LING-L542: Phonological Analysis Final Project

ANDREW DAVIS, LILY KAWAOTO FALL 2021

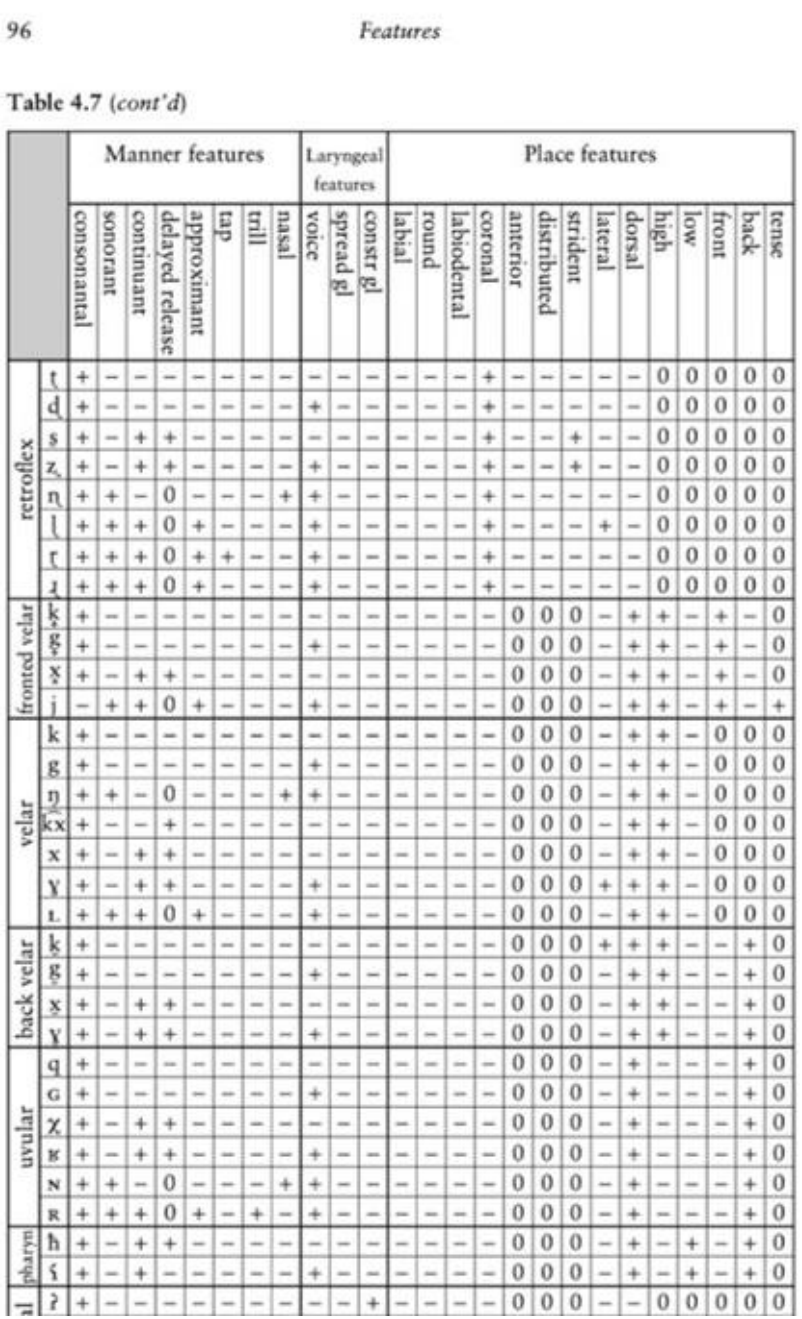
Introduction

Imagine, as a Linguist/Phonologist, you have a body of text or a dataset from the language of your study, and you want to extract the distinctive, phonological features demonstrated in the sounds/text provided...

What are we doing? Why does it matter?

- **Phase I:**
- take data from a language in terms of orthographic/IPA text
- produce the relevant Phonological features & assign relevant features to corresponding IPA symbol
- group the phonemes based on a matrix with qualitative feature weights
- produce an array of this data for the user
- **Phase II – Analyze & Refine:**
- incorporate probabilities (e.g. of bigrams of phonemes) into the model, which will improve our relationship graph
- Train a neural network model with these, compare quantitative frequency with the qualitative data to identify core vs peripharal sounds of a language; Utilizing Deep Learning & Dense Embedding
- **Why does this matter?**
- Bridge the gap in terms of manual labor associated with analyzing a dataset in isolating core vs peripharal sounds of a language
- Determine/predict core sounds from peripheral sounds in languages: L2 Acquisition, Speech Recognition/NLP (AI) (GPT-3 Example), Jenga Example

Methodology



96 *Features*

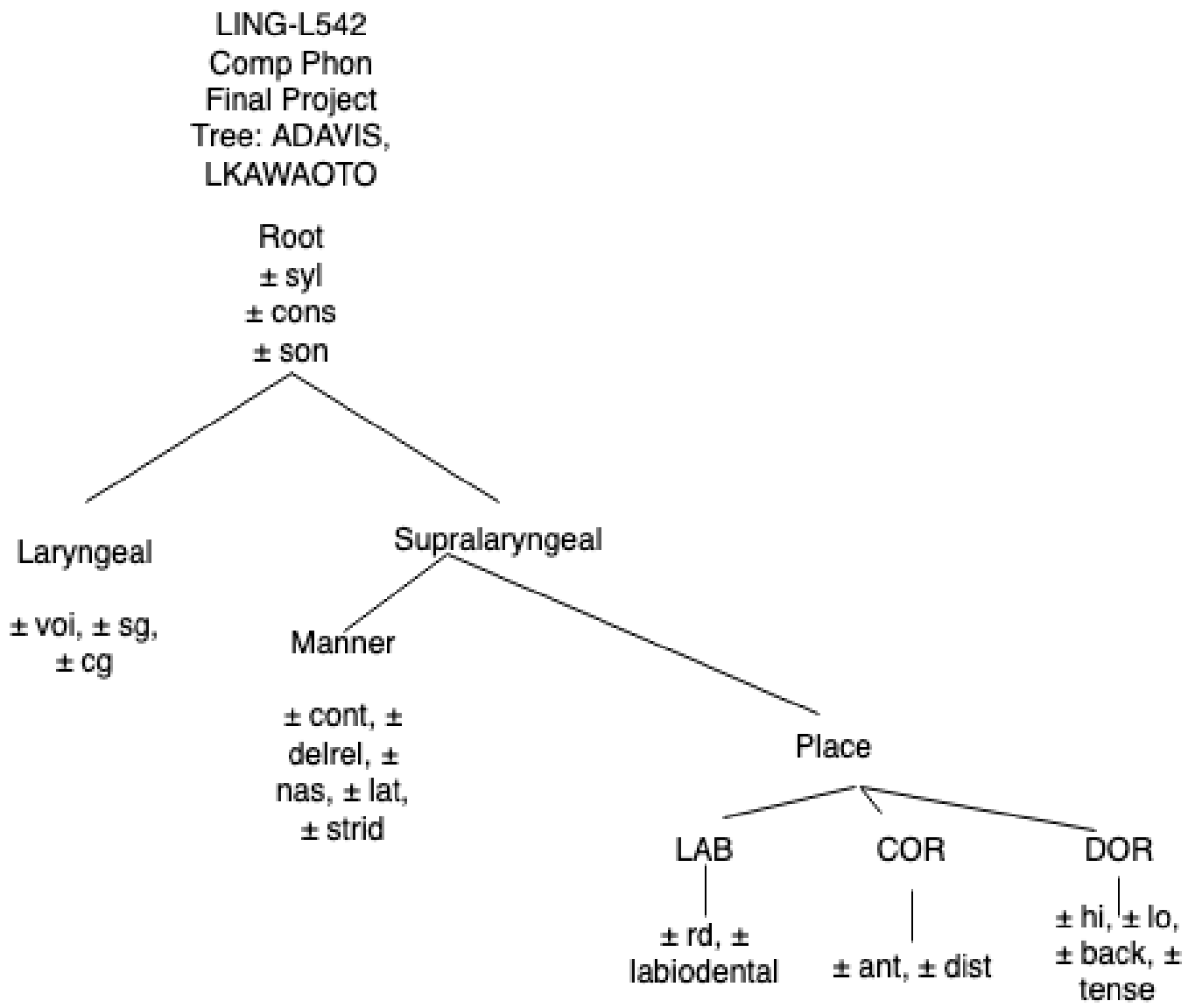
Table 4.7 (*cont'd*)

		Manner features								Laryngeal features			Place features													
		consonantal	sonorant	continuant	delayed release	approximant	tap	trill	nasal	voice	spread gl	constr gl	labial	round	labiodental	coronal	anterior	distributed	strident	lateral	dorsal	high	low	front	back	tense
retroflex	ʈ	+	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−	0	0	0	0	0
	ɖ	+	−	−	−	−	−	−	−	+	−	−	−	−	−	+	−	−	−	−	−	0	0	0	0	0
	ʂ	+	−	+	+	−	−	−	−	−	−	−	−	−	−	+	−	−	+	−	−	0	0	0	0	0
	ʐ	+	−	+	+	−	−	−	−	+	−	−	−	−	−	+	−	−	+	−	−	0	0	0	0	0
	ɳ	+	+	−	0	−	−	−	+	+	−	−	−	−	−	+	−	−	−	−	−	0	0	0	0	0
	ɭ	+	+	+	0	+	−	−	−	+	−	−	−	−	−	+	−	−	−	+	−	0	0	0	0	0
	ɽ	+	+	+	0	+	+	−	−	+	−	−	−	−	−	+	−	−	−	−	−	0	0	0	0	0
	ɻ	+	+	+	0	+	−	−	−	+	−	−	−	−	−	+	−	−	−	−	−	0	0	0	0	0
fronted velar	k̟	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	0	0	0	−	+	+	−	+	−	0
	g̟	+	−	−	−	−	−	−	−	+	−	−	−	−	−	−	0	0	0	−	+	+	−	+	−	0
	x̟	+	−	+	+	−	−	−	−	−	−	−	−	−	−	−	0	0	0	−	+	+	−	+	−	0
	j	−	+	+	0	+	−	−	−	+	−	−	−	−	−	−	0	0	0	−	+	+	−	+	−	+
velar	k	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	0	0	0	−	+	+	−	0	0	0
	g	+	−	−	−	−	−	−	−	+	−	−	−	−	−	−	0	0	0	−	+	+	−	0	0	0
	ŋ	+	+	−	0	−	−	−	+	+	−	−	−	−	−	−	0	0	0	−	+	+	−	0	0	0
	kx	+	−	−	+	−	−	−	−	−	−	−	−	−	−	−	0	0	0	−	+	+	−	0	0	0
	x	+	−	+	+	−	−	−	−	−	−	−	−	−	−	−	0	0	0	−	+	+	−	0	0	0
	ɣ	+	−	+	+	−	−	−	−	+	−	−	−	−	−	−	0	0	0	+	+	+	−	0	0	0
	ɤ	+	+	+	0	+	−	−	−	+	−	−	−	−	−	−	0	0	0	−	+	+	−	0	0	0
back velar	k̠	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	0	0	0	+	+	+	−	−	+	0
	g̠	+	−	−	−	−	−	−	−	+	−	−	−	−	−	−	0	0	0	−	+	+	−	−	+	0
	x̠	+	−	+	+	−	−	−	−	−	−	−	−	−	−	−	0	0	0	−	+	+	−	−	+	0
	ɣ̠	+	−	+	+	−	−	−	−	+	−	−	−	−	−	−	0	0	0	−	+	+	−	−	+	0
uvular	q	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	0	0	0	−	+	−	−	−	+	0
	ɢ	+	−	−	−	−	−	−	−	+	−	−	−	−	−	−	0	0	0	−	+	−	−	−	+	0
	χ	+	−	+	+	−	−	−	−	−	−	−	−	−	−	−	0	0	0	−	+	−	−	−	+	0
	ʁ	+	−	+	+	−	−	−	−	+	−	−	−	−	−	−	0	0	0	−	+	−	−	−	+	0
	ɴ	+	+	−	0	−	−	−	+	+	−	−	−	−	−	−	0	0	0	−	+	−	−	−	+	0
	ʀ	+	+	+	0	+	−	+	−	+	−	−	−	−	−	−	0	0	0	−	+	−	−	−	+	0
pharyn	ħ	+	−	+	+	−	−	−	−	−	−	−	−	−	−	−	0	0	0	−	+	−	+	−	+	0
	ʕ	+	−	+	−	−	−	−	−	+	−	−	−	−	−	−	0	0	0	−	+	−	+	−	+	0
al	ʔ	+	−	−	−	−	−	−	−	−	−	+	−	−	−	−	0	0	0	−	−	0	0	0	0	0

Step 1a: Create a Table of IPA Features

- Based off the IPA feature chart created for PanPhon
- Make the chart more closely aligned with Hayes (2009)
  - Added Dorsal, labiodental
  - Removed high/low tone

Hayes (2009). *Introductory Phonology.*



Step 1b: Create a Table of IPA Features – Create Tree

- Create Feature Geometry Tree of relevant features in this Data to act as standard; keeps it uniform

	ipa	syl	cons	son	cont	delrel	nas	voi	lat	strid	LAR	sg	cg	COR	ant	distr	LAB	round	labiodental	DOR	hi	lo	back	tense
2	g	-	+	-	-	-	-	+	-	-	0	-	-	0	-	-	0	-	-	+	+	-	-	-
3	k	-	+	-	-	-	-	-	-	-	0	-	-	0	-	-	0	-	-	+	+	-	-	-
4	b	-	+	-	-	-	-	+	-	-	0	-	-	0	-	-	+	-	-	0	-	-	-	-
5	d	-	+	-	-	-	-	+	-	-	0	-	-	+	+	-	0	-	-	0	-	-	-	-
6	p	-	+	-	-	-	-	-	-	-	0	-	-	0	-	-	+	-	-	0	-	-	-	-
7	t	-	+	-	-	-	-	-	-	-	0	-	-	+	+	-	0	-	-	0	-	-	-	-
8	d͡ʒ	-	+	-	-	+	-	+	-	+	0	-	-	+	-	+	0	-	-	0	-	-	-	-
9	t͡ʃ	-	+	-	-	+	-	-	-	+	0	-	-	+	-	+	0	-	-	0	-	-	-	-
10	ʃ	-	+	-	+	+	-	-	-	+	0	-	-	+	-	+	0	-	-	0	-	-	-	-
11	ʒ	-	+	-	+	+	-	+	-	+	0	-	-	+	-	+	0	-	-	0	-	-	-	-
12	f	-	+	-	+	+	-	-	-	-	0	-	-	0	-	-	+	-	+	0	-	-	-	-
13	s	-	+	-	+	-	-	-	-	+	0	-	-	+	+	-	0	-	-	0	-	-	-	-
14	v	-	+	-	+	+	-	+	-	-	0	-	-	0	-	-	+	-	+	0	-	-	-	-
15	z	-	+	-	+	-	-	+	-	+	0	-	-	+	+	-	0	-	-	0	-	-	-	-
16	ð	-	+	-	+	-	-	+	-	-	0	-	-	+	+	+	0	-	-	0	-	-	-	-
17	θ	-	+	-	+	+	-	-	-	-	0	-	-	+	+	+	0	-	-	0	-	-	-	-
18	ʔ	-	+	+	-	-	-	-	-	-	+	-	+	0	-	-	0	-	-	0	-	-	-	-
19	ŋ	-	+	+	-	-	+	+	-	-	0	-	-	0	-	-	0	-	-	+	+	-	-	-
20	m	-	+	+	-	-	+	+	-	-	0	-	-	0	-	-	+	-	-	0	-	-	-	-
21	n	-	+	+	-	-	+	+	-	-	0	-	-	+	+	-	0	-	-	0	-	-	-	-
22	h	-	+	-	+	+	-	-	-	-	+	+	-	0	-	-	0	-	-	0	-	-	-	-
23	j	-	-	+	+	-	-	+	-	-	0	-	-	0	-	-	0	-	-	+	+	-	-	+
24	w	-	-	+	+	-	-	+	-	-	0	-	-	0	-	-	+	+	-	+	+	-	+	+
25	ɹ	-	-	+	+	-	-	+	-	-	0	-	-	+	-	+	0	-	-	0	-	-	-	-
26	l	-	+	+	+	-	-	+	+	-	0	-	-	+	+	-	0	-	-	+	-	-	-	-

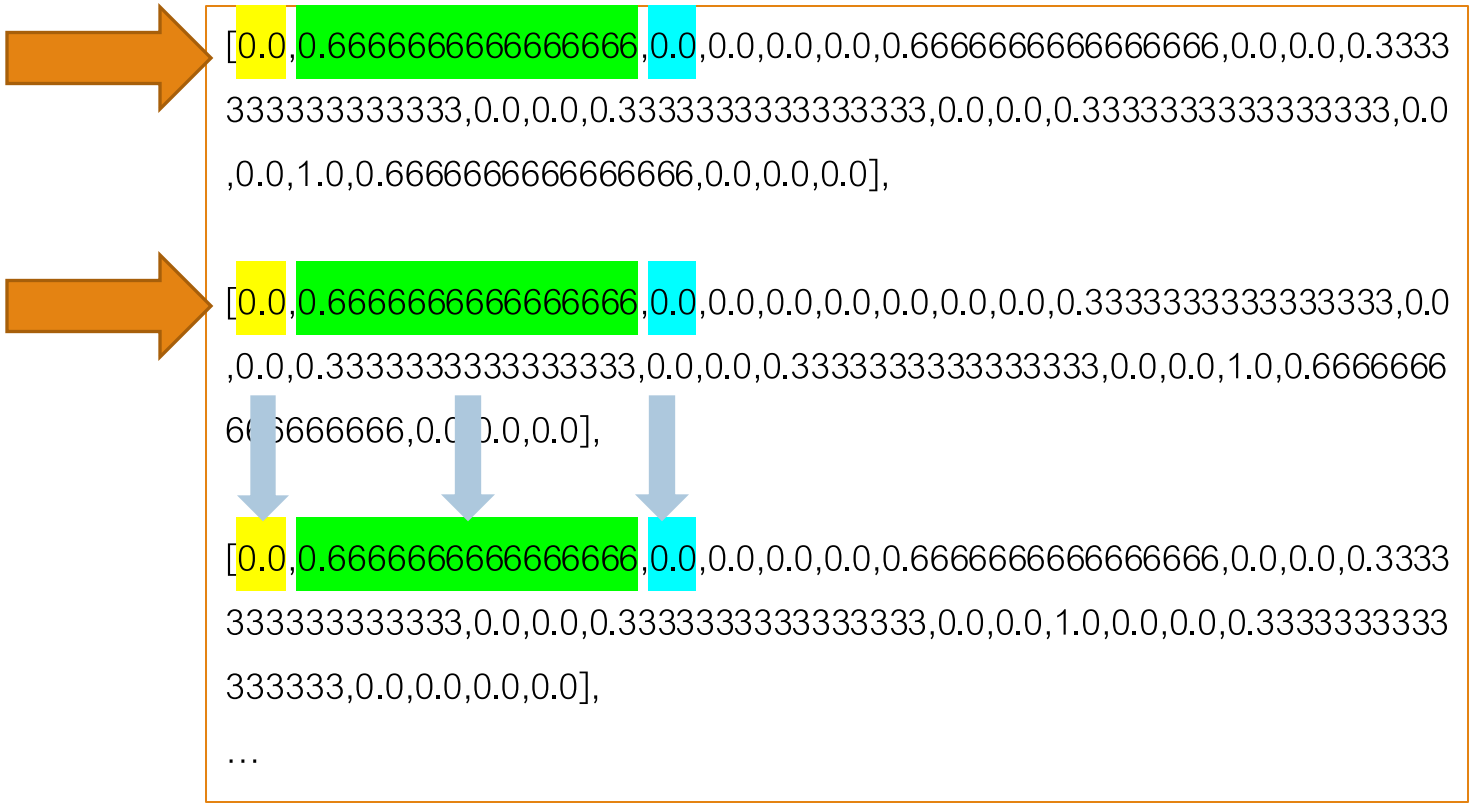
Our final chart of English IPA symbols and their features

Step 1c: Create a Table of IPA Features

- Based off the IPA feature chart created for PanPhon
- Make the chart more closely aligned with Hayes (2009)
  - Added Dorsal, labiodental
  - Removed high/low tone

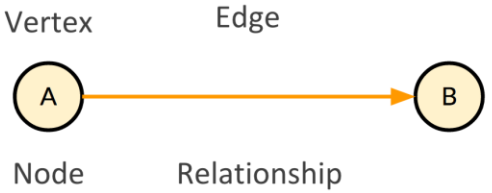
Step 2: Represent features into Computer-Readable Format

- How to represent +/- features into reasonable numbers?
- Feature geometry tree...
- Assign "weights" to each feature
  - +1 --> presence of feature (binary)
    +2 --> presence of feature (unary)
  - -1 --> absence of feature
  - 0 --> lack of feature
- These numbers get stored in an abstract computer "container" called a vector
- Normalized numbers in the vectors to be between 0 and 1



Step 3: Visualize Feature Vectors

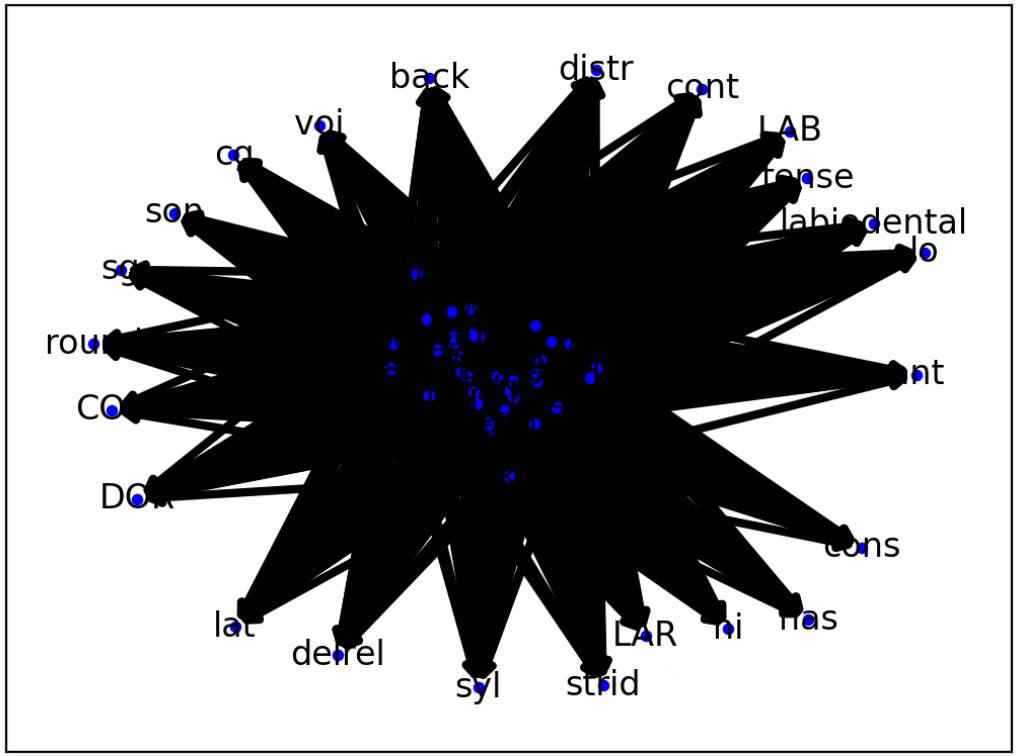
- Used another open-source programming package called NetworkX to make graphs
- How to represent graphs?



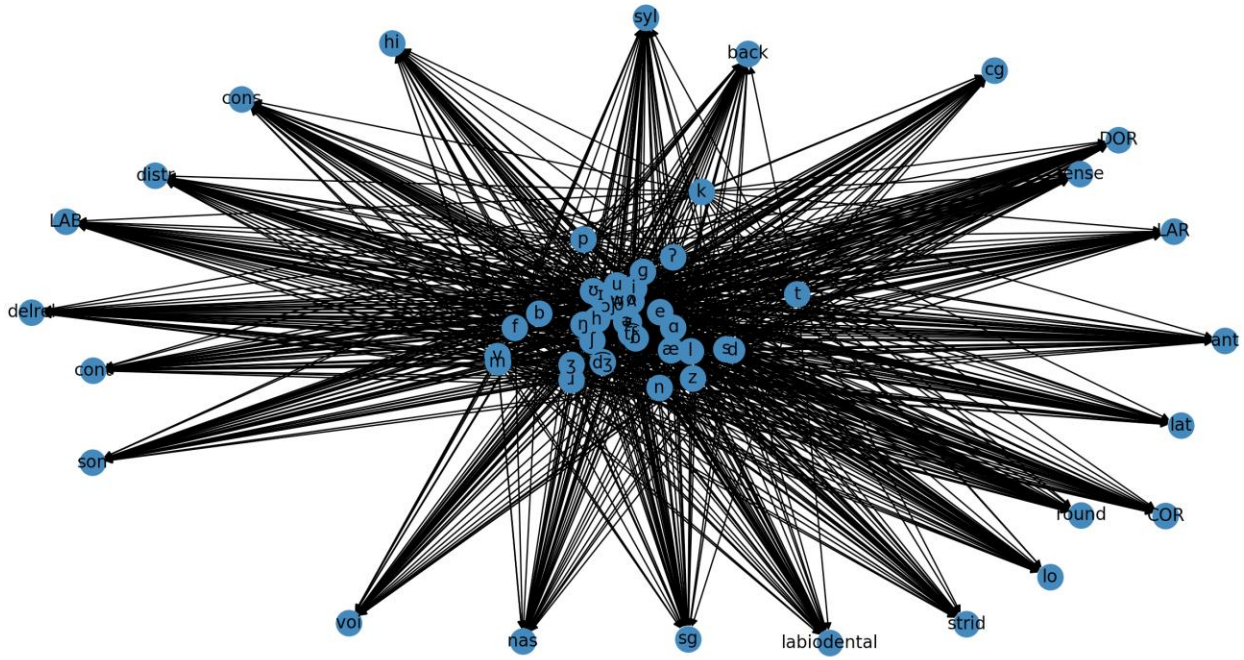
  - An edge connects 2 nodes --> [Node1, Node2, EdgeWeight]
    - Node1 = IPA symbol
    - Node2 = feature
    - EdgeWeight = corresponding number in the feature vector from Step 2

["k", "syl", 0.0], ["k", "cons", 0.6666666666666666], ["k", "son", 0.0], ["k", "cont", 0.0], ["k", "delrel", 0.0], ["k", "nas", 0.0], ["k", "voi", 0.0], ["k", "lat", 0.0], ["k", "strid", 0.0], ["k", "LAR", 0.3333333333333333], ["k", "sg", 0.0], ["k", "cg", 0.0], ["k", "COR", 0.3333333333333333], ["k", "ant", 0.0], ["k", "distr", 0.0], ...

Step 3: Visualize feature vectors



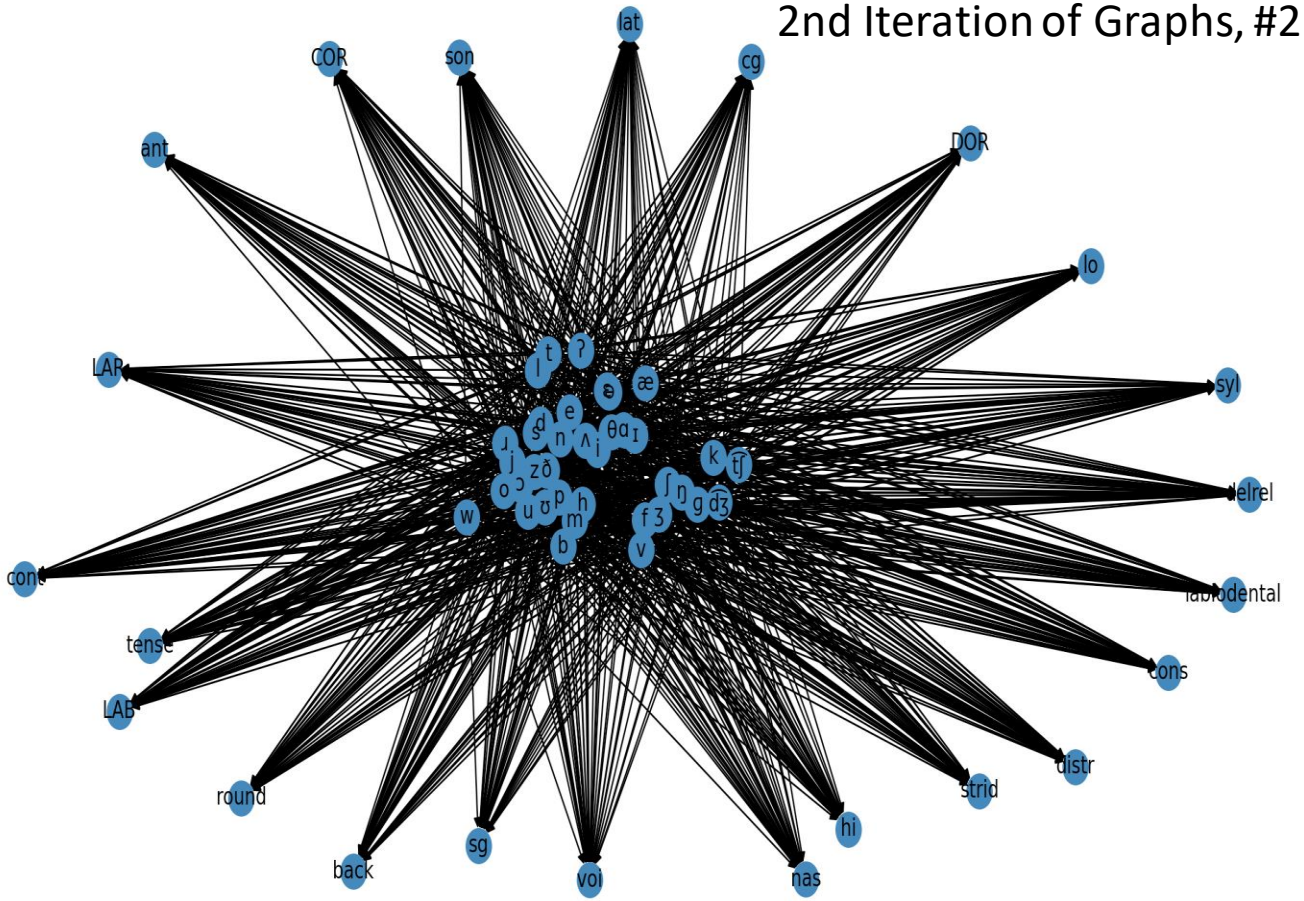
1st Iteration of Graphs

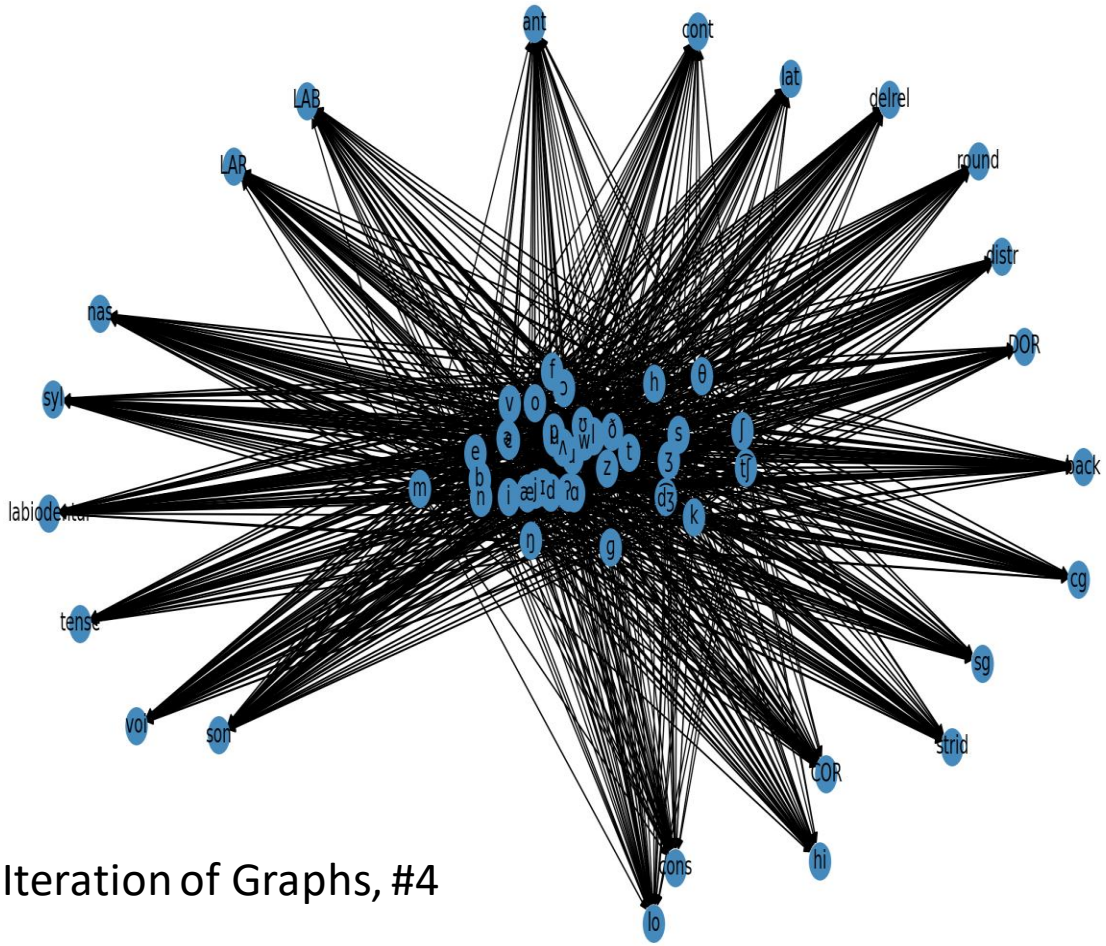


2nd Iteration of Graphs, #1

Progress of graphs after tweaking settings?

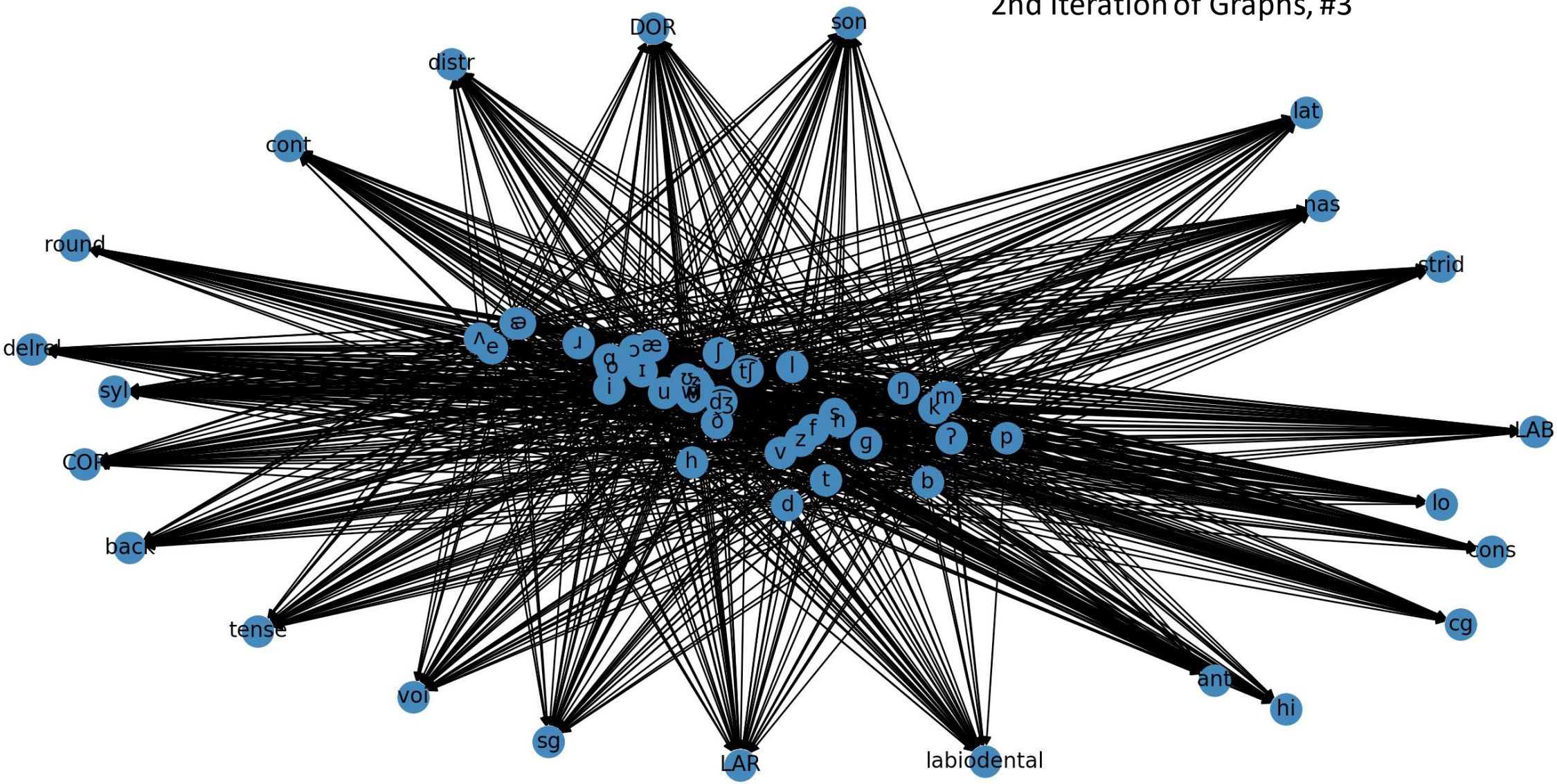
2nd Iteration of Graphs, #2



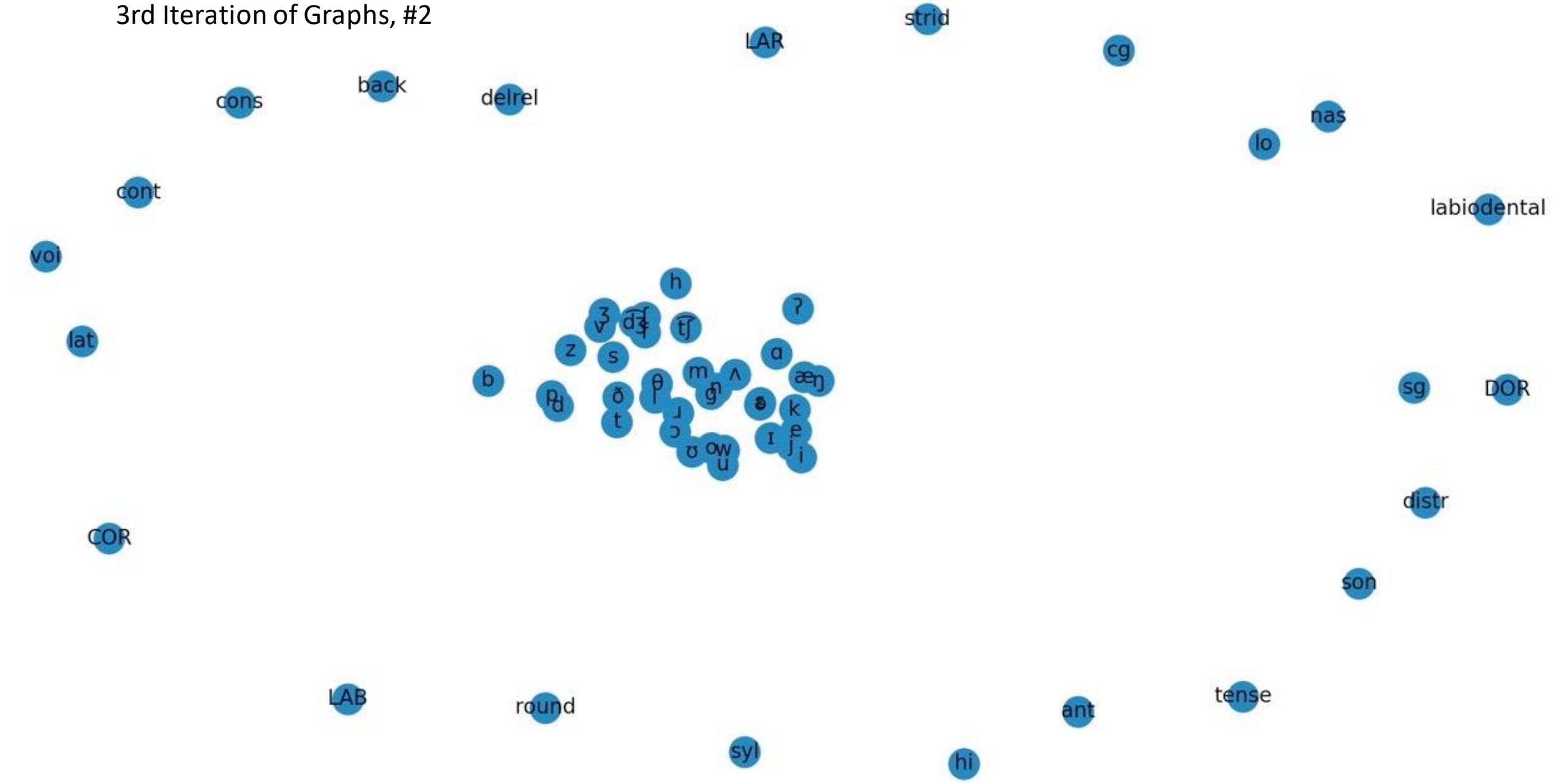


2nd Iteration of Graphs, #4

2nd Iteration of Graphs, #3



3rd Iteration of Graphs, #2



Graph Analysis

Do you agree / disagree with the graph representation?

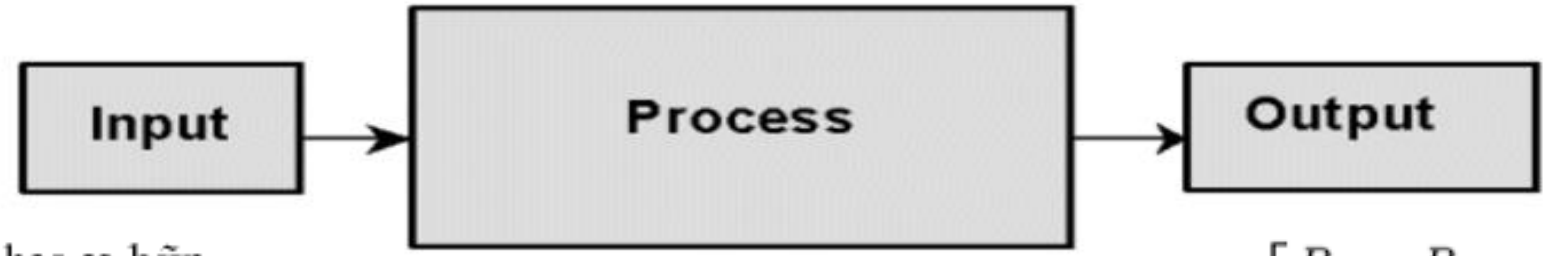
What do you notice about the IPA symbols and their position in relation to the features?

Recall back to our table, did you agree with the changes to Hayes representation in terms binary/unary features that we made?

Phase I Analysis/Observations: manner > place > voice

Next Steps

Our Roadmap



[pʰliːz kɑl stɛlə æsk həɹ ɾə bɹ̃ĩŋ
n̥iːz s̥ĩŋz wɪθ həɹ fɹʌ̃m nə stɔːɹ
sɪks spũːnz əv fɹɛʃ ʃnoʊ piːz
faɪv θɪk slæbz ə̌v blu tʃiːz ə̃m
meɪbi ə snæk fəɹ həɹ bɹʌðəɹ
bɑːb wi ɑlsŏ nid ə smɔl plæstɪk
sneɪk ə̃n ə bɪg tʰɔɪ fɹɔːg fəɹ ðə
kʰɪdz ʃi kə̃n skup ðiːz s̥ĩŋz ĩntə
θɹi ɹɛd bæːgz ə̃n wi wɪl goʊ
miʔ həɹ wɛ̃nzdeɪ ət ðə tɹɛ̃ɪn
steɪʃə̃n]

```
1101010101010100101000010
1001011110000010001001010
0101010101010010111000011
0100001010010101110001001
0101010100001111101010100
0101010101010101001000011
```

$$P = \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,j} & \cdots & P_{1,S} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,j} & \cdots & P_{2,S} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_{i,1} & P_{i,2} & \cdots & P_{i,j} & \cdots & P_{i,S} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_{S,1} & P_{S,2} & \cdots & P_{S,j} & \cdots & P_{S,S} \end{bmatrix}.$$

"gold standard"

Future Applications

Compare qualitative data to quantitative data based on distributional frequency in phonotactic datasets (I.e. frequency of x occurring next to y)

- Scaling weight of relationships based on above; creates a magnetic attraction between like-sounds, phonemic clusters, and entire languages (core elements at the center and peripharal sounds on the fringes)
- Compare this with qualitative data analyze accuracy, refine, re-train model (Dense Embeddings for each sound)
- Create a standard/score based on this etc
- Graphs Validate the Data

Native vs L2 English speakers, L2 Acquisition

- Measure distance (could be score mentioned above) in terms of acoustic and articulatory differ

Tentative Datasets

- Annotated Data Available: http://www.voxforge.org/ , https://sla.talkbank.org/TBB/childes/Other/Arabic/Salama/AbdrahmanFawzy.cha , etc
- Working with locally annotated data

References

Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, “Exploring network structure, dynamics, and function using NetworkX”, in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, Lori Levin (2016). "PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors." *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan, December 11-17 2016.