3. DATA & METHOD

To demonstrate how genre evolution may confound the performance of genre classifiers, I study the case of Chinese Hip-Hop music from the past decade. Hip-Hop has long been a relatively unpopular genre in China since it was introduced to the country in the 1990s [21]. In recent years, however, Chinese Hip-Hop experienced a series of dramatic external shocks, which elevated the genre's prominence extraordinarily. In summer 2017, the first season of *The Rap of China* started to stream on iQiyi, one of the most popular streaming platforms in China. The reality-show-style music competition attracted an unexpectedly expansive audience, securing 2.7 billion views in the course of three months and becoming the most successful online program ever in the history of the sector. The contestants featured in this program, as well as their music, soon became the iconic representatives of an unprecedented fad. Most significantly, many of the high-profile contestants, including the co-winners of the competition, GAI and PG One, were known for their involvement in Trap music, a style that originated in the Southern United States and is characterized by its distinct use of synthesized drums (above all, Roland 808 drums). Moreover, it is remarkably different from the jazzy, pop-oriented Hip-Hop style prevalent in the genre before. In January 2018, Hip-Hop's rise was met with state censorship aimed at blocking Hip-Hop musicians from mainstream media, as the state was concerned about past Hip-Hop figures' problematic practices of including subversive and delinquent content in their songs. Although many musicians had to remove some of their songs from the internet under political pressure, Hip-Hop songs, in general, were still allowed to circulate online via music streaming platforms, and the censorship loosened up six months later without any official announcement [15].

The story of Chinese Hip-Hop music thus provides an interesting case for investigating the issue of genre evolution, as the genre became remarkably popular and potentially turned in a new stylistic direction due to the success of *The Rap of China*. To understand the impact of this stylistic change on machine learning classifiers, I collected songs from one of the most popular music streaming platforms in China. I collected the songs' audio files, as well as their metadata, including their genre and tags. The platform's setup requires each song to have only one identified genre, while the song can have multiple tags that are considered its "subgenres." Because of copyright issues, only a portion of the available songs could be downloaded at the time of the data collection in September 2019. Eventually, I collected 69,427 songs released between 2009 and September 2019 from four genres (Hip-Hop, Pop, Rock, and Folk), constituting the dataset on which I conducted my analysis. The other three genres are also well established in China in that they have distinct fanbases, although the Chinese Hip-Hop community was more connected to the Rock community particularly in the early 2000s for sharing resources and venues [22].

The structure of the dataset is given in Figure 1 below. As shown in the graphic, the number of total song releases consistently increases over the years, with the exception of the year 2019, when the data collection was suspended. The data and the code for the following analysis are publicly available,[1] although the metadata of the songs is left out due to copyright issues.

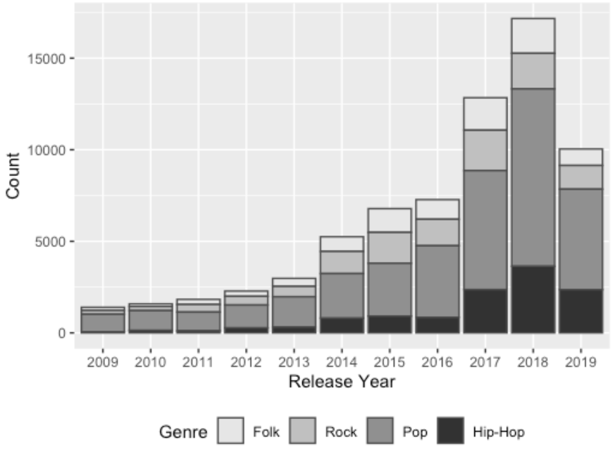

Figure 1. The structure of the dataset. The height of the bar represents the number of releases.

I used librosa [23] to extract sonic features from the audio files to prepare for training the classifiers. As my goal was not to improve classification accuracy, I followed Tzanetakis & Cook's [1] classic approach for computational efficiency. Specifically, I sliced a 1-minute long excerpt of audio signals from each song starting at 30 seconds into the song, which supposedly captures a substantive part of the song, and extracted 11 timbral texture features from this slice, including the chroma features (chroma_stft), root-mean-square of the spectrogram (rms), spectral centroid (spectral_centroid), spectral bandwidth (spectral_bandwidth), spectral roll-off (spectral_rolloff), zero-crossing rate (zero_crossing_rate), and the first five Mel-frequency cepstral coefficients (MFCC). I normalized the value of the features using a 0-1 scale and took the means of the features for each song, resulting in an 11-dimensional feature vector.

I used the songs released in 2009 and 2018, respectively, to train the genre classifiers. They represent the first and the last full year in the dataset, a fact that can be conveniently used to track changes over the year. I also used the widely studied GTZAN dataset [1] — its filtered version [24] due to the errors in the original collection [25] — as a supplementary check of the classifiers' performance trained on a dataset from a different time and region. For each of these three cohorts, I trained four classifiers: a Gaussian Naïve Bayes classifier (hereafter, GNB), a K-Nearest Neighbor classifier (hereafter, KNN), a Random Forest classifier (hereafter, RF), and a classifier trained by Extreme Gradient Boosting, a more recently prevailing tree-based algorithm. They were used in parallel to triangulate the results. For the 2009 cohort, I picked the top 50 most streamed songs of each genre, for a total of 200 songs, to compile the training set, whereas for the 2018 cohort, I picked the top 1900 songs of each genre, totaling 7600 songs. My reason for choosing these two numbers is

[1] The data can be found at https://github.com/norzvic.

that they are the closest number that can be divided by ten to the total number of releases of a genre with the least releases in that year (54 Hip-Hop songs in 2009; 1902 Folk songs in 2018), making it convenient for a 4:1 split for training and testing the classifiers. For the GTZAN dataset, I only used the data for three genres — Pop, Rock, and Hip-Hop — to train the classifiers, as there is no "Folk" in the GTZAN dataset. I extracted the data on the 11 features, normalized them, and used all 300 songs from the three genres in the dataset to train four classifiers similarly. The performance of the classifiers is shown in Table 1. I then used all four sets of classifiers to predict the genre of all the songs in the dataset, the results of which are reported in the next section.

		2009	**2018**	**GTZAN**
GNB	Acc	0.575	0.472	0.750
	AUC	0.832	0.755	0.867
KNN	Acc	0.625	0.584	0.800
	AUC	0.646	0.578	0.608
RF	Acc	0.800	0.580	0.833
	AUC	0.908	0.786	0.927
XGB	Acc	0.700	0.629	0.732
	AUC	0.905	0.864	0.916

Table 1. The performance of the classifiers across cohorts. Acc refers to the accuracy of predicting the test set, whereas AUC refers to the corresponding area under the ROC line for multi-classes.

4. FINDINGS

4.1 Overall Accuracy and Recall

As the focus of this study is Hip-Hop, I will primarily heed two metrics: Hip-Hop accuracy and Hip-Hop recall. Here, Hip-Hop accuracy refers to the rate at which the classifier correctly identifies Hip-Hop versus non-Hip-Hop, whereas Hip-Hop recall refers to the rate at which the classifier correctly identifies Hip-Hop among true Hip-Hop songs. The reason to incorporate recall is that the metric can be useful to detect potential genre evolution: the lower the recall is, the more probable that a Hip-Hop song is classified to other genres, which could mean that the genre is straying away from the sound captured by the training set.

The overall performance of the classifiers is shown in Figure 2. The "Average" column on the right of the graphic takes the average value of the corresponding metrics of the left three classifiers, showing an average trend of the metrics. The graphic demonstrates that the performance of most classifiers follows a fuzzy U-shape trend, with higher values at the start and the end than in the middle years of the dataset. The U-shape trend is statistically tested by a series of polynomial regressions of the metric on year and its quadratic term. In particular, the U-shape is demonstrable in the performance of the averaged classifier. All three recalls from the averaged classifier (bottom right of Figure 2) fit a polynomial regression on year and its quadratic term, where their coefficients are all statistically significant ($p < 0.05$). On the other hand, only the averaged classifier trained from the 2018 cohort songs yields significant results in the regression of accuracy on year and its quadratic term, while the other two averaged classifiers did not.

The results from the recalls, which imply how the true Hip-Hop songs of each year are similar to those of the benchmark cohort, indicate that Hip-Hop deviated from its late 2000s conventions in the middle of the 2010s but bounced back later, in particular after 2017, when the first season of *The Rap of China* aired. The downward trend of the prediction accuracy of the 2009 cohort classifiers may suggest more nuanced changes in the genre: given that the recall of the 2009 cohort is also U-shaped, it implicates that the 2009 cohort classifiers are able to detect the bouncing back of the genre conventions but confused other genres with Hip-Hop when the bounce took place in the late 2010s. This may be primarily because of the relatively small size of the 2009 cohort classifiers' training set, but it also suggests the possibility of other genres' cross-over to Hip-Hip, which I will discuss in 4.3.

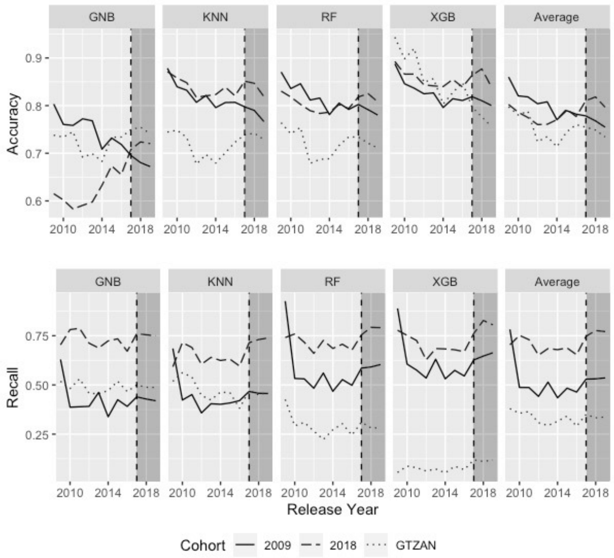

Figure 2. Hip-Hop accuracy and recall across cohorts and classifiers. The darker section on the right of each grid indicates the years since the first season of *The Rap of China* in 2017.

4.2 Metrics and Subgenre Salience

Genres are comprised of "subgenres," which are subsets of their parent genre but can be distinctly different from each other in terms of their sound [1,26]. The idea of distinct subgenres under the same genre category thus suggests that genre evolution may be understood as changes in the most salient subgenres, i.e., subgenres that take up the highest proportion of the genre. It is intuitive, then, to speculate that the classifier's performance is better when it predicts a set of songs whose most salient subgenres are more similar to those of the training set.

Testing this speculation requires, above all, identifying the most salient subgenres of the training set, or the subgenre to which most songs of that genre are affiliated. I thereby designed an original approach to measuring subgenre salience, and I used the approach for measuring the salience of Hip-Hop subgenres as follows. First, I extracted the subgenres from all the Hip-Hop songs in the

		GNB			**KNN**			**RF**			**XGB**			**Average**		
		Subg	Acc	Rec	Subg	Acc	Rec	Subg	Acc	Rec	Subg	Acc	Rec	Subg	Acc	Rec
2009	Top1	OS	0.236* (0.103)	0.625*** (0.095)	OS	0.188* (0.066)	0.636** (0.149)	OS	0.165* (0.067)	1.010*** (0.182)	OS	0.180** (0.050)	0.773** (0.165)	OS	0.154* (0.062)	0.761*** (0.139)
	Top5	OS; I; Con; A; CR	0.143* (0.045)	0.265** (0.075)	OS; I; Con; A; CR	0.123*** (0.021)	0.329** (0.072)	OS; A; U; Pop; I	0.090 (0.065)	0.001 (0.314)	OS; A; U; Pop; I	0.082 (0.058)	-0.049 (0.251)	OS; I; Con; A; CR	0.138*** (0.029)	0.301 (0.168)
2018	Top1	HH	0.211** (0.045)	0.065 (0.264)	HH	-0.001 (0.029)	0.208** (0.055)	HH	0.033 (0.023)	0.149* (0.048)	HH	0.003 (0.028)	0.206* (0.071)	HH	0.071* (0.022)	0.157** (0.047)
	Top5	HH; Pop; T; OS; CU	0.245** (0.058)	0.016 (0.070)	HH; Pop; T; OS; CU	-0.022 (0.034)	0.150 (0.093)	HH; Pop; T; OS; CU	0.245** (0.058)	0.106 (0.074)	HH; Pop; T; OS; CU	-0.013 (0.033)	0.136 (0.109)	HH; Pop; T; OS; CU	0.082** (0.028)	0.102 (0.077)

Note: Standard errors are in parentheses. Among the subtitles, Subg refers to subgenre, Acc refers to accuracy, and Rec refers to recall. Abbreviations in the Subg column: OS for Old-School Hip-Hop; I for Instrumental Hip-Hop; Con for Conscious Hip-Hop; A for Alternative Hip-Hop; CR for Cloud Rap; U for Underground Hip-Hop; Pop for Pop Rap; HH for Hip-Hop; T for Trap; CU for Chinese Underground Hip-Hop.
*p < .05; **p <.01; ***p<.001

Table 2. Performance of genre classifiers on songs released 2010-2019.

dataset and counted their appearances each year. In many cases where a song has multiple subgenres, one more appearance was added for each of them. To account for the fact that prediction performance on each subgenre can significantly impact the prediction performance on the genre as a whole, I used the 2009 cohort classifiers and the 2018 cohort classifiers, respectively, to predict the genre of the Hip-Hop songs released in the year of the same cohort. After that, I calculated the prediction accuracy of each subgenre (which is also identical to recall, as only true Hip-Hop songs are examined here) by computing the proportion of correct predictions among all the songs with the subgenre in question. In this case, the prediction accuracy of each subgenre also indicates how the classifiers were trained to favor certain subgenres over others. I multiplied the prediction accuracy of each subgenre by their number of appearances. The product is the measurement of subgenre salience, which nicely captures the size of the impact that a subgenre can make on the pooled metrics of the genre. I identified the top 1 subgenre and top 5 subgenres with the highest salience score, which are presented in the note in Table 2. Then, I calculated their proportion in the total number of releases each year, respectively. Finally, I employed linear regression of Hip-Hop accuracy and recall on the proportion of the most salient subgenre(s) each year to understand the association between them. The objective of the linear regression is to see if the classifier performs better when the salient subgenres, which are favored by the classifiers by design, are more dominant among all subgenres of a year. The GTZAN cohort was not included in the analysis in this part, as it does not have a corresponding year in the dataset.

	Hip-Hop crossing non-Hip-Hop	**Total Non-Hip-Hop**	**Percentage (%)**
2009	0	1344	0
2010	0	1446	0
2011	0	1717	0
2012	22	1991	1.1
2013	59	2603	2.2
2014	98	4347	2.2
2015	184	5682	3.1
2016	163	6264	2.5
2017	94	10381	0.9
2018	135	13394	1.0
2019	118	7565	1.5

Table 3. The number of Hip-Hop-crossing non-Hip-Hop songs and their percentage in the total non-Hip-Hop releases over the years studied.

The coefficient estimates of the proportions of salient subgenres according to the regression models are presented in Table 2. The overall pattern generally substantiates a positive relationship between performance metrics and the size of the salient subgenres' proportions. In other words, the classifiers perform worse in years where there are fewer salient subgenres among the total releases of the year in question. Specifically, the average classifier trained on 2009 songs performs better when there are more songs in that year that are Old School Hip-Hop, Instrumental Hip-Hop, Conscious Hip-Hop, Alternative Hip-Hop, or Cloud Rap, whereas the average classifier trained on 2018 songs performs better when there are more Pop Rap, Trap, Old School Hip-Hop, or Chinese Underground Hip-Hop present in that year. The results imply that it is possible to detect genre evolution from the performance of the classifiers, which is demonstrable in the metrics when the genre deviates from the salient subgenres of the classifier.

4.3 Metrics and Genre-Crossing

As implied in Figure 2, the evolution of a genre may involve other genres. This could either mean "assimilation," where the genre in question is absorbing musical elements from other genres, or "dispersion," where the musical elements of the genre in question penetrate other genres. In either case, it is reasonable to speculate that the interaction between genres will likely do harm to the performance of the classifiers that are trained on songs where genre-crossing is not prominent.

To investigate this issue, I first identified the number of releases in non-Hip-Hop genres (i.e., Folk, Rock, and Pop) with at least one subgenre that is explicitly associated with Hip-Hop. I did so by searching all the songs whose subgenres incorporate strings such as "Hip-Hop," "Hip Hop," "Rap," "Trap," and their Chinese equivalent. Table 3 presents the number of Hip-Hop-crossing non-Hip-Hop songs and their percentage among the total releases over the years studied. The table has some important implications for understanding classifiers' performance, as illustrated in Figure 2. First, there are no Hip-Hop-crossing non-Hip-Hop songs from 2009-2011. This means that the 2009 cohort classifiers were not well-trained to correctly detect

	GNB		KNN		RF		XGB		Average	
	Diff. Acc	Diff. FN	Diff. Acc	Diff. FN	Diff. Acc	Diff. FN	Diff. Acc	Diff. FN	Diff. Acc	Diff. FN
2009	-0.063*** (0.015)	0.087*** (0.016)	-0.020 (0.017)	0.163*** (0.015)	-0.089*** (0.015)	0.162*** (0.016)	-0.076*** (0.016)	0.181*** (0.016)	-0.062*** (0.008)	0.148*** (0.008)
2018	-0.090*** (0.012)	0.217*** (0.017)	-0.125*** (0.016)	0.270*** (0.017)	-0.131*** (0.016)	0.245*** (0.017)	-0.161*** (0.016)	0.261*** (0.017)	-0.127*** (0.008)	0.249*** (0.008)
GTZAN	-0.079*** (0.011)	0.182*** (0.017)	-0.120*** (0.016)	0.134*** (0.017)	-0.086*** (0.015)	0.263*** (0.015)	-0.081*** (0.014)	0.073*** (0.010)	-0.092*** (0.007)	0.120*** (0.008)

Note: Standard errors are in parentheses.
*p < .05; **p <.01; ***p<.001

Table 4. Two-sample T-tests of key metrics between Hip-Hop-crossing non-Hip-Hop songs and those that do not cross the Hip-Hop boundary. Among the subtitles,

Hip-Hop-crossing non-Hip-Hop songs, which occur more frequently in the subsequent years. This possibly explains why 2009 cohort classifiers have a downward trend in their accuracy, as they are underperformed when identifying crossover songs. Second, the fact that there are more Hip-Hop-crossing non-Hip-Hop songs in terms of their percentage in the total non-Hip-Hop releases in the middle years than in the early and the late 2010s coincides, inversely, with the U-shaped trend of the metrics in Figure 2. This makes sense, as the prominence of genre-crossing is supposed to curb classification performance.

To further understand the relationship between genre-crossing and the performance of the classifiers, I also analyzed the difference in performance between genre-crossing non-Hip-Hop songs and those that did not cross the boundary of Hip-Hop. To do so, I conducted two-sample t-tests on two metrics between the two groups. The first metric was prediction accuracy, indicating the rate at which the non-Hip-Hop genre of the song was correctly identified. The second metric was the False Negative rate, indicating the rate at which the non-Hip-Hop song was incorrectly identified as Hip-Hop. Table 4 shows the results of the test, which unanimously demonstrate how genre-crossing discounts classification performance. Specifically, the consistent and statistically significant negative values under the columns of the difference in prediction accuracy suggest that the classifiers' ability to predict non-Hip-Hop songs that do not cross the boundary of Hip-Hop is always better than their ability to predict those that do so. Similarly, all the positive values under the False Negative columns provide strong evidence that classifiers are better at identifying non-Hip-Hop songs when they do not claim Hip-Hop as one of their subgenres. In short, the performance of genre classifiers may be an indicator of genre evolution, as bad performance points to genre-crossing.

5. DISCUSSION

Genres will evolve, and this will affect the performance of genre classifiers regardless of the data sources or algorithmic approaches they use. By analyzing the case of Chinese Hip-Hop, I demonstrate that the performance of genre classifiers is significantly impacted by genre evolution, particularly when the training set has less songs of the salient subgenres and when genres start to cross boundaries.

The goal of the present study is to highlight genre as a mutable cultural construct and to examine what this means for MIR research on genre classification. This study has at least three implications. First, similar to what has been argued in some of the previous MIR studies of genre, I contend that musical genre is a cultural construct and that the musical elements associated with a genre can change over time. My findings also suggest that the prevalence of a particular music style in a genre may be revived after a period of relative silence, which echoes what music industry scholars call a "revival" of music styles [13].

Second, because genres change and evolve, one possible application of genre classifiers is to understand the patterns of such change. I show it is possible to use genre classifiers, if trained on a particularly designed dataset, to detect the way genres evolve. It is difficult, though, to separate genre evolution from flaws in algorithmic design; therefore, it is important to supplement the analysis with a more detailed investigation of the soundscape of the songs across subgenres and songs that cross genres.

Third, the findings suggest that using accuracy or accuracy-related metrics to train genre classifiers might not be the only best application in practice. Regardless of their algorithmic framework, genre classifiers are almost always better at predicting the genre of songs released within a similar timeframe as those in the training set. As a result, if genre classifiers are meant to help classify newly released songs, it might be futile to aim at improving prediction accuracy, as the genre might evolve. This implies that we need to find an optimal balance between raising the prediction accuracy of the algorithms and understanding the developing trend of genres, which might require data exploration or even qualitative data beyond the design and optimization of machine learning algorithms.

There are obvious limitations in the data analysis; however, I argue that they will not severely affect the findings. Above all, a significant number of songs were inaccessible, especially in very distant years, which may twist the conclusion that Hip-Hop has been "revived" from its convention 10 years ago. It is certainly likely that the actual dominant subgenre in 2009 may not be "Old-School Hip-Hop." However, the finding that the way classifiers are trained can affect their performance still holds. It remains true that, if a genre classifier is trained on the songs of a particular composition of subgenres, the classifier might better predict the genre of a song when it shares commonalities with that subgenre's composition. In that case, classifiers can still be used to understand the evolution of genres, although they would be better "detectors" if more data were available.

6. ACKNOWLEDGEMENT

The author would like to thank Jin Ha Lee, Oriol Nieto, and the anonymous reviewers for their insightful comments and suggestions. The author would also like to acknowledge Leo Y. Yang for his assistance with data collection.

7. REFERENCES

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[2] P. Knees, E. Pampalk, and G. Widmer, "Artist Classification with Web-Based Data.," 2004.

[3] A. J. Craft, G. A. Wiggins, and T. Crawford, "How Many Beans Make Five? The Consensus Problem in Music-Genre Classification and a New Evaluation Method for Single-Genre Categorisation Systems.," in *ISMIR*, 2007, pp. 73–76.

[4] J. Lee, J. Park, K. L. Kim, and J. Nam, "SampleCNN: End-to-end deep convolutional neural networks using very small filters for music classification," *Applied Sciences*, vol. 8, no. 1, p. 150, 2018.

[5] J. Nam, K. Choi, J. Lee, S.-Y. Chou, and Y.-H. Yang, "Deep learning for audio-based music classification and tagging: Teaching computers to distinguish rock from bach," *IEEE signal processing magazine*, vol. 36, no. 1, pp. 41–51, 2018.

[6] S. Oramas, F. Barbieri, O. Nieto Caballero, and X. Serra, "Multimodal deep learning for music genre classification," *Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21.*, 2018.

[7] J. Pons and X. Serra, "Randomly weighted cnns for (music) audio classification," in *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2019, pp. 336–340.

[8] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: a survey," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133–141, 2006.

[9] M. Schedl, E. Gómez Gutiérrez, and J. Urbano, "Music information retrieval: Recent developments and applications," *Foundations and Trends in Information Retrieval. 2014 Sept 12; 8 (2-3): 127-261.*, 2014.

[10] M. Sordo, O. Celma, M. Blech, and E. Guaus, "The quest for musical genres: Do the experts and the wisdom of crowds agree?," in *ISMIR*, 2008, pp. 255–260.

[11] C. McKay and I. Fujinaga, "Musical genre classification: Is it worth pursuing and how can it be improved?," in *ISMIR*, 2006, pp. 101–106.

[12] J. C. Lena and R. A. Peterson, "Classification as culture: Types and trajectories of music genres," *American Sociological Review*, vol. 73, no. 5, pp. 697–718, 2008.

[13] J. C. Lena, *Banding together: how communities create genres in popular music*. Princeton, N.J: Princeton University Press, 2012.

[14] P. DiMaggio, "Classification in Art," *American Sociological Review*, vol. 52, no. 4, pp. 440–455, 1987.

[15] K. Nie, "Disperse and preserve the perverse: computing how hip-hop censorship changed popular music genres in China," *Poetics*, p. 101590, 2021.

[16] D. Diakopoulos, O. Vallis, J. Hochenbaum, J. W. Murphy, and A. Kapur, "21st Century Electronica: MIR Techniques for Classification and Performance.," in *ISMIR*, 2009, pp. 465–470.

[17] B. L. Sturm, "Two systems for automatic music genre recognition: What are they really recognizing?," in *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, 2012, pp. 69–74.

[18] B. L. Sturm, "Classification accuracy is not enough," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 371–406, 2013.

[19] S. Lippens, J.-P. Martens, and T. De Mulder, "A comparison of human and automatic musical genre classification," in *2004 IEEE international conference on acoustics, speech, and signal processing*, 2004, vol. 4, pp. iv–iv.

[20] W. G. Roy, "'Race records' and 'hillbilly music': institutional origins of racial categories in the American commercial recording industry," *Poetics*, vol. 32, no. 3–4, pp. 265–279, Jun. 2004.

[21] J. Sullivan and Y. Zhao, "Rappers as knights-errant: classic allusions in the mainstreaming of Chinese rap," *Popular Music and Society*, pp. 1–18, 2019.

[22] J. De Kloet, *China with a cut: globalisation, urban youth and popular music*, vol. 3. Amsterdam University Press, 2010

[23] B. McFee *et al.*, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, vol. 8.

[24] C. Kereliuk, B. L. Sturm, and J. Larsen, "Deep learning and music adversaries," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2059–2071, 2015

[25] B. L. Sturm, "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use," *Journal of New Music Research*, vol. 43, no. 2, pp. 147–172, Apr. 2014

[26] A. Caparrini, J. Arroyo, L. Pérez-Molina, and J. Sánchez-Hernández, "Automatic subgenre classifi-

cation in an electronic dance music taxonomy," *Journal of New Music Research*, vol. 49, no. 3, pp. 269–284, 2020.

IN SEARCH OF SAÑCĀRAS: TRADITION-INFORMED REPEATED MELODIC PATTERN RECOGNITION IN CARNATIC MUSIC

Thomas Nuttall[1] **Genís Plaja-Roglans**[1] **Lara Pearson**[2] **Xavier Serra**[1]
[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[2] Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany
[1]{thomas.nuttall, genis.plaja, xavier.serra}@upf.edu, [2]lara.pearson@ae.mpg.de

ABSTRACT

Carnatic Music is a South Indian art and devotional musical practice in which melodic patterns (motifs and phrases), known as *sañcāras*, play a crucial structural and expressive role. We demonstrate how the combination of transposition invariant features learnt by a Complex Autoencoder (CAE) and predominant pitch tracks extracted using a Frequency-Temporal Attention Network (FTA-Net) can be used to annotate and group regions of variable-length, repeated, melodic patterns in audio recordings of multiple Carnatic Music performances. These models are trained on novel, expert-curated datasets of hundreds of Carnatic audio recordings and the extraction process tailored to account for the unique characteristics of *sañcāras* in Carnatic Music. Experimental results show that the proposed method is able to identify 54% of all *sañcāras* annotated by a professional Carnatic vocalist. Code to reproduce and interact with these results is available online. [1]

1. INTRODUCTION

Musical styles around the world are often structured in relatively idiosyncratic ways, which can require bespoke tools for MIR tasks and computational music analysis [1]. This is the case in Carnatic Music, a South Indian art and devotional music tradition, in which several characteristic features of the style necessitate tailored MIR approaches.

Relative to other musical styles, Carnatic Music is heavily ornamented. This ornamentation is not superficial decoration but rather is integral to musical meaning [2]. The ornaments, known as *gamakas*, can greatly alter the sound of the notated *svaras* (notes); for example, some *gamakas* do not rest at all on the theoretical pitch of the notated *svara*, and instead involve oscillations between pitches either side of it [3, 4]. This oscillatory movement is particularly characteristic of the style, and can often subsume individual *svaras* [4]. The surface effect on the melodic line is that it typically has fewer stable pitch regions than many other styles. Such qualities makes it impossible for researchers who are not themselves Carnatic musicians to reliably identify *svaras* from audio recordings, although recent work has attempted to automate this task by creating descriptive transcriptions that assign *svara* names to key points in the melodic flow [5–7].

Another important feature of the style is the structural and expressive significance of motifs and phrases known as *sañcāras*, which can be defined as coherent segments of melodic movement that follow the grammar of the *rāga* (melodic framework) [2, 8]. Some musicians use the term in a narrower sense to refer to phrases that are particularly characteristic of the *rāga*, but here we use the term in its broader sense, referring to any melodic segment that is coherent from a Carnatic performer's perspective. These melodic patterns are the means through which the character of the *rāga* is expressed and form the basis of various improvisatory and compositional formats in the style [2,9]. There exists no definitive lists of all possible *sañcāras* in each *rāga*, rather the body of existing compositions and the living oral tradition of *rāga* performance act as repositories for that knowledge.

In this work we take advantage of two deep learning models, CAE [10] and FTA-Net [11], and the Dunya Carnatic corpus [12] for an improved, variable-length, and tradition-informed melodic pattern discovery in Carnatic Music. We first introduce the collections of data we build on top of, and we present our process relating each decision with the tradition-specific characteristics introduced in Section 1.1. We also include an empirical evaluation against expert annotations, brief musicological discussion of results and make available all code, features, models, results and an interactive application to explore them.

1.1 Characteristics of Carnatic Music

Many of the decisions taken in developing the methodology presented here are informed by certain characteristics of Carnatic Music and *sañcāras*, which to varying extents may be shared with other musical traditions around the world. We list some of those most relevant to this task, assigning each a unique code which will be referenced later.

CHAR1. *Sañcāras* conform in some ways to concepts of phrase more widely held; for example in Western Art Music, phrases are understood as influenced by gestalt

[1] https://github.com/MTG/searching_for_sancaras

Attribution: T. Nuttall, G. Plaja-Roglans, L. Pearson, X. Serra, "In Search of Sañcāras: Tradition-informed Repeated Melodic Pattern Recognition in Carnatic Music", in *Proc. of the 23rd Int. Society for Music Information Retrieval Conf.,* Bengaluru, India, 2022.

principles wherein segments are separated by features such as silence and long periods of stasis [13]. Also, as in other musical styles, longer phrases/*sañcāras* can be understood as comprising shorter but still meaningful segments [14].

CHAR2. The ideal boundary between these shorter segments within a longer phrase is not always clear-cut, as there might be no point of rest or silence between them. Based on discussions with the expert annotator during this project, it is clear that often a longer phrase may be plausibly segmented in two or three different ways. However, there are rules of *rāga* grammar, and understandings of typical *sañcāras* that restrict the number of plausible segmentation points. This is similar to the segmentation of Western Art Music, where in any given section, several plausible segmentations may be made, but where the options are not unlimited [15].

CHAR3. A given *sañcāra* when repeated in a performance may be immediately preceded or followed by different melodic material. This feature is relatively common in many musical styles.

CHAR4. When a *sañcāra* is repeated, it is often elaborated on, which involves the insertion of additional *svaras* and *gamakas*. So the same basic or underlying *sañcāra*/phrase may appear many times in a composition with different elaborations. The commonly occurs in the style because the main compositional format, the *kriti*, has a theme and variation structure.

CHAR5. There can be tempo variations between instances of the same *sañcāra*. This can range from minor fluctuations due to extemporisation, to playing a *sañcāra* at double or half the speed it was originally performed, which is a particular feature of some musical formats such as the *varṇam*.

CHAR6. Single performances within a concert can typically range from between approximately 6 to 60 minutes in length. In the longer examples, the presentation is made up of different musical formats, often involving a composition (e.g., a *kriti*) as well as a number of more extemporised formats (*ālāpana*, *niraval*, *kalpana svaram*), based on *sañcāras* and the *rāga* grammar. This freedom for performers to combine different musical formats to create a larger whole is a typical feature of the style.

CHAR7. A final characteristic feature of the style is its instrumentation. The most widely found contemporary ensemble consists of a vocalist accompanied by violin, *mridangam* (double ended drum) and *tambura* (plucked lute, which creates the drone). Other ensembles led by instrumentalists can also be found.

1.2 Related Work

The automatic retrieval of melodic patterns is a prominent problem in MIR given its applications for music analysis [16], segmentation [17] and music theory [18]. Despite efforts to settle consensus in melodic pattern discovery [19, 20], said task may be built on top of diverse representations of audio signals, such as mel-frequency cepstral coefficients (MFCC) [21], chroma [22], or predominant pitch time-series [23]. In a Carnatic Music context, most existing work relies on processes that consider pairwise distance metrics, such as dynamic time-warping (DTW), between sub-sequences of pitch transcriptions of the predominant sang melody [23–29]. However, this can often be computationally expensive and does not necessarily account for some of the idiosyncrasies of the tradition. To reduce the complexity, several works notate the patterns for an optimized similarity computation [30].

In a Carnatic context, gathering musically-relevant pattern annotations for entire recordings is expensive and time-consuming, and requires the involvement of experts. The most common approach is to ask experts to judge the relevance of the retrieved patterns and evaluate via standard classification metrics [9, 23].

2. DATA

For this research we use the Dunya Carnatic corpus [12], which includes $\approx$ 500 hours or music, divided into 2,380 audio recordings and organized in 235 concerts. Up to 259 artists and 227 unique *rāga*s are present in the corpus. The Saraga dataset is extracted from such corpus and is one of the largest and more complete, open-access datasets for research on the Carnatic and Hindustani music traditions [31]. For a total of 168 Carnatic recordings in Saraga, close-microphone tracks are available. Since the vocalist is the soloist and lead performer in these recordings, for this research we create a sub-dataset comprised of such tracks (168 in total) and denote it Saraga Carnatic Vocal (SCV). All tracks have a sampling rate of 44.1kHz.

We also rely on the Saraga-Carnatic-Melody-Synth (SCMS) dataset [29], a large dataset of vocal melody ground-truth for Carnatic Music compiled using a Carnatic-specific Analysis/Synthesis method, which is currently the largest open source collection of such data and has been shown to positively impact the melodic analysis for Carnatic Music [29].

3. METHOD

Our goal is to extract and group regions of repeated melodic patterns from audio recordings of Carnatic Music performances, Figure 1 provides an overview of this process. Many design decisions are informed by the characteristics introduced in Section 1.1, where relevant, we reference these characteristics using their corresponding **CHARx** code.

3.1 Features

Two feature sets are extracted from each recording in SCV: (1) An automated transcription of the predominant sung pitch in Hz (Section 3.1.1), from which we derive a mask corresponding to silent/stable regions (Section 3.1.1), and (2) Transformation-invariant melodic features extracted using a Complex Autoencoder (CAE) from audio in CQT representation (Section 3.1.2), which we use for self-similarity computation.