3

INFORMATION THEORY: DEFINITIONS AND HISTORICAL BACKGROUND

3.1 Towards a theory of information - a historical survey

Although the statistical principles applied in Information Theory are similar to those applied to the relatively older science of thermodynamics, which dates back to the middle of the nineteenth century, it took nearly a hundred years before these principles were actually developed into a theory to measure modes of communication.

As far as the measuring of information in a communication system is concerned, some initial research was done by H. Nyquist (1924), R. A. Fischer (1925), and R. V. L. Hartley (1928). The science of Information Theory is, however, mostly indebted to the work of V. A. Kotel'nikov (1947), N. Wiener (1948) and C. E. Shannon (1948). Shannon with his publication, *A Mathematical Theory of Communication,*[1] is regarded as being responsible for establishing the foundations of Information science as it is known today.

In the following pages a short historical survey of Information Theory is provided. First the development of the concepts of entropy in physics and thermodynamics is discussed. This is followed by a

[1] *Bell Systems Technical Journal*, Vol. 27, pp. 379-423, 623-656. Republished in collaboration with W.W. Weaver as *The Mathematical Theory of Communication*, Urbana: Univ. of Illinois, 1949.

more detailed account of how similar theories were subsequently employed in the development of the science of communications and telecommunications.

3.1.1 Entropy, disorder and the movement of heat molecules

In 1854, the German scholar, Rudolf J. E. Clausius (1822-1888) published a scientific paper[2] in which he formulated what has since become the second law of thermodynamics.[3] In essence the formulation states that in an isolated system

> ... no process is possible whose sole result is the transfer of heat from a colder to a hotter body. (Kendall: 1973, p. 59)

The above law is closely allied to the first law of thermodynamics which states that when two bodies with unequal temperatures are brought into contact, a process of equalisation will take place. Entropy, which is the measure of disorder in a system, reaches its maximum value as the system reaches equilibrium. According to these two laws heat can only transfer from a hotter body to a colder body until the temperature of both is equal. Entropy or disorder in an isolated system can, in other words, only increase and never decrease.

According to Kendall the British physicist, James C. Maxwell (1831-1879), was one of the first to approach the concept of entropy from a probabilistic point of view. (Kendall: 1973, p. 59) Maxwell formulated an equation by which the distribution of velocities of gas molecules may be calculated[4].

Further research in the statistical properties of gas molecules and heat was undertaken by the Austrian physicist Ludwig Boltzmann (1844-1906):

> It was he who linked the thermodynamic concept of entropy with the statistical concept of disorder. (Kendall: 1973, pp. 59 -60)

In his book, *Vorlesungen über die Gastheorie*,[5] Boltzmann provides a formula which expresses the logarithm of the resulting probabilities of the distribution of molecular velocities as a ratio of:

Equation 3-1. Boltzmann's formula

$$n \log_e n$$

This has since become an important component of statistical physics but one which will also recur frequently in this discussion on the application of Information Theory to music.

---

[2] Clausius subsequently published his findings in a book: (1864) *Abhandlungen über die mechanische Wärmtheorie*, Braunschweig: Friedriech Vieweg.

[3] A branch of physics which describes the physical properties of matter and energy.

[4] Funk & Wagnalls, 1983: vol. 17, James Clerk: 'Maxwell'.

Further research with the concept of entropy and statistical physics continued to be done by a number of investigators such as J. von Neumann and L. Szilard.

It was during the late 1920s that several communication engineers began investigating the possibility of applying the concepts of entropy to communication.

3.1.2 Entropy and communications

The first public telegraph line was installed in Britain in 1843. The Italian, Guglielmo Marconi (1874-1937) announced the discovery[6] of wireless telegraphy in 1895. Nearly twenty years before, in 1876, Alexander G. Bell's invention, the telephone, had successfully transmitted human speech. But none of these systems were yet perfect and during the first decade of the twentieth century continuous research relating to the problems that still plagued electronic communication was conducted.

Distortion, signal noise, and inter-symbol interference necessitated techniques to overcome these difficulties and ensure transmission of intelligible messages. This research was predominantly done at the Bell Telephone Laboratories in America where H. Nyquist (1924, 1928) and R.V.L. Hartley (1928) were actively working on these problems. Most of their findings were published during the 1920's in articles such as Hartley's 'Transmission of Information'.[7] More or less at the same time, similar research was done by Norbert Wiener who specialised in the biological application of the transmission of information as, for example, in the nervous system. (Shannon: 1949, p. 3)

3.1.2.1 R.V.L. Hartley

From the sources[8] on the subject it would appear that of the research done at the Bell Telephone Laboratories, it was Hartley's work that stimulated further thought on the application of the concept of entropy in Information and Communication Science the most. In fact it was in the article mentioned above that Hartley expressed the theory that the quantity of information could be measured. Hartley based his theory on a principle simultaneously formulated by Nyquist and the German, Kopfmuller in 1924 which ...

> ... states that for transmitting telegraph signals at a given rate a definite—frequency bandwidth is required. (Reza: 1961, p. 11)

[5] Leipzig: Barth, 1896.

[6] The U.S.A. Supreme Court pronounced in 1943 that Nikola Tesla was the inventor of the radio (Cheney: 1981, pp. 176-184).

[7] 1928. *Bell System Technical Journal*, vol. 7, pp. 535-564.

[8] A more detailed account of the historical background of information theory is provided in E.C. Cherry's article, 'The Communication of Information', *American Scientist*, October, 1952.

Nyquist and Kopfmuller's work was further refined by D. Gabor (1946) and D. M. Mackay (1948), whose contribution to the measurement of information was important to Hartley's work. (Reza: 1961, p. 11)

Hartley's theory is founded on the concept that the information content of a message depends on the successive selection of symbols from a specific set of symbols. In terms of music the seven notes of a diatonic major scale may be used as an example. In an imaginary melody of eleven notes the notes may occur in 7^{11} (7 to the power 11) different ways. In more general terms, where N is the number of notes in a melody with length **L**, there are N^L possible different combinations in which the notes could occur.

Furthermore, Hartley continues by showing that the information in a message needs to be calculated from the actual symbols used in a message against the capacity of the system which transmits it and not from the total possible symbols available. (Kendall: 1961, p. 61) If the imaginary melody is again taken as example, this means that the melody of eleven notes in length, represents the capacity of the melody. If this melody should comprise only five different tones out of the possible seven of the major scale, the selected five represent the capacity of the melody, and not the seven notes of the scale.

3.1.3 Quantifying information.

The formula that Hartley arrives at to express the maximum information of a set with **n** symbols and with **K** being a constant, is:

Equation 3-2. Maximum information of a set of symbols

$$I = K \log n$$

To demonstrate how Hartley arrives at this equation the seven notes (n = 7) of a major scale may again be used. Presuming that a random generator generates one note at a time and with equal probability, the seven notes may be represented as the set, $\{x_1, x_2, x_3, \ldots x_7\}$, each of which has an equiprobable chance of being heard next. The amount of information that the selection of a particular note generates may be expressed as a function of the seven notes, thus:

Equation 3-3. Information of a single note from a set of seven notes

$$I(n_x) = f\left(\tfrac{1}{7}\right)$$

where **I** is the information content and $\mathbf{n_x}$ any one of the seven notes in the set. One of the possible ways of expressing the function above is by using logarithms, thus the amount of information associated with each element in the set is:

Equation 3-4. Information of a single note of seven using logarithms

$$I(n_x) = -\log\left(\tfrac{1}{7}\right)$$

thus:

$$I(n_x) = \log(7)$$

3.1.4 Calculation of information with logarithms base two

Hartley's original suggestion was to calculate information using logarithms on base ten. Nevertheless, calculations of information are often done with logarithms with a base of two, in which case the amount of information is expressed as *bits*[8]. The following paragraphs illustrate why binary units[9] are often preferred for calculations of information contents.

The fact that a single element can be expressed with two equal possibilities (i.e. on/off, sound/silence, yes/no), and as such is the minimum amount of information a system can generate,[10] makes the use of binary counting especially appropriate. As an example the choice of playing a note on a musical instrument may be used: at any moment a performer is presented with the choice of playing a note or not playing a note, a set of two possibilities. The silence may be represented by the number '0' for a note not played (state not true), or '1' for a played note (state true). At any subsequent moment this possibility is repeated, thus the performer is continuously confronted with a situation of two possibilities with a 50% (0.5) probability. In base ten notation the calculation of the information will be:

Equation 3-5. Information of a single element using natural logarithms

$$\begin{aligned} I(x) &= -\log\left(\tfrac{1}{2}\right) \\ &= \log(2) \\ &= 0.3010 \end{aligned}$$

---

8 When natural base logarithms are used the units of information are expressed in nats, when base ten is used the information content is expressed in Hartleys.

9 In contrast to the decimal system which uses ten as its base, the binary system uses two as its base. Compare the following decimal numbers and their binary equivalents:

DECIMAL:	1	2	3	4	5	6	7	8	9	10
BINARY :	01	10	11	100	101	110	111	1000	1001	1010

10 As with the entropy of thermodynamics which can never have a negative value, a communication system can never have a negative information value - it is impossible to have a value smaller than zero.

However, if base two logarithms[12] are used the result of the same calculation is:

Equation 3-6. Calculation of information using binary logarithms

$$\begin{aligned} I(x) &= -\log_2\left(\tfrac{1}{2}\right) \\ &= \log_2(2) \\ &= 1 \end{aligned}$$

The parity of the information content for a single note on a score, as a rudimentary system, may thus be expressed as a single binary unit, customarily contracted into the word 'bit'. Using binary logs as base, Hartley's calculation for x number of symbols is written as:

Equation 3-7. Information for x number of symbols in Hartley's

$$\begin{aligned} I_{\max} &= -\log_2 x \\ &= -\log_2 x^{-1} \\ &= \log_2 x \end{aligned}$$

All the calculations thus far are based on the premise that at any given moment each symbol of a system is equally probable. Such static systems are also called systems without memory—memoryless. Communication systems, including music, are rarely all that simple as there are usually structural principles involved. In music, for instance, various conventions dictate the use of subsequent chords or their inversions in a chordal progression.

According to Kendall, Hartley's formulation had little impact, at first, outside the realm of electronic communication and it was not until ...

> ... Shannon took it up and extended it in a paper (1949) which may be regarded as the effective starting-point of the current interest in the subject. (Kendall: 1973, p. 62)

This was also the point at which the concept of Information Theory as an independent field of investigation began. Since its relatively recent genesis, Information Theory has become an important and active component of Communication Engineering and Cybernetics.[13] The latter refers to the manner in which information is moved, controlled and responded to by living organisms and machines. Because of its extensive and successful application in these sciences it has found extensive application

---

12 Conversion from a logarithm of a base ten number to the logarithm of a base two number is done with the formula (where **n** represents the number of elements): $\log_2 n = \frac{\log n}{\log 2}$

13 Cybernetics emerged as a science in the late 1940's. The auto-pilot of an airliner is an example of the application of Cybernetics. (Longman Dictionary: p. 275)

in other disciplines as well, and has since become an autonomous mathematical science of communication.

3.2 Claude Shannon and the stochastic process

In 1948 Shannon, who at the time was mainly concerned with cryptographic systems, conceived a formula based on the one by Hartley which allows for the unequal distribution values of elements from a discrete source[13] (as apposed to a continuous source) with a finite set of elements. He referred to it as a *stochastic process*, and it is expressed as a ratio of maximum entropy. (Shannon & Weaver: 1949, p. 179) A discrete source implies a limited set of symbols and is described by Shannon as:

> ... as system whereby a sequence of choices from a finite set of elementary symbols **S1** . . . **S**n can be transmitted from one point to another ... We can think of a discrete source as generating the message, symbol by symbol. It will choose successive symbols according to certain probabilities ... (Shannon: 1948)

According to Reza an important but independent contribution to the formulation of Information Theory was made by N. Wiener with his two books, *Cybernetics*[14] and *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*[15]. Reza writes:

> N. Wiener was one of the first scientists who clearly described the stochastic nature of communication problems. Wiener put in focus the fact that communication of information is primarily of a statistical nature. That is, at a given time a message is drawn from a universe of possible messages according to some probability law. At the next moment another message from this universe will be transmitted. (Reza: 1961, p. 375)

The Stochastic process Reza refers to, indicates some important differences in the concepts of Hartley's theory and that of Shannon and Wiener. Whereas Hartley conceived his theory on static systems with any symbol of a set occurring at any given moment with equal probability, Shannon worked with dynamic systems with memory which allows for,

1. the presence of a formal structure in a message in which certain successions of symbols may be excluded, limited or required; and
2. the possibility that each symbol may occur with different relative frequencies (Kendall: 1973, p. 62). This may be explained using a melody as an example in which
3. the appearance of a specific interval may according to convention or choice, preclude it to be followed by another specific interval; and certain intervals will be more predominant than others.

---

13 Shannon also identifies continuous systems, in which the message is treated as a continuous function, and mixed systems, which is a combination of a discrete and continuous system. (Shannon: 1948)

14 Cambridge: Technology Press. 1948.

15 New York: John Wiley. 1949.

3.2.1 The Markov[17] process

Point no. 3 above is a special case or subclass of the stochastic process and is known as a *Markov chain* or a *dependent stochastic process*. The theory that serves as the basis of the Markov chain is that the occurrence of a symbol in a time-dependent message depends on the previous symbol or that certain symbols may dictate or exclude the use of certain other symbols. Shannon explains the process as follows:

> A system which produces a sequence of symbols (which may, of course, be letters or musical notes, say, rather than words) according to certain probabilities is called a *stochastic process, and the special case of a stochastic process in which the probabilities depend on the previous events, is called a Markoff process or a Markoff chain. (Shannon: 1968)*

In other words, the Markov chain refers to the application of 'rules' or conventions in the inherent structure of a message by which it is made coherent, and makes allowances for the relationship between symbols and not only their individual ratios.

Markov processes are also found in conventional music practices, for instance in melodic construction where certain intervals or chords, according to convention, require to be succeeded by other specific intervals or chords or, conversely, preclude certain subsequent intervals or chords. Even in some of the more current musical systems the stochastic process, and especially the Markov chain is strongly in evidence. An example is strict serial music in which the elements of a complete series are dependent on and dictated by preceding elements.

Through the years composers and theorists have continuously relied on systems or conventions to serve as the basis for their compositions. The traditional system of harmony is one example in which the I - IV - V - I chord progression is only too familiar. Arnold Schoenberg's dodecaphonic technique is another example in which preconceived limitations and so-called 'rules' play an important role. For a melody to make sense, its note sequences are usually bound and influenced by structural and tonal elements complemented by idiosyncrasies of convention and 'style'.

Common examples of traditional melodic theory that are essentially stochastic in character include the tendency of the leading tone to resolve to the tonic, or for a melodic leap to be followed by a note within the leap; consecutive leaps in tonal melodies tend to outline specific chords and melodies tend to outline harmonic progressions. Many similar examples may be cited which tend to indicate limitations or precepts of the choice of notes and intervals imposed by convention or so-called 'rules'.

---

[17] This is a phonetic spelling, the name is also spelled Markoff.

3.2.2 Ergodic processes

Perhaps a more difficult concept is the ergodic process, a distinct form of the Markov chain. Weaver describes it as follows:

> Although a rigorous definition of an ergodic process is somewhat involved, the general idea is simple. In an ergodic process every sequence produced by the process is the same in statistical properties. Thus the letter frequencies, digram frequencies, etc., obtained from a particular sequence, will, as the length of the sequence increases, approach definite limits independent of the particular sequence. (Shannon and Weaver: 1986, pp. 45-46)

In simple terms the ergodic process hypothetically means that the relative redundancy values of sections of a message approaches or has the same value as the message as a whole.

Starting with the frequency distribution of single elements in a message, Information Theory, also allows for the calculation of the relationship between the various elements and the manner in which they are organised. Its most valuable application to music is that it calculates those aspects that are conventionally or stylistically fixed as a ratio of the maximum number of possibilities that such a system allows. A composer imposes his personal choice in selecting, a) the parameters such as rhythm, and chromaticism, and, b) allows himself artistic freedom within the chosen system and parameters. This means that the ratio between entropy or redundancy and maximum redundancy can actually represent a measure of the style of music. Hence Entropy is a measure of the creativity and originality applied in a composition within a stylistic framework.

3.3 Entropy

Using Hartley's equation, Shannon demonstrated that the quantity of information produced by a Markov process can be measured, and referred to this quantity as entropy:

> The quantity which meets the natural requirements that one sets up for 'information' turns out to be exactly that which is known in thermodynamics as entropy. (Shannon: 1968, p. 16)

According to Shannon the quantity of information can only be stochastically measured if certain conditions are met. Stochastic entropy is represented by the formulation $H(p_1, p_2, \ldots, p_n)$ where p_i is the probability of the i-th element. The conditions are[18]:

1. **Continuity of H in the p_i.** If the probabilities of the events change slightly, the measure of information (**H**) should similarly vary slightly.

[18] Based on Shannon's paper, 'A Mathematical Theory of Communication', 1948, and Reza's *An Introduction to Information Theory*, 1961, pp. 80-81.

2. **Symmetry of the H function in every p_k.** This means that a change in the order of events should not result in a difference of measurement, or

Equation 3-8. Symmetry of the H function in every p_k

$$H(p_1, p_2, \cdots, p_n) = H(pi_1, pi_2, \cdots, pi_n)$$

where $i_1, i_2, \ldots, i_n$ is any permutation of 1, 2, ... , n

3. **Extremal properties.** When the probabilities of an event are equal, the H function should have its maximum value for that set. In other words, when all events are equally probable there is maximum uncertainty:

Equation 3-9. Extremal properties

$$H_{\max}(p_1, p_2, \cdots, p_n) = H\left(\tfrac{1}{n}, \tfrac{1}{n}, \cdots, \tfrac{1}{2}\right)$$

4. **Additivity.** In a system which contains a choice which itself comprises two choices, the original value of H is equal to the weighted sum of the H values of the individual possibilities.

The formula that Shannon arrived at and which satisfies the above criteria is:

Equation 3-10. Shannon's formula for information in a stochastic process

$$H = -K\sum_{i=1}^{n} p_i \log p_i$$

Figure 3-1 illustrates the entropy value in bits for probabilities of **p** (shaded areas) and **1-p (white areas)**; note that as the probability of either **p** or **1-p** reaches its maximum value—meaning that the outcome of a choice becomes more certain—the entropy tends towards 0. When both **p** and **1-p** are equally probable, the entropy reaches its maximum value of 1 bit.

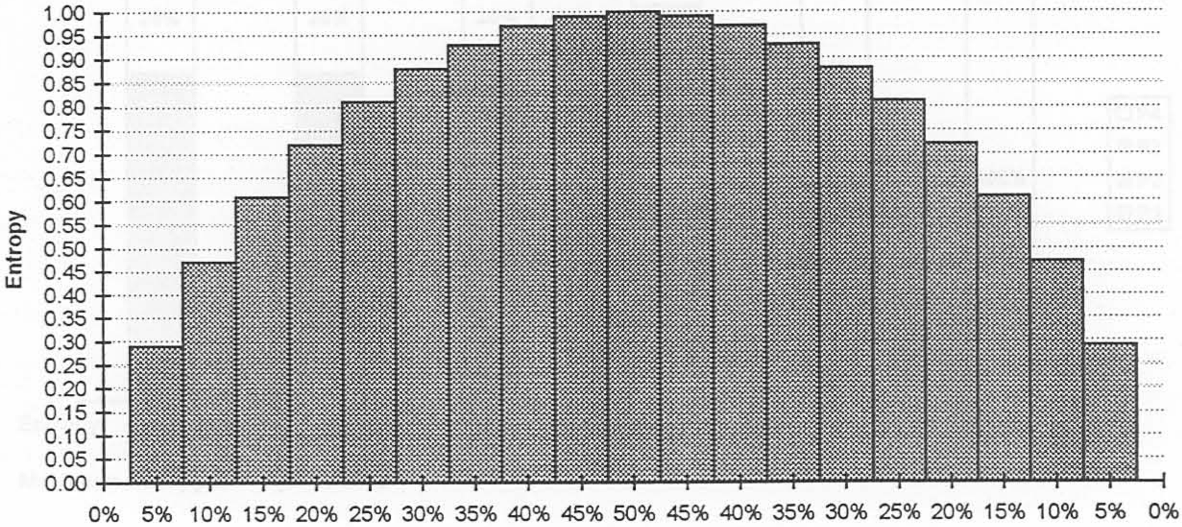

Figure 3-1. Entropy for a set of two symbols with changing probabilities

An important implication of a set with total parity is pointed out by Singh:

> For when all choices are equally probable and none prevails over any other, a person is, of course, completely free to choose among the various alternatives. As soon as one or more messages become more probable than any other, the freedom of choice is restricted and the corresponding information measure ... must naturally decrease. (Singh: 1967, p. 17)

The same principle applies to a set with a greater number of elements with varying probabilities. A message comprising four elements (p1, p2, p3, p4) with varying probabilities is demonstrated in Figure 3-2. Note how the summated entropy changes from maximum entropy when the four elements are equally probable to continuously lower entropy values as the probabilities become more unequal with some elements becoming more probable and others less probable. The first bar in the graph represents equal probability for the four elements, which become progressively more unequal.

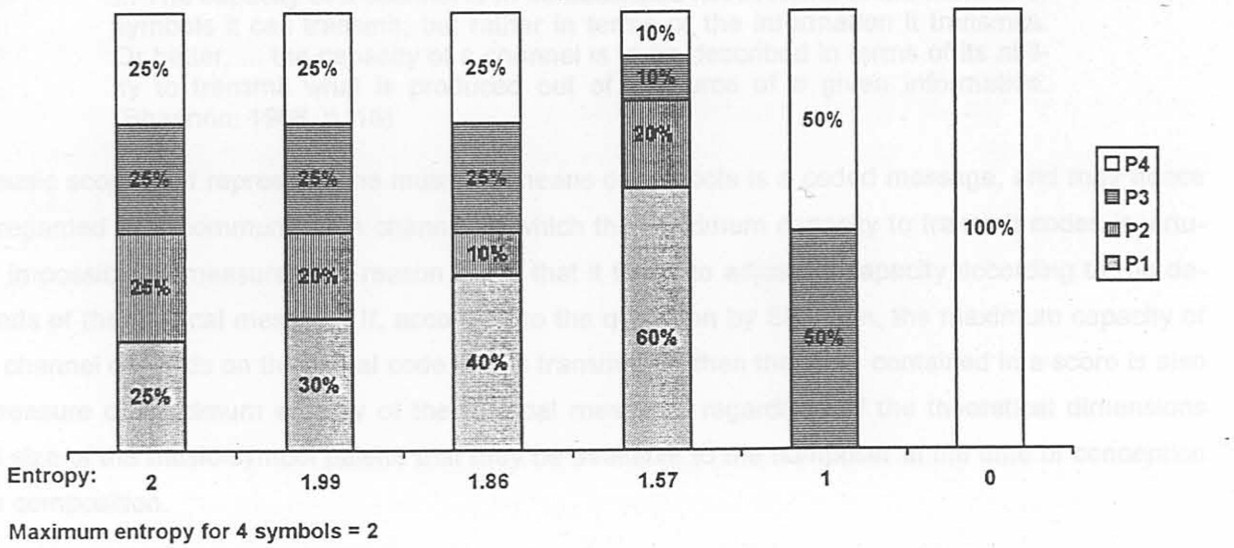

Figure 3-2. Entropy for a set of four symbols with varying illustrative probabilities

When the four elements of the example above are equally probable their entropy is at the maximum (2) and the quantity of information is at its maximum. As soon as the probability of some of the elements increases (reducing that of the other elements) the quantity of information they contain is reduced. Minimum information is generated as soon as one of the elements of the four has a probability of 100%.

3.3.1 Maximum entropy

From the foregoing examples it is clear that when parity exists between the elements of a set of finite symbols in a message there is a definite relationship between the number of symbols and the sum of the information generated by each. As the number of symbols in a message increases the maximum entropy also increases. Using Hartley's formula (Equation 3-2) it is possible to calculate the maximum entropy a symbol-set is capable of generating and, ignoring any noise that may be present, this value would also be the capacity of a transmitting system. This value is expressed by the formula $\mathrm{Log}_2 n$, where n represents the number of elements. As shown in Figure 3-1, a two-element message set has a maximum information value of 1 bit, while a four-element message set has a maximum value of 2 bits (see Figure 3-2). As freedom of choice decreases due to disparity of elements, the amount of information also diminishes and the influence of structural design becomes more prominent.

Shannon makes some important comments in this respect which are equally true when Information Theory is applied to music:

> ... The capacity of a channel is to be described not in terms of the number of symbols it can transmit, but rather in terms of the information it transmits. Or better, ... the capacity of a channel is to be described in terms of its ability to transmit what is produced out of a source of a given information. (Shannon: 1968, p. 16)

A music score, that represents the music by means of symbols is a coded message, and may hence be regarded as a communication channel of which the maximum capacity to transmit codes is virtually impossible to measure. The reason being that it tends to adjust its capacity according to the demands of the musical message. If, according to the quotation by Shannon, the maximum capacity of the channel depends on the actual code that is transmitted, then the code contained in a score is also a measure of maximum entropy of the musical message, regardless of the theoretical dimensions and size of the music-symbol palette that may be available to the composer at the time of conception of a composition.

3.3.2 Relative entropy

The formulae thus far illustrated the calculation of the actual entropy of a message as well as the calculation of the maximum entropy that such a message could generate (when all the symbols have parity value). As such, these independent values merely reflect specific information about the message itself as a discrete instance and the results would not be suitable for comparative study because each different message would produce different maximum entropy values. To make the results more meaningful and to allow it to be used for comparative evaluation the entropy of the message may be expressed as a ratio of the maximum entropy of the elements of the set.

The relationship between the entropy and the maximum entropy of a message is expressed as a percentage calculated with the following formula, where **H** denotes entropy:

Equation 3-11. Relative entropy

$$H_{rel} = \frac{H}{H_{\max}}(100)$$

3.3.3 Absolute and relative redundancy

When the information contents or entropy of a message reaches its maximum value, its predictability decreases diametrically because the selection of the symbols belonging to the set becomes increasingly more equal. Whereas entropy refers to unpredictability or information of a message, predictability is referred to as *redundancy*.

The redundancy of a message is calculated by subtracting the entropy of a message from the maximum entropy. For instance, if H(x) represents the entropy of a message with n symbols, the absolute redundancy is:

Equation 3-12. Absolute redundancy

$$R_{abs} = \log_2 n - H(x)$$

As with relative entropy, if redundancy is to be used for comparative purposes, a common unit base needs to be used. For this purpose relative entropy is subtracted from 1 to obtain relative redundancy. (If relative redundancy is expressed as a percentage, relative entropy is subtracted from 100):

Equation 3-13. Relative redundancy

$$\begin{aligned} R_{rel} &= \frac{\log_2 n - H(2)}{\log_2 n} \\ &= \frac{1 - H(x)}{\log_2 n} \end{aligned}$$

As the information set becomes more ordered, it becomes more predictable with an increasingly smaller information content until maximum redundancy is reached. Any major scale is an example of an element with maximum redundancy as each subsequent note in the scale is always fixed and fully predictable, at least as far as subsequent intervals are concerned. Major scales have a maximum entropy of $\log_2 7$, or 2.81 bits. However, a melody in a major key would not have equal distribution of the seven notes of the scale and could be expected to have an entropy value of less than 2.81 bits.