Machine Learning in Prosody and Meaning

Shirley Fong, Madeleine Guettler, Sofia Hirschmann
Simmons University, Boston, MA, 02115

Abstract

Prosody in spoken language is widely thought to convey meaning, but the complexity and variety of prosodic renditions have made modeling a specific prosody-meaning mapping elusive. In this work, 238 prosodically labeled utterances are used to classify speech as either exclamative or neutral. A first pass was made to cluster accent types (emphasized words) into three groups. This research project captures the importance of a variable number of loosely categorizable elements; clustering is performed; and the percentage of accent type is used as an attribute in the final random forest classification step.

Methods

The raw acoustics of pitch accents matter but are quite varied. In addition, there are more than one pitch accent per sentence and the ultimate goal is to classify the type of sentence. As a result, we seek to obtain a concise subset of pitch accent types that we can use in subsequent classification.

Process:

- ○ Created a data frame with just the numerical attributes
- ○ Scaled the data
- ○ First trial: did not remove outliers
- ○ Second trial: removed three outliers
- ○ Tested important features by dropping features one at a time, while isolating features we thought would be the best performing and least redundant
- ○ Calculated the sum of squares for each cluster and set of attributes
- ○ Obtained the centroid/means for the data points for the best clusters

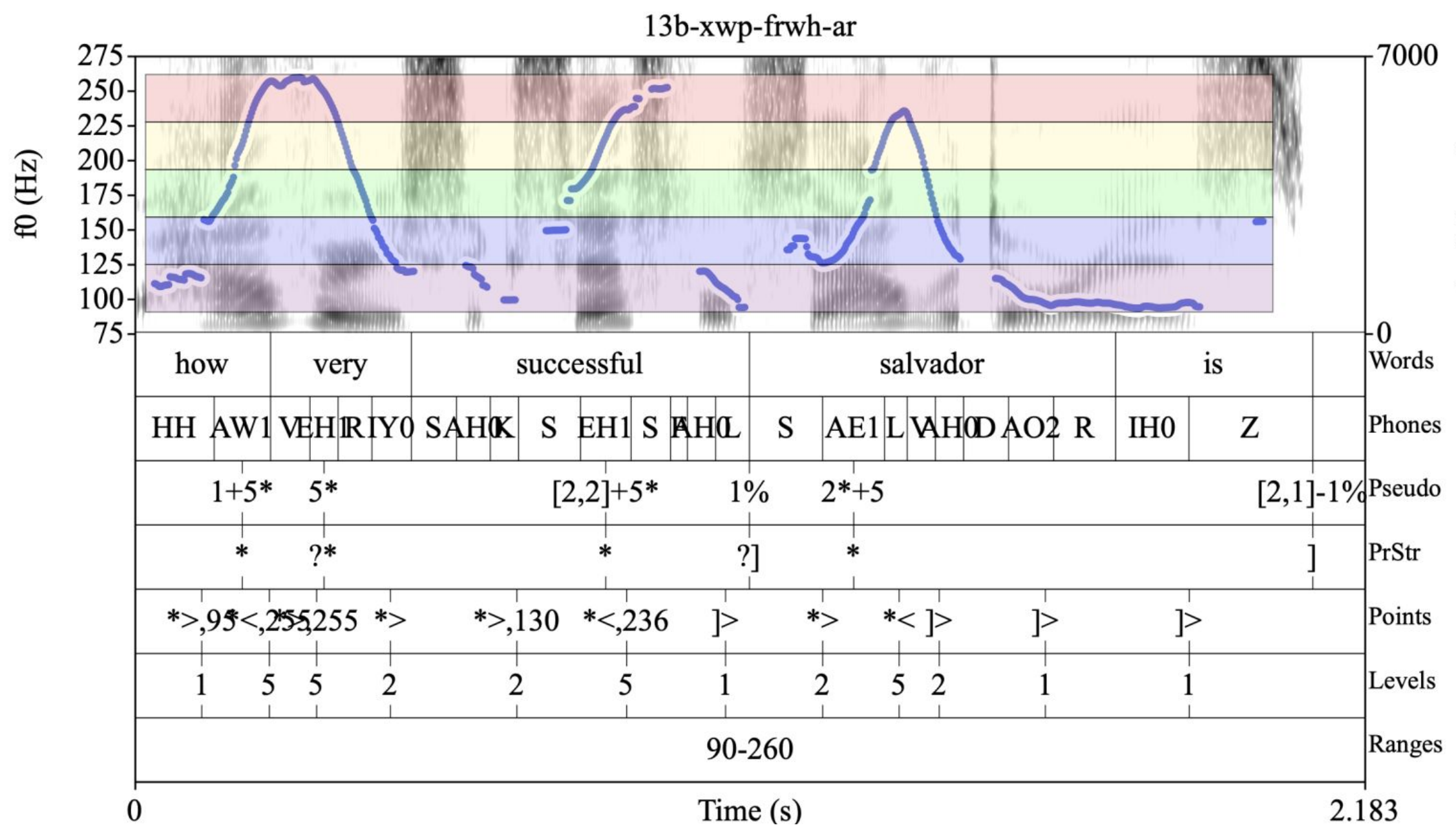
References

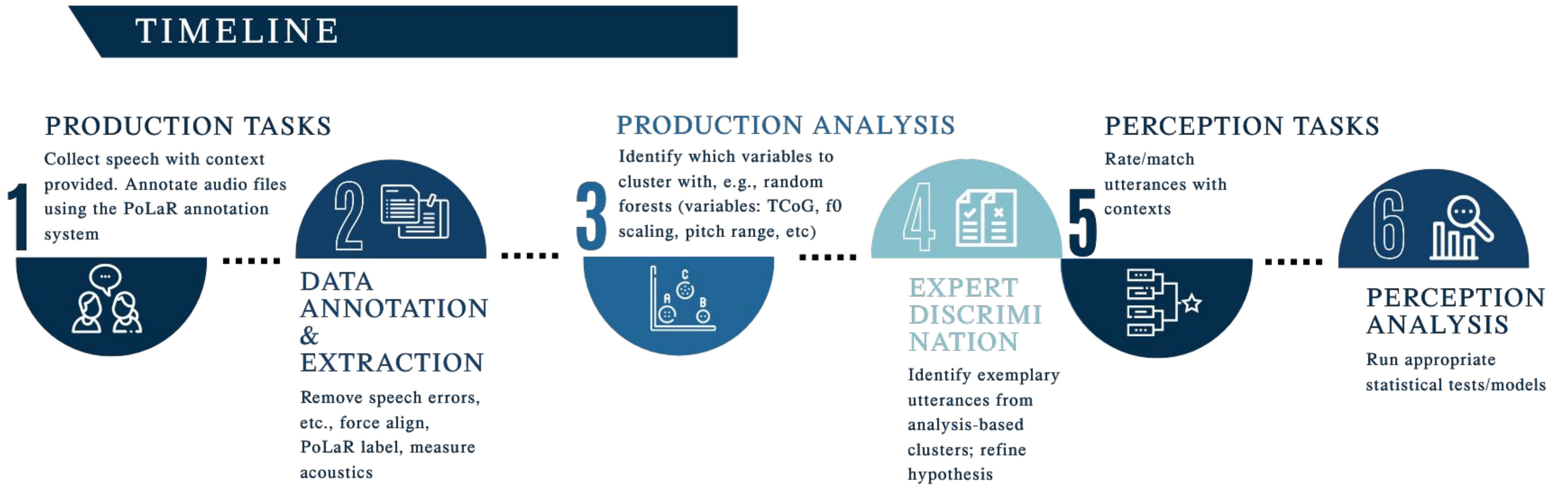
Selected References:

- Beckman & Hirschberg. 1994. The ToBI annotation conventions.
- Ahn et al. 2021. PoLaR Annotation Guidelines (version 1.0). Available at https://osf.io/usbx5.
- Rett & Sturman. 2021. Prosodically marked mirativity. In Proceedings of WCCFL 37.
- Barnes, Veilleux, Brugos, & Shattuck-Hufnagel. 2012. "Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology." *Laboratory Phonology* **3(2)**, pp. 337-383.

Acknowledgments: This material is based upon work supported by the National Science Foundation under Grant No. 2042694, 2042702, 204274. Additionally, we would like to thank Nanette Veilleux, Byron Ahn, Alejna Brugos, and Stefanie Shattuck-Hufnagel and Simmons University for guidance and support.

Background





TIMELINE

1 PRODUCTION TASKS
Collect speech with context provided. Annotate audio files using the PoLaR annotation system

2 DATA ANNOTATION & EXTRACTION
Remove speech errors, etc., force align, PoLaR label, measure acoustics

3 PRODUCTION ANALYSIS
Identify which variables to cluster with, e.g., random forests (variables: TCoG, f0 scaling, pitch range, etc)

4 EXPERT DISCRIMI NATION
Identify exemplary utterances from analysis-based clusters; refine hypothesis

5 PERCEPTION TASKS
Rate/match utterances with contexts

6 PERCEPTION ANALYSIS
Run appropriate statistical tests/models

What is prosody?

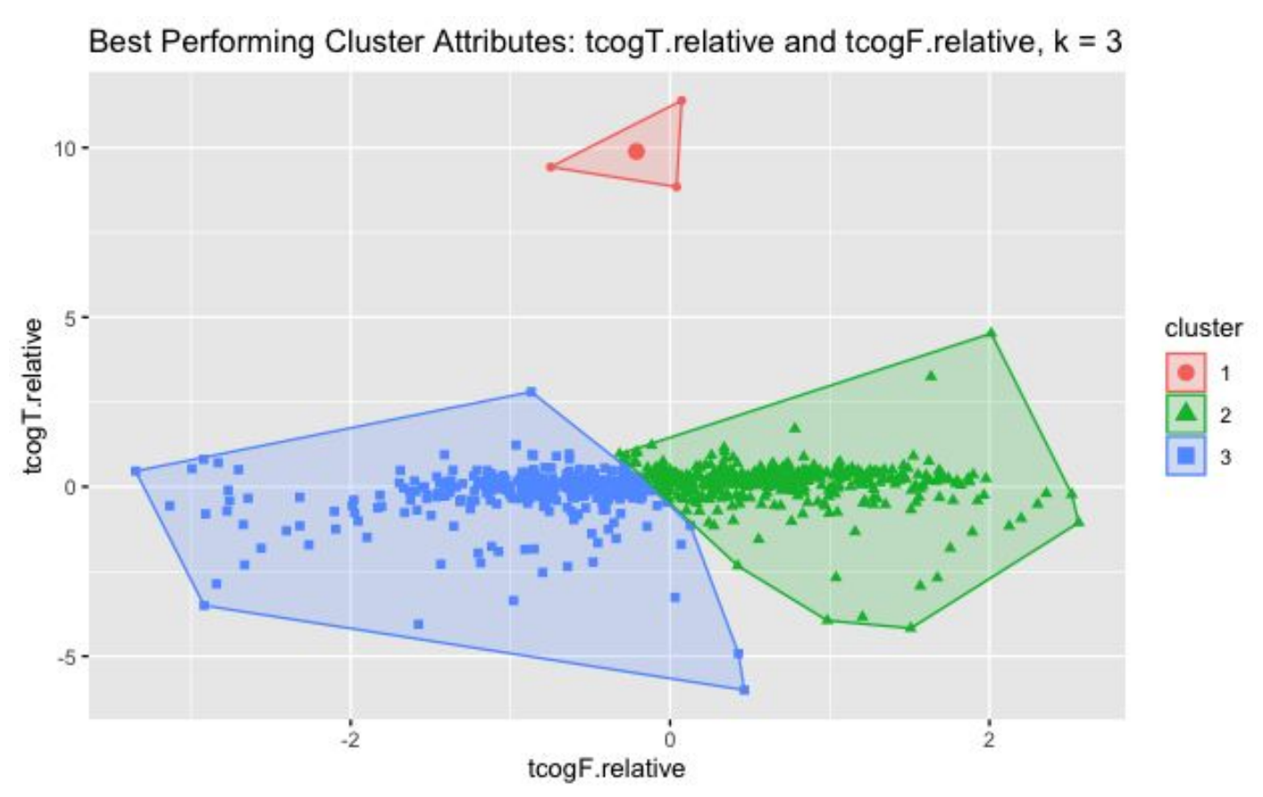
- In spoken language it's thought to convey meaning
- It's not what you say, but *how* you say it through alterations in pitch, duration, and intensity
- Prosody maps to meaning, and here our meaning is what we call mirativity: the idea of being surprised

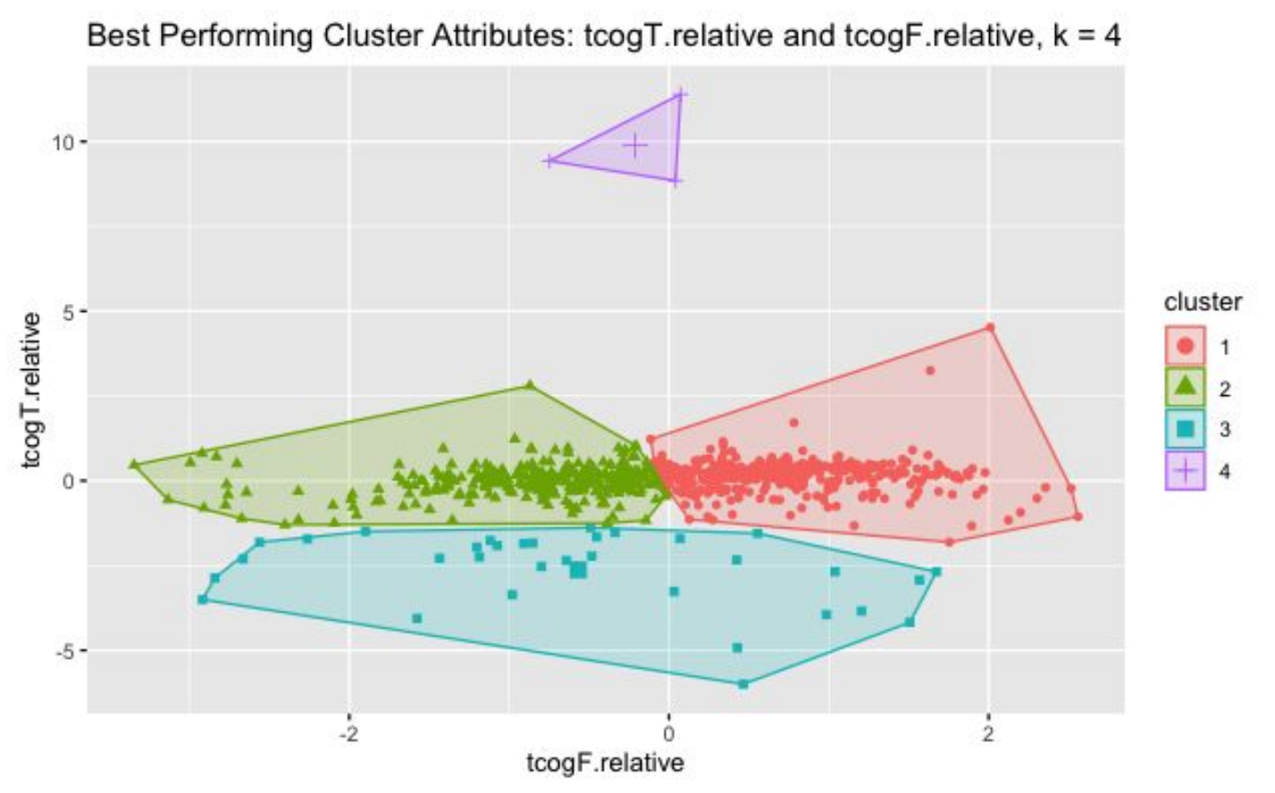
Question

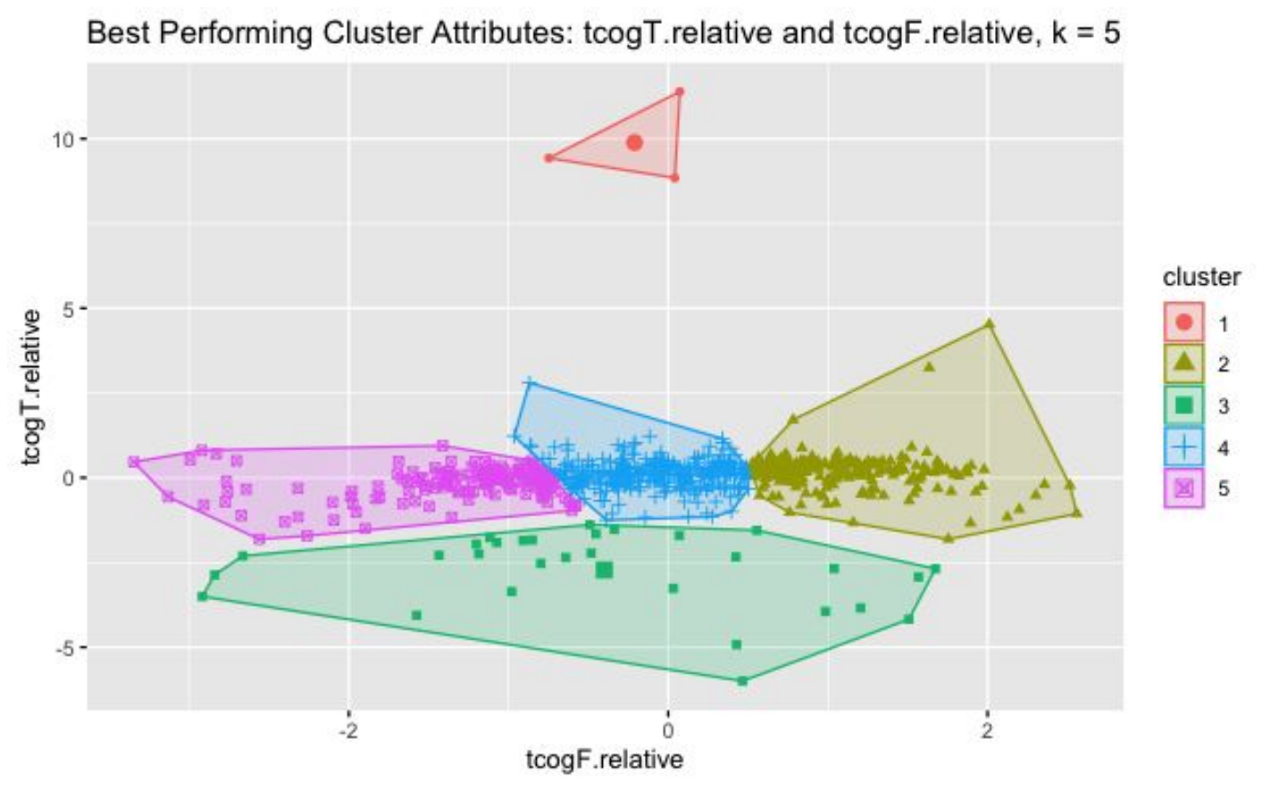
- What are the best attributes and optimal amount of clusters for our k-means model to predict mirativity?

Results

With Outliers


Best Performing Cluster Attributes: tcogT.relative and tcogF.relative, k = 3


Best Performing Cluster Attributes: tcogT.relative and tcogF.relative, k = 4


Best Performing Cluster Attributes: tcogT.relative and tcogF.relative, k = 5

With Outliers

Features	3 Clusters Sum of Squares	4 Clusters Sum of Squares	5 Clusters Sum of Squares
tcogF.relative, tcogT.relative	Clustering SS: 686.3996 BW/TOT: 0.5292184	Clustering SS: 494.671 BW/TOT: 0.6607195	Clustering SS: 380.9342 BW/TOT: 0.7387283
Rise.time, Rise.change.Lvl, Rise.slope.Lvl	Clustering SS: 927.7017 BW/TOT: 0.5758108	Clustering SS: 697.1438 BW/TOT: 0.6812328	Clustering SS: 577.348 BW/TOT: 0.7360092

Features	Centroids	Centroids	Centroids
tcogF.relative, tcogT.relative	tcogF.relative tcogT.relative 1 -0.9098167 -0.2367018 2 0.6893877 0.1073036 3 -0.2112869 9.8879987	tcogF.relative tcogT.relative 1 -0.5676015 -2.6264091 2 -0.2112869 9.8879987 3 -0.8341122 -0.0135954 4 0.7319573 0.1603384	tcogF.relative tcogT.relative 1 -0.40032233 -2.7216077 2 -1.31226824 -0.1630675 3 -0.07574489 0.1191110 4 -0.21128688 9.8879987 5 1.09379953 0.1798647
Rise.time, Rise.change.Lvl, Rise.slope.Lvl	Rise.time Rise.change.Lvl Rise.slope.Lvl 1 0.8926504 0.55897411 -0.2221773 2 -0.6765719 -0.02309292 0.4924669 3 0.2740121 -2.48985343 -2.2213488	Rise.time Rise.change.Lvl Rise.slope.Lvl 1 1.0157173 0.6105793 -0.2330180 2 -0.5494376 -0.7703226 -0.4412892 3 -0.5887689 0.3515331 0.8475355 4 0.3344775 -3.2705839 -2.8836897	Rise.time Rise.change.Lvl Rise.slope.Lvl 1 0.1789796 -3.2455686 -2.9128409 2 -0.6220142 0.3398741 0.8800706 3 0.8709192 0.6437083 -0.1844767 4 -0.5551231 -0.7727897 -0.4401627 5 4.4921169 -1.1768068 -1.3253212

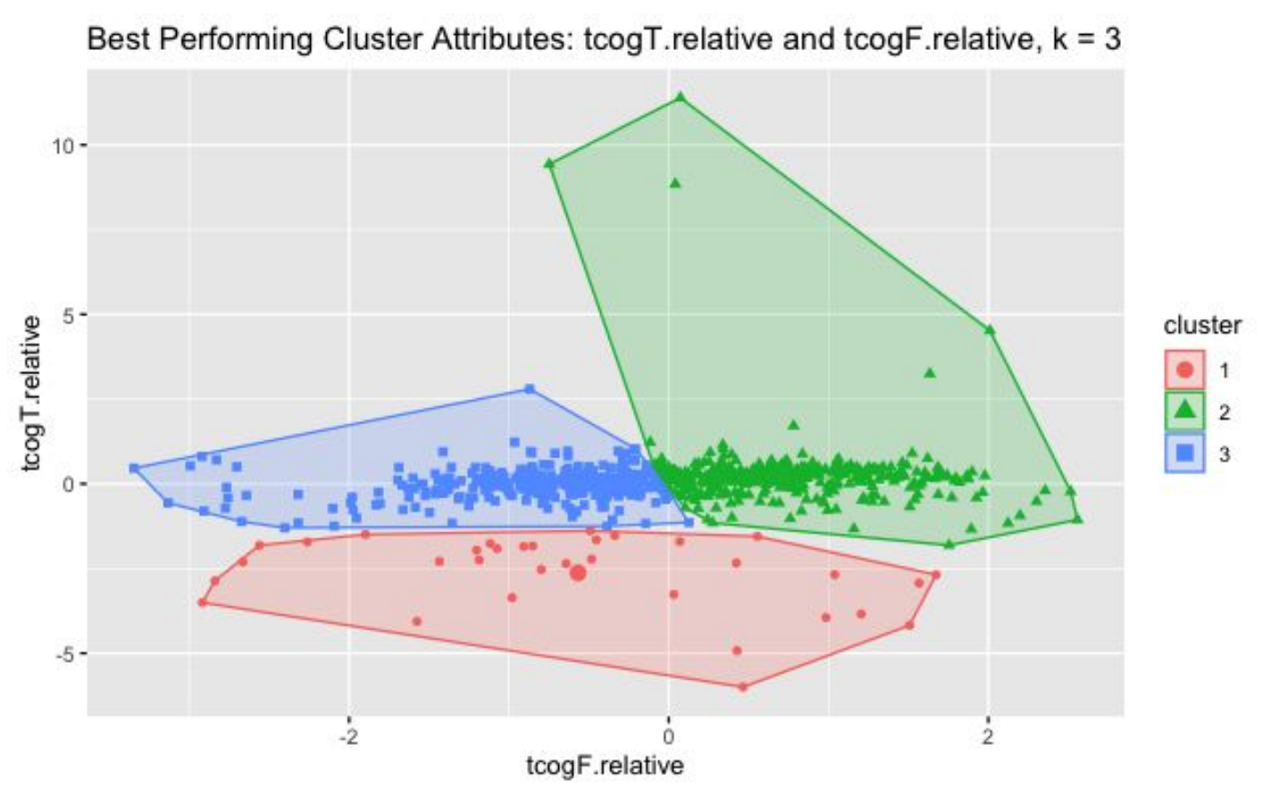
Without Outliers

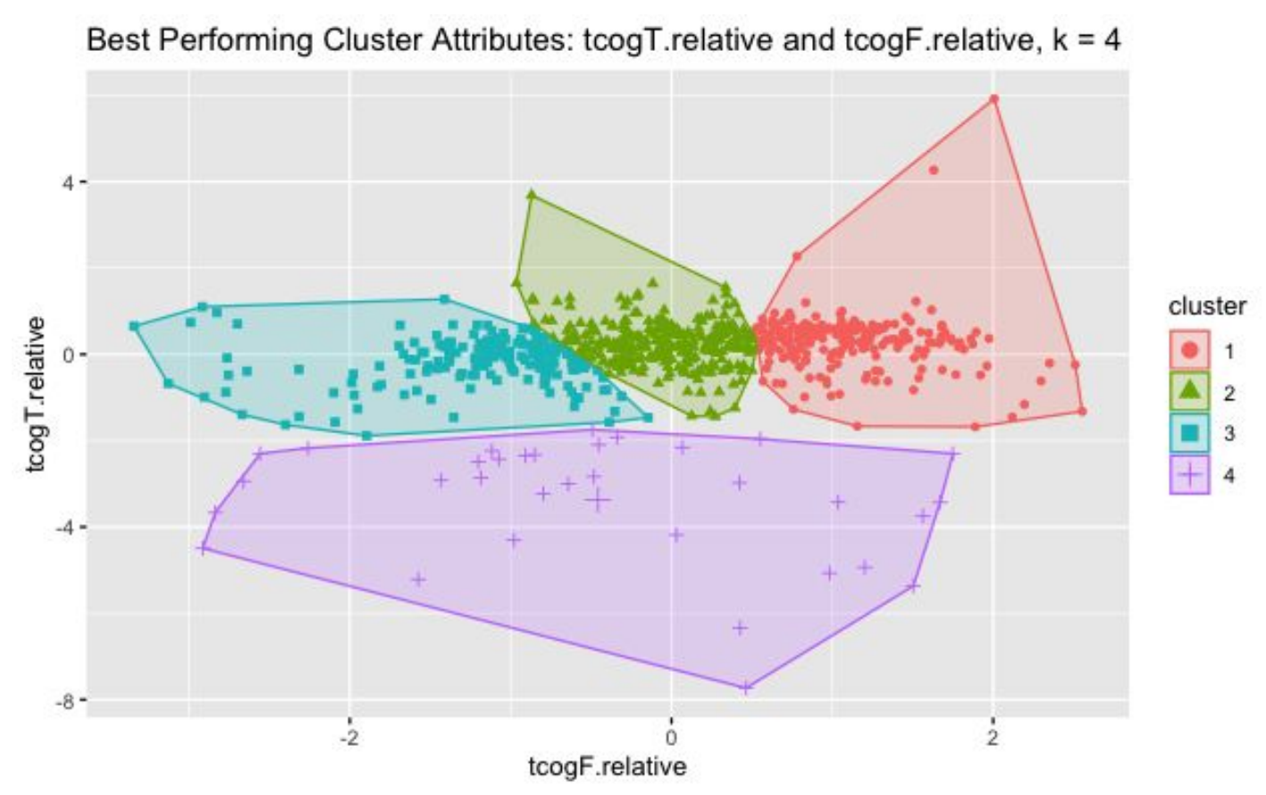
Features	3 Clusters Sum of Squares	4 Clusters Sum of Squares	5 Clusters Sum of Squares
tcogF.relative, tcogT.relative	Clustering SS: 620.9624 BW/TOT: 0.5723399	Clustering SS: 505.8289 BW/TOT: 0.651633	Clustering SS: 445.0025 BW/TOT: 0.6935245
Rise.time, Rise.change.Lvl, Rise.slope.Lvl	Clustering SS: 925.0094 BW/TOT: 0.5752941	Clustering SS: 694.3222 BW/TOT: 0.6812111	Clustering sum of squares: 588.7712 BW/TOT: 0.7296735

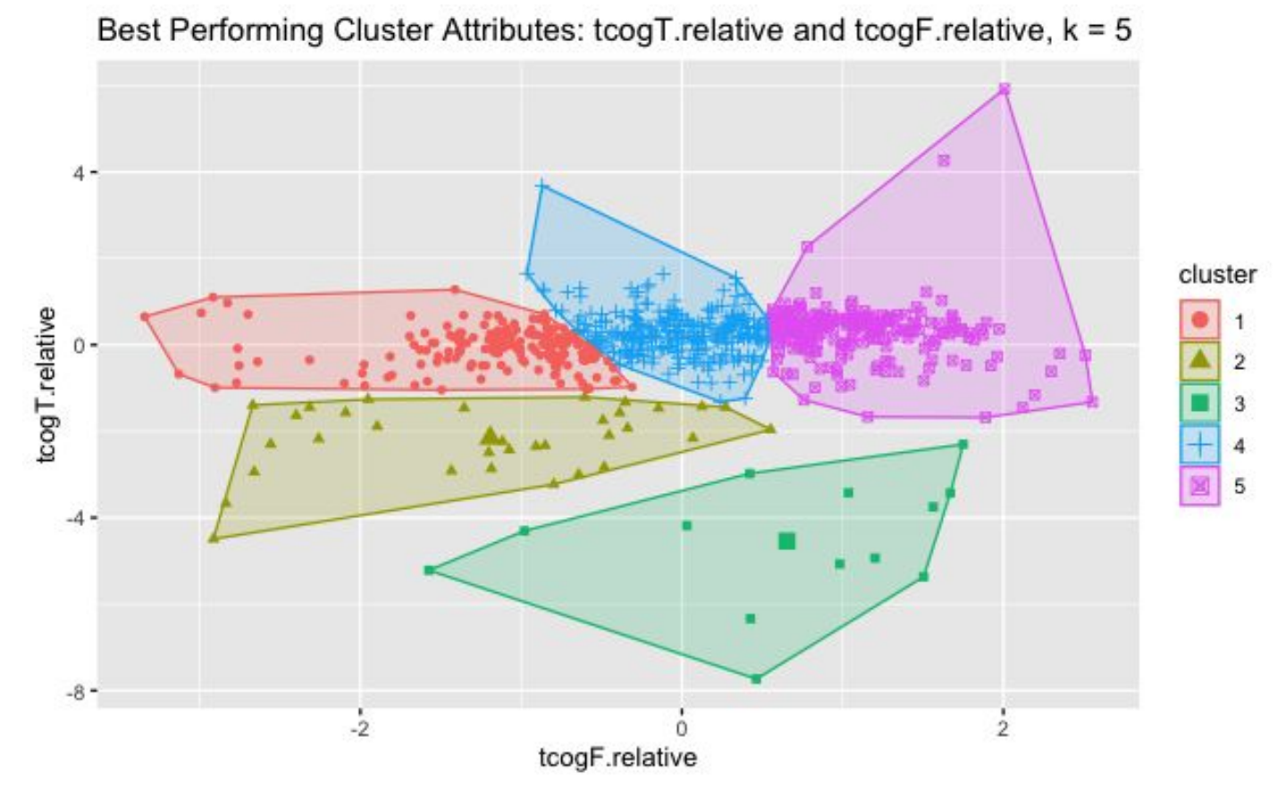
Features	Centroids	Centroids	Centroids
tcogF.relative, tcogT.relative	tcogF.relative tcogT.relative 1 0.7078622 0.287413028 2 -0.8499363 0.004108605 3 -0.4993463 -3.325078930	tcogF.relative tcogT.relative 1 1.09328493 0.2983102 2 -0.03672014 0.2636620 3 -1.19247052 -0.1776622 4 -0.45703841 -3.3686564	tcogF.relative tcogT.relative 1 1.10106395 0.2965778 2 0.65357473 -4.5408500 3 -1.19225560 -2.1583629 4 -0.01965895 0.2787908 5 -1.15657988 -0.0838336
Rise.time, Rise.change.Lvl, Rise.slope.Lvl	Rise.time Rise.change.Lvl Rise.slope.Lvl 1 0.8946883 0.55884894 -0.2233000 2 -0.6754238 -0.02096228 0.4926861 3 0.2756277 -2.48482888 -2.2167150	Rise.time Rise.change.Lvl Rise.slope.Lvl 1 -0.5881996 0.3543686 0.8488617 2 -0.5467097 -0.7667459 -0.4403205 3 0.3360114 -3.2649019 -2.8776923 4 1.0187766 0.6104737 -0.2344369	Rise.time Rise.change.Lvl Rise.slope.Lvl 1 0.8825757 0.6430770 -0.1913730 2 0.3124053 -3.0066286 -2.7021378 3 4.4880347 -1.1728878 -1.3225323 4 -0.6222473 -0.7304906 -0.3644070 5 -0.5976363 0.3786118 0.8916242

- Linguistics have long postulated several categorically different pitch accents and describe them in terms of the F0 (High or Low) and alignment (* placed early, medially or late with respect to the accented syllable)
- The Tonal Center of Gravity is a global measurement that describes alignment in time (TCoG-T) and F0)TCoG-F)
- In the k=3, there appears to be a late and low accent that would correspond to an L*+H and two medially aligned accents: one relatively low in pitch (L+H*) and the other one relatively high (H*)
- As we increase the clusters, the clusters seem to separate variants of these three well known accent types

Without Outliers


Best Performing Cluster Attributes: tcogT.relative and tcogF.relative, k = 3


Best Performing Cluster Attributes: tcogT.relative and tcogF.relative, k = 4


Best Performing Cluster Attributes: tcogT.relative and tcogF.relative, k = 5