Analysing Discrete Self Supervised Speech Representation for Spoken Language Modeling

Amitay Sicherman , Yossi Adi
The Hebrew University of Jerusalem

Abstract

Discrete self-supervised speech representations (units) through the eyes of Generative Spoken Language Modeling (GSLM)

Analysis

- Interpretation. Visualization. Resynthesis.
- Correlation between units and the phonemes.
- Redundancies ⇔ context.

Unsupervised metric to measure unit redundancies.
Improve the robustness of units' clustering.

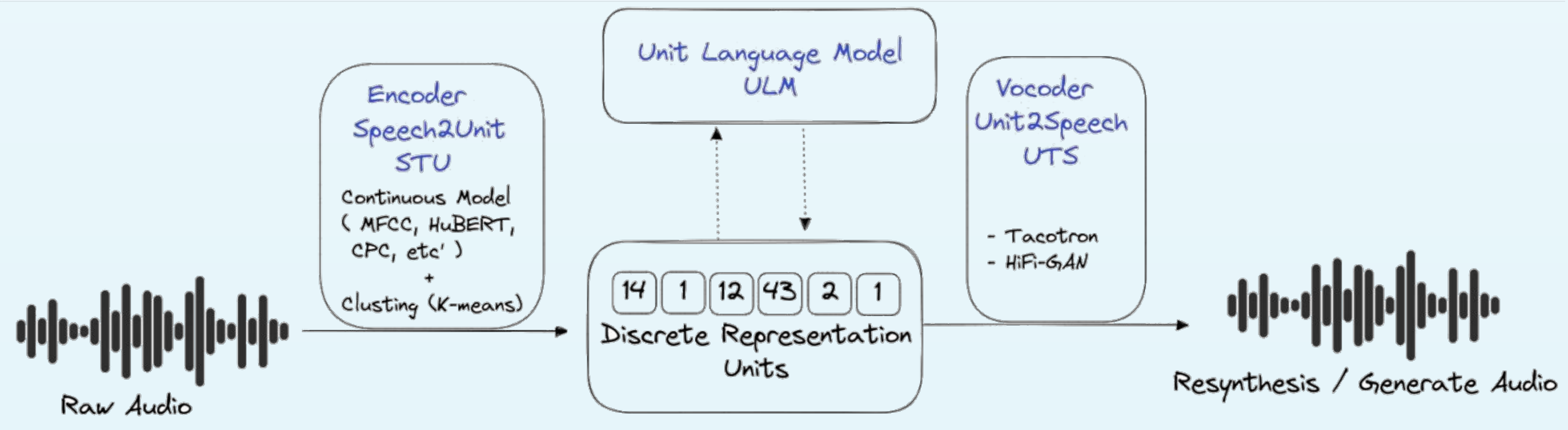
Motivation

- **SSL for speech** - great success, specifically in Generative Spoken Language Modeling (GSLM).
- **GSLM** - learn a discrete representation of the speech signal. generate meaningful and coherent speech.
- **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)**
2. **Unit language model (ULM)**
3. **Unit-to-speech (UTS)**



Analysis of The Discrete Unit -Interpretation

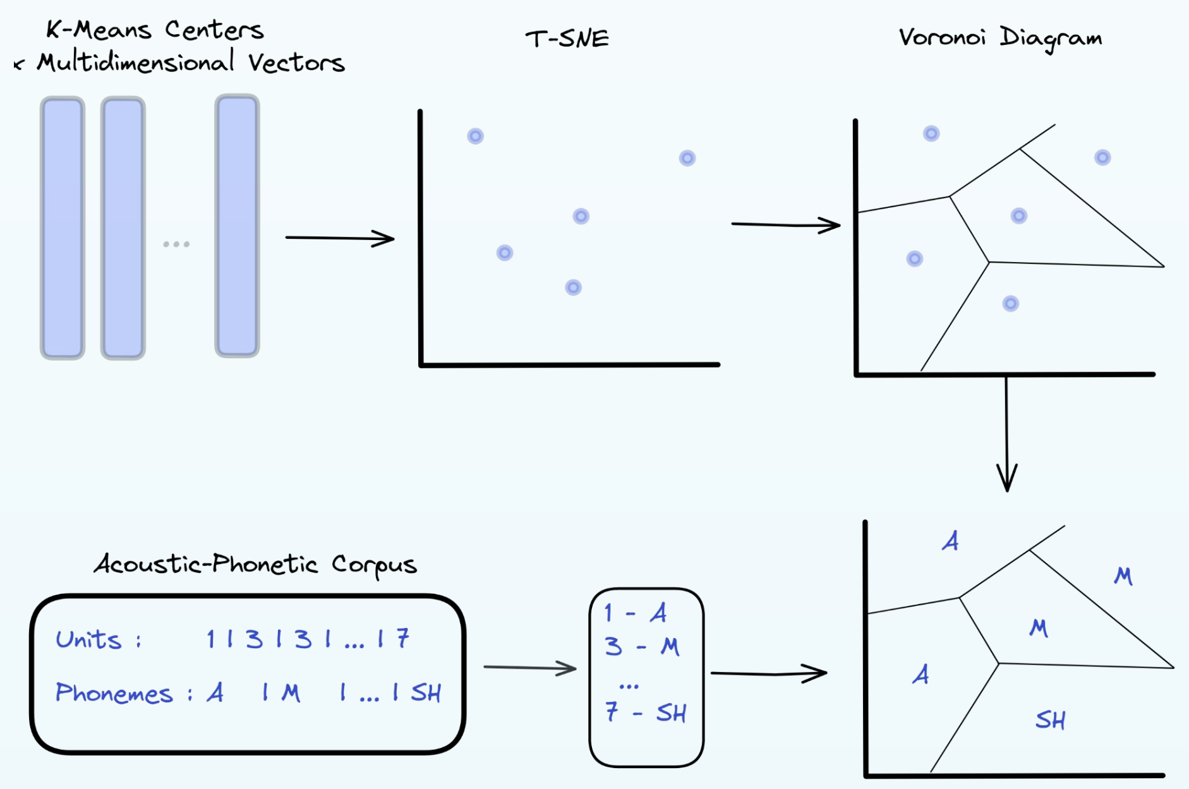
Mutual information between the units and speaker / gender / phoneme.

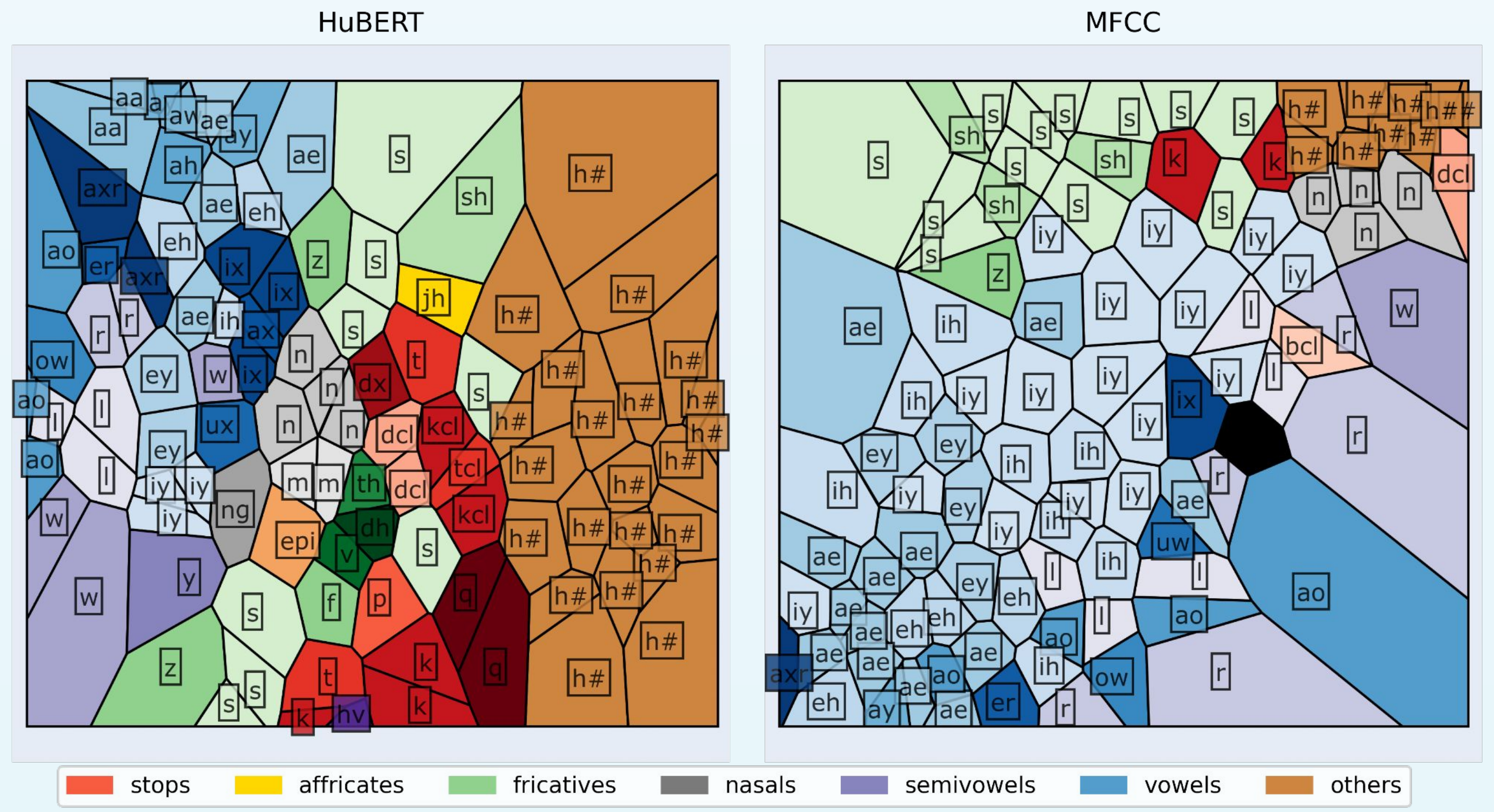
Dense Model	Vocabulary Size	Speaker	Gender	Phoneme
CPC	50	1.35	0.66	47.30
	100	2.35	0.54	**48.45**
	200	3.70	1.62	47.74
HuBERT	50	0.73	0.03	42.49
	100	1.41	0.17	45.48
	200	1.95	0.21	46.64
MFCC	50	9.11	2.90	8.57
	100	11.54	3.97	8.73
	200	13.81	.59	8.96

Analysis of The Discrete Unit -Visualization

Visualization for continuous representation, discrete units, and the phonemes

1. **T-SNE :**
   Continuous representation into 2D.
2. **Voronoi:**
   Scatter plot into an area plot.
3. **Alignment:**
   Unit to phoneme.
4. **Color:**
   Base on the phoneme and phoneme family.


K-Means Centers
< Multidimensional Vectors
T-SNE
Voronoi Diagram
Acoustic-Phonetic Corpus
Units : 1 1 3 1 3 1 ... 1 7
Phonemes : A 1 M 1 ... 1 SH
1 - A
3 - M
7 - SH
A
M
SH



Units representing the same phoneme /phonemes family are usually close.

Analysis of The Discrete Unit -Resynthesis

Intuition: Each unit represent as single 'sound'.

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

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

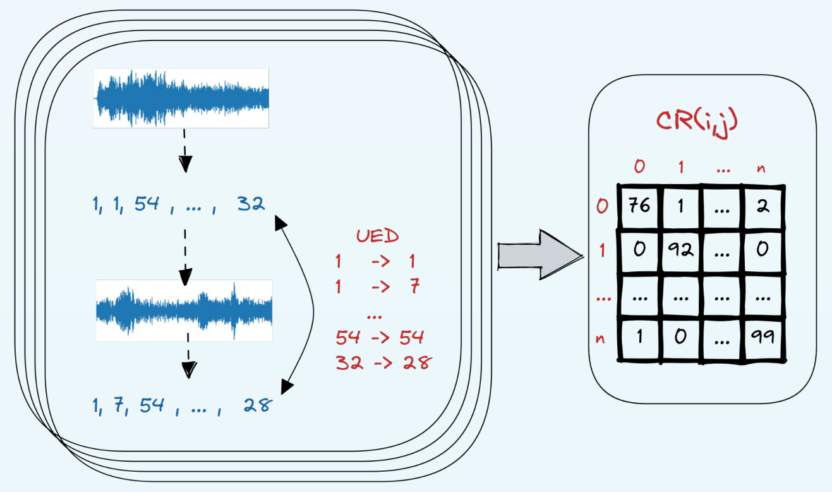
$$F(u_i,l_i) = \begin{cases} T[Key(u_i,l_i)], & \text{if } Key(u_i,l_i) \text{ in T} \\ x_i, & \text{else} \end{cases},$$

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 | 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.

How?

- K-means (K-K)
- Agglomerative clustering (K-H)
- Weighted Agglomerative clustering (K-WH)

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

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/