		Performance		*Šāhed*	*Ist*	Pitch Set	*Radif*	
Dastgāh	*Čāhārgāh*	7	9:02	5	3	2	5	5:21
	Homāyun	5	7:35	5	4	2	3	3:50
	Māhur	6	8:37	5	2	2	3	2:03
	Navā	6	9:19	5	4	2	3	4:01
	Rāst-Panjgāh	4	5:56	4	3	2	2	2:15
	Segāh	5	8:57	5	2	2	3	2:12
	Šur	7	9:55	5	2	2	3	2:12
Āvāz	*Abuatā*	5	6:47	5	3	1	2	2:01
	Afšāri	5	9:36	5	4	1	2	1:56
	Bayāt-e Tork	5	8:03	5	3	1	2	1:47
	Dašti	5	6:53	5	4	1	2	2:10
	Esfahān	5	9:21	5	3	2	2	2:40
	Total	65	100:01	59	38	20	31	31:40

Table 1. Number of recordings in KDC according to *dastgāh* and *āvāz*. For performance and *radif* recordings, the duration is also given.

Inst.	Musician	Rec.	Dur.
voice	F. Sahebghalam, R. Zalpour	24	37:50
ney	R. Zalpour, M. Khodadadi	37	45:54
setār	M. Shaari	35	47:57
Total		96	131:41

Table 2. Number of performance and *radif* recordings in KDC according to instrument.

build a first dataset that provides a coherent and comprehensive representation of the *dastgāh* system, this first version of KDC contains the first *guše* of all 7 *dastgāh*s and 5 *āvāz*es, known as *darāmad*, which holds great significance (see section 1.1). Four professional musicians (see Table 2) have contributed to KDC with original recordings of complete performances [1] of these 12 *darāmad*s. Hence, the corpus currently contains five full renditions of the 12 *darāmad*s, namely, two vocal ones, one female and one male (with the exception of the *darāmad* of *dastgāh rāst-panjgāh*, whose female vocal rendition could not be recorded), two by *ney*, and one by *setār*. According to their own decision, the musicians occasionally recorded more than one version of a particular *darāmad*, and all of them are included in the corpus. This results in 65 performance recordings, which is the core of KDC (see Table 1).

Completeness: KDC includes a csv file with metadata and annotations for all recordings, including the corresponding *dastgāh*, *guše*, artist, instrument, recording type, and duration. Considering the methodological purpose of the corpus, the musicians who collaborate with KDC were asked to record the *šāhed* of the performed *darāmad* as a single, stable pitch. The 59 resulting recordings are included in KDC, and they were used to compute the frequency of the *šāhed*, [2] which is added as an annotation in the metadata file. The musicians who collaborate with KDC show a great interest in the project and actively propose contributions to it, according to their understanding of which data could benefit the purpose of KDC. As a consequence, some musicians recorded the *ist* of the performed *darāmad* in a separate file as a single, stable pitch, resulting in 38 recordings, the corresponding pitch set, resulting in 20 recordings, and one or more related performances according to *radif* (see Table 1). Since pattern analysis is a key task for the study of Iranian *dastgāh*s, two of the collaborator musicians manually annotated the characteristic patterns that identify the performed *guše* using the software Praat [23]. Each of them annotated two full renditions of the 12 *darāmad*s, so that only the *setār* recordings are still not annotated in terms for patterns.

Quality: all the recordings in KDC are provided in the lossless compression format flac. The recordings by M. Shaari and F. Sahebghalam were recorded in a professional recording studio at KUG, those by M. Khodadadi were recorded by the first author using a Zoom H4 recorder, and those by R. Zalpour by himself using his cell phone. The musicians, all having previous recording experience, evaluated the quality of their own recordings both in terms of sound and performance quality and agreed to include them in KDC as good representatives of their art. To validate the pitch frequencies of *šāhed*, mp3 files were generated, and R. Zalpour assessed their correctness.

Reusability: All the collaborator musicians contributed their recordings to KDC for them to be published under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0). [3] In order to ensure the reusability of KDC, we follow the FAIR principles. [4] To ensure that all recordings are findable and reusable, they are made available through the KUG's Phaidra repository. [5] Regarding accessibility, the

[1] Given its improvisational nature, there is no standard rendition of a particular *guše*. By complete performance we mean a rendition that completely conveys the performed *guše*.

[2] The pitch of the *šāhed* was computed with the pYin: Notes plugin [22] of Sonic Visualiser

[3] `https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode`

[4] `https://www.go-fair.org/fair-principles/`

[5] The following Phaidra collection is created for KDC: `https://phaidra.kug.ac.at/view/o:127195`

metadata of all recordings are also stored in MusicBrainz [6] in its original Farsi and with English aliases, and linked to their corresponding Phaidra objects. A specific protocol for interoperable structuring of Iranian *dastgāhi* music is still being developed, and is expected to be evolving along with the expansion of the corpus.

3. RESEARCH POTENTIAL OF KDC

The hypothesis that motivated the KDC project is the conviction that musicological study of Iranian *dastgāhi* music can deeply benefit from computational methods. Conversely, we argue that the characteristics of this music raise interesting research tasks for music information retrieval. In this section we point out some of the most relevant of these tasks. Since KDC was created with the purpose of researching the modal aspect of this music, tasks related to this domain are described in more detail.

3.1 Melody and mode

The study of melody in Iranian *dastgāhi* music is one of the most essential tasks, especially for the purpose of KDC, therefore pitch track extraction is essential. According to our preliminary analyses, state-of-the-art algorithms perform satisfactorily for monophonic recordings as those by voice or *ney*. However, further research is still needed to obtain equally satisfactory results for instruments with high resonance as the *setār* (see section 4). A relevant research task based on these pitch tracks is the corpus driven analysis of vibrato and *tahrir* (see Fig. 5). Their automatic identification and measurement would contribute to their deeper understanding, but also to better define school styles, or idiomatic performance for specific instruments.

Tasks conducing to the characterization of *guše*s are also of key importance. A particularly relevant one is automatic pattern analysis. A main challenge for this task from the computational point of view is evaluation. In order to contribute to that, KDC includes manual annotations by expert musicians. Analyses of each *guše*'s tonal hierarchy are also essential, especially those contributing to the understanding of concepts like *šahed* or *ist* (see section 1.1).

Given the classification system of *guše*s, *dastgāh*s and *āvāz*es, this music tradition is well suited for automatic recognition and classification tasks, which in fact are the ones most commonly explored to date (see section 1.2).

Finally, even though Iranian *dastgāhi* music is not easily represented on staff notation, a tradition of transcriptions using this notation system exists. This music then can be also taken as a challenging case for automatic transcription, for which existing scores (currently only in print) can be used for evaluation.

3.2 Other musical aspects

Iranian *dastgāhi* music offers interesting tasks for rhythm research. In metred performance, automatic *dowr-e iqāi* detection becomes a significant challenge, given the heterophonic nature of ensemble performance and the timbral and stylistic characteristic of solo performance. Expressive microtiming deviations from isochrony is a meaningful analysis for characterising performance style. Unmetred performance of *dastgāhi* music offers a great opportunity for multimodal analysis combining audio and lyrics analysis (this would require incorporating text data to KDC). Natural Language Processing methods can be applied to model the prosodic features of *aruz*, that can later be mapped to music performance.

Timbre is a very diverse and expressive feature of Iranian *dastgāhi* music performance. Combined with nuances in dynamics, singers use vowel shades to introduce variability in sustained notes. Equally, *ney* performers use a wide range of timbres from bright, clean ones to raspy, breathy ones. Automatic analysis of the timbral categories can inform about stylistic preferences and if a relationship exists with specific *guše*s. The use of dynamics is another important element in Iranian *dastgāhi* music expressivity, and its computational analysis can shed light about its systematic or stylistic use.

4. PRELIMINARY ANALYSES

In order to test the potential of KDC as described in the previous section, we carried out a series of preliminary analyses. The current size of KDC does not allow to obtain general conclusions about Iranian *dastgāhi* music. Our aim is to explore the possibilities for computational methods, and therefore to suggest further lines of research for future stages of KDC.

Since our main interest is the melodic dimension of *dastgāhi* music, our analysis started with pitch track extraction. To that aim, we applied the CREPE algorithm [24] to all monophonic recordings of KDC, that is, those by voice and *ney*. After visual and aural examination, we concluded that state-of-the-art algorithms, both for monophonic and polyphonic signals, [7] do not produce results usable for further analysis on the recordings of *setār*, an instrument with high, and sought after, resonance and common use of chords. Consequently, the recordings by *setār* were excluded for these analyses. In order to ensure reproducibility, the data, metadata and annotations used for these analyses are shared as a dataset. [8] The code and resulting plots are also shared in a public repository. [9]

4.1 *Šāhed*

In order to study *šāhed*, we compute pitch histograms from the previously extracted pitch tracks using the algorithm developed in [27]. The histograms are computed in cents aligned to the *šāhed* that KDC has annotated for each individual recording. We also compute pitch histograms for the aggregated pitch tracks of all recordings of the same

[6] The following MusicBrainz collection is created for KDC: https://musicbrainz.org/collection/6ee214c7-f116-4691-b46d-36033c8b17b0

[7] Pitch extraction was attempted on the *setār* recordings using the algorithms CREPE [24], pYin [22] and Melodia [25], the last two in their Sonic Visualiser [26] plug-in version.

[8] https://phaidra.kug.ac.at/o:127202

[9] https://github.com/Rafael-Caro/KDC-v1.0

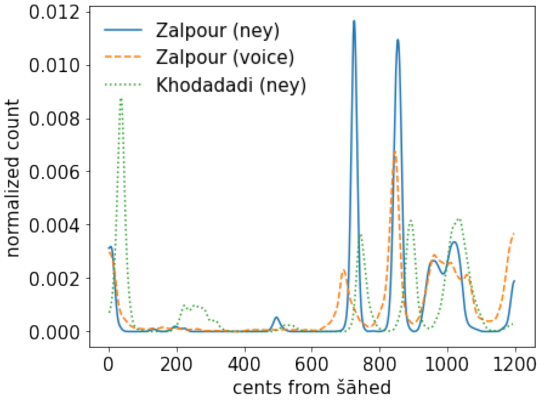

Figure 1. Folded pitch histograms of three recordings of *darāmad abuatā*

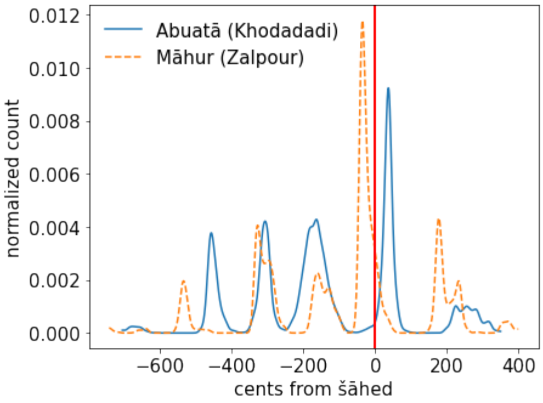

Figure 2. Pitch histograms of two *ney* recordings. The vertical line indicates the *šāhed*

dastgāh. Besides, histograms folded to one octave are also computed for individual recordings and for *dastgāh*s.

One of the most discussed issues in Iranian classical music is the concept of *šāhed*. Different definitions for this concept can be found in the existing literature. According to Farhat, "[i]n most of the Persian modes one tone assumes a conspicuously prominent role. It may or may not be the finalis" [1]. Nooshin argues that *šāhed* is "[t]he most prominent pitch, functioning as the tonal center of the mode" [28], while Talai defines it as "[t]he most frequently repeated pitch" [29]. We argue that corpus driven analyses can help to examine which of the definitions for *šāhed*, namely, "prominent" tone, "tonal center" or "most frequently repeated" tone better represents our results.

Fig. 1 shows the folded pitch histograms of three different performances of the same *guše*, *darāmad abuatā*. It can be seen how the *šāhed* is the most frequent tone in M. Khodadadi's performance, but that is not the case in the two ones by R. Zalpour, both vocal and *ney*. This preliminary result, together with other examples not mentioned here, suggests a more nuanced understanding of *šāhed*, that would escape a unique, universally applicable definition.

Mus. (inst.)	Mean	SD	Lowest	Highest
Saheb. (voice)	4.19	10.28	-10.11	28.10
Zalp. (voice)	6.92	7.26	-5.02	16.46
Zalp. (*ney*)	-3.19	16.87	-35.43	23.99
Khod. (*ney*)	13.95	13.70	-7.69	37.07

Table 3. Mean and SD of the differences in cents between the *šāhed*'s pitch in their isolated recordings and in performance per musician and instrument. The difference between the lowest and highest performance of the *šāhed* compared with the isolated recording is also given in cents.

4.2 Intonation

One interesting issue is the musicians' abstract notion of the *šāhed*'s pitch and its actual intonation in performance. Since all the collaborator musicians recorded the *šāhed* of their performances as an extra, isolated file according to their aforementioned abstract notion, we can compare its pitch in both contexts. Fig. 2 shows how M. Khodadadi performs the *šāhed* 37.07 cents higher than the isolated version he recorded himself, whilst R. Zalpour performs it 35.43 cents lower. We computed the difference between isolated *šāhed* and its performed version (obtained from the value of the closest peak in the histogram) for all recordings (excluding *setār*), and calculated the mean and SD for each musician and instrument, as it can be seen in Table 3. The results show a great variability in the divergence between conceptualized and performed *šāhed*, frequently reaching easily audible deviations.

These observations points to an interesting line of future research regarding the perception of intervallic structures by musicians. The specific width of the intervals of a particular *guše*, an essential element for the conceptualization of Iranian *dastgāhi* music, is part of the implicit knowledge of each individual musician, enacted mainly during performance. Their intellectual formulation, as the examples here analysed suggest, might no correspond with this implicit knowledge.

4.3 Pitch alteration

In some *guše*s specific degrees can be rendered by two neighbouring pitches. For example, in the case of *darāmad afšāri*, the degree above the *šāhed* can be performed at 119 cents above it, which is considered the main tone for this degree, or at 204 cents over the *šāhed*, which is considered an altered version of the same degree, as it can be seen in Fig. 3 (cent values computed according to [1]). These altered pitches are known as *not-e moteġayer* [1,30]. The use of this pitch alteration can be observed in pitch histograms as flat peaks, as it can be seen in Fig. 4 around 200 cents. These results point to another interesting line of research, for the computational detection of *not-e moteġayer* that can contribute to the better characterization of *guše*s.

4.4 Vibrato and *tahrir*

In order to analyse vibrato, we ran the Vibrato algorithm available in the Essentia library [31] on all recordings, and

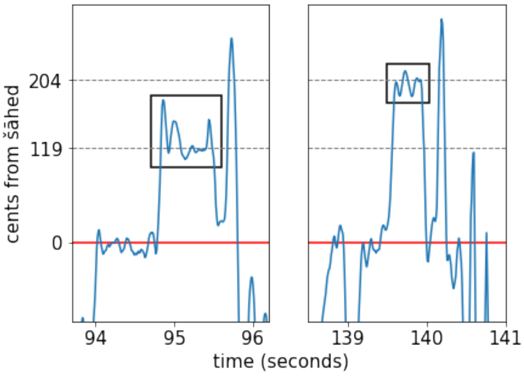

Figure 3. Two excerpts of the the pitch track of the of the vocal recording of *darāmad afšāri* by F. Sahebghalam. The horizontal red line indicates the *šāhed*, the gray line at 119 cents indicates the main pitch, whose corresponding *not-e moteġayer* is indicated by the gray line at 204 cents. The boxes mark the performance of these two notes.

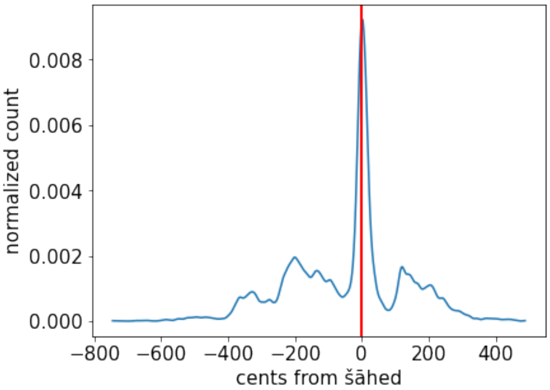

Figure 4. Pitch histogram of the vocal recording of *darāmad afšāri* by F. Sahebghalam. The vertical line indicates the *šāhed*

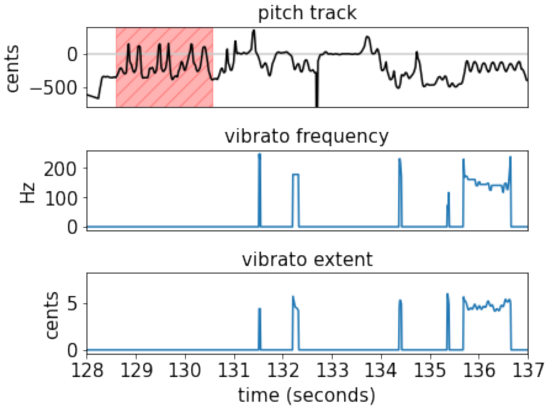

Figure 5. Excerpt of the recording of *darāmad abuatā* by F. Sahebghalam. Upper plot shows pitch track (horizontal line indicates *šāhed*), middle plot detected vibrato frequency, and lower plot detected vibrato extent. Highlighted section shows *tahrir*.

the results were plotted against the pitch track (see Fig. 5 for an example). These results, stored as csv files, together with the plots for all recordings are available in the code repository. Through visual evaluation of the results we argue that further research is required in order to improve automatic detection of vibrato for this tradition.

Dastgāhi music is an idiomatic tradition. Vocal and instrumental performances are characterized by specific techniques and sound qualities, being vibrato a relevant one. Its extent, frequency and occurrence are idiomatically specified by instrument, school or personal style. Automatic vibrato detection and measurement can contribute to characterizing idiomatic styles. Another important vocal technique is *tahrir*, which consists in wide cyclic pitch alteration of a particular tone in terms of pitch and timber. The pitch fluctuation is not realized as glissandi but jumps, and this might explain why it is not detected by the used algorithm, as shown in Fig. 5. Analysing the difference between these two techniques would help to understand the idiomatic practices in Iranian *dastgāhi* music.

5. CONCLUSIONS AND FUTURE WORK

With KDC we set a very ambitious goal: to create an open, well curated, ever growing corpus of Iranian *dastgāhi* music data, metadata and expert annotations which can contribute to the development of a coherent line of computational research for this music tradition. This is a task that requires collaborative efforts from musicologists, musicians and computer engineers. KDC is still in its infancy, but its first version presented in this paper is complete enough to test state-of-the-art methodologies, obtain preliminary results, and consequently propose new directions for future research.

The expansion of KDC is and is going to be a constant future task. As any other aspect of our lives in the past two years, the COVID-19 pandemic has hindered the development of KDC with the suspension of international travels (economic sanctions against Iran prevent working with musicians living in the country from outside). Besides the addition of recordings that cover new *gušes*, instruments and schools, increasing the number of expert annotations is one of our short term goals. In a first stage, we aim at having the 5 renditions of the whole set of 12 *daramāds* annotated at least by three different musicians, so that we can start studying this phenomenon from the perception of the performers.

Regarding computational analysis, a short term goal is to seek collaborations to perform automatic pattern analysis. This is a fundamental element for the characterization of *gušes*, and most of our collaborator musicians have highlighted the importance of this task and their interest in it. The annotations currently existing in KDC are aimed to this goal. Other tasks that are of our interest, and for which collaboration with computer engineers is also necessary, is the improvement of pitch track extraction for instruments with great resonance as *setār*. Finally, collaboration will also be sought for the development of methods for automatic *tahrir* detection and analysis.

6. ACKNOWLEDGEMENTS

The authors would like to express their sincerest gratitude and acknowledgement to their collaborator musicians Masoud Shaari, Farahnaz Sahebghalam, Reza Zalpour and Mohammad Khodadadi, not only for contributing their recordings and ideas to KDC, but also for their enthusiastic support to the project. We would also like to thank Dr. Negar Bouban and Reza Zalpour for their very valuable manual annotations of melodic patterns. Finally, we thank Kunstuniversität Graz for the financial support that makes KDC possible.

7. REFERENCES

[1] H. Farhat, *The Dastgāh Concept in Persian Music*. Cambridge and New York: Cambridge University Press, 1990.

[2] T. Tsuge, “Rhythmic aspects of the Âvâz in persian music,” *Ethnomusicology*, vol. 14, no. 2, pp. 205–227, 1970.

[3] F. Amoozegar-Fassie, “The poetics of persian music: The intimate correlation between prosody and persian classical music,” Ph.D. dissertation, Faculty of Graduate Studies, University of British Columbia, UK, 2008.

[4] M. Azadehfar, *Rhythmic structure in Iranian music*. Tehran: Tehran University Press, 2017.

[5] N. Bouban, “Barresi-ye Āvāi-ye ritm dar pāyehā-ye vājegan-e š’er va nasr-e fārsi-ye rasmi [the phonetic study of rhythm in units of poems and writings in standard farsi],” *Majaleh-ye pajuheš-hā-ye zabānšenasi*, vol. 1, no. 1, pp. 101–122, 2010.

[6] P. Heydarian, “Music note recognition for santoor,” Master’s thesis, Tarbiat Modarres University, Iran, 2000.

[7] P. Heydarian and J. D. Reiss, “The persian music and the santur instrument,” in *Proc. of the 6th Int. Society for Music Information Retrieval Conf.*, London, UK, 2005, pp. 524–527.

[8] P. Heydarian and L. Jones, “Measurement and calculation of the parameters of santur,” in *Proc. of the Annual Conf. of the Canadian Acoustical Association (CAA)*, Vancouver, Columbia, 2008.

[9] ——, “Tonic and scale recognition in persian audio musical signals,” in *12th Int. Conf. on Signal Processing (ICSP)*, Hangzhou, China, 2014.

[10] P. Heydarian, L. Jones, and A. Seago, “Automatic mode estimation of Persian musical signals,” in *133rd Audio Engineering Society Convention*, San Francisco, USA, 2012.

[11] P. Heydarian, “Automatic recognition of persian musical modes in audio musical signals,” Ph.D. dissertation, Faculty of Art, Architecture and Design, London Metropolitan University, UK, 2016.

[12] P. Heydarian and D. Bainbridge, “Dastgàh recognition in Iranian music: Different features and optimized parameters,” in *6th International Conference on Digital Libraries for Musicology (DLfM ’19), November 9, 2019, The Hague, Netherlands.*, 2019, pp. 53–57.

[13] N. Darabi, N. H. Azimi, and H. Nojumi, “Recognition of dastgah and maqam for persian music with detecting skeletal melodic models,” in *Proc. SPS-DARTS 2006-The second annual IEEE Benelux/DSP Valley Signal Processing Symposium*, Antwerp, Belgium, 2006.

[14] S. Abdoli, “Iranian traditional music dastgah classification,” in *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011*, 2011, pp. 275–280.

[15] M. A. Layegh, S. Haghipour, and Y. N. Sarem, “Classification of the radif of mirza abdollah a canonic repertoire of persian music using svm method,” *Gazi University Journal of Science*, vol. 1, no. 4, pp. 57–256, 2013.

[16] L. Ciamarone, B. Bozkurt, and X. Serra, “Automatic Dastgah Recognition Using Markov Models,” in *Proceedings of the 14th International Symposium on Computer Music Multidisciplinary Research*, 2019, pp. 51–58.

[17] S. Shafiei, “Analysis of vocal ornamentation in iranian classical music,” in *Proc. of the 16th Sound and Music Computing Conf.*, Málaga, Spain, 2019, pp. 437–441.

[18] F. Sanati, “An investigation on the value of intervals in persian music,” Master’s thesis, Department of Music, University of Jyväskylä, Finland, 2020.

[19] S. Malekzadeh, M. Samami, S. R. Azar, and M. Rayegan, “Classical music generation in distinct dastgahs with alimnet ACGAN,” *CoRR*, vol. abs/1901.04696, 2019.

[20] P. Heydarian and J. D. Reiss, “A database for persian music,” in *Proc. of the Digital Music Research Network Summer Conference (DMRN 2005)*, Glasgow, UK, 2005.

[21] X. Serra, “Creating research corpora for the computational study of music: the case of the compmusic project,” in *53rd AES International Conference on Semantic Audio*, New York, USA, 2014, pp. 1–9.

[22] M. Mauch and S. Dixon, “pYIN: A fundamental frequency estimator using probabilistic threshold distributions,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, Florence, Italy, 2014, pp. 659–663.

[23] P. Boersma and D. Weenink, “Praat: doing phonetics by computer [computer program],” 1992–2022. [Online]. Available: https://www.praat.org

[24] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 161–165.

[25] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, pp. 1759–1770, 2012.

[26] C. Cannam, C. Landone, and M. Sandler, "Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files," in *Proceedings of the ACM Multimedia 2010 International Conference*, Firenze, Italy, October 2010, pp. 1467–1468.

[27] G. K. Koduri, V. Ishwar, J. Serrà, and X. Serra, "Intonation analysis of rāgas in carnatic music," *Journal of New Music Research*, vol. 43, pp. 73–94, 2014.

[28] L. Nooshin, "The song of the nightingale: Processes of improvisation in dastgāh Segāh (Iranian classical music)," *British Journal of Ethnomusicology*, vol. 7, no. 1, pp. 69–116, 1998.

[29] D. Talai, "A New Approach to the Theory of Persian Art Music: The Radīf and the Modal System," in *Garland Encyclopedia of World Music Volume 6: The Middle East*, V. Danielson, S. Marcus, and D. Reynolds, Eds. New York: Routledge, 2001, pp. 894–903.

[30] H. Asadi, "Theoretical foundation of persian classical music: Dastgāh as a muti-modal cycle," *Mahoor*, vol. 6, pp. 43–56, 2004.

[31] D. Bogdanov, N. Wack, E. Gómez Gutiérrez, S. Gulati, H. Boyer, O. Mayor, G. Roma Trepat, J. Salamon, J. R. Zapata González, and X. Serra, "Essentia: An audio analysis library for music information retrieval," in *Proc. of the 14th Int. Society for Music Information Retrieval Conf.*, Curitiba, Brazil, 2013, pp. 493–498.

INACCURATE PREDICTION OR GENRE EVOLUTION? RETHINKING GENRE CLASSIFICATION

Ke Nie
University of California, San Diego
knie@ucsd.edu

ABSTRACT

The existing MIR research on genre classification primarily focuses on how to classify a song into the "correct" genre while downplaying the fact that genres mutate over time and in response to social change in terms of their musical properties. Songs claiming the same genre can sound very different if they are released years apart, and genres may revive musical traditions from the past. In this paper, I show that the performance of genre classifiers fluctuates as genres evolve. Unsatisfactory performance of the classifiers may not indicate algorithmic flaws but rather the change of genre characteristics. I demonstrate this by studying the case of Chinese Hip-Hop music. Specifically, I collected and analyzed 69,427 songs from four genres (Hip-Hop, Pop, Rock, and Folk) released on a Chinese music platform between 2009 and 2019. Using classifiers trained from the songs in different year cohorts to predict the genre of all the songs, I show how genre classifiers can be used to detect the stylistic shift in Hip-Hop that happened during this period. The paper thus offers a novel, sociological perspective on contending with the much-challenged idea of improving genre classification accuracy for its own sake. However, instead of questioning the effort, I argue that MIR research on genre classification can be helpful for studying genre as a social construct and cultural phenomenon if the pursuit of prediction performance and the cultural meaning of inaccurate prediction are carefully balanced.

1. INTRODUCTION

Past MIR research on genre classification has primarily strived to find ways for algorithms to correctly detect music genres. Numerous studies have experimented with various data sources, algorithmic approaches, and evaluation metrics to improve algorithmic attempts to grasp the characteristics of music genres [1-7]. Most of the existing studies rely on fixed datasets and metrics to evaluate the performance of classifiers for the convenience of comparison.

One major concern in this area, however, is that genre is an ambiguous concept [8]. After all, what distinguishes one genre from another is *subjective*, since it is "based upon subjective responses with little inter-participant consensus" [3]. This is because the boundaries according to which genres are distinguished are not entirely rooted in the musical or sonic elements; rather, they are based on people's understanding of the difference between genres, which depends on their cognitive perception and cultural knowledge [3,9]. In fact, research has found that human evaluators may also be ambivalent about the classification of genres, which problematizes the very idea of genre and challenges the purpose of classification tasks [10, 11].

Adding to this line of questioning, this paper underscores the social dimension of genres that further complicates the concept by exploring critical implications for the design and implementation of genre classifiers. I argue that genres are social constructs that constantly change over time and in response to social, economic, and political pressure [12-15]; therefore, if genres evolve, the performance of genre classifiers trained on songs from precedented cohorts will fluctuate in predicting future cohorts. Given this feature, I argue that the "inaccurate" prediction of genre classifiers can be used to detect and map the evolution of genres. I demonstrate this point by presenting a case study of Chinese Hip-Hop music. Using songs from different year cohorts as the training set, I show that the performance of the classifiers mutates over time as the genre evolves. The findings demonstrate that prediction accuracy may not be the most valuable feature of algorithmic classifiers and that inaccurate predictions may suggest genre evolution.

This paper is organized as follows. In Section 2, I review the current MIR research on genre classification tasks and identify the common concern shared in this research area regarding the ambiguity of the genre concept. I propose taking a deeper look at the social dimension of genres, a generally understudied subject in the MIR field, as a way to advance our understanding of the evaluation of genre classifiers. In Section 3, I present the case of Hip-Hop music in China, a genre that has enjoyed a dramatic turn of events in the past decade. Drawing from different pieces of evidence, in Section 4 I show how the evolution of genre could complicate the performance of genre classifiers and how classifiers can be used to help understand the development of a genre. I then conclude the paper by discussing the impacts of genre evolution on the design and performance of genre classifiers.

2. RELATED WORK

The existing MIR literature on genre classification is primarily aimed at exploring, devising, and experimenting with algorithmic designs that could be used to understand and predict music genres automatically. These studies can be roughly categorized into three main themes.

First and foremost, researchers have attempted to identify different kinds of *data sources* that are useful for earmarking the characteristics of music genres. Two types of sources that are deemed significant are the music content — e.g., music and sonic features extracted from the digital signals of the songs [1] — and the music context — e.g., information on the performers, musicians, communities, and other social-cultural features associated with the genre [2]. Thus, discovering relevant data sources is the foundational work for performing genre classification tasks.

Second, contingent upon the data sources mentioned above, researchers have been trying to improve on the *algorithmic approaches* to achieve better performance in genre classification tasks. For example, researchers who use music content for the task have been experimenting with various types of acoustic representations (MFCCs, spectrograms, etc.) and machine learning architectures (SVM, CNN, etc.) [4,5,7,8,16]. There are also attempts to create a multimodal framework that juxtaposes both music content and music context data, such as the lyrics, to enhance the performance classification task [6]. The primary goal here is to improve the accuracy of the classification algorithms, which are usually tested on an existing dataset, such as GTZAN [1].

Studies aligned with the above two themes comprise much of the MIR research on genre classification. Their common goal is to find ways to enhance genre classification performance, which is usually evaluated based on metrics such as prediction accuracy. While these studies have explored pathbreaking methods through which genre classification tasks are performed at higher accuracy, some scholars have also reflected on the *evaluation metrics* of genre classifiers, arguing that "accuracy is not enough," as genre classification is essentially a task of musical recognition rather than classification accuracy [17,18]. Furthermore, researchers highlight the ambiguity of the concept of "genre" itself, pointing out that the notion is essentially "subjective" [3]; therefore, human evaluation needs to be included in the process [19]. While there is suspicion about the validity of genre classification tasks due to the ambiguity of the genre concept, most scholars still acknowledge the relevance of the research while looking for instruments to mitigate the problem of subjectivity [11]. Previous studies from this third, reflective approach have shared incisive thoughts on how to refine algorithmic frameworks for better classification performance.

Following past reflections on this issue, I want to point out another dimension that could further complicate but also help the task. As many contend, the concept of genre is subjective in the sense that the exact boundaries between genres are subject to human evaluation, which diverges across different human evaluators, leading to ambivalent classification outcomes in the first place [3,19]. Genre is, therefore, a human construct that depends on mutual agreement among evaluators or the community as a whole. This view is close to that of the social scientists and business scholars of the music industry, who have long argued that genres are social constructs that are classificatory apparatuses used by producers, consumers, and intermediaries to make sense of a distinct type of music [12-14]. This sociological view of genre claims that while music content may matter in the identification of genre, genres are, above all, divided by communities. This helps explain cases where different genres sound similar to each other or where the same genre may sound very different in different regions. For example, sociologist William Roy pointed out that "hillbilly music" and "race records" in the 1920s America were two genres that were associated with the same type of music, although the former was used as a market label for African American audience while the latter was for the "rural whites" due to segregations at the time [20]. They are, nevertheless, considered two disparate "genres" as they are tied to different communities.

A more important component of the sociological view of genre that is usually underplayed in the MIR research is that genres can evolve. In other words, the way in which genre classification systems upon which producers, consumers, and intermediaries agree to distinguish different types of music can change across time and localities [12,13]. For example, the "Rock" genre has experienced significant changes in terms of its associated sound since its inception in the 1950s, so music labeled as "Rock" today sounds very different from music given the same label 70 years ago. Moreover, genres may also influence each other in terms of how they sound. For example, "Nu-Metal," which prevailed in the late 1990s and early 2000s, arguably sounds more similar to Hip-Hop contemporaries than their Metal predecessors from the 1970s.

Additionally, genre categories may also evolve in response to external shocks to the industry or the social environment in general. One example is the case of censorship, in which censored genres have to change their content in response to political pressure [15]. Similarly, advancements in technology or sound devices also help to define or shift music genres, such as the use of the Roland TR-808 as the defining sound of early-1980s Hip-Hop music. The idea here is that genres are constantly shaped and reshaped by social, economic, and political forces; therefore, they cannot be seen as fixed, immutable entities.

The mutability of genre adds a further question to the design of genre classification algorithms in terms of their data sources, algorithmic architecture, and evaluation metrics. This paper attempts to address the issue of evaluation specifically. If genres evolve over time, one should then expect that the classification accuracy of a genre classifier will drop when it is used to predict the songs that are released a long time apart from the songs that were used to train the classifier. Specifically, we may expect that a classifier that was trained on a set of songs released in the 1980s will have a harder time predicting the correct genre of songs from the 2010s than songs from the 1990s. In this case, the accuracy metric is not useful for understanding the performance of the classifier.

In this case, evaluation metrics may have a richer meaning than simply as a criterion for making the genre distinction: they may be used to understand how genre evolves or how genres influence each other. For example, inaccurate predictions may not necessarily indicate algorithmic flaws; they may imply that the actual genre is moving toward the predicted genre in terms of its sound. The trickiest part of making such a claim, however, is how to distinguish genre evolution from the drawbacks of the algorithmic design when prediction accuracy is low. This will be discussed further in the analysis and discussion sections below.