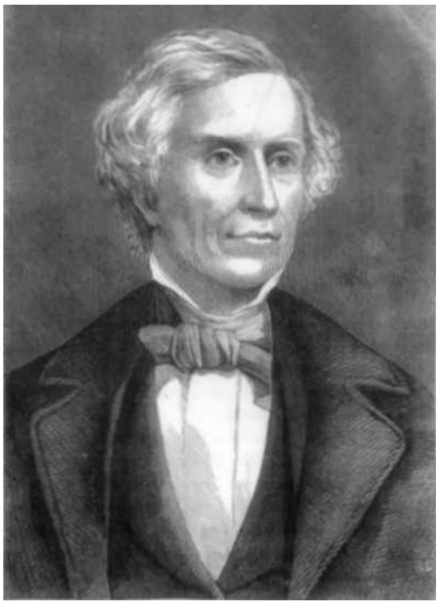
Morse meets Modern Communication Theory

"Science and Art are not opposed" - Samuel Morse c. 1838



Morse
An American Painter and Inventor
1791-1872

Samuel Morse was both a celebrated artist and passionate communications engineer, but it was his love and contribution to communications (Morse code) for which he is most remembered even though he was a most accomplished painter and a member of the Royal Academy for Art and many of his paintings are to be found in America located in the Washington, DC area. Morse devised his code together with two others; Joseph Henry and Alfred Vail. The physicist Henry provided the science behind their invention and Vail provided what today we would call the statistics, but more about that later.

The trio of inventors had jointly constructed and devised their code in around 1832, which was clearly long before any structured and rigorous communication theory had been developed and refined to those in use today, the most famous of which are the theorems by Huffman and Shannon. In this comparison I will compare the code produced by Morse, Henry and Vail to see how they compare by producing 'Morse' with today's analysis techniques.

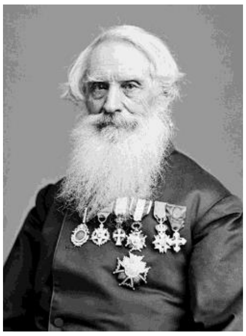


Photo by
Mathew Brady
c.1865-1880

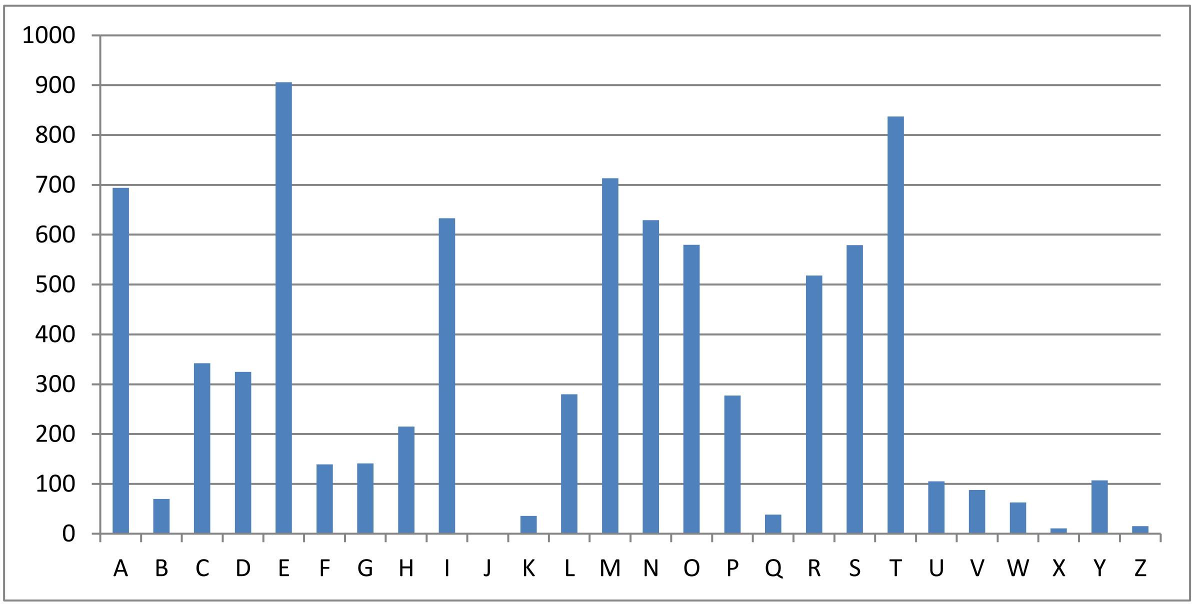
We shall see that Morse in conjunction with his co-inventors developed what emerges as a highly optimal codification of the English alphabet and when compared against what today we would term an optimum code derivation, it remains so – a truly remarkable outcome.

Modern communication theory is generally agreed to be based on the Huffman theorem or entropy encoding algorithm, a term that refers to the use of codes that vary in length as a function of their probability of occurrence and one which can be summarised as 'highly probable characters should have short codes e.g. DIT (e) and DAH (t) and the least probable characters should have progressively longer codes e.g. DIT-DIT-DIT-DAH (v) '.

Let's begin our comparison of Morse versus modern communication theories by going through Huffman's theory, which measures the disorder that exists in a system, in this case English alphabet characters, a term for which is better known by engineers as 'entropy'. In this context this is a measure of the smallest codeword length that is theoretically possible for a given alphabet with associated weighting factors applied to each. If we calculate Entropy for Morse Code the weighted average codeword length is 2.25 bits per symbol, which turns out to be only slightly larger than the calculated entropy of 2.205 bits per symbol, meaning Morse Code is not only optimal in the sense that no other feasible code could perform better, but its entropy is extremely close to the theoretical limit established by Shannon.

This was a truly remarkable achievement for an artist who devised the code on a crossing between England and America, albeit almost certainly derived through as a result of chance, good intellect and perhaps the application of common sense.

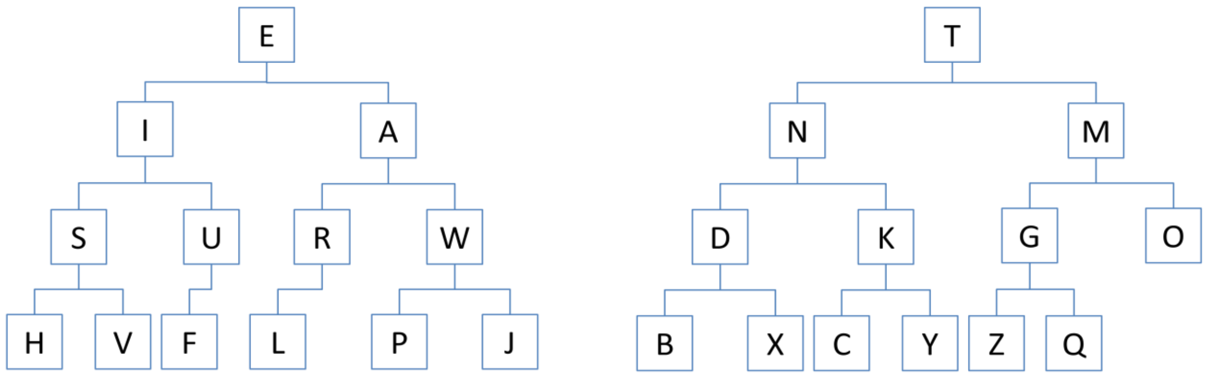
So why is the match between Morse and the Huffman theory so close? First let's examine a typical sentence or paragraph or book of English, for this purpose I've used an English dictionary and counted the number of pages for each letter of the English alphabet, here's the result:



Beginning to see a pattern? Well, take a look at the 'E' and 'T' character counts, see how they are larger than the rest, so let's roughly apply the Huffman Theory, which says assign the shortest codes to these characters, which for Morse is 'DIT - E' and 'DAH-T' to maximise efficiency and increase the chances of our message getting through whilst reducing the code throughput to a minimum.

And so, this is the basis of how Morse is a near perfect match to modern communications theory, his code is close to perfection for optimum communications throughput and efficiency, a statement that requires little convincing agreement for most Amateur radio operators!

We can deconstruct the whole Morse character code into the form of a 'Morse' (or more correctly a binary) tree and compare the results:



First of all, to use the 'Morse' tree, move left for a DIT and right for a DAH, so for example 'C' is DAH-DIT-DAH-DIT or RIGHT to the T, then LEFT to the N, then RIGHT to the K and finally LEFT to the C. The end of the character is denoted by the slightly longer inter-character delay. Trees like these are useful for software decoding of Morse as the reception of DIT's and DAH's is relatively easy as is the measurement of time gaps between characters, it is therefore (relatively) easy t o move down the tree as each character is received until the longer inter-character gap is received and hence we display the character at that point – easy!

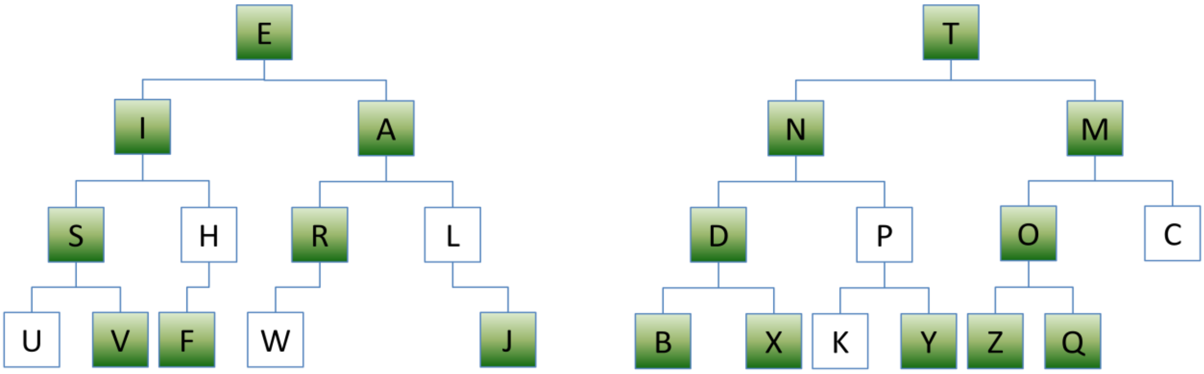
In Huffman terms we have constructed the code-tree from a knowledge of the code (Morse), but in using Huffman's coding technique we would normally start by analysing a piece of text, the larger the better to determine the frequency of occurrence of each character and once derived we would move further up the tree as the frequency of character occurrence increased, just like More did!

Now to compare code lengths to the characters counted in a random article, I used NASA's approach to software quality management as my source. From the article I determined the character distribution in the table below, where you will see 'E' and 'T' occur most frequently as expected. Now let's compare with what Morse devised, indeed E and T are at the top of the Morse tree making them the shortest code sequences and therefore are in accord with the Huffman Theorem. At the bottom of the tree are characters such as H, V, F through to Q which occur less frequently and Morse codified these with the longest code sequences, again exactly in-line with Huffman.

The number and frequency of occurrence of each character in the chosen article is as follows:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
694	70	342	325	906	139	141	215	633	0	36	280	713	629	580	277	38	518	579	837	105	88	63	11	107	15
8.32%	0.84%	4.1%	3.9%	10.86%	1.67%	1.69%	2.58%	7.59%	0.0%	0.43%	3.36%	8.55%	7.54%	6.95%	3.32%	0.46%	6.21%	6.94%	10.03%	1.26%	1.06%	0.76%	0.13%	1.28%	0.18%

Now to construct a binary tree using the Huffman technique to see how close our modern day code is to Morse code, here's the outcome, noting characters with the lowest probably occur at the bottom of the tree and conversely those with the highest at the top. Begin by listing out all those characters with (an arbitrary) frequency of occurrence of less than 2%, next move up the tree a level and repeat the exercise for less than 10% (arbitrary) and so-on until all characters have been analysed.



In our newly constructed binary tree derived from character occurrence / frequencies, we can see the shape and position and correlation denoted by the shaded boxes is very similar. The Huffman generated 'code' gives a 70% correlation with the original Morse code and assuming Morse used a similar, albeit empirical, approach to his code generation and also given that language constructs of the day (word usage) were almost certainly different that those is use today, it is reasonable to conclude we get the same result in constructing Morse's original code today with the benefit of a communication theorem.

The comparison between the distribution of character occurrences in the chosen article and the code sequences chosen by Morse is enough to demonstrate that Morse applied reasoned thought to his code, that they were certainly not random assignments by any means and he appears to have used his judgment, intellect together with perhaps some good luck to devise a code that went on to become one of the most versatile and widely used code in world history.

Samuel Morse is highly revered in the USA and if you ever visit New York City Central Park at 'Inventors Gate' there is a status of the American inventor and artist –holding a key in one hand and ticker-tape in the other, the latter being the original code medium.

Although an accomplished code inventor he was an equally accomplished and prolific artist able to produce what transpires to be a huge collection of paintings, samples of which can be found at the following locations in the USA, many of which I have been privileged to visit:

a. Samuel F.B. Morse at the Metropolitan Museum of Art, New York City
b. Museum of Fine Arts, Boston
c. Samuel F.B. Morse at the National Gallery of Art, Washington D.C.
d. Samuel F.B. Morse at the National Portrait Gallery, Washington D.C.
e. Addison Gallery of American Art, Andover, Massachusetts
f. Brooklyn Museum, New York City
g. Columbia Museum of Art, South Carolina
h. Corcoran Gallery of Art, Washington D.C.
i. Crystal Bridges Museum of American Art, Bentonville, Arkansas
j. Currier Gallery of Art, New Hampshire
k. Fenimore Art Museum, Cooperstown, New York
l. Gibbes Museum of Art, Charleston, South Carolina
m. Harvard University Art Museums, Massachusetts
n. Hunter Museum of American Art, Tennessee
o. Indianapolis Museum of Art, Indiana
p. Long Island Museum of American Art, History and Carriages, Stony Brook, New Yor
q. Mattatuck Museum, Waterbury, Connecticut
r. Museum of the National Academy of Design, New York City
s. Pennsylvania Academy of the Fine Arts, Philadelphia
t. Princeton University Art Museum, New Jersey
u. Samuel F.B. Morse at the Smithsonian American Art Museum, Washington D.C.
v. Samuel F.B. Morse at the Smithsonian Archives of American Art, Washington D.C.
w. Terra Foundation for American Art, Chicago
x. Worcester Art Museum, Massachusetts

So in conclusion, we have all probably taken Morse code for granted, undoubtedly many of us initially thinking the code was some random or ad-hoc use of DIT's and DAH's that we just learnt to achieve the Class-A licence, when actually it's a highly sophisticated and optimal code. Furthermore when we compare it against modern analytical and highly optimised reconstructions of Morse code, the outcome would be almost exactly the same, which is testament to the quality of the original invention and why it has stood the test of time for 175 years and more.

~

David Bird CEng MIET G6EJD