Digital Modes for Weak-Signal Communications

Presentation at the VERON VHF-Day 2007 in Dwingeloo

Klaus von der Heide, DJ5HG

Contents

1. Introduction: Digital Modes vs. Analog Modes

1.1 Analog Modes

In an analog mode a time-dependent physical value, i.e. the voltage of a microphone, is directly modulated on a carrier wave and demodulated at the receiving end resulting in a voltage that is more or less the same as that transmitted. With increasing noise on the radio path the quality of the decoded signal degrades continuously. Figure 1 shows the classic amplitude modulation as an example. The upper curve is the signal of the microphone, the lower curve is the signal that is reconstructed from the amplitude modulated wave in the middle by demodulation.

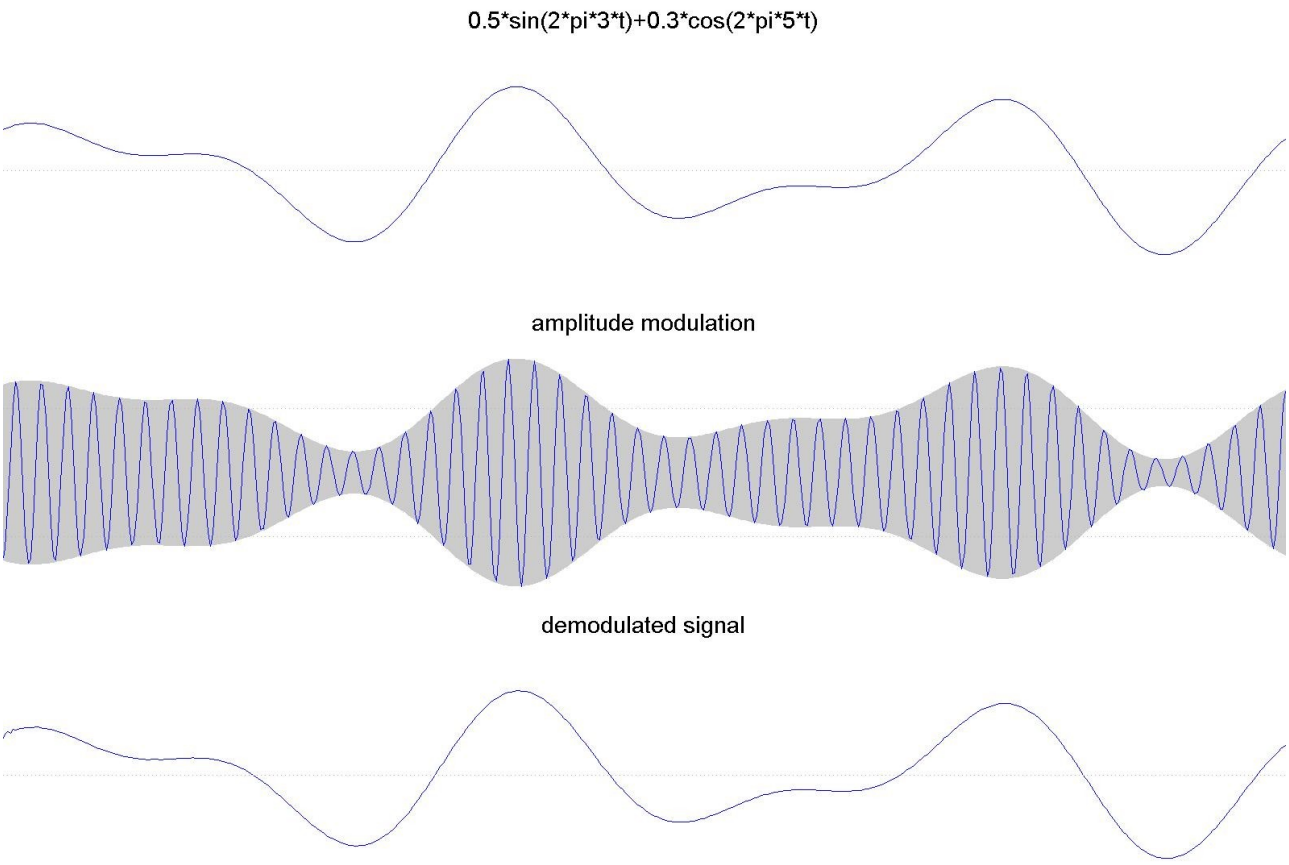


Figure 1. Amplitude modulation (AM) is a simple example for an analog transmission mode. The upper curve shows the time-dependent voltage of the audio signal, the curve in the middle is the modulated wave. For better readability the envelope of the modulated signal is shown grey. The lower curve is what a good demodulator gets out of the modulated signal. If the received signal is noisy then also the demodulated signal is corrupted by noise.

1.2 Digital Modes

Digital modes use a limited set of predefined symbols that are known at both ends of the transmission path. Furthermore, the information must be coded into a sequence of these symbols. A good example for a digital transmission is a written text that uses the latin alphabet, the decimal digits, and some special symbols. The reader must know the alphabet and the complex rules of a natural language to code some abstract information into an unambiguous sentence.

Figure 2 shows a typical analog signal of a digital transmission in baseband (i.e. unmodulated). Figure 3 shows two modulated binary signals (ASK and FSK).

While with the square wave pulses in the examples the difference between analog and digital modes seems to be obvious, this is not the case if more complex forms of the symbols are used. An example is the use of symbols that are the audio signals of the spoken alphabet „alpha“, „bravo“, ... A computer can easily translate a textually written information into a sequence of these symbols. If transmitted via FM for example, a listener cannot detect that this originally is a digital transmission. Indeed, the transmission from the transmitter to him is an analog one while it may be a digital one to others.

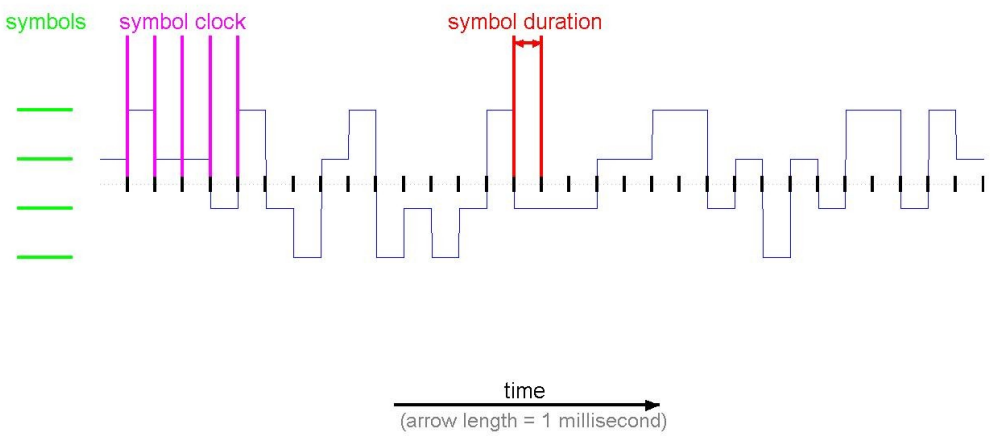


Figure 2. A typical digital transmission that uses 4 different symbols and a fixed symbol clock. Square waves are used in this simple case to generate the analog signal. There is a great variety of other pulse forms (mostly used to narrow the spectrum). In the case of a radio path this base band signal has to be modulated.

Generally, it is very important to realize that many parameters of a transmission path are determined at the receiving end. R. W. Hamming in his book „Coding and Information Theory“ says:
... we must learn to look at the system from the receiving end, where the decisions are to be made, instead of looking at the front end and following what happens to the signal as it goes through the system.

In practise natural language does not fit well to the demands of digital communication. It is only used for demonstration here.

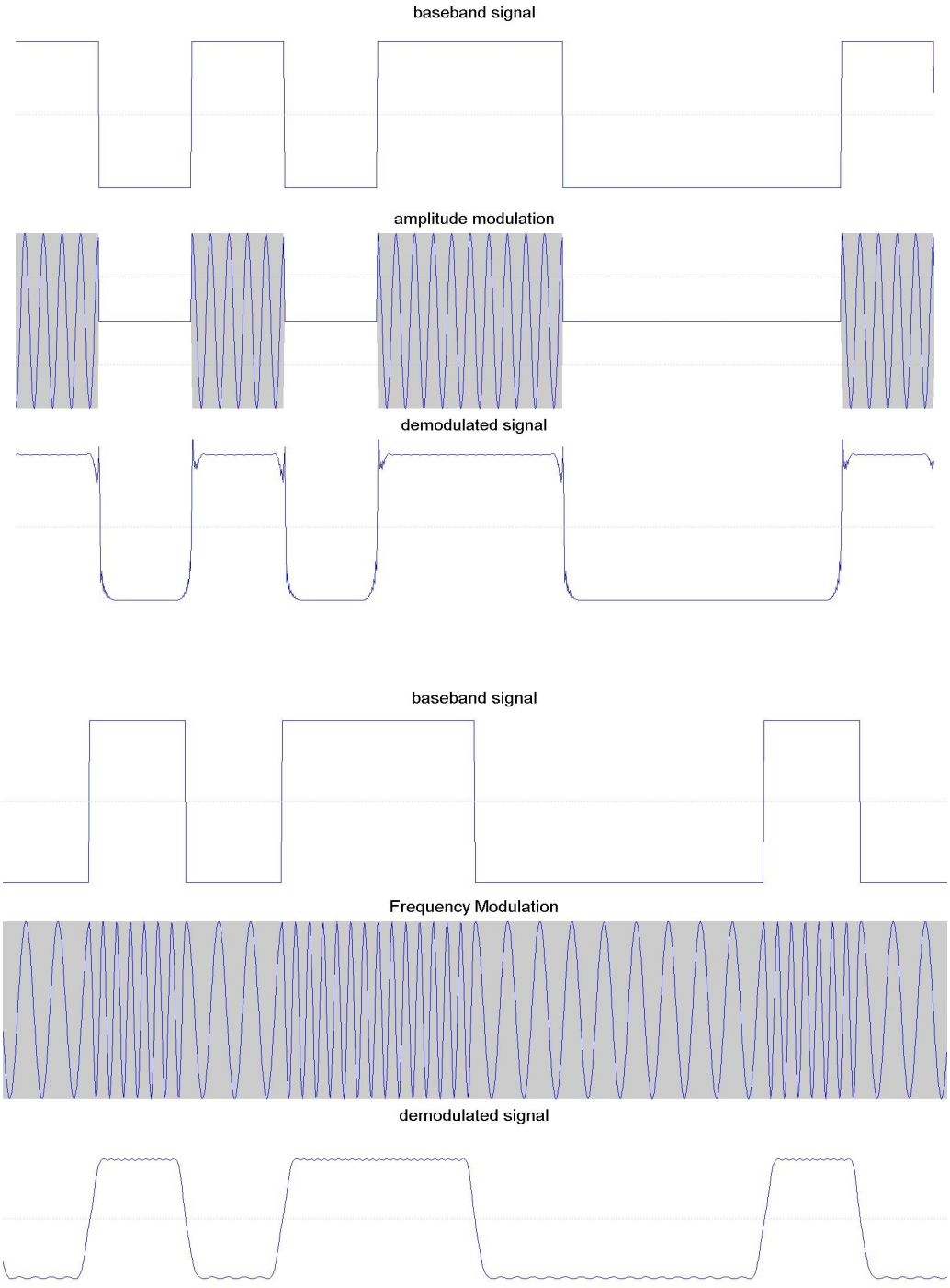


Figure 3. The upper curve shows Amplitude Shift Keying (ASK) of a binary sequence using a square wave pulse. The lower curve demonstrates Frequency Shift Keying (FSK), also with square wave pulses.

2. Technical Aspects of Weak Signal Modes

2.1. Additive White Gaussian Noise

On VHF and above a very good model of the noise is the Additive White Gaussian Noise (AWGN). Additive means that it does not distort the signal but simply is added in the antenna. The power spectral density of the AWGN (the power per Hz of bandwidth) is

$$\mathbf{N_0} = \boldsymbol{k}_\mathbf{B} \cdot \mathbf{T}$$

$\boldsymbol{k}_\mathbf{B} = \mathbf{1.38*10^{-23}\ W / (K \cdot Hz)}$ is the Boltzmann-constant, and **T** is the equivalent noise temperature in degrees Kelvin. The noise temperature varies with frequency, but it is assumed constant within the received bandwidth (that is the meaning of the term "white"). On 2m, **T** lies between 200 K and 800 K. It especially depends on which part of our galaxy radiates into the antenna. At zero elevation also the earth surface contributes (and man-made wideband noise).

Let **w** be the bandwidth of a receiver then the received noise power is

$$\mathbf{P_n = w \cdot N_0}$$

2.2. The Shannon-Limit

C. E. Shannon published a famous paper in which he proved that there is a limit for the rate of information transfer depending on the signal to noise ratio. Usually this limit ist written as

$$\mathbf{E_b / N_0 \geq ln\ 2}$$

$\mathbf{E_b}$ is the energy neccessary to receive one information bit. The same in dB:

$$\mathbf{E_b / N_0 \geq 10 \cdot \log_{10} (ln\ 2) = -1.5917\ dB}$$

Shannon concluded that the bit error rate theoretically could be made arbitrarily low by the use of some good error correcting codes (which he did not know!) if $\mathbf{E_b}$ is above the limit. Since then every communication system is evaluated by its value of $\mathbf{E_b / N_0}$ that is neccessary to reach a standard word error rate (usually 10^{-4}).

Unfortunately, the low limit only is valid for infinite amount of information. Figure 4 shows what can be reached in practise. Deep space communication with modern turbo codes reaches the limit within less than 2 dB. But a minimal QSO does not transfer thousands of bits in a single pass. That is the reason for the fact that the best weak signal mode for ham radio in principle looses about 4 dB compared to deep space transmission. Fortunately, there is no demand for such a low error rate as 0.0001 in a weak signal mode. In ham radio weak signal QSOs the error rate generally is above 0.1 . With this value the loss is nearly compensated.

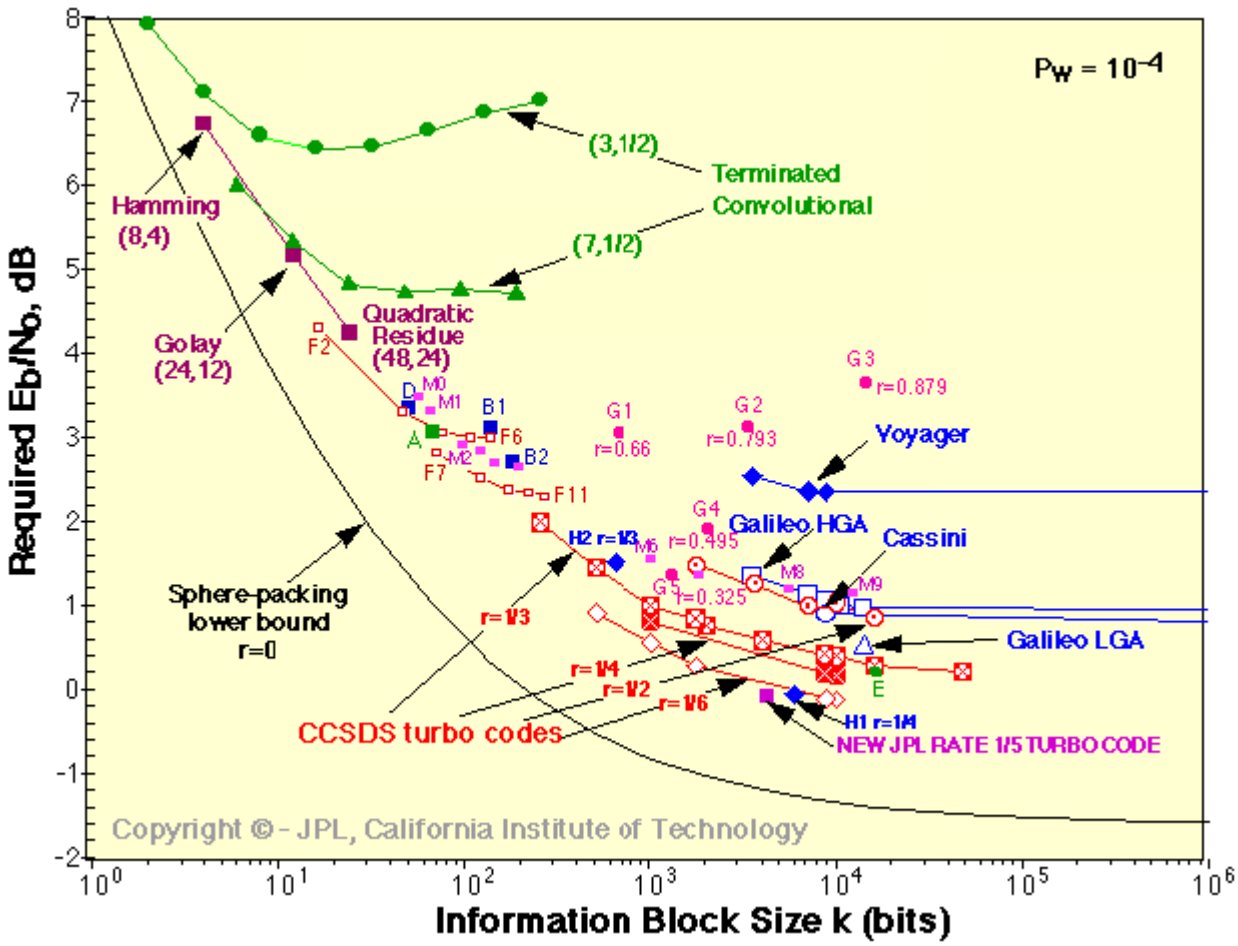


Figure 4. Practical values of E_b / N_0 for some codes to get a word error rate of 0.0001 . The so called sphere-packing lower bound is a lower limit for E_b / N_0 in case of finite information block size (i.e. the number of information bits in a single transmission). It reaches the shannon limit asymptotically to the right of this figure.

2.3. Weak Signal QSOs

In ham radio, a weak signal QSO not only means that it runs near the Shannon limit. It furthermore is so weak that the neccessary energy per information bit only is reached at a very low bit rate:

$$E_b = P_s \cdot t_b$$

where P_s is the received signal power and t_b is the time to get a bit. E_b can be increased to the neccessary value by increasing P_s or by increasing t_b . The signal power is determined by the transmitted power, the antenna gains at both ends, and the channel attenuation. Usually, an amateur station is run at a given limit. So P_s is fixed. Increasing t_b on the other hand has the consequence that the transfer of information may be very slow. This also demands for a severe reduction of the transmitted information. This raises the question what is the minimum of the information that has mutually to be exchanged in a contact to be accepted as a valid QSO.

We speak of a weak signal when t_b is of the order of one second. From the formulas above and with the assumptions $t_b = 1$ s and $T = 400$ K we get

$$P_s \geq 3.83 \cdot 10^{-21} \text{ W}$$

So on 144 MHz the signal power must be of the order of 10^{-20} W.

2.4. The Information Content

The information block size or information content of a message is the least number of bits that a message can be compressed. If the receiver knows that the received message is one out of **n** known messages then the information content is $\log_2 n$. Since there are less than 4 million licensed radio amateurs in the world the information content of a callsign is not larger than 22 bits. On the other hand, if letters, decimal digits, and the blank are used to write down a callsign of length 6 then there are $(26+10+1)^6 = 2565726409$ possibilities which correspond to 32 bits.

Using a more restrictive rule to generate callsigns it is easy to find a coding scheme that only needs 28 bits:

position	symbol	number of symbols
1	digit or letter or blank	37
2	digit or letter	36
3	digit	10
4	letter	26
5	letter oder blank	27
6	letter oder blank	27

The total number of possible calls following this rule is the product of all numbers of symbols in the right column: `#calls = 252467280`. The base-2-logarithm of this number is the number of bits neccessary to code callsigns by this rule: $\log_2(37*36*10*26*27*27) = 27.912$.

It is important to realize that any information that does not satisfy the rules used for this source coding cannot be transmitted this way. If callsigns are encoded into 28 bits then „**HA2004EU“** and „**DJ5HG/P**“ cannot be encoded. Indeed, JT65 uses such a source code, but it provides additional encodings to allow for special callsigns.

It also is very important to realize that optimized source coding has the consequence that every garbage will be decoded into well-looking information. That is not a fault of the digital mode, but believing such result is a fault of the operator. The digital mode should provide confidence values.

2.4. Modulation and Demodulation

The modulated symbols used to code the information should mutually be as different as possible. Figure 5 shows ASK, PSK, and FSK for the binary case and rectangular pulse shape, all for the same mean transmitted energy. PSK is better than the other modulations by 3 dB. This only is true for binary transmission. The distance of m-PSK rapidly decreases with increasing m while the distance of m-FSK remains constant. Figure 6 plots the symbol error of m-FSK over $\mathbf{E_b / N_0}$. Near to the Shannon limit binary PSK remains the best mode. But as was pointed out above, a digital mode may not come so near to the limit. Therefore m-FSK is a good choice for weak signals if bandwidth is not an issue.

A coherent demodulation gains 3 dB over an incoherent demodulation. A coherent demodulation means that the phase of the carrierwave is known, at least approximately. Since the channel may disturb the phase it usually has to be reconstructed from the incoming signal. This is a severe problem in weak signal applications. If the rate of the phase changes is slow compared to the information bit rate as is mostly true for 2m EME then the phase can be reconstructed, and nearly

all of the 3 dB of coherent decoding can be won. If on the other hand the rate of the phase changes is of the order of the information bit rate or higher then it is not possible to get the information on the phase in addition to the message information. The demodulation then must be incoherent. This is the case with EME on the GHz-bands.

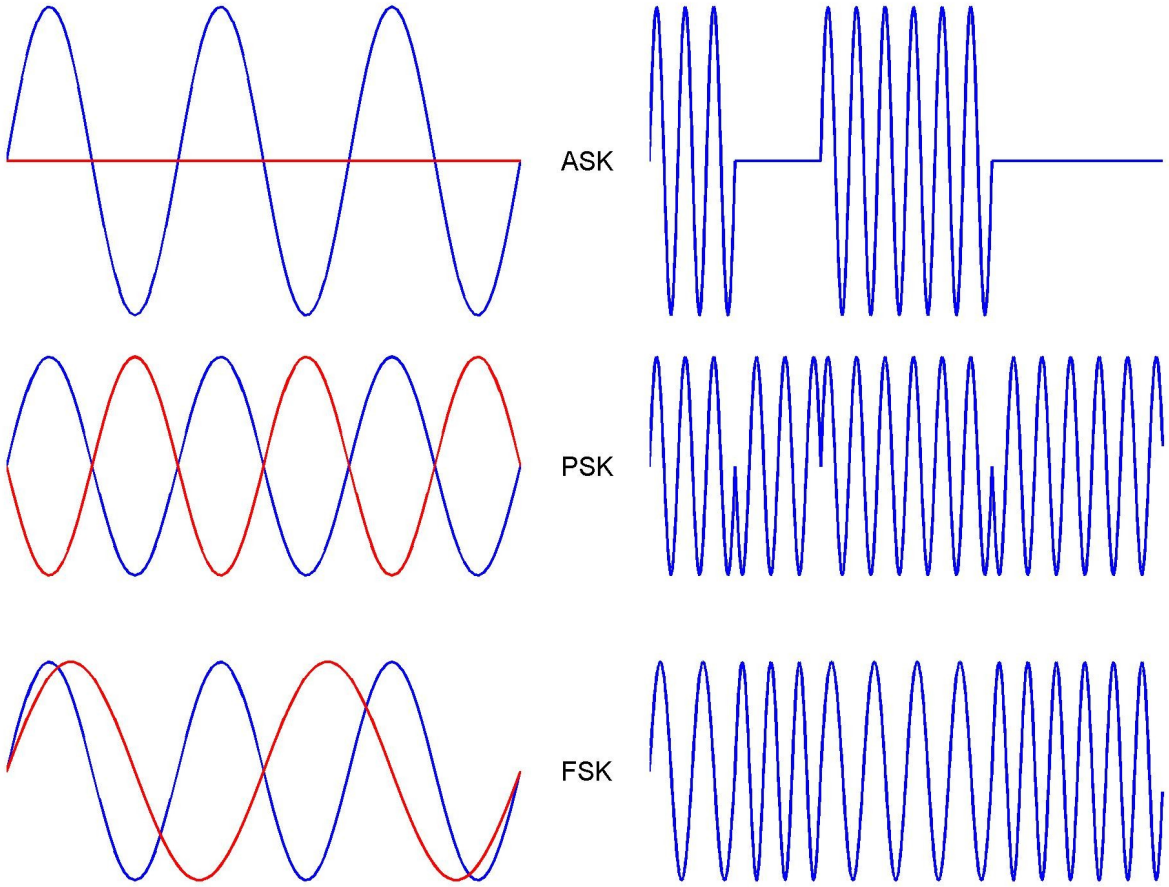


Figure 5. A rectangular pulse binary modulated by ASK, PSK, and FSK (left). The blue line is the signal transmitting a binary 1, the red line transmitting a binary 0. The amplitude of the ASK signal is greater by a factor 1.414 to guarantee the same mean transmitted power. The distance of the two signals is the squareroot of the mean of squares of all vertical distances between the blue line and the red line. The distances of ASK and FSK are 1.0, and that of PSK is 1.414 . So PSK gains 3 dB over the other modulations.

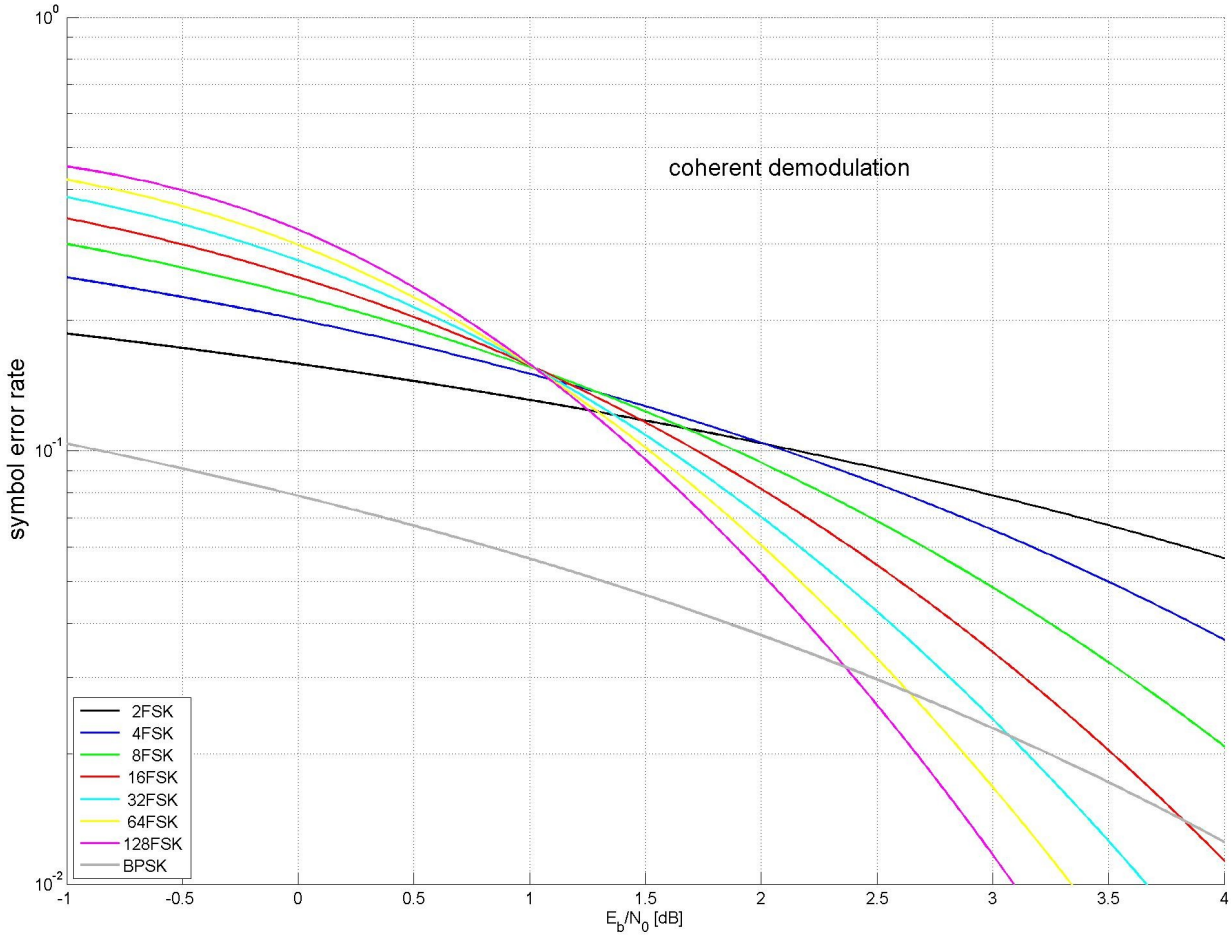


Figure 6. Symbol error rates for m-FSK compared to binary PSK for coherent demodulation. Incoherent demodulation is shifted to the right by 3 dB.

2.6. Detection of Symbols by Correlation

The theoretically optimal method of detecting a symbol in Additive White Gaussian Noise (AWGN) is correlation. A correlator is a filter with an impulse response equal to the shape of the symbol but mirrored in time. Figure 7 shows three examples.

Correlation is not restricted to simple symbols. Figure 8 shows the correlation of two very noisy audiosignals that contain the spoken sequence **"CQ Delta Juliet Five Hotel Golf"** or only the spoken sequence **"Five Hotel"** with the noise-free full CQ-call.

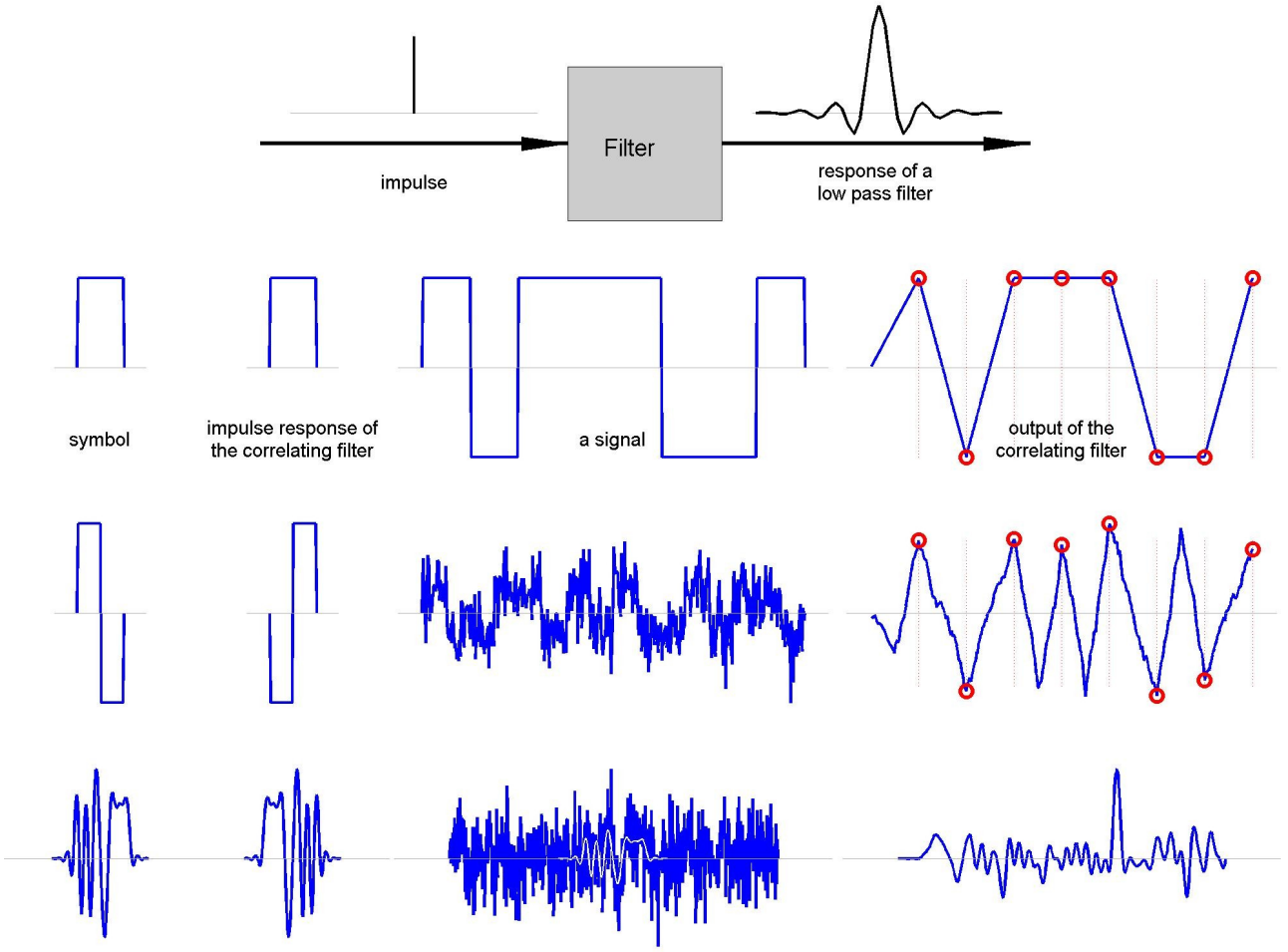
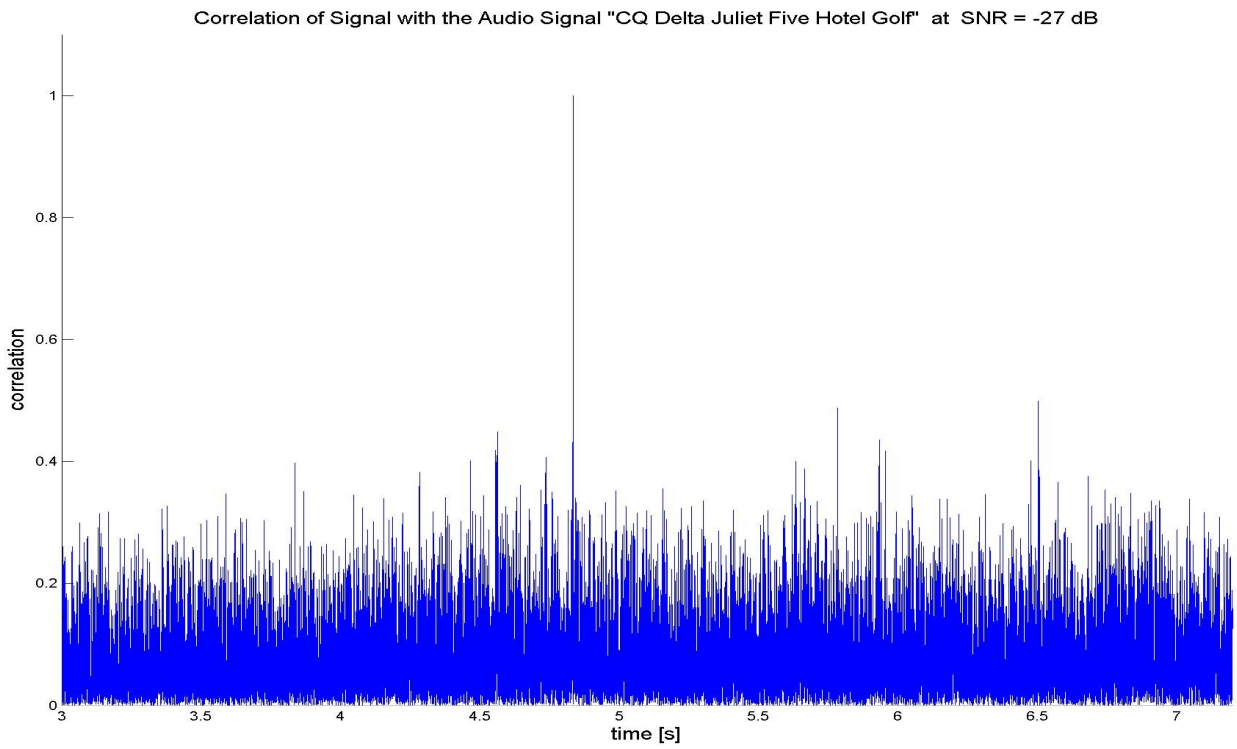


Figure 7. A correlator is a linear filter with an impulse response equal to the symbol that is searched for but mirrored in time. The upper part of the figure explains that the impulse response is what a time-invariant system answers when the input signal is a sharp impulse. The lower part shows three examples with (from left to right): symbol, its mirror symbol, an example of a signal, and the response of the correlator to the given signal (before the modulator, or after the demodulator resp.). The lowest example is the complex pulse of a RADAR (a Barker code). The correlator here finds a single weak signal marked as a white line within the noise at a high resolution. Example 1 and 2 both use two different symbols, the symbol shown and its negative for binary transmission. Eight symbols are transmitted in sequence here. The correlator therefore yields a sequence of positive and negative values (mapped to the binary digits 1 and 0). The small circles in the right figures mark which values of the correlator are taken for the decision of the received bits. It is obvious in the second example that these values are very well usable although the signal is noisy. The detection of the symbol clock is not discussed in this paper.



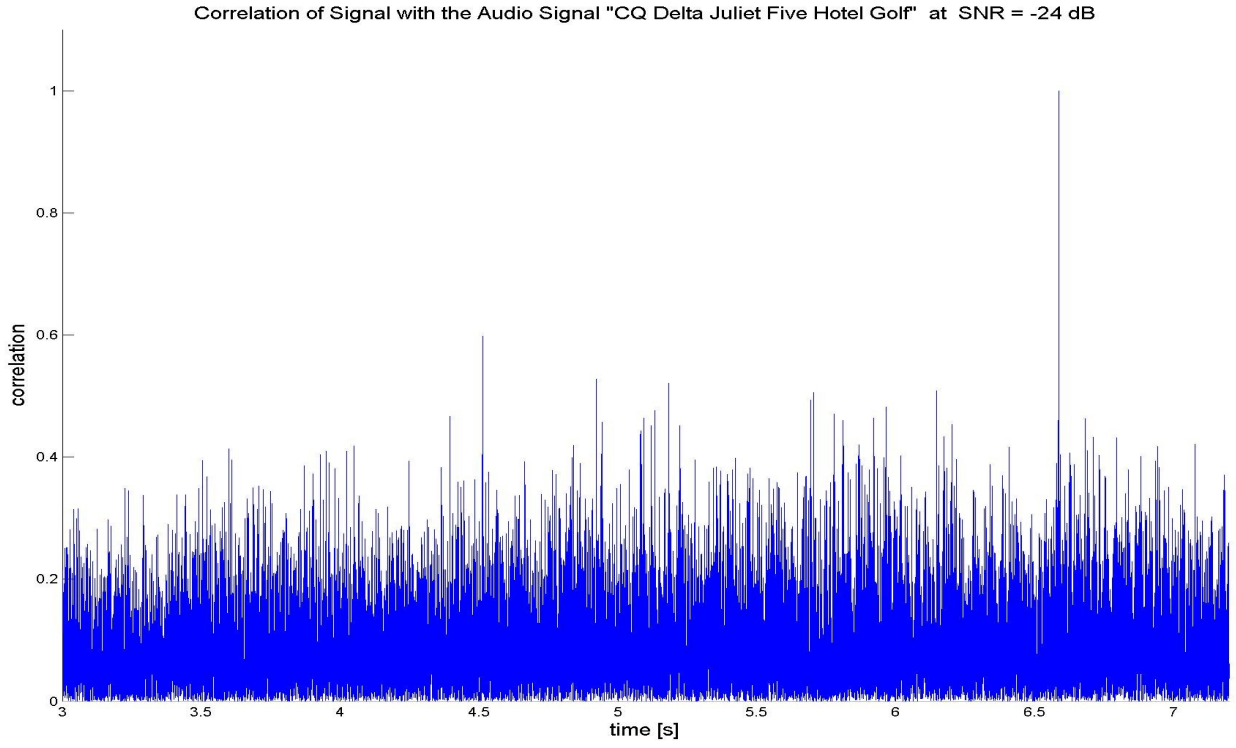


Figure 8. Upper figure: Correlation of a noisy signal that contains the spoken sequence "CQ Delta Juliet Five Hotel Golf" with the original noise-free signal. White Gaussian Noise was added at an SNR = -27 db. Lower figure: The signal only contains the sequence „Five Hotel“ at SNR = -24 db. The correlation is done with the full CQ call as in the upper figure.

2.7. Encoding and Decoding

Usual text is written in lower case and upper case latin letters, the ten decimal digits and some special symbols. For the radio transmission of text this alphabet cannot directly be used. But it is possible to find a coresponding number of different symbols such that the letters can be mapped to these symbols and vice versa. JT44 did it with FSK using 43 sine waves of different frequencies as its symbols (plus one frequency for synchronization). The decoder correlates the incoming signal with all symbols. If the symbol synchronization succeeds the decoder gets 43 real values from the correlators at the correct symbol clock. For each symbol clock the decoder must decide which symbol fits best. It takes the symbol with the largest correlation as „received". To be more precise: The decoder only gets an array of 43 real values, and it finds the index of the maximum value. Then this index is mapped to the known symbol.

If for example the 26 capital letters A...Z are followed by the digits 0...9 in the alphabet then my callsign DJ5HG uniquely is defined by the index array [4 10 32 8 7]. This is what really is received. While all symbols are known to the receiver this index array represents the received message. Before the transmission, all these indices are unknown at the receiving end.

It's the same with natural language. We never precisely hear a predefined word. Everything we detect with our senses is matched to what we already know, and the best fit is taken. If for example I say „Delta Juliet five Hotel Golf" what goes via the channel and what do you receive? It is not the symbols in all details like „Delta" etc.. If I say „Demta", this also will be decoded into „D".

If the probability of all symbols is equal then the information content of a message is measured as

$$\mathbf{n} \log_2 \mathbf{m} \quad \text{in bits}$$

n is the number of symbol positions in the message (the length of the meaasge), and **m** is the number of symbols the decoder looks for.

Example: The message „DJ5HG" with **m = 43** choices between different symbols and **n = 5** symbols in the message has the content of **27.13** bits.

The information content is independent of the symbols actually used. That is the reason why it is possible to transfer a callsign with a spoken alphabet at a similar energy as with JT44. The spoken alphabet even has the advantage not to loose 3 db by a synchronization tone (but there are other striking problems).

2.8. Confidence into a Decoded Message

Under the assumption of AWGN it is relatively easy to compute the probability **p** that the maximum of a set of correlation values used for the decision to a decoded symbol is reached by pure noise. The confidence into a decoded symbol then is **1-p.** This is the probability that the decoded symbol is not the result of noise. In case of non-Gaussian noise the confidence values may be entirely wrong. One should be aware of this fact if there are birdies or other man-made QRM.

2.9. Error Correcting Codes

An error correcting code combines **k** information symbols to one information block and adds some parity symbols such that **n** > **k** symbols have to be transferred. **m** is the number of different symbols. The advantage is that the decoder does not make a decision symbol by symbol but it takes the array of $\mathbf{n} \times \mathbf{m}$ real correlation values to find the best-fitting information block at once. As a consequence, there is only one confidence value for the received message. The possible gain of error correcting codes is shown in figure 9. The maximum possible gain increases with increasing information block size. The actual gain heavily depends on the code used.

The number of choices the decoder has is given by $\mathbf{m}^{\mathbf{k}}$.

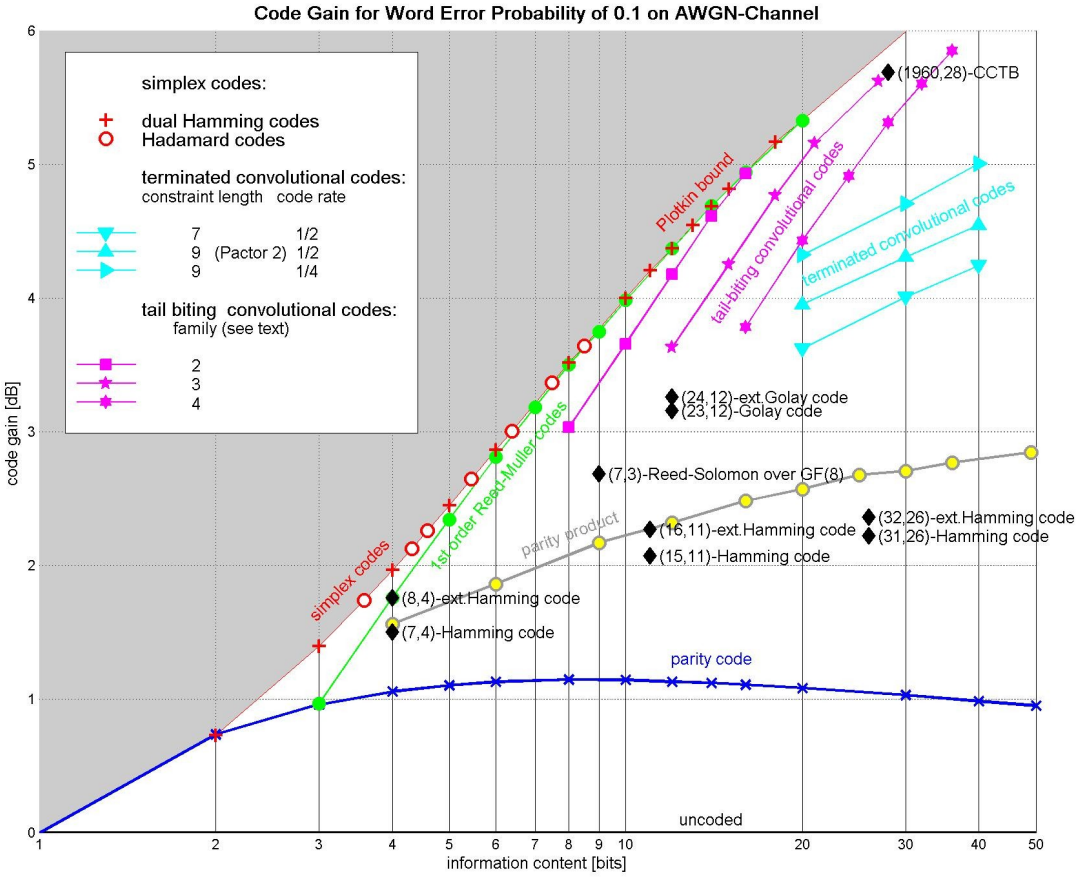


Figure 9. Gain of some codes over the uncoded transmission (AWGN). There is no possible gain if only one single bit must be transmitted (a final RRR for example). The maximum possible gain of an encoded transmission of 6 (30) information bits is 3 (6) dB resp.. The gain is considerably larger when large information content is encoded (4 k bits for ex.). This figure shows the region relevant for a minimal QSO with transmission of callsigns, reports, and rogers only.

3. The Problem of Validity of a QSO

3.1. Minimal QSOs

The classic rule that define a valid QSO demands for the copy of both callsigns at both stations. One should be aware of the fact that receiving the other callsign serves for its identification while receiving the own callsign serves for the confirmation that the QSO has started. The latter is a one-bit information while the identification of a callsign without using prior knowledge at least needs an information transfer given by the base-2-logarithm of the number of licensed radio amateurs. Indeed, this is the number of bits that no lossless digital compression technique for callsigns ever can reach (from higher values).

In analog modes in an answer to a CQing station the transmission of that callsign very often is entirely omitted. The correctly working change-over usually gives enough confidence in the QSO. If the other station does not ask for a repetition of my call I assume it has been received there. That is daily practise, especially in good conditions. As a consequence, QSO-rules for digital modes should respect the fact that there must be a confirmation that the other station got my callsign correctly. Sending the full callsign back for this purpose is waste of channel capacity. There are better methods, and analog modes make use of them.

Additionally to the callsign of the other station, both ends must receive a report and a confirmation. In pileups and in expeditions, QSOs usually are not terminated by the time-wasting handshake that guarantees meta knowledge at both ends upon what is known at the other end:

STN A STN B

RRR ==>	B receives RRR, so B knows C., but A thinks N.C. (C. = completeted)
<== 73	A receives 73, so both know C., but B does'nt know that A knows this
73 ==>	B receives 73, so both know C. and they know that the other knows this (but of course A does not know that B knows that A knows C.)
	if B does not receive the 73 from A, then both know C., and A knows that B knows this, but B does'nt know that A knows that B knows C..

3.2. The Problem of the DS-Decoder of JT65

Digital modes use coding techniques. Copy then means a correct decode. If the decoding algorithm fails then we can guess some possible messages, encode them and compare the guessed codewords with the received signal by correlation. If, for example, we made four guesses and now declare the message that fits best as „received“ then – in the sense of Shannons famous Information Theory – only 2 bits of information (base-2-logarithm of 4) went the entire path from the transmitter to the output of the decoder.

In digital communication it is the receiving end, namely that stage where the decisions are to be made, which determines the amount of information that can be retrieved from a signal. What the DS-decoder of JT65 really does is to declare the encoded messages generated from the list as the new symbols. Then a one-symbol message is received by the DS-decoder. The number of choices the decoder has is $\mathbf{m^k}$ with $\mathbf{k = 1}$ and $\mathbf{m}$ being the number of messages tried. Therefore the information content transported via the radio path is $\mathbf{\log_2 m \approx 14\ bits}$ which is not sufficient to encode a single callsign.

The situation may be explained by Figure 10.

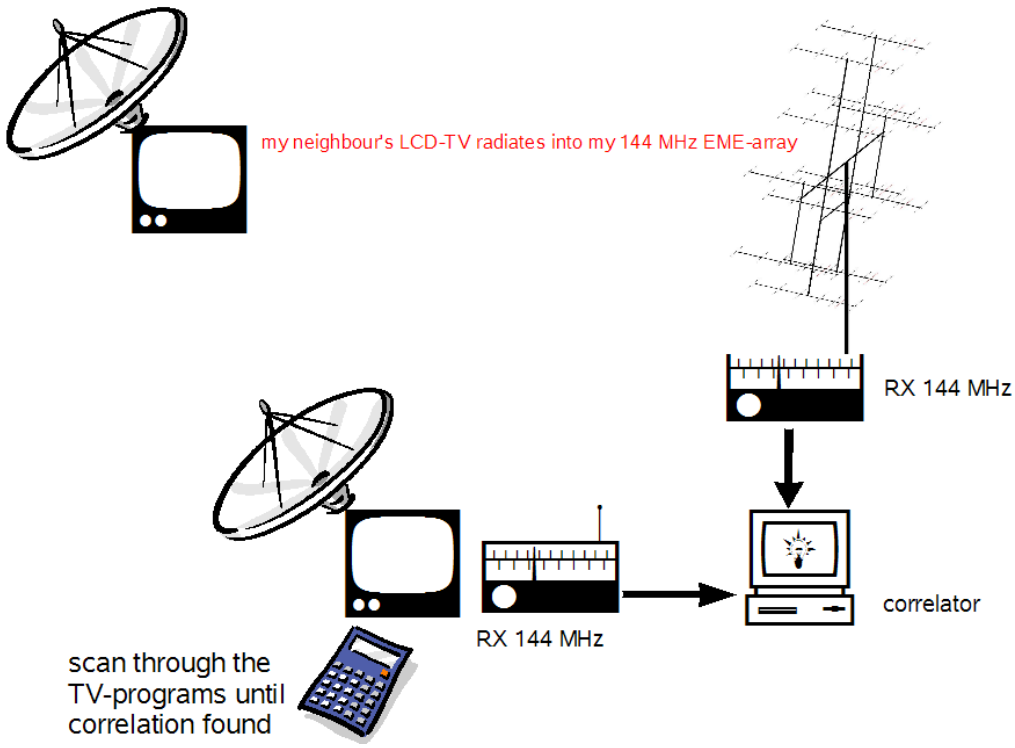


Figure 10. My neighbour's LCD-TV radiates into my EME-array. The noise seems to be white but it has a slight signature depending on the program he looks at. The noise is fed to the left channel of my sound card. A second receiver for 144 MHz directly takes the noise from my own TV and feeds it to the right channel of the sound card. If I now scan all the TV programs, the correlator will say which of the programs simultaneously runs on the neighbour's TV. If we want to believe the contention the DS decoder is based on, then we must believe that what I see on my TV-screen comes from the neighbour's TV via the noise radiation on 144 MHz, via myself over my remote control via the infrared path to my TV. In reality, it comes from my satellite antenna. What goes via the noise on 144 MHz only is the information which program I should choose. This information content does not increase in time. It is $\log_2$ (#programs) . So, nearly nothing of the actual TV scene goes via the noise path, and it is a presumption to say that all I see on my TV comes from my EME receiver. We only have to replace the satellite antenna here by the database and the TV by a JT65 transmission to get the situation of JT65 with the DS decoder.

The Deep-Search decoder of JT65 uses a freely editable list of callsigs for the „guesses“. Although the transmitting station sends the callsigns, the deciding stage in the decoder only offers the line number of the list as its result (or „no line fits“). Depending on the number of callsigns in the list, the amount of information received by the Deep-Search decoder is about 14 bits or less.

Figure 11 shows the Shannon limit computed from Joe's measurements concerning correctly copied symbols and applied to a channel with 64 equally likely symbols. It is clear from this figure that there is no chance to transfer the full message of 72 bits at an SNR lower than about -24 dB by a single path. But the DS-decoder can receive the necessary 14 bits at down to -28 dB .

The problem of the DS-decoder is that it only receives an index (out of some thousand possible indices) that is mapped to a message which is already known to the decoder. So not the callsigns are transported (they all are already there). As a consequence, QSOs based on the DS-decoder do not satisfy the classic QSO rules.

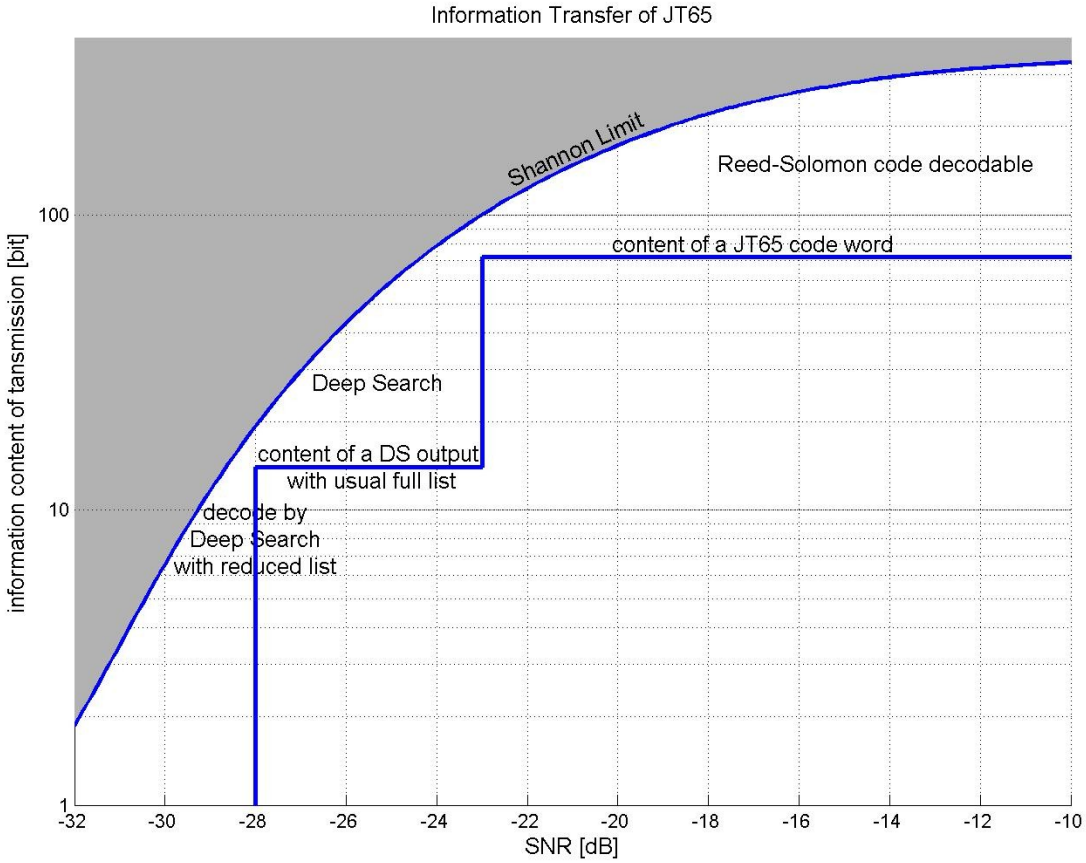


Figure 11. As in most digital modes the information throughput of JT65 does not degrade continuously with decreasing SNR but degrades in steps. The actual information transfer of JT65 is 72 bits or 14 bits or less than 14 bits or 0. The Shannon limit is computed here based on Joe's measurements concerning correctly copied symbols and applied to a channel with 64 equally likely symbols. This limit cannot be reached with finite information content as is used in QSOs.

3.3. A Proposal for a Validity Rule

I propose to use the following formulation of a QSO-rule as a basis for discussion:

A valid contact is one where both operators have
(1) mutually identified each other,
(2) received a report, and
(3) received a confirmation of (a) the successful identification and (b) the reception of the report, and – necessary at one end only - (c) the confirmation.

This formulation avoids the problematic term *copy*. Without loosing any confidence in the QSO, it also respects the fact of the large difference between the two calls that in some sense must be received. Of course, the term *identification* has to be quantified. My suggestion is:

The identification process must be based on, or equal to, a decision out of a number of equally likely possible choices that is larger than the number of licensed radio amateurs throughout the world.

This formulation by no means does restrict any future coding schemes. It simply excludes lossy source coding of callsigns, and guarantees fairness. The formulation also solves the well-known problem of the validity of skeds: The decoder must take into account all possible callsigns. If it only looks for the sked partner that renders the contact invalid.

The proposal also avoids any formulation directed to digital or analog methods. It is entirely independent from those parameters.

4. Two Case Studies

4.1. JT65

The features of JT65 are:

64-FSK (good, but problem with more than one station calling)

incoherent demodulation (easy, but loss of 3 dB)

packet and symbol synchronization by an extra tone with quasi-Barker code (easy, but loss of 3 dB)

Read-Solomon code (no soft decoder exists: loss of 3 dB, use of a patented (!) algorithm reduces the loss to ca. -1.5 dB)

possible gain of up to 6 dB by averaging of subsequent passes

4.2. CWP

CWP primarily was designed for teaching purposes. Although it can well compete with JT65 it was not designed to defend CW against other digital methods. It will not be able to compete with future digital methods. But it retains the charme of CW to transmit variable-length uncoded text. And the human ear even cannot hear any difference to CW.

The differences to CW are:
(1) the speed is precisely fixed
(2) the packet length precisely is one minute
(3) uni/bipolar keying (PSK) is used instead of unipolar keying (ASK) (see figure 12)
(4) demodulation is coherent

CWP signals that are at least 80 Hz apart can be decoded separately from the same signal. Therefore, one should not stop calling a station that answers to another station (see figure 15).

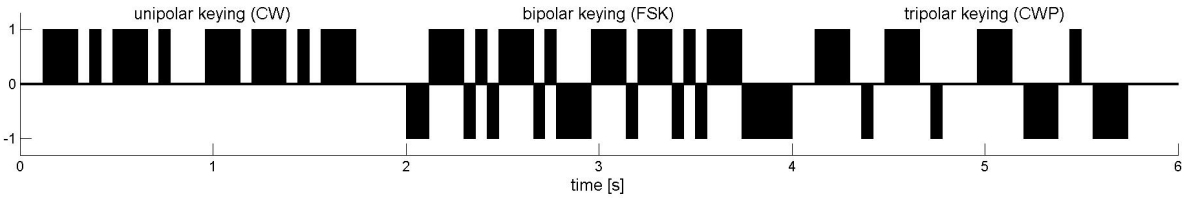


Figure 12. Unipolar, bipolar, and uni/bipolar keying. When modulated, a 0 results in an unkeyed carrier, a 1 is the keyed carrier, and a -1 is the keyed carrier with a phase shift by 180° (BPSK).

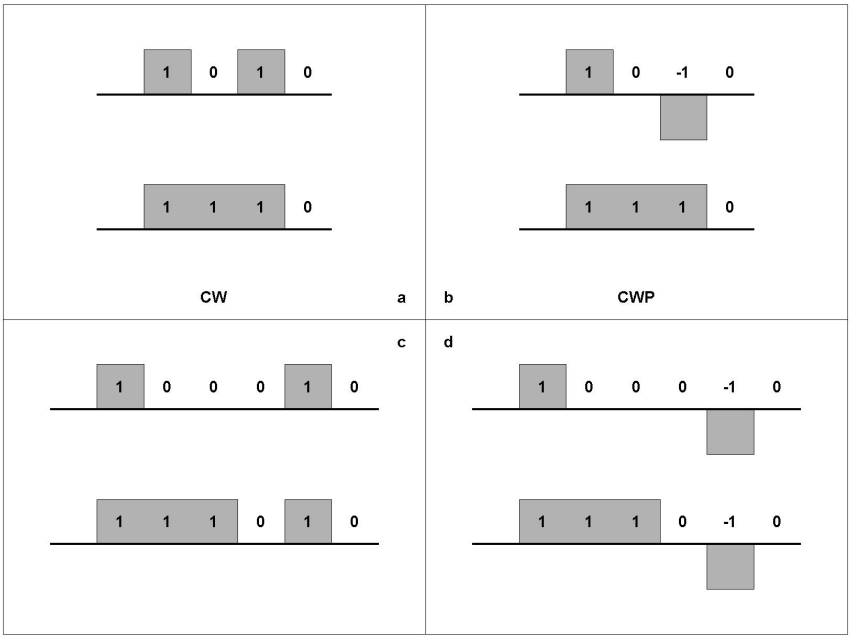


Figure 13. This figure demonstrates why CWP gains over CW. Two common situations are shown: (a) and (b) show the difference between two dots and a dash (type I<=>T) while (c) and (d) show the difference between two separated dots vs. a dash and a dot (type EE<=>N). The distance in each case is the squareroot taken from the sum of the squares of the differences of the values in all binary clock cycles. The relation of the smallest distances of CWP and CW is 6.99 dB.

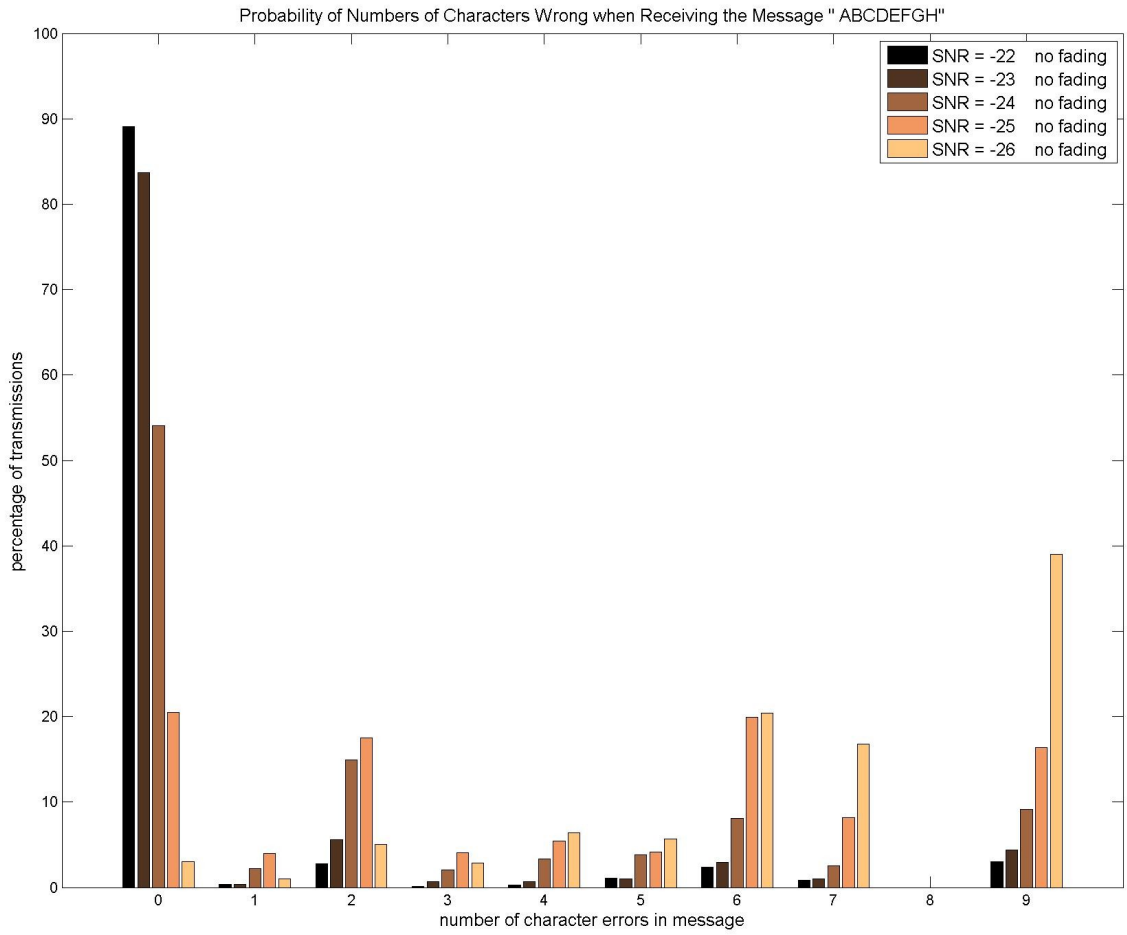


Figure 14. This figure shows the percentage of zero character errors up to all 9 characters in error for the testmessage „ ABCDEFGH“ depending on the SNR. It confirms the very low probability of single character errors. The total packet loss (#errors=9) increases with the SNR decreasing from -22dB to -26dB, while the probability of entirely correct packets decreases from 89% to 3%. The SNR is the relation of received signal energy to received noise energy in a bandwidth of 2500 Hz. The graph is based on 7500 simulated transmissions per SNR value. Fading very slightly increases the number of correct messages.

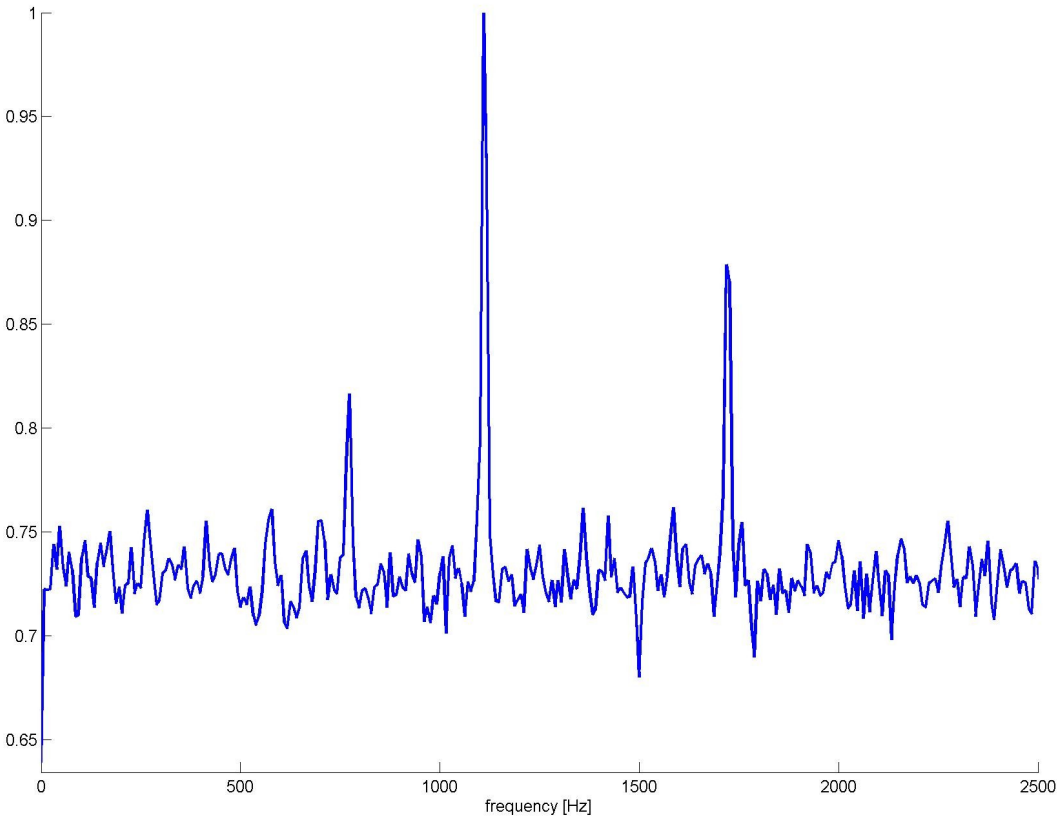


Figure 15. Spectrum of a signal with three CWP-signals at SNR-levels -27 dB, -22 dB, -24 dB (left to right). The drift is 0.0, -3.0, +1.8 Hz per minute, respectively.