Spectral Efficiency

Saso Tomazic
Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia

Abstract
Spectral efficiency is a measure of the performance of channel coding methods. It refers to the ability of a given channel encoding method to utilize bandwidth efficiently. It is defined as the average number of bits per unit of time (bit rate) that can be transmitted per unit of bandwidth (bits per second per Hertz).

The term channel encoding refers to the procedure of mapping a bit stream to an analog signal, which can be transferred through a physical channel and later, after reception, decoded to yield the original bit stream, as shown in Fig. 1.

The received signal may be distorted in various ways and corrupted by noise and other interference on the channel. This can cause the bit stream at the output of the channel decoder to be different from the bit stream at the input of the channel encoder. The ratio of erroneously received bits to all transmitted bits is called the BER and is used as a measure of performance. A small BER is acceptable in practical systems. The value of acceptable BER varies from 10^{-10} to 10^{-1} depending on the system concerned.

The transmitted analog signal occupies a certain bandwidth (frequency range), which depends on the bit rate and on the way that the binary sequence is mapped to the continuous signal, i.e., on the channel encoding method. The ratio of bit rate to bandwidth is called spectral efficiency. Spectral efficiency cannot be made arbitrarily large. The maximal spectral efficiency depends on the power of the transmitted signal, on the acceptable BER, and on the characteristics of the channel (distortion, noise, interference). The coding method that achieves maximal spectral efficiency on a given channel is considered optimal for this channel.

THEORETICAL BOUNDS

The upper bound of the bit rate r_b over a noisy band-limited channel for error-free transmission was first stated by Shannon.[1] One way to express this bound is:

$$r_b \leqslant \int_{f_l}^{f_u} C(f)\,\mathrm{d}f \tag{1}$$

where f_l and f_u are the lower and upper frequency bounds respectively, and *C(f)* is the frequency-dependant capacity of the channel (bits per second per Hertz).

The performance of a channel coding method can be evaluated by comparing the bit rate obtained by that method to the Shannon bound. However, no method for determining *C(f)* of an arbitrary channel is known. It can only be determined for a small number of special cases, for a channel with additive Gaussian noise, e.g, where it can be expressed as:

$$C(f) = \log_2(1 + \mathrm{SNR}(f)) \tag{2}$$

where SNR*(f)* is the frequency-dependant signal-to-noise ratio at the receiver input. Even in this case it can be hard or even impossible to determine SNR*(f)* at the receiver input when the channel is not linear.

Thus, in general one cannot tell how close to the Shannon bound different coding methods are. The performance of different channel coding methods is usually evaluated on an additive white Gaussian noise (AWGN) channel. On an AWGN channel, SNR*(f)* is constant and Eq. 1 can be rewritten as:

$$r_b \leqslant B \log_2\left(1 + \frac{S}{N}\right) \tag{3}$$

where B is the total bandwidth needed for transmission, S is the total signal power and N is the total noise power over bandwidth B. The above inequality is also known as the Shannon-Hartley theorem.

Signal power S is equal to the average bit energy E_b multiplied by the bit rate r_b:

$$S = E_b r_b \tag{4}$$

while noise power N over bandwidth B can be expressed as:

$$N = N_0 B \tag{5}$$

where N_0 is the noise level, i.e., the one-sided power spectral density of AWGN.

Encyclopedia of Wireless and Mobile Communications DOI: 10.1081/E-EWMC-120043448

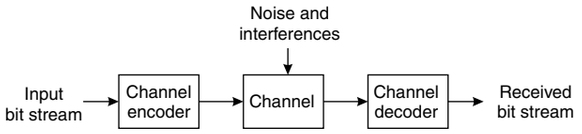

Fig. 1 Channel coding. The input bit stream is mapped to an analog signal (channel encoding) transferred through a channel with interference and decoded at the receiver (channel decoding). Ideally the received bit stream would be identical to the input bit stream.

Spectral efficiency u is given by the ratio:

$$u = \frac{r_b}{B} \tag{6}$$

Substituting Eqs. 4–6 into Eq. 3, after rearrangement, yields:

$$u \leqslant \log_2\left(1 + u\frac{E_b}{N_0}\right) \tag{7}$$

and

$$\frac{E_b}{N_0} \geqslant \frac{2^u - 1}{u} \tag{8}$$

The theoretically minimal ratio E_b/N_0 needed for error-free transmission over an AWGN channel is obtained when equality holds in the above expression. Denoting with BENR (bit energy to noise level ratio) the minimal E_b/N_0 in decibels, we can write:

$$\text{BENR} = 10\log_{10}\left(\frac{2^u - 1}{u}\right)\text{B}3u - 10\log_{10}(u) \tag{9}$$

BENR as a function of spectral efficiency u is shown in Fig. 2. Note that the BENR increases almost linearly with spectral efficiency u. By assigning more bandwidth to the transmission, the spectral efficiency drops and less power is needed for transmission. The minimal BENR i.e., the minimum energy required to transmit one bit of information over an AWGN channel, is obtained when the entire frequency range is assigned to the transmission. In this case, spectral efficiency approaches 0 and BENR approaches -1.592 dB, which is also known as the Shannon bound for an AWGN channel.

On the other hand, to increase spectral efficiency the power of the transmitter must be increased. To transmit one bit per second per Hertz ($u = 1$) the BENR is 0 dB, which means that the bit energy E_b must be at least equal to the noise level N_0. Further increasing the spectral efficiency can be very expensive in terms of transmitted power. To achieve a spectral efficiency of 57 Kbps for dial-up modems (u *B* 18.5), the BENR must already be more than

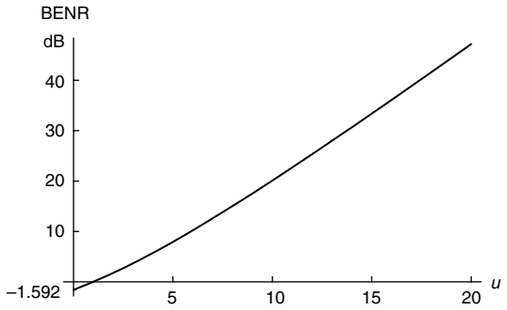

Fig. 2 The bit energy to noise level ratio (BENR) on an additive white Gaussian noise (AWGN) channel as a function of spectral efficiency u. To increase spectral efficiency, more energy is needed to transfer each bit, which also implies more transmitter power. The minimal BENR is -1.592 dB, which is also known as the Shannon bound.

43 dB, which means that the average bit energy must be approximately 20,000 times greater than the noise level.

The above results hold for an AWGN channel only; however, they are often used to estimate the bounds of spectral efficiency on other channels when the exact bounds and/or exact channel characteristics are not known.

CODING AND MODULATION

For practical purposes and also to simplify design and analysis of transmission systems, the process of channel coding (encoding/decoding) is usually divided into two parts: coding (in a narrow sense of the word) and modulation. It is important to note that within the scope of this article, the term encoding is used for mapping a bit stream to a real-valued symbol stream and the term modulation is used for mapping the symbol stream to an analog signal. Similarly, de-modulation maps the analog signal to a symbol stream and decoding maps the symbol stream to a bit stream. Other definitions of coding and modulation can be found in other sources. The divided system is shown in Fig. 3.

The input bit stream $b[n]$ at bit rate r_b is first encoded to a real valued symbol stream $s[m]$ at symbol rate r_s. The modulator maps the symbols $s[m]$ to an analog signal $s(t)$, which is then transmitted through the channel.

Speaking of spectral efficiency implicitly assumes band-limited modulation. If the modulation was not band-limited, spectral efficiency would be zero. When the channel is non-linear it produces out-of-band frequency components. These components can also carry information. Since we assume that the transmission is band-limited, the information transmitted out-of-band should be irrelevant for the receiver. To ensure that no information

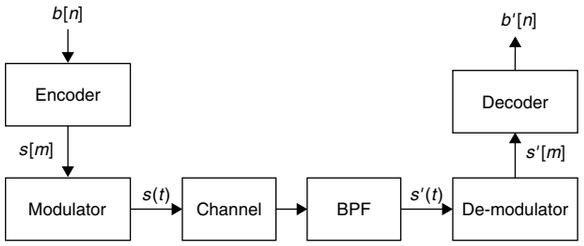

Fig. 3 Division of channel coding into coding and modulation. The input bit stream $b[n]$ is first encoded to a symbol stream $s[m]$ and then modulated to an analog signal $s(t)$. The received signal $s'(t)$ is de-modulated to a received symbol stream $s'[m]$ and then decoded to a received bit stream $b'[n]$. Ideally $b[n]$ and $b'[n]$ would be identical. BPF, band-pass filter.

is carried out of the transmission band, an ideal band pass filter (BPF) of bandwidth B is included at the output of the channel in front of the receiver. The band-limited signal $s'(t)$ at the receiver input is first de-modulated to the received symbol stream $s'[m]$ and then decoded to $b'[n]$. Ideally the received bit stream $b'[n]$ would be identical to the input bit stream $b[n]$. In practice a small number of errors is tolerated as long as BER does not exceed an acceptable level.

The purpose of encoding is twofold: to add redundancy (a dependency among the symbols at the encoder output) and to establish proper levels (values of the symbols) for modulation. The purpose of modulation is to map the values of the symbols to some properties of the analog signal in such a way that the symbol stream can be recovered from the received signal at the receiver.

Although Shannon determined the bounds for error-free transmission, he did not indicate how to achieve them. Different coding (encoding/decoding) techniques are in use to approach these bounds. Redundancy is added in the encoding process. It is used in the decoder to detect errors (error-detecting codes), to correct errors (error-correcting codes), or to minimize the probability of error (forward error correction codes). To approach the bound on a given channel, the code should also be adapted to the channel. Different adaptive techniques, such as adaptive equalization, adaptive inter-symbol interference canceling, adaptive filtering, and others are used for this purpose. There is no single optimal coding method for all channels, nor is any method currently known for determining the optimal code for an arbitrary channel.

The modulator encodes the values of the symbols at its input by varying different properties of the harmonic signal (the carrier) at its output: amplitude (AM, amplitude shift keying—ASK), frequency (FM, frequency shift keying—FSK), phase (PM, phase shift keying—PSK), or amplitude and phase at the same time (quadrature amplitude modulation—QAM). The spectral content of the modulated signal is centered around the frequency of the carrier. Its bandwidth greatly depends on both encoding and modulation. In general, the bandwidth of frequency- and phase-modulated signals is greater than the bandwidth of amplitude-modulated signals. Sometimes, when the noise is not white and/or the channel characteristics are not flat, it is beneficial to modulate multiple carriers (multiple carrier modulation—MCM and orthogonal frequency division modulation—OFDM). Modulated signals with more bandwidth are usually more resistant to noise and other interference, thus robustness is obtained at the expense of lower spectral efficiency.

Spectral efficiency depends on coding and modulation, on one hand, and also depends on channel characteristics on the other. Thus, another way to improve spectral efficiency is to improve the channel characteristics. In wireless communications this can be done by using directional antennas at the transmitter and/or at the receiver. Another method is to use multiple antennas at both sides in so-called multiple input multiple output (MIMO) systems.

OPTIMAL BAND-LIMITED MODULATION

As stated in the previous section, no method is known for determining the optimal channel coding for an arbitrary channel. We also mentioned that spectral efficiency depends on coding and modulation. The latter has led to different attempts to discover a new modulation technique which would improve on all currently known methods in terms of spectral efficiency, e.g., FQPSK (Feher-patented quadrature phase shift keying),[2] VPSK (variable phase shift keying),[3] VMSK (very minimum shift keying)[4] and many other ultra-narrow band modulations. A simple proof that ultra-narrow band modulations may not be as spectrally efficient as claimed is given in Ref. 5.

Although there is no encoding method that would be optimal for every channel, this does not hold for band-limited modulation. If we accept the definition of modulation from the previous section, then QAM at symbol rate $r_s = 2B$ is optimal, at least from a theoretical point of view.

To be more specific: any band-limited modulation can be performed using a QAM modulator preceded by the appropriate encoder. In other words, the signal space at the output of the QAM modulator covers all possible band-limited signals. If it holds that any band-limited modulation can be performed using a QAM modulator, then it also holds that optimal band-limited modulation can be performed as well, which makes QAM itself optimal. An ideal QAM modulator is shown in Fig. 4.

The input symbol stream is split into two half symbol rate symbol streams that are used to modulate the amplitudes of two orthogonal carriers. The modulated carriers are summed up to form the output signal of the modulator. At the receiver the procedure is repeated in reverse order. The received analog signal is first multiplied by two

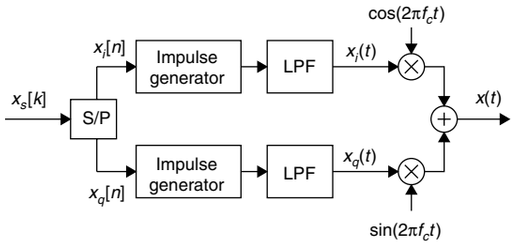

Fig. 4 Ideal QAM modulator. The input symbol stream $x_s[k]$ is split into two symbol streams $x_i[n]$ and $x_q[n]$ at half the symbol rate of the input symbol stream. The ideal Dirac impulses generated at the impulse generators are shaped at the ideal LPFs of bandwidth $B/2$. The in-phase signal $x_i(t)$ and $x_q(t)$ the quadrature signal are multiplied by two orthogonal carriers $\cos(2\pi f_c t)$ and $\sin(2\pi f_c t)$, respectively, and summed up to yield the output signal $s(t)$. S/P, serial-to-parallel converter.

coherent orthogonal carriers, one in each branch of the de-modulator. High frequencies are filtered out by two ideal LPFs and the signals at the output of the filters are sampled at half the symbol rate. The obtained symbol streams are combined into a single symbol stream at the output of the de-modulator. The QAM de-modulator is shown in Fig. 5.

To verify that QAM modulation is optimal, we should first recognize that the QAM de-modulator in Fig. 5 can be used as a band-limited signal sampler. The signals $x_i(t)$ and $x_q(t)$ are base-band signals with bandwidth $B/2$. If the sampling rate of both samplers is B samples per second, then the signals $x_i(t)$ and $x_q(t)$ can be perfectly re-constructed from the samples $x_i[n]$ and $x_q[n]$ respectively. This is known as the sampling theorem. A proof can be found in almost any elementary text on communications theory (see, e.g., Refs. 5, 6, or 7).

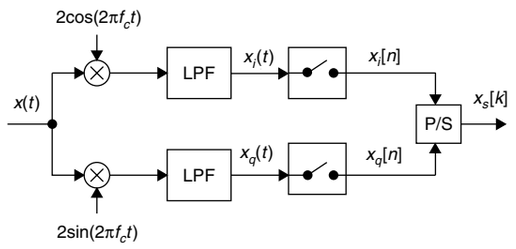

Fig. 5 QAM de-modulator. The input signal $x(t)$ is fed into two branches of the de-modulator. In one branch it is multiplied by $2\cos(2\pi f_c t)$ and by $2\sin(2\pi f_c t)$ in the other. The signals in both branches are then filtered with ideal LPFs and sampled at half the symbol rate. The symbol streams $x_i[n]$ and $x_q[n]$ at the output of the samplers are then combined in the P/S to yield the received symbol stream $x_s[k]$.

From the above it is then obvious that $x(t)$ can be perfectly re-constructed from $x_s[k]$ as well. The QAM modulator in Fig. 4 can be used for re-construction. The signals $x_i(t)$ and $x_q(t)$ are re-constructed from the samples $x_i[n]$ and $x_q[n]$ on ideal LPFs and then multiplied by orthogonal carriers and summed up to yield the original signal $x(t)$. The QAM de-modulator and QAM modulator at symbol rate $r_s = 2B$ can thus used for sampling and perfect re-construction of any band-limited signal with bandwidth less than or equal to B.

Without any change in performance, except for additional delay, we can add sampling with perfect re-construction at both sides of the transition system from Fig. 3, as shown in Fig. 6.

Suppose now that the original encoder and modulator in Fig. 6 were optimal for a given channel with regard to some criterion, e.g., maximal spectral efficiency. Since the performance of the system was not altered by the addition of QAM modulator/de-modulator pairs, the system is still optimal. In this new system the original encoder, original modulator, and QAM de-modulator form a new encoder (dotted box at the left-hand side of Fig. 6) which maps the input bit stream $b[n]$ to a symbol stream $s_s[k]$. The QAM modulator is then used for modulation. At the receiver the QAM de-modulator is used for de-modulation, and a new decoder, which includes the QAM modulator, original de-modulator and original decoder (dotted box at the

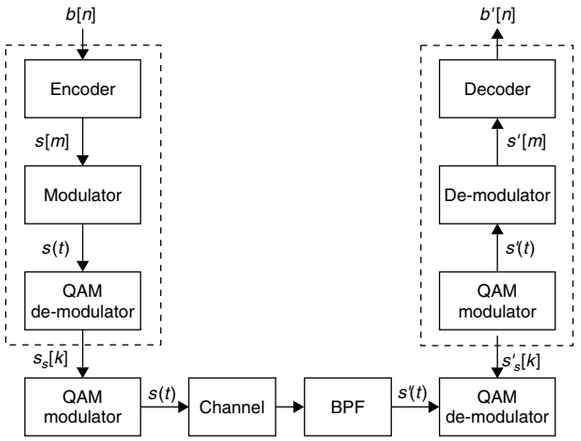

Fig. 6 Modified transmission system. The QAM de-modulator and QAM modulator are added at both sides. QAM de-modulators acts as samplers and QAM modulators perform perfect re-construction, thus the performance of the system is not altered by these additions, except for the delay of the filters. The original encoder, original modulator and QAM de-modulator form a new encoder, which encodes the bit stream $b[n]$ to the symbol stream $s_s[k]$. The QAM modulator, original de-modulator and original decoders form a new decoder, which decodes $s'_s[k]$ to $b'[n]$. BPF, band pass filter.

right-hand side of Fig. 6), is used to decode the received symbol stream $s'_s[k]$ to the received bit stream $b[n]$.

We suppose that the whole system is optimal. The modulation method used in this system is QAM. We can thus conclude that QAM at symbol rate $r_s = 2B$ is an optimal modulation method. It is important to note that the new encoder must be adapted to the channel characteristics; however, this does not hold for the QAM modulator. The same modulator can be used on any channel.

This result is important mainly from a theoretical point of view. We can only adapt to the characteristics of the given channel (e.g., mobile channel) through a suitable choice of coding, which indicates that research should be focused on finding optimal coding methods instead of being focused on discovering new modulation techniques.

Although, at least theoretically, any band-limited modulation can be implemented using QAM, this may not be the most practical solution. Other modulation methods may have practical advantages and thus will continue to be used in transmission systems.

SUMMARY

Spectral efficiency is an important measure of the performance of a digital transmission system, especially in wireless communications. The upper bound of spectral efficiency depends on the channel characteristics and is not known for all channels. To improve spectral efficiency, additional energy is needed to transmit one bit of information and/or channel coding must be adapted to the channel. If channel coding is split into coding and modulation, only the coding needs adaptation to the channel and QAM can be used for modulation. Other modulation methods can have practical advantages and will continue to be used in digital transmission systems.

REFERENCES

1. Shannon, C.E. Communication in presence of noise. IRE. **1949**, *37* (1), 10–21.
2. Feher, K. et al., Feher's quadrature phase shift keying (FQPSK). US Patents 5,784,402 and 5,491,457.
3. Walker, H.R. VPSK and VMSK modulation transmits audio and video at 15 b/s/Hz. IEEE Trans. Broadcast. **1997**, *43* (1), 96–103.
4. Walker, H.R.; Stryzak, B.; Walker, M.L. Attain high bandwidth efficiency with VMSK modulation. Microw. RF. **1997**, *36* (13), 173–186.
5. Tomazic, S. Comments on spectral efficiency of VMSK. IEEE Trans. Broadcast. **2002**, *48* (1), 61–62.
6. Papoulis, A. *The Fourier Integral and its Applications*; McGraw-Hill Book Co.: New York, 1962.
7. Proakis, G.; Salehi, M. *Communications Systems Engineering*; McGraw-Hill: MI, 2001.