Variance of Audio and Lyric 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.” I take the same 100-song sample used in that study. The question I examine here is how do the songs from widespread languages vary in comparison to songs from non-widespread languages?

In this chapter, I 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 conform more or less to 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, Dutch artists? It would make sense, as Spanish is spoken all over the world, heightening the potential that this language has been split into different dialects by divergent evolution. Meanwhile, Dutch is only spoken, officially, in the Netherlands, a region so small that it doesn’t provide the Dutch language very much potential for forming individual dialects.

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)

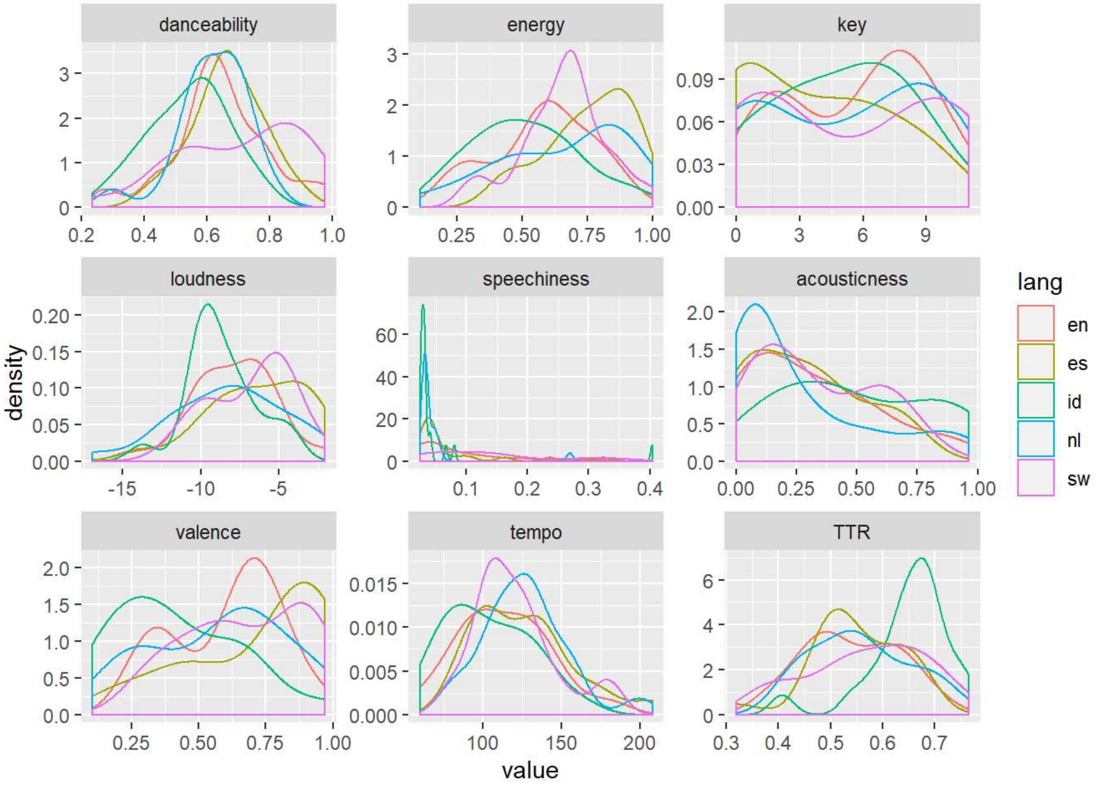
The distinction between the first and second groups is decided by whether a language is an official national language on at least two continents.

I’ve used the natural language processing package $quanteda$ to calculate the following metrics 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.)

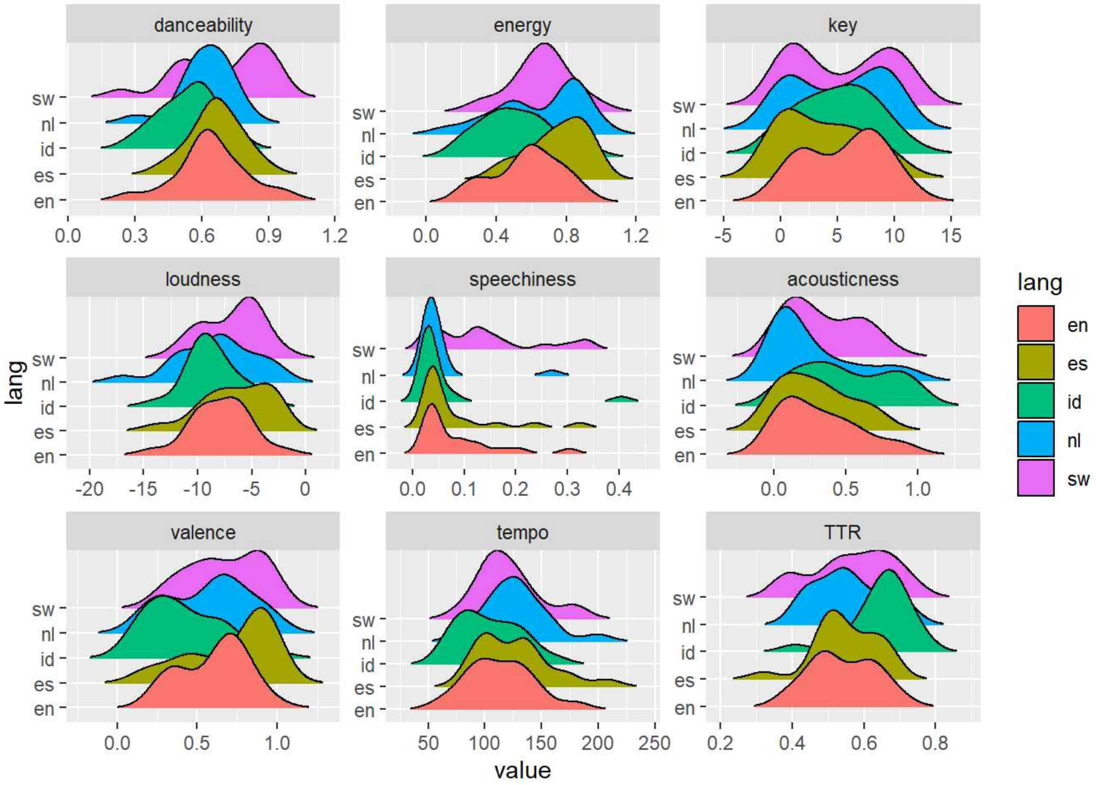
- Danceability (number between 0 and 1)
- Energy (number between 0 and 1)
- Key (number between 0 and 12)
- Loudness (number between -25 and 0)
- Speechiness (number between 0 and 1)
- Acousticness (number between 0 and 1)
- Valence (number between 0 and 1)**
- Tempo (number between 50 and 250)
- Type/Token Ratio, or TTR –> (# distinct words / # total words) in the first few lines of lyrics

*I’m not entirely sure how Spotify measures valence, but I have the understanding that is a measure of how possitively or negatively charged a song decidedly is based on other features.

The plots below show the density curves, with each color-coded by language.



There are some interesting differences between indonesian music and music in other languages (loudness, speechiness, TTR, key). Swahili music also exibits some interesting behavior (energy, danceability). Let’s get a better look…



It would seem that English and Spanish follow similar distributions in danceability, speechiness, acousticness, tempo, and TTR. Meanwhile, Swahili’s distribution always tends to march to the beat of its own drum. It’s not super clear that There is a unifying difference between the more widely-spoken and less widely-spoken languages. Let’s run a statistical test to evaluate the variances associated with each langauge.

Other observations:

- Indonesian music has is not as inclined to use the key most popular with other languages.

- Indonesian music tends to be less positively charged on average than music in other languages (valence).
-

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

lang	danceability	energy	key	loudness	speechiness	acousticness	valence	tempo	TTR	is_widespread
en	0.1573641	0.1937762	3.362330	2.605192	0.0732296	0.2695946	0.1993041	29.14349	0.0906996	TRUE
es	0.1144002	0.1642343	3.495862	3.106435	0.0781860	0.2361951	0.2600962	31.87556	0.0862066	TRUE
id	0.1276873	0.2039098	3.244428	2.254101	0.0836937	0.3118669	0.2326139	27.33171	0.0794105	FALSE
nl	0.1131354	0.2469996	3.923948	3.621842	0.0525505	0.2937826	0.2515147	27.10288	0.0943007	FALSE
sw	0.2002997	0.1653345	4.259973	2.684877	0.1007523	0.2423239	0.2232064	26.67659	0.1134366	FALSE

lang	danceability	energy	key	loudness	speechiness	acousticness	valence	tempo	TTR	is_widespread
en	4	3	2	2	2	3	1	4	3	TRUE
es	2	1	3	4	3	1	5	5	2	TRUE
id	3	4	1	1	4	5	3	3	1	FALSE
nl	1	5	4	5	1	4	4	2	4	FALSE
sw	5	2	5	3	5	2	2	1	5	FALSE

Hypothesis: the more widely-spoken languages will vary to either a greater or a lesser extent than the less widely-spoken languages.

Now let’s run a Wilcoxon Rank Sum test to see if my hypothesis is true:

```
## 
##  Wilcoxon rank sum test with continuity correction
## 
## data:  rank by cat
## W = 279, p-value = 0.4013
## alternative hypothesis: true location shift is not equal to 0
```

That’s a very large p-value, which means there is not a one-way spectrum of variablity in audio/ lyric features associated with how widely spoken a language is. Or at least there is not such a spectrum that is detectable on this level.

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, and ggridges.