Analysing Discrete Self Supervised Speech Representation for Spoken Language Modeling

Amitay Sicherman , Yossi Adi
The Hebrew University of Jerusalem

Abstract

This work profoundly analyzes discrete self-supervised speech representations (units) through the eyes of Generative Spoken Language Modeling (GSLM). Following the findings of such an analysis, we propose practical improvements to the discrete unit for the GSLM. First, we start comprehending these units by analyzing them in three axes: interpretation, visualization, and resynthesis. Our analysis finds a high correlation between the speech units to phonemes and phoneme families, while their correlation with speaker or gender is weaker. Additionally, we found redundancies in the extracted units and claim that one reason may be the units' context. Following this analysis, we propose a new, unsupervised metric to measure unit redundancies. Finally, we use this metric to develop new methods that improve the robustness of units' clustering and show significant improvement considering zero-resource speech metrics such as ABX.

Motivation

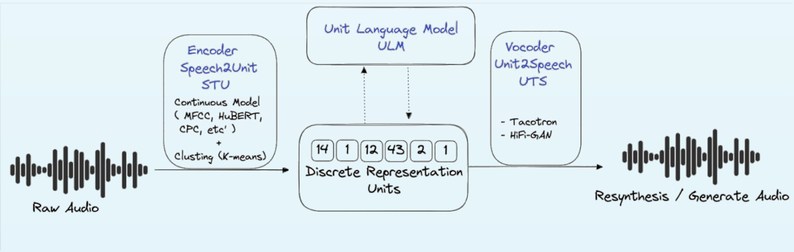
Self-Supervised Learning methods for speech have shown great success on many downstream tasks. Specifically, these Self-Supervised Learning models allow recent success in Generative Spoken Language Modeling. The basic idea behind Generative Spoken Language Modeling is to learn a discrete representation (also called “units”) of the speech signal. Note that using this method, we can use all the NLP models for the audio domain, and in addition, this setting is completely Self-Supervised - which means that we do not need any textual data. Although these models can generate meaningful and coherent speech utterances, little is known about the properties captured by these units.

Generative Spoken Language Modeling

The general pipeline consists of three main modules:

1. Speech-to-unit (STU) - the model encodes the raw speech signal into a continuous representation and then quantized the representation to a sequence of discrete units.
2. Unit language model - a standard language model that gets the discrete unit as input and can be used to generate speech conditionally or unconditionally.
3. Unit-to-speech -converts the discrete speech representation to a raw waveform

We can ignore the unit language model step to make a Speech resynthesis.



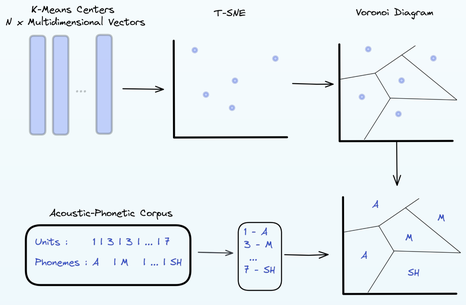
Analysis of The Discrete Unit -Interpretation

Mutual information between the units and speaker / gender / phoneme.

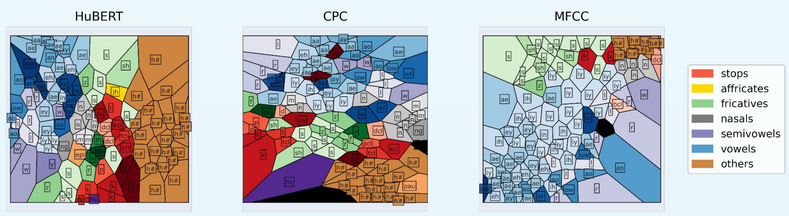
Dense Model	Vocabulary Size	Speaker	Gender	Phoneme
CPC	50	1.35	0.66	47.30
	100	2.35	0.54	**48.45**
	200	3.70	1.62	47.74
	2000	10.39	4.14	44.06
HuBERT	50	0.73	0.03	42.49
	100	1.41	0.17	45.48
	200	1.95	0.21	46.64
	2000	5.15	0.65	43.32
MFCC	50	9.11	2.90	8.57
	100	11.54	3.97	8.73
	200	13.81	4.59	8.96

Analysis of The Discrete Unit -Visualization

Create a 2D spatial view that contains information regarding the relationship between continuous representation, discrete units, and corresponding phonemes.



1. project the high-dimensional speech representation into 2D using the T-SNE
2. Use the Voronoi diagram that converts the scatter plot into an area plot.
3. Create a single label to represent each cluster using unit-phonemes alignment from the TIMIT.
4. Replace the unit id with their corresponding phonemes.
5. Color the area base on the phoneme and phoneme family.



- Units representing the same phoneme are usually close.
- Phonemes from the same phonemes family tend also to be close to each other.

Analysis of The Discrete Unit -Resynthesis

Intuition: Each unit cat represent as single ‘sound’. Unit To Speech by concatenate there sound pieces.

Key Function		repeats length	
		✔	✖
context	✔	Context-Full	Context-Single
	✖	Local-Full	Local-Single

$$LookupVocoder(u,l) = \texttt{concat}(F(u_1,l_1),\ldots,F(u_n,l_n)),$$

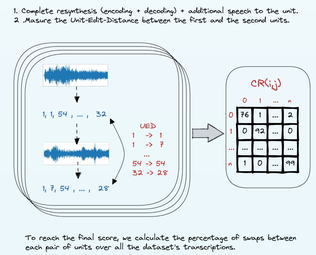
$$F(u_i,l_i) = \begin{cases} T[Key(u_i,l_i)], & \text{if } Key(u_i,l_i) \text{ in T} \\ x_i, & \text{else} \end{cases},$$

Dense Model	Vocabulary Size	Hifi-GEN	Context-Full	Context-Single	Local-Full	Local-Single
CPC	50	5.95	9.12	25.36	39.57	60.98
	100	5.67	6.52	15.21	22.51	53.59
	200	5.37	5.12	10.16	15.18	40.65
HuBERT	50	7.31	10.31	14.96	47.24	58.42
	100	4.39	5.24	6.26	26.55	57.49
	200	4.10	4.25	4.69	15.56	19.88
MFCC	50	50.47	33.85	57.60	71.43	69.22
	100	44.68	15.79	46.55	67.54	66.13
	200	41.67	6.22	30.47	61.46	61.31

High scores - Units express fixed sounds | Context ⇧⇧ length ⇧ | Context ⇔Redundancies

Circular Resynthesis

An unsupervised evaluation metric that measures discrete units' redundancies.



Robust Clustering

Step 1: K-means with k=2000.
Step 2: Merge the clusters centroids using :
Additional k-means (K-K) | Agglomerative clustering (K-H) | weighted Agglomerative clustering (K-WH)

Weighted Agglomerative Clustering: $D(i,j) = L2(c_i,c_j)\cdot[1-\frac{CR(u_i,u_j)+CR(u_j,u_i)}{2}]$

Model	Size	ABX within				ABX across				Speaker probing			
		K	K-K	K-H	K-WH	K	K-K	K-H	K-WH	K	K-K	K-H	K-WH
CPC	50	5.66	5.38	9.62	8.80	7.83	6.77	11.46	10.56	42.22	32.96	19.26	18.15
	100	5.42	5.44	6.66	6.04	7.07	7.13	8.26	7.49	52.96	45.19	20.37	15.56
	200	5.53	5.27	5.61	5.68	7.35	7.10	7.28	7.13	63.70	49.63	26.30	22.59
HuBERT	50	7.23	5.67	5.94	6.12	8.93	6.83	7.43	7.67	30.37	36.30	36.67	31.85
	100	5.82	**5.01**	5.30	5.29	7.47	6.50	6.54	6.32	48.15	48.89	48.15	46.67
	200	5.79	5.24	5.18	5.05	7.49	6.42	6.46	**6.07**	65.19	61.11	54.81	62.96

Acknowledgements

We would like to acknowledge support for this research from the Israeli Science Foundation (ISF grant 2049/22).

Project Page : https://amitaysicherman.github.io/SLM-discrete-representation/