Comparing k-means and OPTICS clustering algorithms for identifying vowel categories

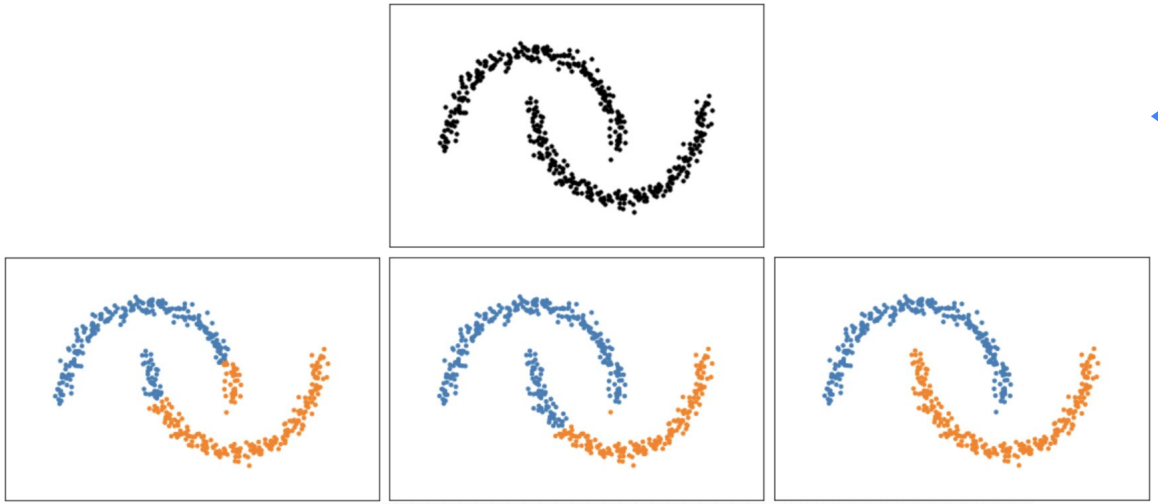
Emily Grabowski and Jennifer Kuo
UC Berkeley and UCLA

Background: clustering algorithms

- Clustering techniques group data into **clusters** without any human input, based on their **similarities/differences**
  - **“Similarity”** can be based on different metrics, depending on the clustering algorithm.
- Possible methods:
  - **Centroid-based** clustering (Forgy 1965); most commonly used
    - K-means (simplest, easiest to implement)
    - Gaussian Mixture Models
  - **density-based** clustering
  - **hierarchical** clustering

Background: clustering algorithms

Different clustering algorithms have different assumptions about how clusters are shaped, and will assign different results to the same data.



Illustrative example: Input data vs. clusters assigned by three different algorithms

Background: clustering algorithms

- In descriptive phonetics, centroid-based clustering is commonly used (e.g. De Boer & Kuhl 2003; Czoska et al 2015; Shi et al 2019).

Question: Is OPTICS (or more generally, density-based clustering) a viable alternative to k-means for learning patterns in vowel data?

- Method: compare OPTICS and k-means on two datasets:
  - Hillenbrand et al. (1994): data in a controlled lab setting
  - Buckeye corpus of conversational speech (Pitt et al. 2005): naturalistic, noisier data

K-means

How it (generally) works (Forgy 1965):

1. Randomly chooses *k* center points
2. Assign each point to the nearest center point.
3. Update center points to mean of points in each cluster.
4. Repeat.

Advantages:

- Relatively simple to implement.
- Scales to large data sets
- one parameter (k)

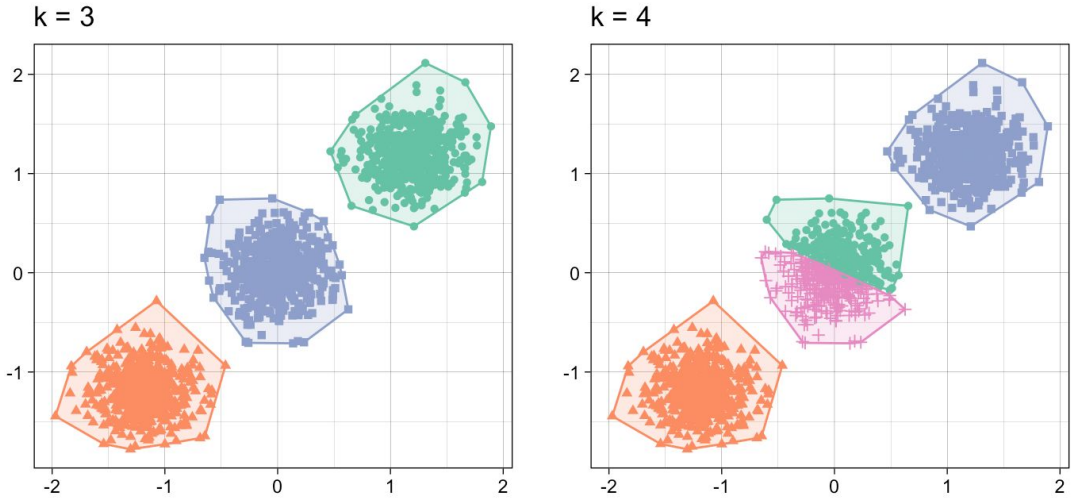
Disadvantage:

- Requires specifying number of clusters (k)
- Assumptions about how clusters are shaped

K-means

Disadvantage: requires specifying the number of clusters

- Example: same dataset, with k=3 vs. k=4

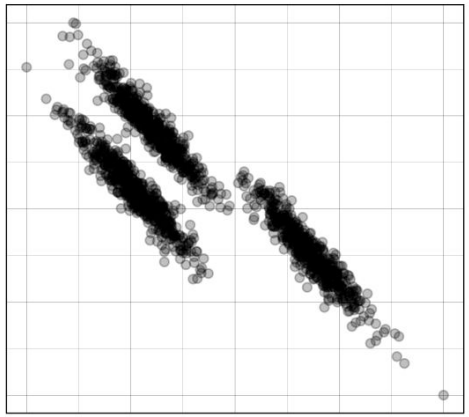


K-means

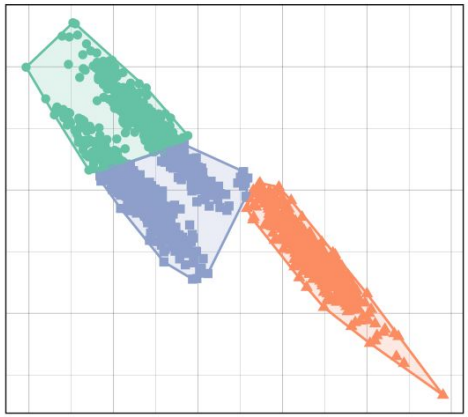
Disadvantage: assumes roughly **circular clusters** of the same size & variance

- Example: non-circular clusters

raw data



k = 3



K-means

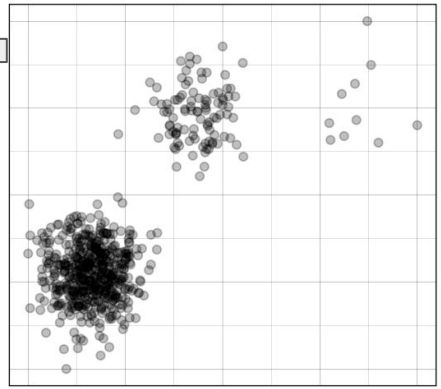
Disadvantage: assumes roughly circular clusters of…

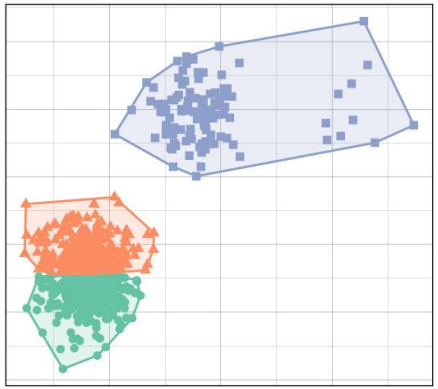
- **same size (n points)**
- **same variance**

raw data (uneven clusters)

k = 3

Different sizes

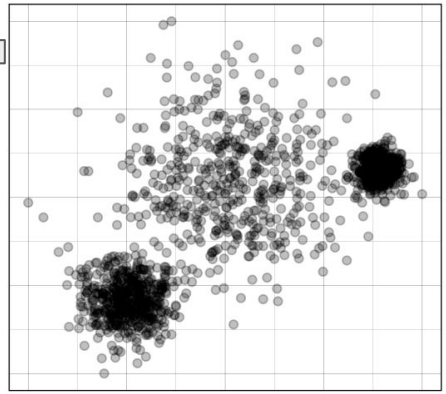


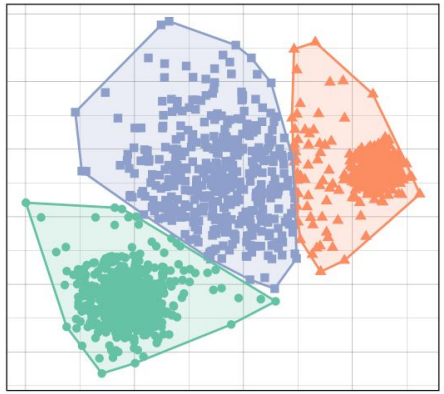


raw data (varied clusters)

k = 3

Different variance





OPTICS (Ordering Points to Identify the Clustering Structure)

What it does: identifies “dense” clusters of points.

How it works: Ankerst et al (1999)

1. Iterate through each point
2. Identify nearest neighbors and record index in list
3. Check if the point is noise
4. Repeat for remaining points

Advantages

- Adaptable to different geometries (cluster “shapes”)
- No set number of clusters

Disadvantages

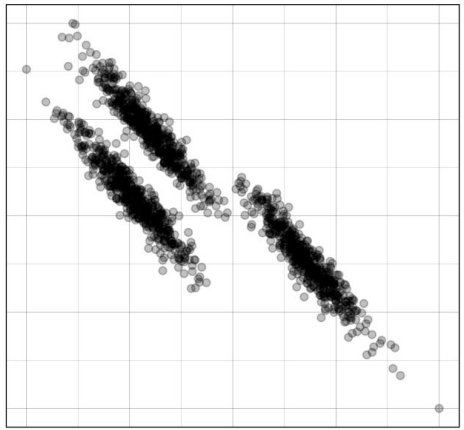
- Several parameters to tune
- Slower
- Requires variation in density to delineate clusters

OPTICS

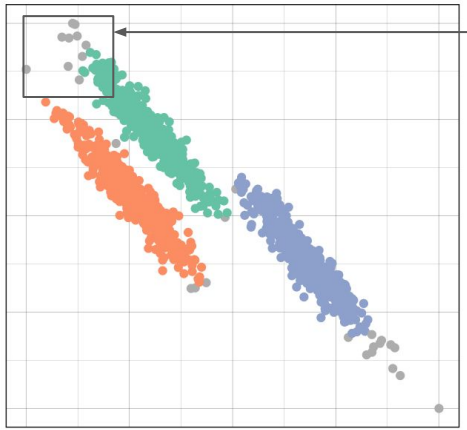
Advantage: accounts for clusters of varying **shape**, size, and density

- Example: non-circular clusters

non-circular clusters



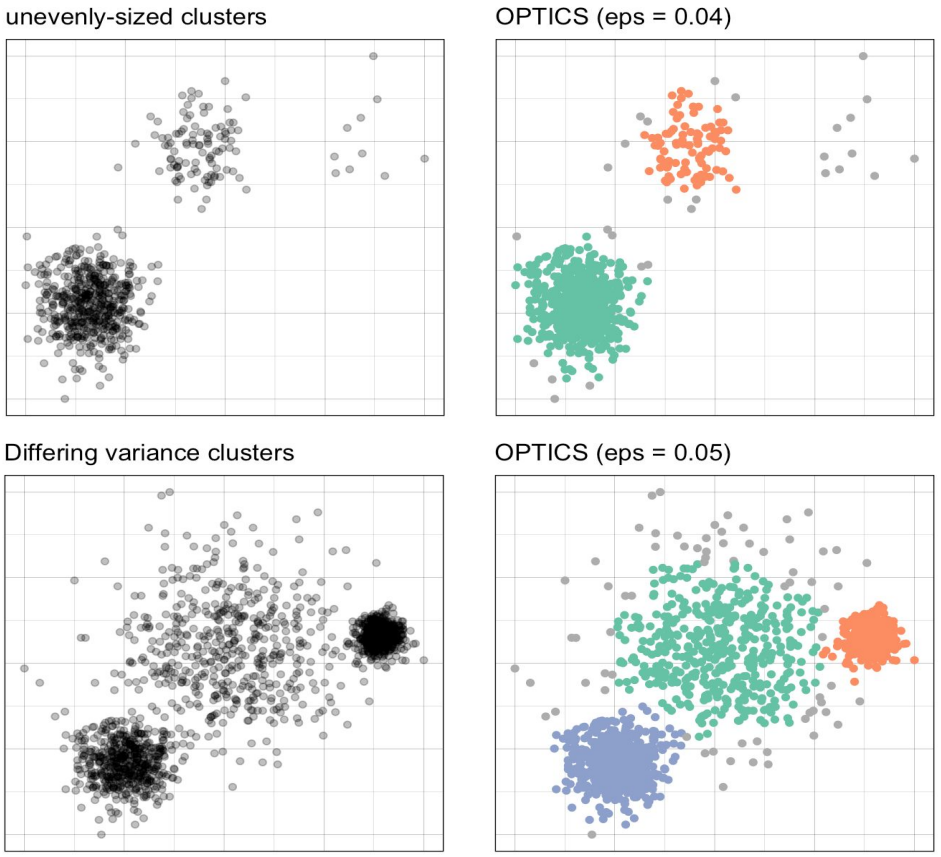
OPTICS (eps = 0.035)



Note: OPTICS separates **noise** from non-noise. In the following figures, noise are colored in grey.

OPTICS

Advantage: accounts for clusters of varying shape, **size (n points), and density**



Summary: k-means vs. OPTICS

K-means	**OPTICS**
Assumes circular clusters of equal shape/variance	Does not assume size/shape/variance of cluster (accounts for different geometries)
Classifies all points	Excludes some points as noise
Only things to optimize is the number of clusters.	Number of clusters is not specified, and other parameters must be optimized.

More on this later!

Why this might be relevant for vowel spaces

- Are vowels:
  - Circular?
  - Same size?
  - Same variance?
  - Overlap?
- Example: cardinal vowels in naturalistic speech (Buckeye corpus)
  - /i/ <iy> has lower variance than other vowels
  - Clusters of differing sizes
  - Significant overlap

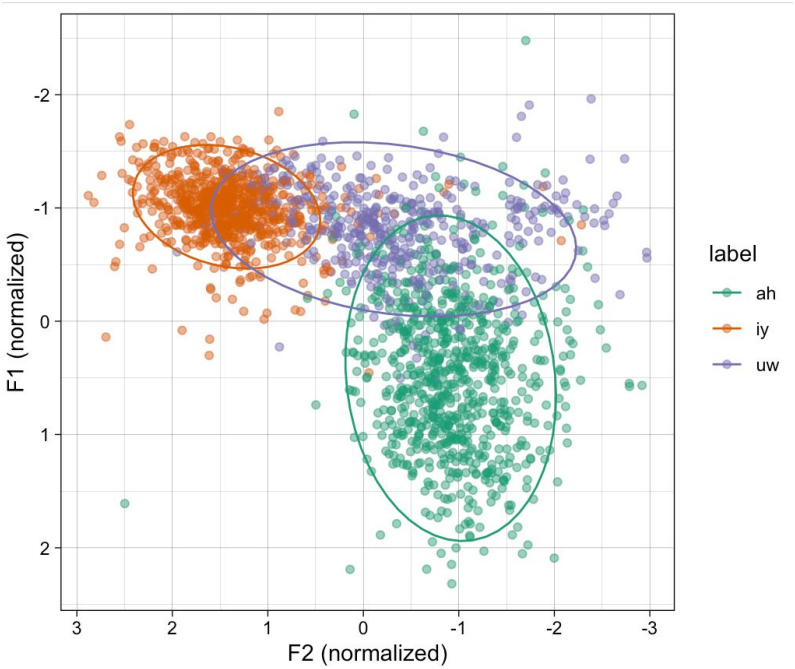
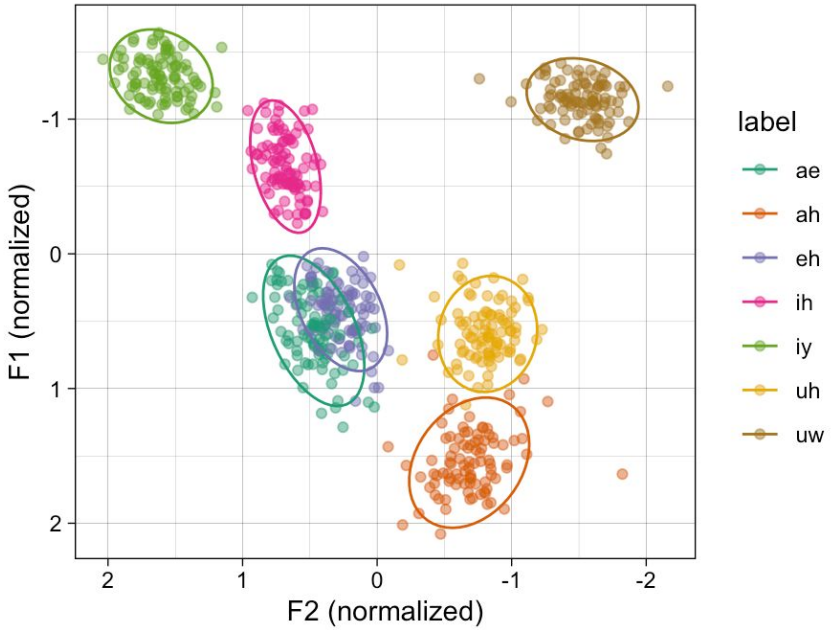


Illustration - Hillenbrand et al. (1994)

- 93 speaker, 14 tokens per speaker (=1302 points)
- Monophthongs (/i ɪ ɛ æ a ʊ u/)
- Recorded in the hVd context
- Measures:
  - F1 and F2 midpoints
  - Normalized by speakers (z-score)
  - Measures were chosen for interpretability/visualization/exploration



Optimization

K-means

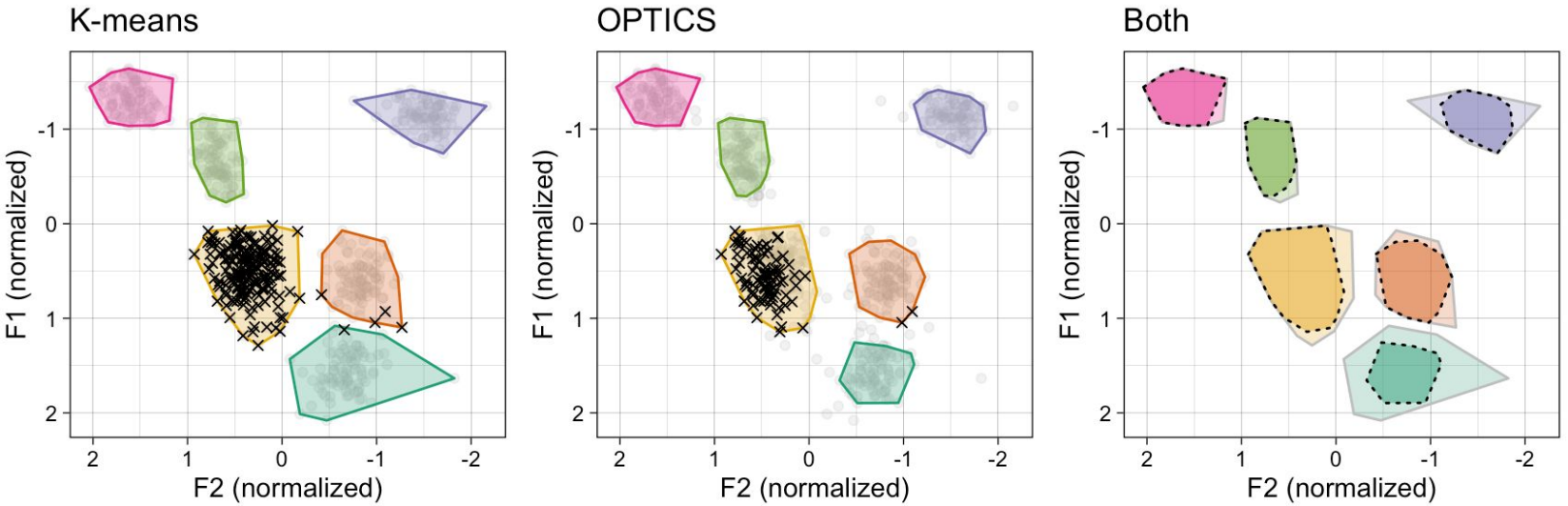
- Optimize number of clusters
- Choose a range of number of clusters to test.
- “Good”: **Fewest** number of clusters, where **clusters are far apart**, and within a cluster **points are close to each other**.
- Two metrics of between/within cluster distance (silhouette and inertia)

OPTICS

- Points ordered into groups of nearby points
- Choose a cutoff that determines how far apart a two points in the same cluster can be
  - Above the cutoff = noise
  - Below the cutoff = clusters
- Number of clusters are determined by the algorithm rather than a parameter

K-means vs OPTICS clustering on Hillenbrand

Note: adding duration improves classification of /ɛ/ vs. /æ/ (Appendix)

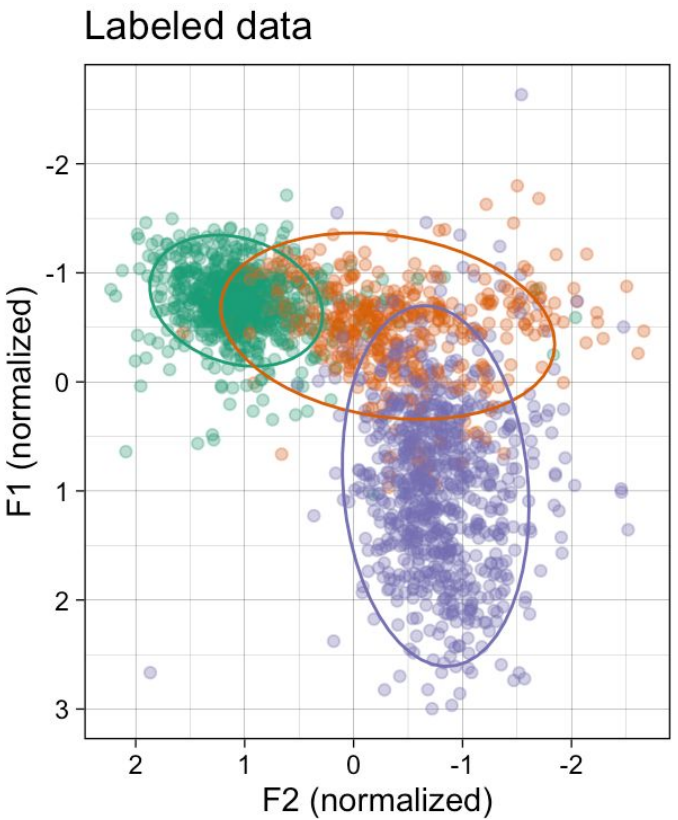


X = mislabeled points

Corpus - Data

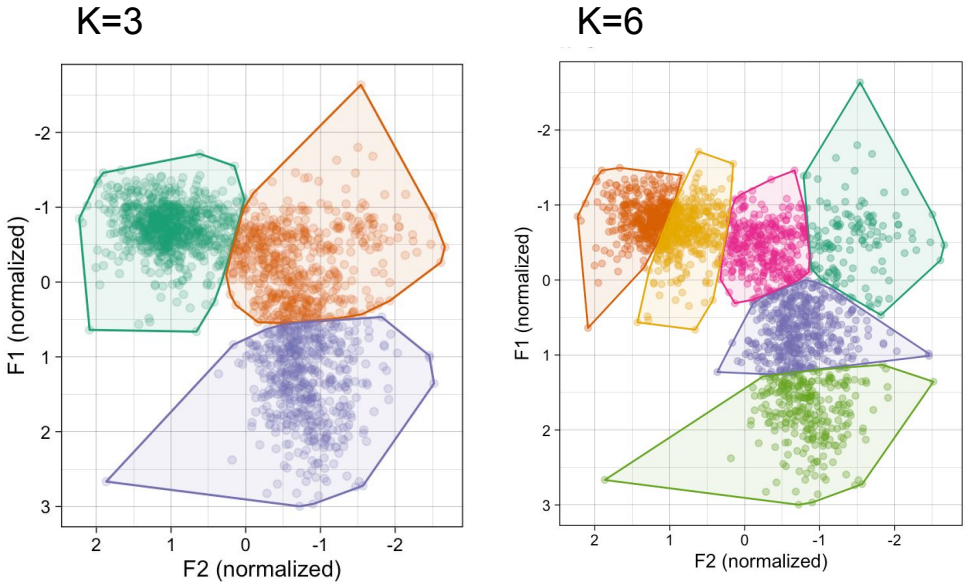
Buckeye corpus of conversational speech (Pitt et al. 2005)

- Data from naturalistic face-to-face interviews.
- Subset of 7 speakers, randomly selected
- For this analysis, included only the vowels /i/, /a/, and /u/
- Used same parameters and preprocessing as Hillenbrand data.



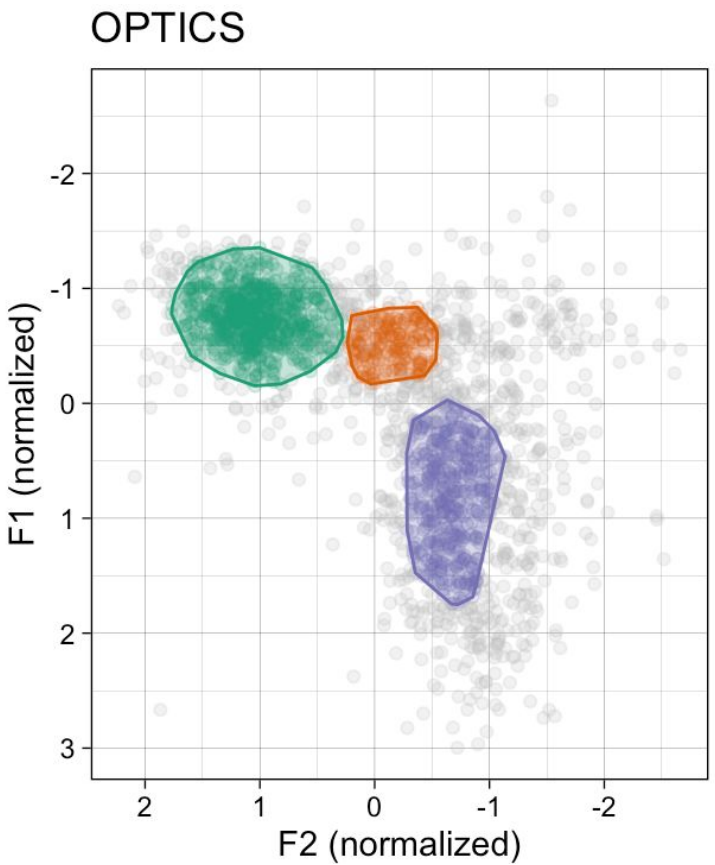
K-means

- Optimizing identifies 6 clusters, while *a priori* we might expect 3 clusters
- Divides up the space into roughly equal areas
- Includes all of the data in a cluster

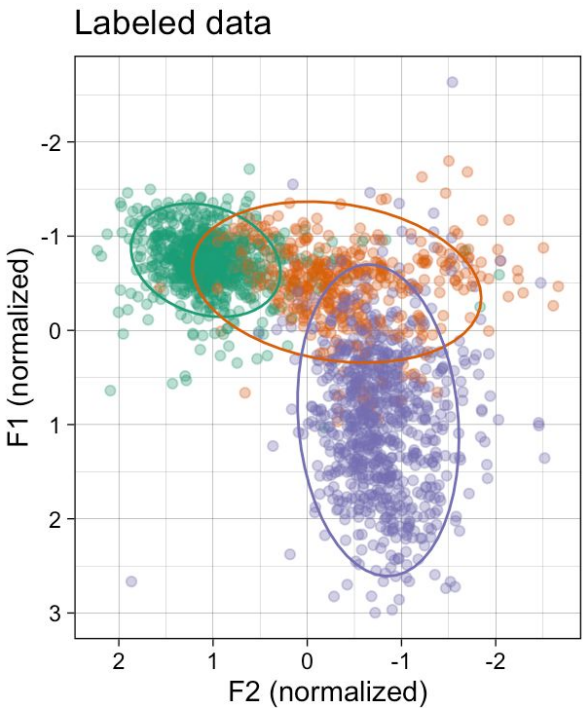


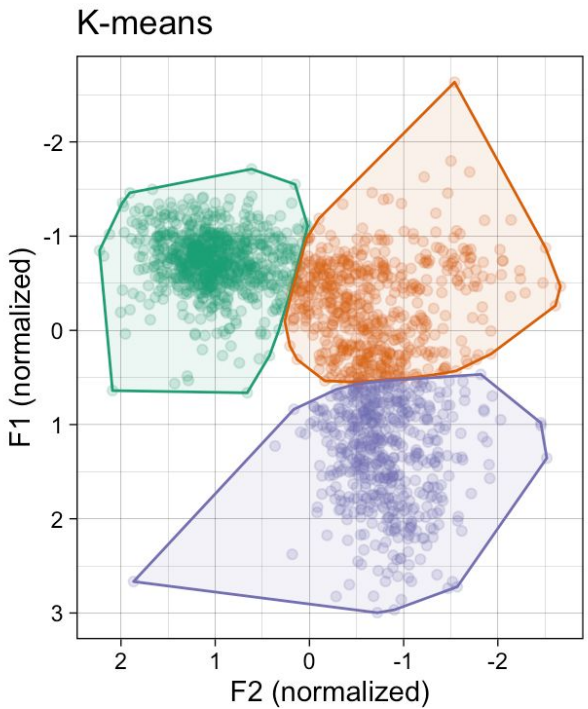
OPTICS

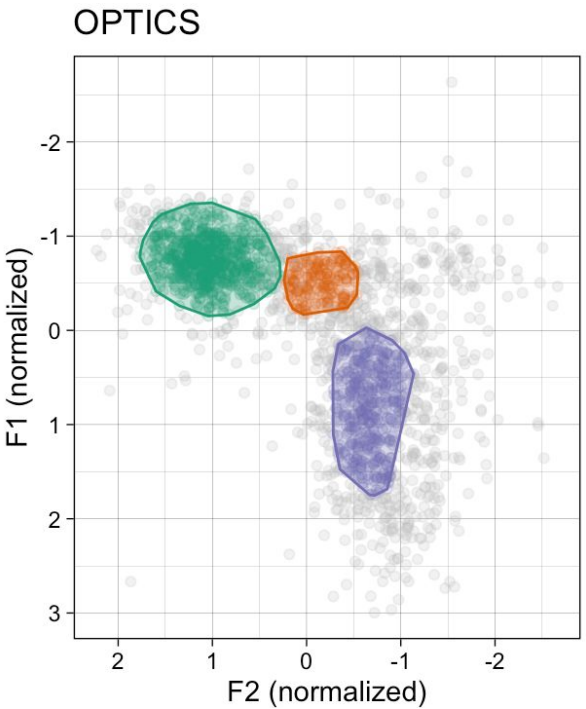
- Identifies three clusters of different sizes / shapes
- Captures the relative frontess of /u/ in this data
- Significant number of tokens excluded as noise - not optimal for prediction



Comparing Clustering Results







Conclusion

- There are advantages and disadvantages to each clustering method
  - K-means focuses on dispersion rather than density
  - OPTICS captures density but excludes noise
  - Both methods struggle with overlapping clusters
- Assumptions about vowel distribution are often based on controlled lab speech, and it is important to look at whether these assumptions hold in naturalistic speech.
- While clusters are not intended to replace other methods of analysis, it can be a useful tool for identifying patterns in data
- Future work:
  - higher dimensional input, including dynamic measures
  - more vowel categories (e.g. diphthongs, subphonemic variation)

Slides, code, and more:

https://github.com/jenniferxkuo/vowel-clustering

References

Ankerst, M., Breunig, M., Kriegel, H-P., and Sander, J. (1999). OPTICS: ordering points to identify the clustering structure. *ACM SIGMOD*, 28(2) (1999): 49-60.

Czoska, A., Katarzyna K., & Karpinski, M. (2015). Polish infant directed vs. adult directed speech: Selected acoustic-phonetic differences. *18th International Congress of Phonetic Sciences,* Glasgow, UK.

De Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, 4(4).

Forgy, E W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*. 21 (3).

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, 97(5).

Miyazawa, K., Kikuchi, H., & Mazuka, R. (2010). Unsupervised learning of vowels from continuous speech based on self-organized phoneme acquisition model. In Interspeech 2010, pp. 2914-2917

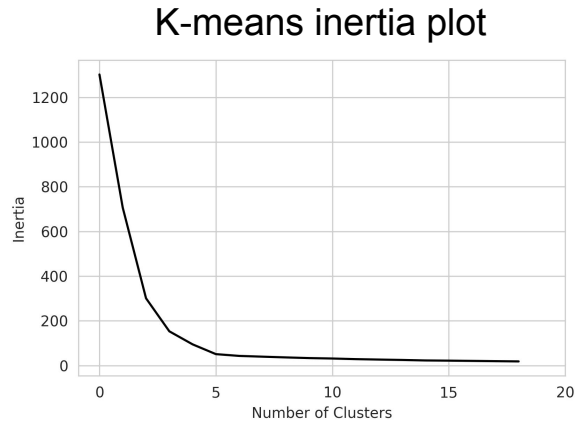
Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1).

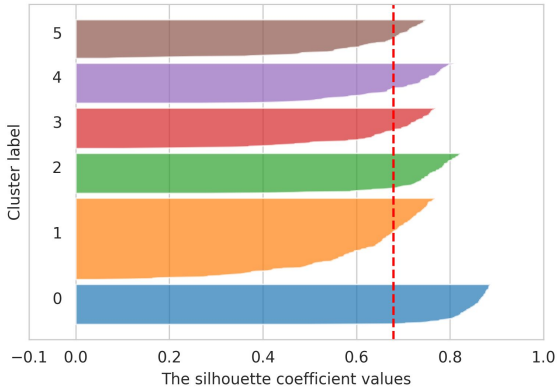
Shi, Y., Renwick M., & Maier, F. (2019). Improved vowel labeling for prenasal merger using customized forced alignment. Poster presented at the *178th Meeting of the Acoustical Society of America*, San Diego, CA.

Optimization: K-means

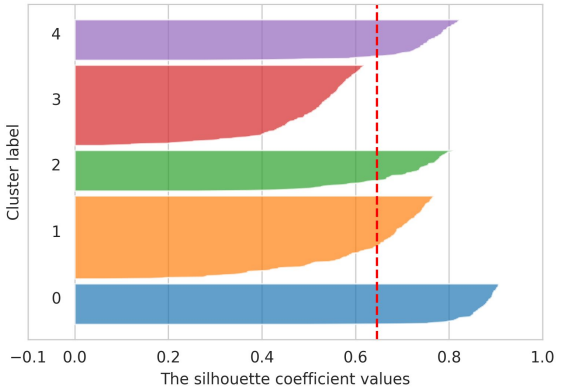
1. Elbow of inertia plot:
   a. sum square distance between points and centroids
   b. Optimal is just after the elbow (in this case around 5-6 clusters)
2. Silhouette plot
   a. another measure of within/between cluster distance
   b. More uniform silhouettes (left) preferred

Based on these, 6 clusters seems optimal





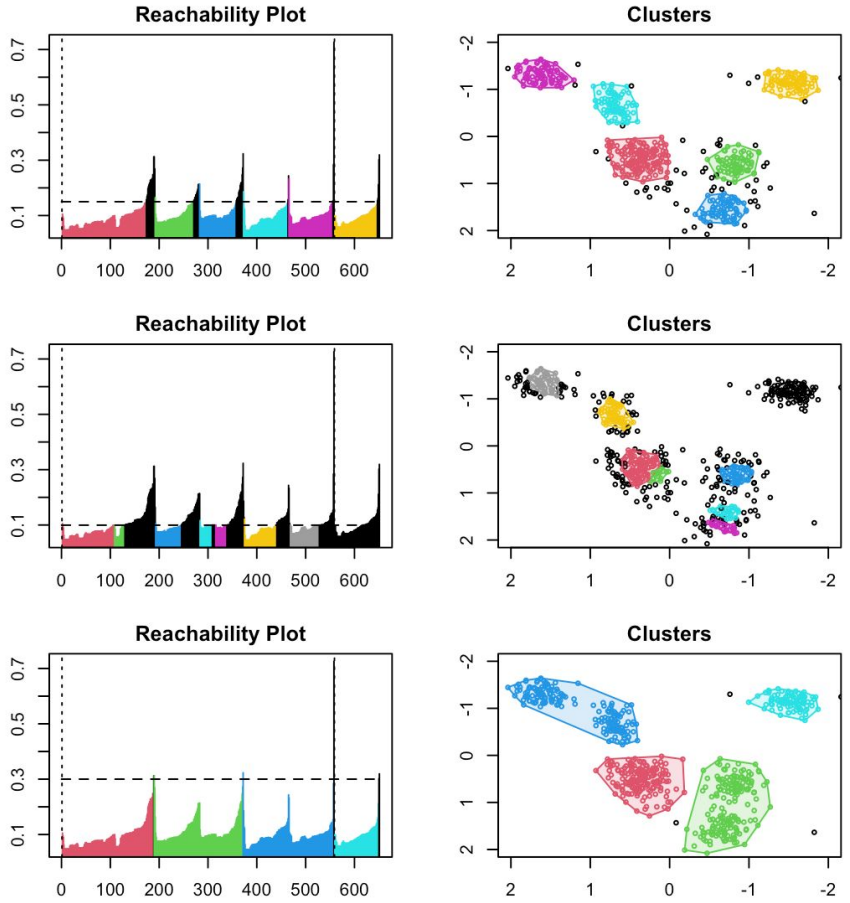
6 clusters



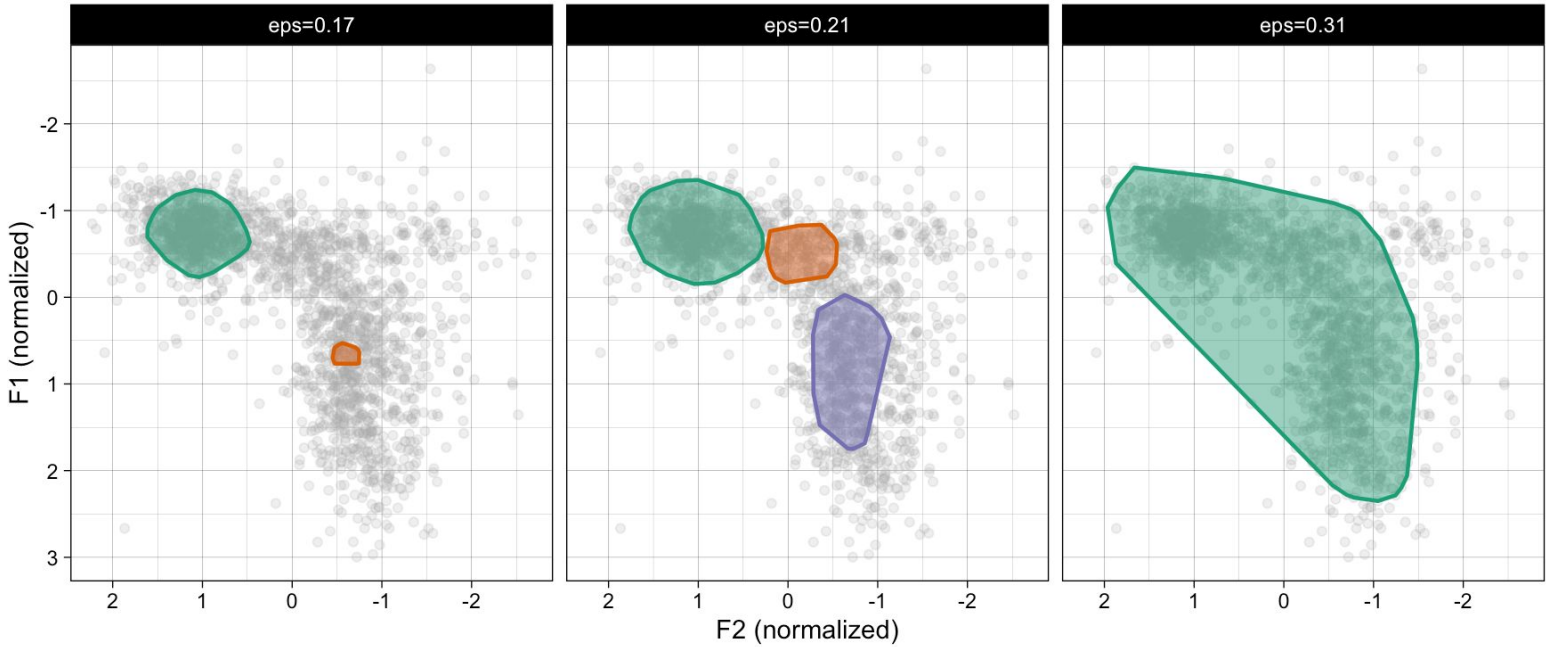
5 clusters

Tuning for OPTICS

- Reachability: a measure of distance between points.
- Valleys in reachability plot indicate clusters, while peaks indicate far away points (often noise)
- Cutoff decides what points are excluded as noise.
- Optimal point is just below the lowest peak in the reachability plot



OPTICS predictions at different cutoffs for Buckeye cardinal vowels



Hillenbrand data (with F1, F2, duration)