A similarity search for monophonic melodies

Shashank Srivastava, Snigdha Chaturvedi, Arnab Bhattacharya
Department of Computer Science and Engineering, IIT Kanpur
ssriva@iitk.ac.in, snigdhac@gmail.com, arnabb@iitk.ac.in

The Problem

Given a sample query, return the *k* most similar melodies from an unlabeled music corpus

We suggest a sequential process of preliminary clustering on global features, followed by subsequent low level feature matching

This approach can significantly reduce search overheads in a large corpus, and lead to better retrieval

Review

- Feature based methods using global signal characteristics, and summary statistical features of waveforms useful for genre identification
- Transportation metrics such as Edit distance and Earth Mover Distance extensively used for temporal and spatial matching, especially in image searches
  - incorporate continuity & partial matching
  - computationally prohibitive for large datasets

Datasets

The following datasets were used in the study

- QSBH midi dataset (Chinese and English melodies)
- Finnish folk songs
- Midi keyboard inputs

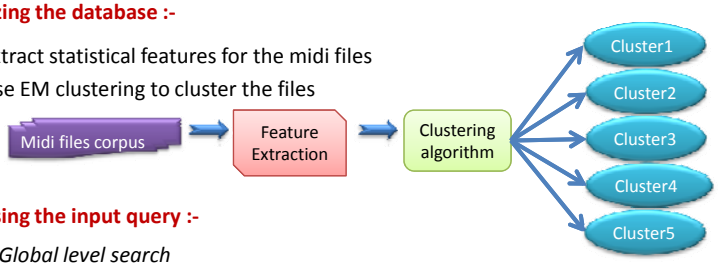
ONSET (Beats)	DURATION (Beats)	MIDI CHANNEL	MIDI PITCH	VELOCITY	ONSET (Sec)	DURATION (Sec)
0.00	0.9000	1.0000	64.00	82.00	0.00	0.5510
1.00	0.9000	1.0000	71.00	89.00	0.61	0.5510
2.00	0.4500	1.0000	71.00	82.00	1.22	0.2755
2.50	0.4500	1.0000	69.00	70.00	1.53	0.2755
3.00	0.4528	1.0000	67.00	72.00	1.83	0.2772

The Midi Representation

Our Approach

Organizing the database :-

- Extract statistical features for the midi files
- Use EM clustering to cluster the files



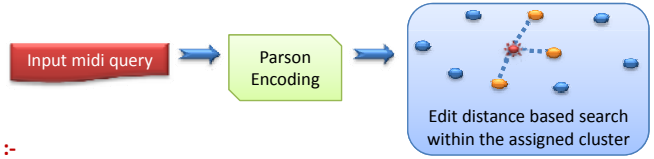
Processing the input query :-

Step 1: *Global level search*

- Extract statistical features for the input midi query
- Use a classification algorithm (SVM) to determine its cluster with best matching high level statistics.

Step 2: Lower level search

- Return scores with the minimum transportation distances, EMD or Edit Distance, from *within* the assigned cluster are returned as similar songs



Details :-

Feature selection for global level search:

- Probable feature vectors
  - Mel Frequency Cepstral coefficients (MFCC)
  - Statistical features: Pitch mean, beat duration, fraction of silent time , fraction of sharp notes, mean note-duration and velocity, note rate etc
- Optimal feature set chosen by validation by classification task on labeled data
  - Accuracy of 66% for MFCC
  - Accuracy of 82% for a subset of 7 statistical features

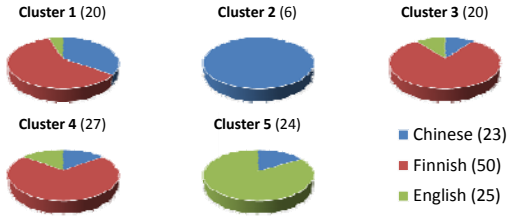
Implementation details in lower level search:

- Parson encoding
  - [62, 63, 64, 64, 66, 65, 62] → [up up same up down down]
- Flexible window size based approach for search within the cluster
- Modified edit distance to allow for occasional mistakes in playing, and domain specific tuning of cost matrix
  - High replacement cost

Results

Clustering :-

On a sample dataset of 98 melodies, EM clustering led to creation of 5 clusters. Cluster-formation by this method was seen to have a strong correlation with the class of melody, with clusters showing a largely homogeneous composition



Similarity Search :-

The system was tested with human test subjects, for evaluation of similarity results for random queries. On average, about 70% of output sequences are judged as similar to the Midi input. In case of inputs trying to play an existing melody, a rendition of the melody is reported among the top three in 6 out of 10 cases. In 3 cases, the file is reported as the first hit.

Clustering is seen to lead to retrievals within the same genre, while avoiding dissonant melodies from alternative genres, but with the same note patterns.

Conclusion

Results are encouraging, especially since most input sequences are played by amateur players. Through this approach, temporal structure can be matched while avoiding exorbitant computational overheads. The approach is also seen to be robust to poor playing, or occasional mistakes in input from a Midi Keyboard.

The approach could be extended for the Earth Mover Distance, which has a much higher complexity than O(n^2) for Edit Distance.