Zero-Shot Cross-lingual Phoneme Recognition from Yoruba to English

Aaron Bahr and **Nikita L. Beklemishev** and **Haejin Cho** and **Kai Seidenspinner** and **Ilinca Vandici**
Universtität Tübingen

Abstract

In this paper, we pre-train an acoustic phoneme recognition model on the TIMIT dataset and evaluate its performance on the Yoruba portion of the Common voice data. We make use of the ResNet-BiLSTM together with CTC loss function architecture, allowing us to forego the need for time-aligned input data, and examine the models performance after transfer through a linguistically motivated feature weighting metric, reaching a 0.36 average Phoneme Error Rate on the Yoruba set. Leveraging the predictions produced by our model, we take an in-depth look at the effects of learning and transfer. All codes are available in our Github.

1 Introduction

Efficiently training an ASR system requires a rich, ideally time-aligned dataset. For low-resource languages, despite documentation efforts, exploiting the properties of transfer learning by pre-training on another, high-resource, language remains a sensible option. (Xu et al., 2021) For the purpose of zero-shot evaluation, picking a set of languages with similar phoneme inventories remains the practice yielding the best performance. We chose to work on transferring American English (< West Germanic < Indo-European) to Nigerian Yoruba (< Volta-Niger < Atlantic-Congo), whose phoneme inventories overlap to a large extent, despite different areas and *phyla* (Appendices A B).

In our case, we focus on producing a consistent, generalized phonemic transcription, conditioned only on acoustic segments. Along with obtaining good performance on both languages, we aim to propose an efficient evaluation metric by employing a linguistically sound feature-weighted version of Phoneme Error Rate (PER). We attempt to disentangle training errors from transfer errors by presenting a holistic view of our results, analyzing the confusion patterns in our predictions through metrics like cross-entropy and normalized PMI.

In recent years, multilingual ASR has gained more and more attention. Models like Radford et al. (2022) perform well on low resource languages, given the diversity and size of the training dataset which allows them to form representations for a variety of phonemes. However, as models grow in size, interpretability becomes a more delicate task. Acknowledging this, we choose to focus on one language pair that instead allows us to gain a deeper insight of the mechanisms at work during transfer.

In terms of the model architecture, using the ResNet-BiLSTM model along with *Connectionist Temporal Classification Loss* ensured that we could evaluate on non time-aligned data, with the additional benefit of enhancing adaptability to context, which will be discussed in section 3.

1.1 Background

Transfer learning occurs when using a model to carry out a task different from the one it has originally been trained on. This can be done in different ways, for instance by using the pre-defined representations a model has formed on downstream tasks, or by fine-tuning the model weights in order to skew the distribution towards that of the new task (Yadav and Sitaram, 2022). This last variant is particularly popular when training multilingual ASR models, which due to compute progress, have becoming more common (Radford et al., 2022). In our case, we examine zero-shot transfer by running inference directly on the new task, without the intermediary of pre-defined feature mappings.

With architectures and compute power improving, models utilizing the property of transfer at the pre-training stage in order to achieve multilingual representations have become increasingly common. Xu et al. (2021) is one example. However, this paper intends to enable transfer-learning without explicitly providing pre-defined feature mappings. We also believe that focusing on separate languages

for training and evaluation allows us to gain a deeper insight of the mechanisms at work during transfer.

Theoretical capabilities It is commonly understood that the more similar the probability distribution under the new task is to that of the original task, the higher the likelihood of successful transfer learning. For the case at hand, it is facilitated by the fact that languages sound similar, and the knowledge of human vocalizations is shared. Sound is just the input part of the model. A common **function** from the sound to the target sequence is the necessary condition for transfer. However, the arbitrary variation between different orthographies makes for a poor generalization target. Luckily, for linguistics, the relevant representation of speech is phonemes, which tend to correspond to sound types—which are, by and large, commonly shared. Out of all the sounds that the vocal tract can produce, languages use only a small subset, and some sounds are much more common than others. (Hayes, 2009, p. 6) The shared knowledge of sound types can be described using IPA symbols. (Ladefoged, 2011) The IPA presents an established framework through which this similarity is defined, and enables us to unify transcriptions across languages, ensuring that the learned function maps from the acoustic domain to a common representation.

Furthermore, we analyze the **confusion and correspondence** patterns that appear in the model's transfer performance. Seeing where the generalizing function from acoustics to phonemes inherited from English fails and withstands being applied to Yoruba, we speculate on the variation and universals in these phonological systems.

1.2 The inventories problem

As identified, the codomain of the sounds to phonemes function will need to at least partly overlap across both tasks in order to obtain an interpretable output. The phonemic systems of Yoruba and English differ (Appendices A, B; (Adesola, 2006; Moran et al., 2014; Akinbo et al., 2022)), which poses a task of relating English IPA predictions to the Yoruba IPA gold standard: What is a matching sequence, given that the alphabets only partially correspond? Feature theory (Chomsky and Halle, 1968) supplies a graded notion of segment similarity. Each sound has a feature representation indicating which natural classes it belongs and does not belong to. Although sounds vary in how many features from the overall set they specify, it is convenient to represent each sound as a vector over the entire feature inventory, with each element coded as positive (+1), negative (1), or unspecified (0). The Hamming distance between two such vectors, restricted to non-zero dimensions and normalized by dimensionality, then serves as the dissimilarity measure (see Appendix C). This approach has been used for a long time in fields like cognate detection (according to Jäger (2013, p. 283)) and dialectometry. (Nerbonne and Heeringa, 2010) Among many notable databanks aggregating the knowledge of features, such as PHOIBLE (Moran et al., 2014) and SoundVector (Rubehn, 2024), we chose PanPhon (Mortensen et al., 2016) for its balance between the quality of representation and availability. The details of the implementation are discussed in Section 4.2.

1.3 Possible disruptors of English to Yoruba transfer

Apart from the differences in inventories discussed above, we hypothesise that several other factors will further complicate the downstream performance. (i) Above all, the IPA does not aim to fully describe language's phonemic system, in a structuralist sense, but to represent "universal" phonological features. (van der Hulst, 2017, p. 40) Hence, for instance, even identical IPA symbols may partition the vowel space differently in English and Yoruba. (ii) The set of universal phonological features is not universally agreed upon. The symbols /ɔ, ɛ/ occur in both languages, but in English they instantiate the traditional [–tense] contrast (Moran et al., 2014), whereas in Yoruba they pattern with [–ATR] (Allen et al., 2013). PanPhon (Mortensen et al., 2016) equates the Germanic lax/tense with ATR, although English is generally analysed as controlling centralization/height rather than tongue-root advancement, and many ATR languages show no such centralization (Ladefoged and Maddieson, 1996; Przezdziecki, 2005). Even though Przezdziecki (2005) shows that Yoruba [–ATR] vowels do appear centralized, we predict the transfer of /ɔ, ɛ/ to be somewhat noisy, as in fact they constitute different features in the source and target languages. (iii) Perceptually, a single phoneme may be decomposed into a sequence in context of another phonological system. Yoruba has double-articulated sounds, i.e. /k͡p/,

/ĩ/, which could be mapped to sequences like /k p/ or /i n/). (iv) Phonological and especially phonotactic constraints are less universal than inventories (Maddieson, 2010), yielding different sound sequence distributions. This hinders transfer-learnin of temporal features. For example, English frequent diphthongs might make the vowel–glide sequences over-predicted relative to Yoruba, which instead has constraints on codaless syllables and traits of tongue root vowel harmony. (Przezdziecki, 2005; Allen et al., 2013; Akinbo et al., 2022) (v) Ideally, the model should learn the contrasts of the target language, but it is in fact trained for distinctions of the training language. For example, Yoruba treats [n ∼ l], [ɾ ∼ ɹ] as allophonic, and the shallow English transcription distinguishes allophones like [ɨ ∼ i], and such cases will noise the correspondence between the predictions and Yoruba labels.

2 Dataset

Training set For training we use the TIMIT ASR corpus. Garofolo et al. (1993) includes 6300 utterances recorded by 630 speakers of 8 major English dialects across the Unites States. Annotations were done according to the customized IPA convention based on ARPAbet (Carnegie Mellon University, 1998) (please refer to Appendix A). As both English and Yoruba are pluricentric languages, training the model on a variety of dialects will likely improve the performance in downstream tasks. Even though TIMIT confines itself to the 1980s US language roof, it represents the existing dialectal variation well. This phonetic variation is necessary for learning to generalize over the variation in sounds, which is particularly relevant for transfer learning. Another advantage of TIMIT is its ubiquity in ASR studies, which gives us confidence in yielding baseline results, comparable with other works in the area.

Preprocessing the training set For usage, we first concatenated the train, validation, and test splits before again randomly dividing it into train and validation splits (with a 75 to 25 ratio). We perform a Fast Fourier Transform on the audio data, collecting 39 log-scale MFCC features, including first and second-order derivatives.

The TIMIT alphabet contained 63 unique labels in total. In order to reduce prediction complexity and ensure the compatibility of the phonemic representation between both languages, making way for the IPA mapping process, we merged or split several labels. This included allophones not annotated in the Yoruba corpus: <ax-h> /ə̥/ → /ə/, syllabic sonorants, e.g. <eng> /n̩/ → /ŋ/. Closures <dcl> /dˈ / and the following releases /d/ were joined into one label: we did not expect systematic unreleased closures in **open-syllable** Yoruba. (Adesola, 2006) In the end, 15 label types were merged. As for the splitting the combinatorially large inventory of English diphthongs, we split them into vowel–glide sequences, keeping the vowels from the IPA convention: thus <oy> /ɔɪ/ became <ao y> /ɔ j/. This step was necessary since Yoruba does not have diphthongs. (Przezdziecki, 2005) We also concluded that splitting them will not perplex the prediction given that the CTC decoding does not need time alignment, and that our evaluation ignores word boundaries.

Evaluation set The Yoruba section of Common Voice was used as a test dataset. Common Voice is a multilingual crowd-sourced corpus aimed for Speech Recognition purposes. (Ardila et al., 2020) The audios were recorded by certified native speakers of each language. Annotations are suggested and later validated by other native speaker users via votes. There are 3.4k samples in total, with each sample including a MP3 file, speaker ID and audio transcription written in Yoruba orthography. The dataset also includes data from different dialects. To an extent, this accounts for the performance gaps that are shown to arise between standard Yoruba and other dialects in NLP tasks. (Ahia et al., 2024)

Processing the evaluation set The sets of train (1.4k), validation (913) and test (1.1k) were again concatenated and used together for testing. We kept the original threshold of downvotes for invalidated samples. As for the audio data, features were extracted under the same configuration as for the TIMIT dataset. Since Common Voice only provides an orthographic representation of the sentence, it was necessary to implement a grapheme-to-phoneme to obtain an IPA representation that could then be compared with the model output. For this purpose, we used `Epitran` package (Mortensen et al., 2018) (for discussion see Section 6) and pre-processed the resulting strings. Word boundaries and pauses were removed. We also ignored the tone annotation (˧, ˥, ˩). We have removed the marginally phonemic /ɔ̃, ŋ/ from the inventory, merging with their positional allophones

/ã/ and /n̩/. (Allen et al., 2013; Akinbo et al., 2022): 2 Another marginal /ɛ̃/ remained. The data also turned out to contain one occurrence of dialectal <ụ> /ʊ/ (Przezdziecki, 2005), which is too small to generalize from, so we removed it as well from the evaluation for clarity. The resulting Yoruba IPA inventory is available in Appendix B.

3 Model and Traning

3.1 ResNet-Bi-LSTM model

Our model is based on that of Dhakal et al. (2022), which used ResNet-BiLSTM model for Nepali Speech Recognition. The reason why we chose this architecture as our base reference was (1) neural-network-based models are relatively easy to train and light-weight in terms of memory, (2) we wanted to include residual connections which is largely used in transformers as well as deep neural network models, and (3) to explore whether a model that is trained from scratch using monolingual data can also perform well on zero-shot cross-lingual speech recognition task unlike Xu et al. (2021). We will first briefly describe overall architecture and setting of the original model and then list our adjustments.

The model starts with an initial CNN layer and 5 consecutive ResNet blocks. These make up the 'ResNet Encoder', which aims to capture local dependencies. Each block consists of 2 unit blocks, with each unit block containing an initial convolution, Batch Normalization, using PReLU in lieu of the activation function. Every convolution layer has 50 feature maps with kernel size 15. The key idea behind residual connections is that adding the original input at deeper stages in the forward pass will result in grounding the representations. Ravanelli and Bengio (2019) have attested that residual connections noticeably stabilize and optimize deep neural network training, which Dhakal et al. (2022) confirms therefore extending to audio data. A Bi-LSTM encoder part then follows the Residual Encoder. Bi-LSTM is designed to reflect distinctive temporally-related features in two opposite directions. Dhakal et al. (2022) has two RNN layers with its dimension being both 170. As final layers, two dense layers and ReLU activation takes the output of Bi-LSTM and projects the 340-dimensional output onto the phonemic embedding space.

We have made several crucial changes based on multiple experiments in order to prevent overfitting and to restrict the depth of the model under

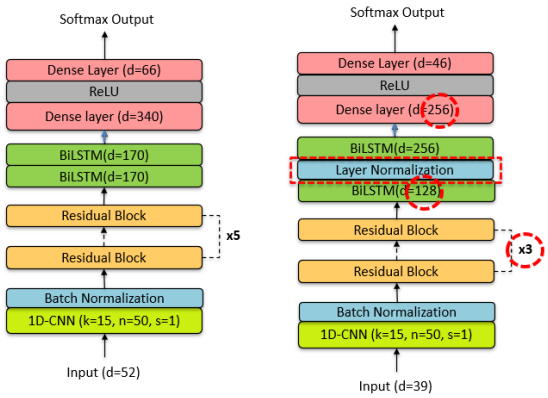

Figure 1: Left: model architecture of (Dhakal et al., 2022) , Right: Adjusted model architecture for English-Yoruba cross-lingual phoneme recognition. Changes highlighted with a red circle.

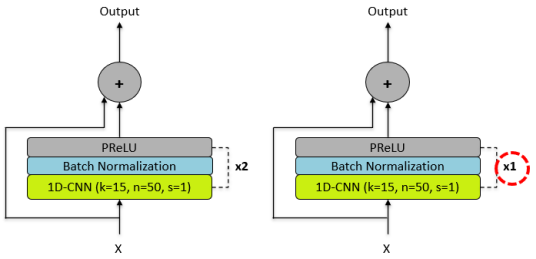

Figure 2: Left: residual block of (Dhakal et al., 2022), Right: Adjusted residual block for English-Yoruba cross-lingual phoneme recognition. Changes highlighted with a red circle.

zero-shot cross-lingual task setting. (1) We downsize the number of ResNetEncoder layers from 5 to 3, and change the number of unit blocks from 2 to 1, meaning that we reduced the original model to a third of its depth. All other ResNet encoder settings remain fixed as in Dhakal et al. (2022). (2) We also diminished hidden dimensions of Bi-LSTM encoder and dense layer, from 170 to 128 as to hidden dimension of RNN layer and from 340 to 256 as to dense layer dimension. (3) Finally, we introduced a stronger dropout rate to the dense layer, with the aim of facilitating cross-lingual application. Refer to the following visualization that marks key differences between Dhakal et al. (2022) and our model.

We employed the Connectionist Temporal Classification (CTC) loss as our training criterion. Unlike conventional frame-level cross-entropy, CTC does not require pre-aligned input-label pairs,

which makes it especially suitable for low-resource languages such as Yoruba, where alignment information is often unavailable. The key idea of CTC is to introduce a blank symbol and permit label repetitions so that multiple frame-level alignments can correspond to the same output sequence. During training, CTC computes the negative log-likelihood of the target sequence from total probabilities of all valid alignments:

$$L_{CTC} = \sum_{(X,Y)\in D} -\log P_{CTC}(Y \mid X)$$

where the probability of a target sequence Y given the input X is defined as

$$\begin{aligned} P_{CTC}(Y \mid X) &= \sum_{A\in B^{-1}(Y)} P(A \mid X) \\ &= \sum_{A\in B^{-1}(Y)} \prod_{t=1}^{T} p(a_t \mid h_t) \end{aligned}$$

with $A = (a_1, \ldots, a_T)$ being a frame-level alignment, B the collapse function that removes blanks and repeated labels, and h_t the hidden representation at time t. The blank symbol ϵ needs to be manually added to the alphabet so that the model can output blanks during training and inference.

At decoding time, we select the output sequence with the highest conditional probability, obtained by summing over all valid alignments, and the CTC collapse operation removes duplicates and blanks to produce the final prediction sequence $\hat{Y}$.

$$\hat{Y} = \arg\max_{Y} P_{CTC}(Y \mid X).$$

In our experiments, however, we use plain greedy CTC decoding, without an external language model. This choice is motivated by the zero-shot setting, where we want the model to remain agnostic to language-specific temporal patterns. It is worth mentioning that, along with the purely practical lack of an aligned corpus, CTC can also in some cases present an advantage over architectures that rely on alignment. Indeed, Hannun (2017) argues that aligned data might induce some biases, especially if the corpus is not diverse enough in regards to the kind of speech contexts available.

3.2 Train

Train settings were set through multiple experiments. The number of epochs is `50` and batch size was `64`. We have adopted Adam as our stochastic gradient optimizer and the weight decay set to `1e-4`. A plateau-based learning rate scheduler was used, with an initial learning rate of `1e-3`. The total number of parameters is `0.8` million. With the (hyper)parameter settings given above, training was noticeably stable and robust to overfitting. Both training and validation loss and PER consistently decreased and we halted training at the 16th epoch where Train PER reached `0.0223` and Validation PER reached a similar figure, `0.0339`.

4 Results

As seen from the final PER results, the model shows a promising performance on the English dataset. We now introduce the framework we used for evaluation, before delving deeper into the results produced by transfer.

4.1 Evaluation Method

We evaluate the output sequence with relation to the corresponding gold labels sequence, by computing its Feature-Weighted Phoneme Error Rate (fwPER). Like the regular PER, this measure is a Levenshtein distance between the predicted and golden sequences, normalized by the length of the golden sequence. Unlike in classic PER, the substitution cost is not fixed, but based on the phonological information vectors we built using PanPhon for each pair of sounds. For two phonemes, it is the Hamming distance of their feature vectors, normalized from 0 to 1 (Appendix C). Thus one can roughly think of it as expected dissimilarity per phoneme token. We took the liberty of removing some of the PanPhon features, those that reflect the distinctions not relevant in describing the English and Yoruba inventories: [`sg`, `velaric`, `long`, `hitone`, `hireg`]. We built on and reworked the original PanPhon source code for fwPER to be able to extract the alignment and also optimized its alignment calculation algorithm for speed. The main assumption behind our evaluation scheme is that the alignments achieved with this method are not only minimally costly, but also intuitively correct. As we see in Discussion, it is not always the case.

4.2 Inference

During inference, we observed that our model had a tendency to over-generate, making the output consistently longer than the gold labels. A look at

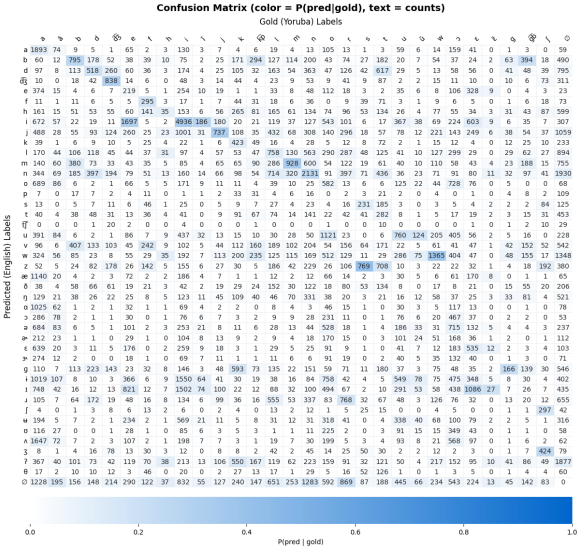

Figure 3: English-Yoruba Confusion Matrix with Posterior Probability

the Appendix D reveals that this is a transfer defect of the model. Some of the worst performances arise from this over-generation. Whenever this happens, the inserted sounds tend to be somewhat rare phones, repeating: /ʔ, ɨ, ʉ, k, t, n/. This mildly leads to the conjecture that it reflects some extra-linguistic noise that was not present in the training TIMIT recordings. Indeed, we noticed that the Common Voice audio data often included long pauses preceding and following the utterance. While for future investigations, it might be relevant to include noise reduction techniques, we chose to tackle this issue at the decoding / evaluation step, by lowering the deletion cost to `0.5` and insertion cost to `0.75`.

Results Default PER achieved on Yoruba was `0.34`. It ranged from `0.18` to `1.11`, with median `0.30`. The best and worst predictions, as well as the alignment of the median prediction are in Appendix D. With the original Levenshtein costs for deletion and insertion, median PER increases to `0.35`.

4.3 Performance by phoneme

Alignment counts (weighted Levenshtein) are depicted below, in the form of a confusion matrix: Figure 3. Color encodes the posterior probabilities of the prediction labels given the gold labels, so for example, a gold Yoruba label /ɛ/ is most likely to be predicted as /ɪ/, followed by /i/ in English IPA inventory.

Confusion entropy We used conditional entropy to find which Yoruba phonemes are successfully transferred (Table 1), assuming that a well-transferred sound will have a less uniform distribution over predicted sounds: most alignments would belong to a couple of phonetically close classes. Most of the formulae we used were drawn from Bentz and Gutiérrez-Vásquez (2022) (Appendix D).

For all of the phonemes the entropy was relatively high. It is evident from the results that phonemically similar symbols in terms of feature distance were the most common alignments for more accurately predicted phonemes—except for, for example, voiced fricatives, which Yoruba does not contrast. It seems, front vowels and consonants articulated in the front were broadly predicted correctly, whereas back vowels and consonants were more frequently confused. We found no possible phonetic explanation to this other than that it could be an artifact of the feature system.

Mapping entropy A perhaps more interesting question in context of using the model for transfer is whether we can restore the Yoruba phonemes reliably from our TIMIT-based model predictions. To see how unambiguous this mapping is, and how much signal our model contains after transfer, we use conditional entropy of the golden label probabilities given the predicted label (Appendix C). The overall entropy was `3.5` bits, i.e. about `11` Yoruba labels exist as viable options per prediction. However, for some of the labels it was much smaller, thus signifying almost one-to-one mapping from predicted labels to Yoruba (Table 2). The number would have likely been even smaller if not for deletions, draw a large share from the probability mass for most sounds. Another factor raising the expected uncertainty is that the sounds that can map to multiple Yoruba labels, e.g. /m, l, w/, are among the most frequent ones. (Figure 15)

All of the back vowels, whose frequent confusions we mentioned above, are now among the most certainly transferrable. This means that the high confusion entropy observed actually stems from the sheer amount of possible vowel segments in English. When English partitions the vowel space to /a, ɑ, æ, ʌ/, they all are included in Yoruba /a/. Ambiguous mappings (indicated by high entropy) also occur for sounds absent in Yoruba: IPA /ʔ, ɹ, ɚ, ð, θ/ have no clear correspondence in Yoruba to gravitate to. This mostly

Table 1: Confusion entropy, empirical and theoretical (feature-based) correspondences.

Gold	Confusions	Similar	H
s	[z, s, n, v]	[s, z, t, ʃ]	3.40
m	[m, n, l, w]	[m, b, n, p]	3.67
d͡ʒ	[d͡ʒ, d, n, z]	[d͡ʒ, t͡ʃ, ʒ, ʃ]	3.69
b	[b, v, m, n]	[b, p, m, v]	3.76
ʃ	[ʒ, ʃ, z, h]	[ʃ, ʒ, s, t͡ʃ]	3.80
i	[i, ɨ, ɪ, j]	[i, e, ɨ, ɪ]	3.81
w	[w, j, ʔ, u]	[w, ʍ, u, ɚ]	3.82
e	[i, ɪ, ɨ, j]	[e, i, æ, ɛ]	3.87
j	[j, i, w, ʔ]	[j, i, ɪ, w]	3.92
ĩ	[i, ɪ, ɨ, u]	[i, e, ɨ, ɪ]	3.93
ɛ̃	[ɪ, ɛ, n, ʔ]	[ɛ, e, ə, ɪ]	3.99
d	[d, n, g, b]	[d, t, n, z]	4.00
n	[n, m, l, d]	[n, d, l, m]	4.03
r	[ɹ, n, j, l]	[r, l, d, n]	4.07
ɛ	[ɪ, i, ɛ, ɨ]	[ɛ, e, ə, ɪ]	4.09
k	[g, ʔ, k, h]	[k, g, h, ŋ]	4.10
g͡b	[b, m, w, v]	[g, b, k, p]	4.13
f	[f, v, z, h]	[f, v, p, s]	4.14
u	[u, ɨ, i, ʉ]	[u, ʉ, o, ɨ]	4.16
t	[z, d, n, t]	[t, d, s, θ]	4.16
k͡p	[b, w, ʔ, v]	[k, p, g, b]	4.18
g	[g, b, w, v]	[g, k, ŋ, b]	4.18
ũ	[u, ɨ, w, ɪ]	[u, ʉ, o, ɨ]	4.21
l	[l, n, ɹ, j]	[l, d, n, z]	4.26
h	[ʔ, h, w, l]	[h, k, ʔ, j]	4.33
o	[u, ɨ, o, i]	[o, ɔ, ʌ, a]	4.41
a	[a, ʌ, æ, ɑ]	[a, æ, ɑ, ʌ]	4.52
ɔ	[o, ə, ʌ, ɨ]	[ɔ, o, ə, ʊ]	4.53
ã	[ɨ, o, u, ə]	[a, æ, ɑ, ʌ]	4.55

Table 2: Mapping entropy and frequentmost correspondences.

Gold	Corr	H	Gold	Corr	H
a	[a, ã]	1.96	ɨ	[i, a]	3.41
ɑ	[a, ã]	2.01	ʉ	[u, ũ]	3.41
æ	[a, e]	2.32	ɪ	[ɪ, i]	3.48
ʃ	[ʃ, s]	2.33	ɚ	[ʊ, ɔ]	3.52
ʌ	[a, e]	2.49	n	[n, d]	3.59
t͡ʃ	[d͡ʒ, ʃ]	2.63	z	[s, d]	3.63
d͡ʒ	[d͡ʒ, ʃ]	2.80	ɹ	[j, l]	3.66
i	[i, e]	2.82	t	[t, d]	3.67
o	[o, ɔ]	3.05	θ	[s, t]	3.69
ʊ	[ʊ, ɔ]	3.06	ʔ	[h, k]	3.73
ɔ	[ɔ, o]	3.06	w	[w, u]	3.75
ɛ	[ɛ, e]	3.07	m	[m, b]	3.78
k	[k, g]	3.09	ŋ	[ŋ, g]	3.80
ʒ	[ʃ, d͡ʒ]	3.10	d	[d, t]	3.82
ə	[ɔ, ɛ]	3.28	l	[l, d]	3.91
ɝ	[ʊ, ɔ]	3.28	b	[b, m]	3.91
s	[s, t]	3.28	j	[j, i]	3.95
e	[e, i]	3.31	g	[g, k]	4.05
u	[u, ũ]	3.32	ð	[d, l]	4.16
f	[f, s]	3.33	h	[h, k]	4.25
p	[b, f]	3.36	v	[f, b]	4.26

applies to consonants, which coincides with a commonly stated (Hayes, 2009; Ladefoged, 2011) intuition that vowel realizations distribute continuously in the acoustic space, while consonant allophones belong to one of some discrete targets.

4.4 Phone embeddings space

After having fully trained the model, we can take a look into the representations it has formed for each item of the vocabulary by collecting phone embeddings from the last layer. We attempt to plot these phone vectors in an easily interpretable manner, in order to investigate what the internal model representations for each language might look like. Ideally, the partition in the space would reflect similarities and differences for each target language, possibly even extending to reflect a scale of articulatory features. We run inference to produce phone embeddings from the last layer output (log probabilities in the phoneme classes space) and plot these using dimensionality reduction.

Phone posteriogram embeddings Note that to obtain the representations below, we only selected items that were correctly predicted, after the CTC collapse. It is entirely possible that some matching predictions are skipped in the process, but since we have not yet established a way to decode the noise that was generated, we limit ourselves to accurately aligned pairs. Unfortunately this excludes the phonemes unique to each language, like /ũ/ and /k͡p/.

To represent each phoneme token, we take only the first time stamp in which it is predicted. It runs against the intuition we have from spectrograms that the early time stamps of a phone are not the most characteristic, but we consider it a suitable representation. By extending the extraction to all the time frames in one example (including frames where the highest probability lies on the blank token), we can obtain a spectrogram-like

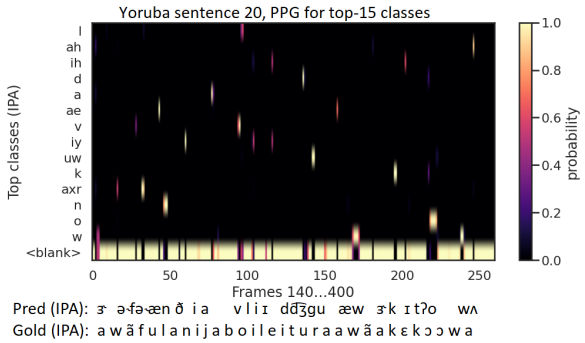

Figure 4: Phone posteriograms for one Yoruba sentence, after softmax, first 14 classes

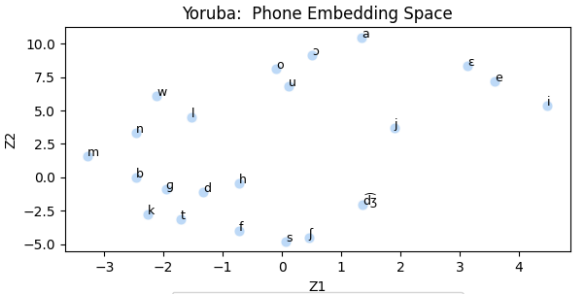

Figure 5: Yoruba Phone Embeddings (with English correspondences only)

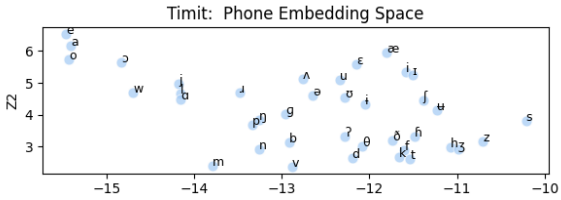

Figure 6: Timit Phone Embeddings

visualization of the shifting of the conditional probabilities over time (Figure 4). As a product of CTC decoding, these posteriograms appear to be different from spectrograms. Since the model is allowed to predict blank symbols, it tends to allocate blank symbols for majority of frames, while producing occasional spikes for non-blank phonemes. These spikes often appear at the time steps where phoneme predictions first become feasible: these are used as embeddings.

We employ *UMAP* to obtain 2-dimensional embeddings from the 42-dimensional TIMIT and 23-dimentional Yoruba predictions, for which we obtain the mean by phoneme. We settled on UMAP since it maintains the non-linear assumption of t-SNE regarding the data while being less dependent on initialization. (McInnes et al., 2018) While both plots seem to group similar sounds together to an extent, comparing inference simultaneously on both datasets remains tricky due to 1) the different probability distributions and the difference in dataset sizes 2) while UMAP remains an efficient dimensionality reduction method, it is unclear whether plotting always reflects similarity through clustering. (McInnes, 2023) We can see that in some cases, similar sounds do tend to cluster, but the extent to which these patterns can be identified different in both datasets.

Yoruba phones In the reduced Yoruba space (Figure 5), cardinal vowels cluster in the same space in the graph, with their distance mostly corresponding to their feature weighted equivalent. We see how natural classes of phonemes group together in space: the sibilants /s/ and /ʃ/, plosives, sonorants, palatals. Incidentally, in that arrangement, the *y*-axis clearly corresponds to sonority hierarchy. (Hayes, 2009) It can also be argued that the *x*-axis loosely represents a gradual backness scale, when including acoustically related parameters like lip protrusion, in the case of labials.

English phones As for the TIMIT space (Figure 6), these patterns are less clear. It is not easy to assign one linguistic feature to the axes, like sonority or backness. The natural groupings of sounds are still observable, but exact composition and number of clusters differs from Yoruba: nasals group separately from approximants and /l/; /p/ is separated from other voiceless stops. As for vowels, while /ə, ʌ/ and /i, ɪ/ (respectively positional allophones and a tense-lax pair vowel) are grouped together, it can be pointed out that /a, o, e/ appear as a detached cluster. Oddly, all of these labels belong only to nuclei of diphthongs, split after data processing, which has to be linked as to why they occupy their own space in the plot. It is thus probable that the differing annotation schemes for both datasets (with more fine-grained distinctions for TIMIT) partly contribute to this.

Yoruba and English The fact that the clusters differ is somewhat trivial, considering that the UMAP reduces space differently for different subsets of labels, and Yoruba clustering did not have to account for relations with all TIMIT sounds. Hence or whence we also look into the simultaneous projection of Yoruba and English embeddings in one space, by first fitting the transform based on the English data, and then reducing the Yoruba

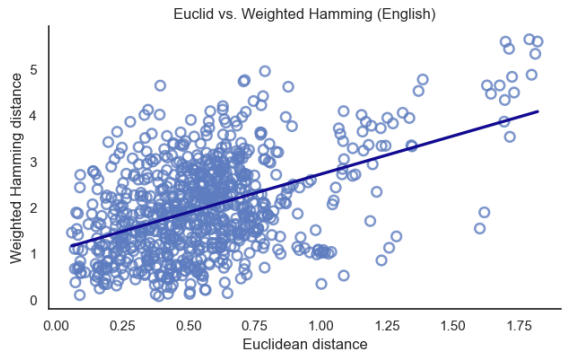

Figure 7: Correlation between Euclidian distance of embeddings and Weighted Hamming distance (English). (Scatter: phone tokens in the subsample, jittered)

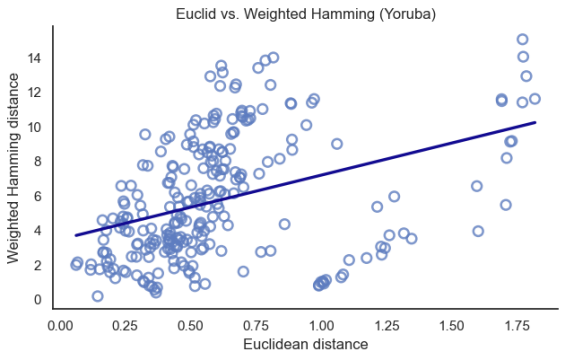

Figure 8: Correlation between Euclidian distance of embeddings and Weighted Hamming distance (Yoruba). (Scatter: phone tokens in the subsample, jittered)

data, collecting Euclidian distances between the common phonemes. After running the projection several times, we find that the distances we obtain are consistent across trials, but it is difficult to extract a common pattern (Figure 17, Table 8).

Distances To investigate this further, we decided to run a correlation test between the Euclidian distance between phones after dimensionality reduction and feature weight. Figures 7, 8 show that there is a significant correlation between the two variables, with the correlation coefficient being lower for Yoruba than it is for English (Table 7). This reflects the fact that the English label inventory represents more varied sounds than the Yoruba predicted values, helping establish the correlation pattern during fitting. While this possibly shows that our presentations are reliable and in line with the framework we establish, this remains obscured by the fact that the plot for English is not easily interpretable.

5 Discussion

5.1 Closer look on confusion and correspondence patterns

One of the pros of the conditional entropy metrics from Results 4.3 for assessing transfer is that they does not require the correspondences to be phonologically accurate—the only condition is that they are unambiguous. The drawback of this is that low entropy could arise from systematic bias toward over-represented classes rather than accurate transfer. Let us consider in detail what phonologically non-trivial associations are found in alignments and how they reflect the expected differences between English and Yoruba discussed in Introduction 1.

PMI To measure association we employ a symmetric measure related to the conditional entropies. Pointwise Mutual Information (PMI) computes how much it is not a coincidence for two values to appear together. (Jurafsky and Martin, 2023, p. 14), (Jäger, 2013, p. 263–265) It is calculated by dividing the joint probability by the product of two marginal probabilities (Appendix C). In a cross-lingual transfer-learning setting, since two distributions do not have identical event space, we used the co-ocurrence matrix (Section 4.3) to find the joint and marginal probabilities. The probability of phonemes varies greatly, therefore we normalize PMI by dividing it with negative log-likelihood of the joint probability (the theoretical maximum (Bentz, 2018, p. 61)). NPMI shows what correspondences are statistically meaningful even when the raw counts are extremely low and can be taken as a measure of correlation. Values between `-1` and `0` imply that the co-occurrence of the phonemes from two variables in alignments is more scarce than random, and values between `0` and `1` suggest that two values co-occur more often than random, so a successful transfer shows each class in predicted map to a certain Yoruba label, and in a trivial transfer the association would be with the closest sound.

Interpretation The first look reveals a strong cyan diagonal in a lilac field: these are the correspondences between predicted phones and their trivial Yoruba counterparts. No suspicious associations dwell in the off-diagonal areas, even among the least frequent phones: phonological closeness rather than noise accounts for the information-

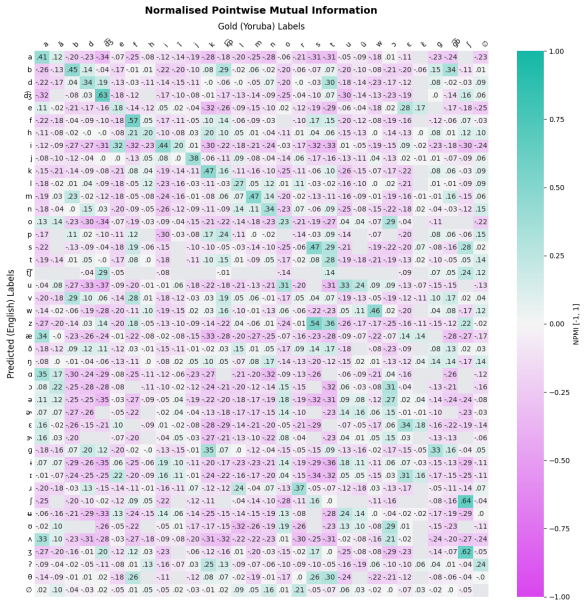

Figure 9: Normalized PMI. Turquoise denotes the encoding efficiency gain from aligning, magenta denotes the gain from not aligning. Grey stands for the alignments that have never occurred (we can assume negative values).

carrying phone pairs.

The rows and columns labelled Ø stand for alignments with deletion and insertion, respectively. All vowels strongly disassociate with deletion. This is likely be due to the CV syllable structure in Yoruba (ref. hypothesis (iv) in Section 1.3). The model generates a lot of consonant clusters, in accordance with English phonotactics, which then undergo deletion in order to align the vowels. The frequency of insertions, on the opposite, seems to be unrelated to the phoneme type. One strange exception is Yoruba /ɾ∼r∼ɹ/. The only *ad-hoc* explanations that come to mind are that our model misses the Yoruba tap realisations, and they are short, or that it predicts an English rhoticized vowel instead of it. English phonemes similar to Yoruba's double-articulated stops, /p, b, k, g/, are occasionally deleted, but no more than other consonants. This likely disproves the hypothesis (iii) (Section 1.3) derived from our cognition, that double-articulated stops will be recognized as sequences. A frequent correspondence with labio-velar stops /g͡b, k͡p/ unexpectedly, but logically turned out to be the labio-velar approximant /w/. Other labial consonants were also associated.

Generally all consonants without close place/manner counterparts show higher entropy. Thus, /θ/ is divided equally between /f, s, t/. The rather sonorant /ð, ŋ/ are pretty much random. The glottal stop is split between /k, h/ and being deleted; /p/ is between /k͡p, k, f, b/. These are all examples of migrating to some sort of place/manner compromise. Contrast this with vowels: all vowels, regardless if they exist in Yoruba, have high entropy (see also Section 4.2).

A clear tendency is that the VOT is transferred not one-to-one, but regularly: Yoruba lacks voiced fricatives, but when /z, ʒ, v/ appear in predictions, they consistently associate with their voiceless pairs, and sometimes with place pairs. Predicted voiced stops map equally to voiceless and voiced Yoruba stops, while voiceless always map to voiceless.

Liquids and nasals form a "connected cluster" of confusions; it is possibly but not necessarily related to their partial allophony in Yoruba, where nasality often spreads (see hypothesis (v) 1.3). In general, more sonorant phones in TIMIT have higher entropy here, and vowels are even more interrelated, slightly grouping by backness.

Yoruba nasal ã is associated with o-like vowels, since ã and ɔ̃ are allophones in Yoruba (see Section 2 Datasets). This probably indicates that in the varieties prominent in Common Voice the mid allophone occurs more often (something not stated in Allen et al. (2013); Adesola (2006); Akinbo et al. (2022); Przedziecki (2005) —the body of works we examined). Other nasal vowels associate with their oral bases (/ĩ/ with i-like, etc.), but with weaker certainty. They are aligned with approximants more often: /ũ/ with /w/, and /ĩ/ with /j/.

5.2 How to benchmark our results

A challenge in evaluating our model and drawing conclusions about phonology lies in circular nature of alignments. We first use theoretical assumptions about phoneme closeness for our alignment algorithm. Then we treat these alignments like the actual correspondence (e.g., as if the model assigned mentioned the predicted label exactly during the time span of the golden Yoruba label) to speculate which sounds the model was not able to recognize correctly. These alignments are not in fact related to the traceable causal/computational link between in- and output during inference. In a way, the alignment being theoretically sound (cf. Section 5.1) is not surprising, because it is delivered via *theory-based* edit distance.

Refer to Appendix D for an example of alignment. Impressionistically, although the model output is clearly linked to the input, the alignments are not always intuitively correct. Indeed, they are proven minimal given the feature model, but it does not guarantee to correspond to a linguists intuition. This is a common issue in the pure feature-based *Lev* alignments (Dellert, p.c.). More traceable options include alignment with some kind of supervised training on a human-aligned data. Nerbonne and Heeringa (2010) reports good alignment results with log-transformed feature weights. Another option is to embed some pre-knowledge of phonotactics in the cost functions by assigning weights to different features: say, to improve resolution of consonant clusters, knowing the Yoruba **CV** syllable structure, we can make deleting consonants cheap, while changing consonants to vowels costly. Jäger (2013) argues for an empirical approach free from feature-theory assumptions.

Combining a feature model with minimal distance leads to the overall fwPER being hard to interpret. For example, we do not know the expected cost for two random phonemes, so, unlike in the regular PER, there is no intuition such as "every third letter is wrong". It is also not clear how much minimizing the alignment lowers the fwPER on average.

In order to establish a theory-light frame of reference within which we can benchmark our model's fwPER performance, we compare it with the fwPER of random strings. We evaluated 3 different randomly generated "predictions" with the same cost settings as in Results 4.2.

First, we took the exact lengths of the sentences from our predictions, to simulate some understanding of time, then we tested: (1) generating random sequences of TIMIT labels of that length, sampling from a uniform distribution ("nulligrams") (2) generating that using the unigram probabilities of the TIMIT labels in our predictions. This way we had the same predictions, but filled with random letters, unrelated to the audio input whatsoever. (3) making a classic bigram model based on actual predictions, generating strings of random length until the EOS symbol.

The fwPER of the equal probabilities and unigrams is surprisingly close to ours: `0.4` and `0.39`. The difference of the distribution and median from the model's performance (Figure 10) is noteworthy: in predictions the positively-skewed bell curve has

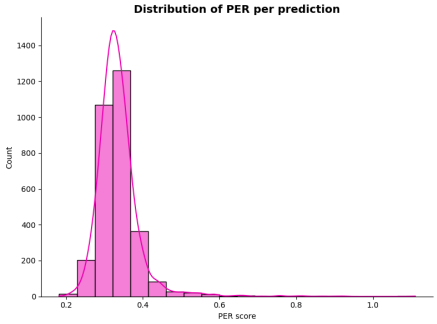

Figure 10: PER Distribution

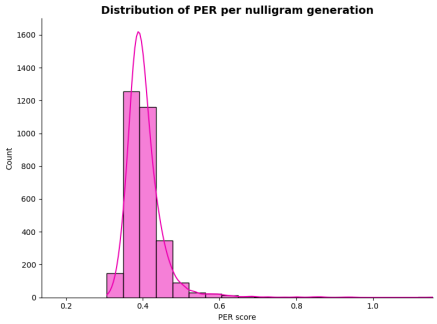

Figure 11: Equal probability PER Distribution

longer tails on both sides; the random alignments gravitate more towards the average. (Figures 11, 12) The intuition allows us to say that the model predictions are not random: it seems, the random simulations never achieve low PER values like `0.2` to `0.3`. A qualitative look at the alignments in Appendix D suggests that the prediction outcomes are better and indeed linked to the input sound.

The length of the produced output is critical for good result. You can see how PER for the birgam generation (Figure 13), made without fixed sequence length, blew up drastically. This suggests that fixing the model's tendency to over-generate in noisy circumstances is the most relevant way to reduce the error.

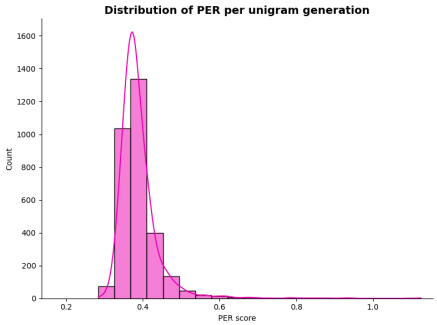

Figure 12: Unigram PER Distribution

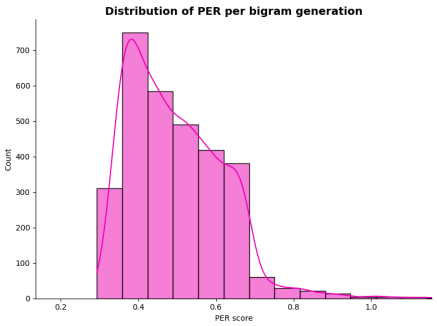

Figure 13: Bigram PER Distribution

6 Limitations

For future reference, we believe that our model could be further improved by dedicating more attention to the over-generation problem, and possibly address it at the pre-training stage with a modified pre-training objective (or simply a denoising procedure at the data preprocessing stage), or by investigating whether there are some repeating patterns linked to over-generation. This would then in turn help improve the quality of alignment for the evaluation step. Moreover, having datasets that agree on the granularity on the description would be helpful, as this would facilitate the mapping and reduce the risk of inducing biases. Although the pre-training stage should in principle be as general as possible, introducing phonotactic constraints through a LM model would likely also contribute to reduce noise during generation. In regards with the Common Voice dataset specifically, we can also mention the size of the corpus in comparison to the English one, since the 0.25 split we used for English for evaluation is almost twice the size of all the Yoruba splits combined. Moreover, although we deem Epitran to be suitable for the G2P step (its paper mentions it being competitive with the FST-based baselines for ASR), it would be interesting to see whether a more-linguistically informed model, perhaps a WFST would yield a different performance. Another direction we could was to use the embeddings obtained to see whether they latently contain some information about Yoruba which was left out from the annotations we used. For instance, it would be interesting to see whether tone can be accurately predicted by feeding the final representations into a classifier, investigating whether the architecture was sufficient for the model to internalize this additional information.

We of course encourage the reproduction of this study on different datasets, be it for Yoruba or other languages.

7 Conclusion

We trained an audio to IPA ResNetBiLSTM CTC recognizer on TIMIT only (`0.0339` on English validation set), taking inspiration from Dhakal et al. (2022), to probe zero-shot transfer to Yoruba under a linguistically-motivated feature-weighted mapping scheme. Tackling the problem of different inventories, we applied a modified version of feature-weighted Levenshtein algorithm for evaluation. The system reaches ≈`0.34` PER on Yoruba (median `0.30`). We reduced the original model to a third of its size (from 1.55 to 0.8 M. parameters), allowing for faster training and inference, which still preserved the signal needed for cross-lingual transfer.

We introduce two evaluation metrics based on conditional entropy: confusion and mapping entropy. Despite relatively high confusion entropy, indicating inexact predictions, the mapping entropy seems low, indicating that Yoruba phonemes are unambiguously inferred from the predictions. This can reflect that English's finely split vowel space and overly narrow transcription tradition collapses toward Yorubas coarser categories. Our analysis of (N)PMI confirms the assumption of phonology-driven transfer: correspondences track place/manner similarity, voicelessness transfers more reliably than voicing (due to VOT differences), labio-velar stops are often perceived as /w/. At the same time, the model over-generates, partly because of pauses outside of utterance and extra-linguistic noise in the evaluation set, partly because the transfer keeps the tendency for consonant clusters, systematically breaking the syllable structure of Yoruba.

Finally, we noted a methodological circularity: alignments are theory-driven (based on feature phonology) and not causal chain-driven, which biases the further conclusions about phonology into confirming the pre-existing notions. This nature of the alignments remains a problem, as long as they are are not obtained through the time-aligned test set or checked for intuitiveness by supervised training or curation. The fwPER evaluation lacks the easy interpretation of PER. Nonetheless, random baselines and analysis of PPGs suggest that our model's predictions are genuinely signal-bearing.

References

Oluseye Adesola. 2006. *Yoruba: A Grammar Sketch, Version 1.0*. Online PDF edition.

Orevaoghene Ahia, Anuoluwapo Aremu, Diana Abagyan, Hila Gonen, David Ifeoluwa Adelani, Daud Abolade, Noah A. Smith, and Yulia Tsvetkov. 2024. Voices unheard: Nlp resources and models for yorùbá regional dialects. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 24)*, pages 4392–4409. Association for Computational Linguistics.

Samuel K. Akinbo, Olanrewaju Samuel, Iyabode B. Alaga, and Olawale Akingbade. 2022. An acoustic study of vocal expression in two genres of yoruba oral poetry. *Frontiers in Communication*, 7:1029400.

Blake Allen, Douglas Pulleyblank, and ládiípò Ajíbóyè. 2013. Articulatory mapping of yoruba vowels: an ultrasound study. *Phonology*, 30(2):183–210.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Reuben Henretty, Michael Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4218–4222, Marseille, France. European Language Resources Association (ELRA).

Christian Bentz. 2018. *Adaptive Languages: An Information-Theoretic Account of Linguistic Diversity*. Empirically Oriented Theoretical Morphology and Syntax. Language Science Press, Berlin.

Christian Bentz and Ximena Gutiérrez-Vásquez. 2022. Information-theoretic analyses of natural languages. In *DGfS Workshop: Information-Theoretic Analyses of Natural Languages*, Tübingen, Germany. Presented at the Annual Conference of the German Linguistic Society (DGfS), February 22, 2022.

Carnegie Mellon University. 1998. Cmu pronouncing dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Harper & Row, New York.

Manish Dhakal, Arman Chhetri, Aman Kumar Gupta, Prabin Lamichhane, Suraj Pandey, and Subarna Shakya. 2022. Automatic speech recognition for the nepali language using cnn, bidirectional lstm and resnet. In *2022 International Conference on Inventive Computation Technologies (ICICT)*, pages 515–521. IEEE.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, and David S. Pallett. 1993. Darpa timit acoustic-phonetic continuous speech corpus. LDC Catalog No. LDC93S1, ISBN 1-58563-014-5.

Awni Hannun. 2017. Sequence modeling with ctc. *Distill*. Https://distill.pub/2017/ctc.

Bruce Hayes. 2009. *Introductory Phonology*. Wiley-Blackwell.

Gerhard Jäger. 2013. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change*, 3(2):245–291.

Daniel Jurafsky and James H. Martin. 2023. *Speech and Language Processing*, 3rd edition. Prentice Hall.

Peter Ladefoged. 2011. *A Course in Phonetics*, 6th edition. Wadsworth, Cengage Learning.

Peter Ladefoged and Ian Maddieson. 1996. *The Sounds of the Worlds Languages*. Blackwell Publishers, Oxford & Cambridge, MA.

Ian Maddieson. 2010. Phonotactics. In Marc van Oostendorp, Colin Ewen, Elizabeth Hume, and Keren Rice, editors, *The Blackwell Companion to Phonology*, pages 80–105. Wiley-Blackwell.

Leland McInnes. 2023. Using umap for clustering. `https://umap-learn.readthedocs.io/en/latest/clustering.html`.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas GroSSberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Steven Moran, Daniel McCloy, and Richard Wright. 2014. Phoible online. *Leipzig: Max Planck Institute for Evolutionary Anthropology*.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2016. Panphon: A resource for mapping ipa segments to articulatory feature vectors. In *Proceedings of COLING 2016*, pages 3475–3484.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2390–2397. European Language Resources Association (ELRA).

John Nerbonne and Wilbert Heeringa. 2010. *Measuring dialect differences*, pages 550–567.

Marek A. Przezdziecki. 2005. *Vowel Harmony and Coarticulation in Three Dialects of Yorùbá: Phonetics Determining Phonology*. Ph.d. dissertation, Cornell University.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

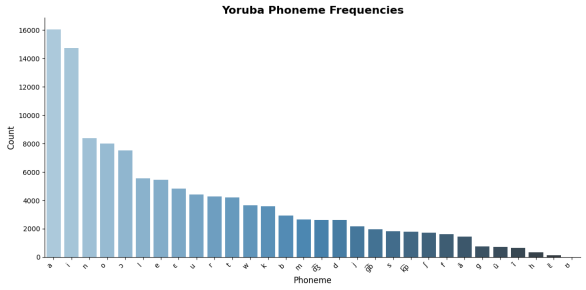

Figure 14: Original Yoruba Dataset Phoneme Frequency

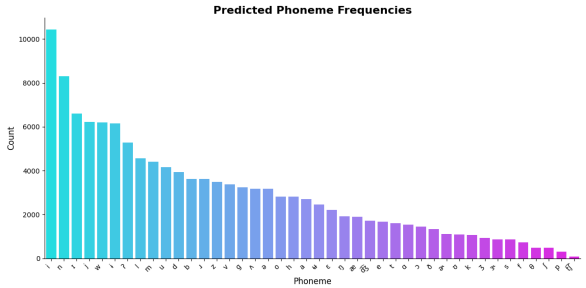

Figure 15: After Evaluation Phoneme Frequency

Mirco Ravanelli and Yoshua Bengio. 2019. Speaker recognition from raw waveform with sincnet. *Preprint*, arXiv:1808.00158.

Arne Rubehn. 2024. Generating phonological feature vectors with soundvectors and clts. *Computer-Assisted Language Comparison in Practice*, 7(2):59–67.

Harry van der Hulst. 2017. Phonological typology. In Alexandra Y. Aikhenvald and R. M. W. Dixon, editors, *The Cambridge Handbook of Linguistic Typology*, pages 39–77. Cambridge University Press.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2021. Simple and effective zero-shot cross-lingual phoneme recognition. *Preprint*, arXiv:2109.11680.

Hemant Yadav and Sunayana Sitaram. 2022. A survey of multilingual models for automatic speech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5071–5079, Marseille, France. European Language Resources Association.

A English inventory

B Yoruba inventory

C Formulae

Formula for the substitution cost of two IPA vectors from PanPhon inventory: English $\vec{e}$ and Yoruba $\vec{y}$.

$$\text{dist}(\vec{e},\vec{y}) = \frac{\left|\{\, i \mid (e_i \neq y_i) \wedge (e_i \neq 0 \vee y_i \neq 0)\,\}\right|}{\left|\{\, i \mid e_i \neq 0 \vee y_i \neq 0\,\}\right|}$$

Table 3: TIMIT to IPA mapping (folded into two columns)

TIMIT	IPA	Our	TIM	IPA	Our
aa	ɑ	ɑ	ch	t͡ʃ	t͡ʃ
ae	æ	æ	d	d	d
ah	ʌ	ʌ	dh	ð	ð
ao	ɔ	ɔ	dx	ɾ	r
aw	aw	a + w	el	l̩	l
ay	aj	a + j	em	m̩	m
ax	ə	ə	en	n̩	n
axr	ɚ	ə	f	f	f
eh	ɛ	ɛ	g	g	g
er	ɝ	ɝ	hh	h	h
ey	ej	e + j	h	h	h
ih	ɪ	ɪ	jh	d͡ʒ	d͡ʒ
ix	ɨ	ɨ	k	k	k
iy	i	i	l	l	l
ow	ow	o + w	m	m	m
oy	ɔj	ɔ + j	n	n	n
uh	ʊ	ʊ	nx	ɾ̃	n
uw	u	u	ng	ŋ	ŋ
ux	ʉ	ʉ	p	p	p
ax-h	ə̥	ə	q	ʔ	ʔ
bcl	b˺	b	r	ɹ	r
dcl	d˺	d	s	s	s
eng	ŋ̍	ŋ	sh	ʃ	ʃ
gcl	g˺	g	t	t	t
hv	ɦ	h	th	θ	θ
kcl	k˺	k	v	v	v
pcl	p˺	p	w	w	w
tcl	t˺	t	wh	ʍ	ʍ
pau	|	–	y	j	j
epi	||	–	z	z	z
h#	/	–	zh	ʒ	ʒ
b	b	b			

(Mortensen et al., 2016)

Formulae for confusions. Formally, let Y denote the gold Yoruba phoneme and $\hat{Y}$ the predicted TIMIT phoneme var. Each row of the confusion matrix is a probability distribution over predicted labels, conditioned on a given Yoruba label.

$$P(\hat{Y} = j \mid Y = i) = \frac{|\{(x : Y(x) = i, \hat{Y}(x) = j)\}|}{|\{(x : Y(x) = i)\}|},$$

Then the confusion entropy is $H(\hat{Y} \mid Y = i) = -\sum_j P(\hat{Y} = j \mid Y = i) \log_2 P(\hat{Y} = j \mid Y = i)$.

The metric that we call mapping entropy is the opposite. With the posterior probabilities of gold

Table 4: Yoruba IPA inventory

Yoruba	IPA	IPA (adjusted)
m	m	m
i	i	i
k	k	k
y	j	j
u	u	u
a	a	a
w	w	w
n	n	n
t	t	t
l	l	l
s	s	s
b	b	b
e	e	e
o	o	o
g	g	g
h	h	h
d	d	d
r	ɾ	r
f	f	f
ẹ	ɛ	ɛ
ṣ	ʃ	ʃ
ọ	ɔ	ɔ
j	d͡ʒ	d͡ʒ
´	˥	–
`	˩	–
in	ĩ	ĩ
un	ũ	ũ
gb	g͡b	g͡b
p	k͡p	k͡p
ọn	ɔ̃	ã
ẹn	ɛ̃	ɛ̃
an	ã	ã
–	˧	–
n	ŋ	–
ụ	ʊ	–
ị	ɪ	–

labels given a certain prediction

$$P(Y = i \mid \hat{Y} = j) = \frac{|\{(x : Y(x) = i, \hat{Y}(x) = j)\}|}{|\{(x : \hat{Y}(x) = j)\}|}$$

the mapping entropy is $H(Y \mid \hat{Y} = j) = -\sum_i P(Y = i \mid \hat{Y} = j) \log_2 P(Y = i \mid \hat{Y} = j)$. With the expectation $H(Y \mid \hat{Y}) = \sum_j P(\hat{Y} = j) H(Y \mid \hat{Y} = j)$, and perplexity $P(Y \mid \hat{Y}) = 2^{H(Y|\hat{Y})}$. (Bentz and Gutiérrez-Vásquez, 2022)

Table 5: Some of the worst predictions (Pred vs. Gold).

	Pred	Gold
S1	ʔ ɪ t ð ə ʔ ɪ ə b i p i ɪ n d u ə b i ɹ ɛ n ə d i ɹ ɪ k i ɨ t ɪ b ʉ p l ɪ n i n t l ʊ d ʉ i ɹ ɪ t i ŋ b ʉ d b l i b ʔ e j ɹ t ʉ z d i d d n g i n p i p	g͡b o g͡b o a w ã d ɔ k i t a t i k͡p ĩ u l a t i w a ɔ n a a b a j ɔ s i a a r u n
S2	ɔ w a w v ə e ʔ j z ʉ w ɪ m ʉ ɹ u u ʔ j i ɹ i ɪ z z ɪ ɪ ə z ʉ ɪ i g u d l ʌ ʌ ʔ ɪ n ɹ i ɪ ɪ z ʉ ɹ h j o	i k u n i e r e ɛ ʃ ɛ f u n ɛ̃ i t o b a k͡p a e e j a n
S3	n ɛ ə ʔ ə t æ ʔ ɛ ʔ ə ɹ n p ʊ ɪ t ɪ m p ɤ m i ɹ i z b ɹ a j t o i n t a t o w t ə ɹ i p i n ɛ ɹ o w u t ɛ t ɹ i ɹ n ə ʔ æ t k w h i p i n b ə p ɹ ʉ z k ʉ n z i w i d i ŋ i p n æ t ɪ n k i t ə ə n	o l u ɔ m ɔ l o f a r a g͡b a ɔ b ɛ f u n a a r ɛ n i b i i k͡p o l o n g o i b o

Table 6: Two best predictions

Distance	6.9
Pred	u i l e j d i ɑ d u l a w ə m æ d ɑ l æ l i ɹ a j i z s ɔ i ɹ u b ɑ m
Gold	o r i l ɛ e d e a d u l a w ɔ m ɛ t a l a l o m a a n s ɔ e d e j o r u b a
Distance	6.0
Pred	o l m i ɑ j ɹ i b i ʌ l w k ʌ b ʌ d͡ʒ u l ɪ b ʊ s i i b ɔ w u
Gold	o m i j a l e n i ɔ l ɔ k͡p a a r ɔ d͡ʒ ɔ i b ã s i i g͡b o h o

Formula for normalized PMI for a gold sound y and predicted sound e in the co-occurrence counts matrix obtained from alignments.

$$\text{nPMI}(y, e) = log_2(y, e) - log_2(y) - log_2(e)$$

(Jurafsky and Martin, 2023)

D Yoruba Prediction Examples

E More on PPGs

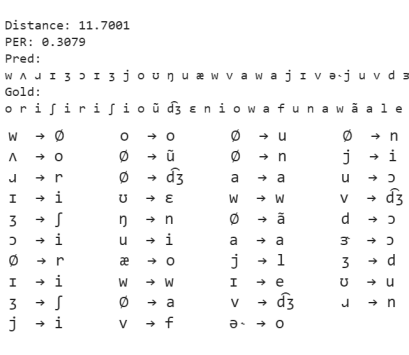

Figure 16: Automatic alignment of the median prediction

Table 7: Pearson correlation between vowel distance and another variable

Language	**Correlation Coefficient** (r)	p**-value**
English	0.47	0.0002
Yoruba	0.49	0.004

Table 8: Vowel distances between English and Yoruba (sorted by similarity)

Vowel	Distance
u	0.095
g	0.170
k	1.119
d	1.280
m	1.424
t	1.571
j	1.633
s	1.800
ɔ	1.947
ɛ	1.971
b	1.993
h	2.549
n	2.887
w	2.907
e	3.059
ʃ	3.278
o	3.291
l	3.363
i	3.590
a	3.643
f	5.945

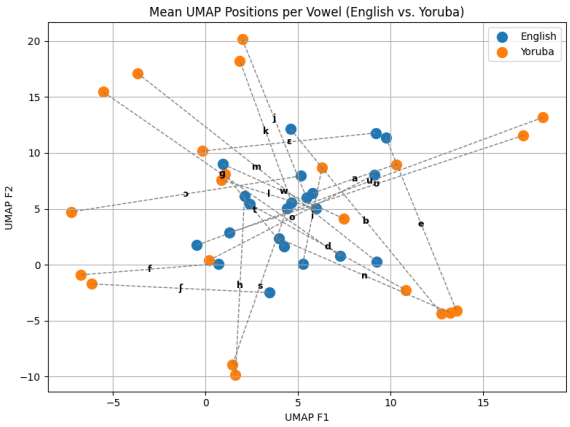

Figure 17: Yoruba and English consonant embeddings plotted in the same space