparent distances may be very slow-changing, just like real targets. By using stagger, a radar designer can force the "jamming" to jump around erratically in apparent range, inhibiting integration and reducing or even suppressing its impact on true target detection.

Without staggered PRF, any pulses originating from another radar on the same radio frequency might appear stable in time and could be mistaken for reflections from the radar's own transmission. With staggered PRF the radar’s own targets appear stable in range in relation to the transmit pulse, whilst the 'jamming' echoes may move around in apparent range (uncorrelated), causing them to be rejected by the receiver. Staggered PRF is only one of several similar techniques used for this, including jittered PRF (where the pulse timing is varied in a less-predictable manner), pulse-frequency modulation, and several other similar techniques whose principal purpose is to reduce the probability of unintentional synchronicity. These techniques are in widespread use in marine safety and navigation radars, by far the most numerous radars on planet Earth today.

Clutter

Clutter (also termed *ground clutter*) is a form of radar signal contamination. It occurs when fixed objects close to the transmitter—such as buildings, trees, or terrain (hills, ocean swells and waves)—obstruct a radar beam and produce echoes. The echoes resulting from ground clutter may be large in both areal size and intensity. The effects of ground clutter fall off as range increases usually due to the curvature of the earth and the tilt of the antenna above the horizon. Without special processing techniques, targets can be lost in returns from terrain on land or waves at sea.

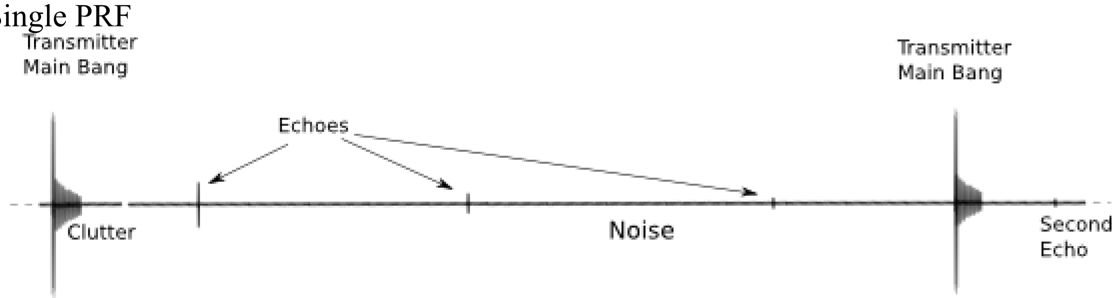
Clutter is used by the military to jam radars by the use chaff. Chaff is small reflective material used to hide troop, ship, or aircraft movements by creating many returns and overwhelming the radar's receiver with spurious targets.

Sensitivity time control (STC)

STC is used to avoid saturation of the receiver from close in ground clutter by adjusting the attenuation of the receiver as a function of distance. More attenuation is applied to returns close in and is reduced as the range increases.

Unambiguous range

Single PRF



In simple systems, echoes from targets must be detected and processed before the next transmitter pulse is generated if range ambiguity is to be avoided. Range ambiguity occurs when the time taken for an echo to return from a target is greater than the pulse repetition period (T); if the interval between transmitted pulses is 1000 microseconds, and the return-time of a pulse from a distant target is 1200 microseconds, the apparent distance of the target is only 200 microseconds. In sum, these 'second echoes' appear on the display to be targets closer than they really are.

Consider the following example : if the radar antenna is located at around 15 m above sea level, then the distance to the horizon is pretty close, (perhaps 15 km). Ground targets further than this range cannot be detected, so the PRF can be quite high; a radar with a PRF of 7.5 kHz will return ambiguous echoes from targets at about 20 km, or over the horizon. If however, the PRF was doubled to 15 kHz, then the ambiguous range is reduced to 10 km and targets beyond this range would only appear on the display after the transmitter has emitted another pulse. A target at 12 km would appear to be 2 km away, although the strength of the echo might be much lower than that from a genuine target at 2 km.

The maximum non ambiguous range varies inversely with PRF and is given by:

$$Range_{maxunambiguous} = \left(\frac{c}{2\,PRF}\right)$$

If a longer unambiguous range is required with this simple system, then lower PRFs are required and it was quite common for early search radars to have PRFs as low as a few hundred Hz, giving an unambiguous range out to well in excess of 150 km. However, lower PRFs introduce other problems, including poorer target painting and velocity ambiguity in Pulse-Doppler systems.

Multiple PRF

Modern radars, especially air-to-air combat radars in military aircraft, may use PRFs in the tens-to-hundreds of kilohertz and stagger the interval between pulses to allow the correct range to be determined. With this form of staggered PRF, a *packet* of pulses is transmitted with a fixed interval between each pulse, and then another *packet* is transmitted with a slightly different interval. Target reflections appear at different ranges for each *packet*; these differences are accumulated and then simple arithmetical techniques may be applied to determine true range. Such radars may use repetitive patterns of *packets*, or more adaptable *packets* that respond to apparent target behaviors. Regardless, radars that employ the technique are universally coherent, with a very stable radio frequency, and the pulse *packets* may also be used to make measurements of the Doppler shift (a velocity-dependent modification of the apparent radio frequency), especially when the PRFs are in the hundreds-of-kilohertz range. Radars exploiting Doppler effects in this manner typically determine relative velocity first, from the Doppler effect, and then use other techniques to derive target distance.

Maximum Unambiguous Range

At its most simplistic, MUR (Maximum Unambiguous Range) for a Pulse Stagger sequence may be calculated using the TSP (Total Sequence Period). TSP is defined as the total time it takes for the Pulsed pattern to repeat. This can be found by the addition of all the elements in the stagger sequence. The formula is derived from the speed of light and the length of the sequence:

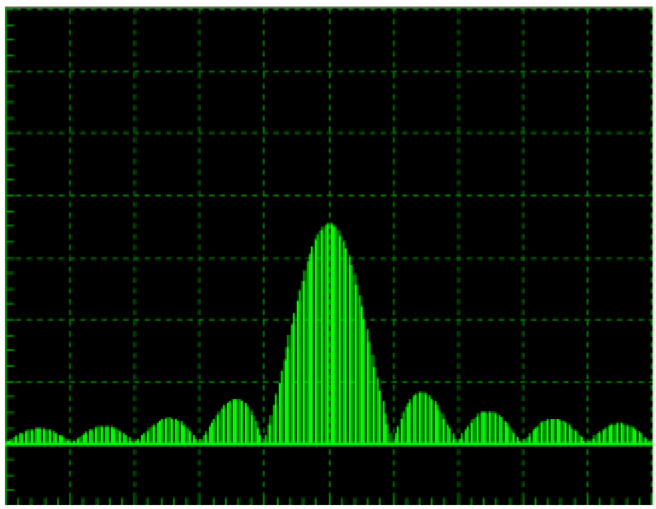
$$MUR = (c * 0.5 * TSP)$$

where c is the speed of light, usually in metres per microsecond, and TSP is the addition of all the positions of the stagger sequence, usually in microseconds. However, it should be noted that in a stagger sequence, some intervals may be repeated several times; when this occurs, it is more appropriate to consider TSP as the addition of all the **unique** intervals in the sequence.

Also, it is worth remembering that there may be vast differences between the MUR and the maximum range (the range beyond which reflections will probably be too weak to be detected), and that the maximum *instrumented* range may be *much* shorter than either of these. A civil marine radar, for instance, may have user-selectable maximum *instrumented* display ranges of 72, or 96 or rarely 120 nautical miles, in accordance with international law, but maximum unambiguous ranges of over 40,000 nautical miles and maximum detection ranges of perhaps 150 nautical miles. When such huge disparities are noted, it reveals that the primary purpose of staggered PRF is to reduce "jamming", rather than to increase unambiguous range capabilities.

The radar signal in the frequency domain

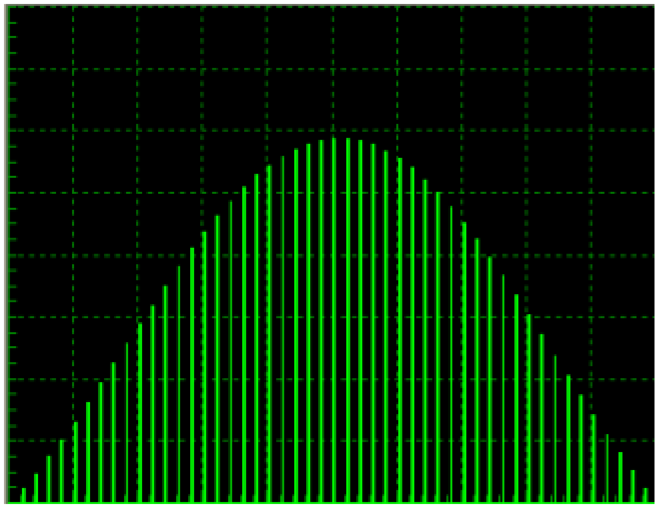
Pure CW radars appear as a single line on a Spectrum analyser display and when modulated with other sinusoidal signals, the spectrum differs little from that obtained with standard analogue modulation schemes used in communications systems, such as Frequency Modulation and consist of the carrier plus a relatively small number of sidebands. When the radar signal is modulated with a pulse train as shown above, the spectrum becomes much more complicated and far more difficult to visualise.



Basic Fourier analysis shows that any repetitive complex signal consists of a number of harmonically related sine waves. The radar pulse train is a form of square wave, the pure form of which consists of the fundamental plus all of the odd harmonics. The exact composition of the pulse train will depend on the pulse width and PRF, but mathematical analysis can be used to calculate all of the frequencies in the spectrum. When the pulse train is used to modulate a radar carrier, the typical spectrum shown on the left will be obtained.

Examination of this spectral response shows that it contains two basic structures. The Coarse Structure; (the peaks or 'lobes' in the diagram on the left) and the Fine Structure which contains the individual frequency components as shown below. The Envelope of the lobes in the Coarse Structure is given by: $\frac{1}{\pi f \tau}$.

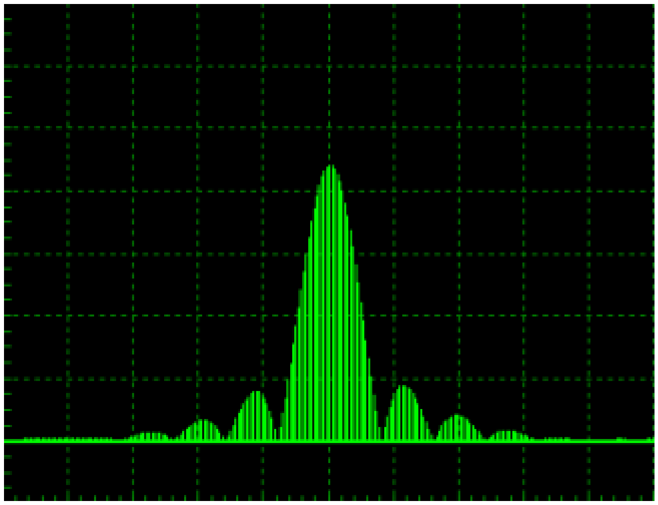
Note that the pulse width (τ) appears on the bottom of this equation and determines the lobe spacing. Smaller pulse widths result in wider lobes and therefore greater bandwidth.



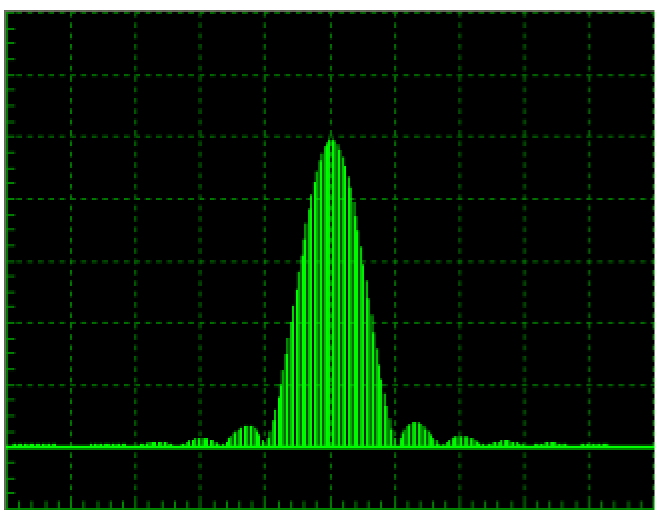
Examination of the spectral response in finer detail, as shown on the right, shows that the Fine Structure contains individual lines or spot frequencies. The formula for the fine structure is given by $\frac{T}{\pi f \tau}$ and since the period of the PRF (T) appears at the top of the fine spectrum equation, there will be fewer lines if higher PRFs are used. These facts affect the decisions made by radar designers when considering the trade-offs that need to be made when trying to overcome the ambiguities that affect radar signals.

Pulse profiling

If the rise and fall times of the modulation pulses are zero, (e.g. the pulse edges are infinitely sharp), then the sidebands will be as shown in the spectral diagrams above. The bandwidth consumed by this transmission can be huge and the total power transmitted is distributed over many hundreds of spectral lines. This is a potential source of interference with any other device and frequency-dependent imperfections in the transmit chain mean that some of this power never arrives at the antenna. In reality of course, it is impossible to achieve such sharp edges, so in practical systems the sidebands contain far fewer lines than a perfect system. If the bandwidth can be limited to include relatively few sidebands, by rolling off the pulse edges intentionally, an efficient system can be realised with the minimum of potential for interference with nearby equipment. However, the trade-off of this is that slow edges make range resolution poor. Early radars limited the bandwidth through filtration in the transmit chain, e.g. the waveguide, scanner etc., but performance could be sporadic with unwanted signals breaking through at remote frequencies and the edges of the recovered pulse being indeterminate. Further examination of the basic Radar Spectrum shown above shows that the information in the various lobes of the Coarse Spectrum is identical to that contained in the main lobe, so limiting the transmit and receive bandwidth to that extent provides significant benefits in terms of efficiency and noise reduction.



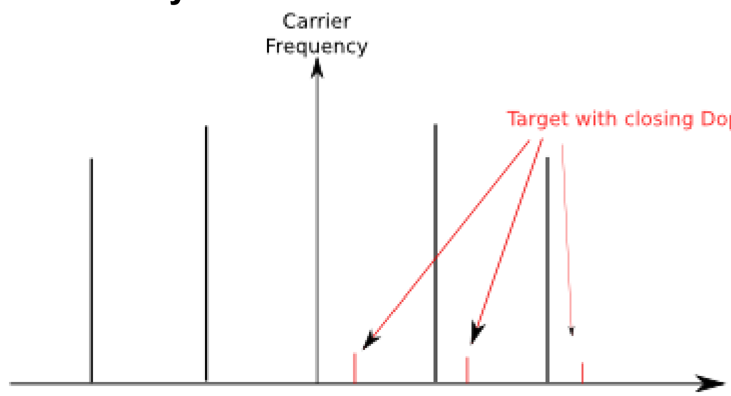
Recent advances in signal processing techniques have made the use of pulse profiling or shaping more common. By shaping the pulse envelope before it is applied to the transmitting device, say to a cosine law or a trapezoid, the bandwidth can be limited at source, with less reliance on filtering. When this technique is combined with pulse compression, then a good compromise between efficiency, performance and range resolution can be realised. The diagram on the left shows the effect on the spectrum if a trapezoid pulse profile is adopted. It can be seen that the energy in the sidebands is significantly reduced compared to the main lobe and the amplitude of the main lobe is increased.



Similarly, the use of a cosine pulse profile has an even more marked effect, with the amplitude of the sidelobes practically becoming negligible. The main lobe is again increased in amplitude and the sidelobes correspondingly reduced, giving a significant improvement in performance.

There are many other profiles that can be adopted to optimise the performance of the system, but cosine and trapezoid profiles generally provide a good compromise between efficiency and resolution and so tend to be used most frequently.

Unambiguous velocity



This is an issue only with a particular type of system; the Pulse-Doppler radar, which uses the Doppler effect to resolve velocity from the apparent change in frequency caused by targets that have net radial velocities compared to the radar device. Examination of the spectrum generated by a pulsed transmitter, shown above, reveals that each of the sidebands, (both coarse and fine), will be subject to the Doppler effect, another good reason to limit bandwidth and spectral complexity by pulse profiling.

Consider the positive shift caused by the closing target in the diagram which has been highly simplified for clarity. It can be seen that as the relative velocity increases, a point will be reached where the spectral lines that constitute the echoes are hidden or aliased by the next sideband of the modulated carrier. Transmission of multiple pulse-packets with different PRF-values, e.g. staggered PRFs, will resolve this ambiguity, since each new PRF value will result in a new sideband position, revealing the velocity to the receiver. The maximum unambiguous target velocity is given by:

$$\pm\frac{c\,PRF}{4\,f}$$

Typical system parameters

Taking all of the above characteristics into account means that certain constraints are placed on the radar designer. For example, a system with a 3 GHz carrier frequency and a pulse width of 1us will have a carrier period of approximately 333ps. Each transmitted pulse will contain about 3000 carrier cycles and the velocity and range ambiguity values for such a system would be:

PRF	Velocity Ambiguity	Range Ambiguity
Low (2 kHz)	50 m/s	75 km
Medium (12 kHz)	300 m/s	12.5 km
High (200 kHz)	5000 m/s	750 m

Chapter-9

Johnson–Nyquist Noise

Johnson–Nyquist noise (**thermal noise**, **Johnson noise**, or **Nyquist noise**) is the electronic noise generated by the thermal agitation of the charge carriers (usually the electrons) inside an electrical conductor at equilibrium, which happens regardless of any applied voltage.

Thermal noise is approximately white, meaning that the power spectral density is nearly equal throughout the frequency spectrum. Additionally, the amplitude of the signal has very nearly a Gaussian probability density function.

History

This type of noise was first measured by John B. Johnson at Bell Labs in 1928. He described his findings to Harry Nyquist, also at Bell Labs, who was able to explain the results.

Noise voltage and power

Thermal noise is distinct from shot noise, which consists of additional current fluctuations that occur when a voltage is applied and a macroscopic current starts to flow. For the general case, the above definition applies to charge carriers in any type of conducting medium (e.g. ions in an electrolyte), not just resistors. It can be modeled by a voltage source representing the noise of the non-ideal resistor in series with an ideal noise free resistor.

The one-sided power spectral density, or voltage variance (mean square) per hertz of bandwidth, is given by

$$\bar{v_n^2} = 4k_BTR$$

where k_B is Boltzmann's constant in joules per kelvin, T is the resistor's absolute temperature in kelvins, and R is the resistor value in ohms (Ω). Use this equation for quick calculation:

$$\sqrt{\bar{v_n^2}} = 0.13\sqrt{R}\ \text{nV}/\sqrt{\text{Hz}}.$$

For example, a 1 kΩ resistor at a temperature of 300 K has

$$\sqrt{\bar{v_n^2}} = \sqrt{4 \cdot 1.38 \cdot 10^{-23}\ \text{J/K} \cdot 300\ \text{K} \cdot 1\ \text{k}\Omega} = 4.07\ \text{nV}/\sqrt{\text{Hz}}.$$

For a given bandwidth, the root mean square (RMS) of the voltage, v_n, is given by

$$v_n = \sqrt{\bar{v_n^2}}\sqrt{\Delta f} = \sqrt{4k_B T R \Delta f}$$

where Δf is the bandwidth in hertz over which the noise is measured. For a 1 kΩ resistor at room temperature and a 10 kHz bandwidth, the RMS noise voltage is 400 nV. A useful rule of thumb to remember is that 50 Ω at 1 Hz bandwidth correspond to 1 nV noise at room temperature.

A resistor in a short circuit dissipates a noise power of

$$P = \bar{v_n^2}/R = 4k_B T \Delta f.$$

The noise generated at the resistor can transfer to the remaining circuit; the maximum noise power transfer happens with impedance matching when the Thévenin equivalent resistance of the remaining circuit is equal to the noise generating resistance. In this case each one of the two participating resistors dissipates noise in both itself and in the other resistor. Since only half of the source voltage drops across any one of these resistors, the resulting noise power is given by

$$P = k_B T \Delta f$$

where P is the thermal noise power in watts. Notice that this is independent of the noise generating resistance.

Noise current

The noise source can also be modeled by a current source in parallel with the resistor by taking the Norton equivalent that corresponds simply to divide by R. This gives the root mean square value of the current source as:

$$i_n = \sqrt{\frac{4k_B T \Delta f}{R}}.$$

Thermal noise is intrinsic to all resistors and is not a sign of poor design or manufacture, although resistors may also have excess noise.

Noise power in decibels

Signal power is often measured in dBm (decibels relative to 1 milliwatt). From the equation above, noise power in a resistor at room temperature, in dBm, is then:

$$P_{\mathrm{dBm}} = 10\ \log_{10}(k_B T \Delta f \times 1000)$$

where the factor of 1000 is present because the power is given in milliwatts, rather than watts. This equation can be simplified by separating the constant parts from the bandwidth:

$$P_{\mathrm{dBm}} = 10\ \log_{10}(k_B T \times 1000) + 10\ \log(\Delta f)$$

which is more commonly seen approximated as:

$$P_{\mathrm{dBm}} = -174 + 10\ \log_{10}(\Delta f)$$

Noise power at different bandwidths is then simple to calculate:

Bandwidth (Δf)	Thermal noise power	Notes
1 Hz	−174 dBm	
10 Hz	−164 dBm	
100 Hz	−154 dBm	
1 kHz	−144 dBm	
10 kHz	−134 dBm	FM channel of 2-way radio
100 kHz	−124 dBm	
180 kHz	−121.45 dBm	One LTE resource block
200 kHz	−120.98 dBm	One GSM channel (ARFCN)
1 MHz	−114 dBm	
2 MHz	−111 dBm	Commercial GPS channel
6 MHz	−106 dBm	Analog television channel
20 MHz	−101 dBm	WLAN 802.11 channel

Thermal noise on capacitors

Thermal noise on capacitors is referred to as *kTC* noise. Thermal noise in an RC circuit has an unusually simple expression, as the value of the resistance (*R*) drops out of the equation. This is because higher *R* contributes to more filtering as well as to more noise. The noise bandwidth of the RC circuit is 1/(4*RC*), which can substituted into the above formula to eliminate *R*. The mean-square and RMS noise voltage generated in such a filter are:

$$\bar{v_n^2} = k_B T / C$$

$$v_n = \sqrt{k_B T/C}.$$

Thermal noise accounts for 100% of *kTC* noise, whether it is attributed to the resistance or to the capacitance.

In the extreme case of the *reset noise* left on a capacitor by opening an ideal switch, the resistance is infinite, yet the formula still applies; however, now the RMS must be interpreted not as a time average, but as an average over many such reset events, since the voltage is constant when the bandwidth is zero. In this sense, the Johnson noise of an RC circuit can be seen to be inherent, an effect of the thermodynamic distribution of the number of electrons on the capacitor, even without the involvement of a resistor.

The noise is not caused by the capacitor itself, but by the thermodynamic equilibrium of the amount of charge on the capacitor. Once the capacitor is disconnected from a conducting circuit, the thermodynamic fluctuation is *frozen* at a random value with standard deviation as given above.

The reset noise of capacitive sensors is often a limiting noise source, for example in image sensors. As an alternative to the voltage noise, the reset noise on the capacitor can also be quantified as the electrical charge standard deviation, as

$$Q_n = \sqrt{k_B T C}.$$

Since the charge variance is $k_B TC$, this noise is often called *kTC noise*.

Any system in thermal equilibrium has state variables with a mean energy of $kT/2$ per degree of freedom. Using the formula for energy on a capacitor ($E = ½CV^2$), mean noise energy on a capacitor can be seen to also be $½C(kT/C)$, or also $kT/2$. Thermal noise on a capacitor can be derived from this relationship, without consideration of resistance.

The *kTC* noise is the dominant noise source at small capacitors.

Noise of capacitors at 300 K

Capacitance	$\sqrt{k_B T/C}$	**Electrons**
1 fF	2 mV	12.5 e⁻
10 fF	640 µV	40 e⁻
100 fF	200 µV	125 e⁻
1 pF	64 µV	400 e⁻
10 pF	20 µV	1250 e⁻
100 pF	6.4 µV	4000 e⁻
1 nF	2 µV	12500 e⁻

Noise at very high frequencies

The above equations are good approximations at any practical radio frequency in use (i.e. frequencies below about 80 gigahertz). In the most general case, which includes up to optical frequencies, the power spectral density of the voltage across the resistor *R*, in V^2/Hz is given by:

$$\Phi(f) = \frac{2Rhf}{e^{\frac{hf}{k_B T}} - 1}$$

where *f* is the frequency, *h* Planck's constant, k_B Boltzmann constant and *T* the temperature in kelvins. If the frequency is low enough, that means:

$$f \ll \frac{k_B T}{h}$$

(this assumption is valid until few terahertz at room temperature) then the exponential can be expressed in terms of its Taylor series. The relationship then becomes:

$$\Phi(f) \approx 2Rk_B T.$$

In general, both *R* and *T* depend on frequency. In order to know the total noise it is enough to integrate over all the bandwidth. Since the signal is real, it is possible to integrate over only the positive frequencies, then multiply by 2. Assuming that *R* and *T* are constants over all the bandwidth Δf, then the root mean square (RMS) value of the voltage across a resistor due to thermal noise is given by

$$v_n = \sqrt{4k_B T R \Delta f},$$

that is, the same formula as above.

Chapter-10

Coherence

In physics, **coherence** is a property of waves that enables stationary (i.e. temporally and spatially constant) interference. More generally, coherence describes all properties of the correlation between physical quantities of a wave.

When interfering, two waves can add together to create a larger wave (**constructive interference**) or subtract from each other to create a smaller wave (**destructive interference**), depending on their relative phase. Two waves are said to be coherent if they have a constant relative phase. The degree of coherence is measured by the interference visibility, a measure of how perfectly the waves can cancel due to destructive interference.

Introduction

Coherence was originally introduced in connection with Young's double-slit experiment in optics but is now used in any field that involves waves, such as acoustics, electrical engineering, neuroscience, and quantum mechanics. The property of coherence is the basis for commercial applications such as holography, the Sagnac gyroscope, radio antenna arrays, optical coherence tomography and telescope interferometers (astronomical optical interferometers and radio telescopes).

Coherence and correlation

The coherence of two waves follows from how well correlated the waves are as quantified by the cross-correlation function. The cross-correlation quantifies the ability to predict the value of the second wave by knowing the value of the first. As an example, consider two waves perfectly correlated for all times. At any time, if the first wave changes, the second will change in the same way. If combined they can exhibit complete constructive interference/superposition at all times. It follows that they are perfectly coherent. As will be discussed below, the second wave need not be a separate entity. It could be the first wave at a different time or position. In this case, the measure of correlation is the autocorrelation function (sometimes called **self-coherence**). Degree of correlation involves correlation functions

Examples of wave-like states

These states are unified by the fact that their behavior is described by a wave equation or some generalization thereof.

- Waves in a rope (up and down) or slinky (compression and expansion)
- Surface waves in a liquid
- Electric signals (fields) in transmission cables
- Sound
- Radio waves and Microwaves
- Light waves (optics)
- Electrons, atoms and any other object (such as a baseball, as described by quantum physics)

In most of these systems, one can measure the wave directly. Consequently, its correlation with another wave can simply be calculated. However, in optics one cannot measure the electric field directly as it oscillates much faster than any detector's time resolution. Instead, we measure the intensity of the light. Most of the concepts involving coherence which will be introduced below were developed in the field of optics and then used in other fields. Therefore, many of the standard measurements of coherence are indirect measurements, even in fields where the wave can be measured directly.

Temporal coherence

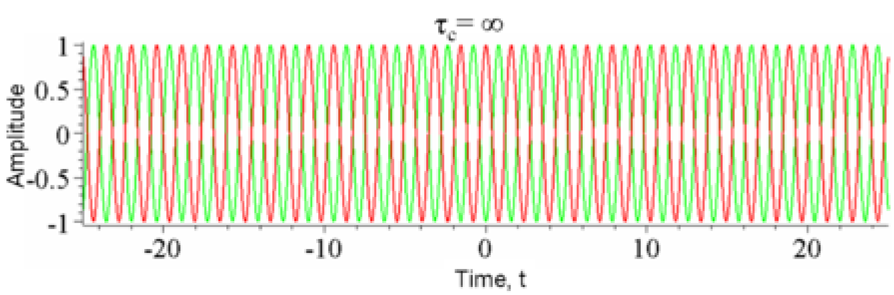


Figure 1: The amplitude of a single frequency wave as a function of time *t* (red) and a copy of the same wave delayed by τ(green). The coherence time of the wave is infinite since it is perfectly correlated with itself for all delays τ.

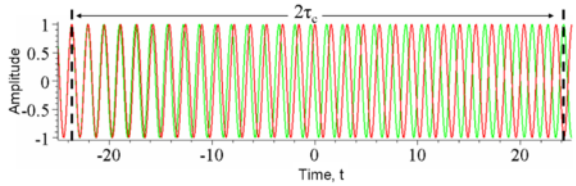


Figure 2: The amplitude of a wave whose phase drifts significantly in time τ_c as a function of time *t* (red) and a copy of the same wave delayed by $2\tau_c$(green). At any particular time t the wave can interfere perfectly with its delayed copy. But, since half the time the red and green waves are in phase and half the time out of phase, when averaged over t any interference disappears at this delay.

Temporal coherence is the measure of the average correlation between the value of a wave at any pair of times, separated by delay τ. Temporal coherence tells us how monochromatic a source is. In other words, it characterizes how well a wave can interfere with itself at a different time. The delay over which the phase or amplitude wanders by a significant amount (and hence the correlation decreases by significant amount) is defined as the coherence time τ_c. At $\tau=0$ the degree of coherence is perfect whereas it drops significantly by delay τ_c. The coherence length L_c is defined as the distance the wave travels in time τ_c.

One should be careful not to confuse the coherence time with the time duration of the signal, nor the coherence length with the coherence area.

The relationship between coherence time and bandwidth

It can be shown that the faster a wave decorrelates (and hence the smaller τ_c is) the larger the range of frequencies Δf the wave contains. Thus there is a tradeoff:

$$\tau_c \Delta f \approx 1.$$

Formally, this follows from the convolution theorem in mathematics, which relates the Fourier transform of the power spectrum (the intensity of each frequency) to its autocorrelation.

Examples of temporal coherence

We consider four examples of temporal coherence.

- A wave containing only a single frequency (monochromatic) is perfectly correlated at all times according to the above relation.
- Conversely, a wave whose phase drifts quickly will have a short coherence time.
- Similarly, pulses (wave packets) of waves, which naturally have a broad range of frequencies, also have a short coherence time since the amplitude of the wave changes quickly.
- Finally, white light, which has a very broad range of frequencies, is a wave which varies quickly in both amplitude and phase. Since it consequently has a very short coherence time (just 10 periods or so), it is often called incoherent.

The most monochromatic sources are usually lasers; such high monochromaticity implies long coherence lengths (up to hundreds of meters). For example, a stabilized helium-neon laser can produce light with coherence lengths in excess of 5 m. Not all lasers are monochromatic, however (e.g. for a mode-locked Ti-sapphire laser, $\Delta\lambda \approx 2$ nm - 70 nm). LEDs are characterized by $\Delta\lambda \approx 50$ nm, and tungsten filament lights exhibit $\Delta\lambda \approx 600$ nm, so these sources have shorter coherence times than the most monochromatic lasers.

Holography requires light with a long coherence time. In contrast, Optical coherence tomography uses light with a short coherence time.

Measurement of temporal coherence

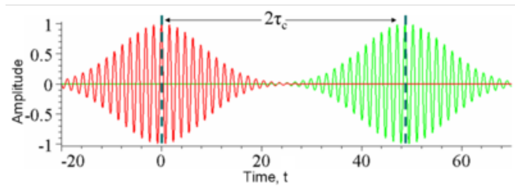


Figure 3: The amplitude of a wavepacket whose amplitude changes significantly in time τ_c (red) and a copy of the same wave delayed by $2\tau_c$(green) plotted as a function of time *t*. At any particular time the red and green waves are uncorrelated; one oscillates while the other is constant and so there will be no interference at this delay. Another way of looking at this is the wavepackets are not overlapped in time and so at any particular time there is only one nonzero field so no interference can occur.

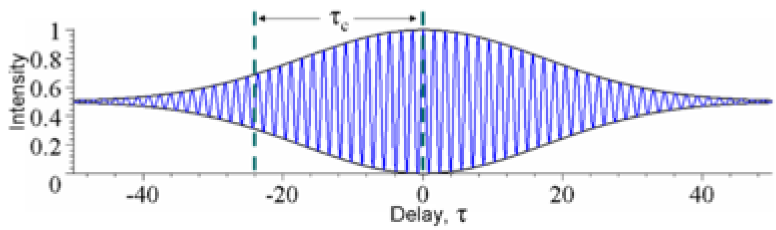


Figure 4: The time-averaged intensity (blue) detected at the output of an interferometer plotted as a function of delay τ for the example waves in Figures 2 and 3. As the delay is changed by half a period, the interference switches between constructive and destructive. The black lines indicate the interference envelope, which gives the degree of coherence. Although the waves in Figures 2 and 3 have different time durations, they have the same coherence time.

In optics, temporal coherence is measured in an interferometer such as the Michelson interferometer or Mach-Zehnder interferometer. In these devices, a wave is combined with a copy of itself that is delayed by time τ. A detector measures the time-averaged intensity of the light exiting the interferometer. The resulting interference visibility gives the temporal coherence at delay τ. Since for most natural light sources, the coherence time is much shorter than the time resolution of any detector, the detector itself does the time averaging. Consider the example shown in Figure 3. At a fixed delay, here $2\tau_c$, an infinitely fast detector would measure an intensity that fluctuates significantly over a time *t* equal to τ_c. In this case, to find the temporal coherence at $2\tau_c$, one would manually time-average the intensity.

Spatial coherence

In some systems, such as water waves or optics, wave-like states can extend over one or two dimensions. Spatial coherence describes the ability for two points in space, x_1 and x_2, in the extent of a wave to interfere, when averaged over time. More precisely, the spatial coherence is the cross-correlation between two points in a wave for all times. If a wave has only 1 value of amplitude over an infinite length, it is perfectly spatially coherent. The range of separation between the two points over which there is significant interference is called the coherence area, A_c. This is the relevant type of coherence for the Young's double-slit interferometer. It is also used in optical imaging systems and particularly in various types of astronomy telescopes. Sometimes people also use "spatial coherence" to refer to the visibility when a wave-like state is combined with a spatially shifted copy of itself.

Examples of spatial coherence

- Spatial coherence

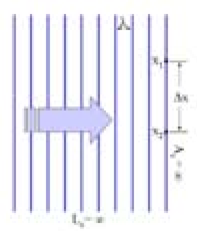


Figure 5: A plane wave with an infinite coherence length.

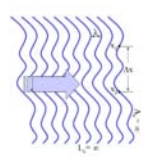


Figure 6: A wave with a varying profile (wavefront) and infinite coherence length.

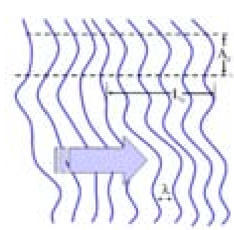


Figure 7: A wave with a varying profile (wavefront) and finite coherence length.

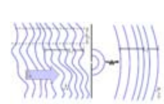


Figure 8: A wave with finite coherence area is incident on a pinhole (small aperture). The wave will diffract out of the pinhole. Far from the pinhole the emerging spherical wavefronts are approximately flat. The coherence area is now infinite while the coherence length is unchanged.

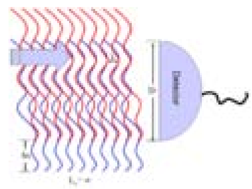


Figure 9: A wave with infinite coherence area is combined with a spatially-shifted copy of itself. Some sections in the wave interfere constructively and some will interfere destructively. Averaging over these sections, a detector with length D will measure reduced interference visibility. For example a misaligned Mach-Zehnder interferometer will do this.

Consider a tungsten light-bulb filament. Different points in the filament emit light independently and have no fixed phase-relationship. In detail, at any point in time the profile of the emitted light is going to be distorted. The profile will change randomly over the coherence time τ_c. Since for a white-light source such as a light-bulb τ_c is small, the filament is considered a spatially incoherent source. In contrast, a radio antenna array, has large spatial coherence because antennas at opposite ends of the array emit with a fixed phase-relationship. Light waves produced by a laser often have high temporal and spatial coherence (though the degree of coherence depends strongly on the exact properties of the laser). Spatial coherence of laser beams also manifests itself as speckle patterns and diffraction fringes seen at the edges of shadow.

Holography requires temporally and spatially coherent light. Its inventor, Dennis Gabor, produced successful holograms more than ten years before lasers were invented. To produce coherent light he passed the monochromatic light from an emission line of a mercury-vapor lamp through a pinhole spatial filter.

In February 2011, Dr Andrew Truscott, leader of a research team at the ARC Centre of Excellence for Quantum-Atom Optics at Australian National University in Canberra, Australian Capital Territory, showed that helium atoms cooled to near absolute zero / Bose-Einstein condensate state, can be made to flow and behave as a coherent beam as occurs in a laser.

Spectral coherence

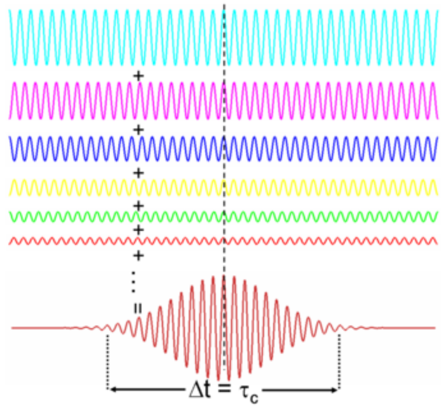


Figure 10: Waves of different frequencies (i.e. colors) interfere to form a pulse if they are coherent.

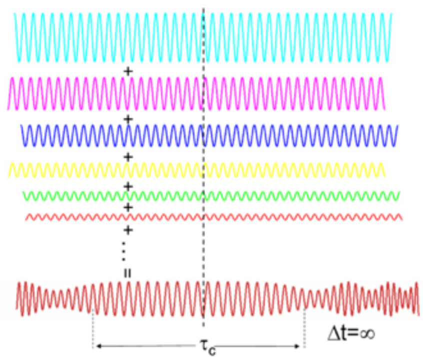


Figure 11: Spectrally incoherent light interferes to form continuous light with a randomly varying phase and amplitude

Waves of different frequencies (in light these are different colours) can interfere to form a pulse if they have a fixed relative phase-relationship. Conversely, if waves of different frequencies are not coherent, then, when combined, they create a wave that is continuous in time (e.g. white light or white noise). The temporal duration of the pulse Δt is limited by the spectral bandwidth of the light Δf according to:

$$\Delta f \Delta t \geq 1,$$

which follows from the properties of the Fourier transform (for quantum particles it also results in the Heisenberg uncertainty principle).

If the phase depends linearly on the frequency (i.e. $\theta(f) \propto f$) then the pulse will have the minimum time duration for its bandwidth (a *transform-limited* pulse), otherwise it is chirped.

Measurement of spectral coherence

Measurement of the spectral coherence of light requires a nonlinear optical interferometer, such as an intensity optical correlator, frequency-resolved optical gating (FROG), or Spectral phase interferometry for direct electric-field reconstruction (SPIDER).

Polarization coherence

Light also has a polarization, which is the direction in which the electric field oscillates. Unpolarized light is composed of two equally intense incoherent light waves with orthogonal polarizations. The electric field of the unpolarized light wanders in every direction and changes in phase over the coherence time of the two light waves. A polarizer rotated to any angle will always transmit half the incident intensity when averaged over time.

If the electric field wanders by a smaller amount the light will be partially polarized so that at some angle, the polarizer will transmit more than half the intensity. If a wave is combined with an orthogonally polarized copy of itself delayed by less than the coherence time, partially polarized light is created.

The polarization of a light beam is represented by a vector in the Poincare sphere. For polarized light the end of the vector lies on the surface of the sphere, whereas the vector has zero length for unpolarized light. The vector for partially polarized light lies within the sphere

Applications

Holography

Coherent superpositions of *optical wave fields* include holography. Holographic objects are used frequently in daily life in bank notes and credit cards.

Non-optical wave fields

Further applications concern the coherent superposition of *non-optical wave fields*. In quantum mechanics for example one considers a probability field, which is related to the wave function $\psi(\mathbf{r})$(interpretation: density of the probability amplitude). Here the applications concern, among others, the future technologies of quantum computing and the already available technology of quantum cryptography. Additionally the problems of the following subchapter are treated.

Quantum coherence

In quantum mechanics, all objects have wave-like properties. For instance, in Young's double-slit experiment electrons can be used in the place of light waves. Each electron can go through either slit and hence has two paths that it can take to a particular final position. In quantum mechanics these two paths interfere. If there is destructive interference, the electron never arrives at that particular position. This ability to interfere is called quantum coherence.

The quantum description of perfectly coherent paths is called a pure state, in which the two paths are combined in a superposition. The correlation between the two particles exceeds what would be predicted for classical correlation alone. If this two-particle system is decohered (which would occur in a measurement via Einselection), then there is no longer any phase relationship between the two states. The quantum description of imperfectly coherent paths is called a mixed state, described by a density matrix (also called the "statistical operator") and is entirely analogous to a classical system of mixed probabilities (the correlations are classical).

Large-scale (macroscopic) quantum coherence leads to novel phenomena. For instance, the laser, superconductivity, and superfluidity are examples of highly coherent quantum systems, whose effects are evident at the macroscopic scale. These examples of quantum coherence are Bose–Einstein condensates. Here, all the particles that make up the condensate are in-phase; they are thus necessarily all described by a single quantum wavefunction. On the other hand, the Schrödinger's cat thought experiment, highlights the fact that quantum coherence is not typically seen at the macroscopic scale but has been observed in the motion of mechanical resonator.

Chapter-11

Heterodyne

In radio and signal processing, **heterodyning** is the generation of new frequencies by combining two frequencies. Heterodyning is useful for **frequency shifting** information of interest into a useful frequency range following modulation or prior to demodulation. The two frequencies are combined in a vacuum tube, transistor, diode, or other non-linear signal processing device. Heterodyning creates two new frequencies, according to the properties of the sine function; one is the sum of the two frequencies mixed, the other is their difference. These new frequencies are called *heterodynes*. Typically only one of the new frequencies is desired—the higher one after modulation and the lower one after demodulation. The other signal is filtered out of the output of the mixer.

History

The word *heterodyne* is derived from the Greek roots *hetero-* "different", and *dyn-* "power" (cf. dynamis). The original heterodyne technique was pioneered by Canadian inventor-engineer Reginald Fessenden in 1901, but was not pursued very far because local oscillators were not very stable at the time. Heterodyning was invented by Fessenden as a technique to make the Morse code radiotelegraph (CW) signals used during the wireless telegraphy era audible. A "heterodyne" receiver had an oscillator circuit, called a local oscillator, that produced a radio signal that was adjusted to be close in frequency to the signal being received, so that when the two signals were mixed the difference or "beat" frequency was in the audible range. This produced a musical tone in the speaker, so the "dots" and "dashes" of Morse code were audible as beeping sounds. This technique is still used in radio telegraphy; the oscillator is now called a beat frequency oscillator.

Later, the superheterodyne receiver (superhet) was invented by Edwin Howard Armstrong in 1918, converting the incoming Radio Frequency (RF) to a fixed Intermediate Frequency (IF), using the heterodyne technique. The difference between the superhet and the original heterodyne is the use of a tunable RF filter on the front end, a mixer circuit, a stable local oscillator (LO) and a fixed frequency high-gain band-pass amplifier. The original heterodyne technique tried to accomplish all of this in one stage thus producing an unstable amplifier.

Applications

Heterodyning is used very widely in communications engineering to generate new frequencies and move information from one frequency channel to another. Besides its use in the superheterodyne circuit which is found in almost all radio and television receivers, it is used in radio transmitters, modems, satellite communications and set-top boxes, radar, radio telescopes, telemetry systems, cell phones, cable television converter boxes and headends, microwave relays, metal detectors, atomic clocks, and military electronic countermeasures (jamming) systems.

Up and down converters

In digital communications signals can be transmitted in baseband or passband. On receiver side usually a downconverter is used to transform the signal from the passband back to the baseband for further processing. The local oscillator frequency is $\sqrt{2}e^{j2\Pi f_c t}$, with f_c being the carrier frequency. This results in a scaling of the received signal by $\sqrt{2}$ and a phase shifting by f_c to the left, so that the resulting signal is located in the baseband.

A **radio frequency upconverter** is a device that takes an input of radio frequency energy of a specific frequency range and outputs it on a higher frequency. Likewise, downconverters take an input frequency and reduce it to a lower output frequency. Both converters are commonly used in transverters and satellite communications. Upconverters achieve this frequency conversion via heterodyning, the same principle as modern receivers and transmitters to offset the frequency.

Analog videotape recording

Many analog videotape systems rely on a downconverted color subcarrier in order to record color information in their limited bandwidth. These systems are referred to as "heterodyne systems" or "color-under systems". For instance, for NTSC video systems, the VHS (and S-VHS) recording system converts the color subcarrier from the NTSC standard 3.58 MHz to ~629 kHz. PAL VHS color subcarrier is similarly downconverted (but from 4.43 MHz). The now-obsolete 3/4" U-matic systems use a heterodyned ~688 kHz subcarrier for NTSC recordings (as does Sony's Betamax, which is at its basis a 1/2" consumer version of U-matic), while PAL U-matic decks came in two mutually incompatible varieties, with different subcarrier frequencies, known as Hi-Band and Low-Band. Other videotape formats with heterodyne color systems include Video-8 and Hi8.

The heterodyne system in these cases is used to convert quadrature phase-encoded and amplitude modulated sine waves from the broadcast frequencies to frequencies recordable in less than 1 MHz bandwidth. On playback, the recorded color information is heterodyned back to the standard subcarrier frequencies for display on televisions and for interchange with other standard video equipment.

Some U-matic (3/4") decks feature 7-pin mini-DIN connectors to allow dubbing of tapes without a heterodyne up-conversion and down-conversion, as do some industrial VHS, S-VHS, and Hi8 recorders.

Music synthesis

The theremin, an electronic musical instrument, uses the heterodyne principle to produce a variable audio frequency in response to the movement of the musician's hands in the vicinity of some antennas. The output of a fixed radio frequency oscillator is mixed with that of an oscillator whose frequency is affected by the variable capacitance between the antenna and the thereminist as that person moves her or his hand near the pitch control antenna. The difference between the two oscillator frequencies produces a tone in the audio range.

The Ring modulator is a type of heterodyne incorporated into some synthesizers or used as a stand-alone audio effect.

Optical heterodyning

Optical heterodyne detection is a current active area of research is extension of the heterodyning technique to higher frequencies, particularly to light frequencies, which are above radio frequencies in the electromagnetic spectrum. This technique could greatly improve optical modulators, increasing the density of information carried by optical fibers. It is also being applied in the creation of more accurate atomic clocks based on directly measuring the frequency of a laser beam.

Since optical frequencies are far beyond any feasible electronic circuit all photon detectors are inherently energy detectors not oscillating electric field detectors. However since energy detection is inherently "square-law" detection, it intrinsically mixes any optical frequencies present on the detector. Thus sensitive detection of specific optical frequencies is possible by optical heterodyne detection when two different (close-by) wavelengths of light illuminate the detector so that the oscillating electrical output corresponds to their difference frequency. This allows extremely narrow band detection (much narrower band than any possible color filter can achieve) as well as precision measurements of phase and frequency of a signal light relative to a reference light source, as in Laser Doppler Vibrometry.

This phase sensitive detection has been applied for Doppler measurements of wind speed, and imaging through dense media. The high sensitivity against background light is especially useful for LIDAR.

In optical Kerr effect (OKE) spectroscopy, optical heterodyning of the OKE signal and a small part of the probe signal produces a mixed signal consisting of probe, heterodyne OKE-probe and homodyne OKE signal. The probe and homodyne OKE signals can be filtered out, leaving the heterodyne signal for detection.

Mathematical principle

Heterodyning is based on the trigonometric identity:

$$\sin\theta\sin\varphi = \frac{1}{2}\cos(\theta-\varphi) - \frac{1}{2}\cos(\theta+\varphi)$$

The product on the left hand side represents the multiplication ("mixing") of a sine wave with another sine wave. The right hand side shows that the resulting signal is the difference of two sinusoidal terms, one at the sum of the two original frequencies, and one at the difference, which can be considered to be separate signals.

Using this trigonometric identity, the result of multiplying two sine wave signals, $\sin(2\pi f_1 t)$ and $\sin(2\pi f_2 t)$ can be calculated:

$$\sin(2\pi f_1 t)\sin(2\pi f_2 t) = \frac{1}{2}\cos[2\pi(f_1 - f_2)t] - \frac{1}{2}\cos[2\pi(f_1 + f_2)t]$$

The result is the sum of two sinusoidal signals, one at the sum $f_1 + f_2$ and one at the difference $f_1 - f_2$ of the original frequencies

The two signals are multiplied in a device called a mixer. In order to multiply the signals, the mixer must be a nonlinear component, that is, its output current or voltage must be a nonlinear function of its input. Most circuit elements in communications circuits are designed to be linear. This means they obey the superposition principle; if *F(v)* is the output of a linear element with an input of *v*:

$$F(v_1 + v_2) = F(v_1) + F(v_2)$$

So if two sine wave signals are applied to a linear device, the output is simply the sum of the outputs when the two signals are applied separately, with no product terms. So the function *F* must be nonlinear. Examples of nonlinear components that are used as mixers are vacuum tubes and transistors biased near cutoff (class C), and diodes. For lower frequencies, IC analog multipliers can be used which multiply signals precisely. Ferromagnetic core inductors driven into saturation can also be used. In nonlinear optics, crystals that have nonlinear characteristics are used to mix laser light beams to create heterodynes at optical frequencies.

To demonstrate mathematically how a nonlinear component can multiply signals and generate heterodyne frequencies, the nonlinear function *F* can be expanded in a power series (MacLaurin series):

$$F(v) = \alpha_1 v + \alpha_2 v^2 + \alpha_3 v^3 + \ldots$$

To simplify the math, the higher order terms above α_2 will be indicated by an ellipsis (". . .") and only the first terms will be shown. Applying the two sine waves at frequencies $\omega_1 = 2\pi f_1$ and $\omega_2 = 2\pi f_2$ to this device:

$$v_{out} = F(A_1 \sin\omega_1 t + A_2 \sin\omega_2 t)$$
$$v_{out} = \alpha_1(A_1 \sin\omega_1 t + A_2 \sin\omega_2 t) + \alpha_2(A_1 \sin\omega_1 t + A_2 \sin\omega_2 t)^2 + \ldots$$
$$v_{out} = \alpha_1(A_1 \sin\omega_1 t + A_2 \sin\omega_2 t) + \alpha_2(A_1^2 \sin^2\omega_1 t + 2A_1 A_2 \sin\omega_1 t \sin\omega_2 t + A_2^2 \sin^2\omega_2 t) + \ldots$$

It can be seen that the second term above contains a product of the two sine waves. Simplifying with trigonometric identities:

$$v_{out} = \alpha_1(A_1 \sin\omega_1 t + A_2 \sin\omega_2 t) + \alpha_2\left(\frac{A_1^2}{2}[1 - \cos 2\omega_1 t] + A_1 A_2[\cos(\omega_1 t - \omega_2 t) - \cos(\omega_1 t + \omega_2 t)] + \frac{A_2^2}{2}[1 - \cos 2\omega_2 t]\right) + \ldots$$
$$v_{out} = \alpha_2 A_1 A_2 \cos(\omega_1 - \omega_2)t - \alpha_2 A_1 A_2 \cos(\omega_1 + \omega_2)t + \ldots$$

So the output contains sinusoidal terms with frequencies at the sum $\omega_1 + \omega_2$ and difference $\omega_1 - \omega_2$ of the two original frequencies. It also contains terms at the original frequencies and at multiples of the original frequencies $2\omega_1$, $2\omega_2$, $3\omega_1$, $3\omega_2$, etc.; the latter are called harmonics. These unwanted frequencies, along with the unwanted heterodyne frequency, must be filtered out of the mixer output to leave the desired heterodyne.

Chapter-12

Photoacoustic Doppler Effect

The **photoacoustic Doppler effect**, as its name implies, is one specific kind of Doppler effect, which occurs when an intensity modulated light wave induces a photoacoustic wave on moving particles with a specific frequency. The observed frequency shift is a good indicator of the velocity of the illuminated moving particles. A potential biomedical application is measuring blood flow.

Specifically, when an intensity modulated light wave is exerted on a localized medium, the resulting heat can induce an alternating and localized pressure change. This periodic pressure change generates an acoustic wave with a specific frequency. Among various factors that determine this frequency, the velocity of the heated area and thus the moving particles in this area can induce a frequency shift proportional to the relative motion. Thus, from the perspective of an observer, the observed frequency shift can be used to derive the velocity of illuminated moving particles..

Theory

To be simple, consider a clear medium firstly. The medium contains small optical absorbers moving with velocity vector $\vec{v}$. The absorbers are irradiated by a laser with intensity modulated at frequency f_0. Thus, the intensity of the laser could be described by:

$$I = I_0 \left[1 + cos\left(2\pi f_0 t\right)\right] /2$$

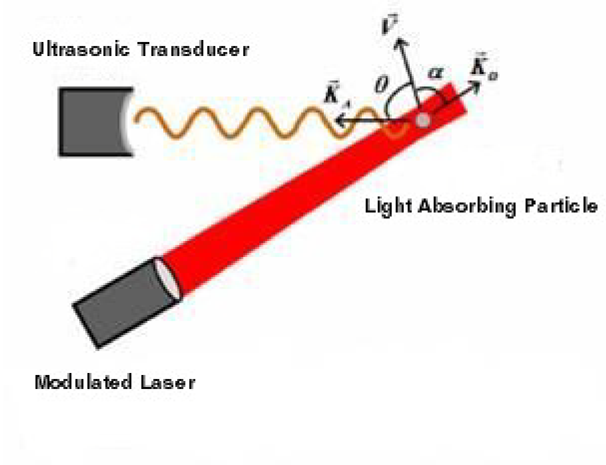


Figure 1: Overview of PAD Effect

When $\vec{v}$ is zero, an acoustic wave with the same frequency f_0 as the light intensity wave is induced. Otherwise, there is a frequency shift in the induced acoustic wave. The magnitude of the frequency shift depends on the relative velocity $\vec{v}$, the angle α between the velocity and the photon density wave propagation direction, and the angle θ between the velocity and the ultrasonic wave propagation direction. The frequency shift is given by:

$$f_{PAD} = -f_0 \frac{v}{c_0} cos\alpha + f_0 \frac{v}{c_a} cos\theta$$

Where c_0 is the speed of light in the medium and c_a is the speed of sound. The first term on the right side of the expression represents the frequency shift in the photon density wave observed by the absorber acting as a moving receiver. The second term represents the frequency shift in the photoacoustic wave due to the motion of the absorbers observed by the ultrasonic transducer.

In practice, since $\frac{c_0}{c_a} \sim 10^5$ and $v \ll c_a$, only the second term is detectable. Therefore, the above equation reduces to:

$$f_{PAD} = f_0 \frac{v}{c_a} cos\theta = \frac{v}{\lambda} cos\theta$$

In this approximation, the frequency shift is not affected by the direction of the optical radiation. It is only affected by the magnitude of velocity and the angle between the velocity and the acoustic wave propagation direction.

This equation also holds for a scattering medium. In this case, the photon density wave becomes diffusive due to light scattering. Although the diffusive photon density wave has a slower phase velocity than the speed of light, its wavelength is still much longer than the acoustic wave.

Experiment

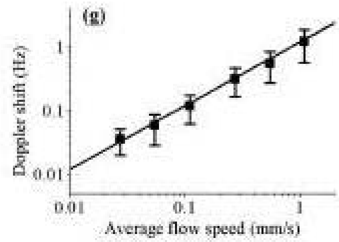


Figure 2: Average Photoacoustic Doppler Shift vs. Velocity for a Scattering Medium

In the first demonstration of the Photoacoustic Doppler effect, a continuous wave diode laser was used in a photoacoustic microscopy setup with an ultrasonic transducer as the detector. The sample was a solution of absorbing particles moving through a tube. The tube was in a water bath containing scattering particles

Figure 2 shows a relationship between average flow velocity and the experimental photoacoustic Doppler frequency shift. In a scattering medium, such as the experimental phantom, fewer photons reach the absorbers than in an optically clear medium. This affects the signal intensity but not the magnitude of the frequency shift. Another demonstrated feature of this technique is that it is capable of measuring flow direction relative to the detector based on the sign of the frequency shift. The reported minimum detected flow rate is 0.027 mm/s in the scattering medium.

Application

One promising application is the non-invasive measurement of flow. This is related to an important problem in medicine: the measurement of blood flow through arteries, capillaries, and veins. Measuring blood velocity in capillaries is an important component to clinically determining how much oxygen is delivered to tissues and is potentially important to the diagnosis of a variety of diseases including diabetes and cancer.However, a particular difficulty of measuring flow velocity in capillaries is caused by the low blood flow rate and micrometre-scale diameter.Photoacoustic Doppler effect based imaging is a promising method for blood flow measurement in capillaries.

Existing techniques

Based on either ultrasound or light there are several techniques currently being used to measure blood velocity in a clinical setting or other types of flow velocities.

Doppler ultrasound

The Doppler ultrasound technique uses Doppler frequency shifts in ultrasound wave. This technique is currently used in biomedicine to measure blood flow in arteries and veins. It is limited to high flow rates (> 1cm/s) generally found in large vessels due to the high background ultrasound signal from biological tissue.

Laser doppler flowmetry

Laser Doppler Flowmetry utilizes light instead of ultrasound to detect flow velocity. The much shorter optical wavelength means this technology is able to detect low flow velocities out of the range of Doppler ultrasound. But this technique is limited by high background noise and low signal due to multiple scattering. Laser Doppler flowmetry can measure only the averaged blood speed within $1mm^3$ without information about flow direction.

Doppler optical coherence tomography

Doppler Optical coherence tomography is an optical flow measurement technique that improves on the spatial resolution of laser Doppler flowmetry by rejecting multiple scattering light with coherent gating. This technique is able to detect flow velocity as low as 100μm/s with the spatial resolution of $5 \times 5 \times 15 \mu m^3$. The detection depth is usually limited by the high optical scattering coefficient of biological tissue to < 1mm.

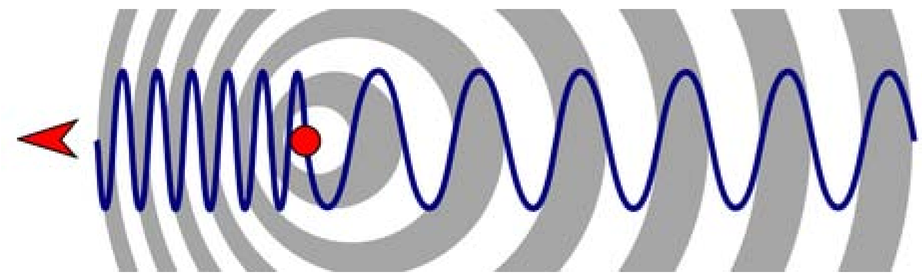
Photoacoustic doppler flowmetry

Photoacoustic Doppler effect can be used to measure the blood flow velocity with the advantages of Photoacoustic imaging. Photoacoustic imaging combines the spatial resolution of ultrasound imaging with the contrast of optical absorption in deep biological tissue. Ultrasound has good spatial resolution in deep biological tissue since ultrasonic scattering is much weaker than optical scattering, but it is insensitive to biochemical properties. Conversely, optical imaging is able to achieve high contrast in biological tissue via high sensitivity to small molecular optical absorbers, such as hemoglobin found in red blood cells, but its spatial resolution is compromised by the strong scattering of light in biological tissue. By combining the optical imaging with ultrasound, it is possible to achieve both high contrast and spatial resolution.

The photoacoustic Doppler flowmetry could use the power of photoacoustics to measure flow velocities that are usually inaccessible to pure light-based or ultrasound techniques. The high spatial resolution could make it possible to pinpoint only a few absorbing particles localized to a single capillary. High contrast from the strong optical absorbers make it possible to clearly resolve the signal from the absorbers over the background.

Chapter-13

Doppler Effect



Change of wavelength caused by motion of the source.

An Doppler effect causes a car engine or siren to sound higher in pitch when it is approaching than when it is receding. The pink circles are sound waves. When the car is moving to the left, each successive wave is emitted from a position further to the left than the previous wave. So for an observer in front *(left)* of the car, each wave takes slightly less time to reach him than the previous wave. The waves "bunch together", so the time between arrival of successive wavefronts is reduced, giving them a higher frequency. For an observer in back *(right)* of the car, each wave takes a slightly longer time to reach him than the previous wave. The waves "stretch apart", so the time between the arrival of successive wavefronts is increased slightly, giving them a lower frequency.

The **Doppler effect** (or **Doppler shift**), named after Austrian physicist Christian Doppler who proposed it in 1842 in Prague, is the change in frequency of a wave for an observer moving relative to the source of the wave. It is commonly heard when a vehicle sounding a siren or horn approaches, passes, and recedes from an observer. The received frequency is higher (compared to the emitted frequency) during the approach, it is identical at the instant of passing by, and it is lower during the recession.

The relative increase in frequency can be explained as follows. When the source of the waves is moving toward the observer, each successive wave crest is emitted from a position closer to the observer than the previous wave. Therefore each wave takes slightly less time to reach the observer than the previous wave. Therefore the time between the arrival of successive wave crests at the observer is reduced, causing an increase in the frequency. While they are traveling, the distance between successive wavefronts is reduced; so the waves "bunch together". Conversely, if the source of waves is moving away from the observer, each wave is emitted from a position farther from the observer than the previous wave, so the arrival time between successive waves is increased, reducing the frequency. The distance between successive wavefronts is increased, so the waves "spread out".

For waves that propagate in a medium, such as sound waves, the velocity of the observer and of the source are relative to the medium in which the waves are transmitted. The total Doppler effect may therefore result from motion of the source, motion of the observer, or motion of the medium. Each of these effects is analyzed separately. For waves which do not require a medium, such as light or gravity in general relativity, only the relative difference in velocity between the observer and the source needs to be considered.

Development

Doppler first proposed the effect in 1842 in his treatise "*Über das farbige Licht der Doppelsterne und einiger anderer Gestirne des Himmels*" (On the coloured light of the binary stars and some other stars of the heavens). The hypothesis was tested for sound waves by Buys Ballot in 1845. He confirmed that the sound's pitch was higher than the emitted frequency when the sound source approached him, and lower than the emitted frequency when the sound source receded from him. Hippolyte Fizeau discovered independently the same phenomenon on electromagnetic waves in 1848 (in France, the effect is sometimes called "effet Doppler-Fizeau" but was not adopted by the rest of the world as Fizeau's discovery was three years after Doppler's; therefore, it was only an addition to a previously established discovery). In Britain, John Scott Russell made an experimental study of the Doppler effect (1848).

An English translation of Doppler's 1842 treatise can be found in the book '*The Search for Christian Doppler* by Alec Eden.

General

In classical physics, where the speeds of source and the receiver relative to the medium are lower than the velocity of waves in the medium, the relationship between observed frequency f and emitted frequency f_0 is given by:

$$f = \left(\frac{v + v_r}{v + v_s}\right) f_0$$

where

v is the velocity of waves in the medium

v_r is the velocity of the receiver relative to the medium; positive if the receiver is moving towards the source.

v_s is the velocity of the source relative to the medium; positive if the source is moving away from the receiver.

The frequency is decreased if either is moving away from the other.

The above formula works for sound wave if and only if the speeds of the source and receiver relative to the medium are slower than the speed of sound.

The above formula assumes that the source is either directly approaching or receding from the observer. If the source approaches the observer at an angle (but still with a constant velocity), the observed frequency that is first heard is higher than the object's emitted frequency. Thereafter, there is a monotonic decrease in the observed frequency as it gets closer to the observer, through equality when it is closest to the observer, and a continued monotonic decrease as it recedes from the observer. When the observer is very close to the path of the object, the transition from high to low frequency is very abrupt. When the observer is far from the path of the object, the transition from high to low frequency is gradual.

In the limit where the speed of the wave is much greater than the relative speed of the source and observer (this is often the case with electromagnetic waves, e.g. light), the relationship between observed frequency f and emitted frequency f_0 is given by:

Observed frequency	Change in frequency
$f = \left(1 - \frac{v_{s,r}}{c}\right) f_0$	$\Delta f = -\frac{v_{s,r}}{c} f_0 = -\frac{v_{s,r}}{\lambda_0}$

where

$v_{s,r} = v_s - v_r$ is the velocity of the source relative to the receiver: it is positive when the source and the receiver are moving further apart.

c is the speed of wave (e.g. 3×10^8 m/s for electromagnetic waves travelling in a vacuum)

λ_0 is the wavelength of the transmitted wave in the reference frame of the source.

These two equations are only accurate to a first order approximation. However, they work reasonably well when the speed between the source and receiver is slow relative to the speed of the waves involved and the distance between the source and receiver is large relative to the wavelength of the waves. If either of these two approximations are violated, the formulae are no longer accurate.

Analysis

The frequency of the sounds that the source *emits* does not actually change. To understand what happens, consider the following analogy. Someone throws one ball every second in a man's direction. Assume that balls travel with constant velocity. If the thrower is stationary, the man will receive one ball every second. However, if the thrower is moving towards the man, he will receive balls more frequently because the balls will be less spaced out. The inverse is true if the thrower is moving away from the man. So it is actually the *wavelength* which is affected; as a consequence, the received frequency is also affected. It may also be said that the velocity of the wave remains constant whereas wavelength changes; hence frequency also changes.

If the source moving away from the observer is emitting waves through a medium with an actual frequency f_0, then an observer stationary relative to the medium detects waves with a frequency f given by

$$f = \left(\frac{v}{v + v_s}\right) f_0$$

where v_s is positive if the source is moving away from the observer, and negative if the source is moving towards the observer.

A similar analysis for a moving *observer* and a stationary source yields the observed frequency (the receiver's velocity being represented as v_r):

$$f = \left(\frac{v + v_r}{v}\right) f_0$$

where the similar convention applies: v_r is positive if the observer is moving towards the source, and negative if the observer is moving away from the source.

These can be generalized into a single equation with both the source and receiver moving.

$$f = \left(\frac{v + v_r}{v + v_s}\right) f_0$$

With a relatively slow moving source, $v_{s,r}$ is small in comparison to v and the equation approximates to

$$f = \left(1 - \frac{v_{s,r}}{v}\right) f_0$$

where $v_{s,r} = v_s - v_r$.

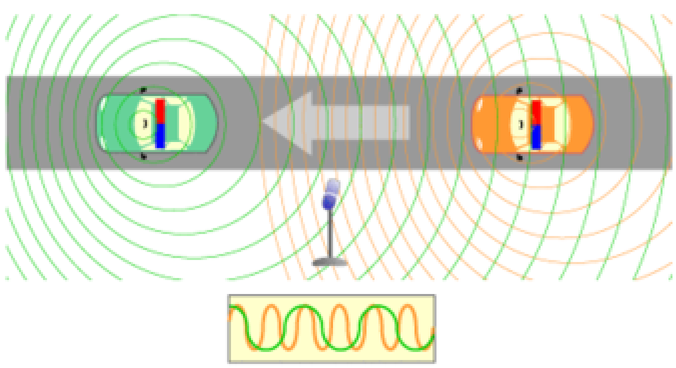
However the limitations mentioned above still apply. When the more complicated exact equation is derived without using any approximations (just assuming that source, receiver, and wave or signal are moving linearly relatively to each other) several interesting and perhaps surprising results are found. For example, as Lord Rayleigh noted in his classic book on sound, by properly moving it would be possible to hear a symphony being played backwards. This is the so-called "time reversal effect" of the Doppler effect. Other interesting conclusions are that the Doppler effect is time-dependent in general (thus we need to know not only the source and receivers' velocities, but also their positions at a given time), and in some circumstances it is possible to receive two signals or waves from a source, or no signal at all. In addition there are more possibilities than just the receiver approaching the signal and the receiver receding from the signal.

All these additional complications are derived for the classical, i.e., non-relativistic, Doppler effect, but hold for the relativistic Doppler effect as well.

A common misconception

Craig Bohren pointed out in 1991 that some physics textbooks erroneously state that the observed frequency *increases* as the object approaches an observer and then decreases only as the object passes the observer. In most cases, the observed frequency of an approaching object declines monotonically from a value above the emitted frequency, through a value equal to the emitted frequency when the object is closest to the observer, and to values increasingly below the emitted frequency as the object recedes from the observer. Bohren proposed that this common misconception might occur because the *intensity* of the sound increases as an object approaches an observer and decreases once it passes and recedes from the observer and that this change in intensity is misperceived as a change in frequency. Higher sound pressure levels make for a small decrease in perceived pitch in low frequency sounds, and for a small increase in perceived pitch for high frequency sounds.

Application



A stationary microphone records moving police sirens at different pitches depending on their relative direction.

Sirens

The siren on a passing emergency vehicle will start out higher than its stationary pitch, slide down as it passes, and continue lower than its stationary pitch as it recedes from the observer. Astronomer John Dobson explained the effect thus:

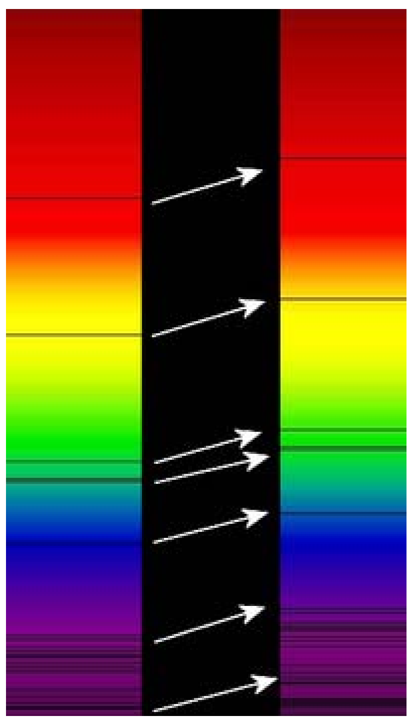
> "The reason the siren slides is because it doesn't hit you."

In other words, if the siren approached the observer directly, the pitch would remain constant (as $v_{s,\,r}$ is only the radial component) until the vehicle hit him, and then immediately jump to a new lower pitch. Because the vehicle passes by the observer, the radial velocity does not remain constant, but instead varies as a function of the angle between his line of sight and the siren's velocity:

$$v_r = v_s \cdot \cos\theta$$

where v_s is the velocity of the object (source of waves) with respect to the medium, and θ is the angle between the object's forward velocity and the line of sight from the object to the observer.

Astronomy



Redshift of spectral lines in the optical spectrum of a supercluster of distant galaxies (right), as compared to that of the Sun (left).

The Doppler effect for electromagnetic waves such as light is of great use in astronomy and results in either a so-called redshift or blue shift. It has been used to measure the speed at which stars and galaxies are approaching or receding from us, that is, the radial velocity. This is used to detect if an apparently single star is, in reality, a close binary and even to measure the rotational speed of stars and galaxies.

The use of the Doppler effect for light in astronomy depends on our knowledge that the spectra of stars are not continuous. They exhibit absorption lines at well defined frequencies that are correlated with the energies required to excite electrons in various elements from one level to another. The Doppler effect is recognizable in the fact that the absorption lines are not always at the frequencies that are obtained from the spectrum of a stationary light source. Since blue light has a higher frequency than red light, the spectral lines of an approaching astronomical light source exhibit a blue shift and those of a receding astronomical light source exhibit a redshift.

Among the nearby stars, the largest radial velocities with respect to the Sun are +308 km/s (BD-15°4041, also known as LHS 52, 81.7 light-years away) and -260 km/s (Woolley 9722, also known as Wolf 1106 and LHS 64, 78.2 light-years away). Positive radial velocity means the star is receding from the Sun, negative that it is approaching.

Temperature measurement

Another use of the Doppler effect, which is found mostly in plasma physics and astronomy, is the estimation of the temperature of a gas (or ion temperature in a plasma) which is emitting a spectral line. Due to the thermal motion of the emitters, the light emitted by each particle can be slightly red- or blue-shifted, and the net effect is a broadening of the line. This line shape is called a Doppler profile and the width of the line is proportional to the square root of the temperature of the emitting species, allowing a spectral line (with the width dominated by the Doppler broadening) to be used to infer the temperature.

Radar

The Doppler effect is used in some types of radar, to measure the velocity of detected objects. A radar beam is fired at a moving target — e.g. a motor car, as police use radar to detect speeding motorists — as it approaches or recedes from the radar source. Each successive radar wave has to travel farther to reach the car, before being reflected and re-detected near the source. As each wave has to move farther, the gap between each wave increases, increasing the wavelength. In some situations, the radar beam is fired at the moving car as it approaches, in which case each successive wave travels a lesser distance, decreasing the wavelength. In either situation, calculations from the Doppler effect accurately determine the car's velocity. Moreover, the proximity fuze, developed during World War II, relies upon Doppler radar to explode at the correct time, height, distance, etc.

Medical imaging and blood flow measurement

An echocardiogram can, within certain limits, produce accurate assessment of the direction of blood flow and the velocity of blood and cardiac tissue at any arbitrary point using the Doppler effect. One of the limitations is that the ultrasound beam should be as parallel to the blood flow as possible. Velocity measurements allow assessment of cardiac valve areas and function, any abnormal communications between the left and right side of the heart, any leaking of blood through the valves (valvular regurgitation), and calculation of the cardiac output. Contrast-enhanced ultrasound using gas-filled microbubble contrast media can be used to improve velocity or other flow-related medical measurements.

Although "Doppler" has become synonymous with "velocity measurement" in medical imaging, in many cases it is not the frequency shift (Doppler shift) of the received signal that is measured, but the phase shift (*when* the received signal arrives).

Velocity measurements of blood flow are also used in other fields of medical ultrasonography, such as obstetric ultrasonography and neurology. Velocity measurement of blood flow in arteries and veins based on Doppler effect is an effective tool for diagnosis of vascular problems like stenosis.

Flow measurement

Instruments such as the laser Doppler velocimeter (LDV), and acoustic Doppler velocimeter (ADV) have been developed to measure velocities in a fluid flow. The LDV emits a light beam and the ADV emits an ultrasonic acoustic burst, and measure the Doppler shift in wavelengths of reflections from particles moving with the flow. The actual flow is computed as a function of the water velocity and phase. This technique allows non-intrusive flow measurements, at high precision and high frequency.

Velocity profile measurement

Developed originally for velocity measurements in medical applications (blood flows), Ultrasonic Doppler Velocimetry (UDV) can measure in real time complete velocity profile in almost any liquids containing particles in suspension such as dust, gas bubbles, emulsions. Flows can be pulsating, oscillating, laminar or turbulent, stationary or transient. This technique is fully non-invasive.

Underwater acoustics

In military applications the Doppler shift of a target is used to ascertain the speed of a submarine using both passive and active sonar systems. As a submarine passes by a passive sonobuoy, the stable frequencies undergo a Doppler shift, and the speed and range from the sonobuoy can be calculated. If the sonar system is mounted on a moving ship or another submarine, then the relative velocity can be calculated.

Audio

The Leslie speaker, associated with and predominantly used with the Hammond B-3 organ, takes advantage of the Doppler Effect by using an electric motor to rotate an acoustic horn around a loudspeaker, sending its sound in a circle. This results at the listener's ear in rapidly fluctuating frequencies of a keyboard note.

Vibration measurement

A laser Doppler vibrometer (LDV) is a non-contact method for measuring vibration. The laser beam from the LDV is directed at the surface of interest, and the vibration amplitude and frequency are extracted from the Doppler shift of the laser beam frequency due to the motion of the surface.