Audio Engineering Society

Convention Paper

Presented at the 150th Convention
2021 June 1–4, Online

This convention paper was selected based on a submitted abstract and 750-word precis that have been peer reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This convention paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Timbre-based machine learning of clustering Chinese and Western Hip Hop music

Rolf Bader[1], Axel Zielke[1], and Jonas Franke[1]

[1]*Institut of Systematic Musicology, University of Hamburg, Germany*

Correspondence should be addressed to Rolf Bader (`r_bader@t-online.de`)

摘要

Chinese and Western Hip Hop musical pieces are clustered using timbre-based Music Information Retrieval (MIR) and machine learning (ML) algorithms. Psychoacoustically motivated algorithms extracting timbre features such as spectral centroid, roughness, sharpness, sound pressure level (SPL), flux, etc. were extracted form 38 contemporary Chinese and 38 Western 'classical' (USA, Germany, France, Great Britain) Hip Hop pieces. All features were integrated over the pieces with respect to mean and standard deviation. A Kohonen self-organizing map, as integrated in the Computational Music and Sound Archive (COMSAR[6]) and apollon[1] framework was used to train different combinations of feature vectors in their mean and standard deviation integrations. No mean was able to cluster the corpora. Still SPL standard deviation perfectly separated Chinese and Western pieces. Spectral flux, sharpness, and spread standard deviation created two sub-cluster within the Western corpus, where only Western pieces had strong values there. Spectral centroid std did sub-cluster the Chinese Hip Hop pieces, where again only Chinese pieces had strong values. These findings point to different production, composition, or mastering strategies. E.g. the clear SPL-caused clusters point to the loudness-war of contemporary mastering, using massive compression to achieve high perceived loudness.

Keywords: Hip-Hop, Music Information Retrieval, Machine Learning

1 Introduction

Although Hip Hop is a global phenomenon, many subgenres like old school, trap, etc., national and local styles, as well as different production environments and end-user equipment adaptations exist, next to different languages and rap styles. This paper is a first attempt to find similarities and dissimilarities between Western and China based Hip Hop pieces. As a first attempt, timbre comparison is performed. Further analysis will include other musical features, as well as comparison between Chinese Hip Hop and Chinese traditional music. The investigation is part of the Computational Music and Sound Archive (COMSAR) project which is implemented online with the Ethnographic Sound Recording Archive (ESRA) of collections based at the Institute of Systematic Musciology in Hamburg, Germany, and offline as a jupyter notebook framework COMSAR based on an apollon MIR and ML toolbox.

2 Methods

2.1 Musical corpora

38 Chinese and 38 Western Hip Hop pieces were selected by the second author of this paper as an expert on Hip Hop music, both, in the West and in China. Western pieces are based in USA (10), Germany (10), France (10), and Great Britain (8), carefully selected with respect to criteria discussed below. A larger corpus will be used in the future, still experience with other coropra show that starting from a small set gives often better hints with respect to differences compared to a larger one. Still, future investigations will include more and more differentiated pieces, including traditional Chinese music.

2.1.1 Western Hip Hop pieces

The collection of Western Hip Hop music focuses on 'classical' pieces from different countries in a musical style that can best be described as Boom Bap or Golden Era Hip Hop music. The US-based selection consists in large part of classic 90s tracks (such as 'C.R.E.A.M by *Wu-Tang Clan* and 'Scenario' by *A Tribe Called Quest*) and a few selected tracks recorded after 2000, which are sonically closely related to the classic Boom Bap sound, as for instance 'King Kunta' by *Kendrick Lamar*.

Though there have been German Hip Hop releases in the 80s, the evolution of Hip Hop in Germany proceeded with delay in comparison the Hip Hop from the USA, where the culture was originated. Hence, a first culmination of now classic songs heavily influenced by Boom Bap emerged in the late 90s and early 2000s which constitute the majority of German songs in the collection, as for instance 'Wir waren mal Stars' by *Torch* or 'Cruisen' by *Massiven Töne*.

French Hip Hop was much ahead of the German scene, and developed an own, now classic, style of Hip Hop, as well influenced by US-based Boom Bap as practically all Hip Hop outside the USA. The great deal of the French songs is taken from the second half of the 90s (such as 'Tout n’ est pas si facile' by *Supreme NTM*) including one earlier classic ('Qui seme le vent recolte le tempo' by *MC Solaar* from 1991) and a few songs from the early 2000s like 'L’ ombre sur la mesure' by *La Rumeur*.

While UK Hip Hop also bears similarities to US Hip Hop, the sound differs quite a bit, not least because of a heavy Caribbean influence. UK artists developed a much grimier and faster style (not uncommonly above 100 bpm) of Hardcore-Rap labelled Britcore. A good portion of the UK collection are Britcore classics like for example ’Mind Of An Ordinary Citizen’ by *Blade*, ’20 Seconds To Comply’ by *Silver Bullet* or ’Terrorist Group’ by *Hijack*. A few songs are post 2000, for their part today viewed as classics from the UK, as for instance ’Witness The Fitness’ by *Roots Manuva*.

Generally, the collection of Western Hip Hop does not contain more recent implementations of Hip Hop music like Drill, Trap, or Cloud Rap, but focuses on a classic sound of the respective countries and areas.

2.1.2 Chinese Hip Hop pieces

Though Hip Hop arrived in China in the 80s (played in nightclubs by local DJs) it was not until 1992 that the first song with a Rap on it was released. Hip Hop remained more or less a subculture through the nineties, where artists often recreating standards from the USA. The first record to pioneer a more unique style was the debut by Beijing based Hip Hop group *Yin Ts’ ang* called ’Serve The People’. Though more and more records were released and numerous Hip Hop venues were established by the late 2000s, Hip Hop did not become a mass phenomenon until 2017, with the emergence of the reality-TV-show ’The Rap of China’.

The collection of Chinese Hip Hop songs used here focuses on a more current sound, with the majority of songs being produced after 2015. It compiles songs from various different cities / provinces throughout the country, involving songs from Taiwan and Hong Kong, in order to represent the versatility of the current sound of Chinese Hip Hop.

More specifically, the collection consists of five songs from the Guangdong province (such as ’广东西安说唱/饶舌’ by *Tizzy T* from Guangzhou), four songs from Beijing (such as ’Real Rapper’ by *Saber*), three songs from Hunan (such as ’Manta’ by *Lexie Lu* from Changsha), nine songs from Sichuan (such as ’Made in China’ by one of China’ s most poular Hip Hop Acts *The Higher Brothers* from Chengdu, and ’My New Swag’ by famous female rapper *VAVA* from Ya’ an), one song from Xinjiang (’The Luxury Life’ by *Afterjourney & Boom*), two songs from Chongqing (such as ’Endless Flow’ by *Damnshine, GAI & Bridge*), two songs from Nanjing (such as ’喜新戀舊’ by *Jonie J*), two songs in each case from Shanghai and Fuzhou (such as ’说唱大帝’ by *Kozay* from Shanghai), and one song from Xi’ an which is ’中二病’ by *PG One*. The collection is completed by three songs from Taipei (including ’我的生活’ by famous Taiwanese rapper *MC Hotdog*), and three songs from Hong Kong, including a song by *MC Jin* who was the first Asian rapper to be signed by a US based major label.

2.2 Feature extraction and machine learning

The pieces were analyzed with respect to several timbre features. For the sake of brevity of this paper, only those showing best clustering are discussed below. Together seven features (spectral centroid, spectral spread, spectral flux, roughness, sharpness, kurtosis, and sound pressure level (SPL)) were analyzed for frames of length 2^{15} with overlap of 2^{14} sample points, where all pieces had CD quality of a sample rate of dt = 44.1 kHz.

All feature extraction and machine learning was performed using the COMSAR[6] and apollon[1] framework developed at the Institute of Systematic Musicology.

2.2.1 Spectral Centroid

The spectral centroid C is the center of a spectrum, where the sum of amplitudes of frequencies above and below this center are equal, and is calculated as

$$C = \frac{\sum_{i=0}^{N} f_i A_i}{\sum_{i=0}^{N} A_i} \ . \tag{1}$$

This corresponds to psychoacoustic brightness perception.

2.2.2 Sharpness

Perceptual sharpness is related to the work of Bismarck[5] and followers [2][3][7]. It corresponds to small frequency-band energy. According to [7] it is measured in acum, where 1 acum is a small-band noise within one critical band around 1 kHz at 60 dB loudness level. Sharpness increases with frequency in a nonlinear way. If a small-band noise increases its center frequency from about 200 Hz to 3 kHz sharpness increases slightly, but above 3 kHz strongly, according to perception that very high small-band sounds have strong sharpness. Still, sharpness is mostly independent of overall loudness, spectral centroid, or roughness, and therefore qualifies as a parameter on its own.

To calculate sharpness, the spectrum A is integrated with respect to 24 critical or Bark bands, as we are considering small-band noise. With loudness L_B at each Bark band B sharpness is

$$S = 0.11 \frac{\sum_{B=0}^{24Bark} L_B g_B B}{\sum_{B=0}^{24Bark} L_B} \ acum, \tag{2}$$

where a weighting function g_B is used strengthening sharpness above 3 kHz like[9]

$$g_B = \begin{cases} 1 \ if \ B < 15 \\ 0.066 e^{0.171B} \ if \ z \geq 15 \end{cases} \tag{3}$$

2.2.3 Spectral Spread

The spectral spread is defined as the standard deviation of a spectrum around its mean like

$$Sp = \sqrt{\left(\sum_{i=0}^{N} (f_i - C)^2 \, A_i \right)} \tag{4}$$

with spectral centroid C. Broadband spectra therefore have a large spread, while narroband signals are low with respect to this measure.

2.2.4 Spectral flux

Spectral flux is calculated as a second derivative of a spectrum using a central difference derivation like

$$Sf = \sum_i \frac{-2\, A_i^t + A_i^{t-1} + A_i^{t+1}}{dt^2} \ , \tag{5}$$

with sample rate dt.

2.2.5 Sound pressure level (SPL)

Although several algorithms of sound loudness have been proposed[7], for music, still no satisfying results have been obtained[10]. Most loudness algorithms aim for industrial noise, and it appears that musical content considerably contributes to perceived loudness. Also, loudness is found to statistically significantly differ between male and female subjects due to the different constructions of the outer ears between the sexes. Therefore a very simple estimation of loudness is used, and further investigations in the subject are needed. The algorithm used is

$$L = 20\log_{10}\frac{1}{N}\sqrt{\sum_{i=0}^{N}\frac{A_i^2}{A_{ref}^2}}\ . \tag{6}$$

For each feature, mean and standard deviation over the whole piece was calculated, resulting in 2 × 7 values. From these, by trail-and-error, different combinations were fed into a Kohonen self-organizing map[Kohonen 2001] (SOM) of dimensions 18 × 18. In all cases of feature combinations, a SOM was trained, where 500 iterations showed convergence. Successful, clustering SOMS were calculated several times, leading to the same results.

3 Results

Training SOMs only with mean values of the timbre feature vectors did not result in any clustering of Chinese and Western Hip Hop pieces. Only the standard deviations were successful. Combination of feature standard deviations were performed, starting with a single feature and using up to six features. Only perfect clusters are discussed below, which appeared in combinations of up to three features.

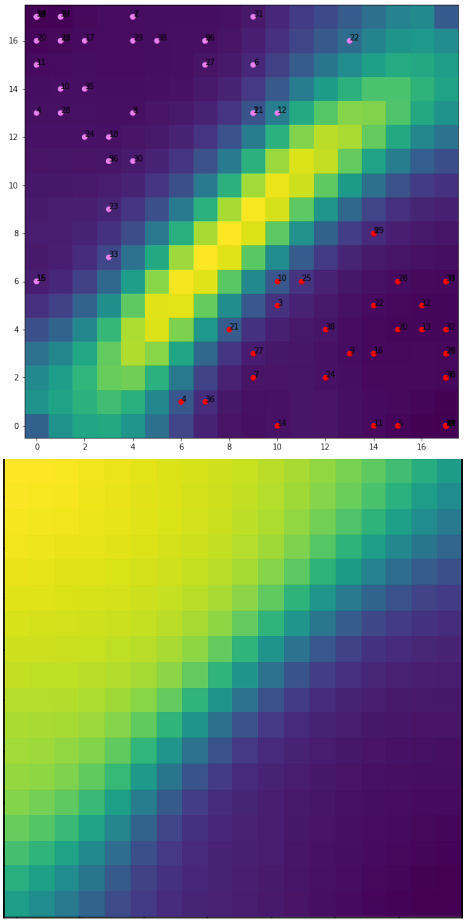

Fig. 1: top: u-matrix of the trained Kohonen self-organizing map (SOM) with training pieces placed at best-fitting neurons, trained by SPL standard deviation only. Red: Chinese Hip Hop pieces, Violet: Western Hip Hop pieces. The SOM is able to cluster Hip Hop pieces of Chinese and Western origin. Background color: similarity between neighboring neurons (blue: most similar, yellow: most dissimilar). Numbers indicate musical pieces. Bottom: SOM SPL feature, indicating that Western Hip Hop has generally a much larger SPL standard deviation compared to Chinese Hip Hop.SPL alone is perfectly able to distinguish between both corpora.

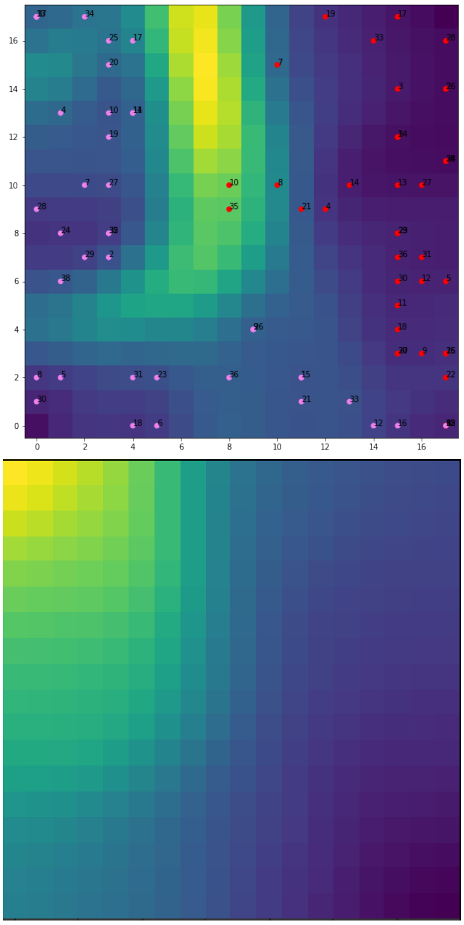

Fig. 2: Top: u-matrix of trained SOM for three featues: SPL, roughness, and sharpness standard deviation. Both corpora cluster perfectly, while Western Hip Hop is further divided into an upper and lower cluster on the left. Second from top to bottom: features SPL, roughness, and sharpness std respectively. While again SPL std perfectly splits the corpora, both, roughness and sharpness show strong std for the upper left Western Hip Hop pieces only.

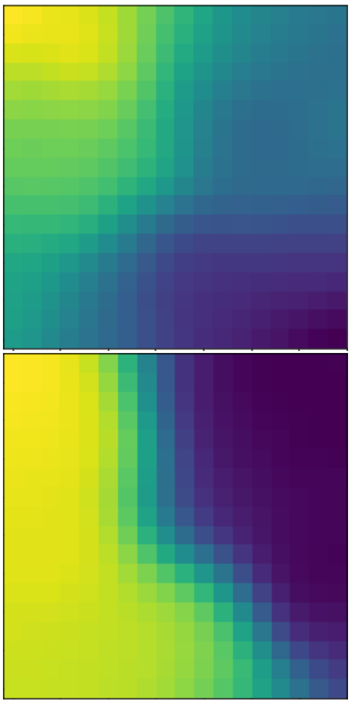

Fig. 3: Fig 2 cont

Fig. 1 (top) shows the u-matrix of the trained SOM with only one parameter, the sound pressure level (SPL), where the training pieces are best-fitted. The Chinese Hip Hop pieces (red) and the Western Hip Hop songs perfectly separate and cluster already with SPL alone. The blue and green regions of the background color indicate similar neurons, the yellow regions point to dissimilar neurons. Therefore, both corpora are very consistent within themselves, and clearly split apart. Thereby, the bottom plot in Fig. 1 shows the SPL feature. The Western Hip Hop pieces all have a much higher SPL standard deviation compared to the Chinese pieces. This correlates with the so-called loudness-war of compression in song mastering. This is nearly not applied to classical Western Hip Hop and is strongly present in Chinese Hip Hop of more contemporary kind.

No other feature alone was able to separate the two corpora perfectly. In Fig. 2 three features are used, roughness, sharpness, and SPL std. SPL again perfectly clusters, while roughness and sharpness std split the Western pieces in two clusters. A upper left cluster has strong values in all features, while a lower left and bottom cluster still has a strong SPL std but roughness and sharpness std similar to those of Chinese Hip Hop. Still the lower cluster is not corresponding to a country.

Other combinations are differentiating the picture. The combination of spectral centroid and SPL, again both std, also perfectly cluster the two corpora. This is caused by the SPL. Still the centroid differentiates the Chinese Hip Hop pieces into a cluster with high and a cluster with low centroid std. Still all Western pieces show a low std in spectral centroid. Therefore, all Western pieces have a more constant brightness over the course of each piece, while brightness might or might not change strongly in Chinese Hip Hop.

The combination of spectral centroid and spread, both std, are also able to perfectly cluster the corpora. Still here it is a combination of the two features leading to the clusters, where spread std is less with Chinese Hip Hop pieces. When combining spectral flux, sharpness, and SPL std, the clustering is again perfect, where SPL is the cause of the cluster. Both, sharpness and flux split the Western corpus into two clusters, one with strong sharpness and flux alike, and one with little so. For the Chinese songs, both., sharpness and flux are low.

In summery, SPL is the main feature separating the corpora perfectly. The Western corpus is further subdivided by spectral sharpness, flux, and spread std, where large values only appear in this Western and not in the Chinese corpus. With spectral centroid the situation is vice versa. Although this feature splits the Chinese corpus in two sub-clusters, Western pieces have an overall lower brightness SPL.

Differentiating the Western pieces as not possible with any combination of the SOM trained with all pieces. Investigating differences within Western Hip Hop might be subject to future studies.

4 Discussion

This study is a fist attempt to discover characteristics of Hip Hop pieces with respect to timbre features. Although the music corpora are small, as the pieces are chosen carefully by a Hip Hop expert the results are expected to show a general trend.

The strong SPL standard deviation split between the corpora is pointing to the loudness-war of music mastering, strongly present today and not much used with 'classical' Hip Hop pieces. The sub-clustering of Western pieces with respect to sharpness, flux, and spread need to be investigated further. Sharpness is energy in higher frequency bands, where variations might be caused by use of cymbals, or high-pitched instruments. Flux is associated with a strong use of audio effects such as flanging, phasing, or chorus. Spread is pointing to enhance use of noise in the music. When all these features do not deviate in terms of their mean values, as found, this means that both corpora use these features to the same amount. Still, Western pieces seam to variate these timbre aspects more over the course of the pieces compared to contemporary Chinese Hip Hop pieces. The opposite is true for spectral centroid. Again the mean of this feature is not clustering the corpora, but the standard deviation does, where Chinese pieces tend more to deviate stronger. This means a stronger deviation of bright and dark passages, compared to Western pieces.

All these findings might be caused by performance, composition, mixing, or mastering techniques. Future investigations need to discuss this in interviews with musicians, producers and sound engineers. Also, differentiating the results further with respect to subgenres, comparison to Chinese traditional music, as well as K-Pop, or Japanese music need to be done.

参考文献

[1] https://github.com/ifsm/apollon, Docs can be found here: https://apollon.readthedocs.io/

[2] Aures, W.: Der sensorische Wohlklang als Funktion psychoakustischer Empfindungsgrös̈sen [Sensory pleasing sounds as function of psychoacoustic perception parameters], Acustica 58, 282-290, 1985.

[3] Aures, W.: Berechnungsverfahren für den sensorischen Wohlklang beliebiger Schallsignale [Calculation methods of sensory pleasing sounds for arbitrary sound signals], Acustica 59, 130-141, 1985.

[4] Bader, R. (ed.): Computational Phonogram Archiving, Springer Series 'Current Research in Systematic Musicology', Vol. 5, 2019.

[5] Bismarck, G. v.: Sharpness as an attribute of the timbre of steady-state sounds, Acustica 30, 159-172, 1974.

[6] https://comsar.fbkultur.uni-hamburg.de/ Docs can be found here: https://comsar.readthedocs.io/

[7] Fastl H. & Zwicker, E.: *Psychoacoustics. Facts and Models.* 3rd edition, Springer 2007.

[Kohonen 2001] Kohonen, T.: *Self-organizing maps.* 3^{rd} edition, Springer, Berlin 2001.

[9] Peeters, G.: A large set of audio features for sound description (similarity and classification) in the CUIDADO project. http://recherche.ircam.fr/anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf Ircam, 2004.

[10] Ruschkowski, A. v.: *Lautheit von Musik: eine empirische Untersuchung zum Einflus̈s von Organismusvariablen auf die Lautstärkewahrnehmung von Musik. [Loudness of music: an empirical investigation on the impact of Organism variables on loudness perception of music.]* https://katalogplus.sub.uni-hamburg.de/vufind/Record/78110422X?rank=1 Hamburg 2013.