Data Clustering Techniques for Gender Recognition by Voice and Speech Analysis

José Andrés Carvajal Bautista
Department of Mathematical Sciences
Mathematical Engineering
Universidad EAFIT
Medellín, Colombia
jacarvajab@eafit.edu.co

David Andrés Romero Millán
Department of Mathematical Sciences
Mathematical Engineering
Universidad EAFIT
Medellín, Colombia
dromero1@eafit.edu.co

***Abstract*—This work applies five different clustering techniques (Mountain Clustering, Subtractive Clustering, K-means, Fuzzy C-means, and Hierarchical Clustering) with several norms (Euclidean, Manhattan, Infinity, and Mahalanobis) to a *Voice Gender* dataset that was built to distinguish a voice as male or female, based upon their speech's acoustic properties. Case in point, the workflow builds two additional spaces, a high-dimensional space with polynomial features and a 2-dimensional space using Barnes Hut t-SNE algorithm, to assess which exploration and learning setup renders the most suitable clustering properties as measured by several internal validation indexes. Results show that clustering results are highly dependant on the elected space-norm combination. In addition, evidence shows that the proposed polynomial transformation does not furnish the expected benefits. Contrastingly, the embedding procedure seems to induce new partitions in the dataset that might not be related to the ground truth.**

I. Introduction

Detecting the gender (male or female) of a voice signal poses a challenging task to computerized systems. In contrast to humans, computer programs do not have the inherent capacity to extract distinguishable features from voice samples to derive personal attributes such as gender, nationality, dialect, emotion, age, or language fluency. Undoubtedly, there is a significant social need to classify gender accurately even in conditions of deprived or less than impeccable sensible input. In fact, gender classification is, in many cases, a trivial problem for the human brain as people learn over time to classify males and females from peculiarities like pitch, timbre, frequency, and breathiness [1].

Gender plays a significant role in the day-to-day lives of the members of society. Given the rise of voice recognition applications, gender classification has arisen as a notable subject of examination as it increases the interpretability of voice signals [2]. Being that the case, gender identification has various potential use-cases. In surveillance, voice gender perception might help enhance security systems' ability to unmask criminals hiding their identity [3]. Such a technique could be used to reduce search efforts to speed up critical investigations. Additionally, in digital marketing, gender classification can provide indicative data to afford customized services to the users of a platform.

This work provides new insights regarding how different clustering techniques could be used to group gender-labeled voice samples. The idea is to apply five different clustering algorithms with several norms in three different dimensions to assert how each arrangement procures different quality attributes.

II. State of the Art

To date, several publications have investigated the application of artificial intelligence techniques for gender classification in voice samples [2], [4], [5]. For instance, the work in [2] proposes a stacked gender classification algorithm using the acoustic parameters of a set of voice samples. In fact, this publication introduces a scheme that mixes classification trees (CTs), support vector machines (SVMs), and neural networks (NNs) and ensembles them together to derive sounder classification metrics. Verily, the scholars secure a 96.74% accuracy score. However, the authors admit that stacking an SVM did not procure notable benefits.

A similar approach was followed in [6] to achieve the same goal. In this case, 3000 voice samples in `.WAV` format are pre-processed by extracting 22 acoustic properties of each signal. The authors implement a Multilayer Perceptron Network (MLP) with one input layer, four hidden layers, and one output layer. For validation purposes, a 5-fold cross-validation setup was implemented to obtain an average classification score. The scholars use a softmax activation function and a 0.25 dropout at each hidden layer. Positively, the authors obtain a 97.64% accuracy score. Nonetheless, the researchers claim that a larger dataset is needed to minimize incorrect classifications.

Lastly, a comparable strategy is implemented in [4] with a different methodology. In this instance, 3168 American English voice samples are used to develop a Deeper Long Short Term Memory (LSTM) network that is able to achieve a 98.4% accuracy rate. Likewise, 20 features are extracted from each sample to build a 20-dimensional dataset. Remarkably, the authors apply several Relief-based methods to obtain the top 10 features of the dataset. In addition, the scholars admit that more data is needed to achieve even better results.

Overall, these studies highlight the interest in developing new methods to increase gender classification accuracy in

audio samples. These investigations indicate that many machine learning algorithms have been proven effective in many related scenarios. Despite that, these papers suggest that the problem is far from being fully resolved. Given all that has been mentioned so far, some common flaws were detected, mainly concern with the exclusive use of euclidean norms. Around that, as mentioned beforehand, this work will assess the effects of using various norms in three speech spaces with different dimensions.

III. Methods

As mentioned earlier, five data clustering techniques were implemented to analyze how recorded voice samples are grouped using different unsupervised methods with various norms. The following subsections present the task's specifications in greater detail.

A. Dataset

The *Voice Gender* dataset was built to distinguish a voice as male or female, based upon their speech's acoustic properties. In particular, the dataset consists of 3,168 recorded voice samples collected from male and female talkers with no further details regarding age, language, or any other distinctive features [7]. Moreover, its author pre-processed the beforementioned instances using acoustic analysis in the human vocal range (0 Hz - 280 Hz). Besides, very few details are given of the pre-processing phase. Notwithstanding, the scholar mentions she measured 20 acoustic parameters in each sample using the `WarbleR` R package.

B. Features

The following 20 acoustic properties of each voice sample are available for analysis:

- *meanfreq*: mean frequency (in kHz)
- *sd*: standard deviation of frequency
- *median*: median frequency (in kHz)
- *Q25*: first quantile (in kHz)
- *Q75*: third quantile (in kHz)
- *IQR*: interquantile range (in kHz)
- *skew*: skewness
- *kurt*: kurtosis
- *sp.ent*: spectral entropy
- *sfm*: spectral flatness
- *mode*: mode frequency
- *centroid*: frequency centroid
- *peakf*: peak frequency (frequency with highest energy)
- *meanfun*: average of fundamental frequency measured across acoustic signal
- *minfun*: minimum fundamental frequency measured across acoustic signal
- *maxfun*: maximum fundamental frequency measured across acoustic signal
- *meandom*: average of dominant frequency measured across acoustic signal
- *mindom*: minimum of dominant frequency measured across acoustic signal
- *maxdom*: maximum of dominant frequency measured across acoustic signal
- *dfrange*: range of dominant frequency measured across acoustic signal

C. Statistical Methods

A significant concern with the examined dataset is that it biases the modeler to select two prototypes for clustering. However, there are no additional classes in the dataset to perform a more in-depth analysis. Furthermore, we cannot assure the soundness of the labeling procedure. Nevertheless, one might perform some simple statistical tests to verify if any other categories may exist within each group.

The proposed approach is described as follows. First, a *one-way analysis of variance* is performed on the mean frequency feature to assert males and females are drawn from populations with different means. Later, under the assumption that each group's acoustic frequency comes from the same distribution, each class is set aside to inspect if the top and bottom half points comply with that premise. For that purpose, the non-parametric *Kruskal-Wallis* test is used to inquire whether both samples originate from the same unknown distribution. Finally, the p-value of each group is inspected to check if the null hypothesis is rejected at a 5% significance level. If that is the case, we would have gathered sufficient evidence to show that at least one other unspecified class exists in the dataset.

D. Workflow

The proposed pipeline for the clustering task includes the following sequence of steps. Firstly, the dataset is pre-processed using conventional scaling and cleaning techniques. Such procedure procures a 20-dimensional feature space in the unit hypercube. Verily, normalization is deemed particularly useful in many machine learning algorithms [8]. Indeed, it reduces the chance of a particular feature governing over the others in the context of the objective function [9]; which implies that the method transforms each component so that it contributes proportionately to its relative distances.

Secondly, the original space is transformed to obtain a higher-dimensional Hilbert space. Feature augmentation is expected to make the classes easier to set apart, thus increasing the possibility of better groupings. In addition, the technique is regarded as being considerably powerful for improving the performance in many applications such as image and text classification [10]. Thirdly, the **Barnes Hut t-SNE** algorithm is applied to reduce the original space's dimension and explore how the clustering methods operate in that scenario. The aforementioned approach is credited as an exceptional procedure for visualizing high-dimensional data [11].

In the fourth step, once the three spaces are built, two exploratory clustering methods, **Mountain** and **Subtractive**, are used to estimate the best number of clusters to use in each case according to the validation indexes. In particular, an estimation will be made to each space-distance pair since several norms are being examined. Later, having each setup, three clustering algorithms, **K-means**, **Fuzzy C-means**, and

Hierarchical Clustering are run several times to learn a fixed number of prototypes associated with each space-distance setup. Finally, results are compared to determine which space and distance combination derives the most suitable clustering according to the several internal and external evaluation metrics.

E. Pre-processing and Feature Augmentation

As mentioned earlier, pre-processing is made by scaling each feature to the $[0, 1]$ interval. Such a state is achieved by applying Equation 1.

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad \forall i = 1, \ldots, n \tag{1}$$

where, $x = (x_1, \ldots, x_n)$ is the feature vector and z_i is the i^{th} normalized data point. No supplementary cleaning techniques are used since the dataset was found in excellent conditions.

Turning now to the procurement of the higher-dimensional space, some problems were detected. Bear in mind that the dataset was initially pre-processed; hence only summarized statistics from the speech signal were given. That meant that the original recordings were not available, making it unrealistic to apply *Fourier* or *Wavelet* transforms. Without the source, we could not use signal processing techniques.

That being the case, a more straight-forward approach was followed. For convenience, we implemented a polynomial transformation up to second power. That meant that polynomial features were extracted from the initial 20 features. Each new characteristic is, in fact, the first or second power of one of the originals or any of its interactions. More precisely, the resulting dataset is 231-dimensional.

F. Validation Metrics

Four validation metrics are used to assess the three fundamental properties of each clustering experiment: its compacity, its connectedness, and spatial separation [12]. Remarkably, three internal indexes and one external index are implemented. Namely, the *Silhouette*, *Davies-Bouldin*, *Calinski-Harabasz*, and *Purity* figures are employed to determine each algorithm's quality concerning its ability to generate non-trivial partitions and to identify existing groups.

The *Silhouette* index captures the consistency inside clusters of data [13]. Unquestionably, it is computed as the maximum mean silhouette coefficient $S(i)$ over all the data points for a specific number of clusters k. As it happens, the silhouette value of a data point measures, in the -1 to 1 range, how similar that object is to its cluster compared to other clusters [14]. Values close to 1 reveal that each object significantly resembles its group and differs meaningfully from the other neighboring partitions. Furthermore, Equations 2 and 3 present its specifications, where $a(i)$ is the mean distance between point i and all the other data points in the same cluster, and $b(i)$ is the smallest mean distance from i to all the points in any other cluster.

$$SI = \max_i \tilde{s}(i) \tag{2}$$

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases} \tag{3}$$

The *Davies-Bouldin* index is an internal evaluation metric defined as the ratio of within-cluster scatter and between-cluster separation [15]. An ideal configuration yields a low value in the numerator (high intra-cluster similarity) and a high value in the denominator (low inter-cluster similarity). Having that, the algorithm that procures a clustering scheme with the smallest Davies-Bouldin index is held as the most suitable arrangement. Moreover, Equation 4 presents its formulation, where k is the number of clusters, c_i is the i^{th} centroid, and μ_i is the mean distance of all elements in cluster i to c_i.

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\mu_i + \mu_j}{d(c_i, c_j)} \right) \tag{4}$$

The *Calinski-Harabasz Criterion* is determined as the weighted ratio of overall between-cluster variance SS_B and within-cluster variance SS_W [16]. Well-defined clusters have a sizeable between-cluster variance and a small within-cluster variance. The larger the Calinski-Harabasz ratio, the better the data partition. Besides, Equations 5, 6, and 7 present its expression, where n_i is the number of items in the i^{th} cluster, and m is the mean of the sample.

$$CH = \frac{SS_B}{SS_W} \times \frac{(n-k)}{(k-1)} \tag{5}$$

$$SS_B = \sum_{i=1}^{k} n_i \left\| c_i - m \right\|^2 \tag{6}$$

$$SS_W = \sum_{i=1}^{k} \sum_{x \in C_i} \left\| x - c_i \right\|^2 \tag{7}$$

Lastly, the *Purity* index is an external evaluation criterion of cluster quality that measures the degree to which clusters contain a single class. Also, it can be interpreted as the percent of data points that were classified correctly [17]. A correct classification implies that each cluster C_i has identified a group of objects as the same class that the ground truth has indicated. In general, higher Purity values are favored in a particular clustering experiment. However, this index does not penalize a large number of clusters. In addition to this, Equation 8 showcases its formulation, where D is the set of ground truth classes.

$$PU = \frac{1}{n} \sum_{i=1}^{k} \max_{d \in D} |C_i \cap d| \tag{8}$$

IV. Results

This section presents the results of the proposed workflow in the *Voice Gender* dataset. In addition, the same methodology is applied to the well-known *Iris* dataset by means of validating each procedure.

A. Statistical Methods

The *one-way analysis of variance* is performed on the mean frequency feature with MATLAB's `anova1` function. Indeed, a p-value of 0 is obtained, suggesting that males' and females' mean frequencies are significantly different. Evidence indicates that males have a mean frequency around 0.1708 in contrast to women that hold one closer to 0.1910.

Regarding the same feature, the male samples are split up into two groups, the upper half and the lower half. Under the assumption that all male speeches come from the same frequency distribution, we hope that applying a *Kruskal-Wallis* homogeneity test yields both halves originate from the same distribution. However, a p-value of 0 reveals that the premise is not valid, and truly, not every male sample comes from the same unknown frequency distribution. The evidence implies that at least one additional division exists within males data points.

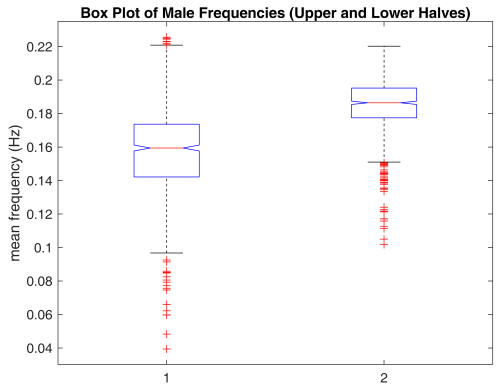

Figure 1: Box Plot of Male Frequencies (Upper and Lower Halves)

The same procedure is employed in the female group delivering the same conclusions. Further, Figures 1 and 2 present the corresponding box plots of male and female homogeneousness in both halves. Therefore, different clustering methods must be implemented to determine if any other partitions might exist in the dataset.

B. Iris' Exploration

As stated before, the *Mountain* and *Subtractive Clustering* algorithms are used to resolve an adequate number of clusters to use in each space-norm combination. Such a procedure involves exploring the parameter space of both methods to affirm which setup is most favorable in each case. Figures 3, 4, 5, and 6 showcase the effect of tuning the parameters with the Euclidean norm in the Iris 4-d dataset.

In this case, as suggested in [18], the r_b parameter is taken as 1.5 times r_a, meaning that only one parameter is calibrated in the Subtractive Clustering scenario. From the available evidence, it is clear that the highest Silhouette and Calinski-Harabasz indexes are obtained with $\alpha = 0.2$ and $\beta = 1$ under Mountain Clustering using the Euclidean norm. On the other

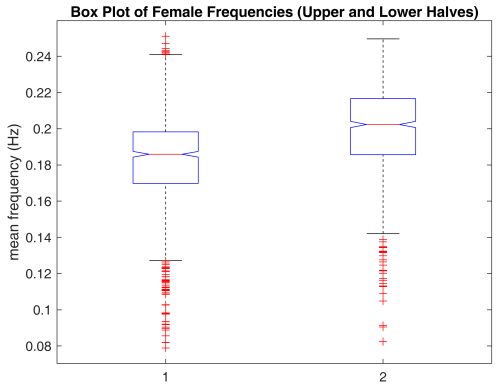

Figure 2: Box Plot of Female Frequencies (Upper and Lower Halves)

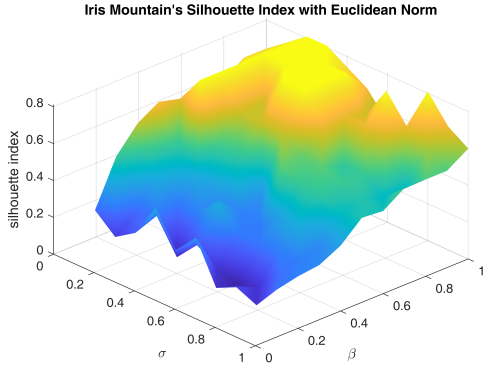

Figure 3: Iris 4-d Mountain's Silhouette Index with Euclidean Norm (Parameters)

hand, Figures 5 and 6 reveal no agreement concerning the three internal validation indexes in the Subtractive Clustering setting. In particular, Davies-Bouldin and Silhouette coincide in a $r_a = 0.1$ selection. Nonetheless, Calinski-Harabasz suggests it could be $r_a = 1$. As expected, internal validation indexes do not always concur in their results since they measure different clustering properties. For instance, Calinski-Harabasz's disagreement might be related to several factors such as features having different variances, between-cluster variance increasing nearby $r_a = 1$, or within-cluster variance dropping close to the same value.

Table I presents the identified most fitting parameter selections for Iris' space-norm combinations. Additional figures are not included to facilitate the reader's comprehension. Results suggest that the selected space-norm setup has a meaningful impact on the most suitable clustering parameters. This determination was made by weighting the validation indexes with a particular interest in consistency and similarity. Evidence hints that a combination of lower α values with higher β figures yield the most promising clustering properties when applying the Mountain Algorithm to the three different representations of Iris' dataset. Notwithstanding, a similar pattern is not

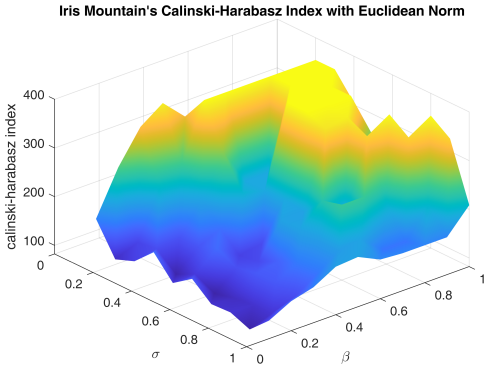

Figure 4: Iris 4-d Mountain's Calinski-Harabasz Index with Euclidean Norm (Parameters)

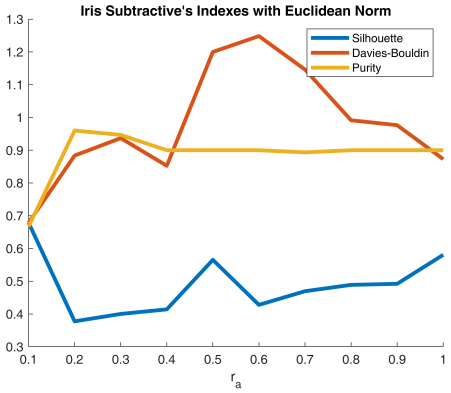

Figure 5: Iris 4-d Subtractive's Indexes with Euclidean Norm (Parameters)

detected concerning the r_a parameter. Moreover, no selections were made for the α and β parameters in the high-dimensional space because the Mountain algorithm requires vast amounts of computational space to build an n-dimensional grid of points. Several experiments were performed in those scenarios, and every single one crashed due to memory constraints.

Having identified each algorithm's parameters, the number of prototypes for each space-distance setting must be settled. Such a decision is made by running both algorithms and

Table I: Most Favorable Parameters for Iris' Exploration

Space	Norm	α	β	r_a
Original (4-d)	euclidean	0.2	1.0	0.1
	manhattan	0.4	1.0	0.4
	infinity	0.1	1.0	1.0
	mahalanobis	0.6	0.5	0.1
Higher-Dimension (15-d)	euclidean	-	-	0.8
	manhattan	-	-	0.1
	infinity	-	-	1.0
	mahalanobis	-	-	0.9
Embedded (2-d)	euclidean	0.1	1.0	1.0
	manhattan	0.3	1.0	0.2
	infinity	0.1	1.0	1.0
	mahalanobis	0.1	1.0	0.4

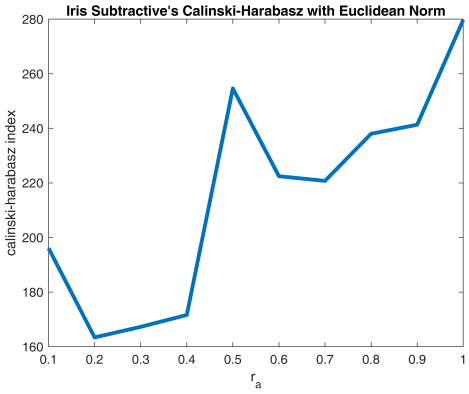

Figure 6: Iris 4-d Subtractive's Calinski-Harabasz Index with Euclidean Norm (Parameters)

grouping the measured validation metrics by the number of obtained clusters. For instance, Figures 7, 8, 9, and 10 display the index figures gathered with different prototypes in the 4-dimensional Iris dataset using the Euclidean norm.

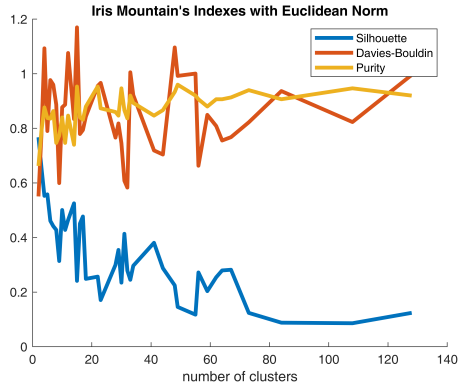

Figure 7: Iris 4-d Mountain's Silhouette Index with Euclidean Norm (Number of Prototypes)

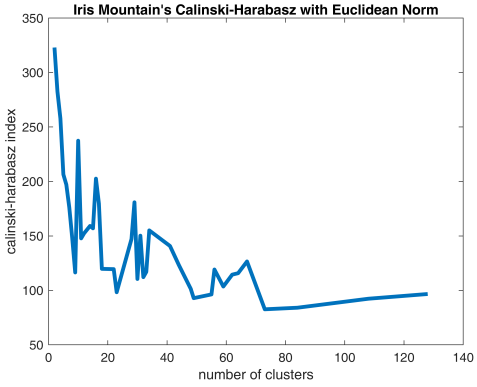

Figure 8: Iris 4-d Mountain's Calinski-Harabasz Index with Euclidean Norm (Number of Prototypes)

Again, as foreseen, the clustering metrics do not fully agree

with the ideal number of clusters to use in this situation. If we consider all the available information, we get that the 4-dimensional dataset could be grouped into 2, 3, or 4 classes. As a result, the former set's median is applied to make a fair and robust choice, and a 3-partition is picked. Coincidentally, three is also the number of classes in the ground truth. What is more, this same procedure is employed on the twelve space-norm combinations. For that purpose, Table II shows the number of clusters procured for each setup.

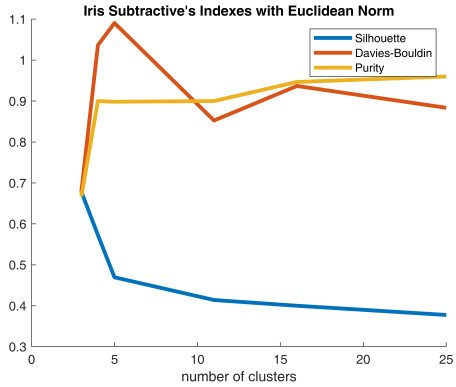

Figure 9: Iris 4-d Subtractive's Indexes with Euclidean Norm (Number of Prototypes)

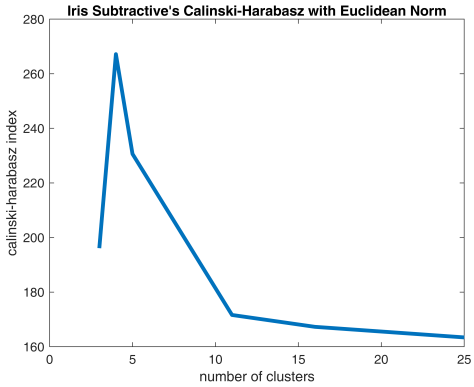

Figure 10: Iris 4-d Subtractive's Calinski-Harabasz Index with Euclidean Norm (Number of Prototypes)

Table II: Number of Clusters obtained in Iris' Exploration

Space	Norm	k
Original (4-d)	euclidean	3
	manhattan	5
	infinity	3
	mahalanobis	3
Higher-Dimension (15-d)	euclidean	4
	manhattan	9
	infinity	3
	mahalanobis	4
Embedded (2-d)	euclidean	2
	manhattan	4
	infinity	2
	mahalanobis	14

Table II's figures exhibit substantial discrepancies in the number of elected prototypes. For instance, the data points out that the number of clusters highly depends on the examined space and norm. As anticipated, not all distances have the same performance in each space. Remarkably, the Euclidean and Infinity norms are the ones that seem closer to the ground truth. Contrastingly, the Manhattan and Mahalanobis norms do not seem to perform appropriately in the 15-d and 2-d spaces, respectively. In summary, transforming the space through augmenting it or reducing it provides a new perspective on the number partitions within the original dataset. Thereupon, such operations are presumed to influence significantly the end outcome of the clustering task.

C. Voice Gender's Exploration

By the same means, the exploration of the Voice Gender dataset is conducted to ascertain the most favorable set of parameters and a reasonable number of clusters for each space-norm setup. Figures 11, 12, 13, and 14 showcase the effect of calibrating the parameters with the Euclidean norm in the 20-d dataset.

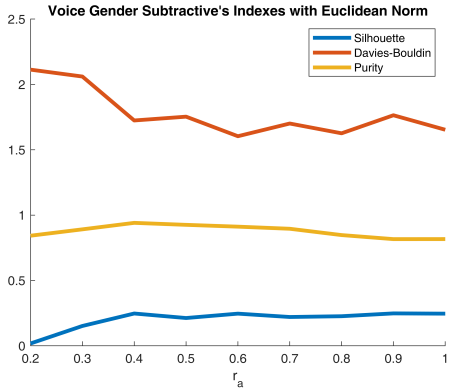

Figure 11: Voice Gender 20-d Subtractive's Indexes with Euclidean Norm (Parameters)

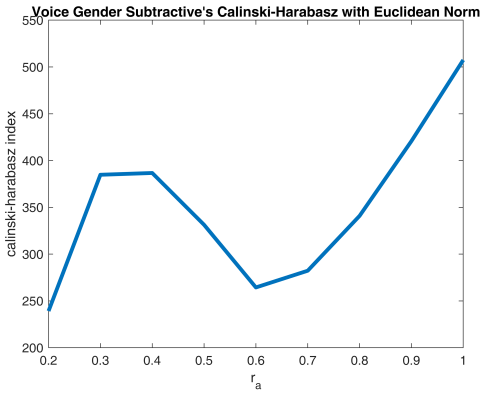

Figure 12: Voice Gender 20-d Subtractive's Calinski-Harabasz Index with Euclidean Norm (Parameters)

Keep in mind that the Mountain Clustering algorithm cannot be applied satisfactorily in higher-dimensions due to memory

constraints. Therefore, the examination in the 20-d and 231-d spaces are only delivered with the Subtractive scheme. Case in point, Figures 11 and 12 agree that an $r_a = 1$ assignment leads to more attractive clustering properties with the Euclidean norm in the original space.

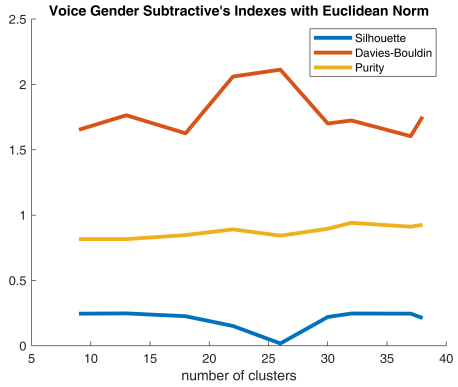

Figure 13: Voice Gender 20-d Subtractive's Indexes with Euclidean Norm (Number of Prototypes)

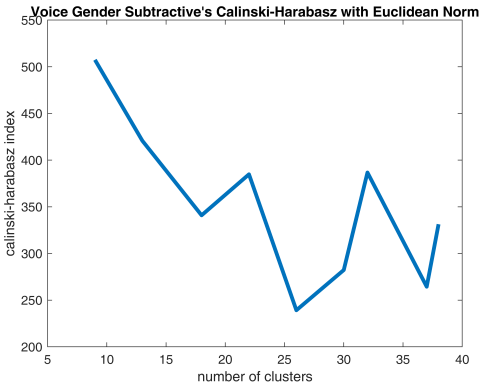

Figure 14: Voice Gender 20-d Subtractive's Calinski-Harabasz Index with Euclidean Norm (Number of Prototypes)

As expected, the assessment is replicated with each combination and Table III is attained. What stands out in the table is that some distances did not procure a valid clustering setup. Namely, the Mahalanobis and Manhattan norms seem to have difficulties assembling congruent partitions in the 20-d and 231-d spaces. To be specific, the ? value is assigned if at least one of the validation metrics exhibits abnormal behavior.

Further evidence is needed to confirm if either larger or smaller radiuses are preferred, since proof attests that might be closely linked to the experiment's settings. Another interesting fact is that the 20-d and 231-d samples have ill-conditioned covariance matrices, which adds up to the issue of finding reliable results with the Mahalanobis distance.

Having finished the parameter identification procedure, we now focus on establishing the number of prototypes to use in each space-norm setup. As mentioned before, such a judgment is made by running the exploration algorithms several times and aggregating the metrics by the number of derived clusters. If the internal validation metrics do not match, the median of the number of groups associated with each index is selected. For example, Figures 13 and 14 show the index figures collected with different prototypes in the 20-dimensional Voice Gender dataset using the Euclidean norm.

Table III: Most Favorable Parameters for Voice Gender's Exploration

Space	Norm	α	β	r_a
Original (20-d)	euclidean	-	-	1.0
	manhattan	-	-	0.4
	infinity	-	-	0.8
	mahalanobis	-	-	?
Higher-Dimension (231-d)	euclidean	-	-	0.8
	manhattan	-	-	?
	infinity	-	-	0.7
	mahalanobis	-	-	?
Embedded (2-d)	euclidean	0.5	1.0	0.1
	manhattan	1.0	0.1	0.1
	infinity	0.1	0.2	0.1
	mahalanobis	0.2	0.7	0.4

For illustrative purposes, Figures 13 and 14 show that a 9-cluster assignment is pertinent for that scenario, as the three indexes agree on the same spot. Though, it is striking that near the 25-prototype mark, things seem to go sideways concerning inner-cluster consistency. These operations must be analyzed in greater detail in future endeavors.

The specified procedure is extended to the twelve space-norm combinations as unfurled in Table IV. Similar to Iris' results, the three spaces show contrasting figures verifying that space transformations and the norm selection are decisive towards the clustering outcomes. From this data, we can see that the Manhattan distance in 20-d and the Infinity norm in 231-d procure the same number of classes designated in the ground truth. Still, one cannot guarantee that is the correct number of partitions since the statistical analysis revealed that at least one additional class might exist within each group. Furthermore, the embedded space's results exhibit a divergent situation. Evidence shows that the Barnes Hut t-SNE algorithm transmuted the dataset so that a larger number of clusters are distinguished.

Table IV: Number of Clusters obtained in Voice Gender's Exploration

Space	Norm	k
Original (20-d)	euclidean	9
	manhattan	2
	infinity	4
	mahalanobis	-
Higher-Dimension (231-d)	euclidean	6
	manhattan	-
	infinity	2
	mahalanobis	-
Embedded (2-d)	euclidean	16
	manhattan	32
	infinity	13
	mahalanobis	33

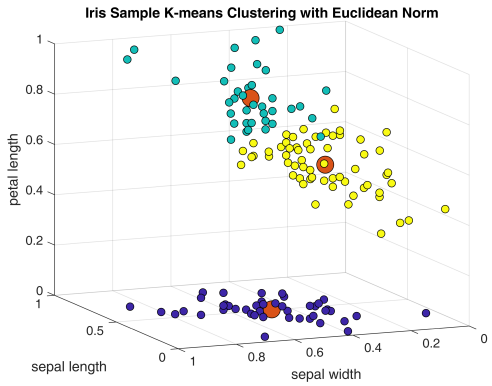

Figure 15: Iris 4-d K-means Clustering with Euclidean Norm

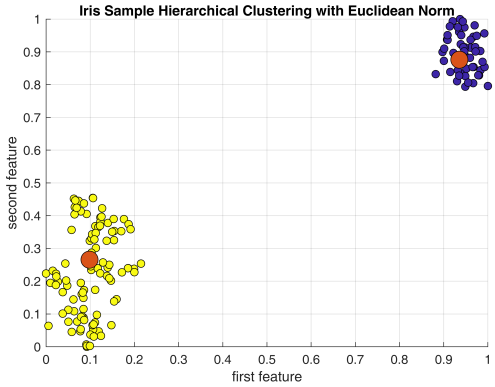

Figure 16: Iris 2-d Hierarchical Clustering with Euclidean Norm

D. Iris's Learning

Having completed the exploration phase, three clustering algorithms are employed to obtain the prototypes that partition each space-norm experiment. Namely, K-means, Fuzzy C-means, and Hierarchical Clustering are run several times to learn the fixed number of clusters associated with each combination. At each run, the four validation metrics are computed to obtain a mean estimation of each quality attribute.

Figures 15 and 16 display two sample clustering outcomes with Iris' 4-d and 2-d spaces with the Euclidean norm. As expected, both algorithms perform satisfactorily in each case, suggesting that the clustering algorithms were implemented appropriately. Surprisingly, the two samples show how the Barnes Hut t-SNE algorithm's application leads to a linearly separable space where the three species are morphed into two classes.

The complete set of Iris' Clustering results is presented in Table V. Evidence reveals that the best clustering properties are manifested in the Embedded 2-d space with the Euclidean and Infinity norms. As a matter of fact, these space-norm setups happen to shape two clusters with high consistency, low within-cluster scatter, high between-cluster separation, and a sizeable ratio of inter-cluster and intra-cluster variances. However, recall from Table II that these settings are clustered with two prototypes, which differs from the 3-clusters ground truth.

Interestingly, the Mahalanobis distance seems to perform poorly in the three reference spaces, in some cases even exhibiting negative silhouette figures. Surprisingly, there is no sufficient evidence to claim that the feature augmentation operation derived better outcomes than the original space in the internal validation indexes. Not to mention, that the lousier results are attached to the Mahalanobis distance in the 15-d space setting. Again, as foreseen, the higher purity values are linked to the setups with the highest number of prototypes. Yet, the classification accuracy, without incorporating Mahalanobis' outcomes, places in between 0.70 and 0.85. Moreover, the most striking result to emerge from the Iris' clustering results is that the Fuzzy C-means algorithm is, on average, less effective than the other two.

E. Voice Gender's Learning

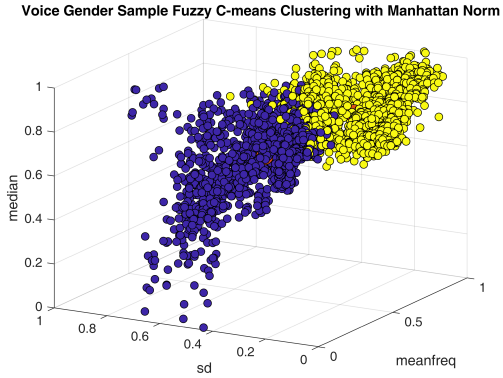

Figure 17: Voice Gender 20-d Fuzzy C-means Clustering with Manhattan Norm

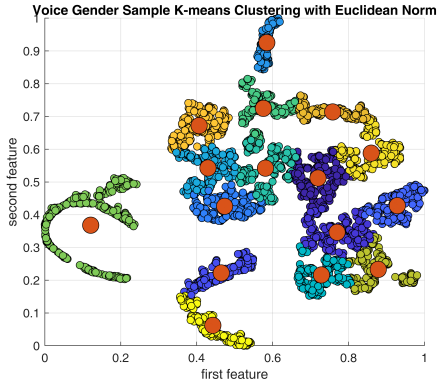

Figure 18: Voice Gender 2-d K-means Clustering with Euclidean Norm

Having finished the Voice Gender dataset exploration process, an identical learning scheme is followed. Similarly, the three reference clustering methods are employed to group the

Table V: Iris's Clustering Results

Space	Norm	Method	SI	DB	CH	PU
Original (4-d)	euclidean	K-means	0.6654	0.7388	53.7568	0.7778
		FC-means	0.6376	0.7276	54.2062	0.7739
		Hierarchical	0.6427	0.7300	53.7568	0.7704
Original (4-d)	manhattan	K-means	0.5197	0.8220	41.8386	0.8033
		FC-means	0.4481	0.9328	39.7962	0.8012
		Hierarchical	0.4797	0.8214	41.8386	0.7896
Original (4-d)	infinity	K-means	0.6172	0.8091	49.2057	0.7771
		FC-means	0.5972	0.7706	50.1577	0.7767
		Hierarchical	0.6120	0.6976	49.2057	0.7433
Original (4-d)	mahalanobis	K-means	0.2414	1.8697	21.5260	0.6987
		FC-means	0.2550	1.8808	19.0136	0.7063
		Hierarchical	-0.1085	2.5110	21.5260	0.5484
Higher-Dimension (15-d)	euclidean	K-means	0.5964	0.7412	44.9668	0.7850
		FC-means	0.5970	0.7539	49.3602	0.7945
		Hierarchical	0.5642	0.7100	44.9668	0.7900
Higher-Dimension (15-d)	manhattan	K-means	0.4354	0.7699	38.8244	0.8461
		FC-means	0.3789	0.8686	35.0771	0.8352
		Hierarchical	0.4512	0.7421	38.8244	0.8246
Higher-Dimension (15-d)	infinity	K-means	0.6518	0.6869	47.7055	0.7650
		FC-means	0.5809	0.7478	43.8382	0.7673
		Hierarchical	0.5380	0.7416	47.7055	0.7296
Higher-Dimension (15-d)	mahalanobis	K-means	-0.1186	3.3969	2.1105	0.5556
		FC-means	-0.2017	3.6072	1.8208	0.5734
		Hierarchical	-0.1593	4.5434	2.1105	0.5146
Embedded (2-d)	euclidean	K-means	0.9877	0.1164	910.5302	0.7588
		FC-means	0.9858	0.1195	816.9741	0.7531
		Hierarchical	0.9870	0.1182	910.5302	0.7521
Embedded (2-d)	manhattan	K-means	0.7513	0.5212	857.3871	0.7883
		FC-means	0.6801	0.6390	875.7389	0.7973
		Hierarchical	0.7553	0.5122	857.3871	0.8004
Embedded (2-d)	infinity	K-means	0.9874	0.1166	860.6034	0.7550
		FC-means	0.9874	0.1169	855.7925	0.7543
		Hierarchical	0.9871	0.1178	860.6034	0.7554
Embedded (2-d)	mahalanobis	K-means	0.3886	0.8262	585.4084	0.8333
		FC-means	0.4356	0.8057	649.6557	0.8202
		Hierarchical	0.3375	0.9012	585.4084	0.8421

data points in each space-distance setting. Furthermore, the same three mean internal validation indexes and one external validation figure are estimated with Monte Carlo simulation.

Figures 17 and 18 display two sample clustering outcomes with Voice Gender's 20-d and 2-d spaces with the Manhattan and Euclidean norm, respectively. As envisioned, both algorithms perform competently in each case, implying that the clustering algorithms do work in a non-trivial dataset. Despite this, there are notable differences in each case. Take for example Figure 17, where the Fuzzy C-means algorithm is able to split the dataset into two groups, presumably most men in blue and most women in yellow. Contrastingly, as specified in Table IV, the embedding procedure induced a considerable amount of non-elementary partitions, as signaled in Figure 18.

The complete set of Voice Gender's Clustering results is presented in Table VI. Evidence points out that the most favorable clustering properties are associated with the Embedded 2-d space under the Mahalanobis distance. Nevertheless, that does not mean the clustering properties are ideal. To illustrate, see that the table's maximum silhouette value is around 0.6088, which is not necessarily bad but is unquestionably far from the 1 paragon. In other words, on average, the obtained clustering patterns are not entirely consistent, which means that some objects within a given cluster are not amply similar to each other. In contrast, a maximum Calinski-Harabasz index of 838.83 reveals that the previously identified best setup can render low within-cluster and high between cluster variances. In summary, the best settings regarding the internal validation indexes build clusters with low intra-cluster scatter. Yet, those groups are not sufficiently distant from their neighboring partitions, as exhibited in Figure 18.

Although, the Embedded 2-d space secures the most solid validation indexes, that certainly does not indicate this is the best transformation. Recall Table IV, where you can observe that the minimum number of clusters in this space is 13. In effect, such a number is the farthest from the 2-cluster ground truth. Furthermore, one cannot choose which clustering method performs best in all situations, as the data shows that it is highly subordinate to the learning setup. But then again, on average, the Fuzzy C-means algorithm reveals substandard performance as reflected by low Silhouette values and high Davies-Bouldin figures. Not to mention, that there is not sufficient evidence to judge if the polynomial transformation furnished benefits concerning the evaluated metrics.

V. Conclusions

This study set out to ascertain the effects of applying five clustering algorithms in three speech spaces with different dimensions (20-d, 231-d, and 2-d) while inducing several

Table VI: Voice Gender's Clustering Results

Space	Norm	Method	SI	DB	CH	PU
Original (20-d)	euclidean	K-means	0.3029	1.5383	119.1938	0.8495
		FC-means	-0.0675	2.3295	61.9660	0.7719
		Hierarchical	0.1628	1.6048	119.1938	0.7862
Original (20-d)	manhattan	K-means	0.4884	1.2918	268.2836	0.6453
		FC-means	0.4578	1.3078	264.0528	0.6907
		Hierarchical	0.5243	1.1076	268.2836	0.6221
Original (20-d)	infinity	K-means	0.2168	1.9622	135.9601	0.7966
		FC-means	-0.0022	2.4452	95.9358	0.7165
		Hierarchical	0.0923	1.6967	135.9601	0.6604
Higher-Dimension (231-d)	euclidean	K-means	0.3140	1.5824	123.4260	0.8510
		FC-means	-0.0246	2.2314	75.1230	0.7726
		Hierarchical	0.1825	1.7130	123.4260	0.7759
Higher-Dimension (231-d)	infinity	K-means	0.3861	1.5685	176.9732	0.6479
		FC-means	0.2895	1.9907	129.9227	0.6560
		Hierarchical	0.3102	2.2310	176.9732	0.6304
Embedded (2-d)	euclidean	K-means	0.6088	0.7398	693.4495	0.7314
		FC-means	0.2276	0.9369	300.8345	0.7057
		Hierarchical	0.6067	0.6909	693.4495	0.7298
Embedded (2-d)	manhattan	K-means	0.5945	0.7164	823.5428	0.7442
		FC-means	-0.0661	2.3621	239.2519	0.6978
		Hierarchical	0.5664	0.7082	823.5428	0.7511
Embedded (2-d)	infinity	K-means	0.5899	0.7477	632.7916	0.7287
		FC-means	0.0939	1.4162	269.9736	0.7063
		Hierarchical	0.5515	0.7295	632.7916	0.7333
Embedded (2-d)	mahalanobis	K-means	0.6065	0.7051	838.8363	0.7446
		FC-means	0.2436	0.9288	325.1271	0.7113
		Hierarchical	0.6079	0.6821	838.8363	0.7519

norms. In particular, this paper has argued that clustering results are highly dependant on the elected exploration or learning setup.

An opening insight is that the Mountain Clustering algorithm has significant limitations regarding the computational space required to build the n-dimensional grid. Such constraints make the method above impractical in many real-life situations where a broad set of features must be probed. Later, evidence showed under the evaluated scenarios that not all the examined norms, specifically Mahalonobis and Manhattan, can put together a proper clustering arrangement. In particular, the ill-conditioned property of the covariance matrix is undoubtedly a setback while employing the former.

A central still at the same time naive finding is that internal validation metrics do not always agree with their results. As explained before, each index measures different quality attributes that may or may not match in their judgments. The intuition above is quite noteworthy since the results might trick the modeler into selecting an incorrect clustering setting. For instance, in both datasets, the Embedding transformation derived dominant outcomes concerning the validation metrics. However, as duly noted, that same alteration leads to groupings, in some cases quite distinct from the ground truth.

The learning procedure in the twelve space-norm combinations prompted essential verdicts. In particular, records showed that K-means, Fuzzy C-means, and Hierarchical Clustering could not fabricate a set of deeply consistent clusters as evinced through the Silhouette figures. The characteristics of each space appeared to favor a high ratio of intra-cluster and inter-cluster variances. Simultaneously, a striking fact is that the Fuzzy C-means algorithm performed poorly on average compared to the other two. Thus, hinting that each clustering algorithm might have some particular set of conditions where it performs at its best.

In conclusion, evidence shows that the proposed polynomial transformation does not furnish the expected benefits. Contrastingly, the embedding procedure went the other way around and induced new partitions in the dataset that might not be related to the ground truth. Unquestionably, employing clustering algorithms in different dimensions helps derive fresh perspectives about the structure, shape, relationships, and complexity of the source data. Even though the proposed transformations did not provide optimal outcomes, they did afford new valuable intuitions about the examined data.

REFERENCES

[1] S. Watson, "The unheard female voice: Women are more likely to be talked over and unheeded. but slps can help them speak up and be heard.," 2019.

[2] P. Gupta, S. Goel, and A. Purwar, "A stacked technique for gender recognition through voice," in *2018 Eleventh International Conference on Contemporary Computing (IC3)*, IEEE, 2018, pp. 1–3.

[3] M. Demirkus, K. Garg, and S. Guler, "Automated person categorization for video surveillance using soft biometrics," in *Biometric Technology for Human Identification VII*, International Society for Optics and Photonics, vol. 7667, 2010, 76670P.

[4] F. Ertam, "An effective gender recognition approach using voice data via deeper lstm networks," *Applied Acoustics*, vol. 156, pp. 351–358, 2019.

[5] S. Lakra, J. Singh, and A. K. Singh, "Automated pitch-based gender recognition using an adaptive neuro-fuzzy inference system," in *2013 International Conference on Intelligent Systems and Signal Processing (ISSP)*, IEEE, 2013, pp. 82–86.

[6] M. Buyukyilmaz and A. O. Cibikdiken, "Voice gender recognition using deep learning," in *2016 International Conference on Modeling, Simulation and Optimization Technologies and Applications (MSOTA2016)*, Atlantis Press, 2016, pp. 409–411.

[7] K. Becker, *Identifying the gender of a voice using machine learning*, https : / / www . kaggle . com / primaryobjects/voicegender, 2016.

[8] A. B. Graf and S. Borer, "Normalization in support vector machines," in *Joint pattern recognition symposium*, Springer, 2001, pp. 277–282.

[9] J.-M. Jo, "Effectiveness of normalization pre-processing of big data to the machine learning performance," *The Journal of the Korea institute of electronic communication sciences*, vol. 14, no. 3, pp. 547–552, 2019.

[10] S. Wu, H. Zhang, G. Valiant, and C. Ré, "On the generalization effects of linear transformations in data augmentation," in *International Conference on Machine Learning*, PMLR, 2020, pp. 10 410–10 420.

[11] L. Van Der Maaten, "Accelerating t-sne using tree-based algorithms," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221–3245, 2014.

[12] J. Handl, J. Knowles, and D. B. Kell, "Computational cluster validation in post-genomic data analysis," *Bioinformatics*, vol. 21, no. 15, pp. 3201–3212, 2005.

[13] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009, vol. 344.

[14] T. Thinsungnoena, N. Kaoungkub, P. Durongdumronchaib, K. Kerdprasopb, and N. Kerdprasopb, "The clustering validity with silhouette and sum of squared errors," *learning*, vol. 3, no. 7, 2015.

[15] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[16] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[17] D. Marutho, S. H. Handaka, E. Wijaya, *et al.*, "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news," in *2018 International Seminar on Application for Technology of Information and Communication*, IEEE, 2018, pp. 533–538.

[18] K. Hammouda and F. Karray, "A comparative study of data clustering techniques," *University of Waterloo, Ontario, Canada*, vol. 1, 2000.