**Audio dataset clustering**

In this section, we assess the performance of the clustering algorithm on audio datasets. three different datasets from multiple sources were selected to conduct the experiments. The performance of each of the algorithms is presented in terms of accuracy measure. The detail of the selected audio dataset and experimental results are presented below

1. Music Audio Benchmark Data Set [1]

The Dataset contains 1886 songs all being encoded in mp3 format. The frequency and bitrate of the audio files are 44,100 Hz and 128 kb. The dataset contains 9 music genres: Pop, Rock, Folk/Country, Alternative, Jazz, Electronic, Blues, Rap/HipHop, Funk/Soul. The following table shows the overview of the distribution of the dataset.

Table Music genre distribution on the music audio dataset

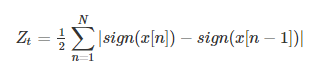
|  |  |
| --- | --- |
| Genre | Number of Samples |
| alternative | 145 |
| blues | 120 |
| electronic | 113 |
| Folk/country | 222 |
| Funk soul rnb | 47 |
| jazz | 319 |
| pop | 116 |
| Rap hip hop | 300 |
| rock | 504 |

* 1. Feature extraction

Every audio signal consists of several features. However, we have to extract meaningful characteristics that are relevant to classify our audio clips. In this study, we have selected 5 features: Mel-Frequency Cepstral Coefficients, Spectral Centroid, Zero Crossing Rate, Chroma Frequencies, Spectral Roll-off.

1. Zero Crossing Rate

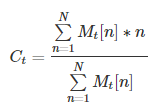
The zero-crossing rate is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or back. This feature has been used heavily in both speech recognition and music information retrieval. It usually has higher values for highly percussive sounds like those in metal and rock.



where the sign function is 1 for positive arguments and 0 for negative arguments and x[n] is the time domain signal for frame t. Time domain zero crossings provide a measure of the noisiness of the signal.

1. Spectral Centroid

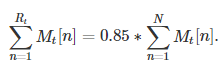
It indicates where the “center of mass” for a sound is located. spectral Centroid is calculated as the weighted mean of the frequencies present in the sound.



where Mt[n] is the magnitude of the Fourier transform at frame t and frequency bin n. The centroid is a measure of spectral shape and higher centroid values correspond to “brighter” textures with more high frequencies.

1. Spectral Rolloff

It is a measure of the shape of the signal. It represents the frequency below which a specified percentage of the total spectral energy



1. Mel-Frequency Cepstral Coefficients

The Mel frequency cepstral coefficients (MFCCs) of a signal are a small set of features (usually about 10–20) which concisely describe the overall shape of a spectral envelope. It models the characteristics of the human voice

1. Chroma features

Chroma features are an interesting and powerful representation for music audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave.

* 1. Experimental results

We have conducted three experiments by varying the number of music genres. Table 2 presents the experimental results using the accuracy metric.

Table Clustering results in terms of accuracy measure

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Geners | No.of genres | kmeans | FCM | POCS |
| alternative, blues, electronic, folk/country, funk/soul/rnb, jazz, pop, rap/hiphop | 8 | 0.1975 | 0.1975 | 0.2329 |
| alternative, blues, electronic, folkcountry, funksoulrnb, | 6 | 0.2267 | 0.2194 | 0.2732 |
| alternative, blues, electronic, folkcountry | 4 | 0.3133 | 0.31 | 0.34 |

1. GTZAN Genre Collection dataset [2]

This dataset was used for the well-known paper in genre classification " Musical genre classification of audio signals " by G. Tzanetakis and P. Cook in IEEE Transactions on Audio and Speech Processing 2002. The dataset consists of 1000 audio tracks each 30 seconds long. It contains 10 genres, each represented by 100 tracks. The tracks are all 22050Hz Mono 16-bit audio files in .wav format. We have conducted the same feature extraction technique presented in section. The clustering experimental results of this dataset are presented in table 3.

Table Clustering results in terms of accuracy measure

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Geners | No.of genres | kmeans | FCM | POCS |
| blues, classical, country, disco ,hiphop, jazz, metal ,pop ,reggae | 9 | 0.37 | 0.3822 | 0.37 |
| blues, classical, country, disco ,hiphop, jazz, metal ,pop | 8 | 0.38 | 0.42625 | 0.4075 |
| blues, classical, country, disco ,hiphop, jazz | 6 | 0.3966 | 0.4066 | 0.39 |
| blues, classical, country, disco | 4 | 0.5325 | 0.5225 | 0.5375 |

1. Gender voice dataset [3]

In order to analyze gender by voice and speech, a training database was required. A database was built using thousands of samples of male and female voices, each labeled by their gender of male or female. The dataset has the following features:

* duration: length of signal
* meanfreq: mean frequency (in kHz)
* sd: standard deviation of frequency
* median: median frequency (in kHz)
* Q25: first quantile (in kHz)
* Q75: third quantile (in kHz)
* IQR: interquantile range (in kHz)
* skew: skewness
* kurt: kurtosis
* sp.ent: spectral entropy
* sfm: spectral flatness
* mode: mode frequency
* centroid: frequency centroid
* peakf: peak frequency (frequency with highest energy)
* meanfun: average of fundamental frequency measured across acoustic signal
* minfun: minimum fundamental frequency measured across acoustic signal
* maxfun: maximum fundamental frequency measured across acoustic signal
* meandom: average of dominant frequency measured across acoustic signal
* mindom: minimum of dominant frequency measured across acoustic signal
* maxdom: maximum of dominant frequency measured across acoustic signal
* dfrange: range of dominant frequency measured across acoustic signal
* modindx: modulation index. Calculated as the accumulated absolute difference between

Table Clustering results in terms of accuracy measure

|  |  |
| --- | --- |
|  | accuracy |
| kmeans | 0.6338 |
| FCM | 0.6767 |
| POCS | 0.6284 |

References

1. Homburg, Helge, et al. "A Benchmark Dataset for Audio Classification and Clustering." *ISMIR*. Vol. 2005. 2005.
2. Tzanetakis, George, and Perry Cook. "Musical genre classification of audio signals." *IEEE Transactions on speech and audio processing* 10.5 (2002): 293-302.
3. http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning/