How to Calculate Entropy-Rates from Match-Count Profiles

Addendum to
"Search-Effectiveness Measures for Symbolic Music Queries " in Very Large Databases" Sapp, Liu, & Selfridge-Field
ISMIR 2004, Barcelona, Spain

Let:

- M = database size (number of entries).
- $E(n)$ = the average expected number of matches given a query symbol length of n.
- $R = 2^H$ where H is the entropy (rate) of the data feature being examined. The random features of the database are assumed to have a constant entropy rate.

In general,

$$E(n) = \frac{M}{R^n}$$

For example, a random variable with uniform distribution over three possible symbol states has this expected number of matches:

$$E(n) = \frac{M}{3^n}$$

since the entropy rate will be $\log_2 3$ and $R^n = 3^n$ since $R = 2^{\log_2 3} = 3$. If the length of the query is one symbol, then $M/3$ (anchored) matches are expected. If the length of the query is two symbols, then $M/3^2 = M/9$ matches are expected, and so on.

However, the Match-Count Profiles are generated with this expectation function:

$$E(n) = \frac{M-1}{R^n} + 1$$

since one match is guarenteed to be found in the database because the query string is generated from an actual target match found in the database. This guarenteed match yields the $+1$ term in the above equation. The other term is just the expectation of finding the query string in the rest of the database excluding the known match.

Notice that for small n, the $(M-1)/R^n$ term dominates the match expectation function. This is the useful term since it contains the entropy (rate), H.

Also note that for large n, the $+1$ term dominates. The $+1$ term is not very interesting, so cancel it out and display the log of the difference between two adjacent expectation functions: $E(n) - E(n+1)$.

$$E(n) - E(n+1) = \frac{M-1}{R^n} - \frac{M-1}{R^{n+1}}$$

Now simplify the right side of the above equation, and take the $\log_2$ of both sides:

$$E(n) - E(n+1) = \frac{R(M-1) - (M-1)}{R^{n+1}}$$

$$E(n) - E(n+1) = \frac{(R-1)(M-1)}{R\ R^n}$$

$$\log_2[E(n) - E(n+1)] = \log_2\left[\frac{(R-1)(M-1)}{R}\right] - \log_2 R^n$$

Let $y = \log_2[E(n) - E(n+1)]$ which is the y-axis of the plot. Let $b = \log_2[(R-1)(M-1)/R]$ which is a constant (assuming H is constant).

Therefore, the equation now becomes:

$$y = b - \log_2 R^n$$

But $R = 2^H$, so the equation becomes:

$$y = b - \log_2 2^{Hn}$$

Now, let $x = n$ where x is the x-axis of the plot, so that the equation becomes:

$$y = -Hx + b$$

This is the equation of a line, where the slope of the line is $-H$.

Thus the entropy rate can be calculated directly from the slope of the plot for match-count profiles when they are plotted using the formula:

$$y = \log_2[E(n) - E(n+1)]$$

where $E(n)$ is the averaged experimentally measured match counts for a length-n symbol query.

The slope of the resulting plots can be aquired by averaging the differences between y-values of points at the beginning of the curve down to the point where $y = 0$. Below this level, the slope of the line may deviate significantly due to quantization noise.