Analysis and Clustering of Musical Compositions using Melody-based Features

Isaac Caswell Erika Ji

December 13, 2013

Abstract

This paper demonstrates that melodic structure fundamentally differentiates musical genres. We use two methods: the k-Means algorithm as an unsupervised learning method for understanding how to cluster unorganized music and a Markov Chain Model which determines relative probabilities for the next note's most likely value, and evaluate our algorithms' accuracy in predicting correct genre. Our experiments indicate that the k-Means approach is modestly successful for separating out most genres, whereas the Markov Chain Model tends to be very accurate for music classification.

1 Objective

This paper demonstrates that melodic structure, i.e. note subsequences and which notes are likely to follow other notes, can fundamentally differentiate musical genres, without additional information about instrumentation, chord structure, language, etc.

This idea is inspired in part by the concept in Indian classical music that each raga, or scale, is distinguished by its own characteristic melodic phrases, or melodic idioms. This occurs in Western classical music also; for instance, consider the typical third, trilled second, root, root to end phrases in Baroque pieces in the major scale.

Potential applications of our models include: using the k-Means algorithm to define musical clusters for melodies without specified genres; using either k-Means cluster centroids or the Markov Chain Model to determine known melodies similar to a new melody, using the either technique for automated genre prediction, and using the Markov Chain Model to generate new melodies within a musical genre or tradition.

2 Data

Songs were stored as arrays of integers, with each integer representing a musical note. Rhythm was ignored. The sources of data were the following:

- 2,000 Irish traditional songs scraped from thesession.org in the Dorian, Mixolydian, Ionian (Major) and Aeolian (Minor) modes and in the time signatures 2/4, 3/4, 4/4, 6/8 and 9/8. The majority (70%) were major.
- 27 Carnatic (South Indian classical) compositions in equivalent scales to the Irish modes: Shankarabharanam (Major), Kharaharapriya (Dorian), Harikambhoji (Mixolydian), Bhairavi (Minor), and Malahari (incomplete Minor).
- Smaller data sets: a variety of children's songs and 13 Sarali Varasai (Carnatic vocal exercises) in ragam Mayamalavagowla.

All data are expressed relative to the root: the key is disregarded.

3 Methods

3.1 k-Means Clustering

3.1.1 Rationale

Given a dataset of melodies with unknown genre, can we identify which melodies are similar? To answer this question, we looked for an unsupervised learning algorithm with a non-probabilistic model to identify song clusters. Although we believed that our data would fit subspaces better than clusters, we did not want to obscure the data's original features in our results. Therefore, we decided to use the k-Means algorithm as opposed to alternatives such as the PCA model.

In order to test the success of our clustering, we ran k-Means on two melody genres with only two clusters, used maximum recall probability to determine the correct cluster assignment, and calculated an F-score to account for both precision and recall error. A higher F-score indicates less error and better success. To account for variability in k-Means, we averaged the F-scores for multiple iterations of k-Means.

Eventually, to determine the ideal cluster for a new song, one could compare the song's distance in the feature space to the cluster centroids, and the cluster with the centroid a minimum distance away would be considered the cluster of best fit.

3.1.2 Feature Definition

We define our features to be a sequence of absolute notes of a specified length, such as C-E-G#. We characterize each composition by the frequency of each feature in that composition, and cluster the compositions based on their location in the resulting feature space.

3.1.3 Feature Subset Selection

Since using all possible note sequences as features would result in a too-sparse feature space for k-Means clustering, we selected a subset of features to serve as the axes for the feature space.

We considered two methods for selecting features: 1) selecting the total most frequent features across all songs in the two genres, and 2) selecting the features with the highest variance in relative frequency, i.e. features that are very frequent in some categories and very infrequent in other categories.

3.1.4 Number of Features

We varied the number of features selected for k-Means clustering from 1 feature to 200 features. 1 feature would be equivalent to seeing if a single note sequence is more prevalent in some genres than others. The maximum number of features is $number_of_notes^{feature_length}$.

3.1.5 Feature Length

We examined features from length 1 to 5. For feature length 1, our algorithm is equivalent to analyzing differences in note frequency distributions.

3.2 Markov Chain Model

For our second model, we modeled each genre with a Markov chain model for a variety of levels k. This model makes intuitive sense as a way to capture melodic idioms, because it explicitly models each note as being drawn from a probability distribution dependent on the k notes directly preceding it.

For a level k Markov chain model, we model the probability $P(d^{(i)}|g)$ that a held out document $d^{(i)}$ of length n belongs to a genre g with the following formula:

$$P(d^{(i)}|g) = \prod_{j=k+1}^{n} p(d_j^{(i)}|d_{j-k-1\ldots j-1}^{(i)}, g)$$

where each term $p(d_i|d_{i-k-1\ldots i-1})$ is the smoothed probability given by the Markov chain that the k-note subsequence $d_{i-k-1\ldots i-1}$ (henceforth also: feature) is followed by the note d_i. This equation is the result of a slightly stronger variant of Naive Bayes assumption: it is derived by assuming that note $d_j^{(i)}$ is independent of all notes further than k notes before it. Therefore: Because of the varying sizes of the data sets, we used a standard 70/30% hold out split for the large data sets and leave out one cross validation (LOOCV) for the smaller ones. The Markov Chain Model tended to perform the quite well, with training error of around 1% for $k = 3$ for the entire data set. The following data involve representations of the songs in terms of relative degree.

Observing the performance of the model as a function of k provides important insight into the data (Figure 2b). Over a variety of parameter values and data subsets, Markov models of level 3 and 4 showed the least average training error.

The failure of $k = 1$ to predict genre well demonstrates that looking only at one note before, suggesting melodic idioms of length two (i.e. intervals), is too myopic. (Note that it still performs significantly better than chance, however.) Longer features also lead to poorer models. This also makes intuitive sense, in that longer subsequences begin to be characteristic of the overall melody of a specific song, and are consequently long enough to be easily consciously recognized. The drive to be unique will therefore discourage songs from developing similar features of this length. Equally importantly, the data becomes sparser for these k-values, because the size of the feature space is exponential in the length of the feature.

We postulate moreover that levels three and four showed the best genre categorization because they are similar to the most common lengths of measures, which are natural structural breaks in melody. Irish songs in particular tend to be strongly rhythmic, and furthermore are a robuster dataset. We therefore separated the Irish tunes into four categories, based on their time signatures: 2/4, 3/4, 4/4, 6/8 and 9/8, predicting that $k = 3$ would predict better for 3/4, 6/8 and 9/8, whereas $k = 4$ would predict 4/4 better (Figure 3).

This certainly turned out to be true for $k = 4$. For $k = 3$ the odd-numbered time signatures fared much better, but still had higher error than 4/4. This could either indicate that songs in the tempo 4/4 tend to be more self similar and therefore predictable; or it could be a result of the fact that this category has more songs.

The effect of the length of feature chosen can therefore be viewed as, to some extent, implicitly modeling low level structural elements of the songs. This illuminates the inevitable bias of a simple model like this over a domain, music, renowned for its complexity. This is supported by the fact that increasing the size of the dataset did not have a significant effect on its predictive capability. A model which explicitly takes the structure into account, such as a hierarchical feature model, would certainly be excellent for this task, but it was beyond the resources of the authors to implement.

By considering the relative degree of notes in a melody instead of the absolute displacement in half steps of a note from the root, as we do with the Markov Chain Model, we demonstrate something even more surprising: within the same musical tradition and the same genre, one can distinguish songs from different scales, even when projected onto the same relative scale. It is easy to suppose, for instance, that major melodies and minor songs are fundamentally the same, differing only in that the latter have flat thirds, sixths and sevenths. This paper, however, demonstrates that at least for Irish folk tunes, this is not the case. Each mode instead appears to be characterized by particular relative melodic idioms that are independent of the absolute difference in half steps from the root.

This trend is especially visible in Irish folk music. One reason why this trend might be particularly clear for this genre is a result of the instruments that the songs are played on. Many traditional instruments, such the penny whistle (feadg) and the harp, are tuned diatonically (i.e. the white keys on the piano), and so different modes are also necessarily in different keys. This can affect which melodies are the easiest or most possible to play. On the penny whistle, for instance, it is particularly easy to go to the seventh (all fingers down) before hitting the root (one finger up); whereas this pattern is impossible in the major scale except for in the octave above.

Table 1: F-Scores for Feature Subset Selection by Highest Frequency

GENRE	malahari	children's tunes	harikambodhi	shankarabharanam	bhairavi	kharaharapriya	irish major	irish minor	irish dorian	irish mixylodian
sarali varasai	0.8833	0.8198	0.8077	0.8539	0.9167	0.9042	0.2790	0.5747	0.5983	0.8228
malahari		0.8542	0.8333	0.8452	0.8667	0.8542	0.2456	0.5531	0.4053	0.4494
children's tunes			0.7471	0.6392	0.8875	0.8750	0.2482	0.4246	0.4405	0.5118
harikambodhi				0.6198	0.8667	0.6889	0.2521	0.3906	0.4160	0.5420
shankarabharanam					0.8786	0.8661	0.2517	0.4428	0.3816	0.6126
bhairavi						0.6467	0.2636	0.5680	0.5159	0.7406
kharaharapriya							0.2509	0.4734	0.4875	0.7510
irish major								0.5043	0.4459	0.3519
irish minor									0.5983	0.6684
irish dorian										0.5600

Mean F-score: 60.32

Table 2: F-Scores for Feature Subset Selection by Highest Variance

GENRE	malahari	children's tunes	harikambodhi	shankarabharanam	bhairavi	kharaharapriya	irish major	irish minor	irish dorian	irish mixylodian
sarali varasai	0.7266	0.7566	0.6701	0.7855	0.8328	0.6466	0.3163	0.9582	0.9594	0.9314
malahari		0.7689	0.8333	0.8452	0.8667	0.8542	0.2450	0.5170	0.4008	0.4052
children's tunes			0.6410	0.6606	0.6990	0.6761	0.2463	0.4464	0.3854	0.5453
harikambodhi				0.6750	0.8667	0.6889	0.2392	0.3850	0.3407	0.4590
shankarabharanam					0.8786	0.8661	0.2478	0.4250	0.3626	0.4593
bhairavi						0.7075	0.2589	0.5344	0.4327	0.6249
kharaharapriya							0.2520	0.4581	0.3857	0.5129
irish major								0.4172	0.4372	0.3386
irish minor									0.6197	0.6813
irish dorian										0.5619

Mean F-score: 57.70

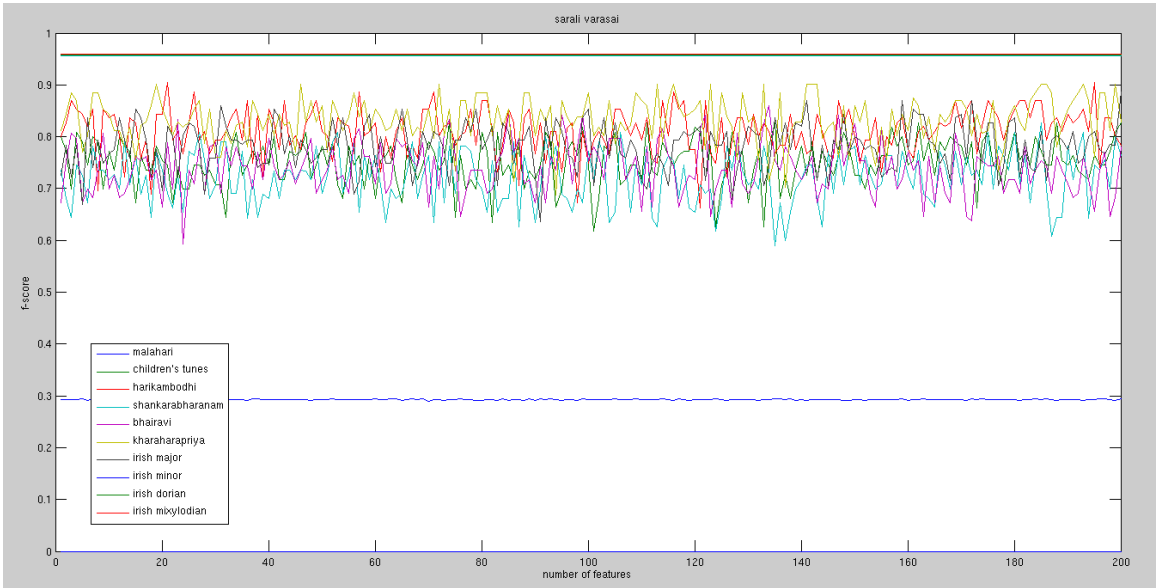
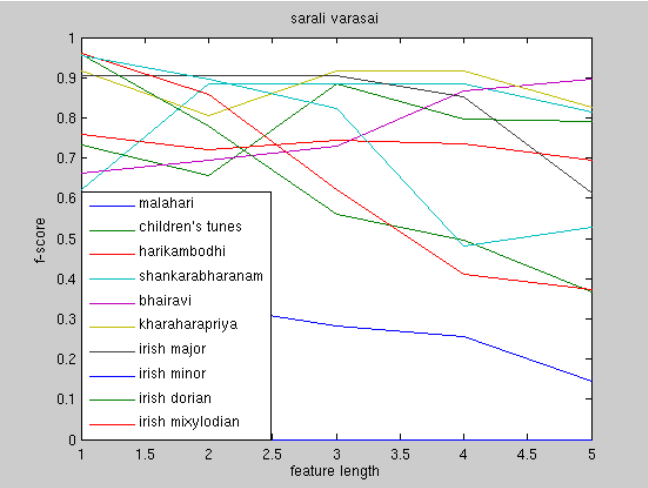


Figure 1: F-score by Number of Features

Figure 2

(a) F-score by Number of Features



(b) Average classification error over held-out documents in all genres, for the complete dataset (minus the smaller data sets), 9 categories.

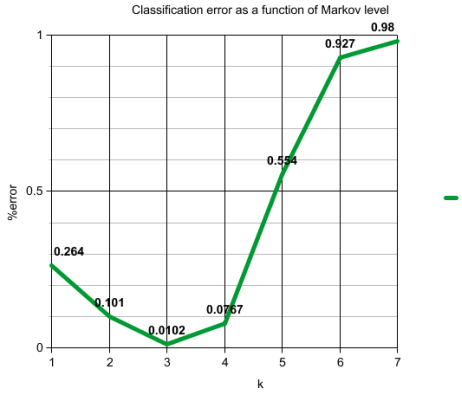


Figure 3: Training error versus time signature (beats per measure). When training with features of length 4 (right), songs in 4/4 are much more accurately classified. A feature of length 3 (left) improves the classification of odd-valued tempos significantly, but are still not classified as well as 4/4.