Telecommunication Theory in Signal Processing

Lois Epstein

First Edition, 2012

ISBN 978-81-323-4036-2

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

Table of Contents

Chapter-1

Spectral Efficiency

Spectral efficiency, **spectrum efficiency** or **bandwidth efficiency** refers to the information rate that can be transmitted over a given bandwidth in a specific communication system. It is a measure of how efficiently a limited frequency spectrum is utilized by the physical layer protocol, and sometimes by the media access control (the channel access protocol).

Link spectral efficiency

The **link spectral efficiency** of a digital communication system is measured in *bit/s/Hz*, or, less frequently but unambiguously, in *(bit/s)/Hz*. It is the net bitrate (useful information rate excluding error-correcting codes) or maximum throughput divided by the bandwidth in hertz of a communication channel or a data link. Alternatively, the spectral efficiency may be measured in in *bit/symbol*, which is equivalent to *bits per channel use* (*bpcu*), implying that the net bit rate is divided by the symbol rate (modulation rate) or line code pulse rate.

Link spectral efficiency is typically used to analyse the efficiency of a digital modulation method or line code, sometimes in combination with a forward error correction (FEC) code and other physical layer overhead. In the latter case, a "bit" refers to a user data bit; FEC overhead is always excluded.

The **modulation efficiency** in bit/s is the gross bitrate (including any error-correcting code) divided by the bandwidth.

Example 1: A transmission technique using one kilohertz of bandwidth to transmit 1,000 bits per second has a modulation efficiency of 1 (bit/s)/Hz.
Example 2: A V.92 modem for the telephone network can transfer 56,000 bit/s downstream and 48,000 bit/s upstream over an analog telephone network. Due to filtering in the telephone exchange, the frequency range is limited to between 300 hertz and 3,400 hertz, corresponding to a bandwidth of 3,400 − 300 = 3,100 hertz. The spectral efficiency or modulation efficiency is 56,000/3,100 = 18.1 (bit/s)/Hz downstream, and 48,000/3,100 = 15.5 (bit/s)/Hz upstream.

An upper bound for the attainable modulation efficiency is given by the Nyquist rate or Hartley's law as follows: For a signaling alphabet with M alternative symbols, each symbol represents $N = \log_2 M$ bits. N is the modulation efficiency measured in *bit/symbol* or *bpcu*. In the case of baseband transmission (line coding or pulse-amplitude modulation) with a baseband bandwidth (or upper cut-off frequency) B, the symbol rate can not exceed $2B$ symbols/s in view to avoid intersymbol interference. Thus, the spectral efficiency can not exceed $2N$ (bit/s)/Hz in the baseband transmission case. In the passband transmission case, a signal with passband bandwidth W can be converted to an equivalent baseband signal (using undersampling or a superheterodyne receiver), with upper cut-off frequency $W/2$. If double-sideband modulation schemes such as QAM, ASK, PSK or OFDM are used, this results in a maximum symbol rate of W symbols/s, and in that the modulation efficiency can not exceed N (bit/s)/Hz. If digital single-sideband modulation is used, the passband signal with bandwidth W corresponds to a baseband message signal with baseband bandwidth W, resulting in a maximum symbol rate of $2W$ and an attainable modulation efficiency of $2N$ (bit/s)/Hz.

Example 3: An 16QAM modem has an alphabet size of $M = 16$ alternative symbols, with $N = 4$ bit/symbol or bpcu. Since QAM is a form of double sideband passband transmission, the spectral efficiency cannot exceed $N = 4$ (bit/s)/Hz.
Example 4: The 8VSB (8-level vestigial sideband) modulation scheme used in the ATSC digital television standard gives N=3 bit/symbol or bpcu. Since it can be described as nearly single-side band, the modulation efficiency is close to $2N = 6$ (bit/s)/Hz. In practice, ATSC transfers a gross bit rate of 32 Mbit/s over a 6 MHz wide channel, resulting in a modulation efficiency of 32/6 = 5.3 (bit/s)/Hz.
Example 5: The downlink of a V.92 modem uses a pulse-amplitude modulation with 128 signal levels, resulting in $N = 7$ bit/symbol. Since the transmitted signal before passband filtering can be considered as baseband transmission, the spectral efficiency cannot exceed $2N = 14$ (bit/s)/Hz over the full baseband channel (0 to 4 kHz). As seen above, a higher spectral efficiency is achieved if we consider the smaller passband bandwidth.

If a forward error correction code is used, the spectral efficiency is reduced from the uncoded modulation efficiency figure.

Example 6: If a forward error correction (FEC) code with code rate 1/2 is added, meaning that the encoder input bit rate is one half the encoder output rate, the spectral efficiency is 50% of the modulation efficiency. In exchange for this reduction in spectral efficiency, FEC usually reduces the bit-error rate, and typically enables operation at a lower signal to noise ratio (SNR).

An upper bound for the spectral efficiency possible without bit errors in a channel with a certain SNR, if ideal error coding and modulation is assumed, is given by the Shannon-Hartley theorem.

Example 7: If the SNR is 1 times expressed as a ratio, corresponding to 0 decibel, the link spectral efficiency can not exceed 1 (bit/s)/Hz for error-free detection (assuming an ideal error-correcting code) according to Shannon-Hartley regardless of the modulation and coding.

Note that the goodput (the amount of application layer useful information) is normally lower than the maximum throughput used in the above calculations, because of packet retransmissions, higher protocol layer overhead, flow control, congestion avoidance, etc. On the other hand, a data compression scheme, such as the V.44 or V.42bis compression used in telephone modems, may however give higher goodput if the transferred data is not already efficiently compressed.

The link spectral efficiency of a wireless telephony link may also be expressed as the maximum number of simultaneous calls over 1 MHz frequency spectrum in erlangs per megahertz, or *E/MHz*. This measure is also affected by the source coding (data compression) scheme. It may be applied to analog as well as digital transmission.

In wireless networks, the *link spectral efficiency* can be somewhat misleading, as larger values are not necessarily more efficient in their overall use of radio spectrum. In a wireless network, high link spectral efficiency may result in high sensitivity to co-channel interference (crosstalk), which affects the capacity. For example, in a cellular telephone network with frequency reuse, spectrum spreading and forward error correction reduce the spectral efficiency in (bit/s)/Hz but substantially lower the required signal-to-noise ratio in comparison to non-spread spectrum techniques. This can allow for much denser geographical frequency reuse that compensates for the lower link spectral efficiency, resulting in approximately the same capacity (the same number of simultaneous phone calls) over the same bandwidth, using the same number of base station transmitters. As discussed below, a more relevant measure for wireless networks would be *system spectral efficiency* in bit/s/Hz per unit area. However, in closed communication links such as telephone lines and cable TV networks, and in noise-limited wireless communication system where co-channel interference is not a factor, the largest link spectral efficiency that can be supported by the available SNR is generally used.

System spectral efficiency or area spectral efficiency

In digital wireless networks, the *system spectral efficiency* or area spectral efficiency is typically measured in (bit/s)/Hz per unit area, (bit/s)/Hz per cell, or (bit/s)/Hz per site. It is a measure of the quantity of users or services that can be simultaneously supported by a limited radio frequency bandwidth in a defined geographic area. It may for example be defined as the maximum throughput or goodput, summed over all users in the system, divided by the channel bandwidth. This measure is affected not only by the single user transmission technique, but also by multiple access schemes and radio resource management techniques utilized. It can be substantially improved by dynamic radio resource management. If it is defined as a measure of the maximum goodput, retransmissions due to co-channel interference and collisions are excluded. Higher-layer protocol overhead (above the media access control sublayer) is normally neglected.

Example 8: In a cellular system based on frequency-division multiple access (FDMA) with a fixed channel allocation (FCA) cellplan using a frequency reuse factor of 4, each base station has access to 1/4 of the total available frequency spectrum. Thus, the maximum possible system spectral efficiency in *(bit/s)/Hz per site* is 1/4 of the link spectral efficiency. Each base station may be divided into 3 cells by means of 3 sector antennas, also known as a 4/12 reuse pattern. Then each cell has access to 1/12 of the available spectrum, and the system spectral efficiency in *(bit/s)/Hz per cell* or *(bit/s)/Hz per sector* is 1/12 of the link spectral efficiency.

The system spectral efficiency of a cellular network may also be expressed as the maximum number of simultaneous phone calls per area unit over 1 MHz frequency spectrum in E/MHz per cell, E/MHz per sector, E/MHz per site, or (E/MHz)/m^2. This measure is also affected by the source coding (data compression) scheme. It may be used in analog cellular networks as well.

Low link spectral efficiency in (bit/s)/Hz does not necessarily mean that an encoding scheme is inefficient from a system spectral efficiency point of view. As an example, consider Code Division Multiplexed Access (CDMA) spread spectrum, which is not a particularly spectral efficient encoding scheme when considering a single channel or single user. However, the fact that one can "layer" multiple channels on the same frequency band means that the system spectrum utilization for a multi-channel CDMA system can be very good.

Example 9: In the W-CDMA 3G cellular system, every phone call is compressed to a maximum of 8,500 bit/s (the useful bitrate), and spread out over a 5 MHz wide frequency channel. This corresponds to a link throughput of only 8,500/5,000,000 = 0.0017 *(bit/s)/Hz*. Let us assume that 100 simultaneous (non-silent) simultaneous calls are possible in the same cell. Spread spectrum makes it possible to have as low a frequency reuse factor as 1, if each base station is divided into 3 cells by means of 3 directional sector antennas. This corresponds to a system spectrum efficiency of over 1 × 100 × 0.0017 = 0.17 *(bit/s)/Hz per site*, and 0.17/3 = 0.06 *(bit/s)/Hz per cell or sector*.

The spectral efficiency can be improved by radio resource management techniques such as efficient fixed or dynamic channel allocation, power control, link adaptation and diversity schemes.

A combined fairness measure and system spectral efficiency measure is the fairly shared spectral efficiency.

Comparison table

Examples of numerical spectral efficiency values of some common communication systems can be found in the table below.

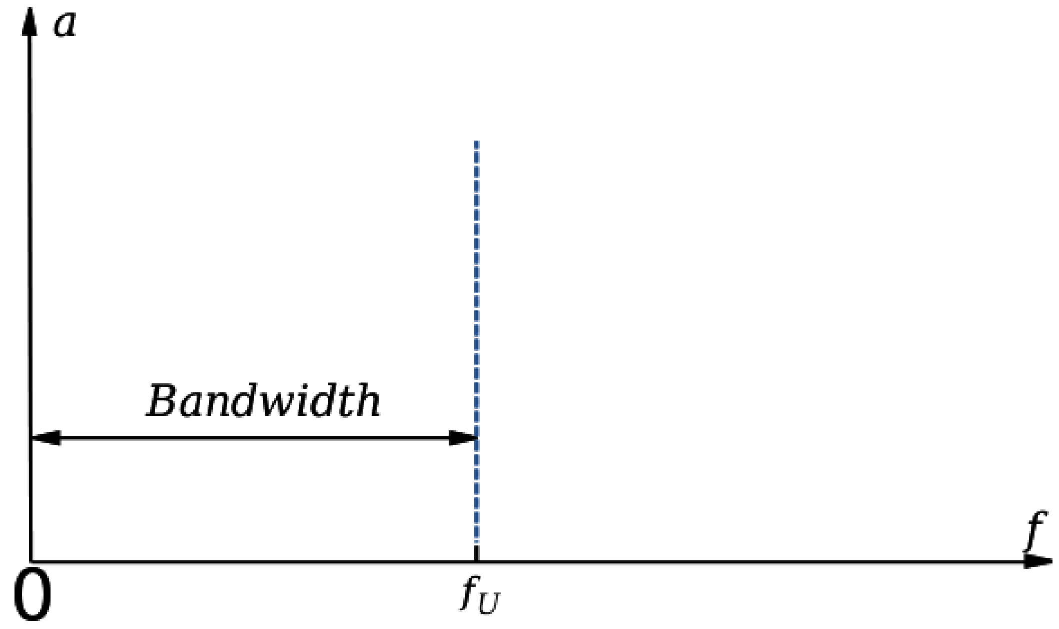
Spectral efficiency of common communication systems.

Service	**Standard**	**Launched year**	**Net bitrate *R* per carrier** (Mbit/s)	**Bandwidth *B* per carrier** (MHz)	**Link spectral efficiency *R/B*** ((bit/s)/Hz)	**Typical reuse factor *1/K***	**System spectral efficiency** **Approx. ((*R*/*B*)/*K*)** ((bit/s)/Hz per site)
1G cellular	**NMT 450 modem**	1981	0.0012	0.025	0.45	⅐	0.064
1G cellular	**AMPS modem**	1983	0.0003	0.030	0.001	⅐	0.0015
2G cellular	**GSM**	1991	0.013 × 8 timeslots = 0.104	0.2	0.52	⅑ (⅓ in 1999)	0.17 (in 1999)
2G cellular	**D-AMPS**	1991	0.013 × 3 timeslots = 0.039	0.030	1.3	⅑ (⅓ in 1999)	0.45 (in 1999)
2.75G cellular	**CDMA2000 1× voice**	2000	Max. 0.0096 per phone call × typ 22 calls per carrier	1.2288	0.0078 per mobile × typ 22 calls per carrier	1	0.172 (fully loaded)
2.75G cellular	**GSM + EDGE**	2003	Max.: 0.384; Typ.: 0.20;	0.2	Max.: 1.92; Typ.: 1.00;	⅓	0.33
2.75G cellular	**IS-136HS + EDGE**		Max.: 0.384; Typ.: 0.27;	0.2	Max.: 1.92; Typ.: 1.35;	⅓	0.45
3G cellular	**WCDMA FDD**	2001	Max.: 0.384 per mobile;	5	Max.: 0.077 per mobile;	1	Max 0.51
3G cellular	**CDMA2000 1x PD**	2002	Max.: 0.153 per mobile;	1.2288	Max.: 0.125 per mobile;	1	Max 0.1720 (fully loaded)
3G cellular	**CDMA2000 1×EV-DO Rev.A**	2002	Max.: 3.072 per mobile;	1.2288	Max.: 2.5 per mobile;	1	Max 1.3 average loaded

							sector
Fixed WiMAX	**IEEE 802.16d**	2004	96	20 (1.75, 3.5, 7, ...)	4.8	¼	1.2
3.5G cellular	**HSDPA**	2007	Max.: 42.2 per mobile;	5	Max.: 8.44 per mobile;	1	Max 8.44
3.9G MBWA	**iBurst HC-SDMA**	2005	Max.: 3.9 per carrier;	0.625	Max.: 7.23 per carrier;	1	Max 7.23
3.9G cellular	**LTE**	2009	Max.: 326.4 per mobile;	20	Max.: 16.32 per mobile;	1	Max.: 16.32;
Wi-Fi	**IEEE 802.11a/g**	2003	Max.: 54;	20	Max.: 2.7;	⅓	0.9
Wi-Fi	**IEEE 802.11n Draft 2.0**	2007	Max.: 144.4;	20	Max.: 7.22;	⅓	Max 2.4
TETRA	**ETSI**	1998	4 timeslots = 0.036	0.025	1.44		
Digital radio	**DAB**	1995	0.576 to 1.152	1.712	0.34 to 0.67	⅕	0.08 to 0.17
Digital radio	**DAB with SFN**	1995	0.576 to 1.152	1.712	0.34 to 0.67	1	0.34 to 0.67
Digital TV	**DVB-T**	1997	Max.: 31.67; Typ.: 22.0;	8	Max.: 4.0; Typ.: 2.8;	⅕	0.55
Digital TV	**DVB-T with SFN**	1996	Max.: 31.67; Typ.: 22.0;	8	Max.: 4.0; Typ.: 2.8;	1	Max.: 4.0; Typ.: 2.8;
Digital TV	**DVB-H**	2007	5.5 to 11	8	0.68 to 1.4	⅕	0.14 to 0.28
Digital TV	**DVB-H with SFN**	2007	5.5 to 11	8	0.68 to 1.4	1	0.68 to 1.4
Digital cable TV	**DVB-C 256-QAM mode**		38	6	6.33	N/A	N/A
Broadband modem	**ADSL2 downlink**		12	0.962	12.47	N/A	N/A
Telephone modem	**V.92 downlink**	1999	0.056	0.004	14.0	N/A	N/A

Chapter-2

Bandwidth (signal processing)



Baseband bandwidth. Here the bandwidth equals the upper frequency.

Bandwidth is the difference between the upper and lower frequencies in a contiguous set of frequencies. It is typically measured in hertz, and may sometimes refer to *passband bandwidth*, sometimes to *baseband bandwidth*, depending on context. **Passband bandwidth** is the difference between the upper and lower cutoff frequencies of, for example, an electronic filter, a communication channel, or a signal spectrum. In case of a low-pass filter or baseband signal, the bandwidth is equal to its upper cutoff frequency. The term **baseband bandwidth** always refers to the upper cutoff frequency, regardless of whether the filter is bandpass or low-pass.

Bandwidth in hertz is a central concept in many fields, including electronics, information theory, radio communications, signal processing, and spectroscopy. A key characteristic of bandwidth is that a band of a given width can carry the same amount of information, regardless of where that band is located in the frequency spectrum (assuming equivalent noise level). For example, a 5 kHz band can carry a telephone conversation whether that band is at baseband (as in your POTS telephone line) or modulated to some higher (passband) frequency.

In computer networking and other digital fields, the term bandwidth often refers to a data rate measured in bits per second, for example network throughput, sometimes denoted *network bandwidth*, *data bandwidth* or *digital bandwidth*. The reason is that according to Hartley's law, the digital data rate limit (or channel capacity) of a physical communication link is proportional to its bandwidth in hertz, sometimes denoted **radio frequency (RF) bandwidth**, **signal bandwidth**, **frequency bandwidth**, **spectral bandwidth** or **analog bandwidth**. For *bandwidth* as a computing term, less ambiguous terms are bit rate, throughput, maximum throughput, goodput or channel capacity.

Overview

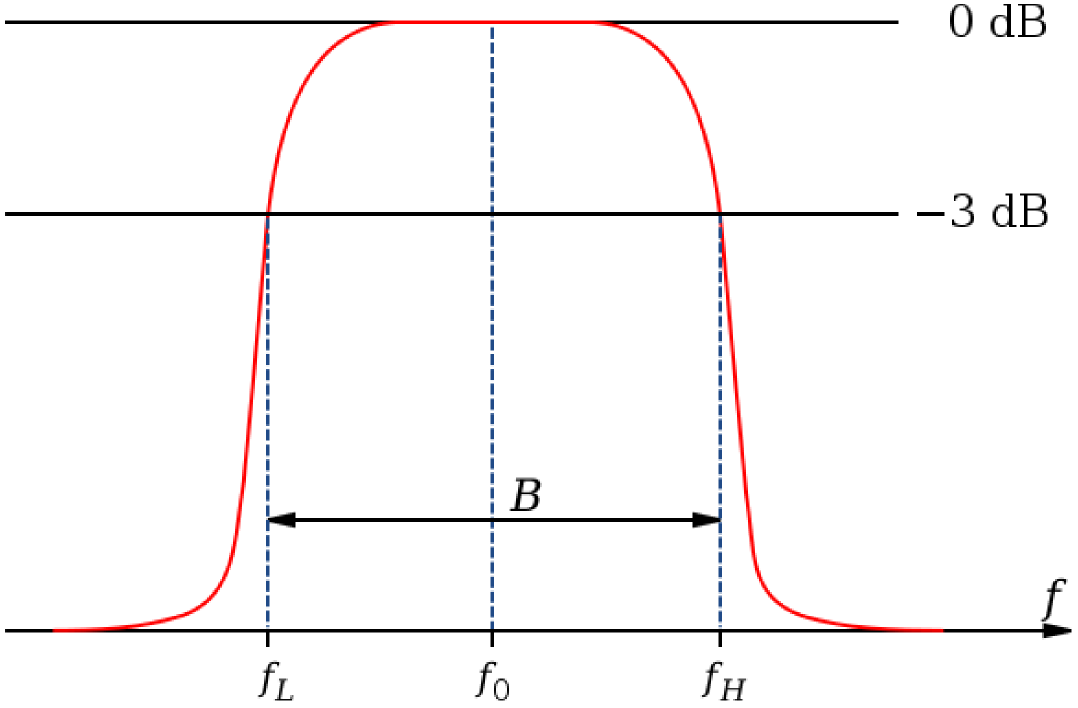
Bandwidth is a key concept in many telephony applications. In radio communications, for example, bandwidth is the frequency range occupied by a modulated carrier wave, whereas in optics it is the width of an individual spectral line or the entire spectral range.

In many signal processing contexts, bandwidth is a valuable and limited resource. For example, an FM radio receiver's tuner spans a limited range of frequencies. A government agency (such as the Federal Communications Commission in the United States) may apportion the regionally available bandwidth to licensed broadcasters so that their signals do not mutually interfere. Each transmitter owns a slice of bandwidth, a valuable (if intangible) commodity.

For different applications there are different precise definitions. For example, one definition of bandwidth could be the range of frequencies beyond which the frequency function is zero. This would correspond to the mathematical notion of the support of a function (i.e., the total "length" of values for which the function is nonzero). A less strict and more practically useful definition will refer to the frequencies where the frequency function is *small*. Small could mean less than 3 dB below (i.e., less than half of) the maximum value, or more rarely 10 dB below, or it could mean below a certain absolute value. As with any definition of the *width* of a function, many definitions are suitable for different purposes.

Bandwidth typically refers to baseband bandwidth in the context of for example sampling theorem and Nyquist sampling rate, while it refers to passband bandwidth in the context of Nyquist symbol rate or Shannon-Hartley channel capacity for communication systems.

X-dB bandwidth



A graph of a bandpass filter's gain magnitude, illustrating the concept of −3 dB bandwidth at a gain of 0.707. The frequency axis of this symbolic diagram can be linear or logarithmically scaled.

In some contexts, the signal bandwidth in hertz refers to the frequency range in which the signal's spectral density is nonzero or above a small threshold value. That definition is used in calculations of the lowest sampling rate that will satisfy the sampling theorem. Because this range of non-zero amplitude may be very broad or infinite, this definition is typically relaxed so that the bandwidth is defined as the range of frequencies in which the signal's spectral density is above a certain threshold relative to its maximum. Most commonly, bandwidth refers to the 3-dB bandwidth, that is, the frequency range within which the spectral density (in W/Hz or V^2/Hz) is above half its maximum value (or the spectral amplitude, in V or V/Hz, is more than 70.7% of its maximum); that is, above −3 dB relative to the peak.

The word bandwidth applies to signals as described above, but it could also apply to *systems*, for example filters or communication channels. To say that a system has a certain bandwidth means that the system can process signals of that bandwidth, or that the system reduces the bandwidth of a white noise input to that bandwidth.

The 3 dB bandwidth of an electronic filter or communication channel is the part of the system's frequency response that lies within 3 dB of the response at its peak, which in the passband filter case is typically at or near its center frequency, and in the lowpass filter is near 0 hertz. If the maximum gain is 0 dB, the 3 dB gain is the range where the gain is more than -3dB, or the attenuation is less than + 3dB. This is also the range of frequencies where the amplitude gain is above 70.7% of the maximum amplitude gain, and above half the maximum power gain. This same "half power gain" convention is also used in spectral width, and more generally for extent of functions as full width at half maximum (FWHM).

In electronic filter design, a filter specification may require that within the filter passband, the gain is nominally 0 dB +/- a small number of dB , for example within the +/- 1 dB interval. In the stopband(s), the required attenuation in dB is above a certain level, for example >100 dB. In a transition band the gain is not specified. In this case, the filter bandwidth corresponds to the passband width, which in this example is the 1dB-bandwidth. If the filter shows amplitude ripple within the passband, the x dB point refers to the point where the gain is x dB below the nominal passband gain rather than x dB below the maximum gain.

A commonly used quantity is *fractional bandwidth*. This is the bandwidth of a device divided by its center frequency. E.g., a passband filter that has a bandwidth of 2 MHz with center frequency 10 MHz will have a fractional bandwidth of 2/10, or 20%.

In communication systems, in calculations of the Shannon–Hartley channel capacity, bandwidth refers to the 3dB-bandwidth. In calculations of the maximum symbol rate, the Nyquist sampling rate, and maximum bit rate according to the Hartley formula, the bandwidth refers to the frequency range within which the gain is non-zero, or the gain in dB is below a very large value.

The fact that in equivalent baseband models of communication systems, the signal spectrum consists of both negative and positive frequencies, can lead to confusion about bandwidth, since they are sometimes referred to only by the positive half, and one will occasionally see expressions such as $B = 2W$, where B is the total bandwidth (i.e. the maximum passband bandwidth of the carrier-modulated RF signal and the minimum passband bandwidth of the physical passband channel), and W is the positive bandwidth (the baseband bandwidth of the equivalent channel model). For instance, the baseband model of the signal would require a lowpass filter with cutoff frequency of at least W to stay intact, and the physical passband channel would require a passband filter of at least B to stay intact.

In signal processing and control theory the bandwidth is the frequency at which the closed-loop system gain drops 3 dB below peak.

In basic electric circuit theory, when studying band-pass and band-reject filters, the bandwidth represents the distance between the two points in the frequency domain where the signal is $\frac{1}{\sqrt{2}}$of the maximum signal amplitude (half power).

Antenna systems

In the field of antennas, two different methods of expressing relative bandwidth are used for narrowband and wideband antennas. For either, a set of criteria is established to define the extents of the bandwidth, such as input impedance, pattern, or polarization.

Percent bandwidth, usually used for narrowband antennas, is used defined as $\%B = \frac{f_H - f_L}{f_c} = 2\frac{f_H - f_L}{f_H + f_L}$. The theoretical limit to fractional bandwidth is 200%, which occurs for $f_L = 0$.

Fractional bandwidth, usually used for wideband antennas, is defined as $B = f_H / f_L$, and is typically presented in the form of B:1. Fractional bandwidth is used for wideband antennas because of the compression of the percent bandwidth that occurs mathematically with percent bandwidths above 100%, which corresponds to a fractional bandwidth of 3:1.

Photonics

In photonics, the term *bandwidth* occurs in a variety of meanings:

- the bandwidth of the output of some light source, e.g., an ASE source or a laser; the bandwidth of ultrashort optical pulses can be particularly large
- the width of the frequency range that can be transmitted by some element, e.g. an optical fiber
- the gain bandwidth of an optical amplifier
- the width of the range of some other phenomenon (e.g., a reflection, the phase matching of a nonlinear process, or some resonance)
- the maximum modulation frequency (or range of modulation frequencies) of an optical modulator
- the range of frequencies in which some measurement apparatus (e.g., a powermeter) can operate
- the data rate (e.g., in Gbit/s) achieved in an optical communication system.

A related concept is the spectral linewidth of the radiation emitted by excited atoms.

Chapter-3

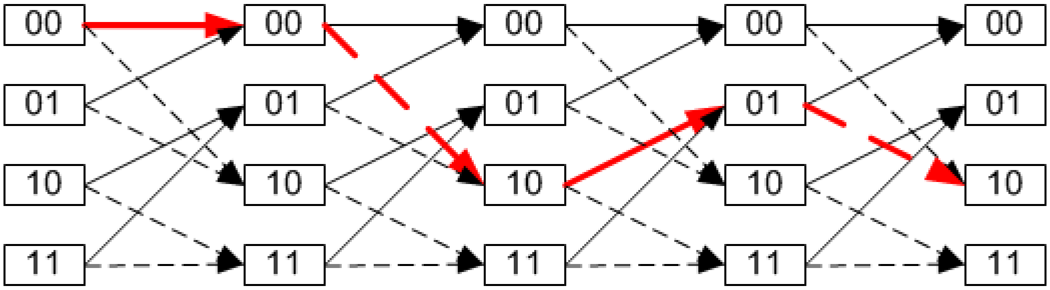
Trellis Modulation and Turbo Equalizer

Trellis modulation

In telecommunication, **trellis modulation** (also known as **trellis coded modulation**, or simply **TCM**) is a modulation scheme which allows highly efficient transmission of information over band-limited channels such as telephone lines. Trellis modulation was invented by Gottfried Ungerboeck working for IBM in the 1970s, and first described in a conference paper in 1976; but it went largely unnoticed until he published a new detailed exposition in 1982 which achieved sudden widespread recognition.

In the late 1980s, modems operating over plain old telephone service (*POTS*) typically achieved 9.6 kbit/s by employing 4 bits per symbol QAM modulation at 2,400 baud (symbols/second). This bit rate ceiling existed despite the best efforts of many researchers, and some engineers predicted that without a major upgrade of the public phone infrastructure, the maximum achievable rate for a POTS modem might be 14 kbit/s for two-way communication (3,429 baud × 4 bits/symbol, using QAM). However, 14 kbit/s is only 40% of the theoretical maximum bit rate predicted by Shannon's Theorem for POTS lines (approximately 35 kbit/s).

A new modulation method



Trellis diagram.

The name *trellis* was coined because a state diagram of the technique, when drawn on paper closely resembles the trellis lattice used in rose gardens. The scheme is basically a convolutional code of rates (r,r+1). Ungerboeck's unique contribution is to apply the parity check on a per symbol basis instead of the older technique of applying it to the bit stream then modulating the bits. The key idea he termed Mapping by Set Partitions. This idea was to group the symbols in a tree like fashion then separate them into two limbs of equal size. At each limb of the tree, the symbols were further apart. Although in multi-dimensions, it is hard to visualize, a simple one dimension example illustrates the basic procedure. Suppose the symbols are located at [1, 2, 3, 4, ...]. Then take all odd symbols and place them in one group, and the even symbols in the second group. This is not quite accurate because Ungerboeck was looking at the two dimensional problem, but the principle is the same, take every other one for each group and repeat the procedure for each tree limb. He next described a method of assigning the encoded bit stream onto the symbols in a very systematic procedure. Once this procedure was fully described, his next step was to program the algorithms into a computer and let the computer search for the best codes. The results were astonishing. Even the most simple code (4 state) produced error rates nearly 1,000 times lower than an equivalent uncoded system. For two years Ungerboeck kept these results private and only conveyed them to close colleagues. Finally, in 1982, Ungerboeck published a paper describing the principles of trellis modulation.

A flurry of research activity ensued, and by 1990 the International Telecommunication Union had published modem standards for the first trellis-modulated modem at 14.4 kbit/s (2,400 baud and 6 bits per symbol). Over the next several years further advances in encoding, plus a corresponding symbol rate increase from 2,400 to 3,429 baud, allowed modems to achieve rates up to 34.3 kbit/s (limited by maximum power regulations to 33.8 kbit/s). Today, the most common trellis-modulated V.34 modems use a 4-dimensional set partition which is achieved by treating two 2-dimensional symbols as a single lattice. This set uses 8, 16, or 32 state convolutional codes to squeeze the equivalent of 6 to 10 bits into each symbol sent by the modem (for example, 2,400 baud × 8 bits/symbol = 19,200 bit/s).

Once manufacturers introduced modems with trellis modulation, transmission rates increased to the point where interactive transfer of multimedia over the telephone became feasible (a 200 kilobyte image and a 5 megabyte song could be downloaded in less than 1 minute and 30 minutes, respectively). Sharing a floppy disk via a BBS could be done in just a few minutes, instead of an hour. Thus Ungerboeck's invention played a key role in the Information Age.

Turbo equalizer

In digital communications, a **turbo equalizer** is a type of receiver used to receive a message corrupted by a communication channel with intersymbol interference (ISI). It approaches the performance of a maximum a posteriori (MAP) receiver via iterative message passing between a soft-in soft-out (SISO) equalizer and a SISO decoder. It is closely related to turbo codes, as a turbo equalizer may be considered a turbo decoder if the channel is viewed as a convolutional code.

History

In 1993, turbo codes were introduced by Berrou, Glavieux, and Thitimajshima. In 1995, the turbo principle, which was developed for turbo codes, was applied to an equalizer by Douillard, Jézéquel, and Berrou. They formulated the ISI receiver problem as a turbo code decoding problem, where the channel is thought of as a rate 1 convolutional code and the error correction coding is the second code. In 1997, Glavieux, Laot, and Labat demonstrated that a linear equalizer could be used in a turbo equalizer framework. This discovery made turbo equalization computationally efficient enough to be applied to a wide range of applications.

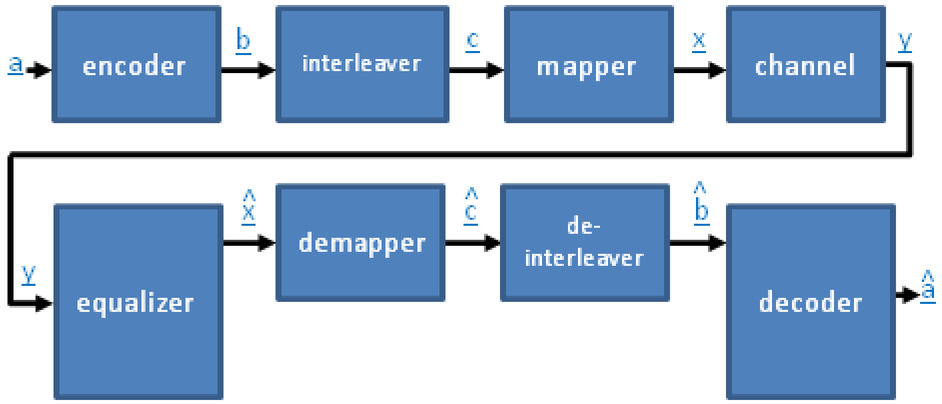
Overview

Standard Communication System Overview

Before discussing turbo equalizers, it is necessary to understand the basic receiver in the context of a communication system. At the transmitter, information bits a are encoded. Encoding adds redundancy by mapping the information bits *a* to a longer bit vector--the code bit vector *b*. The encoded bits *b* are then interleaved. Interleaving permutes the order of the code bits *b* resulting in bits *c*. The main reason for doing this is to insulate the information bits from bursty noise. Next, the symbol mapper maps the bits *c* into complex symbols *x*. These digital symbols are then converted into analog symbols with an D/A converter. Typically the signal is then up-converted to pass band frequencies by mixing it with a carrier signal. This is a necessary step for complex symbols. The signal is then ready to be transmitted through the channel.

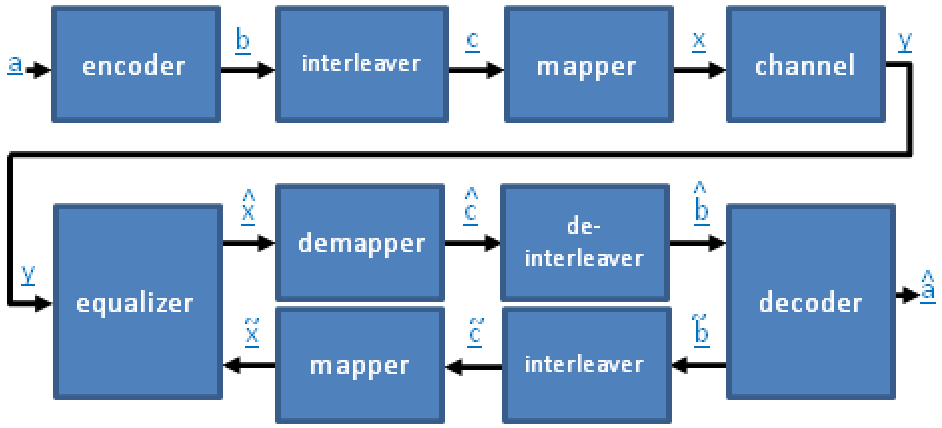
At the receiver, the operations performed by the transmitter are reversed to recover $\hat{a}$, an estimate of the information bits. The down-converter mixes the signal back down to baseband. The A/D converter then samples the analog signal, making it digital. At this point, *y* is recovered. The signal *y* is what would be received if *x* were transmitted through the digital baseband equivalent of the channel plus noise. The signal is then equalized. The equalizer attempts to unravel the ISI in the received signal to recover the transmitted symbols. It then outputs the bits $\hat{c}$associated with those symbols. The vector $\hat{c}$may represent hard decisions on the bits or soft decisions. If the equalizer makes soft decisions, it outputs information relating to the probability of the bit being a 0 or a 1. If the equalizer makes hard decisions on the bits, it quantizes the soft bit decisions and outputs either a 0 or a 1. Next, the signal is deinterleaved which is a simple permutation transformation that undoes the transformation the interleaver executed. Finally, the bits are decoded by the decoder. The decoder estimates $\hat{a}$from $\hat{b}$.

A diagram of the communication system is shown below. In this diagram, the channel is the equivalent baseband channel, meaning that it encompasses the A/D, the up converter, the channel, the down converter, and the D/A.



Turbo Equalizer Overview

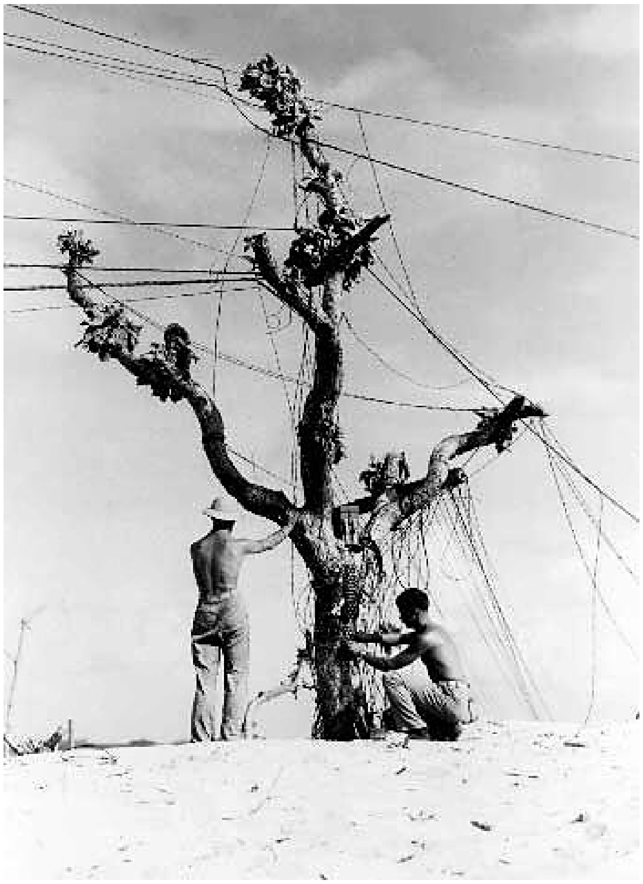
The block diagram of a communication system employing a turbo equalizer is shown below. The turbo equalizer encompasses the equalizer, the decoder, and the blocks in between.



The difference between a turbo equalizer and a standard equalizer is the feedback loop from the decoder to the equalizer. Due to the structure of the code, the decoder not only estimates the information bits *a*, but it also discovers new information about the coded bits *b*. The decoder is therefore able to output extrinsic information, $\tilde{b}$about the likelihood that a certain code bit stream was transmitted. Extrinsic information is new information that is not derived from information input to the block. This extrinsic information is then mapped back into information about the transmitted symbols *x* for use in the equalizer. These extrinsic symbol likelihoods, $\tilde{x}$, are fed into the equalizer as *a priori* symbol probabilities. The equalizer uses this *a priori* information as well as the input signal *y* to estimate extrinsic probability information about the transmitted symbols. The *a priori* information fed to the equalizer is initialized to 0, meaning that the initial estimate $\hat{a}$made by the turbo equalizer is identical to the estimate made by the standard receiver. The information $\hat{x}$is then mapped back into information about *b* for use by the decoder. The turbo equalizer repeats this iterative process until a stopping criterion is reached.

Chapter-4

Channel (communications)



Old telephone wires are a challenging communications channel for modern digital communications.

In telecommunications and computer networking, a **communication channel**, or **channel**, refers either to a physical transmission medium such as a wire, or to a logical connection over a multiplexed medium such as a radio channel. A channel is used to convey an information signal, for example a digital bit stream, from one or several *senders* (or transmitters) to one or several *receivers*. A channel has a certain capacity for transmitting information, often measured by its bandwidth in Hz or its data rate in bits per second

In information theory, a channel refers to a theoretical *channel model* with certain error characteristics. In this more general view, a storage device is also a kind of channel, which can be sent to (written) and received from (read).

Examples

A channel can take many forms. Examples of communications channels include:

1. A connection between initiating and terminating nodes of a circuit.
2. A single path provided by a transmission medium via either
    - physical separation, such as by multipair cable or
    - electrical separation, such as by frequency-division or time-division multiplexing.
3. A path for conveying electrical or electromagnetic signals, usually distinguished from other parallel paths.
    - A storage which can communicate a message over time as well as space
    - The portion of a storage medium, such as a track or a band, that is accessible to a given reading or writing station or head.
    - A buffer from which messages can be 'put' and 'got'.
4. In a communications system, the physical or logical link that connects a data source to a data sink.
5. A specific radio frequency, pair or band of frequencies, usually named with a letter, number, or codeword, and often allocated by international agreement. Examples:
    - Marine VHF radio uses some 88 channels in the VHF band for two-way FM voice communication. Channel 16, for example, is 156.800 MHz. In the US, seven additional channels, WX1 - WX7, are allocated for weather broadcasts.
    - Television channels such as North American TV Channel 2 = 55.25 MHz, Channel 13 = 211.25 MHz. Each channel is 6 MHz wide. Besides these "physical channels", television also has "virtual channels".
    - Wi-Fi consists of unlicensed channels 1-13 from 2412 MHz to 2484 MHz in 5 MHz steps.
    - The radio channel between an amateur radio repeater and a ham uses two bands often 600 kHz (0.6 MHz) apart. For example, a repeater that transmits on 146.94 MHz typically listens for a ham transmitting on 146.34 MHz.
6. A room in the Internet Relay Chat (IRC) network, in which participants can communicate with each other.

All of these communications channels share the property that they transfer information. The information is carried through the channel by a signal.

Channel models

A channel can be modelled physically by trying to calculate the physical processes which modify the transmitted signal. For example in wireless communications the channel can be modelled by calculating the reflection off every object in the environment. A sequence of random numbers might also be added in to simulate external interference and/or electronic noise in the receiver.

Statistically a communication channel is usually modelled as a triple consisting of an input alphabet, an output alphabet, and for each pair *(i, o)* of input and output elements a transition probability *p(i, o)*. Semantically, the transition probability is the probability that the symbol *o* is received given that *i* was transmitted over the channel.

Statistical and physical modelling can be combined. For example in wireless communications the channel is often modelled by a random attenuation (known as fading) of the transmitted signal, followed by additive noise. The attenuation term is a simplification of the underlying physical processes and captures the change in signal power over the course of the transmission. The noise in the model captures external interference and/or electronic noise in the receiver. If the attenuation term is complex it also describes the relative time a signal takes to get through the channel. The statistics of the random attenuation are decided by previous measurements or physical simulations.

Channel models may be continuous channel models in that there is no limit to how precisely their values may be defined.

Communication channels are also studied in a discrete-alphabet setting. This corresponds to abstracting a real world communication system in which the analog->digital and digital->analog blocks are out of the control of the designer. The mathematical model consists of a transition probability that specifies an output distribution for each possible sequence of channel inputs. In information theory, it is common to start with memoryless channels in which the output probability distribution only depends on the current channel input.

A channel model may either be digital (quantified, e.g. binary) or analog.

Digital channel models

In a digital channel model, the transmitted message is modelled as a digital signal at a certain protocol layer. Underlying protocol layers, such as the physical layer transmission technique, is replaced by a simplified model. The model may reflect channel performance measures such as bit rate, bit errors, latency/delay, delay jitter, etc. Examples of digital channel models are:

- Binary symmetric channel (BSC), a discrete memoryless channel with a certain bit error probability
- Binary bursty bit error channel model, a channel "with memory"
- Binary erasure channel (BEC), a discrete channel with a certain bit error detection (erasure) probability
- Packet erasure channel, where packets are lost with a certain packet loss probability or packet error rate
- Arbitrarily varying channel (AVC), where the behavior and state of the channel can change randomly

Analog channel models

In an analog channel model, the transmitted message is modelled as an analog signal. The model can be a linear or non-linear, time-continuous or time-discrete (sampled), memoryless or dynamic (resulting in burst errors), time-invariant or time-variant (also resulting in burst errors), baseband, passband (RF signal model), real-valued or complex-valued signal model. The model may reflect the following channel impairments:

- Noise model, for example
    - Additive white Gaussian noise (AWGN) channel, a linear continuous memoryless model
    - Phase noise model
- Interference model, for example cross-talk (co-channel interference) and intersymbol interference (ISI)
- Distortion model, for example a non-linear channel model causing intermodulation distortion (IMD)
- Frequency response model, including attenuation and phase-shift
- Group delay model
- Modelling of underlying physical layer transmission techniques, for example a complex-valued equivalent baseband model of modulation and frequency response
- Radio frequency propagation model, for example
    - Log-distance path loss model
    - Fading model, for example Rayleigh fading, Ricean fading, log-normal shadow fading and frequency selective (dispersive) fading
    - Doppler shift model, which combined with fading results in a time-variant system
    - Ray tracing models, which attempt to model the signal propagation and distortions for specified transmitter-receiver geometries, terrain types, and antennas
    - Mobility models, which also causes a time-variant system

Types of communications channels

- Digital (discrete) or analog (continuous) channel
- Baseband and passband channel

- Transmission medium, for example a fibre channel
- Multiplexed channel
- Computer network virtual channel
- Simplex communication, duplex communication or half duplex communication channel
- Return channel
- Uplink or downlink (upstream or downstream channel)
- Broadcast channel, unicast channel or multicast channel

Multi-terminal channels, with application to cellular systems

In networks, as opposed to point-to-point communication, the communication media is shared between multiple nodes (terminals). Depending on the type of communication, different terminals can cooperate or interfere on each other. In general, any complex multi-terminal network can be considered as a combination of simplified multi-terminal channels. The following channels are the principal multi-terminal channels which was first introduced in the field of information theory:

- A point-to-multipoint channel, also known as broadcasting medium (not to be confused with broadcasting channel): In this channel, a single sender transmits multiple messages to different destination nodes. All wireless channels except radio links can be considered as broadcasting media, but may not always provide broadcasting service. The downlink of a cellular system can be considered as a point-to-multipoint channel, if only one cell is considered and inter-cell co-channel interference is neglected. However, the communication service of a phone call is unicasting.
- Multiple access channel: In this channel, multiple senders transmit multiple possible different messages over a shared physical medium to one or several destination nodes. This requires a channel access scheme, including a media access control (MAC) protocol combiend with a multiplexing scheme. This channel model has applications in the uplink of the cellular networks.
- Relay channel: In this channel, one or several intermediate nodes (called relay, repeater or gap filler nodes) cooperate with a sender to send the message to an ultimate destination node. Relay nodes are considered as a possible add-on in the upcoming cellular standards like 3GPP Long Term Evolution (LTE).
- Interference channel: In this channel, two different senders transmit their data to different destination nodes. Hence, the different senders can have a possible cross-talk or co-channel interference on the signal of each other. The inter-cell interference in the cellular wireless communications is an example of the interference channel. In spread spectrum systems like 3G, interference also occur inside the cell if non-orthogonal codes are used.
- A unicasting channel is a channel that provides a unicasting service, i.e. that sends data addressed to one specific user. An established phone call is an example.
- A broadcasting channel is a channel that provides a broadcasting service, i.e. that sends data addressed to all users in the network. Cellular network examples are the paging service as well as the Multimedia Broadcast Multicast Service.

- A multicasting channel is a channel where data is addressed to a group of subscribing users. LTE exampels are the Physical Multicast Channel (PMCH) and MBSFN (Multicast Broadcast Single Frequency Network).

From the above 4 basic multi-terminal channels, multiple access channel is the only one whose capacity region is known. Even for the special case of the Gaussian scenario, the capacity region of the other 3 channels except the broadcast channel is unknown in general.

Chapter-5

Detection Theory

Detection theory, or **signal detection theory**, is a means to quantify the ability to discern between signal and noise. According to the theory, there are a number of determiners of how a detecting system will detect a signal, and where its threshold levels will be. The theory can explain how changing the threshold will affect the ability to discern, often exposing how adapted the system is to the task, purpose or goal at which it is aimed.

When the detecting system is a human being, experience, expectations, physiological state (e.g. fatigue) and other factors can affect the threshold applied. For instance, a sentry in wartime will likely detect fainter stimuli than the same sentry in peacetime.

Much of the early work in detection theory was done by radar researchers. The psychological theory was first published by Wilson P. Tanner, David M. Green, and John A. Swets in 1954. Detection theory was used in 1966 by John A. Swets and David M. Green for psychophysics. Green and Swets criticized the traditional methods of psychophysics for their inability to discriminate between the real sensitivity of subjects and their (potential) response biases.

Detection theory has applications in many fields such as diagnostics of any kind, quality control, telecommunications, and psychology. The concept is similar to the signal to noise ratio used in the sciences and confusion matrices used in artificial intelligence. It is also usable in alarm management, where it is important to separate important events from background noise.

Psychology

Signal detection theory (SDT) is used when psychologists want to measure the way we make decisions under conditions of uncertainty, such as how we would perceive distances in foggy conditions. SDT assumes that the decision maker is not a passive receiver of information, but an active decision-maker who makes difficult perceptual judgements under conditions of uncertainty. In foggy circumstances, we are forced to decide how far away from us an object is, based solely upon visual stimulus which is impaired by the fog. Since the brightness of the object, such as a traffic light, is used by the brain to discriminate the distance of an object, and the fog reduces the brightness of objects, we perceive the object to be much farther away than it actually is.

To apply signal detection theory to a data set where stimuli were either present or absent, and the observer categorized each trial as having the stimulus present or absent, the trials are sorted into one of four categories:

	Respond "Absent"	**Respond "Present"**
Stimulus Present	Miss	Hit
Stimulus Absent	Correct Rejection	False Alarm

Based on the proportions of these types of trials, numerical estimates of sensitivity can be obtained with statistics like the sensitivity index d' and A', and response bias can be estimated with statistics like β.

Signal detection theory can also be applied to memory experiments, where items are presented on a study list for later testing. A test list is created by combining these 'old' items with novel, 'new' items that did not apear on the study list. On each test trial the subject will respond 'yes, this was on the study list' or 'no, this was not on the study list'. Items presented on the study list are called Targets, and new items are called Distractors. Saying 'Yes' to a target constitutes a Hit, while saying 'Yes' to a distractor constitutes a False Alarm.

	Respond "No"	**Respond "Yes"**
Target	Miss	Hit
Distractor	Correct Rejection	False Alarm

Applications

Signal Detection Theory has wide application, both in humans and other animals. Topics include memory, stimulus characterists of schedules of reinforcement, etc.

Sensitivity or discriminability

Conceptually, sensitivity refers to how hard or easy it is to detect that a target stimulus is present from background events. For example, in a recognition memory paradigm, having longer to study to-be-remembered words makes it easier to recognize previously seen or heard words. In contrast, having to remember 30 words rather than 5 makes the discrimination harder. One of the most commonly used statistics for computing sensitivity is the so-called sensitivity index, or *d'*. There are also non-parametric measures.

Bias

Bias is the extent to which one response is more probable than another. That is, a receiver may be more likely to respond that a stimulus is present or more likely to respond that a stimulus is not present. Bias is independent of sensitivity. For example, if there is a penalty for either false alarms or misses, this may influence bias. If the stimulus is a bomber, then a miss (failing to detect the plane) may increase deaths, so a liberal bias is likely. In contrast, crying wolf (a false alarm) too often may make people less likely to respond, grounds for a conservative bias.

Mathematics

P(H1|y) > P(H2|y) / MAP Testing

In the case of making a decision between two hypotheses, *H1*, absent, and *H2*, present, in the event of a particular observation, *y*, a classical approach is to choose *H1* when *p(H1|y) > p(H2|y)* and *H2* in the reverse case. In the event that the two *a posteriori* probabilities are equal, one typically defaults to a single choice, say *H2*. One could also flip a coin although the expected number of errors would be the same.

When taking this approach, usually what one knows are the conditional probabilities, *p(y|H1)* and *p(y|H2)*, and the *a priori* probabilities $p(H1) = \pi_1$ and $p(H2) = \pi_2$. In this case,

$$p(H1|y) = \frac{p(y|H1) \cdot \pi_1}{p(y)},$$

$$p(H2|y) = \frac{p(y|H2) \cdot \pi_2}{p(y)}$$

where *p(y)* is the total probability of event *y*,

$$p(y|H1) \cdot \pi_1 + p(y|H2) \cdot \pi_2.$$

H2 is chosen in case

$$\frac{p(y|H2) \cdot \pi_2}{p(y|H1) \cdot \pi_1 + p(y|H2) \cdot \pi_2} \geq \frac{p(y|H1) \cdot \pi_1}{p(y|H1) \cdot \pi_1 + p(y|H2) \cdot \pi_2}$$

$$\Rightarrow \frac{p(y|H2)}{p(y|H1)} \geq \frac{\pi_1}{\pi_2}$$

and *H1* otherwise.

Often, the ratio $\frac{\pi_2}{\pi_1}$ is called τ_{MAP} and $\frac{p(y|H2)}{p(y|H1)}$ is called $L(y)$, the *likelihood ratio*.

Using this terminology, *H2* is chosen in case $L(y) \geq \tau_{MAP}$. This is called MAP testing, where MAP stands for "maximum *a posteriori*").

Taking this approach minimizes the expected number of errors one will make.

Bayes Criterion

In some cases, it is far more important to respond appropriately to *H1* than it is to respond appropriately to *H2*. For example, if one is trying to detect an incoming bomber known to be carrying a nuclear weapon, it is much more important to shoot down the bomber if it is there, than it is not to send a fighter squadron to inspect a false alarm (assuming a large supply of fighter squadrons). The Bayes criterion is an approach suitable for such cases.

Here a utility is associated with each of four situations:

- U_{11}: One responds with behavior appropriate to H1 and H1 is true: fighters destroy bomber, incurring fuel, maintenance, and weapons costs, take risk of some being shot down;
- U_{12}: One responds with behavior appropriate to H1 and H2 is true: fighters sent out, incurring fuel and maintenance costs, bomber location remains unknown;
- U_{21}: One responds with behavior appropriate to H2 and H1 is true: city destroyed;
- U_{22}: One responds with behavior appropriate to H2 and H2 is true: fighters stay home, bomber location remains unknown;

As is shown below, what is important are the differences, $U_{11} - U_{21}$ and $U_{22} - U_{12}$.

Similarly, there are four probabilities, P_{11}, P_{12}, etc., for each of the cases (which are dependent on one's decision strategy).

The Bayes criterion approach is to maximize the expected utility:

$$U = P_{11} \cdot U_{11} + P_{21} \cdot U_{21} + P_{12} \cdot U_{12} + P_{22} \cdot U_{22}$$

$$U = P_{11} \cdot U_{11} + (1 - P_{11}) \cdot U_{21} + P_{12} \cdot U_{21} + (1 - P_{12}) \cdot U_{22}$$

$$U = U_{12} + U_{21} + P_{11} \cdot (U_{11} - U_{21}) - P_{12} \cdot (U_{22} - U_{12})$$

Effectively, one may maximize the sum,

$$U' = P_{11} \cdot (U_{11} - U_{21}) - P_{12} \cdot (U_{22} - U_{12}),$$

and make the following substitutions:

$$P_{11} = \pi_1 \cdot \int_{R_1} p(y|H1)\, dy$$

$$P_{12} = \pi_2 \cdot \int_{R_1} p(y|H2)\, dy$$

where π_1 and π_2 are the *a priori* probabilities, $P(H1)$ and $P(H2)$, and R_1 is the region of observation events, y, that are responded to as though *H1* is true.

$$\Rightarrow U' = \int_{R_1} \left\{ \pi_1 \cdot (U_{11} - U_{21}) \cdot p(y|H1) - \pi_2 \cdot (U_{22} - U_{12}) \cdot p(y|H2) \right\} dy$$

U' and thus *U* are maximized by extending R_1 over the region where

$$\pi_1 \cdot (U_{11} - U_{21}) \cdot p(y|H1) - \pi_2 \cdot (U_{22} - U_{12}) \cdot p(y|H2) > 0$$

This is accomplished by deciding H2 in case

$$\pi_2 \cdot (U_{22} - U_{12}) \cdot p(y|H2) \geq \pi_1 \cdot (U_{11} - U_{21}) \cdot p(y|H1)$$

$$\Rightarrow L(y) \equiv \frac{p(y|H2)}{p(y|H1)} \geq \frac{\pi_1 \cdot (U_{11} - U_{21})}{\pi_2 \cdot (U_{22} - U_{12})} \equiv \tau_B$$

and H1 otherwise, where *L(y)* is the so-defined *likelihood ratio*.

Chapter-6

Filter (Signal Processing)

In signal processing, a **filter** is a device or process that removes from a signal some unwanted component or feature. Filtering is a class of signal processing, the defining feature of filters being the complete or partial suppression of some aspect of the signal. Most often, this means removing some frequencies and not others in order to suppress interfering signals and reduce background noise. However, filters do not exclusively act in the frequency domain; especially in the field of image processing many other targets for filtering exist.

There are many different bases of classifying filters and these overlap in many different ways; there is no simple hierarchical classification. Filters may be:

- analog or digital
- discrete-time (sampled) or continuous-time
- linear or non-linear
- time-invariant or time-variant, also known as shift invariance. If the filter operates in a spatial domain then the the characterization is space invariance.
- passive or active type of continuous-time filter
- infinite impulse response (IIR) or finite impulse response (FIR) type of discrete-time or digital filter.

Linear continuous-time filters

Linear continuous-time circuit is perhaps the most common meaning for filter in the signal processing world, and simply "filter" is often taken to be synonymous. These are filters that are designed to remove certain frequencies and allow others to pass. Such a filter is, of necessity, a linear filter. Any non-linearity will result in the output signal containing components of frequency which were not present in the input signal.

The modern design methodology for linear continuous-time filters is called network synthesis. Some important filter families designed in this way are:

- Chebyshev filter, has the best approximation to the ideal response of any filter for a specified order and ripple.
- Butterworth filter, has a maximally flat frequency response.
- Bessel filter, has a maximally flat phase delay.
- Elliptic filter, has the steepest cutoff of any filter for a specified order and ripple.

The difference between these filter families is that they all use a different polynomial function to approximate to the ideal filter response. This results in each having a different transfer function.

Another older, less-used methodology is the image parameter method. Filters designed by this methodology are archaically called "wave filters". Some important filters designed by this method are:

- Constant k filter, the original and simplest form of wave filter.
- m-derived filter, a modification of the constant k with improved cutoff steepness and impedance matching.

Terminology

Some terms used to describe and classify linear filters:

- The frequency response can be classified into a number of different bandforms describing which frequencies the filter passes (the passband) and which it rejects (the stopband):
    - Low-pass filter – low frequencies are passed, high frequencies are attenuated.
    - High-pass filter – high frequencies are passed, low frequencies are attenuated.
    - Band-pass filter – only frequencies in a frequency band are passed.
    - Band-stop filter or band-reject filter – only frequencies in a frequency band are attenuated.
    - Notch filter – rejects just one specific frequency - an extreme band-stop filter.
    - Comb filter – has multiple regularly spaced narrow passbands giving the bandform the appearance of a comb.
    - All-pass filter – all frequencies are passed, but the phase of the output is modified.

- Cutoff frequency is the frequency beyond which the filter will not pass signals. It is usually measured at a specific attenuation such as 3dB.
- Roll-off is the rate at which attenuation increases beyond the cut-off frequency.
- Transition band, the (usually narrow) band of frequencies between a passband and stopband.
- Ripple is the variation of the filters insertion loss in the passband.

- The order of a filter is the degree of the approximating polynomial and in passive filters corresponds to the number of elements required to build it. Increasing order increases roll-off and brings the filter closer to the ideal response.

Technologies

Filters can be built in a number of different technologies. The same transfer function can be realised in several different ways, that is the mathematical properties of the filter are the same but the physical properties are quite different. Often the components in different technologies are directly analogous to each other and fulfill the same role in their respective filters. For instance, the resistors, inductors and capacitors of electronics correspond respectively to dampers, masses and springs in mechanics. Likewise, there are corresponding components in distributed element filters.

- Electronic filters were originally entirely passive consisting of resistance, inductance and capacitance. Active technology makes design easier and opens up new possibilities in filter specifications.
- Digital filters operate on signals represented in digital form. The essence of a digital filter is that it directly implements a mathematical algorithm, corresponding to the desired filter transfer function, in its programming or microcode.
- Mechanical filters are built out of mechanical components. In the vast majority of cases they are used to process an electronic signal and transducers are provided to convert this to and from a mechanical vibration. However, examples do exist of filters that have been designed for operation entirely in the mechanical domain.
- Distributed element filters are constructed out of components made from small pieces of transmission line or other distributed elements. There are structures in distributed element filters that directly correspond to the lumped elements of electronic filters, and others that are unique to this class of technology.
- Waveguide filters consist of waveguide components or components inserted in the waveguide. Waveguides are a class of transmission line and many structures of distributed element filters, for instance the stub (electronics), can be implemented in waveguides also.
- Acoustic filters
- Optical filters were originally developed for purposes other than signal processing such as lighting and photography. With the rise of optical fiber technology, however, optical filters increasingly find signal processing applications and signal processing filter terminology, such as longpass and shortpass, are entering the field.

The transfer function

The transfer function $H(s)$of a filter is the ratio of the output signal $Y(s)$to that of the input signal $X(s)$as a function of the complex frequency s:

$$H(s) = \frac{Y(s)}{X(s)}$$

with $s = \sigma + j\omega$.

The transfer function of all linear time-invariant filters generally share certain characteristics:

- For filters which are constructed of discrete components, their transfer function must be the ratio of two polynomials in s, i.e. a rational function of s. The order of the transfer function will be the highest power of $s$encountered in either the numerator or the denominator.
- The polynomials of the transfer function will all have real coefficients. Therefore, the poles and zeroes of the transfer function will either be real or occur in complex conjugate pairs.
- Since the filters are assumed to be stable, the real part of all poles (i.e. zeroes of the denominator) will be negative, i.e. they will lie in the left half-plane in complex frequency space.

Distributed element filters do not, in general, produce rational functions but can often approximate to them.

The proper construction of a transfer function involves the Laplace transform, and therefore it is needed to assume null initial conditions, because

$$\mathcal{L}\left\{\frac{df}{dt}\right\} = s \cdot \mathcal{L}\left\{f(t)\right\} - f(0),$$

And when f(0)=0 we can get rid of the constants and use the usual expression

$$\mathcal{L}\left\{\frac{df}{dt}\right\} = s \cdot \mathcal{L}\left\{f(t)\right\}$$

An alternative to transfer functions is to give the behavior of the filter as a convolution. The convolution theorem, which holds for Laplace transforms, guarantees equivalence with transfer functions.

Classification

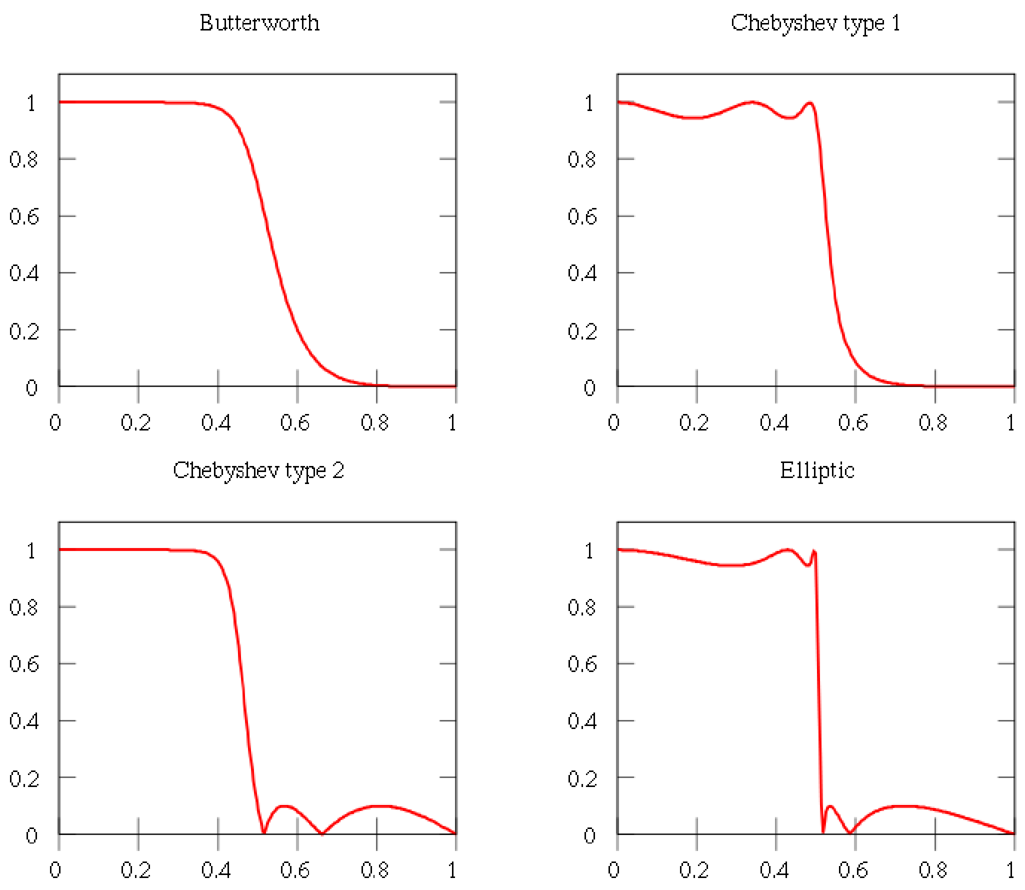
Filters may be specified by family and bandform. A filter's family is specified by the approximating polynomial used and each leads to certain characteristics of the transfer function of the filter. Some common filter families and their particular characteristics are:

- Butterworth filter – no gain ripple in pass band and stop band, slow cutoff

- Chebyshev filter (Type I) – no gain ripple in stop band, moderate cutoff
- Chebyshev filter (Type II) – no gain ripple in pass band, moderate cutoff
- Bessel filter – no group delay ripple, no gain ripple in both bands, slow gain cutoff
- Elliptic filter – gain ripple in pass and stop band, fast cutoff
- Optimum "L" filter
- Gaussian filter – no ripple in response to step function
- Hourglass filter
- Raised-cosine filter

Each family of filters can be specified to a particular order. The higher the order, the more the filter will approach the "ideal" filter; but also the longer the impulse response is and the longer the latency will be. An ideal filter has full transmission in the pass band, complete attenuation in the stop band, and an abrupt transition between the two bands, but this filter has infinite order (i.e., the response cannot be expressed as a linear differential equation with a finite sum) and infinite latency (i.e., its compact support in the Fourier transform forces its time response to be ever lasting).

Here is an image comparing Butterworth, Chebyshev, and elliptic filters. The filters in this illustration are all fifth-order low-pass filters. The particular implementation – analog or digital, passive or active – makes no difference; their output would be the same.



As is clear from the image, elliptic filters are sharper than all the others, but they show ripples on the whole bandwidth.

Any family can be used to implement a particular bandform of which frequencies are transmitted, and which, outside the passband, are more or less attenuated. The transfer function completely specifies the behavior of a linear filter, but not the particular technology used to implement it. In other words, there are a number of different ways of achieving a particular transfer function when designing a circuit. A particular bandform of filter can be obtained by transformation of a prototype filter of that family.

Impedance matching

Impedance matching structures invariably take on the form of a filter, that is, a network of non-dissipative elements. For instance, in a passive electronics implementation, this would likely take the form of a ladder topology of inductors and capacitors. The design of matching networks shares much in common with filters and the design invariably will have a filtering action as an incindental consequence. Although the prime purpose of a matching network is not to filter, it is often the case that both functions are combined in the same circuit. The need for impedance matching does not arise while signals are in the digital domain.

Some filters for specific purposes

- Audio filter
- Line filter
- Texture filtering

Filters for removing noise from data

- Wiener filter
- Kalman filter
- Savitzky–Golay smoothing filter

Chapter-7

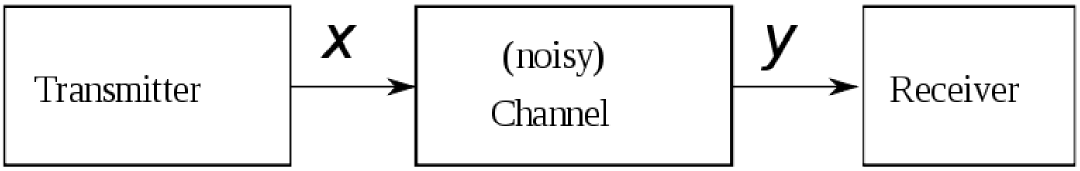
Channel Capacity and Frequency Mixer

Channel capacity

In electrical engineering, computer science and information theory, **channel capacity** is the tightest upper bound on the amount of information that can be reliably transmitted over a communications channel. By the noisy-channel coding theorem, the channel capacity of a given channel is the limiting information rate (in units of information per unit time) that can be achieved with arbitrarily small error probability.

Information theory, developed by Claude E. Shannon during World War II, defines the notion of channel capacity and provides a mathematical model by which one can compute it. The key result states that the capacity of the channel, as defined above, is given by the maximum of the mutual information between the input and output of the channel, where the maximization is with respect to the input distribution.

Formal definition



Let *X* represent the space of signals that can be transmitted, and *Y* the space of signals received, during a block of time over the channel. Let

$$p_{Y|X}(y|x)$$

be the conditional distribution function of *Y* given *X*. Treating the channel as a known statistic system, $p_{Y|X}(y \mid x)$ is an inherent fixed property of the communications channel (representing the nature of the noise in it). Then the joint distribution

$$p_{X,Y}(x,y)$$

of *X* and *Y* is completely determined by the channel and by the choice of

$$p_X(x) = \int_y p_{X,Y}(x,y)\,dy$$

the marginal distribution of signals we choose to send over the channel. The joint distribution can be recovered by using the identity

$$p_{X,Y}(x,y) = p_{Y|X}(y|x)\,p_X(x)$$

Under these constraints, next maximize the amount of information, or the message, that one can communicate over the channel. The appropriate measure for this is the mutual information *I*(*X*;*Y*), and this maximum mutual information is called the **channel capacity** and is given by

$$C = \sup_{p_X} I(X;Y)$$

Noisy-channel coding theorem

The noisy-channel coding theorem states that for any ε > 0 and for any rate *R* less than the channel capacity *C*, there is an encoding and decoding scheme that can be used to ensure that the probability of block error is less than ε for a sufficiently long code. Also, for any rate greater than the channel capacity, the probability of block error at the receiver goes to one as the block length goes to infinity.

Example application

An application of the channel capacity concept to an additive white Gaussian noise (AWGN) channel with *B* Hz bandwidth and signal-to-noise ratio *S/N* is the Shannon–Hartley theorem:

$$C = B\log\left(1+\frac{S}{N}\right)$$

C is measured in bits per second if the logarithm is taken in base 2, or nats per second if the natural logarithm is used, assuming *B* is in hertz; the signal and noise powers *S* and *N* are measured in watts or volts^2, so the signal-to-noise ratio here is expressed as a power ratio, *not* in decibels (dB); since figures are often cited in dB, a conversion may be needed. For example, 30 dB is a power ratio of $10^{30/10} = 10^3 = 1000$.

AWGN channel

If the average received power is $\bar{P}$[W] and the noise power spectral density is N_0 [W/Hz], the AWGN channel capacity is

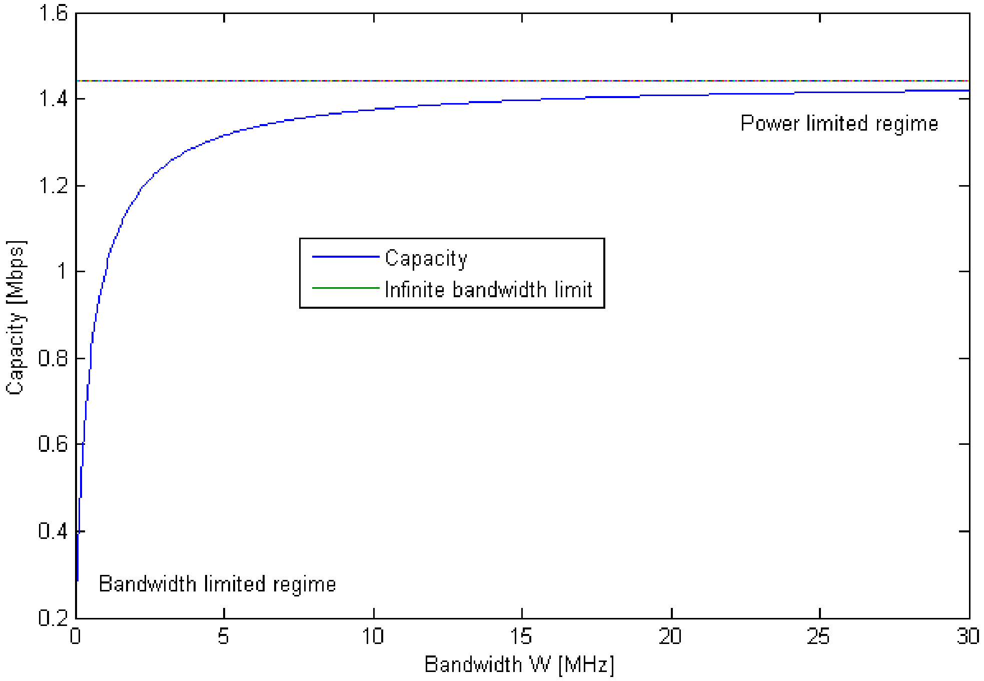
$$C_{awgn} = W \log_2 \left(1 + \frac{\bar{P}}{N_0 W}\right)$$ [bits/Hz],

where $\frac{\bar{P}}{N_0 W}$ is the received signal-to-noise ratio (SNR).

When the SNR is large (SNR >> 0 dB), the capacity $C \approx W \log_2 \frac{\bar{P}}{N_0 W}$ is logarithmic in power and approximately linear in bandwidth. This is called the *bandwidth-limited regime*.

When the SNR is small (SNR << 0 dB), the capacity $C \approx \frac{\bar{P}}{N_0} \log_2 e$ is linear in power but insensitive to bandwidth. This is called the *power-limited regime*.

The bandwidth-limited regime and power-limited regime are illustrated in the figure.



AWGN channel capacity with the power-limited regime and bandwidth-limited regime indicated. Here, $\frac{\bar{P}}{N_o} = 10^6$.

Frequency-selective channel

The capacity of the frequency-selective channel is given by so-called waterfilling power allocation,

$$C_{N_c} = \sum_{n=0}^{N_c-1} \log_2 \left(1 + \frac{P_n^* |\bar{h}_n|^2}{N_0} \right),$$

where $P_n^* = \max \left(\left(\frac{1}{\lambda} - \frac{N_0}{|\bar{h}_n|^2} \right), 0 \right)$ and $|\bar{h}_n|^2$ is the gain of subchannel *n*, with λ chosen to meet the power constraint.

Slow-fading channel

In a slow-fading channel, where the coherence time is greater than the latency requirement, there is no definite capacity as the maximum rate of reliable communications supported by the channel, $\log_2(1 + | h | ^2 SNR)$, depends on the random channel gain $| h | ^2$. If the transmitter encodes data at rate *R* [bits/s/Hz], there is a certain probability that the decoding error probability cannot be made arbitrarily small,

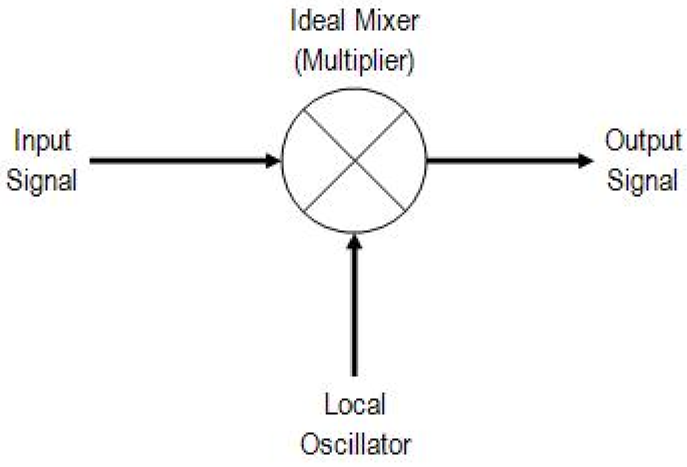
$$p_{out} = \mathbb{P}(\log(1 + |h|^2 SNR) < R),$$

in which case the system is said to be in outage. With a non-zero probability that the channel is in deep fade, the capacity of the slow-fading channel in strict sense is zero. However, it is possible to determine the largest value of *R* such that the outage probability p_{out} is less than ε. This value is known as the ε-outage capacity.

Fast-fading channel

In a fast-fading channel, where the latency requirement is greater than the coherence time and the codeword length spans many coherence periods, one can average over many independent channel fades by coding over a large number of coherence time intervals. Thus, it is possible to achieve a reliable rate of communication of $\mathbb{E}(\log_2(1 + |h|^2 SNR))$[bits/s/Hz] and it is meaningful to speak of this value as the capacity of the fast-fading channel.

Frequency mixer



Frequency Mixer Symbol.

In electronics a **mixer** or **frequency mixer** is a nonlinear electrical circuit that creates new frequencies from two signals applied to it. In its most common application, two signals at frequencies f_1 and f_2 are applied to a mixer, and it produces new signals at the sum $f_1 + f_2$ and difference $f_1 - f_2$ of the original frequencies. Other frequency components may also be produced in a practical frequency mixer.

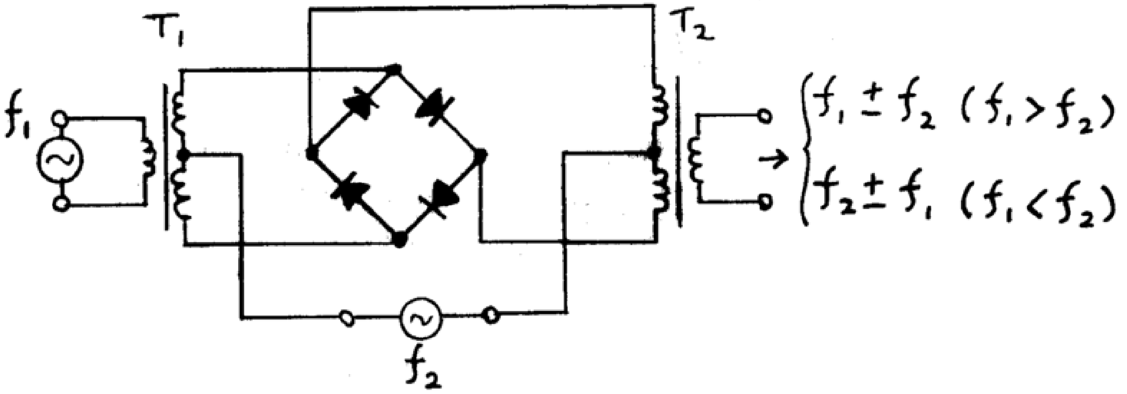
Mixers are widely used to shift signals from one frequency range to another, a process known as heterodyning, for convenience in transmission or further signal processing. For example, a key component of a superheterodyne receiver is a mixer used to move received signals to a common intermediate frequency. Frequency mixers are also used to modulate a carrier frequency in radio transmitters.

Types

Passive mixers use one or more diodes and rely on the non-linear relation between voltage and current to provide the multiplying element. In a passive mixer, the desired output signal is always of lower power than the input signals. Active mixers can increase the strength of the product signal. Active mixers improve isolation between the ports, but may have higher noise and more power consumption; an active mixer can be less tolerant of overload.

Mixers may be built of discrete components, may be part of integrated circuits, or can be delivered as hybrid modules.

Mixers may also be classified by their topology. Unbalanced mixers allow some of the input signal power to pass through to the output. A single-balanced mixer is arranged so that the local oscillator, (or RF) signal port, cancels and cannot pass through to the output. A doubly-balanced mixer has symmetrical paths for both inputs, and will have no output if either input signal is not present.



Schematic diagram of a double-balanced passive diode mixer. There is no output unless both f1 and f2 inputs are present.

Selection of a mixer type is a trade off for a particular application. Mixer circuits are characterized by conversion gain, and noise figure. Balanced and double-balanced designs allow less of the input signals to feed through to the output.

Nonlinear electronic components that are used as mixers include diodes, transistors biased near cutoff, and at lower frequencies, analog multipliers. Ferromagnetic-core inductors driven into saturation have also been used. In nonlinear optics, crystals with nonlinear characteristics are used to mix two frequencies of laser light to create optical heterodynes.

Diode

A diode can be used to create a simple unbalanced mixer. This type of mixer produces the original frequencies as well as their sum and their difference. The importance of the diode is that it is non-linear (or non-Ohmic), which means its response (current) is not proportional to its input (voltage). The diode therefore does not reproduce the frequencies of its driving voltage in the current through it, which allows the desired frequency manipulation. Certain other non-linear devices such as tunnel diodes or Gunn diodes can be utilized similarly.

The current *I* through an ideal diode as a function of the voltage *V* across it is given by

$$I = I_S \left(e^{\frac{qV_D}{nkT}} - 1 \right)$$

where what is important is that *V* appears in *e*'s exponent. The exponential can be expanded as

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

and can be approximated for small *x* (that is, small voltages) by the first few terms of that series:

$$e^x - 1 \approx x + \frac{x^2}{2}$$

Suppose that the sum of the two input signals $v_1 + v_2$ is applied to a diode, and that an output voltage is generated that is proportional to the current through the diode (perhaps by providing the voltage that is present across a resistor in series with the diode). Then, disregarding the constants in the diode equation, the output voltage will have the form

$$v_o = (v_1 + v_2) + \frac{1}{2}(v_1 + v_2)^2 + \ldots$$

The first term on the right is the original two signals, as expected, followed by the square of the sum, which can be rewritten as $(v_1 + v_2)^2 = v_1^2 + 2v_1v_2 + v_2^2$, where the multiplied signal is obvious. The ellipsis represents all the higher powers of the sum which we assume to be negligible for small signals.

Switching

Another form of mixer operates by switching, with the smaller input signal being passed inverted or uninverted according to the phase of the local oscillator (LO). This would be typical of the normal operating mode of a packaged double balanced mixer module such as an SBL-1, with the local oscillator drive considerably higher than the signal amplitude.

The aim of a switching mixer is to achieve linear operation over the signal level, and hard switching driven by the local oscillator. Mathematically the switching mixer is not much different from a multiplying mixer, just because instead of the LO sine wave term we would use the signum function. In the frequency domain the switching mixer operation leads to the usual sum and difference frequencies, but also to further terms e.g. +-3*fLO, +-5*fLO, etc. The advantage of a switching mixer is that it can achieve - with the same effort - a lower noise figure (NF) and larger conversion gain. This come because the switching diodes or transistors act either like a low resistor (switch closed) or large resistor (switch open) and in both cases only minimum noise is added. From the circuit perspective many multiplying mixers can be used as switching mixers, just by increasing the LO amplitude. So RF engineers simply talk about mixers, and mean switching mixers.

Applications

The mixer circuit can be used not only to shift the frequency of an input signal as in a receiver, but also as a product detector, modulator, phase detector or frequency multiplier. For example a communications receiver might contain two mixer stages for conversion of the input signal to an intermediate frequency, and another mixer employed as a detector for demodulation of the signal.

Chapter-8

Hilbert–Huang Transform

The **Hilbert–Huang transform** (**HHT**) is a way to decompose a signal into so-called intrinsic mode functions (IMF), and obtain instantaneous frequency data. It is designed to work well for data that are nonstationary and nonlinear. In contrast to other common transforms like the Fourier transform, the HHT is more like an algorithm (an empirical approach) that can be applied to a data set, rather than a theoretical tool.

Introduction

The **Hilbert–Huang transform** (**HHT**), a NASA designated name, was proposed by Huang et al. (1996, 1998, 1999, 2003). It is the result of the empirical mode decomposition (EMD) and the Hilbert spectral analysis (HSA). The HHT uses the EMD method to decompose a signal into so-called intrinsic mode function, and uses the HSA method to obtain instantaneous frequency data. The HHT provides a new method of analyzing nonstationary and nonlinear time series data.

Introduction to EMD and IMF

The fundamental part of the HHT is the **empirical mode decomposition** (**EMD**) method. Using the EMD method, any complicated data set can be decomposed into a finite and often small number of components, which is a collection of **intrinsic mode functions** (**IMF**). An IMF represents a generally simple oscillatory mode as a counterpart to the simple harmonic function. By definition, an IMF is any function with the same number of extrema and zero crossings, with its envelopes being symmetric with respect to zero. The definition of an IMF guarantees a well-behaved Hilbert transform of the IMF. This decomposition method operating in the time domain is adaptive and highly efficient. Since the decomposition is based on the local characteristic time scale of the data, it can be applied to nonlinear and nonstationary processes.

Introduction to HSA

The Hilbert spectral analysis (HSA) provides a method for examining the IMF's instantaneous frequency data as functions of time that give sharp identifications of embedded structures. The final presentation of the results is an energy-frequency-time distribution, designated as the Hilbert spectrum.

Techniques

The empirical mode decomposition (EMD)

The EMD method is a necessary step to reduce any given data into a collection of intrinsic mode functions (IMF) to which the Hilbert spectral analysis can be applied. An IMF is defined as a function that satisfies the following requirements:

1. In the whole data set, the number of extrema and the number of zero-crossings must either be equal or differ at most by one.
2. At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

Therefore, an IMF represents a simple oscillatory mode as a counterpart to the simple harmonic function, but it is much more general: instead of constant amplitude and frequency in a simple harmonic component, an IMF can have variable amplitude and frequency along the time axis.
The procedure of extracting an IMF is called sifting. The sifting process is as follows:

1. Identify all the local extrema in the test data.
2. Connect all the local maxima by a cubic spline line as the upper envelope.
3. Repeat the procedure for the local minima to produce the lower envelope.

The upper and lower envelopes should cover all the data between them. Their mean is m_1. The difference between the data and m_1 is the first component h_1:

$$X(t) - m_1 = h_1.$$

Ideally, h_1 should satisfy the definition of an IMF, for the construction of h_1 described above should have made it symmetric and having all maxima positive and all minima negative. After the first round of sifting, the crest may become a local maximum. New extrema generated in this way actually reveal the proper modes lost in the initial examination. In the subsequent sifting process, h_1 can only be treated as a proto-IMF. In the next step, it is treated as the data, then

$$h_1 - m_{11} = h_{11}.$$

After repeated sifting up to k times, h_1 becomes an IMF, that is

$$h_{1(k-1)} - m_{1k} = h_{1k}.$$

Then, it is designated as the first IMF component from the data:

$$c_1 = h_{1k}.$$

The stoppage criteria of the sifting process

The stoppage criterion determines the number of sifting steps to produce an IMF. Two different stoppage criteria have been used traditionally:

- 1. The first criterion is proposed by Huang et al. (1998). It similar to the Cauchy convergence test, and we define a sum of the difference, SD, as

  $$SD_k = \frac{\sum_{t=0}^{T} |h_{k-1}(t) - h_k(t)|^2}{\sum_{t=0}^{T} h_{k-1}^2(t)}.$$

  Then the sifting process is stop when SD is smaller than a pre-given value.

- 2. A second criterion is based on the number called the S-number, which is defined as the number of consecutive siftings when the numbers of zero-crossings and extrema are equal or at most differing by one. Specifically, an S-number is pre-selected. The sifting process will stop only if for S consecutive times the numbers of zero-crossings and extrema stay the same, and are equal or at most differ by one.

Once a stoppage criterion is selected, the first IMF, c_1, can be obtained. Overall, c_1 should contain the finest scale or the shortest period component of the signal. We can, then, separate c_1 from the rest of the data by $X(t) - c_1 = r_1$.Since the residue, r_1, still contains longer period variations in the data, it is treated as the new data and subjected to the same sifting process as described above.

This procedure can be repeated to all the subsequent r_j's, and the result is

$$r_{n-1} - c_n = r_n.$$

The sifting process stops finally when the residue, r_n, becomes a monotonic function from which no more IMF can be extracted. From the above equations, we can induce that

$$X(t) = \sum_{j=1}^{n} c_j + r_n.$$

Thus, a decomposition of the data into n-empirical modes is achieved. The components of the EMD are usually physically meaningful, for the characteristic scales are defined by the physical data. Flandrin et al. (2003) and Wu and Huang (2004) have shown that the EMD is equivalent to a dyadic filter bank.

Hilbert spectral analysis

Having obtained the intrinsic mode function components, the instantaneous frequency can be computed using the Hilbert Transform. After performing the Hilbert transform on each IMF component, the original data can be expressed as the real part, Real, in the following form:

$$X(t) = \text{Real}\sum_{j=1}^{n} a_j(t)e^{i\int \omega_j(t)dt}.$$

Current applications

- **Biomedical applications**: Huang et al. [1999b] analyzed the pulmonary arterial pressure on conscious and unrestrained rats.

- **Chemistry and chemical engineering**: Phillips et al. [2003] investigated a conformational change in Brownian dynamics(BD) and molecular dynamics(MD) simulations using a comparative analysis of HHT and wavelet methods. Wiley et al. [2004] used HHT to investigate the effect of reversible digitally filtered molecular dynamics(RDFMD) which can enhance or suppress specific frequencies of motion. Montesinos et al. [2002] applied HHT to signals obtained from BWR neuron stability.

- **Financial applications**: Huang et al. [2003b] applied HHT to nonstationary financial time series and used a weekly mortgage rate data.

- **Image processing**: Hariharan et al. [2006] applied EMD to image fusion and enhancement. Chang et al. [2009] applied an improved EMD to iris recognition, which reported a 100% faster in computational speed without losing accuracy than the original EMD.

- **Meteorological and atmospheric applications**: Salisbury and Wimbush [2002], using Southern Oscillation Index(SOI) data, applied the HHT technique to determine whether the SOI data are sufficiently noise free that useful predictions can be made and whether future El Nino southern oscillation(ENSO) events can be predicted from SOI data. Pan et al. [2002] used HHT to analyze satellite scatterometer wind data over the northwestern Pacific and compared the results to vector empirical orthogonal function(VEOF) results.

- **Ocean engineering**:Schlurmann [2002] introduced the application of HHT to characterize nonlinear water waves from two different perspectives, using laboratory experiments. Veltcheva [2002] applied HHT to wave data from nearshore sea. Larsen et al. [2004] used HHT to characterize the underwater electromagnetic environment and identify transient manmade electromagnetic disturbances.

- **Seismic studies**: Huang et al. [2001] used HHT to develop a spectral representation of earthquake data. Chen et al. [2002a] used HHT to determined the dispersion curves of seismic surface waves and compared their results to Fourier-based time-frequency analysis. Shen et al. [2003] applied HHT to ground motion and compared the HHT result with the Fourier spectrum.

- **Solar Physics**: Barnhart and Eichinger [2010] used HHT to extract the periodic components within sunspot data, including the 11-year Schwabe, 22-year Hale, and ~100-year Gleissberg cycles. They compared their results with traditional Fourier analysis.

- **Structural applications**: Quek et al. [2003] illustrate the feasibility of the HHT as a signal processing tool for locating an anomaly in the form of a crack, delamination, or stiffness loss in beams and plates based on physically acquired propagating wave signals. Using HHT, Li et al. [2003] analyzed the results of a pseudodynamic test of two rectangular reinforced concrete bridge columns.

- **Health monitoring**: Pines and Salvino [2002] applied HHT in structural health monitoring. Yang et al. [2004] used HHT for damage detection, applying EMD to extract damage spikes due to sudden changes in structural stiffness. Yu et al. [2003] used HHT for fault diagnosis of roller bearings.

- **System identification**: Chen and Xu [2002] explored the possibility of using HHT to identify the modal damping ratios of a structure with closely spaced modal frequencies and compared their results to FFT. Xu et al. [2003] compared the modal frequencies and damping ratios in various time increments and different winds for one of the tallest composite buildings in the world.

Limitations

Chen and Feng [undated]] proposed a technique to improve the HHT procedure. The authors noted that the EMD is limited in distinguishing different components in narrow-band signals. The narrow band may contain either (a) components that have adjacent frequencies or (b) components that are not adjacent in frequency but for which one of the components has a much higher energy intensity than the other components. The improved technique is based on beating-phenomenon waves.

Datig and Schlurmann [2004] did the most comprehensive studies on the performance and limitations of HHT with particular applications to irregular waves. The authors did extensive investigation into the spline interpolation. The authors discussed using additional points, both forward and backward, to determine better envelopes. They also performed a parametric study on the proposed improvement and showed significant improvement in the overall EMD computations. The authors noted that HHT is capable of differentiating between time-variant components from any given data. Their study also showed that HHT was able to distinguish between riding and carrier waves.

Chapter-9

Intersymbol Interference and Pulse Shaping

Intersymbol interference

In telecommunication, **intersymbol interference** (**ISI**) is a form of distortion of a signal in which one symbol interferes with subsequent symbols. This is an unwanted phenomenon as the previous symbols have similar effect as noise, thus making the communication less reliable. ISI is usually caused by multipath propagation or the inherent non-linear frequency response of a channel causing successive symbols to "blur" together. The presence of ISI in the system introduces errors in the decision device at the receiver output. Therefore, in the design of the transmitting and receiving filters, the objective is to minimize the effects of ISI, and thereby deliver the digital data to its destination with the smallest error rate possible. Ways to fight intersymbol interference include adaptive equalization and error correcting codes.

Causes

Multipath propagation

One of the causes of intersymbol interference is what is known as multipath propagation in which a wireless signal from a transmitter reaches the receiver via many different paths. The causes of this include reflection (for instance, the signal may bounce off buildings), refraction (such as through the foliage of a tree) and atmospheric effects such as atmospheric ducting and ionospheric reflection. Since all of these paths are different lengths - plus some of these effects will also slow the signal down - this results in the different versions of the signal arriving at different times. This delay means that part or all of a given symbol will be spread into the subsequent symbols, thereby interfering with the correct detection of those symbols. Additionally, the various paths often distort the amplitude and/or phase of the signal thereby causing further interference with the received signal.

Bandlimited channels

Another cause of intersymbol interference is the transmission of a signal through a bandlimited channel, i.e., one where the frequency response is zero above a certain frequency (the cutoff frequency). Passing a signal through such a channel results in the removal of frequency components above this cutoff frequency; in addition, the amplitude of the frequency components below the cutoff frequency may also be attenuated by the channel.

This filtering of the transmitted signal affects the shape of the pulse that arrives at the receiver. The effects of filtering a rectangular pulse; not only change the shape of the pulse within the first symbol period, but it is also spread out over the subsequent symbol periods. When a message is transmitted through such a channel, the spread pulse of each individual symbol will interfere with following symbols.

As opposed to multipath propagation, bandlimited channels are present in both wired and wireless communications. The limitation is often imposed by the desire to operate multiple independent signals through the same area/cable; due to this, each system is typically allocated a piece of the total bandwidth available. For wireless systems, they may be allocated a slice of the electromagnetic spectrum to transmit in (for example, FM radio is often broadcast in the 87.5 MHz - 108 MHz range). This allocation is usually administered by a government agency; in the case of the United States this is the Federal Communications Commission (FCC). In a wired system, such as an optical fiber cable, the allocation will be decided by the owner of the cable.

The bandlimiting can also be due to the physical properties of the medium - for instance, the cable being used in a wired system may have a cutoff frequency above which practically none of the transmitted signal will propagate.

Communication systems that transmit data over bandlimited channels usually implement pulse shaping to avoid interference caused by the bandwidth limitation. If the channel frequency response is flat and the shaping filter has a finite bandwidh, it is possible to communicate with no ISI at all. Often the channel response is not known beforehand, and an adaptive equalizer is used to compensate the frequency response.

Effects on eye patterns

One way to study ISI in a PCM or data transmission system experimentally is to apply the received wave to the vertical deflection plates of an oscilloscope and to apply a sawtooth wave at the transmitted symbol rate R, 1/T to the horizontal deflection plates. The resulting display is called an eye pattern because of its resemblance to the human eye for binary waves. The interior region of the eye pattern is called the eye opening. An eye pattern provides a great deal of information about the performance of the pertinent system.

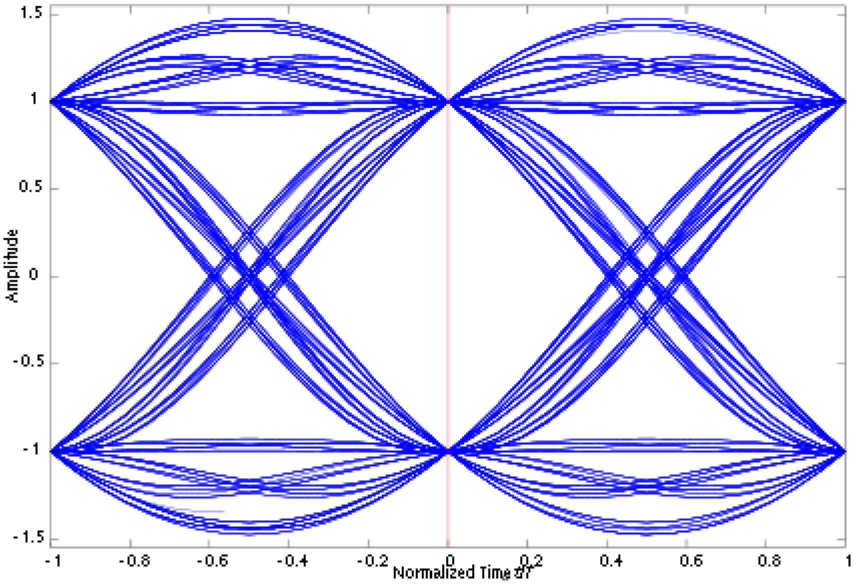
1. The width of the eye opening defines the time interval over which the received wave can be sampled without error from ISI. It is apparent that the preferred time for sampling is the instant of time at which the eye is open widest.
2. The sensitivity of the system to timing error is determined by the rate of closure of the eye as the sampling time is varied.

3. The height of the eye opening, at a specified sampling time, defines the margin over noise.

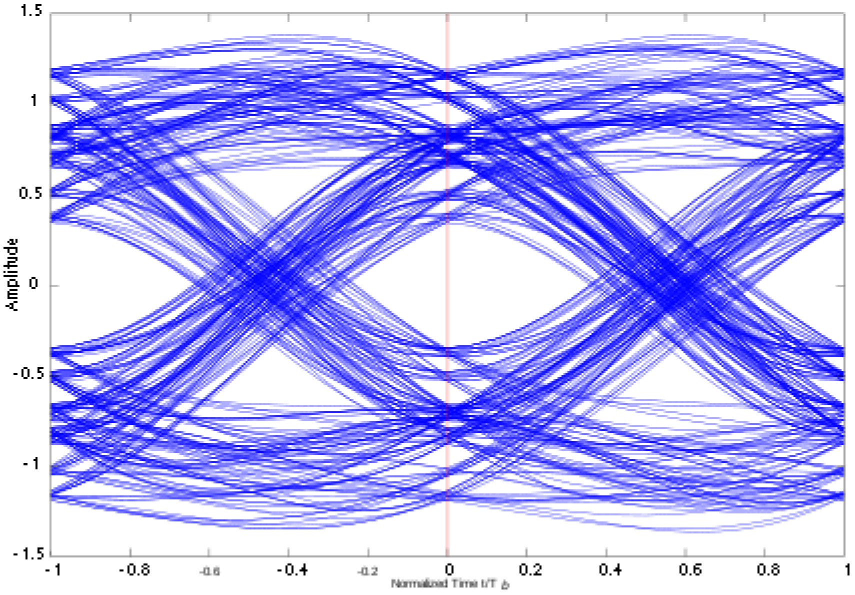
An eye pattern, which overlays many samples of a signal, can give a graphical representation of the signal characteristics. The first image below is the eye pattern for a binary phase-shift keying (PSK) system in which a one is represented by an amplitude of -1 and a zero by an amplitude of +1. The current sampling time is at the center of the image and the previous and next sampling times are at the edges of the image. The various transitions from one sampling time to another (such as one-to-zero, one-to-one and so forth) can clearly be seen on the diagram.

The noise margin - the amount of noise required to cause the receiver to get an error - is given by the distance between the signal and the zero amplitude point at the sampling time; in other words, the further from zero at the sampling time the signal is the better. For the signal to be correctly interpreted, it must be sampled somewhere between the two points where the zero-to-one and one-to-zero transitions cross. Again, the further apart these points are the better, as this means the signal will be less sensitive to errors in the timing of the samples at the receiver.

The effects of ISI are shown in the second image which is an eye pattern of the same system when operating over a multipath channel. The effects of receiving delayed and distorted versions of the signal can be seen in the loss of definition of the signal transitions. It also reduces both the noise margin and the window in which the signal can be sampled, which shows that the performance of the system will be worse (i.e. it will have a greater bit error ratio).



The eye diagram of a binary PSK system

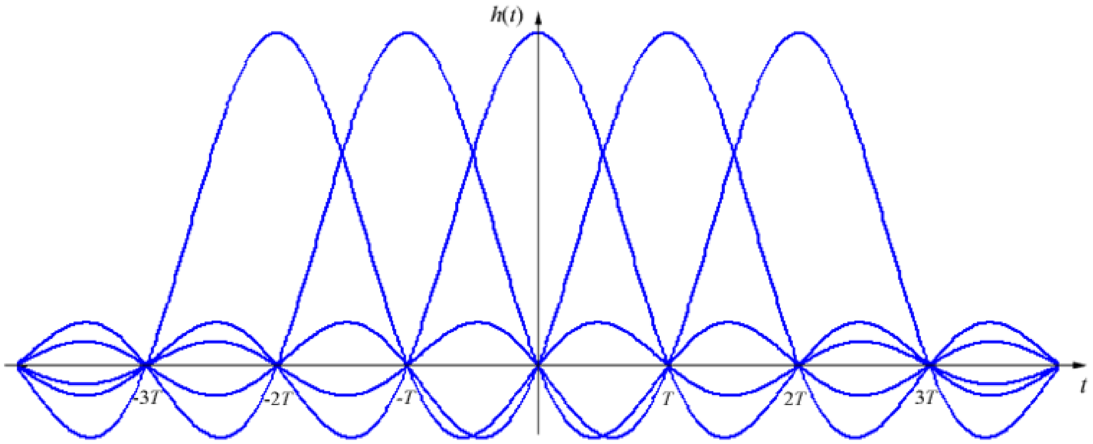


The eye diagram of the same system with multipath effects added

Countering ISI

There are several techniques in telecommunication and data storage that try to work around the problem of intersymbol interference.

- Design systems such that the impulse response is short enough that very little energy from one symbol smears into the next symbol.



Consecutive raised-cosine impulses, demonstrating zero-ISI property

- Separate symbols in time with guard periods.
- Apply an equalizer at the receiver, that, broadly speaking, attempts to undo the effect of the channel by applying an inverse filter.
- Apply a sequence detector at the receiver, that attempts to estimate the sequence of transmitted symbols using the Viterbi algorithm.

Pulse shaping

In digital telecommunication, **pulse shaping** is the process of changing the waveform of transmitted pulses. Its purpose is to make the transmitted signal better suited to the communication channel by limiting the effective bandwidth of the transmission. By filtering the transmitted pulses this way, the intersymbol interference caused by the channel can be kept in control. In RF communication, pulse shaping is essential for making the signal fit in its frequency band.

Typically pulse shaping occurs after line coding and before modulation.

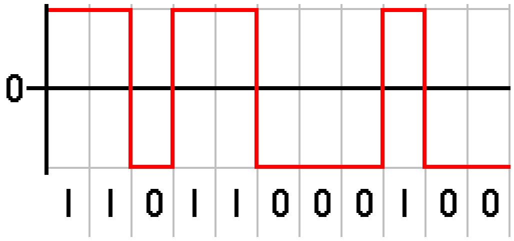
Need for pulse shaping

Transmitting a signal at high modulation rate through a band-limited channel can create intersymbol interference. As the modulation rate increases, the signal's bandwidth increases. When the signal's bandwidth becomes larger than the channel bandwidth, the channel starts to introduce distortion to the signal. This distortion is usually seen as intersymbol interference.

The signal's spectrum is determined by the pulse shaping filter used by the transmitter. Usually the transmitted symbols are represented as a time sequence of dirac delta pulses. This theoretical signal is then filtered with the pulse shaping filter, producing the transmitted signal. The spectrum of the transmission is thus determined by the filter.

In many base band communication systems the pulse shaping filter is implicitly a boxcar filter. Its spectrum is of the form *sin(x)/x*, and has significant signal power at frequencies higher than symbol rate. This is not a big problem when optical fibre or even twisted pair cable is used as the communication channel. However, in RF communications this would waste bandwidth, and only tightly specified frequency bands are used for single transmissions. In other words, the channel for the signal is band-limited. Therefore better filters have been developed, which attempt to minimise the bandwidth needed for a certain symbol rate.

Pulse filters



A typical NRZ coded signal is implicitly filtered with a boxcar filter.

Not all filters can be used as a pulse shaping filter. The filter itself must not introduce intersymbol interference — it needs to satisfy certain criteria. Nyquist ISI criterion is commonly used criterion for evaluation of filters, because it relates the frequency spectrum of the transmitter signal to intersymbol interference.

Examples of pulse-shaping filters that are commonly found in communication systems are:

- The trivial boxcar filter
- Sinc shaped filter
- Raised-cosine filter
- Gaussian filter

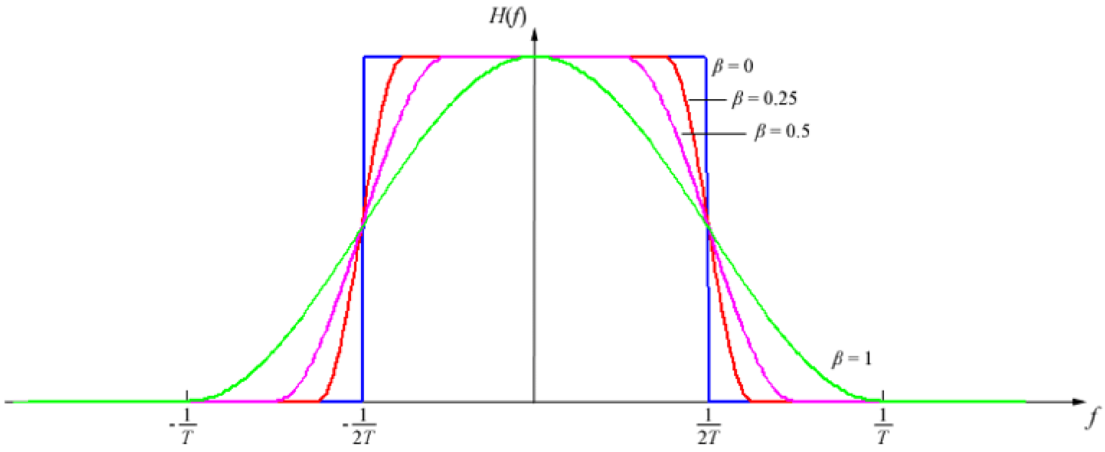
Sender side pulse shaping is often combined with a receiver side matched filter to achieve optimum tolerance for noise in the system. In this case the pulse shaping is equally distributed to the sender and receiver filters. The filters' amplitude responses are thus pointwise square-roots of the system filters.

Other approaches that eliminate complex pulse shaping filters have been invented. In OFDM, the carriers are modulated so slowly that each carrier is virtually unaffected by the bandwidth limitation of the channel.

Boxcar filter

The boxcar filter results in infinitely wide bandwidth for the signal. Thus its usefulness is limited, but it is used widely in wired baseband communications, where the channel has some extra bandwidth and the distortion created by the channel can be tolerated.

Sinc filter



Amplitude response of raised-cosine filter with various roll-off factors

Theoretically the best pulse shaping filter would be the sinc filter, but it cannot be implemented precisely. It is a non-causal filter with relatively slowly decaying tails. It is also problematic from a synchronisation point of view as any phase error results in steeply increasing intersymbol interference.

Raised-cosine filter

Raised-cosine filter is practical to implement and it is in wide use. It has a parametrisable excess bandwidth, so communication systems can choose a trade-off between a more complex filter and spectral efficiency.

Gaussian filter

This gives an output pulse shaped like a Gaussian function.

Chapter-10

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:

$$|a^Hb|^2 \leq (a^Ha)(b^Hb),$$

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:

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

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*:

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

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:

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

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

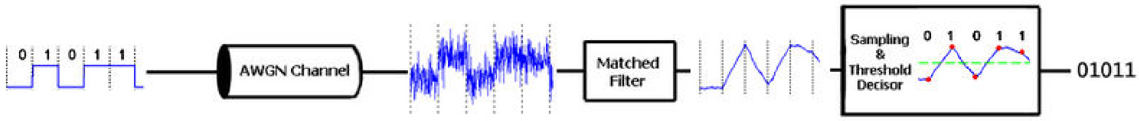
Matched filters are often used in signal detection. As an example, suppose that we wish to judge the distance of an object by reflecting a signal off it. We may choose to transmit a pure-tone sinusoid at 1 Hz. We assume that our received signal is an attenuated and phase-shifted form of the transmitted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

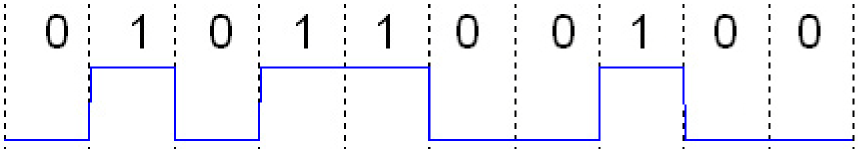
$$a_k = \begin{cases} 1, & \text{if bit } k \text{ is } 1, \\ 0, & \text{if bit } k \text{ is } 0. \end{cases}$$

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

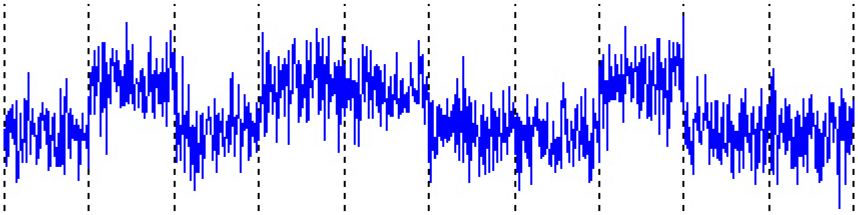
$$M(t) = \sum_{k=-\infty}^{\infty} a_k \times \Pi\left(\frac{t-kT}{T}\right).$$

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

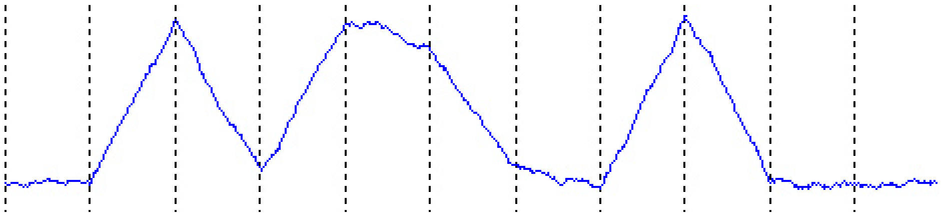
$$h(t) = \Pi\left(\frac{t}{T}\right).$$

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.

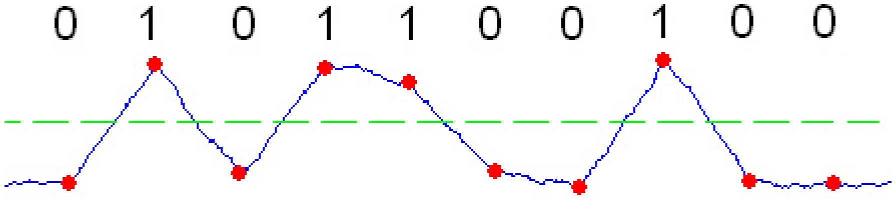
After convolving with the correct matched filter, the resulting signal, $M_{\text{filtered}}(t)$ is,

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

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-11

Modulation

In electronics, **modulation** is the process of varying one or more properties of a high-frequency periodic waveform, called the *carrier signal*, with respect to a *modulating signal* (which typically contains information to be transmitted). This is done in a similar fashion to a musician modulating a tone (a periodic waveform) from a musical instrument by varying its volume, timing and pitch. The three key parameters of a periodic waveform are its amplitude ("volume"), its phase ("timing") and its frequency ("pitch"), all of which can be modified in accordance with a low frequency signal to obtain the modulated signal. Typically a high-frequency sinusoid waveform is used as carrier signal, but a square wave pulse train may also occur.

In telecommunications, modulation is the process of conveying a message signal, for example a digital bit stream or an analog audio signal, inside another signal that can be physically transmitted. Modulation of a sine waveform is used to transform a baseband message signal into a passband signal, for example low-frequency audio signal into a radio-frequency signal (RF signal). In radio communications, cable TV systems or the public switched telephone network for instance, electrical signals can only be transferred over a limited passband frequency spectrum, with specific (non-zero) lower and upper cutoff frequencies. Modulating a sine-wave carrier makes it possible to keep the frequency content of the transferred signal as close as possible to the centre frequency (typically the carrier frequency) of the passband.

A device that performs modulation is known as a modulator and a device that performs the inverse operation of modulation is known as a demodulator (sometimes *detector* or *demod*). A device that can do both operations is a modem (modulator–demodulator).

Aim

The aim of **digital modulation** is to transfer a digital bit stream over an analog bandpass channel, for example over the public switched telephone network (where a bandpass filter limits the frequency range to between 300 and 3400 Hz), or over a limited radio frequency band.

The aim of **analog modulation** is to transfer an analog baseband (or lowpass) signal, for example an audio signal or TV signal, over an analog bandpass channel, for example a limited radio frequency band or a cable TV network channel.

Analog and digital modulation facilitate frequency division multiplexing (FDM), where several low pass information signals are transferred simultaneously over the same shared physical medium, using separate passband channels.

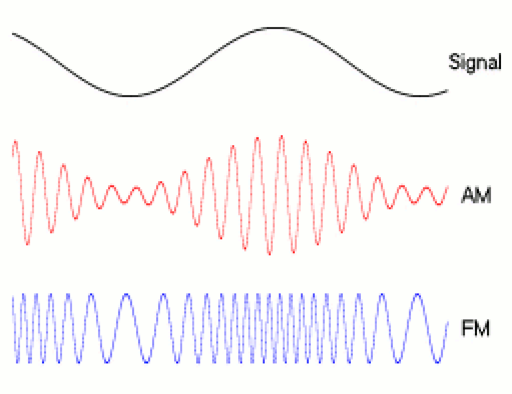
The aim of **digital baseband modulation** methods, also known as line coding, is to transfer a digital bit stream over a baseband channel, typically a non-filtered copper wire such as a serial bus or a wired local area network.

The aim of **pulse modulation** methods is to transfer a narrowband analog signal, for example a phone call over a wideband baseband channel or, in some of the schemes, as a bit stream over another digital transmission system.

In music synthesizers, modulation may be used to synthesise waveforms with a desired overtone spectrum. In this case the carrier frequency is typically in the same order or much lower than the modulating waveform.

Analog modulation methods

In analog modulation, the modulation is applied continuously in response to the analog information signal.



A low-frequency message signal (top) may be carried by an AM or FM radio wave.

Common analog modulation techniques are:

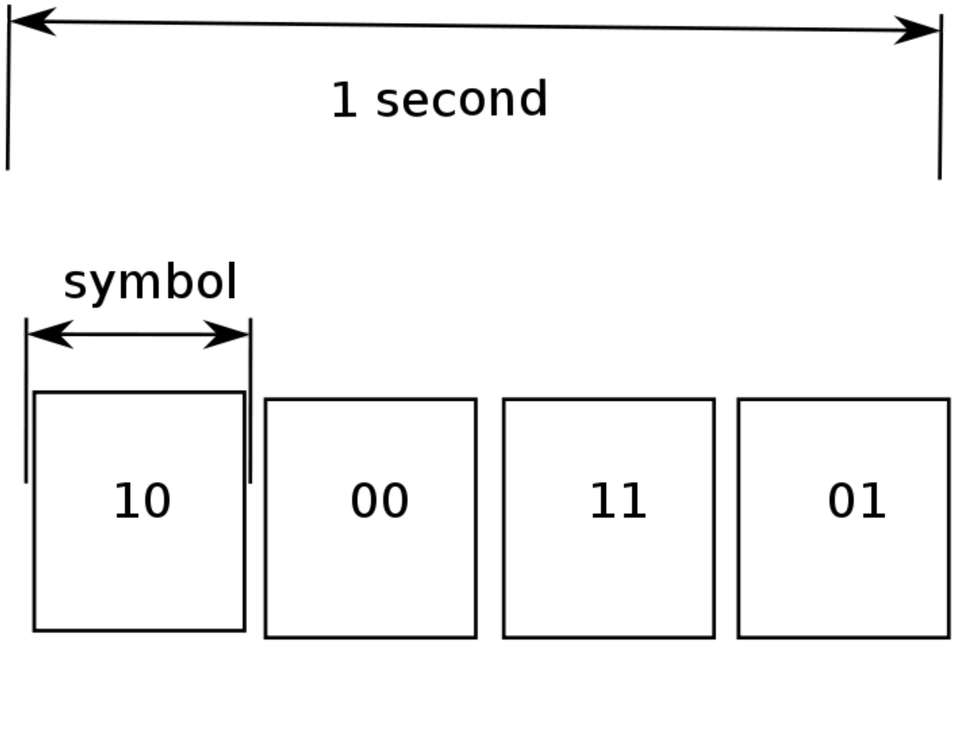
- Amplitude modulation (AM) (here the amplitude of the carrier signal is varied in accordance to the instantaneous amplitude of the modulating signal)
    - Double-sideband modulation (DSB)
        - Double-sideband modulation with carrier (DSB-WC) (used on the AM radio broadcasting band)

        - Double-sideband suppressed-carrier transmission (DSB-SC)
        - Double-sideband reduced carrier transmission (DSB-RC)
    - Single-sideband modulation (SSB, or SSB-AM)
        - SSB with carrier (SSB-WC)
        - SSB suppressed carrier modulation (SSB-SC)
    - Vestigial sideband modulation (VSB, or VSB-AM)
    - Quadrature amplitude modulation (QAM)

- Angle modulation
    - Frequency modulation (FM) (here the frequency of the carrier signal is varied in accordance to the instantaneous amplitude of the modulating signal)
    - Phase modulation (PM) (here the phase shift of the carrier signal is varied in accordance to the instantaneous amplitude of the modulating signal)

Digital modulation methods

In digital modulation, an analog carrier signal is modulated by a digital bit stream. Digital modulation methods can be considered as digital-to-analog conversion, and the corresponding demodulation or detection as analog-to-digital conversion. The changes in the carrier signal are chosen from a finite number of M alternative symbols (the *modulation alphabet*).



Schematic of 4 baud (8 bps) data link.

A simple example: A telephone line is designed for transferring audible sounds, for example tones, and not digital bits (zeros and ones). Computers may however communicate over a telephone line by means of modems, which are representing the digital bits by tones, called symbols. If there are four alternative symbols (corresponding to a musical instrument that can generate four different tones, one at a time), the first symbol may represent the bit sequence 00, the second 01, the third 10 and the fourth 11. If the modem plays a melody consisting of 1000 tones per second, the symbol rate is 1000 symbols/second, or baud. Since each tone (i.e., symbol) represents a message consisting of two digital bits in this example, the bit rate is twice the symbol rate, i.e. 2000 bits per second. This is similar to the technique used by dialup modems as opposed to DSL modems.

According to one definition of digital signal, the modulated signal is a digital signal, and according to another definition, the modulation is a form of digital-to-analog conversion. Most textbooks would consider digital modulation schemes as a form of digital transmission, synonymous to data transmission; very few would consider it as analog transmission.

Fundamental digital modulation methods

The most fundamental digital modulation techniques are based on keying:

- In the case of PSK (phase-shift keying), a finite number of phases are used.
- In the case of FSK (frequency-shift keying), a finite number of frequencies are used.
- In the case of ASK (amplitude-shift keying), a finite number of amplitudes are used.
- In the case of QAM (quadrature amplitude modulation), a finite number of at least two phases, and at least two amplitudes are used.

In QAM, an inphase signal (the I signal, for example a cosine waveform) and a quadrature phase signal (the Q signal, for example a sine wave) are amplitude modulated with a finite number of amplitudes, and summed. It can be seen as a two-channel system, each channel using ASK. The resulting signal is equivalent to a combination of PSK and ASK.

In all of the above methods, each of these phases, frequencies or amplitudes are assigned a unique pattern of binary bits. Usually, each phase, frequency or amplitude encodes an equal number of bits. This number of bits comprises the *symbol* that is represented by the particular phase, frequency or amplitude.

If the alphabet consists of $M = 2^N$ alternative symbols, each symbol represents a message consisting of N bits. If the symbol rate (also known as the baud rate) is f_S symbols/second (or baud), the data rate is Nf_S bit/second.

For example, with an alphabet consisting of 16 alternative symbols, each symbol represents 4 bits. Thus, the data rate is four times the baud rate.

In the case of PSK, ASK or QAM, where the carrier frequency of the modulated signal is constant, the modulation alphabet is often conveniently represented on a constellation diagram, showing the amplitude of the I signal at the x-axis, and the amplitude of the Q signal at the y-axis, for each symbol.

Modulator and detector principles of operation

PSK and ASK, and sometimes also FSK, are often generated and detected using the principle of QAM. The I and Q signals can be combined into a complex-valued signal *I*+*jQ* (where *j* is the imaginary unit). The resulting so called equivalent lowpass signal or equivalent baseband signal is a complex-valued representation of the real-valued modulated physical signal (the so called passband signal or RF signal).

These are the general steps used by the modulator to transmit data:

1. Group the incoming data bits into codewords, one for each symbol that will be transmitted.
2. Map the codewords to attributes, for example amplitudes of the I and Q signals (the equivalent low pass signal), or frequency or phase values.
3. Adapt pulse shaping or some other filtering to limit the bandwidth and form the spectrum of the equivalent low pass signal, typically using digital signal processing.
4. Perform digital-to-analog conversion (DAC) of the I and Q signals (since today all of the above is normally achieved using digital signal processing, DSP).
5. Generate a high-frequency sine wave carrier waveform, and perhaps also a cosine quadrature component. Carry out the modulation, for example by multiplying the sine and cosine wave form with the I and Q signals, resulting in that the equivalent low pass signal is frequency shifted into a modulated passband signal or RF signal. Sometimes this is achieved using DSP technology, for example direct digital synthesis using a waveform table, instead of analog signal processing. In that case the above DAC step should be done after this step.
6. Amplification and analog bandpass filtering to avoid harmonic distortion and periodic spectrum

At the receiver side, the demodulator typically performs:

1. Bandpass filtering.
2. Automatic gain control, AGC (to compensate for attenuation, for example fading).
3. Frequency shifting of the RF signal to the equivalent baseband I and Q signals, or to an intermediate frequency (IF) signal, by multiplying the RF signal with a local oscillator sinewave and cosine wave frequency.

4. Sampling and analog-to-digital conversion (ADC) (Sometimes before or instead of the above point, for example by means of undersampling).
5. Equalization filtering, for example a matched filter, compensation for multipath propagation, time spreading, phase distortion and frequency selective fading, to avoid intersymbol interference and symbol distortion.
6. Detection of the amplitudes of the I and Q signals, or the frequency or phase of the IF signal.
7. Quantization of the amplitudes, frequencies or phases to the nearest allowed symbol values.
8. Mapping of the quantized amplitudes, frequencies or phases to codewords (bit groups).
9. Parallel-to-serial conversion of the codewords into a bit stream.
10. Pass the resultant bit stream on for further processing such as removal of any error-correcting codes.

As is common to all digital communication systems, the design of both the modulator and demodulator must be done simultaneously. Digital modulation schemes are possible because the transmitter-receiver pair have prior knowledge of how data is encoded and represented in the communications system. In all digital communication systems, both the modulator at the transmitter and the demodulator at the receiver are structured so that they perform inverse operations.

Non-coherent modulation methods do not require a receiver reference clock signal that is phase synchronized with the sender carrier wave. In this case, modulation symbols (rather than bits, characters, or data packets) are asynchronously transferred. The opposite is coherent modulation.

List of common digital modulation techniques

The most common digital modulation techniques are:

- Phase-shift keying (PSK):
    - Binary PSK (BPSK), using M=2 symbols
    - Quadrature PSK (QPSK), using M=4 symbols
    - 8PSK, using M=8 symbols
    - 16PSK, using M=16 symbols
    - Differential PSK (DPSK)
    - Differential QPSK (DQPSK)
    - Offset QPSK (OQPSK)
    - π/4–QPSK
- Frequency-shift keying (FSK):
    - Audio frequency-shift keying (AFSK)
    - Multi-frequency shift keying (M-ary FSK or MFSK)
    - Dual-tone multi-frequency (DTMF)
    - Continuous-phase frequency-shift keying (CPFSK)
- Amplitude-shift keying (ASK)

- On-off keying (OOK), the most common ASK form
    - M-ary vestigial sideband modulation, for example 8VSB
- Quadrature amplitude modulation (QAM) - a combination of PSK and ASK:
    - Polar modulation like QAM a combination of PSK and ASK.
- Continuous phase modulation (CPM) methods:
    - Minimum-shift keying (MSK)
    - Gaussian minimum-shift keying (GMSK)
- Orthogonal frequency-division multiplexing (OFDM) modulation:
    - discrete multitone (DMT) - including adaptive modulation and bit-loading.
- Wavelet modulation
- Trellis coded modulation (TCM), also known as trellis modulation
- Spread-spectrum techniques:
    - Direct-sequence spread spectrum (DSSS)
    - Chirp spread spectrum (CSS) according to IEEE 802.15.4a CSS uses pseudo-stochastic coding
    - Frequency-hopping spread spectrum (FHSS) applies a special scheme for channel release

MSK and GMSK are particular cases of continuous phase modulation. Indeed, MSK is a particular case of the sub-family of CPM known as continuous-phase frequency-shift keying (CPFSK) which is defined by a rectangular frequency pulse (i.e. a linearly increasing phase pulse) of one symbol-time duration (total response signaling).

OFDM is based on the idea of frequency-division multiplexing (FDM), but is utilized as a digital modulation scheme. The bit stream is split into several parallel data streams, each transferred over its own sub-carrier using some conventional digital modulation scheme. The modulated sub-carriers are summed to form an OFDM signal. OFDM is considered as a modulation technique rather than a multiplex technique, since it transfers one bit stream over one communication channel using one sequence of so-called OFDM symbols. OFDM can be extended to multi-user channel access method in the orthogonal frequency-division multiple access (OFDMA) and multi-carrier code division multiple access (MC-CDMA) schemes, allowing several users to share the same physical medium by giving different sub-carriers or spreading codes to different users.

Of the two kinds of RF power amplifier, switching amplifiers (Class C amplifiers) cost less and use less battery power than linear amplifiers of the same output power. However, they only work with relatively constant-amplitude-modulation signals such as angle modulation (FSK or PSK) and CDMA, but not with QAM and OFDM. Nevertheless, even though switching amplifiers are completely unsuitable for normal QAM constellations, often the QAM modulation principle are used to drive switching amplifiers with these FM and other waveforms, and sometimes QAM demodulators are used to receive the signals put out by these switching amplifiers.

Digital baseband modulation or line coding

The term **digital baseband modulation** (or digital baseband transmission) is synonymous to line codes. These are methods to transfer a digital bit stream over an analog baseband channel (a.k.a. lowpass channel) using a pulse train, i.e. a discrete number of signal levels, by directly modulating the voltage or current on a cable. Common examples are unipolar, non-return-to-zero (NRZ), Manchester and alternate mark inversion (AMI) codings.

Chapter-12

Noisy-channel Coding Theorem and Raised-cosine Filter

Noisy-channel coding theorem

In information theory, the **noisy-channel coding theorem** (sometimes **Shannon's theorem**), establishes that for any given degree of noise contamination of a communication channel, it is possible to communicate discrete data (digital information) nearly error-free up to a computable maximum rate through the channel. This result was presented by Claude Shannon in 1948 and was based in part on earlier work and ideas of Harry Nyquist and Ralph Hartley.

The **Shannon limit** or **Shannon capacity** of a communications channel is the theoretical maximum information transfer rate of the channel, for a particular noise level.

Overview

Stated by Claude Shannon in 1948, the theorem describes the maximum possible efficiency of error-correcting methods versus levels of noise interference and data corruption. The theory doesn't describe *how to construct* the error-correcting method, it only tells us how good the *best possible* method can be. Shannon's theorem has wide-ranging applications in both communications and data storage. This theorem is of foundational importance to the modern field of information theory. Shannon only gave an outline of the proof. The first rigorous proof is due to Amiel Feinstein in 1954.

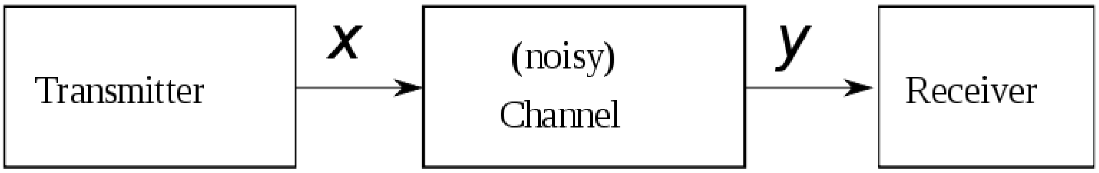
The Shannon theorem states that given a noisy channel with channel capacity C and information transmitted at a rate R, then if $R < C$ there exist codes that allow the probability of error at the receiver to be made arbitrarily small. This means that, theoretically, it is possible to transmit information nearly without error at any rate below a limiting rate, C.

The converse is also important. If $R > C$, an arbitrarily small probability of error is not achievable. All codes will have a probability of error greater than a certain positive minimal level, and this level increases as the rate increases. So, information cannot be guaranteed to be transmitted reliably across a channel at rates beyond the channel capacity. The theorem does not address the rare situation in which rate and capacity are equal.

The channel capacity C can be calculated from the physical properties of a channel; for a band-limited channel with Gaussian noise, using the Shannon–Hartley theorem.

Simple schemes such as "send the message 3 times and use a best 2 out of 3 voting scheme if the copies differ" are inefficient error-correction methods, unable to asymptotically guarantee that a block of data can be communicated free of error. Advanced techniques such as Reed–Solomon codes and, more recently, turbo codes come much closer to reaching the theoretical Shannon limit, but at a cost of high computational complexity. Using low-density parity-check (LDPC) codes or turbo codes and with the computing power in today's digital signal processors, it is now possible to reach very close to the Shannon limit. In fact, it was shown that LDPC codes can reach within 0.0045 dB of the Shannon limit (for very long block lengths).

Mathematical statement



Theorem (Shannon, 1948):

1. For every discrete memoryless channel, the channel capacity

$$C = \sup_{p_X} I(X;Y)$$

has the following property. For any ε > 0 and $R < C$, for large enough N, there exists a code of length N and rate $\geq R$ and a decoding algorithm, such that the maximal probability of block error is $\leq$ ε.
2. If a probability of bit error p_b is acceptable, rates up to $R(p_b)$ are achievable, where

$$R(p_b) = \frac{C}{1 - H_2(p_b)}.$$

and $H_2(p_b)$ is the *binary entropy function*

$$H_2(p_b) = -\left[p_b \log p_b + (1 - p_b) \log(1 - p_b)\right]$$

3. For any p_b, rates greater than $R(p_b)$ are not achievable.

(MacKay (2003), p. 162; cf Gallager (1968), ch.5; Cover and Thomas (1991), p. 198; Shannon (1948) thm. 11)

Outline of proof

As with several other major results in information theory, the proof of the noisy channel coding theorem includes an achievability result and a matching converse result. These two components serve to bound, in this case, the set of possible rates at which one can communicate over a noisy channel, and matching serves to show that these bounds are tight bounds.

The following outlines are only one set of many different styles available for study in information theory texts.

Achievability for discrete memoryless channels

This particular proof of achievability follows the style of proofs that make use of the asymptotic equipartition property (AEP). Another style can be found in information theory texts using error exponents.

Both types of proofs make use of a random coding argument where the codebook used across a channel is randomly constructed - this serves to reduce computational complexity while still proving the existence of a code satisfying a desired low probability of error at any data rate below the channel capacity.

By an AEP-related argument, given a channel, length *n* strings of source symbols X_1^n, and length *n* strings of channel outputs Y_1^n, we can define a *jointly typical set* by the following:

$$\begin{aligned}
&A_\varepsilon^{(n)} = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \\
&2^{-n(H(X)+\varepsilon)} \le p(X_1^n) \le 2^{-n(H(X)-\varepsilon)} \\
&2^{-n(H(Y)+\varepsilon)} \le p(Y_1^n) \le 2^{-n(H(Y)-\varepsilon)} \\
&2^{-n(H(X,Y)+\varepsilon)} \le p(X_1^n, Y_1^n) \le 2^{-n(H(X,Y)-\varepsilon)}\}
\end{aligned}$$

We say that two sequences X_1^n và Y_1^n are *jointly typical* if they lie in the jointly typical set defined above.

Steps

1. In the style of the random coding argument, we randomly generate 2^{nR} codewords of length n from a probability distribution Q.
2. This code is revealed to the sender and receiver. It is also assumed that one knows the transition matrix *p*(*y* | *x*) for the channel being used.
3. A message W is chosen according to the uniform distribution on the set of codewords. That is, $Pr(W = w) = 2^{-nR}, w = 1, 2, \dots, 2^{nR}$.
4. The message W is sent across the channel.

5. The receiver receives a sequence according to $P(y^n|x^n(w)) = \prod_{i=1}^{n} p(y_i|x_i(w))$
6. Sending these codewords across the channel, we receive Y_1^n, and decode to some source sequence if there exists exactly 1 codeword that is jointly typical with Y. If there are no jointly typical codewords, or if there are more than one, an error is declared. An error also occurs if a decoded codeword doesn't match the original codeword. This is called *typical set decoding*.

The probability of error of this scheme is divided into two parts:

1. First, error can occur if no jointly typical X sequences are found for a received Y sequence
2. Second, error can occur if an incorrect X sequence is jointly typical with a received Y sequence.

- By the randomness of the code construction, we can assume that the average probability of error averaged over all codes does not depend on the index sent. Thus, without loss of generality, we can assume $W = 1$.

- From the joint AEP, we know that the probability that no jointly typical X exists goes to 0 as n grows large. We can bound this error probability by ε.

- Also from the joint AEP, we know the probability that a particular $X_1^n(i)$and the Y_1^n resulting from $W = 1$ are jointly typical is $\leq 2^{-n(I(X;Y)-3\varepsilon)}$.

Define: $E_i = \{(X_1^n(i), Y_1^n) \in A_\varepsilon^{(n)}\}, i = 1, 2, \ldots, 2^{nR}$

as the event that message i is jointly typical with the sequence received when message 1 is sent.

$$\begin{aligned} P(\text{error}) = P(\text{error}|W=1) &\leq P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i) \\ &\leq \varepsilon + 2^{-n(I(X;Y)-R-3\varepsilon)}. \end{aligned}$$

We can observe that as *n* goes to infinity, if *R* < *I*(*X*;*Y*) for the channel, the probability of error will go to 0.

Finally, given that the average codebook is shown to be "good," we know that there exists a codebook whose performance is better than the average, and so satisfies our need for arbitrarily low error probability communicating across the noisy channel.

Weak converse for discrete memoryless channels

Suppose a code of 2^{nR} codewords. Let W be drawn uniformly over this set as an index. Let X^n and Y^n be the codewords and received codewords, respectively.

1. $nR = H(W) = H(W|Y^n) + I(W;Y^n)$ using identities involving entropy and mutual information
2. $\leq H(W|Y^n) + I(X^n(W);Y^n)$ since X is a function of W
3. $\leq 1 + P_e^{(n)}nR + I(X^n(W);Y^n)$ by the use of Fano's Inequality
4. $\leq 1 + P_e^{(n)}nR + nC$ by the fact that capacity is maximized mutual information.

The result of these steps is that $P_e^{(n)} \geq 1 - \frac{1}{nR} - \frac{C}{R}$. As the block length *n* goes to infinity, we obtain $P_e^{(n)}$ is bounded away from 0 if R is greater than C - we can get arbitrarily low rates of error only if R is less than C.

Strong converse for discrete memoryless channels

A strong converse theorem, proven by Wolfowitz in 1957, states that,

$$P_e \geq 1 - \frac{4A}{n(R-C)^2} - e^{-n(R-C)}$$

for some finite positive constant *A*. While the weak converse states that the error probability is bounded away from zero as *n* goes to infinity, the strong converse states that the error goes exponentially to 1. Thus, *C* is a sharp threshold between perfectly reliable and completely unreliable communication.

Channel coding theorem for non-stationary memoryless channels

We assume that the channel is memoryless, but its transition probabilities change with time, in a fashion known at the transmitter as well as the receiver.

Then the channel capacity is given by

$$C = \lim\inf \max_{p^{(X_1)},p^{(X_2)},\ldots} \frac{1}{n}\sum_{i=1}^{n} I(X_i;Y_i).$$

The maximum is attained at the capacity achieving distributions for each respective channel. That is, $C = \liminf \frac{1}{n} \sum_{i=1}^{n} C_i$ where C_i is the capacity of the i*th* channel.

Outline of the proof

The proof runs through in almost the same way as that of channel coding theorem. Achievability follows from random coding with each symbol chosen randomly from the capacity achieving distribution for that particular channel. Typicality arguments use the definition of typical sets for non-stationary sources defined in the asymptotic equipartition property article.

The technicality of lim inf comes into play when $\frac{1}{n} \sum_{i=1}^{n} C_i$ does not converge.

Raised-cosine filter

The **raised-cosine filter** is a filter frequently used for pulse-shaping in digital modulation due to its ability to minimise intersymbol interference (ISI). Its name stems from the fact that the non-zero portion of the frequency spectrum of its simplest form (β = 1) is a cosine function, 'raised' up to sit above the *f* (horizontal) axis.

Mathematical description

The raised-cosine filter is an implementation of a low-pass Nyquist filter, i.e., one that has the property of vestigial symmetry. This means that its spectrum exhibits odd symmetry about $\frac{1}{2T}$, where *T* is the symbol-period of the communications system.

Its frequency-domain description is a piecewise function, given by:

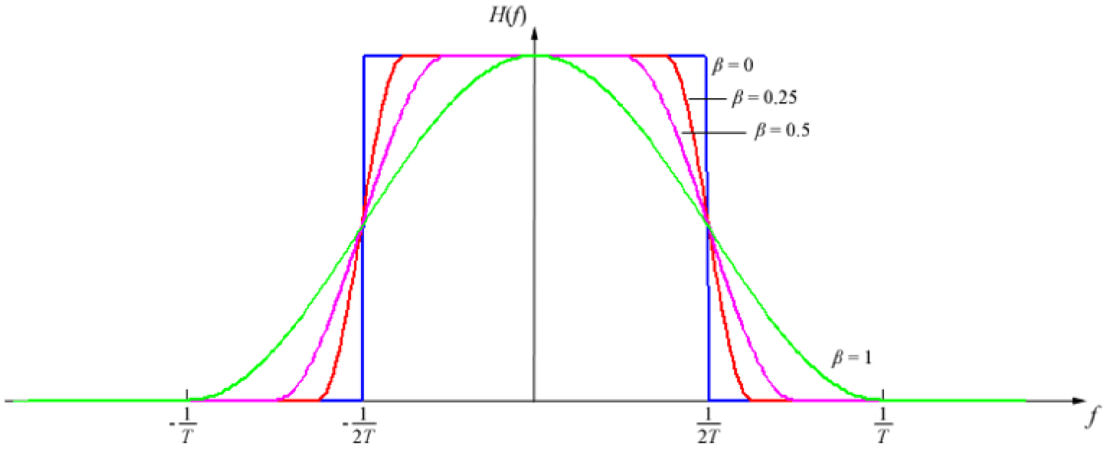
$$H(f) = \begin{cases} T, & |f| \leq \frac{1-\beta}{2T} \\ \frac{T}{2}\left[1 + \cos\left(\frac{\pi T}{\beta}\left[|f| - \frac{1-\beta}{2T}\right]\right)\right], & \frac{1-\beta}{2T} < |f| \leq \frac{1+\beta}{2T} \\ 0, & \text{otherwise} \end{cases}$$

$$0 \leq \beta \leq 1$$

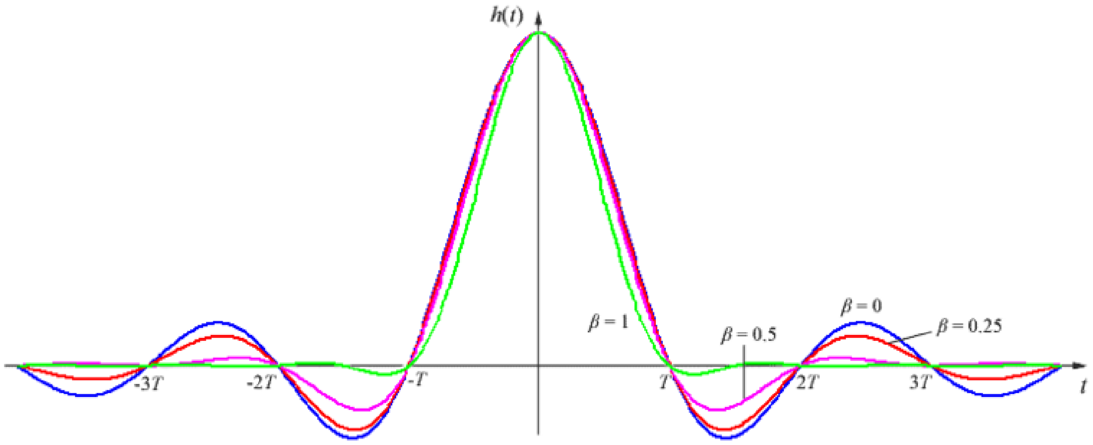
and characterised by two values; β, the *roll-off factor*, and *T*, the reciprocal of the symbol-rate.

The impulse response of such a filter is given by:

$h(t) = \mathrm{sinc}\left(\frac{t}{T}\right)\frac{\cos\left(\frac{\pi\beta t}{T}\right)}{1 - \frac{4\beta^2 t^2}{T^2}}$, in terms of the normalised sinc function.



Frequency response of raised-cosine filter with various roll-off factors



Impulse response of raised-cosine filter with various roll-off factors

Roll-off factor

The roll-off factor, β, is a measure of the *excess bandwidth* of the filter, i.e. the bandwidth occupied beyond the Nyquist bandwidth of $\frac{1}{2T}$. If we denote the excess bandwidth as Δ*f*, then:

$$\beta = \frac{\Delta f}{\left(\frac{1}{2T}\right)} = \frac{\Delta f}{R_S/2} = 2T\Delta f$$

where $R_S = \frac{1}{T}$ is the symbol-rate.

The graph shows the amplitude response as β is varied between 0 and 1, and the corresponding effect on the impulse response. As can be seen, the time-domain ripple level increases as β decreases. This shows that the excess bandwidth of the filter can be reduced, but only at the expense of an elongated impulse response.

β = 0

As β approaches 0, the roll-off zone becomes infinitesimally narrow, hence:

$$\lim_{\beta \to 0} H(f) = \mathrm{rect}(fT)$$

where rect(.) is the rectangular function, so the impulse response approaches $\mathrm{sinc}\left(\frac{t}{T}\right)$. Hence, it converges to an ideal or brick-wall filter in this case.

β = 1

When β = 1, the non-zero portion of the spectrum is a pure raised cosine, leading to the simplification:

$$H(f)|_{\beta=1} = \begin{cases} \frac{1}{2}\left[1 + \cos\left(\pi f T\right)\right], & |f| \leq \frac{1}{T} \\ 0, & \mathrm{otherwise} \end{cases}$$

Bandwidth

The bandwidth of a raised cosine filter is most commonly defined as the width of the non-zero portion of its spectrum, i.e.:

$$BW = \frac{1}{2}R_S(\beta + 1)$$

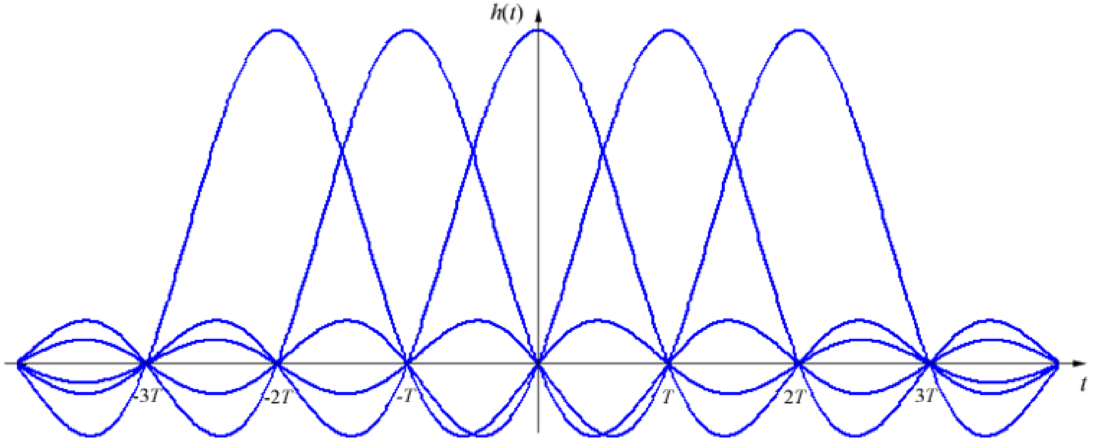
Auto-correlation function

The auto-correlation function of raised cosine function is as follows:

$$R\left(\tau\right) = T\left[sinc\left(\frac{\tau}{T}\right)\frac{cos\left(\beta\frac{\pi\tau}{T}\right)}{1-\left(\frac{2\beta\tau}{T}\right)^2} - \frac{\beta}{4}sinc\left(\beta\frac{\tau}{T}\right)\frac{cos\left(\frac{\pi\tau}{T}\right)}{1-\left(\frac{\beta\tau}{T}\right)^2}\right]$$

The auto-correlation result can be used to analyze various sampling offset results when analyzed with auto-correlation.

Application



Consecutive raised-cosine impulses, demonstrating zero-ISI property

When used to filter a symbol stream, a Nyquist filter has the property of eliminating ISI, as its impulse response is zero at all *nT* (where *n* is an integer), except *n* = 0.

Therefore, if the transmitted waveform is correctly sampled at the receiver, the original symbol values can be recovered completely.

However, in many practical communications systems, a matched filter is used in the receiver, due to the effects of white noise. For zero ISI, it is the net response of the transmit and receive filters that must equal *H*(*f*)**:**

$$H_R(f) \cdot H_T(f) = H(f)$$

And therefore**:**

$$|H_R(f)| = |H_T(f)| = \sqrt{|H(f)|}$$

These filters are called root-raised-cosine filters.

Chapter-13

Shannon–Hartley Theorem

In information theory, the **Shannon–Hartley theorem** (also known as **Shannon's law**) is an application of the noisy channel coding theorem to the archetypal case of a continuous-time analog communications channel subject to Gaussian noise. The theorem establishes Shannon's channel capacity for such a communication link, a bound on the maximum amount of error-free digital data (that is, information) that can be transmitted with a specified bandwidth in the presence of the noise interference, assuming (a) the signal power is bounded; (b)the Gaussian noise process is characterized by a known power or power spectral density. The law is named after Claude Shannon and Ralph Hartley.

Statement of the theorem

Considering all possible multi-level and multi-phase encoding techniques, the Shannon–Hartley theorem states the channel capacity *C*, meaning the theoretical tightest upper bound on the information rate (excluding error correcting codes) of clean (or arbitrarily low bit error rate) data that can be sent with a given average signal power *S* through an analog communication channel subject to additive white Gaussian noise of power *N*, is:

$$C = B \log_2 \left(1 + \frac{S}{N} \right)$$

where

> *C* is the channel capacity in bits per second;
> *B* is the bandwidth of the channel in hertz (passband bandwidth in case of a modulated signal);
> *S* is the total received signal power over the bandwidth (in case of a modulated signal, often denoted *C*, i.e. modulated carrier), measured in watt or volt^2;
> *N* is the total noise or interference power over the bandwidth, measured in watt or volt^2; and

S/N is the signal-to-noise ratio (SNR) or the carrier-to-noise ratio (CNR) of the communication signal to the Gaussian noise interference expressed as a linear power ratio (not as logarithmic decibels).

Historical development

During the late 1920s, Harry Nyquist and Ralph Hartley developed a handful of fundamental ideas related to the transmission of information, particularly in the context of the telegraph as a communications system. At the time, these concepts were powerful breakthroughs individually, but they were not part of a comprehensive theory. In the 1940s, Claude Shannon developed the concept of channel capacity, based in part on the ideas of Nyquist and Hartley, and then formulated a complete theory of information and its transmission.

Nyquist rate

In 1927, Nyquist determined that the number of independent pulses that could be put through a telegraph channel per unit time is limited to twice the bandwidth of the channel. In symbols,

$$f_p \leq 2B$$

where f_p is the pulse frequency (in pulses per second) and *B* is the bandwidth (in hertz). The quantity 2*B* later came to be called the *Nyquist rate*, and transmitting at the limiting pulse rate of 2*B* pulses per second as *signalling at the Nyquist rate*. Nyquist published his results in 1928 as part of his paper "Certain topics in Telegraph Transmission Theory."

Hartley's law

During that same year, Hartley formulated a way to quantify information and its line rate (also known as data signalling rate or gross bitrate inclusive of error-correcting code 'R' across a communications channel). This method, later known as Hartley's law, became an important precursor for Shannon's more sophisticated notion of channel capacity.

Hartley argued that the maximum number of distinct pulses that can be transmitted and received reliably over a communications channel is limited by the dynamic range of the signal amplitude and the precision with which the receiver can distinguish amplitude levels. Specifically, if the amplitude of the transmitted signal is restricted to the range of [–*A* ... +*A*] volts, and the precision of the receiver is ±Δ*V* volts, then the maximum number of distinct pulses *M* is given by

$$M = 1 + \frac{A}{\Delta V}.$$

By taking information per pulse in bit/pulse to be the base-2-logarithm of the number of distinct messages *M* that could be sent, Hartley constructed a measure of the line rate *R* as:

$$R = f_p \log_2(M),$$

where f_p is the pulse rate, also known as the symbol rate, in symbols/second or baud.

Hartley then combined the above quantification with Nyquist's observation that the number of independent pulses that could be put through a channel of bandwidth *B* hertz was 2*B* pulses per second, to arrive at his quantitative measure for achievable line rate.

Hartley's law is sometimes quoted as just a proportionality between the analog bandwidth, *B*, in Hertz and what today is called the digital bandwidth, *R*, in bit/s. Other times it is quoted in this more quantitative form, as an achievable line rate of *R* bits per second:

$$R \leq 2B \log_2(M).$$

Hartley did not work out exactly how the number *M* should depend on the noise statistics of the channel, or how the communication could be made reliable even when individual symbol pulses could not be reliably distinguished to *M* levels; with Gaussian noise statistics, system designers had to choose a very conservative value of *M* to achieve a low error rate.

The concept of an error-free capacity awaited Claude Shannon, who built on Hartley's observations about a logarithmic measure of information and Nyquist's observations about the effect of bandwidth limitations.

Hartley's rate result can be viewed as the capacity of an errorless *M*-ary channel of 2*B* symbols per second. Some authors refer to it as a capacity. But such an errorless channel is an idealization, and the result is necessarily less than the Shannon capacity of the noisy channel of bandwidth *B*, which is the Hartley–Shannon result that followed later.

Noisy channel coding theorem and capacity

Claude Shannon's development of information theory during World War II provided the next big step in understanding how much information could be reliably communicated through noisy channels. Building on Hartley's foundation, Shannon's noisy channel coding theorem (1948) describes the maximum possible efficiency of error-correcting methods versus levels of noise interference and data corruption. The proof of the theorem shows that a randomly constructed error correcting code is essentially as good as the best possible code; the theorem is proved through the statistics of such random codes.

Shannon's theorem shows how to compute a channel capacity from a statistical description of a channel, and establishes that given a noisy channel with capacity C and information transmitted at a line rate *R*, then if

$$R < C$$

there exists a coding technique which allows the probability of error at the receiver to be made arbitrarily small. This means that theoretically, it is possible to transmit information nearly without error up to nearly a limit of C bits per second.

The converse is also important. If

$$R > C$$

the probability of error at the receiver increases without bound as the rate is increased. So no useful information can be transmitted beyond the channel capacity. The theorem does not address the rare situation in which rate and capacity are equal.

Shannon–Hartley theorem

The Shannon–Hartley theorem establishes what that channel capacity is for a finite-bandwidth continuous-time channel subject to Gaussian noise. It connects Hartley's result with Shannon's channel capacity theorem in a form that is equivalent to specifying the *M* in Hartley's line rate formula in terms of a signal-to-noise ratio, but achieving reliability through error-correction coding rather than through reliably distinguishable pulse levels.

If there were such a thing as an infinite-bandwidth, noise-free analog channel, one could transmit unlimited amounts of error-free data over it per unit of time. Real channels, however, are subject to limitations imposed by both finite bandwidth and nonzero noise.

So how do bandwidth and noise affect the rate at which information can be transmitted over an analog channel?

Surprisingly, bandwidth limitations alone do not impose a cap on maximum information rate. This is because it is still possible for the signal to take on an indefinitely large number of different voltage levels on each symbol pulse, with each slightly different level being assigned a different meaning or bit sequence. If we combine both noise and bandwidth limitations, however, we do find there is a limit to the amount of information that can be transferred by a signal of a bounded power, even when clever multi-level encoding techniques are used.

In the channel considered by the Shannon-Hartley theorem, noise and signal are combined by addition. That is, the receiver measures a signal that is equal to the sum of the signal encoding the desired information and a continuous random variable that represents the noise. This addition creates uncertainty as to the original signal's value. If the receiver has some information about the random process that generates the noise, one can in principle recover the information in the original signal by considering all possible states of the noise process. In the case of the Shannon-Hartley theorem, the noise is assumed to be generated by a Gaussian process with a known variance. Since the variance of a Gaussian process is equivalent to its power, it is conventional to call this variance the noise power.

Such a channel is called the Additive White Gaussian Noise channel, because Gaussian noise is added to the signal; "white" means equal amounts of noise at all frequencies within the channel bandwidth. Such noise can arise both from random sources of energy and also from coding and measurement error at the sender and receiver respectively. Since sums of independent Gaussian random variables are themselves Gaussian random variables, this conveniently simplifies analysis, if one assumes that such error sources are also Gaussian and independent.

Implications of the theorem

Comparison of Shannon's capacity to Hartley's law

Comparing the channel capacity to the information rate from Hartley's law, we can find the effective number of distinguishable levels *M*:

$$2B\log_2(M) = B\log_2\left(1+\frac{S}{N}\right)$$
$$M = \sqrt{1+\frac{S}{N}}.$$

The square root effectively converts the power ratio back to a voltage ratio, so the number of levels is approximately proportional to the ratio of rms signal amplitude to noise standard deviation.

This similarity in form between Shannon's capacity and Hartley's law should not be interpreted to mean that *M* pulse levels can be literally sent without any confusion; more levels are needed, to allow for redundant coding and error correction, but the net data rate that can be approached with coding is equivalent to using that *M* in Hartley's law.

Alternative forms

Frequency-dependent (colored noise) case

In the simple version above, the signal and noise are fully uncorrelated, in which case *S* + *N* is the total power of the received signal and noise together. A generalization of the above equation for the case where the additive noise is not white (or that the S/N is not constant with frequency over the bandwidth) is obtained by treating the channel as many narrow, independent Gaussian channels in parallel:

$$C = \int_0^B \log_2 \left(1 + \frac{S(f)}{N(f)}\right) df$$

where

C is the channel capacity in bits per second;
B is the bandwidth of the channel in Hz;
S(f) is the signal power spectrum
N(f) is the noise power spectrum
f is frequency in Hz.

Note: the theorem only applies to Gaussian stationary process noise. This formula's way of introducing frequency-dependent noise cannot describe all continuous-time noise processes. For example, consider a noise process consisting of adding a random wave whose amplitude is 1 or -1 at any point in time, and a channel that adds such a wave to the source signal. Such a wave's frequency components are highly dependent. Though such a noise may have a high power, it is fairly easy to transmit a continuous signal with much less power than one would need if the underlying noise was a sum of independent noises in each frequency band.

Approximations

For large or small and constant signal-to-noise ratios, the capacity formula can be approximated:

- If S/N >> 1, then

  $$C \approx 0.332 \cdot B \cdot \mathrm{SNR\ (in\ dB)}$$

  where

  $$\mathrm{SNR\ (in\ dB)} = 10 \log_{10} \frac{S}{N}.$$

- Similarly, if S/N << 1, then

  $$C \approx 1.44 \cdot B \cdot \frac{S}{N}.$$

  In this low-SNR approximation, capacity is independent of bandwidth if the noise is white, of spectral density N_0 watts per hertz, in which case the total noise power is $B \cdot N_0$.

  $$C \approx 1.44 \cdot \frac{S}{N_0}$$

Examples

1. If the SNR is 20 dB, and the bandwidth available is 4 kHz, which is appropriate for telephone communications, then C = 4 $\log_2(1 + 100)$ = 4 $\log_2$ (101) = 26.63 kbit/s. Note that the value of S/N = 100 is equivalent to the SNR of 20 dB.
2. If the requirement is to transmit at 50 kbit/s, and a bandwidth of 1 MHz is used, then the minimum S/N required is given by 50 = 1000 $\log_2(1+S/N)$ so S/N = $2^{C/B}$ - 1 = 0.035, corresponding to an SNR of -14.5 dB (10 x $\log_{10}(0.035)$).
3. Let's take the example of W-CDMA (Wideband Code Division Multiple Access), the bandwidth = 5 MHz, you want to carry 12.2 kbps of data (AMR voice), then the required SNR is given by $2^{12.2/5000}$ -1 corresponding to an SNR of -27.7 dB for a single channel. This shows that it is possible to transmit using signals which are actually much weaker than the background noise level, as in spread-spectrum communications. However, in W-CDMA the required SNR will vary based on design calculations.
4. There is another interesting fact. It is stated above that channel capacity is proportional to the bandwidth of the channel and to the logarithm of SNR. This means channel capacity can be increased linearly by increasing bandwidth of the channel with same SNR requirement OR with fixed bandwidth, one should invoke higher order modulations that need very high SNR to operate. As one moves to the higher modulation rate, the spectral efficiency also improves but on cost of SNR requirement. eg: there is an exponential rise in SNR requirement if one adopts a 16QAM or 64QAM (see: Quadrature amplitude modulation) however the spectral efficiency is improved.

Chapter-14

Signal-to-noise Ratio

Signal-to-noise ratio (often abbreviated **SNR** or **S/N**) is a measure used in science and engineering to quantify how much a signal has been corrupted by noise. It is defined as the ratio of signal power to the noise power corrupting the signal. A ratio higher than 1:1 indicates more signal than noise. While SNR is commonly quoted for electrical signals, it can be applied to any form of signal (such as isotope levels in an ice core or biochemical signaling between cells).

In less technical terms, signal-to-noise ratio compares the level of a desired signal (such as music) to the level of background noise. The higher the ratio, the less obtrusive the background noise is.

"Signal-to-noise ratio" is sometimes used informally to refer to the ratio of useful information to false or irrelevant data in a conversation or exchange. For example, in online discussion forums and other online communities, off-topic posts and spam are regarded as "noise" that interferes with the "signal" of appropriate discussion.

Definition

Signal-to-noise ratio is defined as the power ratio between a signal (meaningful information) and the background noise (unwanted signal):

$$\mathrm{SNR} = \frac{P_\mathrm{signal}}{P_\mathrm{noise}},$$

where *P* is average power. Both signal and noise power must be measured at the same or equivalent points in a system, and within the same system bandwidth. If the signal and the noise are measured across the same impedance, then the SNR can be obtained by calculating the square of the amplitude ratio:

$$\mathrm{SNR} = \frac{P_\mathrm{signal}}{P_\mathrm{noise}} = \left(\frac{A_\mathrm{signal}}{A_\mathrm{noise}}\right)^2,$$

where *A* is root mean square (RMS) amplitude (for example, RMS voltage). Because many signals have a very wide dynamic range, SNRs are often expressed using the logarithmic decibel scale. In decibels, the SNR is defined as

$$\mathrm{SNR_{dB}} = 10\log_{10}\left(\frac{P_\mathrm{signal}}{P_\mathrm{noise}}\right) = P_\mathrm{signal,dB} - P_\mathrm{noise,dB},$$

which may equivalently be written using amplitude ratios as

$$\mathrm{SNR_{dB}} = 10\log_{10}\left(\frac{A_\mathrm{signal}}{A_\mathrm{noise}}\right)^2 = 20\log_{10}\left(\frac{A_\mathrm{signal}}{A_\mathrm{noise}}\right).$$

The concepts of signal-to-noise ratio and dynamic range are closely related. Dynamic range measures the ratio between the strongest un-distorted signal on a channel and the minimum discernable signal, which for most purposes is the noise level. SNR measures the ratio between an arbitrary signal level (not necessarily the most powerful signal possible) and noise. Measuring signal-to-noise ratios requires the selection of a representative or *reference* signal. In audio engineering, the reference signal is usually a sine wave at a standardized nominal or alignment level, such as 1 kHz at +4 dBu (1.228 $\mathrm{V_{RMS}}$).

SNR is usually taken to indicate an *average* signal-to-noise ratio, as it is possible that (near) instantaneous signal-to-noise ratios will be considerably different. The concept can be understood as normalizing the noise level to 1 (0 dB) and measuring how far the signal 'stands out'.

Alternative definition

An alternative definition of SNR is as the reciprocal of the coefficient of variation, i.e., the ratio of mean to standard deviation of a signal or measurement:

$$\mathrm{SNR} = \frac{\mu}{\sigma}$$

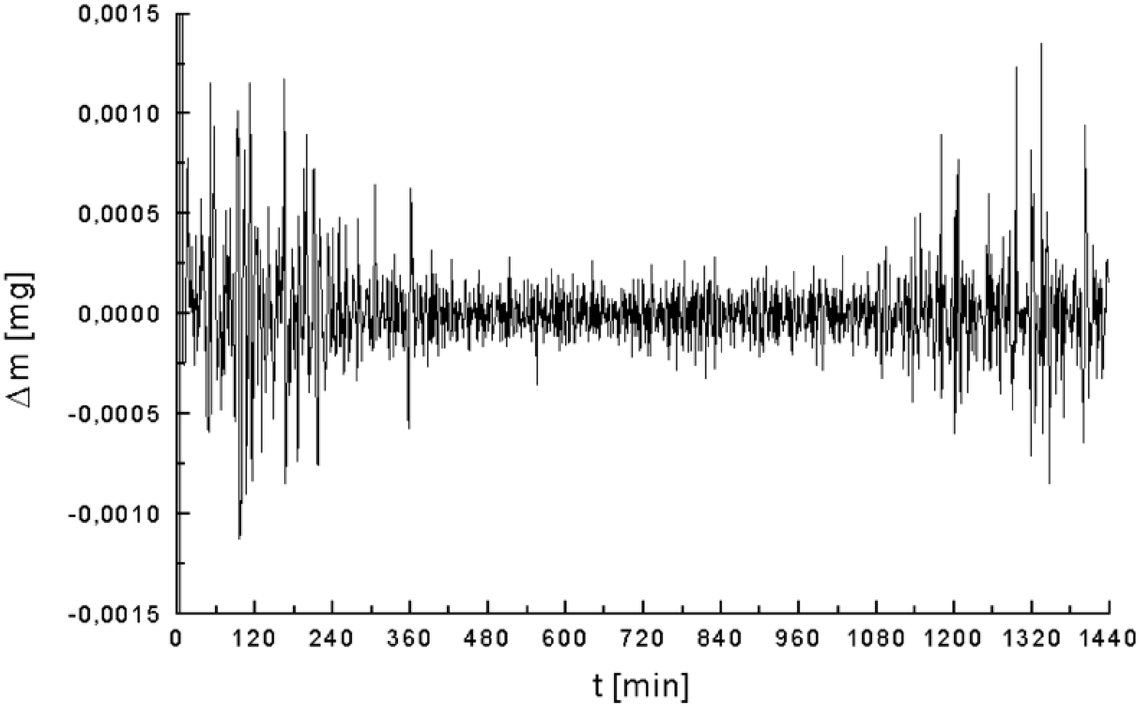
where μ is the signal mean or expected value and σ is the standard deviation of the noise, or an estimate thereof. Notice that such an alternative definition is only useful for variables that are always positive (such as photon counts and luminance). Thus it is commonly used in image processing, where the SNR of an image is usually calculated as the ratio of the mean pixel value to the standard deviation of the pixel values over a given neighborhood. Sometimes SNR is defined as the square of the alternative definition above.

The *Rose criterion* (named after Albert Rose) states that an SNR of at least 5 is needed to be able to distinguish image features at 100% certainty. An SNR less than 5 means less than 100% certainty in identifying image details.

Yet another alternative, very specific and distinct definition of SNR is employed to characterize sensitivity of imaging systems.

Related measures are the "contrast ratio" and the "contrast-to-noise ratio".

Improving SNR in practice



Recording of the noise of a thermogravimetric analysis device that is poorly isolated from a mechanical point of view; the middle of the curve shows a lower noise, due to a lesser surrounding human activity at night.

All real measurements are disturbed by noise. This includes electronic noise, but can also include external events that affect the measured phenomenon — wind, vibrations, gravitational attraction of the moon, variations of temperature, variations of humidity, etc., depending on what is measured and of the sensitivity of the device. It is often possible to reduce the noise by controlling the environment. Otherwise, when the characteristics of the noise are known and are different from the signals, it is possible to filter it or to process the signal. When the signal is constant or periodic and the noise is random, it is possible to enhance the SNR by averaging the measurement.

Digital signals

When a measurement is digitised, the number of bits used to represent the measurement determines the maximum possible signal-to-noise ratio. This is because the minimum possible noise level is the error caused by the quantization of the signal, sometimes called

Quantization noise. This noise level is non-linear and signal-dependent; different calculations exist for different signal models. Quantization noise is modeled as an analog error signal summed with the signal before quantization ("additive noise").

This theoretical maximum SNR assumes a perfect input signal. If the input signal is already noisy (as is usually the case), the signal's noise may be larger than the quantization noise. Real analog-to-digital converters also have other sources of noise that further decrease the SNR compared to the theoretical maximum from the idealized quantization noise, including the intentional addition of dither.

Although noise levels in a digital system can be expressed using SNR, it is more common to use E_b/N_o, the energy per bit per noise power spectral density.

The modulation error ratio (MER) is a measure of the SNR in a digitally modulated signal.

Fixed point

For *n*-bit integers with equal distance between quantization levels (uniform quantization) the dynamic range (DR) is also determined.

Assuming a uniform distribution of input signal values, the quantization noise is a uniformly-distributed random signal with a peak-to-peak amplitude of one quantization level, making the amplitude ratio $2^n/1$. The formula is then:

$$\mathrm{DR_{dB}} = \mathrm{SNR_{dB}} = 20\log_{10}(2^n) \approx 6.02 \cdot n$$

This relationship is the origin of statements like "16-bit audio has a dynamic range of 96 dB". Each extra quantization bit increases the dynamic range by roughly 6 dB.

Assuming a full-scale sine wave signal (that is, the quantizer is designed such that it has the same minimum and maximum values as the input signal), the quantization noise approximates a sawtooth wave with peak-to-peak amplitude of one quantization level and uniform distribution. In this case, the SNR is approximately

$$\mathrm{SNR_{dB}} \approx 20\log_{10}(2^n\sqrt{3/2}) \approx 6.02 \cdot n + 1.761$$

Floating point

Floating-point numbers provide a way to trade off signal-to-noise ratio for an increase in dynamic range. For n bit floating-point numbers, with n-m bits in the mantissa and m bits in the exponent:

$$\mathrm{DR_{dB}} = 6.02 \cdot 2^m$$
$$\mathrm{SNR_{dB}} = 6.02 \cdot (n - m)$$

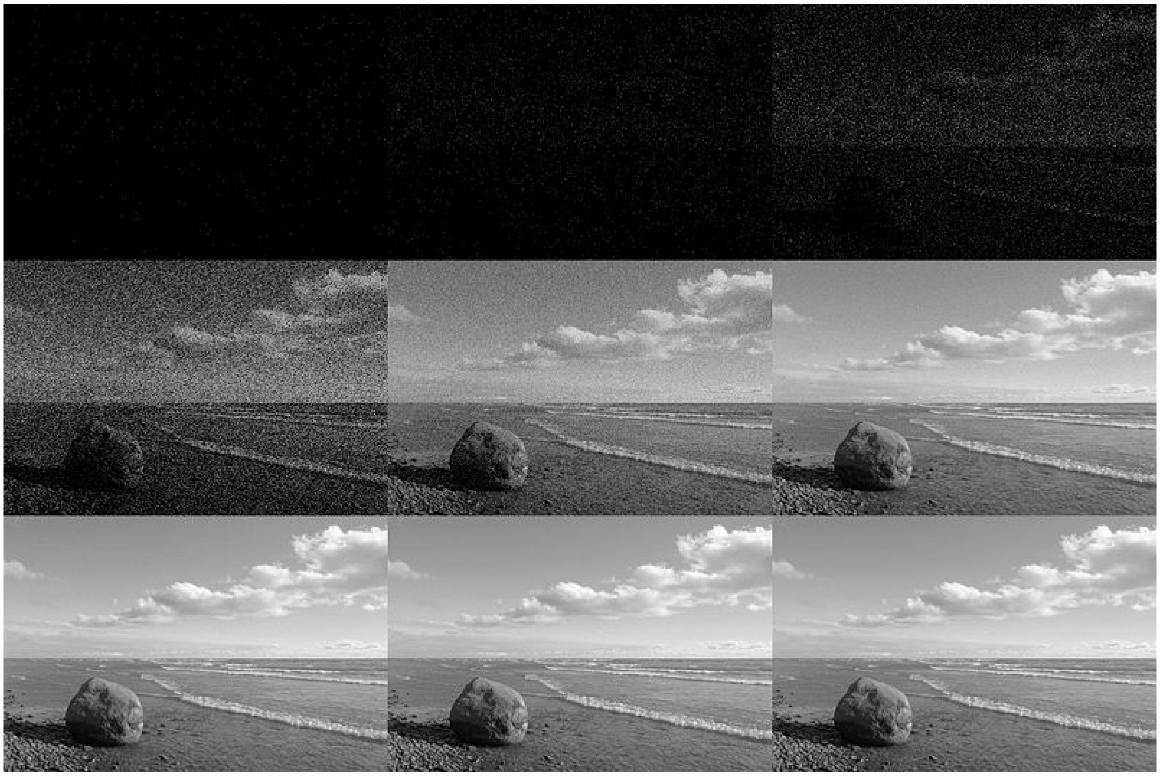
Note that the dynamic range is much larger than fixed-point, but at a cost of a worse signal-to-noise ratio. This makes floating-point preferable in situations where the dynamic range is large or unpredictable. Fixed-point's simpler implementations can be used with no signal quality disadvantage in systems where dynamic range is less than 6.02m. The very large dynamic range of floating-point can be a disadvantage, since it requires more forethought in designing algorithms.

Optical SNR

Optical signals have a carrier frequency, which is much higher than the modulation frequency (about 200 THz and more). This way the noise bandwidth covers a bandwidth which is much wider than the signal itself. The resulting signal influence relies mainly on the filtering of the noise. To describe the signal quality without taking the receiver into account the optical SNR (OSNR) is used. The OSNR is the ratio between the signal power and the noise power in a given bandwidth. Most commonly a reference bandwidth of 0.1 nm is used. This bandwidth is independent from the modulation format, the frequency and the receiver. For instance a OSNR of 20dB/0.1nm could be given, even the signal of 40 GBit DPSK would not fit in this bandwidth. OSNR is measured with a Optical Spectrum Analyzer

Chapter-15

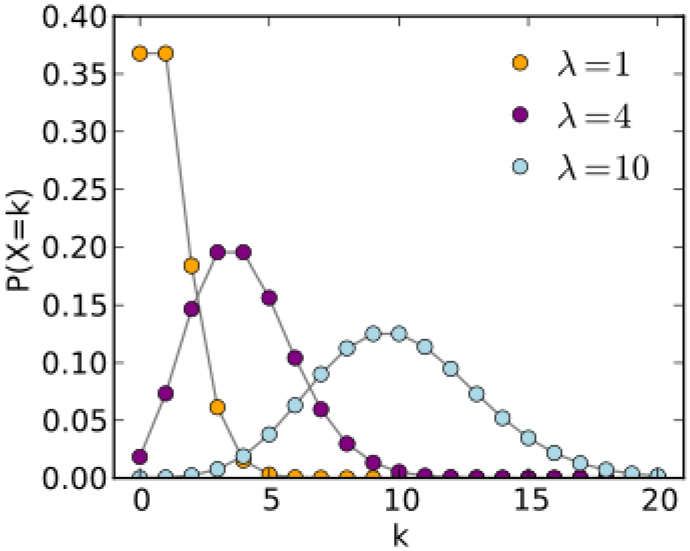
Shot Noise



Photon noise simulation.

Shot noise is a type of electronic noise that occurs when the finite number of particles that carry energy (such as electrons in an electronic circuit or photons in an optical device) is small enough to give rise to detectable statistical fluctuations in a measurement. It is important in electronics, telecommunications, and fundamental physics.

It also refers to an analogous noise in particle simulations, where due to the small number of particles, the simulation exhibits detectable statistical fluctuations not observed in the real-world system. Magnitude of this noise increases with the average magnitude of the current or intensity of the light. However, since the magnitude of the average signal increases more rapidly than that of the shot noise (its relative strength decreases with increasing signal), shot noise is often only a problem with small currents or light intensities.



The number of photons that are collected by a given detector varies, and follows a Poisson distribution, depicted here for averages of 1, 4, and 10.

The intensity of a source will yield the *average* number of photons collected, but knowing the average number of photons which will be collected will not give the actual number collected. The actual number collected will be more than, equal to, or less than the average, and their distribution about that average will be a Poisson distribution.

Since the Poisson distribution approaches a normal distribution for large numbers, the photon noise in a signal will approach a normal distribution for large numbers of photons collected. The standard deviation of the photon noise is equal to the square root of the average number of photons. The signal-to-noise ratio is then

$$\mathrm{SNR} = \frac{N}{\sqrt{N}} = \sqrt{N}$$

where *N* is the average number of photons collected. When *N* is very large, the signal-to-noise ratio is very large as well. It can be seen that photon noise becomes more important when the number of photons collected is small.

Explanation

Intuitive explanation

It is known to everyone that in a statistical experiment such as tossing a fair coin and counting the occurrences of heads and tails, the numbers of heads and tails after a great many throws will differ by only a tiny percentage, while after only a few throws outcomes with a significant excess of heads over tails or vice versa are common; if an experiment with a few throws is repeated over and over, the outcomes will fluctuate a lot. (It can be proved that the relative fluctuations reduce as the square root of the number of throws, a result valid for all statistical fluctuations, including shot noise.)

Shot noise exists because phenomena such as light and electric current consist of the movement of discrete, quantized 'packets'. Consider light—a stream of discrete photons—coming out of a laser pointer and hitting a wall to create a visible spot. The fundamental physical processes that govern light emission are such that these photons are emitted from the laser at random times; but the many billions of photons needed to create a spot are so many that the brightness, the number of photons per unit time, varies only infinitesimally with time. However, if the laser brightness is reduced until only a handful of photons hit the wall every second, the relative fluctuations in number of photons, i.e. brightness, will be significant, just as when tossing a coin a few times. These fluctuations are shot noise.

In electronic devices

Shot noise in electronic devices consists of random fluctuations of the electric current in many electrical conductors, due to the current being carried by discrete charges (electrons) whose number per unit time fluctuates. This is often an issue in p-n junctions. In metal wires this is not an issue, as correlations between individual electrons remove these random fluctuations.

Shot noise is distinct from current fluctuations in thermal equilibrium, which happen without any applied voltage and without any average current flowing. These thermal equilibrium current fluctuations are known as Johnson-Nyquist noise or thermal noise.

Shot noise is a Poisson process and the charge carriers which make up the current will follow a Poisson distribution. The current fluctuations have a standard deviation of

$$\sigma_i = \sqrt{2\,q\,I\,\Delta f}$$

where *q* is the elementary charge, Δ*f* is the bandwidth in hertz over which the noise is measured, and *I* is the average current through the device. All quantities are assumed to be in SI units.

For a current of 100 mA this gives a value of

$$\sigma_i = 0.18\,\mathrm{nA}$$

if the noise current is filtered with a filter having a bandwidth of 1 Hz.

If this noise current is fed through a resistor the resulting noise power will be

$$P = 2\,q\,I\,\Delta f R.$$

If the charge is not fully localized in time but has a temporal distribution of $q\,F(t)$ where the integral of $F(t)$ over t is unity then the power spectral density of the noise current signal will be,

$$S_i(f) = 2\,q\,I\,|\Psi(f)|^2,$$

where $\Psi(f)$ is the Fourier transform of $F(t)$.

Note: Shot noise and Johnson–Nyquist noise are both quantum fluctuations. Some authors treat them as a single unified concept.

In quantum optics

In quantum optics, shot noise is caused by the fluctuations of detected photons, again therefore a consequence of discretization (of the energy in the electromagnetic field in this case). Shot noise is a main part of **quantum noise**.

Shot noise is measurable not only in measurements at the few-photons level using photomultipliers, but also at stronger light intensities measured by photodiodes when using high temporal resolution oscilloscopes. As the photocurrent is proportional to the light intensity (number of photons), the fluctuations of the electromagnetic field are usually contained in the electric current measured.

In the case of a *coherent light* source such as a laser, the shot noise scales as the square-root of the average intensity:

$$\Delta I^2 \overset{\mathrm{def}}{=} \langle (I - \langle I \rangle)^2 \rangle \propto I.$$

A similar lower bound of quantum noise occurs in linear quantum amplifiers. The only exception being if a squeezed coherent state can be formed through correlated photon generation. The reduction of uncertainty of the number of photons per mode (and therefore the photocurrent) may take place just due to the saturation of gain; this is intermediate case between a laser with locked phase and amplitude-stabilized laser.

Space charge

Low noise active electronic devices are designed such that shot noise is suppressed by the electrostatic repulsion of the charge carriers. Space charge limiting is not possible in photon devices.