1 Confusion Entropy

$\mathcal{X} = \{$ m, h, f, θ, ɚ, j, ɝ, ə, ɛ, o, ...ɔ $\}$. — TIMIT inventory

$\mathcal{Y} = \{$ f, m, n, ɛ, k͡p, k, b, t, ũ, d͡ʒ...ɔ $\}$. — Yoruba inventory

The conditional probability mass function for a predictions given a golden in dataset:

$$p(x \mid y) = \Pr[X = x \mid Y = y], \qquad x \in \mathcal{X},\ y \in \mathcal{Y}.$$

The maximum likelihood estimate

$$\widehat{p}(x \mid y) = \frac{n_{x,y}}{\sum_{x \in \mathcal{X}} n_{x,y}}.$$

fills in the correspondence matrix. (each column sums up to 1)

Confusion entropy

$$H_y \;=\; H(X \mid Y{=}y) \;=\; -\sum_{x \in \mathcal{X}} \widehat{p}(x \mid y)\, \log_2 \widehat{p}(x \mid y),$$

generalises how wrong / uncertain on average the model performs per Yoruba sound y, in bits.

Then we can also nicely give an overall assessment of our model with conditional entropy.

$$H(X \mid Y) \;=\; \sum_{y \in \mathcal{Y}} \widehat{p}(y)\, H_y.$$

ofc we'll have to avoid 0 values. Refer to my goat Bentz's workshop (2022)

What's important about this formula is that it has best value (0 bits) if there is a 1 to 1 correspondence to Yoruba–English. At the same time it doesn't matter what in IPA is this correspondence: could be even t͡ʃ to ũ.

To account for closeness the method with least maths and most linguistics is FWPER.

2 Feature-Weighted Phoneme Error Rate

Right now weighted Levenstein Lev(o, g) for sequences o, g is a sum of Costs c of all symbols p, q in the alignment $A(o, g)$:

1. substitutions $c(p \to q) = 1 - feature\ similarity(p, q)$
2. deletions $c(p \to \epsilon) = 1$
3. insertions $c(\epsilon \to q) = 1$

Since it's a sum, it's maximum grows with the phoneme length of the longer sequence. Confusion/correspondece counts $C_{p,q}$ are computed from frequencies of alignments A. Deletions and insertions count as substitutions by an empty symbol.

Then weighted PER is the expected number of operations per phoneme, expected cost per phoneme. So, because Lev is a sum, the easiest way to compute **wPER** for our sequences is $\mathbb{E}_{o,g}[c] = \frac{Lev(o,g)}{|g|}$. The expectation compared to the gold standard sequence g is relevant.

Overall formula for our dataset $\{(o_i, g_i)\}_{i=1}^{M}$:

$$\text{wPER} = \mathbb{E}[\frac{Lev(o, g)}{|g|}] = \frac{1}{M} \sum_{i=1}^{M} (\frac{1}{|g_i|} \sum_{(p,q) \in A(o_i, g_i)} c(p \to q))$$

There is another way to compute it from our confusion matrices, which is more accurate in averaging.

$$\text{wPER} = \frac{\sum_{p \in \mathcal{Y} \cup \{\varepsilon\}} \sum_{q \in \mathcal{X} \cup \{\varepsilon\}} C_{p,q} \cdot c(p \to q)}{\sum_{p \in \mathcal{Y}} \sum_{x \in \mathcal{X} \cup \{\varepsilon\}} C_{p,q}}$$

We weigh counts by the cost of it in all cells, then divide by the total gold tokens.

3 Subtracting the Errors intrinsic to English predictions

These formulas assume that the labels in English and Yoruba are the same. And their values will depend on what values we choose as "the same". Suppose we defined a set of common sounds $\mathcal{T} = \mathcal{X} \cap \mathcal{Y}$ and smh normalised the probablities, just pretending other sounds don't exist. Then two options:

1. KL divergence (entropy-based)
2. Subtract the probabilities we would have found in Yoruba if it was confusing like in English: let X, Y — random vars defined over $\mathcal{T}$ for predicted and gold standard Yoruba, and $p_E(x \mid y),\ x, y \in \mathcal{T}$ — English probabilities of a predicted sound x for a label y from the confusion matrix. Then $p_{exp}(x, y) = P(Y = y) \cdot p_E(x \mid y)$ gives the Yoruba expected probabilities, and the difference between observed and expected is $P(X = x|Y = y) \cdot P(Y = y) - p_{exp}(x, y)$.

4 Mutual Information of Logprobs

Y — Yoruba phone, $E = p(e \mid x)$ — rand var over English phones for a given token x. Now the joint is also not just confusion probabilities, but average of E.

$$\widehat{p}(y, e) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{y_i = y\}\ p(e \mid x_i)$$

$$\widehat{p}(y) = \sum_{e} \widehat{p}(y, e) \quad \widehat{p}(e) = \sum_{y} \widehat{p}(y, e)$$

We can make a heatmap of how much information the assosiation between English prediction e and Yoruba golden y carries.

$$PMI(y, e) = log_2\ \widehat{p}(y, e) - log_2\ \widehat{p}(y) - log_2\ \widehat{p}(e)$$

And normalise it between $[-1; 1]$: $\frac{PMI(y,e)}{log_2\ \widehat{p}(y,e)}$. For one overall score we can find $I(Y, E) = H(E) - H(E \mid Y)$

5 Embedding space

The previous formula already uses results after softmax, averaged over time step. These are PPGs and are viewed as embeddings. For English we compute the average per golden phone in TIMIT. $\mu_x \in \mathbb{R}^{|\mathcal{X}|}$. Or for example per golden phone per batch. Then for each Yoruba gold label y we also average $\mu_y \in \mathbb{R}^{|\mathcal{X}|}$. Then reduce the dimensions with UMAP so that it preserves the distances ***in English***, and transform the Yoruba means the same way. Point Yoruba gold labels in each Yoruba point and in the circles for English clusters.