Identifying Musical Similarities Across Geographical Regions

Eric Su, Michael Valancius, Andrew Cooper

Abstract

- The purpose of the project was to analyze musical similarities across geographical origins
- Audio features were extracted from 1,059 songs with minimal Western influence from around the world
- Dimension reduction was performed on these features using PCA, t-SNE, and an autoencoder
- A Gaussian Mixture Model was chosen to cluster on the reduced dimensions from the autoencoder
- Cluster analysis shows that smaller, island-like regions tend to have more isolated musical identities

Introduction

- Music is a major aspect of most cultures
- Different areas of the world develop unique musical styles
- However, as cultures mix, so do aspects of music composition
- It is unclear how and to what extent aspects of music are shared across the world
- **Can we use clustering techniques to identify musical similarities across geographical regions?**

Music Dataset

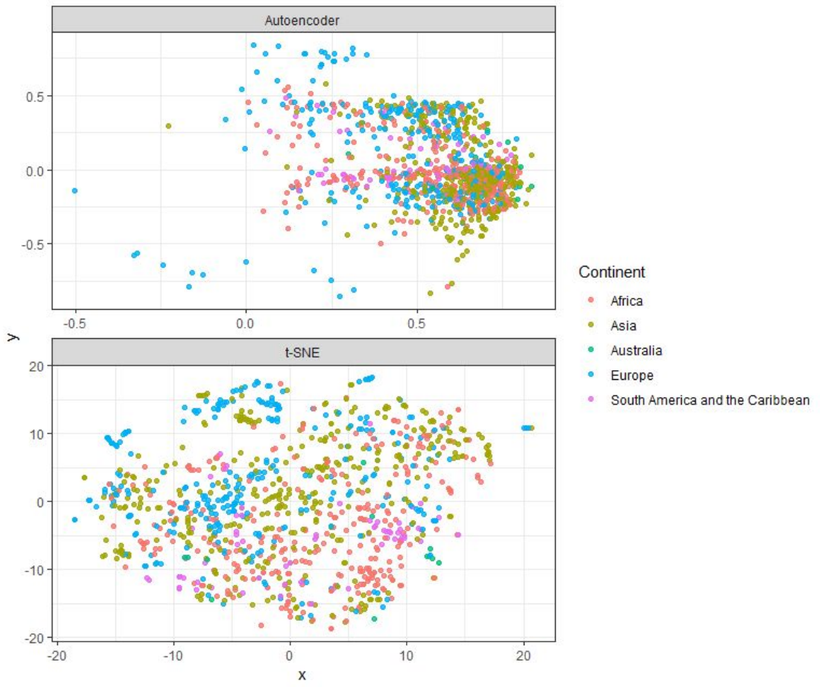
- Dataset comes from **"Predicting the Geographical Origin of Music"** (Zhao, Q, King, 2014)
- 1,059 tracks were chosen, with music originating from 73 unique countries
- Songs were processed through **MARYSAS** (Music Analysis, Retrieval and Synthesis for Audio Signals)
- Each songs contains 68 audio features with additional chromatic information

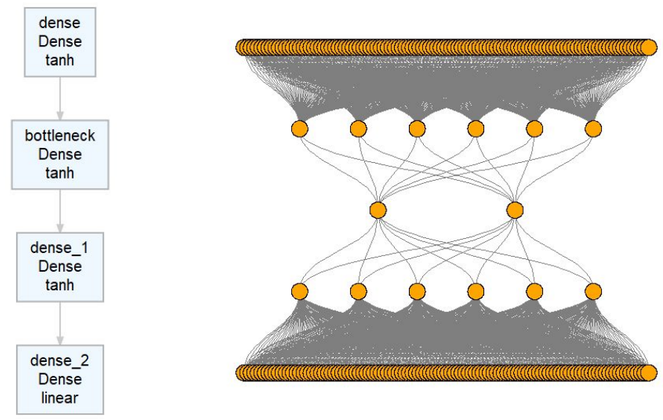
Geographic Origins of 1,059 Songs in Music Dataset

Point size indicates the number of songs from that country

Dimension Reduction

- Dimension reduction was performed to reduce audio features to most prominent components
- Three methods of dimension reduction were considered:
  a. **Principal Component Analysis (PCA)**
  b. **t-Distributed Stochastic Neighbor Embedding (t-SNE)**
  c. **Autoencoder**
- PCA struggled to handle nonlinear, spatial relationships
- While neither embedding shows distinct clusters, the autoencoder proved to have more separated clusters



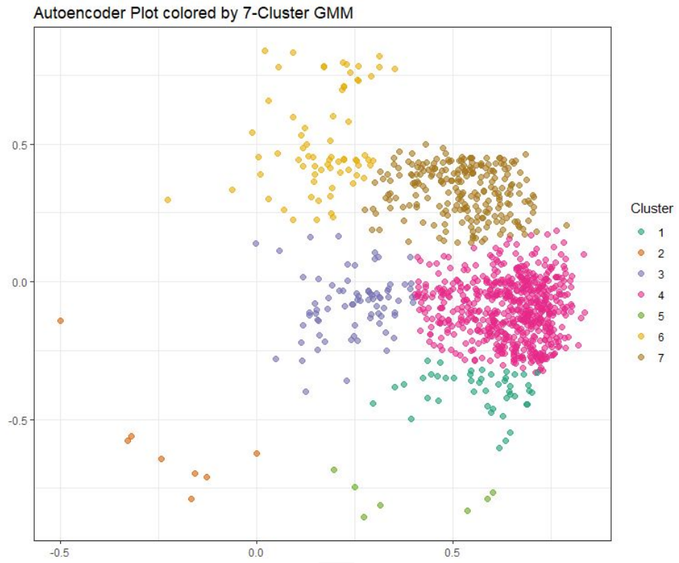


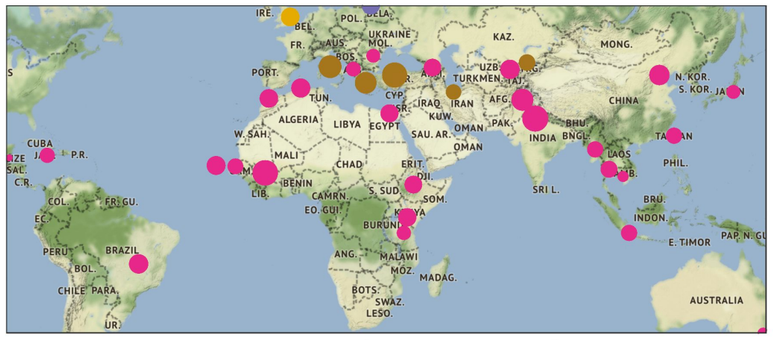
Autoencoder Architecture

Clustering Methodology

- Cluster analysis was performed on the reduced dimensionality created by the autoencoder.
- Two important questions to consider when clustering:
  a. **What method of clustering should be performed?**
  b. **How many clusters should be declared?**
- Comparison of clustering methods indicated a **Gaussian Mixture Model** best identified distinct clusters
- Cluster diagnosis showed **7** to be the most appropriate number of clusters

GMM Clustering Results





Points colored by most frequent cluster classification

- **Cluster purity**: What percentage of each country belongs to the dominant cluster?
- Higher purity level indicates a more distinct, singular music style

Table 1: Countries with Highest Purity

Country	Purity
Japan	0.95
Australia	0.86
Cambodia	0.86
Algeria	0.83
Taiwan	0.80

Table 2: Countries with Lowest Purity

Country	Purity
Iran	0.48
Morocco	0.45
Italy	0.43
Turkey	0.41
Lithuania	0.28

Conclusions

- Purity table shows countries like Japan and Cambodia have the most distinct clusters
- **Small, island-like countries tend to have the most distinct musical identities, perhaps because of the nature of their isolated geography**
- **Conversely, counties close to European trade routes tend to have the least distinct musical identities, perhaps because aspects like music can be shared more easily**
- Future work: can audio features be used to classify the continent of origin for a piece of music?
- Neural network analysis to predict country origin based on musical features