Predicting Continuous Vowel Spaces in the Wilderness

Emily Ahn and **David Mortensen**
Language Technologies Institute
Carnegie Mellon University
{eahn1,dmortens}@cs.cmu.edu

Abstract

We aim to model the acoustic vowel spaces of 24 diverse languages—a subset taken from the CMU Wilderness corpus of Bible recordings. With this model, we test hypotheses from phonological typology. We also expand upon previous work that used formant measurements taken by field linguists, and we use automatic tools to align and extract vowel segments in large-scale recorded speech. This work in progress is at the stage where data has been carefully processed, just prior to implementation of the model.

1 Introduction

Every language has a system of vowels, whether few or many, and understanding how these systems work crosslingually has been a goal in linguistic phonological typology. We are interested in empirically studying vowel spaces from the CMU Wilderness Multilingual Speech Dataset (Black, 2019), a large database of audio recordings from around 700 languages. Cotterell and Eisner (2018) developed a deep generative model to use acoustic formants to predict vowel spaces across a dataset of 223 languages. We aim to expand upon their findings and build a model from a set of languages that contain rich acoustic data per language and many vowel tokens per vowel type.

From this model, we can analyze our results to answer other questions regarding phonological typology. Dispersion Theory predicts that vowel types will be maximally dispersed within the vowel space, and Focalization Theory predicts that vowel types will preferentially be centered around canonical focii (Schwartz et al., 1997). Both theories make predictions about the distribution of centroids of types within formant space, but neither theory makes an explicit prediction about the dispersion of vowel tokens of a given type. In fact, both of these theories suggest that within-type dispersion should be relatively insensitive to other factors, since they treat vowel space as a system of categories. Other theories of vowel spaces, like those based on Evolutionary Phonology and exemplar theory, predict that the dispersion of tokens of each type should be inversely related to the number of contrasting types within a vowel system, since phonological categories can be seen as competing with one another for phonetic space (Vaux and Samuels, 2015). Reduced vowel inventories, in such a theory, are the result of mergers and merged categories take up more phonetic space than either of the categories prior to merger.

The Wilderness corpus provides a unique opportunity to test whether the number of vowel contrasts in a language's phonological inventory predicts the average dispersion of tokens in each vowel type. Rather than just providing idealized tokens of vowels from many languages or many tokens of vowels from few languages, it provides a massive number of tokens from a very large number of languages. If a relationship between number of types and token dispersion does exist on a large scale, it would be important evidence for evolutionary approaches to vowel space typology. If token dispersion is insensitive to the number of vowel types, support would be lent to the dispersion-focalization model.

These findings would be of interest to computational typologists and have implications for low-resource NLP and speech technologies. In addition to this narrow scientific question, this paper would contribute a replicable methodology for extracting vowels in a subset consisting of 24 languages from the public Wilderness corpus of 700 languages. This methdology could be used to extract vowel tokens from the corpus on a much larger scale.

2 Data

The Wilderness corpus comprises of roughly 700 languages of read speech from the New Testament of the Bible, originally scraped from Bible.is.[1]

2.1 Selection of 24 Languages

For this work, we intersected these languages with the PHOIBLE database of crosslingual phonological inventories (Moran and McCloy, 2019). We chose 24 languages where for each integer between 3 and 10, there are three languages (from distinct regions) whose vowel inventory size is that integer. We also used a criterion of choosing languages with the highest automatic alignment scores, as determined by algorithms provided in the Wilderness data, and this list is given in Table 1.

Language	Country	Vowels	Hours
Cebuano	Philippines	3	22
Kabyle	Algeria	3	8
Tena Quechua	Ecuador	3	19
Yupik	United States	4	22
Maranao	Philippines	4	24
Podoko	Cameroon	4	21
Russian	Russia	5	15
Twampa	Ethiopia	5	31
Urarina	Peru	5	31
Hanga	Ghana	6	14
Paumari	Brazil	6	48
Manado Malay	Indonesia	6	25
Komi	Russia	7	17
Sundanese	Indonesia	7	20
Tigrinya	Ethiopia	7	14
Denya	Cameroon	8	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

Table 1: The 24 languages chosen for this analysis, from the Wilderness data. They are sorted by number of vowel types as determined by PHOIBLE, and we attempted to balance the languages by region.

2.2 Preprocessing

We first obtained phone-level alignments via the tool provided from Festvox.[2] We then manually mapped the phoneme lists from the data with the IPA from PHOIBLE, since there is noise in the pronunciation model. Given vowel alignments, we extracted means of the first few formants using DeepFormants,[3] a tool for formant estimation. All preprocessing scripts will be made publicly available for future analyses on any language from the Wilderness.

A limitation of the data is that each language recording is spoken by few and undocumented speakers. We also anticipate challenges with regards to formant normalization across speaker gender, given that females tend to have higher and a greater range of formants than male speakers even when controlling for vocal tract.

3 Hypotheses

We hypothesize that vowel cloud size is inversely correlated with the number of vowels in a language's inventory. Cloud size will be measured as the level of dispersion in the probabilistic distribution of the two-dimensional vowel space.

4 Methodology

We plan to implement the deep generative models using determinantal point processes (DPP) from (Cotterell and Eisner, 2018) to analyze the formants in vowel spaces of our subset of the Wilderness data. We may choose a different method as well, in order to capture the variety of vowel tokens since the prior work used one representative vowel token per phoneme. We may use cross-entropy to evaluate our generative models, and will present our findings with regards to our hypotheses.

References

Alan W Black. 2019. CMU wilderness multilingual speech dataset. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975. IEEE.

Ryan Cotterell and Jason Eisner. 2018. A deep generative model of vowel formant typology. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 37–46.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Jean-Luc Schwartz, Louis-Jean Boë, Nathalie Vallée, and Christian Abry. 1997. The dispersion-

[1] http://www.bible.is/

[2] http://festvox.org/

[3] https://github.com/MLSpeech/DeepFormants

focalization theory of vowel systems. *Journal of phonetics*, 25(3):255–286.

Bert Vaux and Bridget Samuels. 2015. Explaining vowel systems: dispersion theory vs natural selection. *The Linguistic Review*, 32(3):573–599.