K-means and hierarchical clustering of f0 contours

Constantijn Kaland[1]*, Jeremy Steffman*[2]*, Jennifer Cole*[3]

[1]Institute of Linguistics - Phonetics, University of Cologne, Germany
[2]Linguistics and English Language, The University of Edinburgh, United Kingdom
[3]Department of Linguistics, Northwestern University, United States

ckaland@uni-koeln.de, jeremy.steffman@ed.ac.uk, jennifer.cole1@northwestern.edu

Abstract

Cluster analysis on time-series f0 data is an increasingly popular method in intonation research. There are a number of methodological decisions to take when applying cluster analysis. Crucially, these decisions may affect the clustering results, potentially also the conclusions of the research. This paper investigates the extent to which the choice for either K-means or hierarchical clustering, two of the most popular clustering methods, leads to grouping differences that are potentially relevant for intonation research. This is tested using a dataset of f0 measures taken from imitated intonation patterns in American English. The analysis concerns a generic correlation test between K-means and hierarchical clustering outcomes as well as a number of specific measures assessing partitioning quality and f0 contour differences. The results show that both cluster methods generally show very similar outcomes, although considerable differences for specific clusterings might occur.

Index Terms: f0 contour, cluster analysis, intonation, k-means, hierarchical

1. Introduction

Cluster analysis is an unsupervised statistical method to divide a dataset into groups (clusters) of observations that are similar along one or more quantitative measurement dimensions ([1]; [2]). It is a particularly suitable method when little is known about the (number of) possible groups. The applications of cluster analysis range from genetics to marketing and recently it became a popular method in research on intonation (e.g. [3]; [4]). Models of intonation generally assume phonological categories of tones or tonal configurations that make up an intonational feature (e.g., pitch accents and boundary tones; [5]). The simplest intonational pattern consists of single intonational feature, while complex configurations are composed of a sequence of two or more intonational features [6]. Much research has been devoted to the extent to which intonation patterns are indeed categorical in nature. The literature has shown that a strict categorical approach to intonational meaning does not hold [7]: some patterns have multiple meanings and different meanings might be expressed by the same pattern (e.g., [8]; [9]; [10]; [11]; [12]; [13]; [14]). Contemporary empirical methods such as cluster analysis offer a new perspective to the categorical nature of intonation. Studies have applied this method in order to explore the intonation of previously under-researched languages (e.g. [3]; [15]; [16]) as well as to test existing models of intonation in well-studied languages (e.g., [17]; [18]; [4]).

The usefulness of cluster analysis for intonation crucially depends on a myriad of methodological choices that researchers need to make. Roughly, these choices can be divided into ones relating to the contour representation, that is the way an f0 contour is numerically represented *before* clustering, and ones relating to the cluster analysis *itself*, i.e. its specific statistical properties (e.g., [19]). The current paper focuses on the latter, in particular on the type of cluster analysis. To this end, K-means and hierarchical clustering are compared for their performance on the same dataset of American English f0 contours, as used in previous research [4].

K-means clustering (KM) is based on partitioning the data according to *centroids* [2]. This method cannot be applied without a prior choice for the number of clusters (K). Centroids are then identified in an iterative way, by assigning observations to K clusters randomly. Each time all observations are randomly assigned, centroids are calculated by taking the mean value for all observations in a cluster. Then, each observation is assigned to the centroid to which it lies closest, as calculated by Euclidean distance or some other distance metric. The centroid computation and centroid assignment are repeated until the centroids are stable, i.e. when the two most recent centroid assignments are identical.

Hierarchical clustering (HC) is based on partitioning the data using a tree-structure (*dendrogram*) [2]. Bottom-up HC starts with assigning each observation into a cluster. Then, an iterative merging process follows in which the (clusters of) observations with the smallest distance are merged until all observations are in one cluster. Note that the initial and final state are not informative as they are tantamount to single observations or the entire dataset (=no clustering) respectively. The dendrogram visualises the merging process such that the height in the tree structure shows how many clusters were formed (higher in the tree = less clusters). As in KM, HC requires setting the way distances between observations are calculated (distance metric). In addition, HC requires setting the way distances between clusters of observations are calculated (linkage criterion).

Both KM and HC are among the most popular clustering techniques in intonation research, see [19] for an overview. Given the fundamentally different way in which they work, their performance might vary. Previous research has made explicit comparisons of KM and HC using datasets from a variety of scientific disciplines. Concerning the computational costs of KM and HC, several studies concluded that the running time as well as memory usage is larger for HC than for KM ([20]; [21]). It was also found that clustering performance generally increases with larger datasets and that HC performs better than KM for smaller datasets ([22]; [21]; [23]). It was furthermore reported that KM shows overall less accuracy in clustering preknown groups in the data and is more sensitive to noise in the data than HC [22]. Crucially, the latter conclusions were drawn from a comparison of multiple clustering techniques on time-series data. Another study comparing multiple clustering meth-

ods (among which KM and HC with Ward linkage criterion) concluded that performance of the cluster analysis largely depends on the dataset and recommend exploratory methods that compare results from different algorithms [24].

So far, intonation research has not explicitly compared the impact of the choice of clustering technique on the outcome of the analysis. Recent studies did investigate the discriminatory performance of different contour representations and distance metrics [19], which were mainly tailored to HC. Another study used multiple classification techniques, among which KM, to test the categories underlying inventory of nuclear intonation patterns in American English [4]. It was found that the model that predicts eight distinct nuclear intonation patterns formed using simple High and Low tone pitch accents is likely in need of revision given the optimum of five distinct patterns appearing from both production and perception analyses. Thus, it remains to be seen whether the type of clustering method needs to be chosen in a principled way when clustering f0 contours. The current study investigates this by a number of metrics assessing the quality of the KM or HC clustering output based on the same data. The types of tests carried out in the current study are two-fold. On the one hand we perform tests to assess the similarity of the KM and HC clustering output, by comparing their outputs to each other. This is done using correlation metrics and distributions of cluster mismatches, which do not directly take into account any f0 data to assess the clustering methods. On the other hand, we test the respective clustering outputs relative to the data, i.e. assessing the partitioning quality with regard to the f0 contour differences. The latter type of tests concern proportional assignment of contours to clusters, root-mean-square-distances between the contours in the clusters, and a comparison of within and between cluster variance. Further details on the methodological choices are provided in the next section.

2. Methodology

The dataset used in this study is identical to the one used in [4]. The dataset, R-script and additional figures are available as supplementary material: https://osf.io/u3nsz/. For a detailed description of the data collection, see [4]. The relevant data for the current study concerns time-series f0 measures of 240 intonation contours, each specified in terms of three intonational features (pitch accent, phrase accent, boundary tone), produced by 30 speakers of American English. Each contour (observation) is represented by 30 equidistant measurement points in the equivalent rectangular bandwidth (ERB) scale and then speaker-scaled. Each contour is the mean of eight repeated imitations by the same speaker for one out of eight patterns: HHH, HHL, HLH, HLL, LHH, LHL, LLH, LLL.

2.1. Cluster analyses

Both the KM and HC cluster analyses were performed in R [25] and R Studio [26]. KM was done using the `kml` package [27], HC was done using the `stats` package available in base R [25]. KM was done using the algorithm described in [28], using Euclidean distance with Gower adjustment [29]. The algorithms to obtain an initial cluster assignment (starting point for finding stable centroids) were left to the default: 'kmeans-' followed by an alternation of 'kmeans--' and 'randomK', as described in further detail in [27]. HC was performed with Euclidean distance as distance metric and complete linkage as linkage criterion. The latter HC settings have been used by default by the software tool used in previous work [3]. Note that the assessment of different linkage criteria falls beyond the scope of the current study and is planned for future research.

The outcomes of KM and HC were obtained for several *rounds* of cluster analysis, i.e. assuming 2 to 8 clusters. This was done in order to allow for a comparison over the course of different clusterings, potentially revealing differences between the respective methods. Note that multiple (unsupervised) classification techniques in [4] indicated that the optimum for this dataset lies at 5 clusters. Together with the indication of the eight imitated intonation contours, these data provide a reference for the assessment of the clustering results in the current study.

2.2. Similarity assessments

Once the clusterings were obtained, the similarity of the KM and HC outcomes was assessed. Note that this assessment could not be done on the raw cluster assignments of each method. That is, for a clustering round assuming, for example, 3 clusters, cluster 1 obtained from KM does not necessarily match cluster 1 obtained from HC. Therefore, a matching procedure was carried out. This was done by systematically testing all possible mappings and by choosing the mapping that had the highest number of contours that matched between KM and HC (i.e. the optimal exhaustive mapping). Note that this matching procedure tests an exponentially growing number of possible mappings for each subsequent round of clustering. With 2 assumed clusters, KM and HC could map in two ways {1:1,2:2} or {1:2,2:1}. With 3 assumed clusters, KM and HC could map in six ways: {1:1,2:2,3:3}, {1:1,2:3,3:2}, {1:2,2:1,3:3}, {1:2,2:3,3:1}, {1:3,2:1,3:2}, or {1:3,2:2,3:1}. From 4 to 8 clusters assumed, the number of possible mappings tested were 24, 120, 720, 5040, and 40320 respectively. The numbers assigned to the clusters were then recoded according to the optimal mapping of the matching procedure, for each round of clustering, such that the clusters numbers of KM and HC aligned, i.e. KM cluster 1 mapped optimally onto HC cluster 1, KM2 onto HC2, etc. Thereafter, Kendall τ was computed between the KM and HC outcomes of each round. This was done to assess to what extent the clustering methods correlated, i.e. by ranking their cluster number assignments and calculating concordant and discordant rankings. In addition, it was counted for each intonation pattern and for each clustering round which contours occurred in matching clusters and which ones did not. An overall mismatch percentage was obtained for each clustering round.

2.3. Partitioning quality assessments

In order to assess the quality of the partitioning, the cluster assignment for each of the eight intonation patterns was calculated as a proportion. That is, for each pattern it was calculated how often it was assigned to each cluster, in terms of the proportion of productions of that pattern that were assigned to that cluster, maximally 30. For example, for a round assuming 3 clusters the maximum number of LHH patterns assigned to a particular cluster was 19 times (out of 30 possible assignments = 0.63). For each clustering round, the average of the maximum proportions (over all patterns) was taken in order to assess the quality of the pattern-to-cluster assignments. Higher proportions would indicate a more systematic mapping of a given contour to one particular cluster. Thus, lower proportions might indicate either the failure of the analysis to assign patterns to clusters in an optimal way and/or that some patterns are particularly difficult to cluster in general, because of their shape similarity.

In order to disentangle the latter two scenarios further, an

additional measure was taken to assess the difference between two clusters based on their average F0 contours. This was done by taking the RMSD between the two F0 vectors, where the 30 time-normalized F0 measurements in each vector are averaged over 30 participants. RMSD was shown to be a close approximation of how f0 contour differences are perceived by the human ear [30]. For rounds with more clusters, more comparisons were made. Per clustering round, the average of all RMSDs was taken to represent the distance between the clusters in that round.

As a final partitioning quality assessment, the variance within and between clusters was calculated for each round, a method implemented in the software tool [3]. The general assumption for the majority of the clustering methods is that with more clusters, the variance within clusters becomes smaller (observations within clusters are more similar), while the variance between clusters becomes larger (clusters represent more distinct groups). When the two variances no longer significantly diverge, the optimal number of clusters has been reached. Within cluster variance was computed by taking the standard deviation of the f0 measures representing the average contour in a certain cluster, which in turn was averaged over all clusters in a particular round. Between cluster variance was computed by taking the mean f0 for each measurement point in each cluster, over which the absolute difference between the minimum and maximum was calculated as an assessment of the f0 distance between the clusters. The resulting 30 measurement points of f0 distance were then averaged for each round.

3. Results

Table 1: *KM and HC similarity measures per clustering round: number of mismatching observations per intonation pattern (max. per cell is 30), mismatch percentage, and Kendall τ.*

	No. of clusters						
Pattern	2	3	4	5	6	7	8
HHH	0	0	14	9	9	9	14
HHL	0	1	5	0	1	1	9
HLH	0	1	1	1	17	18	13
HLL	0	0	1	3	12	13	13
LHH	0	9	28	0	13	23	16
LHL	1	1	9	0	0	23	23
LLH	4	1	1	0	1	4	3
LLL	0	0	30	0	0	1	1
mismatch %	2.08	5.42	36.25	5.42	22.08	38.33	38.33
τ	0.96	0.85	0.24	0.84	0.54	0.42	0.47

The results of the similarity measures are reported in Table 1. A general trend can be observed in that the number of contours for which KM and HC mismatch increases with higher numbers of clusters. Kendall τ showed a significant correlation for all rounds, which was moderate to strong for all except the round with 4 clusters (weak correlation). As for the individual intonation patterns, LLH had the lowest number of mismatches, whereas LHH had the highest number of mismatches. The round with 4 clusters shows a particularly high number of mismatching contours. This round was further analyzed by an inspection of the contours in the clusters of KM and HC. It can be observed from the contours in Figure 1 that KM and HC showed mismatching cluster assignments for a concave rising pattern (KM1/HC4: HHH vs. LHH), for a rise-fall-rise pattern (KM2/HC3: HLH/HLL vs. LHL/LLH), for an overall rising pattern (KM3/HC4: LHH), a shallow fall (KM4/HC2: LLL vs. HLH/HLL), and a level contour (KM4/HC3: LLL vs. LHL/LLH). Note that KM and HC did not have any matching contours in their fourth cluster.

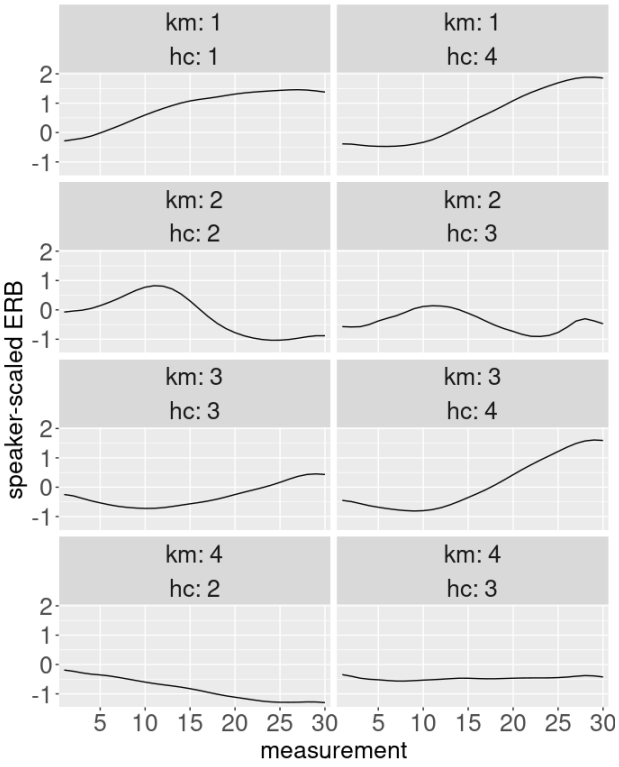

Figure 1: *Average f0 contours in the round with 4 clusters, for each combination (matching/mismatching) of cluster numbers.*

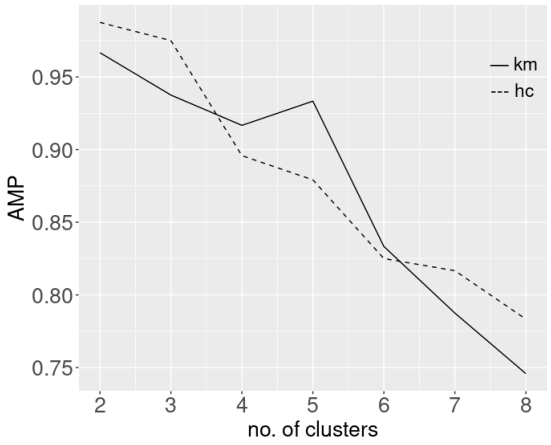

Figure 2: *Average Maximum Proportion (AMP) of cluster assignment per tune in each clustering round.*

The partitioning quality measures showed an overall decline of the proportion maxima per tune with more clusters (AMP, Figure 2). KM shows a clear deviation from this decline for the round with 5 clusters, whereas HC shows an overall meandering decline for increasing numbers of clusters. The RMSD measure showed overall smaller values for higher numbers of clusters, although highly fluctuating for both KM and HC (Fig-

ure 3). KM shows higher RMSD values than HC (larger differences between the average contours of the clusters) for lower number of clusters (<6). For the rounds with 7 or 8 clusters HC shows higher RMSD values than KM. KM and HC show virtually identical RMSD values for the round with 5 clusters. As for the variance as measured within and between clusters (Figure 4), KM and HC show a highly similar pattern: strong divergence up to the round with 5 clusters, then a weak divergence with HC reaching a somewhat higher between cluster variance than KM for the rounds with 6, 7, and 8 clusters respectively.

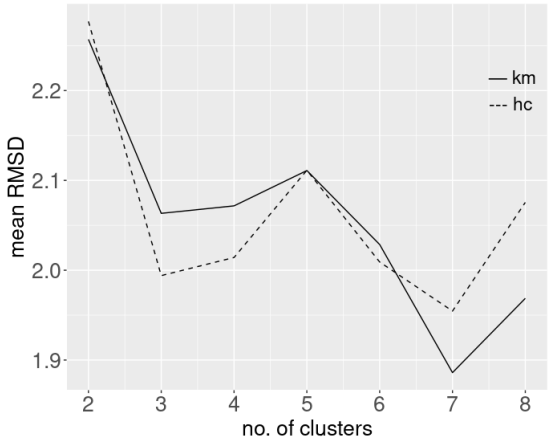

Figure 3: *Mean RMSD between the average contours of all cluster combinations in each clustering round.*

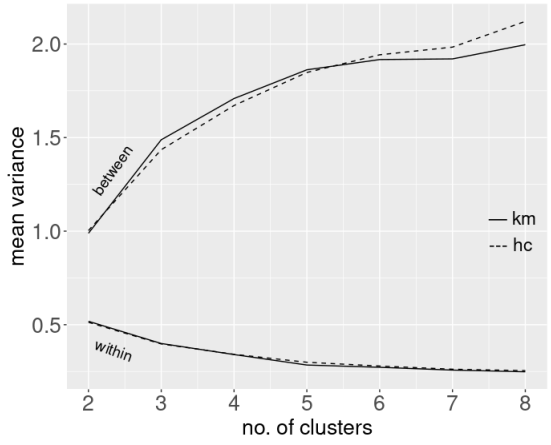

Figure 4: *Within and between cluster variance in each clustering round.*

4. Conclusions

This study has shown that KM and HC clustering methods generally provide highly comparable results on the dataset of f0 contours investigated here. The most remarkable difference between the two methods appeared from the round with 4 clusters, showing a relatively high percentage of contours for which KM and HC mismatched. Further inspection of this particular round revealed that LLL, LHH and HHH were the intonation patterns with the highest number of mismatches respectively (Table 1). The dataset included other patterns that showed a similar shape from the acoustic measures. The difference between HHH and LHH appeared particularly difficult (Figure 1), likely due to the subtle realisation of the initial L tone.

It is furthermore interesting that KM and HC tend to converge on 5 as the optimal number of clusters. This appeared not only from the relatively high τ value and hence the low percentage of mismatches, but also from the RMSD and variance measures. In addition, the AMP measure appeared particularly informative in this respect for the KM method. Note that in [4] a hierarchy of patterns was proposed in which the eight nuclear configurations were merged into five (groups) of contours: {HHH,HHL}, LHH, {LHL,LLH}, {HLL,HLH} and LLL. It is striking that the small number of mismatches between KM and HC for the round with 5 clusters exclusively concerns patterns that are involved in the merge proposed in [4]; HHH, HLH, and HLL.

The observed differences between KM and HC are overall rather small, except for the round with 4 clusters. Crucially, both patterns for which the methods had the most mismatches (LHH and LLL) showed a (near) perfect match in the previous round (3 clusters) and the next round (5 clusters). In this respect it is important to observe that the patterns HHH, HLH, HLL, and LHH had a the highest mismatch rates throughout all clusterings (Table 1). Thus, the observed mismatches in one particular round seem to be a local phenomenon, in the sense that they do not necessarily reveal structural differences between the methods over the course of multiple clusterings. Compared to matching contours, the mismatches in Figure 1 appear to originate from subtle shape differences in some contours. In retrospect, studies that choose either KM or HC to cluster f0 contours are unlikely to have caused a methodological bias in their results. Nevertheless, the current study shows in which ways and due to what kind of contour differences KM and HC might disagree on more than a third of the observations. Crucially, the round with 4 clusters shows that KM and HC had entirely differently fourth clusters (no matches). It cannot be taken for granted, therefore, that KM and HC always lead to the same conclusions. In this respect, it also needs to be mentioned that for the round of clustering that was most likely the optimum, KM and HC were in high agreement. It can thus be reconfirmed from the perspective of comparing clustering methods that finding the optimal number of clusters is key in doing cluster analysis [1]. The degree of convergence between different cluster methods may therefore be taken as an (indirect) indication for the optimal number of clusters (see also [24]). Note that a variance analysis as presented here, or an evaluation based on information cost [31] are more direct and potentially more informative additional evaluation methods that should also be taken into account to obtain a reliable assessment of the ideal number of clusters. It is also important to consider that the current study investigated KM and HC difference on a single dataset. Future work should include multiple datasets and a wider range of parameters to further investigate how KM and HC might differ on f0 contours.

5. Acknowledgements

Research for this paper was funded by the German Research Foundation (DFG) – Project 281511265 – SFB 1252 (CK) and by the U.S. National Science Foundation BCS-1944773 (JC). The dataset, R-script and additional figures are available as supplementary material: https://osf.io/u3nsz/.

6. References

[1] L. Kaufman and P. J. Rousseeuw, Eds., *Finding Groups in Data*, ser. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc., 1990.

[2] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, ser. Springer Texts in Statistics. New York, NY: Springer New York, 2013, vol. 103.

[3] C. Kaland, "Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours," *Journal of the International Phonetic Association*, vol. 53, no. 1, pp. 159–188, 2021.

[4] J. Cole, J. Steffman, S. Shattuck-hufnagel, and S. Tilsen, "Hierarchical distinctions in the production and perception of nuclear tunes in American English." *Laboratory Phonology*, vol. 14, no. 1, pp. 1–51, 2023.

[5] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *Second international conference on spoken language processing*. Banff: ISCA, 1992.

[6] J. Pierrehumbert and J. Hirschberg, "The Meaning of Intonational Contours in the Interpretation of Discourse," in *Intentions in Communication*, P. R. Cohen, J. Morgan, and M. E. Pollack, Eds. Cambridge, MA: MIT Press, 1990.

[7] D. R. Ladd, "The Trouble with ToBI," in *Prosodic Theory and Practice*, J. Barnes and S. Shattuck-Hufnagel, Eds. The MIT Press, 2022, pp. 247–258.

[8] S. Calhoun, "The theme/rheme distinction: Accent type or relative prominence?" *Journal of Phonetics*, vol. 40, no. 2, pp. 329–349, 2012.

[9] P. Prieto, "Intonational meaning," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 6, no. 4, pp. 371–381, 2015.

[10] D. Büring, *Intonation and meaning*, ser. Oxford surveys in semantics and pragmatics. Oxford: Oxford University Press, 2016, no. 3.

[11] M. Grice, S. Ritter, H. Niemann, and T. B. Roettger, "Integrating the discreteness and continuity of intonational categories," *Journal of Phonetics*, vol. 64, pp. 90–107, 2017.

[12] E. Chodroff and J. Cole, "Information Structure, Affect and Prenuclear Prominence in American English," in *Interspeech 2018*. ISCA, 2018, pp. 1848–1852.

[13] ——, "The phonological and phonetic encoding of information structure in American English nuclear accents," in *Proceedings of the 19th International Congress of Phonetic Sciences*. Melbourne, Australia: Australasian Speech Science and Technology Association Inc., 2019, pp. 1570–1574.

[14] X. Xie, A. Buxó-Lugo, and C. Kurumada, "Encoding and decoding of meaning through structured variability in intonational speech prosody," *Cognition*, vol. 211, p. 104619, 2021.

[15] C. Kaland, N. Peck, T. M. Ellison, and U. Reinöhl, "An initial exploration of the interaction of tone and intonation in Kera'a," in *1st International Conference on Tone and Intonation (TAI)*. ISCA, 2021, pp. 132–136.

[16] S. Babinski and C. Bowern, "Automatic categorization of prosodic contours in Bardi," *Proceedings of the Linguistic Society of America*, vol. 7, no. 1, p. 5218, 2022.

[17] H. Seeliger and C. Kaland, "Boundary tones in German wh-questions and wh-exclamatives - a cluster-based approach," in *Proceedings of the 11th International Conference on Speech Prosody 2022*, S. Frota and M. Vigário, Eds., Lisbon, Portugal, 2022, pp. 27–31.

[18] T. J. Laméris, K. K. Li, and B. Post, "Phonetic and Phono-Lexical Accuracy of Non-Native Tone Production by English-L1 and Mandarin-L1 Speakers," *Language and Speech*, vol. 66, no. 4, pp. 974–1006, 2023.

[19] C. Kaland, "Intonation contour similarity: f0 representations and distance measures compared to human perception in two languages," *The Journal of the Acoustical Society of America*, vol. 154, no. 1, pp. 95–107, 2023.

[20] B. Karthikeyan, D. J. George, G. Manikandan, and T. Tony, "A Comparative Study on K-Means Clustering and Agglomerative Hierarchical Clustering," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 5, pp. 1600–1604, 2020.

[21] A. Gupta, H. Sharma, and A. Akhtar, "A comparative analysis of k-means and hierarchical clustering," *EPRA International Journal of Multidisciplinary Research (IJMR)*, pp. 412–418, 2021.

[22] O. A. Abbas, "Comparisons between data clustering algorithms." *International Arab Journal of Information Technology (IAJIT)*, vol. 5, no. 3, 2008.

[23] H. I. Abdalla, "A Brief Comparison of K-means and Agglomerative Hierarchical Clustering Algorithms on Small Datasets," in *Proceeding of 2021 International Conference on Wireless Communications, Networking and Applications*, Z. Qian, M. Jabbar, and X. Li, Eds. Singapore: Springer Nature Singapore, 2022, pp. 623–632, series Title: Lecture Notes in Electrical Engineering.

[24] A. Javed, B. S. Lee, and D. M. Rizzo, "A benchmark study on time series clustering," *Machine Learning with Applications*, vol. 1, p. 100001, 2020.

[25] R Core Team, "R: the R project for statistical computing," 2022, version 4.2.1.

[26] R Studio Team, "RStudio: Integrated Development for R," 2022.

[27] C. Genolini, X. Alacoque, M. Sentenac, and C. Arnaud, "kml and kml3d: R Packages to Cluster Longitudinal Data," *Journal of Statistical Software*, vol. 65, no. 4, pp. 1–34, 2015.

[28] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Applied Statistics*, vol. 28, no. 1, p. 100, 1979.

[29] J. C. Gower, "Some distance properties of latent root and vector methods used in multivariate analysis," *Biometrika*, vol. 53, no. 3-4, pp. 325–338, 1966.

[30] D. J. Hermes, "Measuring the Perceptual Similarity of Pitch Contours," *Journal of Speech, Language, and Hearing Research*, vol. 41, no. 1, pp. 73–82, 1998.

[31] C. Kaland and T. M. Ellison, "Evaluating cluster analysis on f0 contours: an information theoretic approach on three languages," in *Proceedings of the 20th International Congress of Phonetic Sciences*, R. Skarnitzl and J. Volín, Eds. Prague (CZ): Guarant International, 2023, pp. 3448–3452.