Phonetic analysis of HuBERT Self-supervised representation for Yoruba

Opeyemi Osakuade Simon King

CSTR, University of Edinburgh

Abstract

Self-supervised speech representation learning aims to discover representations of unlabeled speech. HuBERT, an English-based self-supervised speech representation learning approach is said to out-perform other state-of-the-art systems on downstream tasks such as ASR and speech synthesis. Its approach utilizes an offline clustering step to provide aligned target labels for a BERT-like prediction loss. It applies the prediction loss over the masked regions only, which forces the model to learn a combined acoustic and language model over the continuous inputs.

These learnt sequences of discrete units have been evaluated to capture sub-phonetic events such as the distinct closure and release portions of plosives in English language. While certain speech sounds are similar to English, the Yoruba language places significant importance on its tonal system, distinguished by high and low tones. Operating as a tonal language, Yoruba utilizes distinct pitch patterns to distinguish between individual words and grammatical variations of those words.

The question here is to what extent does HuBERT sequence of discrete units represent the phonemes and tones in the Yoruba language? We present an analysis of the discrete units discovered extracted from Yoruba language speech using a pre-trained HuBERT model, to see how well HuBERT represen-tations might capture phonemes unseen in its monolingual training data.

Data Preparation

The experiments were carried out using LJ Speech (24 hours) for English and Yoruba BibleTTS (93 hours).

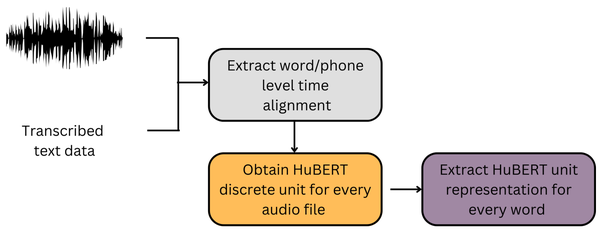


Figure 1. Data Preparation steps

HuBERT Speech representation

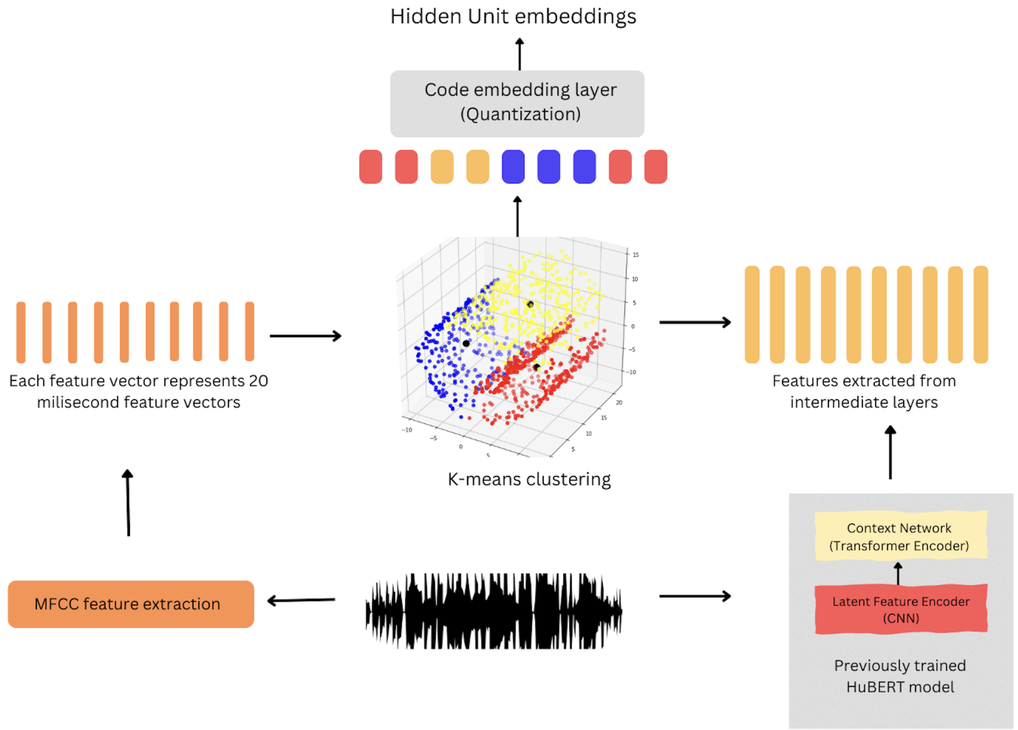


Figure 2. HuBERT approach to discovering hidden units targets through clustering, these hidden unit representations are then quantised into speech codes

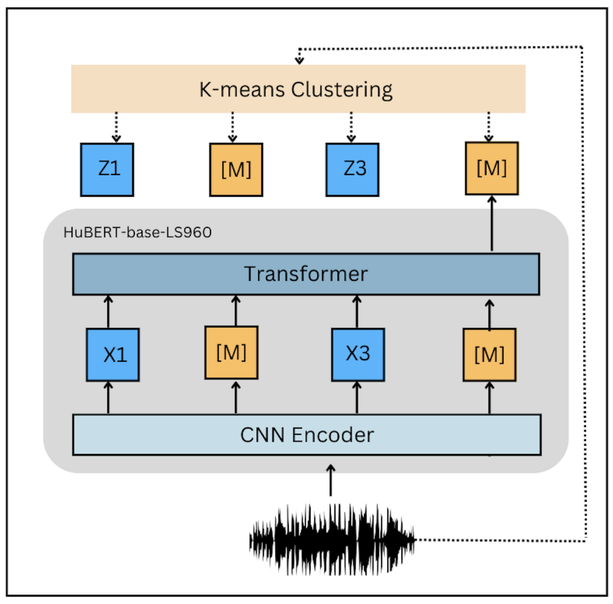


Figure 3. HuBERT masked Prediction step

Yorùbá Phonology

Consonants		Vowels		Tones	
Plosives	b,t,d,k,g,j,kp,gb	Orals	i, e, a, ẹ, o, ọ, u	Low (do)	‘
Nasals	m	Nasals	in, ẹn, ọn, un	High (mi)	´
Fricatives	f, s, ʃ, h			Mid (re)	Unmarked
Flapped	r				
Laterals	l,n				
Semi-vowels	y,w				

Evaluation and analysis

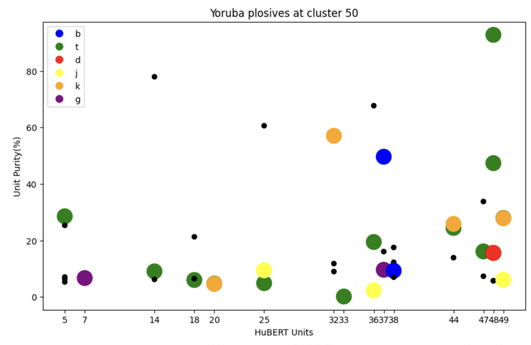
1. **Phone Purity**, Phone Purity quantifies how consistent the phone labels are within a cluster, which means how often the frames in a cluster are assigned the same phone label.Higher phone purity indicates that frames in a cluster mostly share the same phone label
2. **Unit Purity** quantifies how consistent the larger linguistic units are within a cluster. How well the representations group together similar linguistic units low unit purity would indicate that the frames within a cluster are more diverse in terms of the linguistic units they represent.
3. **PNMI**. measures how much uncertainty about the true phone labels (y) is reduced after observing the k-means clustering labels (z). The higher the PNMI value, the better the k-means clustering quality

Units	50	100	200
Phone purity	0.42	0.51	0.56
Unit purity	0.40	0.32	0.23
PNMI	0.47	0.55	0.60

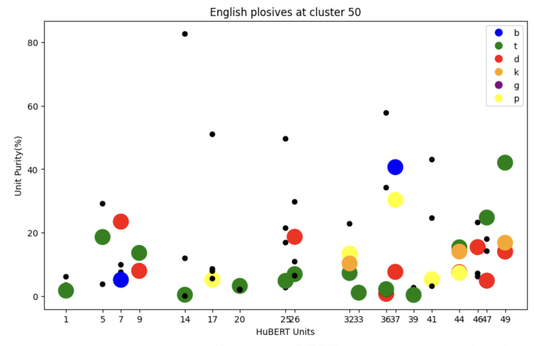
Table 1. Purity measures across different Kmeans clusters for English data

Units	50	100	200
Phone purity	0.36	0.41	0.43
Unit purity	0.38	0.35	0.30
PNMI	0.42	0.46	0.49

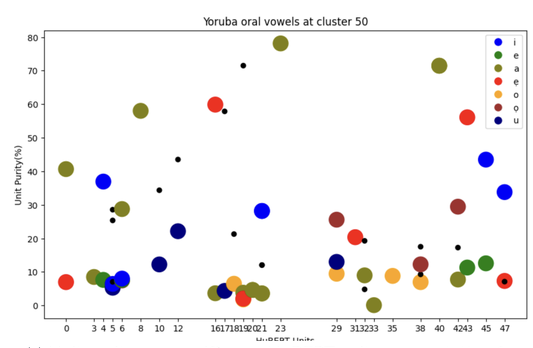
Table 2. Purity measures across different Kmeans clusters for Yoruba data



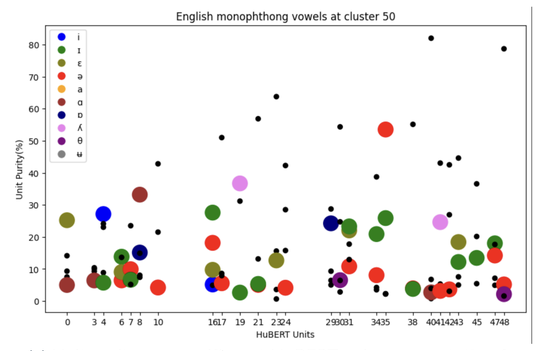
(a) Unit purity across different HuBERT unit representation for Yoruba plosive sounds



(b) Unit purity across different HuBERT unit representation for English plosive sounds



(c) Unit purity across different HuBERT unit representation for Yoruba Oral Vowel sounds



(d) Unit purity across different HuBERT unit representation for English monophthong sounds

Is the context representation layer able to distinguish between the vowel tones?

	Samples	Cluster 50 Discrete Units	ABX result
A	wà	[10, 10, 10, 38, 38, 0, 8, 8, 3, 3]	AX1: 210/AX2: 162
B	wá	[10, 10, 38, 38, 8, 8, 8, 8, 8, 8, 8, 8]	BX1: 255/BX2: 209
X1	wọ̀	[10, 10, 29, 29, 29, 29, 29, 29, 29, 42, 42, 42, 29, 41]	X1 is more similar to A
X2	wó	[45, 10, 10, 10, 10, 10, 38, 38, 35, 35, 7]	X2 is more similar to A

Table 3. ABX phone discrimination test at Cluster 50

	Samples	Cluster 100 Discrete Units	ABX result
A	wà	[70, 70, 14, 14, 24, 24, 13, 58, 58, 1]	AX1: 337/AX2: 120
B	wá	[70, 70, 14, 24, 24, 13, 13, 43, 65]	BX1: 277/BX2: 65
X1	wọ̀	[70, 48, 48, 48, 48, 48, 48, 48, 51, 51, 19, 19, 19, 99]	X1 is more similar to B
X2	wó	[69, 70, 70, 70, 70, 14, 14, 14, 76, 65, 74]	X2 is more similar to B

Table 4. ABX phone discrimination test at Cluster 100

	Samples	Cluster 50 Discrete Units	ABX result
A	sí	[11, 14, 14, 14, 14, 21, 45, 45, 45, 31, 28]	AX1: 101/AX2: 291
B	si	[34, 11, 14, 14, 14, 21, 45, 45]	BX1: 78/BX2: 287
X1	bí	[20, 18, 18, 18, 18, 7, 7, 37, 37, 37, 45, 45]	X1 is more similar to B
X2	bi	[7, 7, 7, 37, 37, 37, 37, 45, 45, 45, 45, 4, 4, 4, 4, 4, 4, 4, 4, 4]	X2 is more similar to B

Table 5. ABX phone discrimination test at Cluster 50

	Samples	Cluster 100 Discrete Units	ABX result
A	sí	[6, 36, 36, 36, 36, 7, 7, 45, 53, 44, 80]	AX1: 319/AX2: 270
B	si	[29, 6, 15, 36, 7, 7, 45, 45]	BX1: 255/BX2: 326
X1	bí	[72, 93, 10, 10, 66, 66, 66, 66, 66, 47, 45, 45]	X1 is more similar to B
X2	bi	[16, 66, 66, 66, 66, 66, 47, 11, 45, 45, 45, 45, 45, 64, 64, 64, 64, 64, 64, 64]	X2 is more similar to A

Table 6. ABX phone discrimination test at Cluster 100