Analysis of Audio and Lyric Variation Metrics

Julia Stelman

11/11/2019

Introduction

This is kind of a spin-off of the investigation done in "A Brief Analysis of Lyric Metrics." We take the same 100-song sample used in that study. The question we examine here is how do the songs from widely spoken languages vary in comparison to songs from non-widely spoken languages?

In this chapter, I'll look at the spreads of a few audio and lyric features from 100 songs in 5 different languages (20 songs written in each language). I want to know if languages spoken across a wider geographical area produce a wider range of patterns in audio features and song lyrics. In other words, do the songs written by, say, Spanish-speaking artists vary more than songs written by, say, Indonesian-speaking artists? less?

The following two languages make up the *widely spoken* group:

- English (en)
- Spanish (es)

The following three languages make up the *not widely spoken* group:

- Indonesian (id)
- Dutch (nl)
- Swahili (sw)

The distinction between the first and second groups is decided by whether a language is an official national language on at least three continents.

I've used the natural language processing package `quanteda` to calculate the following audio/ lyric features for the first few lines of lyrics in each of the 100 sampled songs. (20 songs were randomly sampled from each of the 5 languages.)

- Danceability (number between 0 and 1)
- Energy (number between 0 and 1)
- Key (category represented as a number between 0 and 11)*
- Loudness (number between -25 and 0)
- Speechiness (number between 0 and 1)
- Acousticness (number between 0 and 1)
- Valence (number between 0 and 1)**
- Tempo (number between 50 and 250)
- Type/Token Ratio, or TTR –> (# distinct words / # total words) in the first few lines of lyrics

*Each of the 12 numbers corresponds to one of the 12 keys on the chromatic scale, (e.g. C, C#, D, …, A, A#, B).

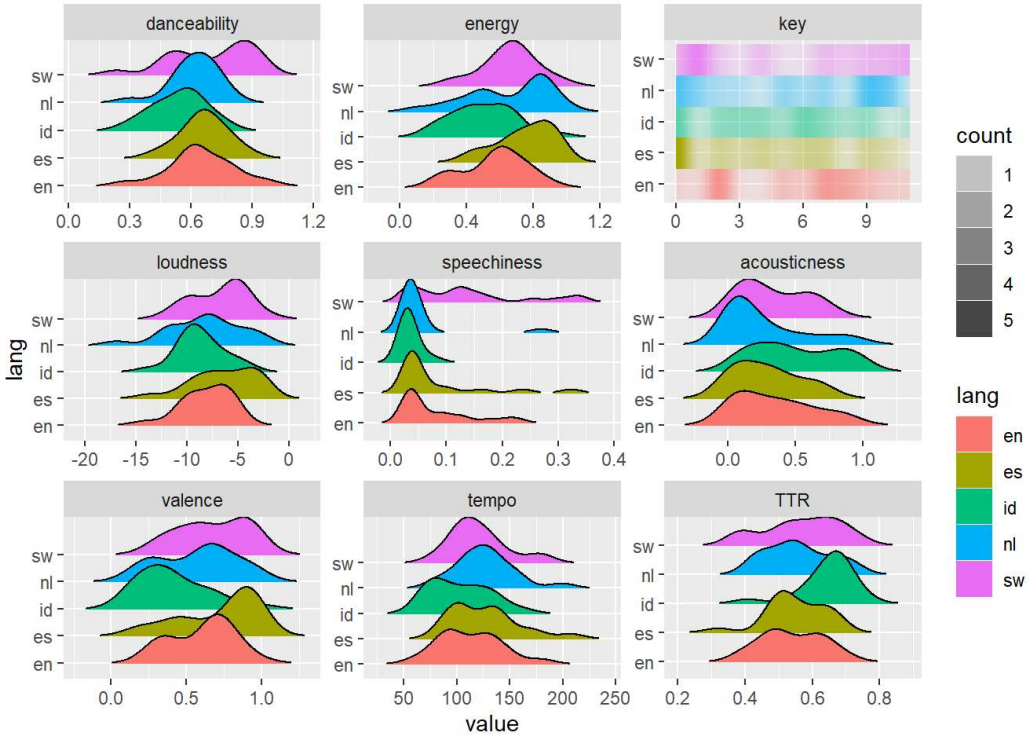
**I'm not entirely sure how Spotify measures valence, but I have the understanding that is an overall measure of how "happy" a song decidedly is based on other features.

EDA

Density plots

The plots below show the density curves, with each color-coded by language. For every measure besides Key, the density curves show us something about center, shape, and spread of their distributions among the songs of each language.

Because key is actually a categorical variable, and not a quantitative variable, I used a different type of density plot for Key. Each key on the chromatic scale is associated with a certain value, 0-11, featured on the x-axis. The darker the shading at a value of x, the more songs written in the key associated with that x-value turned up in the sample. The idea is not to look for similarities in the locations of dark and light shading, but rather to compare and contrast the distribution of pigmentation across whole strips. This helps us get a better look at spread, which is all we are interested in, as shape and center don't have the same kind of meaning when dealing with a categorical variable as with quantitative variables.



In the Key plot (top right), the pink strip seems to be the splotchiest-looking. This means that there appears to be less variation of song key among songs written in English than among songs written in other languages.

It would seem that the distributions of many features of English (pink) and Spanish (olive) music have similar shapes (danceability, loudness, speechiness, acousticness, tempo, and TTR). For all of these but loudness, these distributions seem to have commonalities not just of shape, but also of center and spread (ignoring abnormal tail behavior).

Swahili's distribution (purple) always tends to march to the beat of its own drum (terrible pun intended). Truthfully though, if you look at all nine of the plots, there doesn't seem to be one in which the purple curve (or strip) seems to mimic the distribution of another color.

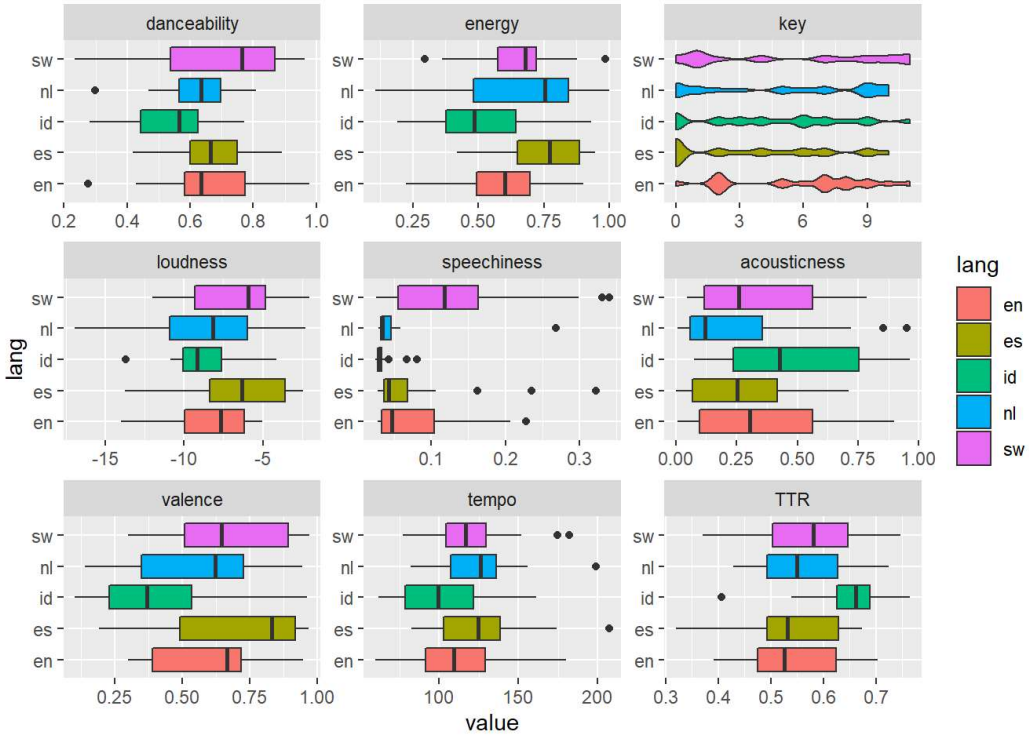
What's interesting about Indonesian music (teal) is that acousticness is basically its only bimodally distributed feature. All its other features are distributed pretty unimodally, which is unusual. Meanwhile, every other language's music has at least two features that are distributed bimodally.

All in all, there doesn't seem to be a super clear unifying difference between the more widely-spoken and less widely-spoken languages.

Box plots

The boxplots below show the five-number-summaries and outliers for all of the same data that was used in the density curves above. Key, the categorical feature, is portrayed in a violin plot instead of a boxplot so that it is easier to see the concentrations at individual values of key.

```
## Coordinate system already present. Adding new coordinate system, which will replace the existing one.
```



There are lots of outliers in speechiness, which, while prominent in the density curves, is even more prominent here, especially for Indonesian.

There don't seem to be any consistent patterns. As far as what these plots portray, there is no clear relationship between song feature distribution characteristics and widespokenness.

Methods

We want to know whether or not there is a difference between extent of variation within the music of more widely spoken languages vs music of less widely spoken languages. We're going to approach exploring this in two ways.

- Running a statistical hypothesis test
- Performing K-means clustering

Rank Sum tests

Rank Sum tests are usually not used in the way I am using them here. The Wilcoxon Rank Sum test is designed to be run on a data set whose observations are all in the same units. In order to deal with this restriction, I have to take some intermediate steps.

I'll first calculate the spread metric (standard deviation or IQR) for each of the nine features for each of the 5 languages (5 * 9 = 45 calculations). Second, within each of the nine features separately, I'll perform ranking on the spread metric. Third, I'll pool together those *ranks* (each will be a value 1-5) to form the set of "observations" on which the Wilcoxon Rank Sum Test will be performed. I'll be giving each one of the 45 "observations" a label specifying which of the two contributing "populations" (widely spoken/ not widely spoken) it pertains to. During the Rank Sum test, each one will get a new overall rank based on where it falls relative to the rest of the "observations".

Those "observations" originating from data of English or Spanish music pertain to the **widely spoken** "population", and those "observations" originating from data of Indonesian, Dutch, or Swahili music pertain to the **not widely spoken** "population."

Kmeans Clustering

Unlike the Wilcoxon Rank Sum test, Kmeans, a clustering method, is built to handle feature-rich data. However, it is not built for computing how much evidence exists in favor of the clustering formation hypothesized. It only determines which of all the possible clustering formations has the most evidence in favor of it. Therefore I will take some additional steps to interpret the results of Kmeans as I would the outcome of a statistical test.

There are $\binom{5}{0} + \binom{5}{1} + \binom{5}{2} = 1 + 5 + 10 = 16$ possible ways that Kmeans could split 5 languages into exactly two groups.

Assuming equal chance of each outcome exists when there is no relationship between varied-ness in audio/ lyric features and how widely spoken a language is, each outcome has exactly a 0.0625 chance of occurring. In other words, under the null hypothesis of no relationship, 0.0625 is the probability that the clustering procedure groups the languages in the following way:

- English (en), Spanish (es)
- Indonesian (id), Dutch (nl), Swahili (sw)

I'll take alpha to be 0.1. This way, if the pattern outlined above results, our p-value of 0.0625 will be small enough to allow us to reject the null hypothesis of no relationship.

I won't go into more detail on that, but I hope this has been enough to convince anyone reading this that some level of statistical thinking has gone into designing the unusual method I'll use in this segment of my analysis.

The need-to-know is this:

H_0: There is *no relationship* between varied-ness in audio/ lyric features and how widely spoken a language is.

H_a: There is a *relationship* between varied-ness in audio/ lyric features and how widely spoken a language is.

A statistical test has two possible outcomes: either the null gets rejected, thus supporting some alternative suspicion, or it doesn't.

If the clustering procedure groups the languages in the following way:

- English (en), Spanish (es)
- Indonesian (id), Dutch (nl), Swahili (sw)

then we reject the null.

If the clustering procedure groups the languages in any other way, then we fail to to reject the null.

Results

Rank Sum tests

We'll perform two Wilcoxon Rank Sum tests. To start, we will use standard deviation as a metric of spread. Following that, we will repeat the same process but with IQR as our spread metric.

For the more detail on how spread metrics of Key were calculated, see the Appendix.

Standard Deviation Rank Sum test

The first of the following tables contains the standard deviation of each feature for each individual language. The second table replaces the standard deviations with their ranks.

Step 1: Calculate the standard deviation for each of the nine features for each of the five individual languages.

lang	danceability	energy	key	loudness	speechiness	acousticness	valence	tempo	TTR	widely spoken?
en	0.1607261	0.1869324	3.183428	2.302330	0.0624885	0.2843769	0.1994177	29.55932	0.0912833	yes
es	0.1144002	0.1642343	3.495862	3.106435	0.0781860	0.2361951	0.2600962	31.87556	0.0862066	yes
id	0.1279928	0.1893246	3.346640	2.234726	0.0142092	0.3069545	0.2251435	27.92545	0.0766881	no
nl	0.1131354	0.2469996	3.923948	3.621842	0.0525505	0.2937826	0.2515147	27.10288	0.0943007	no
sw	0.2002997	0.1653345	4.259973	2.684877	0.1007523	0.2423239	0.2232064	26.67659	0.1134366	no

Step 2: Rank the five languages, within each feature separately, by standard deviation.

lang	danceability	energy	key	loudness	speechiness	acousticness	valence	tempo	TTR	widely spoken?
en	4	3	1	2	3	3	1	4	3	yes
es	2	1	3	4	4	1	5	5	2	yes
id	3	4	2	1	1	5	3	3	1	no
nl	1	5	4	5	2	4	4	2	4	no
sw	5	2	5	3	5	2	2	1	5	no

Step 3: Take the values from Table 2 to be the set of "observations" on which the Wilcoxon Rank Sum Test is to be performed. Determine each one's individual *overall* rank. Keep track of which "observations" belong to which "population."

obs_value	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4

obs_rank	14	14	14	14	14	14	14	14	14	23	23	23	23	23	23	23	23	32	32	32	32	32	32	32	32			
widely spoken?	yes	yes	yes							yes	yes	yes	yes	yes				yes	yes	yes	yes							
				no	no	no	no	no	no						no	no	no	no					no	no	no	no			

obs_value	4	5	5	5	5	5	5	5	5	5	1	1	1	1	1	1	1	1	1									
obs_rank	32	41	41	41	41	41	41	41	41	41	5	5	5	5	5	5	5	5	5									
widely spoken?		yes	yes								yes	yes	yes	yes														
	no			no	no	no	no	no	no	no					no	no	no	no	no									

Above are the values used as observations, their overall ranks, and whether or not they each belong to the widely spoken language group. They're ordered by overall rank.

Now let's run the test to see if the variation behavior of songs' audio/ lyric features (as can be detected at the standard deviation level) is associated with how widely spoken the language of their lyrics:

Alternative Hypothesis: The features of music written in more widely spoken languages will vary to either a greater or a lesser extent than the features of music written in less widely spoken languages.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  rank_sd by cat
## W = 270, p-value = 0.531
## alternative hypothesis: true location shift is not equal to 0
```

As 0.513 is a very large p-value, it seems there is not evidence of a unidirectional relationship between varied-ness within audio/ lyric features and how widely spoken a language is. Or at least there is not evidence of such a relationship that is detectable on the standard deviation level.

IQR Rank Sum test

The first of the following tables contains the IQR of each feature for each individual language. The second table replaces the IQRs with their ranks.

Step 1: Calculate the IQR for each of the nine features for each of the five individual languages.

lang	danceability	energy	key	loudness	speechiness	acousticness	valence	tempo	TTR	widely spoken?
en	0.19150	0.20275	0.0208333	3.80000	0.071275	0.464075	0.32800	37.51150	0.1496629	yes
es	0.14950	0.23650	0.0833333	4.76075	0.031775	0.349625	0.42675	35.80675	0.1355790	yes
id	0.18025	0.26900	0.0000000	2.40550	0.005200	0.516000	0.30625	43.10125	0.0617610	no
nl	0.13375	0.36300	0.0000000	4.94500	0.013475	0.296925	0.37800	28.96450	0.1350869	no
sw	0.32975	0.14675	0.0208333	4.46025	0.108750	0.447300	0.38325	25.14725	0.1429793	no

Step 2: Rank the five languages, within each feature separately, by IQR.

lang	danceability	energy	key	loudness	speechiness	acousticness	valence	tempo	TTR	widely spoken?
en	4	2	3.5	2	4	4	2	4	5	yes
es	2	3	5.0	4	3	2	5	3	3	yes
id	3	4	1.5	1	1	5	1	5	1	no
nl	1	5	1.5	5	2	1	3	2	2	no
sw	5	1	3.5	3	5	3	4	1	4	no

Now let's run a Wilcoxon Rank Sum test on the Inter-quartile ranges:

Step 3: Take the values from Table 2 to be the set of "observations" on which the Wilcoxon Rank Sum Test is to be performed. Determine each one's individual *overall* rank. Keep track of which "observations" belong to which "population."

obs_value	1	1	1	1	1	1	1	1	1.5	1.5	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3		
obs_rank	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	9.5	9.5	14.5	14.5	14.5	14.5	14.5	14.5	14.5	14.5	22.5	22.5	22.5	22.5	22.5	22.5	22.5	22.5		
widely spoken?											yes	yes	yes	yes	yes				yes	yes	yes	yes						
	no	no	no	no	no	no	no	no	no	no						no	no	no					no	no	no	no		

obs_value	3.5	3.5	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5
obs_rank	27.5	27.5	32.5	32.5	32.5	32.5	32.5	32.5	32.5	32.5	41	41	41	41	41	41	41	41	41
widely spoken?	yes	no	yes	yes	yes	yes	yes	no	no	no	yes	yes	yes	no	no	no	no	no	no

Now let's run the test to see if the variation behavior of songs' audio/ lyric features (as can be detected at the IQR level) is associated with how widely spoken the language of their lyrics:

Alternative Hypothesis: The features of music written in more widely spoken languages will vary to either a greater or a lesser extent than the features of music written in less widely spoken languages.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  rank_iqr by cat
## W = 181.5, p-value = 0.1513
## alternative hypothesis: true location shift is not equal to 0
```

It's a smaller p-value than we got from standard deviation. Still, 0.1513 is too large of a p-value to provide evidence supporting the existence of a one-way spectrum of variablity in audio/ lyric features associated with how widely spoken a language is.

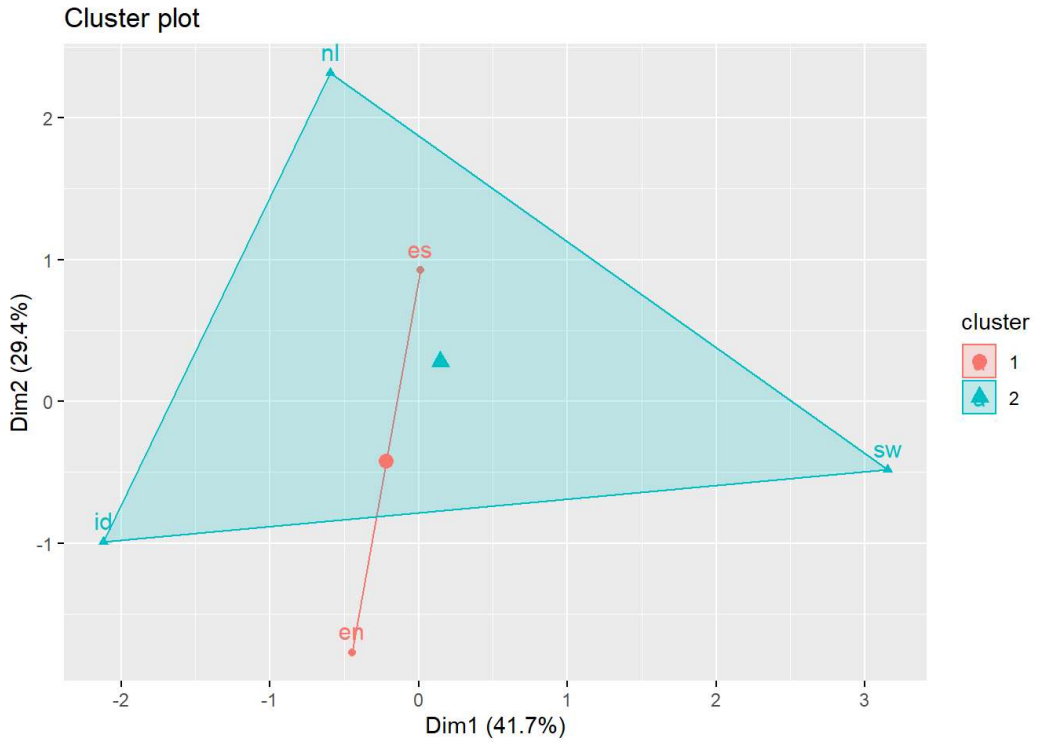
Based on these results, there is *no relationship* between varied-ness in audio/ lyric features and wide-spokenness of lyric language.

Kmeans Clustering

I used the `cluster` and `factoextra` packages to conduct Kmeans clustering analysis, first by Standard Deviation, and then by IQR.

Kmeans Clustering by Standard Deviations

```
## K-means clustering with 2 clusters of sizes 2, 3
##
## Cluster means:
##   danceability    energy      key loudness speechiness acousticness
## 1    0.1375631 0.1755834 3.339645 2.704382  0.07033725    0.2602860
## 2    0.1471426 0.2005529 3.843520 2.847148  0.05583733    0.2810203
##     valence    tempo        TTR
## 1 0.2297569 30.71744 0.08874493
## 2 0.2332882 27.23497 0.09480847
##
## Clustering vector:
## en es id nl sw
##  1  1  2  2  2
##
## Within cluster sum of squares by cluster:
## [1] 3.059045 2.249467
##  (between_SS / total_SS =  73.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```



The clusters actually split the languages into widely and not widely spoken languages! We reject the null and conclude that there is evidence of a relationship between varied-ness in audio/ lyric features and how widely spoken a language is at the standard deviation level!

Take a look at the numbers in Table 5 which is a reformated version of the table shown in the R output chunk above. The row containing cluster means for the cluster which was formed around English and Spanish is labeled “widely spoken.” Seven times out of nine, the features of music from widely spoken languages varied *less*, on average, than features of music from “not widely spoken” languages, judged on the basis of standard deviation.

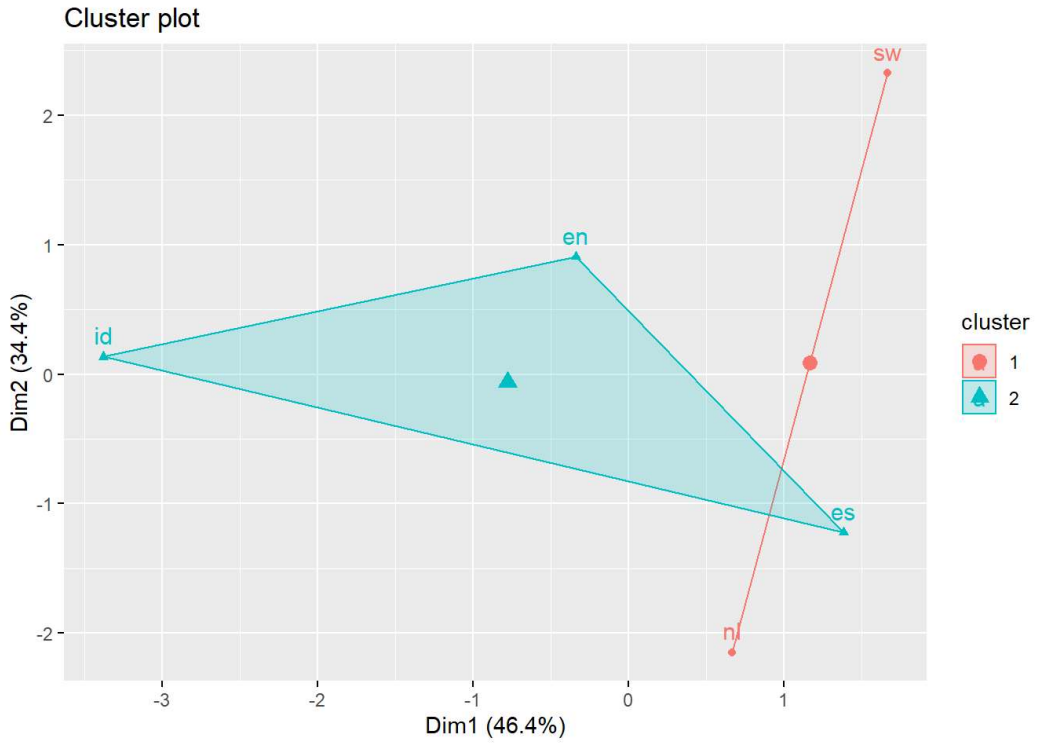
Table 5: Cluster means from Kmeans Clustering on Standard Deviation. Max of each column in bold.

	danceability	**energy**	**key**	**loudness**	**speechiness**	**acousticness**	**valence**	**tempo**	**TTR**
widely spoken	0.138	0.176	3.34	2.704	**0.07**	0.26	0.23	**30.717**	0.089
not widely spoken	**0.147**	**0.201**	**3.844**	**2.847**	0.056	**0.281**	**0.233**	27.235	**0.095**

Kmeans Clustering by IQRs

Let’s try it again with IQR instead of standard deviation.

```
## K-means clustering with 2 clusters of sizes 2, 3
##
## Cluster means:
##   danceability    energy        key loudness speechiness acousticness
## 1      0.23175 0.2548750 0.01041667 4.702625  0.06111250    0.3721125
## 2      0.17375 0.2360833 0.03472222 3.655417  0.03608333    0.4432333
##     valence    tempo       TTR
## 1 0.3806250 27.05588 0.1390331
## 2 0.3536667 38.80650 0.1156676
##
## Clustering vector:
## en es id nl sw
##  2  2  2  1  1
##
## Within cluster sum of squares by cluster:
## [1]  7.461887 31.961671
##  (between_SS / total_SS =  80.9 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```



The algorithm did not split the data into the clusters we suspected. Therefore, we fail to reject the null and conclude that there is not evidence of relationship between varied-ness in audio/ lyric features and how widely spoken a language is at the IQR level.

Discussion

Standard deviation is not outlier-resistant, which is one way IQR differs from it. Standard Deviation is also more inflated by bimodal distribution behavior than IQR is, and as we saw in those density curves, there were quite a few cases of bimodality. I found it interesting that the null was rejected in the standard deviation version of the Kmeans test but not in the IQR version. Yet, in the Rank Sum tests, though in neither version of the test was the null rejected, the standard deviation version resulted in a p-value that was, by far, the least rejectable (0.531).

I have a theory for why this would be. Because ranks were used instead of the original values in the Rank Sum tests, magnitude of difference was not able to have any influence. However, magnitude of difference *was* able to influence the Kmeans test. This difference between Rank Sum test and Kmeans test didn't seem to change much in the case of IQR. However, it made a huge difference in the case of standard deviation. Standard

deviation has more moving parts than IQR, making it more likely that distances between two values in the set of 45 standard deviations would vary more than distances between two values in the set of 45 IQRs. This might be one reason why the conclusions of the two hypothesis tests for standard deviation differed.

Now that there is finally a null rejection to talk about, let's discuss what could actually be causing it. As Table 5 made apparent, for 78% of features, there was less in-language variation in music from widely spoken languages than music from less widely spoken languages.

This, I found quite interesting and unexpected. Initially, I predicited that, if anything, there would be *more* variation among songs from *more* widely spoken languages. The *more* places a language is spoken, the *more* speakers it has. The *more* speakers a language has, the *more* singers it has. The *more* singers that contribute their voices to music in a language, the *more* variation exists among music produced in it.

However, there is another side to all of this. Even if the *list* of songs I began with for this study were truly a *simple random sample* of *all* the songs from each language, I'm still limited to only the songs with:

- audio feature data on Spotify,
- lyrics on Musixmatch,
- lyrics encoded in a way that the software I'm using can process,
- encoded lyrics decoded in a way that is recognized (as non-gibberish) by the machine translation API,
- and decoded lyrics correctly detected by Google Translate.

Probably, the more commercially successful the song, the more likely it is to have remained in the sampleable dataset by the end of all this. Therefore, we can't think just about the music that gets *made* by a lot of people. We have to think also about the music that gets *heard* by a lot of people. There are certain qualities associated with songs that get popular. One of them, undoubtedly, is a wide-spoken lyric language.

I'm no expert, but I would guess that pop artists seeking to maximize their audiences don't opt to write their music in Swahili. On the other hand, those that *do* opt to write their music in Swahili probably aren't shooting for world-wide popularity above all else. Singers who write songs with pop music qualities have more incentive to write lyrics in more widely-spoken languages. Often, common goals result in common products. And if that common goal is commercial success, songs written in English or Spanish with pop music qualities are the common results.

The music that remained in the dataset by the time I began sampling was probably a mix between pop music and non-pop music. However, it probably was not an unbiased mix. Music of the pop genre had probably been lost in the metadata collection process at a lower rate than music of other genres. This probably changed the genre diversity of the sampleable data across all five language groups, but had the highest pop-saturation effect on English and Spanish.

Acknowledgements

Thank you to Spotify, Musixmatch, Everynoise.com, and lang-detect for the help I got from your libraries, websites, and APIs in the data collection and cleaning phase of this project, which I did in Python.

Also thank you to the R libraries `quanteda`, `tidyverse`, `DescTools`, `caret`, `dplyr`, `stringi`, and `knitr`, `ggplot2`, `ggridges`, `plyr`, `cluster`, `factoextra`, and `kableExtra`.

Appendix

How the spread metrics of Key were calculated

```
# turn the 0-11 key variable into 12 bernoull variables with 20 observations each
dummies <- esinw_df[order(esinw_df$lang),] %>% select(key)

# and its levels
dum.levels <- levels(factor(unlist(dummies)))
key <- lapply(dummies,
              function(x) table(sequence(nrow(dummies)), factor(x, levels = dum.levels)))$key


#### For Standard Deviation

# take the standard devation of all 12 of them, then take the mean of those 12 sds
key_sds <- sapply(0:4,function(c){
  stdevskeys = c()
  for (i in 1:12){
    stdevskeys = append(stdevskeys,
                        sd(key[(c*(20) + (1:20)),i]))}
  mean(stdevskeys)})

# store that in the key column of the standard devation data frame
sds$key <- key_sds


#### For IQR

# take the IQR of all 12 of them, then take the mean of those 12 IQRs
key_iqrs <- sapply(0:4,function(c){
  iqrkeys = c()
  for (i in 1:12){
    iqrkeys = append(iqrkeys,
                        IQR(key[(c*(20) + (1:20)),i]))
  }
  mean(iqrkeys)
  })

# store that in the key column of the IQR data frame
iqrs$key <- key_iqrs
```