Predicting Continuous Vowel Spaces in the Wilderness

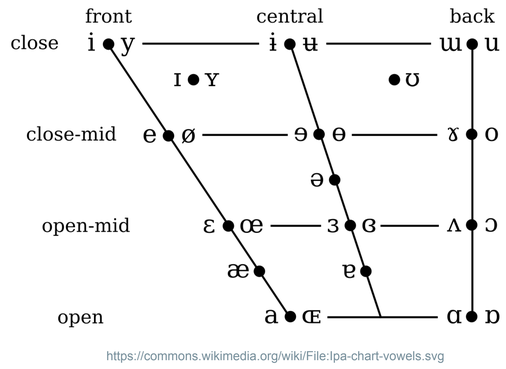
Emily Ahn and **David Mortensen**

Language Technologies Institute
Carnegie Mellon University

https://www.jetsetter.com/uploads/sites/7/2018/04/XiS9bzs5.jpeg

1. Motivation

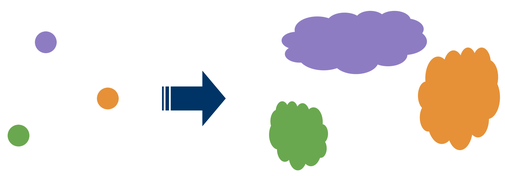
Traditionally, vowel **types** look like this:



https://commons.wikimedia.org/wiki/File:Ipa-chart-vowels.svg

Our Goals:

(1) Analyze vowel features as **continuous** distributions, rather than **discrete** points.



(2) Generate methodology to test phonological typology theories.

Dispersion Theory

*Vowel **types** (centroids) are maximally dispersed within vowel space.*

Focalization Theory

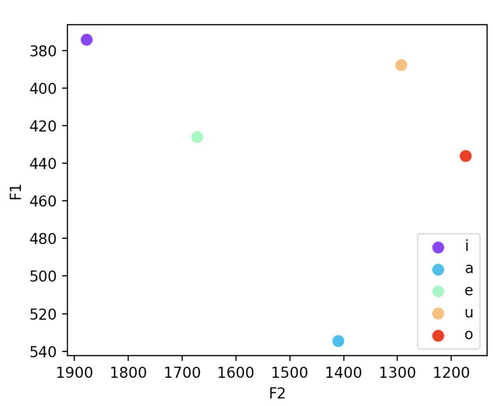
*Vowel **types** (centroids) are centered around canonical focii.*

Exemplar Theory

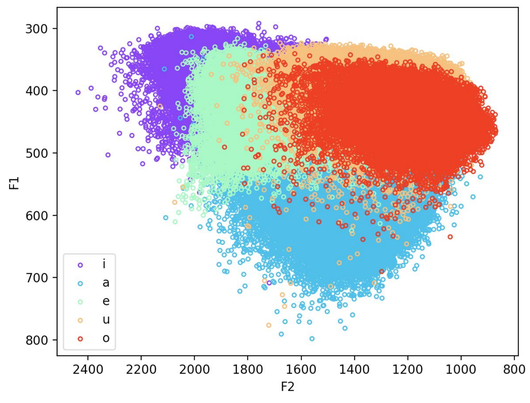
*Vowel **tokens** (clouds) are distributed around vowel **types** with lesser overlap.*

Ex: *Manado Malay in formant space*

Vowel **types** (centroids)



Vowel **tokens** (occurrences)

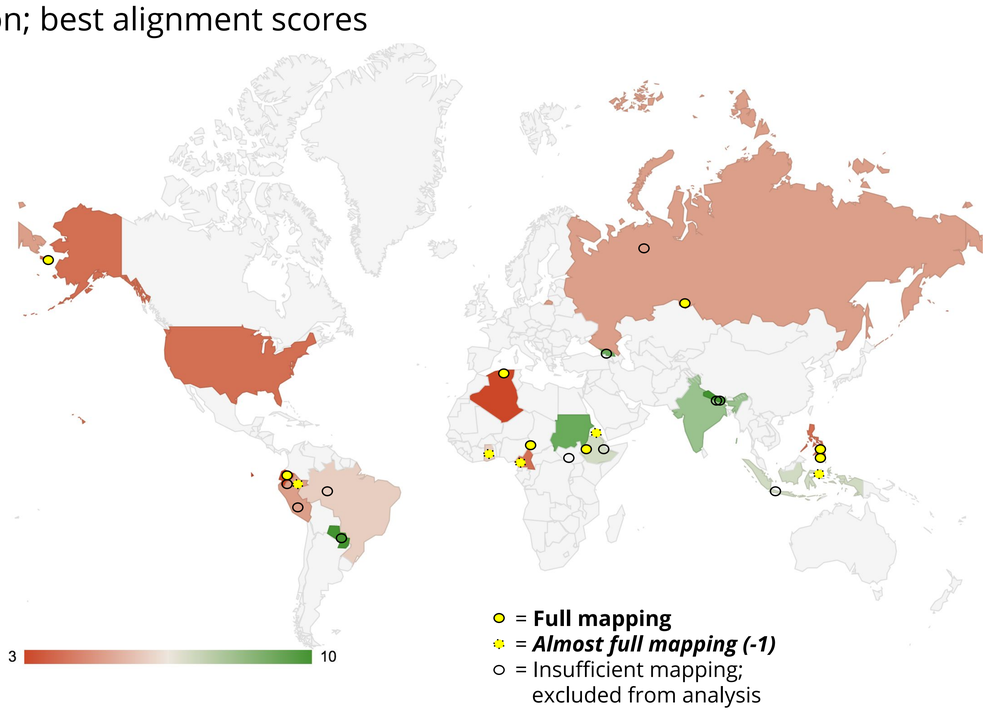


2. Data

We select **24 languages** from the CMU Wilderness Multilingual Speech Dataset (Black, 2019)

- Bible read speech
- Balanced for # vowel types & region; best alignment scores

Language	Country	Vowel types	Vowel tokens	Hours
Cebuano	Philippines	**3**	87,984	22
Kabyle	Algeria	**3**	18,610	8
Tena Quechua	Ecuador	**3**	68,931	19
Maranao	Philippines	**4**	96,813	24
Podoko	Cameroon	**4**	47,279	21
Yupik	United States	**4**	67,798	22
Russian	Russia	**5**	31,851	15
Twampa	Ethiopia	**5**	75,275	31
Urarina	*Peru*	**5**	137,860	31
Hanga	*Ghana*	**6**	44,707	14
Manado Malay	*Indonesia*	**6**	86,631	25
Paumari	Brazil	6	--	48
Komi	Russia	7	--	17
Sundanese	Indonesia	7	--	20
Tigrinya	*Ethiopia*	**7**	21,552	14
Denya	*Cameroon*	**8**	23,441	15
Huambisa	Peru	8	--	28
Maithili	India	8	--	14
Moru	Sudan	9	--	23
Nomatsigenga	Peru	9	--	36
Ossetian	Georgia	9	--	12
Eastern Oromo	Ethiopia	10	--	24
Maka	Paraguay	10	--	29
Tamang	Nepal	10	--	18



Align phones → **Map phones** → **Extract formants**

Automatic:	*Manual*:	*Automatic*:
`Festvox`	`PHOIBLE + Festvox`	`DeepFormants`

3. Methodology

Hypothesis

*Following **Exemplar Theory**, a language with more vowels would have a narrower distribution of vowel tokens.*

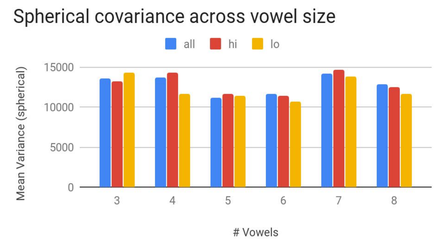
(1) Model each vowel type's set of tokens with a bivariate **gaussian** distribution.
(2) Take **variance** as a measure of the distribution (cloud size).

4. Analysis

Finding

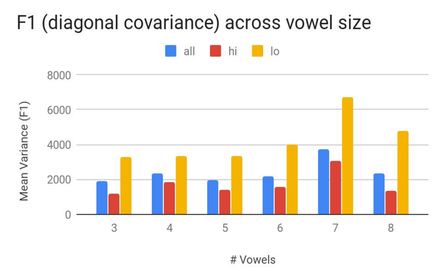
No significant correlation between # vowels and variance.

A.

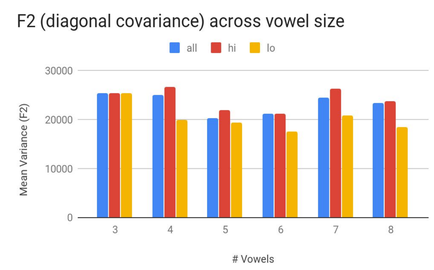


Color = vowel height

B.



C.



Discussion

(1) When discarding Russian, there is significant correlation for low vowels to increase F1 variance ($p < 0.05$) [chart **B**].
→ Exemplar Theory is contradicted.

(2) Token variance (cloud size) is insensitive to size of vowel inventory [charts **A** and **C**].
→ Exemplar Theory is unsupported.

5. Conclusion

(1) Methods to massively analyze vowels distributions, while previously impossible, are still **challenging** due to quality of automated data analysis.

(2) Our Exemplar Theory hypothesis is **not supported** (for now).

Future Work

- Try different models (e.g. Determinantal Point Processes)
- Expand language set
- Examine other effects
  - ➢ Contextual phonological environment
    - Ex: [æ] → [k æ t] vs [p æ t]
  - ➢ Language family / other typological properties
  - ➢ Prosody (e.g. stress, length)