On the Entropy Bound for Optimum Homophonic Substitution

Valdemar C. da Rocha Jr.[1] and James L. Massey[2]

[1]CODEC, Dept. Electronics & Systems, P.O. Box 780, Fed. Univ. Pernambuco, 50711-970 Recife PE, BRAZIL
[2]Signal & Info. Proc. Lab., Swiss Federal Inst. Tech., ETH-Zentrum, CH-8092, Zürich, SWITZERLAND

Abstract — **A proof is given of the fact that, for optimum binary prefix-free homophonic coding, the entropy of a homophone is less than 2 bits greater than that of its antecedent regardless of the alphabet size for this antecedent.**

I. Proof of the Entropy Bound

Homophonic substitution is a cryptographic technique for reducing the redundancy of a message to be enciphered at the cost of plaintext expansion. In 1988, Günther [2] introduced *variable-length homophonic substitution*, in which the homophones (or "substitutes") for a particular letter of the message can have different lengths and different probabilities of selection, and showed that this technique can remove all redundancy from the homophonically coded sequence, producing what Shannon calls a *strongly ideal* cipher [1] while also reducing plaintext expansion.

For simplicity, we consider binary homophonic coding of the output sequence $U_1, U_2, U_3, \ldots$ of an L-ary discrete *memoryless* source, for which the homophonic coding problem reduces to that for a single random variable U taking values in $\{u_1, u_2, \ldots, u_L\}$ where $L \geq 2$ and all L values have nonzero probability. The homophone V for U takes values in the set $\{v_1, v_2, \ldots\ \}$, which may be finite or countably infinite, and is characterized by the fact that for each j there is exactly one i such that $P_{V|U}(v_j|u_i) \neq 0$. The homophone $V = (X_1, X_2, \ldots, X_W)$, where X_i is a binary random variable and where W is also a random variable. It is required that $(X_1, X_2, \ldots, X_W)$ be a *prefix-free* encoding of V, i.e., that the homophones be distinct and none be the prefix of another. The homophonic coding is *perfect* if the components $X_1, X_2, \ldots, X_W$ of the homophone V are independent and coin-tossing, and is *optimum* if it is perfect and minimizes $\mathrm{E}[W]$ among perfect homophonic codings.

It was shown in [3, *Proposition 3*] that a binary homophonic coding scheme is optimum if and only if, for every i, the probabilities $P_V(v_j) = P_{VU}(v_j, u_i) = P_{V|U}(v_j|u_i)P_U(u_i)$, $j \geq 1$, of the homophones are equal (in some order) to the terms in the decomposition of $P_U(u_i)$ as a finite sum of distinct negative integer powers of 2, when this is possible, and as an infinite such sum otherwise.

In [3], the upper bound $H(V|U) < 2$ bits, independent of L, is asserted for optimum binary homophonic substitution, but the simple "proof" given there is fallacious. One purpose of this paper is to present a (valid) proof. Another purpose is to strengthen the upper bound by including its dependence on the number of terms in the expansion of $P_U(u_i)$.

Theorem 1 *For optimum binary homophonic coding of an L-ary random variable U, the conditional entropy $H(V|U)$ of the homophone V given U is less than 2 bits.*

Proof: Let $p_1, p_2, p_3, \ldots$ be the (possibly infinitely many) nonzero terms in decreasing order of the conditional probability distribution $P_{V|U}(.|u_i)$. Because the binary homophonic coding is optimum, $p_{j+1} \leq p_j/2$ for all $j \geq 1$, which implies that $1/2 < p_1 \leq 1$. It suffices to consider $1/2 < p_1 < 1$. Letting $H(V|U = u_i) = h(p_1, p_2, \ldots)$ denote the entropy in bits of the probability distribution $(p_1, p_2, p_3, \ldots)$, we have

$$H(V|U = u_i) = h(p_1) + (1-p_1)h(p_2/(1-p_1), p_3/(1-p_1), \ldots).$$

The condition $1/2 < p_1 < 1$ now implies that

$$H(V|U = u_i) < 1 + \frac{1}{2}h(q_1, q_2, \ldots)$$

where $(q_1, q_2, q_3, \ldots) = (p_2/(1-p_1), p_3/(1-p_1), \ldots)$. Thus $q_{j+1} \leq q_j/2$ for all $j \geq 1$, which implies that $1/2 < q_1 \leq 1$. Iterating this argument gives

$$H(V|U = u_i) < 1 + \frac{1}{2}(1 + \frac{1}{2}(1 + \frac{1}{2}(1 + \ldots\quad))) = 2, \tag{1}$$

which implies that $H(V|U) < 2$ bits. □

Suppose now that $P_U(u_i) = N_i/2^r$ for positive integers N_i and r with N_i odd. It suffices to consider $3 \leq N_i < 2^r$, which implies that $r \geq 2$ and that the number, n_i, of terms in the expansion of N_i as a sum of distinct nonnegative integer powers of 2 satisfies $2 \leq n_i \leq r$. This further implies that the number of terms in the sum in (1) is $n_i - 1$ and hence that this sum is at most $2 - 2^{-(n_i-2)}$. We have proved the following bound which, for $n = 2$, implies $H(V|U) < 1$ bit.

Theorem 2 *For optimum binary homophonic coding of an L-ary random variable U such that $P_U(u_i)$ can be expressed as a sum of at most n distict negative integer powers of 2 for $i = 1, 2, \ldots, L$, the conditional entropy $H(V|U)$ of the homophone V given U satisfies*

$$H(V|U) < 2 - 2^{-(n-2)} \quad \textit{bits,}$$

if $n > 1$ and $H(V|U) = 0$ if $n = 1$.

References

[1] C.E. Shannon, ”Communication Theory of Secrecy Systems”, *Bell System Tech. J.*, vol. 28, pp. 656-715, Oct., 1949.

[2] Ch.G. Günther, "A Universal Algorithm for Homophonic Coding", pp. 405-414 in *Advances in ryptology-Eurocrypt'88*, Lecture Notes in Computer Science, No.330. Heidelberg and New York: Springer, 1988.

[3] H.N. Jendal, Y.J.B. Kuhn and J.L. Massey, "An Information-Theoretic Approach to Homophonic Substitution", pp. 382-394 in *Advances in Cryptology-Eurocrypt'89*, Lecture Notes in Computer Science, No.434. Heidelberg and New York: Springer, 1990.

Acknowledgements

The first author's research was supported by the Brazilian National Council for Scientific and Technological Development (CNPq) under Grant No. 304214/77-9.