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

Emily Grabowski & Jennifer Kuo*

Abstract. The K-means algorithm is the most commonly used clustering method for phonetic vowel description but has some properties that may be sub-optimal for representing phonetic data. This study compares K-means with an alternative algorithm, OPTICS, in two speech styles (lab vs. conversational) in English to test whether OPTICS is a viable alternative to K-means for characterizing vowel spaces. We find that with noisier data, OPTICS identifies clusters that more accurately represent the underlying data. Our results highlight the importance of choosing an algorithm whose assumptions are in line with the phonetic data being considered.

Keywords. phonetics; vowels; unsupervised clustering; K-means; machine learning; corpus methods

1. Introduction. Cluster analysis is a machine learning task that places unlabeled data into groups. This is a common technique for exploratory data analysis because it can reveal patterns that are imperceptible or unexpected. Clustering can also generate objective labels that can be useful in subsequent analysis. In descriptive phonetics, clustering algorithms are primarily used to characterize vowel spaces in a data-driven way. The K-means algorithm (Forgy 1965) is most commonly used for this purpose (e.g. Renwick & Ladd 2016; Shi 2019; Bissell 2021), in part because it is intuitive, computationally efficient, and relatively straightforward to implement. Otherwise, existing work has adopted algorithms that belong more broadly to the family of centroid-based clustering, which K-means is also a part of (e.g. Gaussian Mixture Models; Vallabha et al. 2007).

K-means has been used for for tasks such as confirming previously labeled vowel categories (e.g. Bissell 2021 on Tol; Renwick & Ladd 2016 on Standard Italian; Nadeu and Renwick 2016 on Catalan), evaluating whether specific phonetic parameters improve vowel classification (e.g. Shi 2019; Shi et al. 2019), and comparing the separability of vowel categories across different speech styles (e.g. Czoska et al. 2015). Oftentimes, clustering algorithms are used as one of several analysis routes. For example, Renwick & Ladd (2016) use K-means to supplement other acoustic evidence for a marginal vowel contrast in Standard Italian.

However, K-means makes simplifying assumptions about the underlying data that may not be met by the particular data under consideration. In these cases, results from K-means clustering may be a sub-optimal or even misleading characterization of the data. In this paper, we consider an alternative clustering algorithm, OPTICS (Ordering Points To Identify the Clustering Structure; Ankerst et al. 1999), which uses areas of high density in the data to identify clusters. Compared to K-means, OPTICS makes fewer assumptions about properties of the underlying data, and also separates prototypical members of a cluster from noise. These characteristics make OPTICS a potentially useful algorithm for descriptive phonetic analysis.

In this paper, we compare the two algorithms using data from i) Hillenbrand et al. (1995) and ii) the Buckeye Corpus of Conversational Speech (Pitt et al. 2005), respectively representing lab and conversational speech. We find that while K-means and OPTICS perform similarly in

[*] Thank you to the audience at the LSA 2023 for their insightful comments. Authors: Emily Grabowski, University of California, Berkeley (emily_grabowski@berkeley.edu) & Jennifer Kuo, University of California, Los Angeles (jenniferkuo2018@ucla.edu).

lab speech, they diverge significantly when applied to speech collected in more naturalistic contexts. In conversational speech, the clusters that OPTICS identifies appear to align better with human-identified vowel centers, while K-means primarily maximizes the dispersion of the cluster centers. We argue that this result positions OPTICS as a useful algorithm in the clustering of vowel spaces.

2. Comparison of K-Means and OPTICS: method and assumptions. While many clustering algorithms are similar on the surface, each algorithm has built-in assumptions that can result in different characterizations of the same data. Among these many options, we chose to compare K-means and OPTICS because they represent two major types of clustering—centroid-based and density-based—that differ significantly in their mathematical approach. Within centroid-based algorithms, K-means is most commonly used in phonetics, and a logical starting point for comparison. OPTICS (a close relative of another well-known method DBSCAN) is a type of density-based clustering, and has become a common tool for exploratory analysis more generally due to its relatively few statistical assumptions. While it is certainly also worth considering other clustering algorithms, we will take these two algorithms as a useful starting point for a comparative analysis.

In the following section, we compare some key differences between K-means and OPTICS that are particularly relevant to vowel data. A more general discussion can be found in papers describing the benefits and shortcomings of both K-means and OPTICS (Morissette & Chartier 2013; Kanagala & Krishnaiah 2016; Schubert et al. 2017; Bajal et al. 2022).

2.1. CLUSTER CENTERS. K-means starts with a known number of clusters (where K is the number of clusters). Clusters are randomly initialized, and the location of each cluster center (i.e. centroid) is adjusted iteratively to best fit the data.

The original formulation of K-means is known to be sensitive to the randomly initiated starting point of centroids. This issue can be dealt with by either running the algorithm multiple times on different starting points, or using variants like the K-means++ algorithm (Arthur & Vassilvitskii 2006), which initializes centers to be maximally dispersed from each other. The latter solution builds a new assumption to the algorithm- that the ideal cluster centers are also approximately maximally dispersed. Such an assumption may hold true for some vowel data and is in line with theories that predict maximal dispersion for vowel categories (Liljencrants & Lindblom 1972). However, the algorithm is consequently expected to struggle in cases where cluster centers are not evenly distributed.

OPTICS identifies clusters as areas of relatively high density separated by areas of low density. This means that there is not an *a priori* set number of clusters. Additionally, while cluster centers can be calculated from the resulting clusters, they are not a key part of the algorithm. Because OPTICS relies on variation in density, it is predicted to struggle when the data is of approximately uniformly high or low density, for example with significantly overlapping clusters.

2.2. TREATMENT OF POTENTIAL OUTLIERS. K-means classifies all points as part of a cluster.[1] On the other hand OPTICS will label some points as 'noise' when they do not fit the criteria to be in any cluster. In other words, not all points are assigned a cluster label.

Whether it is desirable to classify all points or not depends on the goal of clustering: either to

[1] While traditionally the K-means algorithm has been known to be sensitive to outliers, a minor adjustment to the algorithm can help account for this by classifying outliers without using them to update the centroid.

classify all points or to identify structure in the data. In the case of exploratory descriptions, we argue that the latter is the priority. In this light, OPTICS may perform better in terms of identifying core structure in the data and distinguishing between prototypical data points and outliers.

2.3. Data geometry. K-means also contains assumptions about the geometry of the underlying clusters that it is attempting to identify. Generally, all centroid-based clustering algorithms assume that clusters radiate out evenly from centroids. K-means further assumes that clusters are roughly circular and of similar size and variance (Murphy 2022). Where data deviates from these assumptions, K-means can identify clusters that mischaracterize the data. In contrast, OPTICS makes no assumptions about cluster geometry. Clusters can be of any shape or size as long as they follow the density constraints discussed above.

2.4. Implications for vowel spaces. Now we consider whether the two algorithms discussed above are suited to the vowel data considered in this study (Figure 1). The data on the left are 7 English monophthongs collected in a controlled environment and context, taken from Hillenbrand et al. (1995). Here, vowels fall into similarly-sized clusters that are roughly circular and of similar densities. With one exception (/æ/ vs. /ɛ/[2]), clusters are well-separated and relatively evenly dispersed across the vowel space. In other words, K-means is expected to work well, as the data aligns with assumptions of the algorithm. OPTICS, which primarily assumes that there are areas of high densities separated by areas of lower densities, should also perform well on the lab speech.

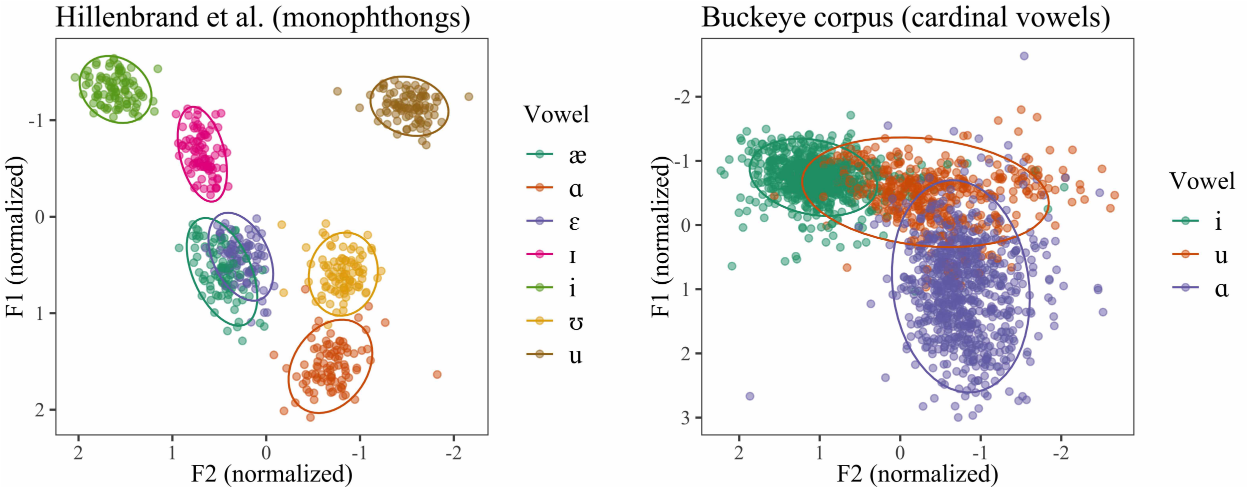

Figure 1. Normalized data for lab (left) and conversational (right) speech. Ellipses show 95% confidence intervals

On the right of this figure is conversational data, specifically a subset of cardinal vowels (/i, ɑ, u/) from the Buckeye corpus (Pitt et al. 2005). In this case, several of the K-means assumptions are not met; the categories are not of the same size, not circular, and significantly overlap. The category centers are also not maximally dispersed within the data space, where /i/ and /u/ in particular are relatively close together. Consequently, K-means will not be expected to perform as well on the data. While variation in shape and size is not an issue for OPTICS, the significant

[2] Higher dimensional representations (i.e. including duration) improves the separability of /æ/ and /ɛ/, but in general higher dimensionality in the data may still result in overlapping categories.

overlap between the categories (and therefore lower variation in density) may also pose issues for this algorithm.

In summary, K-means has several pitfalls: it makes assumptions about optimal cluster dispersion and geometry, and requires an *a priori* determination of the number of clusters (though, as discussed below, several metrics can be used to estimate the 'optimal' number of clusters). While these assumptions are reasonably met in controlled lab speech, they are clearly not met in more naturalistic speech. OPTICS makes fewer assumptions, but the significant overlap of clusters in the naturalistic speech data can potentially pose a problem for this algorithm as well. In the next section, we will investigate how these methods perform in lab vs. naturalistic speech, starting with the lab speech context.

3. OPTICS vs K-Means in controlled lab speech. The first case that we consider is controlled lab speech. Lab speech is recorded in a relatively noise-free environment, and often has more constrained contextual variation. Compared to corpus data, lab speech typically has less data points and speakers and can exclude potentially interesting sources of variation. This trade-off between the amount of data and control over the environment is common in linguistics.

Moreover, findings from lab speech, which is artificially restrictive, may not hold in more naturalistic corpus data. Methodologically, it is therefore productive to treat lab and corpus data as complimentary. For this study, we use controlled lab speech as a first case for comparing the viability of the OPTICS and K-means algorithms. This allows for investigation of clustering in the best case scenario.

3.1. DATA. Lab speech data come from Hillenbrand et al. (1995), which consists of acoustic measurements (duration and F1-F3 at multiple time points) from 93 speakers. Note that original recordings were not available, so our calculations were made from the published acoustic measurements. Stimuli were English vowel phonemes with two repetitions per speaker, all collected in the /hVd/ environment. We used the subset of 7 monophthongs (/i ɛ æ ɑ ʊ u/), for a total of 1302 tokens. This dataset has a relatively low number of tokens per speaker, but since all tokens are in the same environment there is expected to be relatively little variation.

We restricted the data to monophthongs to compare the two clustering methods using static measures of F1 and F2. For monophthongs, midpoint F1/F2 data are the most commonly reported acoustic measures and are considered to be reasonably descriptive of the vowel space. As such, we take the midpoint F1/F2 data as the input for the clustering algorithm. While both K-means and OPTICS can perform well on higher-dimensional data (e.g. by including other measures such as duration and formant trajectories), we use this simpler two-dimensional parameter space for ease of visual interpretation and methodological comparison.

3.2. METHODS. We normalize the data by speaker using Lobanov normalization (Nearey 1965). A token was also excluded if it was less than 50 ms long, or if any of the normalized measures were outside of 3 standard deviations from the mean. Under these criteria, approximately 5% of the data were excluded. The same data cleaning process was used for both clustering algorithms.

The other major methodological step is to optimize each clustering algorithm, by setting parameters to get the best fit for the data. For K-means, optimization means selecting the number of clusters K. When the number of clusters is known *a priori*, this is straightforward, but more frequently linguists are interested in cases where the 'correct' number of clusters is unknown. In these cases, K is generally selected to minimize the number of clusters while maximizing the separation between clusters.

We consider both possibilities for the Hillenbrand data (where *a priori* K=7) and optimize K using two metrics, the inertia plot (Thorndike 1953) and silhouette plot (Rousseeuw 1987). When setting the number of clusters, there is a trade-off between the interpretability of the results (fewer clusters) and the ability to describe the data (distance between points and centers). The inertia plot and silhouette plot visualize this trade-off in different ways, and can be used to identify a value K where these competing constraints are optimized (see also Nanjundan et al. 2019 for a summary of K-means optimization). Note that there is still a level of subjective judgment present in determining the 'best' value for K.

In OPTICS, the number of clusters is not decided beforehand, but is based on some cutoff value which is set specifically for each data set. Setting this cutoff involves a trade-off between inclusion of points and number of clusters. A high cutoff results in less points determined as noise (i.e. more points included), but also may result in large, uninformative clusters. On the other hand, a smaller cutoff increases the separability between clusters, but results in more points being excluded as noise. We set a value for the cutoff that minimized the number of points classified as noise but maintained some cluster structure (i.e. the maximum cutoff that avoided merging all of the data into a single cluster)

3.3. Results. Now, we turn to comparing results of K-means and OPTICS clustering on the Hillenbrand dataset, as visualized in Figure 2. Based on the procedure described in Section 3.2, both algorithms independently identified 6 clusters as optimal.

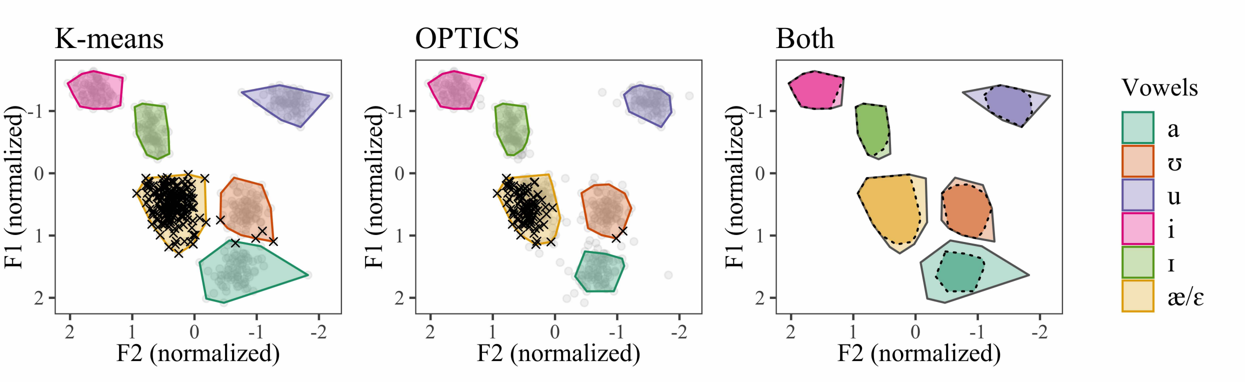

Figure 2. Clusters found by K-means vs. OPTICS using monophthongs from Hillenbrand et al. (1995). Clusters are visualized as convex hulls, where 'x' indicates mislabeled points.

At a high level, OPTICS and K-means appear to perform very similarly. Note that the underlying data had 7 vowel categories; two of them (/æ/ and /ɛ/) have significantly overlapped F1/F2 midpoints in the data, and consequently both K-means and OPTICS are not able to separate the two categories. In this particular situation, the overlap can be resolved by adding an additional dimension (duration). However, overlapping clusters is a very common feature of vowel data and is not always resolved by higher dimensionality, as is the case for the conversational speech in the next section.

In terms of differences between the algorithms, K-means has slightly larger clusters because it does not exclude points as noise and the clusters accommodate potential outliers. In contrast, OPTICS identifies these values as noise and they are not assigned a cluster. This results in more separation between the clusters, and smaller, rounder shapes (despite no constraint on cluster shape existing in the algorithm). These results demonstrate that in an idealized vowel space derived from controlled lab data, both K-means and OPTICS can capture the general patterns in how vowels are structured in phonetic space.

4. OPTICS vs K-means in conversational speech. In this section, we investigate how K-means and OPTICS perform on conversational speech. Corpus data provides insight into speech in more naturalistic settings and often allows for more data for an individual speaker than can be reasonably collected in a lab setting. On the other hand, these data also often have less even distribution across vowel categories and contexts, more overlap, and more variation, which can be expected to affect the results of some clustering algorithms (as discussed in Section 2).

4.1. DATA AND METHODS. Data for this analysis are from the Buckeye Speech Corpus (Pitt et al. 2005), which consists of face-to-face naturalistic interviews. For the purpose of this study seven speakers were randomly selected for analysis. In addition, data were subset to the three vowels /ɑ, i, u/ and constrained for environment by removing vowels that were adjacent to semivowels, nasals, and liquids, resulting in 1953 tokens as summarized in Table 1. We use this relatively restricted set of vowels to minimize overlap between categories and allow for a more interpretable comparison of the algorithms under consideration. Data processing and outlier selection were conducted as described above in Section 3.2.

Vowel	Tokens
/ɑ/	776
/i/	740
/u/	437

Table 1. Tokens distribution in Buckeye subset

4.2. RESULTS.

4.2.1. OPTIMIZATION. As discussed in Section 2, for K-means we consider both a pre-specified K value and one identified using parameter optimization. K-means optimization selected 6 clusters, compared to the 'ground truth' of K=3 clusters.[3] Figure 3 compares the clusters identified by K-means in both cases (K=3 vs. K=6). In both configurations, K-means splits the space into relatively even groups that are maximally dispersed, seemingly regardless of the underlying structure of the data. This suggests that the optimization procedure for K-means is not as effective in cases where the data do not meet the algorithmic assumptions (Schubert 2022). In the rest of this section, for the sake of comparability, we use the K-means K=3 results.

For OPTICS, optimization resulted in the algorithm identifying 3 clusters. Note that a large proportion of points are excluded as noise (36%), which is suboptimal because this may mask meaningful variation. Nevertheless, as discussed in the next section, the OPTICS results appear to better capture broader generalizations about the distribution of phonemic vowel categories.

[3] Note that even though there are 3 vowel clusters, there could easily be more than three clusters of interest, if for example one is interested in subphonemic variation.

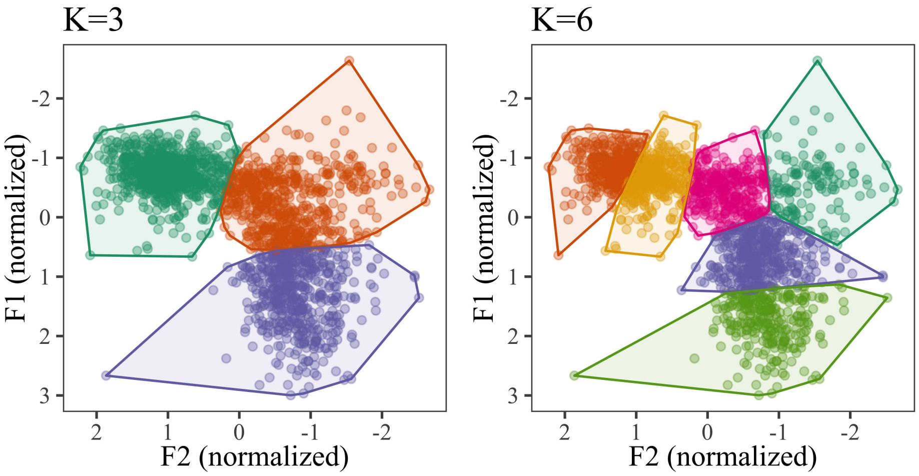

Figure 3. Clusters identified by K-means for conversational speech with three clusters (left) and six clusters (right)

4.2.2. Comparison of K-means and OPTICS. Figure 4 compares the results of K-means and OPTICS against the original hand-labeled data. First, when examining the original data, we can identify the following general patterns:

1. /i/ is front and has distinct boundaries.
2. /u/ is relatively fronted and has significant overlap with the adjacent categories.
3. /ɑ/ is oblong.

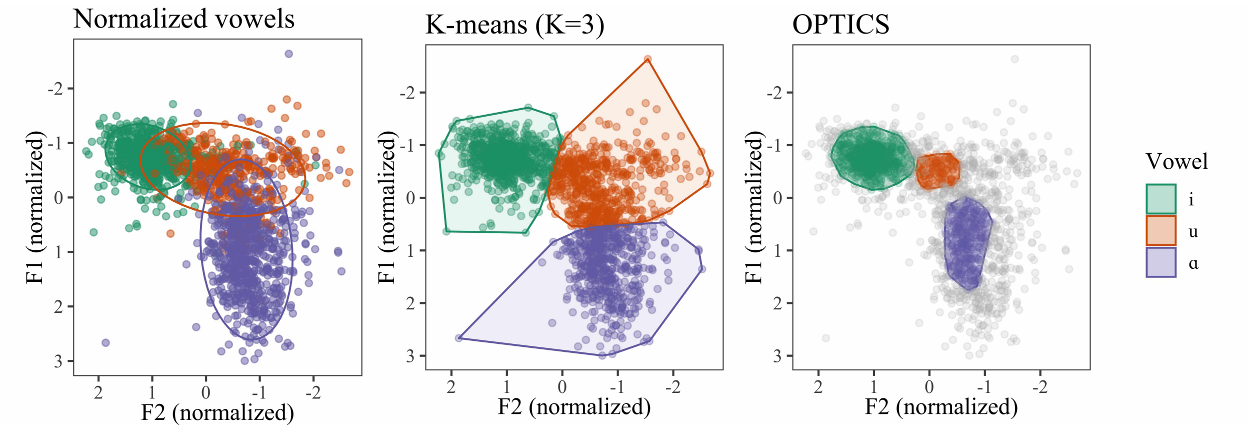

Figure 4. Comparison of K-means (K=3) and OPTICS on Buckeye corpus cardinal vowels

The K-means results fail to capture these generalizations in the data. Instead, the data are partitioned into three clusters of roughly equal size. Notably, clusters also divide areas that may be expected to group together based on the underlying data. In contrast, OPTICS appears to extract the generalizations identified above: /i/ is the largest cluster, not because of the size of the underlying vowel category, but because it has well-defined boundaries, resulting in less points being excluded as noise. The /u/ cluster is relatively front and placed in an area where there is minimal overlap with the other vowels in the original data. Finally, the /ɑ/ cluster is relatively oblong. From these results, it appears that OPTICS is more likely to extract qualitatively meaningful patterns from the data in conversational speech.

To deal with overlapping clusters, K-means appears to enforce sometimes artificial boundaries that cut across these overlapping areas. In contrast, OPTICS identifies areas of high density, and excludes many of these overlapping areas as noise. These results suggest that while OPTICS does not classify every point, it is identifying likely vowel centers. While this may abstract away from interesting and informative variation, it also allows for several key generalizations to be captured in the data, which we argue is the more common usage of clustering.

5. Conclusion. In this paper, we compare two clustering methods (OPTICS and K-means) on lab and conversational speech. We find that in lab speech, both algorithms perform similarly and are able to capture the major patterns in the data. In contrast, in conversational speech, K-means appears to struggle to capture key generalizations, while OPTICS captures those patterns at the cost of excluding many points from the clusters. This difference underlines the need to assess each set of data before choosing a clustering method and highlights the potential of OPTICS for supplementing traditional methods for describing vowels.

A particular challenge for both methods is posed by overlapping clusters in the underlying categories. In conversational speech, even with constraints set on the number of categories and environments, there was significant overlap between the clusters. Such overlap can be expected to increase in even more naturalistic and varying data. One way to help reduce overlap would be to move to a higher-dimensional representation of the vowel space, which is a potential productive direction for future research, although we note that this is unlikely to entirely resolve acoustic overlap between vowel categories.

This paper outlines an initial comparison between two clustering approaches, but there are several beneficial paths to further develop our understanding of the use of clustering in descriptive phonetics. Although we looked exclusively at F1/F2 midpoints in English monophthongs, subsequent work should also increase the type and variety of data being examined, such as by considering dynamic segments, higher-dimensionality data, and non-English vowel systems. In addition, we focused on categorization of phonemic vowel categories in the current analysis, but there is also interesting variation at the subphonemic level that may also be of interest to those using cluster analysis. Finally, we describe only a small subset of the available clustering algorithms, and there are likely others that may also be appropriate for the use case that we are defining here, such as hierarchical clustering methods, which would be useful to explore in further detail.

Both K-means and OPTICS are able to approximate superficial structure in the vowel space, but differ in their underlying mathematical approach. Unlike K-means, density-based methods like OPTICS can identify areas of high density even in noisy data and extract likely vowel centers from that structure, which we propose results in more insight into the structure and patterns of the vowel system. This study emphasizes the importance of choosing an appropriate clustering algorithm and highlights the potential of density-based algorithms like OPTICS for analysis of vowel spaces.

References

Ankerst, Mihael, Markus M Breunig, Hans-Peter Kriegel & Jörg Sander. 1999. OPTICS: Ordering points to identify the clustering structure. *ACM SIGMOD* 28(2). 49–60.

Arthur, David & Sergei Vassilvitskii. 2006. k-means++: The advantages of careful seeding. Tech. rep. Stanford.

Bajal, Eshan, Vipin Katara, Madhulika Bhatia & Madhurima Hooda. 2022. A review of clustering algorithms: comparison of DBSCAN and K-mean with oversampling and t-SNE. *Recent Patents on Engineering* 16(2). 17–31.

Bissell, Marie. 2021. Automatic phonetic classification of vocalic allophones in Tol. *Proceedings of the Linguistic Society of America* 6(1). 403–410.

Czoska, Agnieszka, Klessa Katarzyna & Maciej Karpinski. 2015. Polish infant directed vs. adult directed speech: Selected acoustic-phonetic differences. In *18th international congress of phonetic sciences (ICPhS)*, .

Forgy, Edward W. 1965. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. *biometrics* 21(3). 768–769.

Hillenbrand, James, Laura A Getty, Michael J Clark & Kimberlee Wheeler. 1995. Acoustic characteristics of American English vowels. *The Journal of the Acoustical society of America* 97(5). 3099–3111.

Kanagala, Hari Krishna & VV Jaya Rama Krishnaiah. 2016. A comparative study of K-Means, DBSCAN and OPTICS. In *2016 international conference on computer communication and informatics (ICCCI)*, 1–6. IEEE.

Liljencrants, Johan & Björn Lindblom. 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language* 839–862.

Morissette, Laurence & Sylvain Chartier. 2013. The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology* 9(1). 15–24.

Murphy, Kevin P. 2022. *Probabilistic machine learning: an introduction*. MIT press.

Nadeu, Marianna & Margaret EL Renwick. 2016. Variation in the lexical distribution and implementation of phonetically similar phonemes in Catalan. *Journal of Phonetics* 58. 22–47.

Nanjundan, Sukavanan, Shreeviknesh Sankaran, CR Arjun & G Paavai Anand. 2019. Identifying the number of clusters for K-Means: A hypersphere density based approach. *arXiv preprint arXiv:1912.00643* .

Nearey, Terrance M. 1965. *Phonetic feature systems for vowels*: University of Alberta dissertation.

Pitt, Mark A, Keith Johnson, Elizabeth Hume, Scott Kiesling & William Raymond. 2005. The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication* 45(1). 89–95.

Renwick, Margaret EL & D Robert Ladd. 2016. Phonetic distinctiveness vs. lexical contrastiveness in non-robust phonemic contrasts. *Laboratory Phonology* 7(1).

Rousseeuw, Peter J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20. 53–65.

Schubert, Erich. 2022. Stop using the elbow criterion for k-means and how to choose the number of clusters instead. *arXiv preprint arXiv:2212.12189* .

Schubert, Erich, Jörg Sander, Martin Ester, Hans Peter Kriegel & Xiaowei Xu. 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)* 42(3). 1–21.

Shi, Yuanming. 2019. *An investigation on prenasal merger in southern American English through automatic speech recognition*: University of Georgia dissertation.

Shi, Yuanming, Margaret E. Renwick & Frederick Maier. 2019. Improved vowel labeling for prenasal merger using customized forced alignment. *The Journal of the Acoustical Society of America* 146(4). 2957–2957.

Thorndike, Robert L. 1953. Who belongs in the family? *Psychometrika* 18(4). 267–276.

Vallabha, Gautam K, James L McClelland, Ferran Pons, Janet F Werker & Shigeaki Amano. 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences* 104(33). 13273–13278.