music programs, to private lessons, to music conservatories [29]. A vibrant music scene also requires equitable access to funding, the fair enforcement of city ordinances, and the cultivation of shared social spaces [28]. Are the small business loans needed to support music infrastructure equitably available across demographic sectors? Are noise ordinances and venue regulations enforced and policed evenly across neighborhoods? Does the local government sponsor arts festivals and programs available to and welcoming of all citizens?

Demographics play a key role in establishing and invigorating a music community; for example, young people such as university students are more likely to go out to music events [4, 11, 12]. Denser populations often mean stronger social networking, associated with vibrant music scenes [19]. Lastly, as the population increases, the demand for live music also often increases, which can strengthen a music community [18, 21].

3. MUSIC EVENT DATA

To obtain a rough estimate of the strength of a city's music scene, we calculated its *live music event rate* (LMER). Specifically, we calculate the number of music events per year per 100,000 residents in a given geographic area.

We compiled the LocalifyMusicEvents-USA-2019 dataset [3] that consists of information for 308,051 music events from 1,139 cities in the United States, each with a population of 10,000 or more residents. The music event information was collected using custom web scrapers in late 2018 and all of 2019. We developed two web scrapers, one for BandsInTown, a comprehensive music event website, and one for Facebook, a popular social network. The BandsInTown scraper relied on snowball sampling [30] to build growing lists of artist and venue pages. Each page was scraped for events every 30 days. New artists and venues were continually added based on similarities between artists, artists playing at different venues, and artists playing events with other artists. By contrast, our Facebook scraper collected music event information by cycling repeatedly over a large list of cities continually.

To calculate LMER, we first count the total number of music events listed in the city during 2019 by matching the city and state names. We then divided by the city population according to the US Census population estimates for 2019.

Our dataset also contains 28 city-level socioeconomic indicators obtained from three different websites: Census Reporter [4], DataUSA [5], and Census Quickfacts [6]. These indicators are partitioned into six groups (transportation, population, economics, age, education, race) and are summarized in Table 2.

3.1 Music Havens

We divide the cities into three subsets grouped by population: small (pop. 10K-100K), medium (pop. 100K-500K), and large (pop. 500K or more). For each subset, the top ten "music havens" (cities with the largest LMER) are shown in Table 3.

Size	Principal City	Pop.	LMER
Small	South Burlington, VT	19162	0.036
	Steamboat Springs, CO	12928	0.020
	Asheville, NC	92859	0.019
	Santa Cruz, CA	64605	0.019
	Key West, FL	24843	0.017
	Burlington, VT	42545	0.017
	Ithaca, NY	30569	0.016
	Fredericksburg, TX	11245	0.016
	Lahaina, HI	12776	0.014
	Rutland, VT	15398	0.013
Medium	Salt Lake City, UT	200546	0.023
	Berkeley, CA	121353	0.016
	New Orleans, LA	390144	0.015
	Richmond, VA	230436	0.015
	Cambridge, MA	118925	0.013
	Boulder, CO	105670	0.010
	Orlando, FL	287435	0.009
	Fort Collins, CO	170245	0.009
	Minneapolis, MN	429605	0.009
	Charleston, SC	143151	0.008
Large	Denver, CO	727211	0.016
	Washington, DC	705749	0.011
	Austin, TX	979263	0.011
	Atlanta, GA	506804	0.010
	Seattle, WA	753655	0.010
	Portland, OR	653467	0.009
	Las Vegas, NV	651297	0.008
	San Francisco, CA	881549	0.007
	Philadelphia, PA	1584064	0.005
	Dallas, TX	1343565	0.005

Table 3: Top 10 "Music Havens" for Small (pop. 10K-100K), Medium (pop. 100K-500K), and Large Cities (pop. 500K+) ranked by Live Music Evert Rate (LMER)

In these three lists, we find several popular tourist towns, e.g., Steamboat Springs, CO, Key West, FL, Las Vegas, NV, as well as college towns, e.g., Burlington & South Burlington, VT (U. of Vermont), Ithaca, NY (Cornell U.), Berkeley, CA (U. of California). We also find several cities, like Austin, TX, Asheville, NC, Atlanta, GA, and New Orleans, LA which all frequently appear on lists of top destinations for "music tourism" in the United States. [7]

Many of the medium and large cities that we identify as music havens also appear on a list of American cities with the most concerts per capita based on data from Seat-

[3] Data can be found at `https://github.com/JimiLab/LocalifyMusicEventData`.

[4] https://censusreporter.org/

[5] https://datausa.io/

[6] https://www.census.gov/quickfacts/

[7] https://www.thrillist.com/travel/nation/best-cities-for-live-music-new-york-memphis-asheville-and-austin

Indicator	Description	Source
Transportation		
Mean Travel Time	Mean travel time to work (minutes)	Census Reporter
Public Transit	% of population who took public transit (e.g., buses)	Census Reporter
Bicycle	% of population who biked	Census Reporter
Walkability	% of population who walked	Census Reporter
Public Transit+Bicycle+Walkability	% of population who took public transit, biked, or walked	Census Reporter
Population		
Population Density	Population per square mile	Census Reporter
10 Year Population Growth	Population growth from 2010 (acc. to Census Quickfacts, as of April 1, 2010) to 2019 (acc. to ACS 2019).	Census Quickfacts, Census Reporter
Migration Rate Since Previous Year	Geographic mobility - % of population who moved to city since last year.	Census Reporter
Economics		
Per Capita Income	Average income per person in city ($)	Census Reporter
Median Household Income	Median income per household in city ($)	Census Reporter
Poverty Rate	% of city population below poverty line	Census Reporter
Median Property Value	Median value of owner-occupied housing units ($)	Census Reporter
Employment Rate	Number of people employed (DataUSA) divided by 2019 population (Census Reporter)	DataUSA, Census Reporter
Median Gross Rent	Median gross rent ($), 2015-2019	Census Quickfacts
Median Owner Cost With Mortgage	Median selected monthly owner costs -with a mortgage ($), 2015-2019	Census Quickfacts
Median Owner Cost Without Mortgage	Median selected monthly owner costs -without a mortgage ($), 2015-2019	Census Quickfacts
Owner-Occupied Housing Unit Rate	Owner-occupied housing unit rate (%), 2015-2019	Census Quickfacts
Age		
Median Age	Median age of city	Census Reporter
Percent Under 18	Percentage of population under age 18	Census Reporter
Percent 18-29	Percentage of population between ages 18 and 29	Census Reporter
Percent Under 30	Percentage of population under age 30	Census Reporter
Percent 20-29	Percentage of population between ages 20 and 29	Census Reporter
Percent 10-29	Percentage of population between ages 10 and 29	Census Reporter
Age Diversity Index	Simpson's diversity index for age (0-9, 10-19, ..., 70-79, 80+)	Census Reporter
Education		
High School Or Higher	Percentage of population that are high school grads or higher	Census Reporter
Bachelor Or Higher	Percentage of population with Bachelor'\s degree or higher	Census Reporter
Postgrad Degree	Percentage of population with post-grad degree	Census Reporter
Race		
Race Diversity Index	Simpson's diversity index for ethnicity (White, Black, Native, Asian, Islander, Other, Two+, Hispanic)	Census Reporter

Table 2: A list of all 28 city-level socioeconomic indicators used and their corresponding descriptions.

Geek [8], a large ticket reselling site [31]. Their top cities include Las Vegas, NV, Nashville, TN, Austin, TX, and Denver, CO. This overlap gives us some confidence in our approach but it should be noted that the SeatGeek reports only examines large events from the top 100 grossing artists in the top 100 market areas (i.e., cities).

3.2 Music Deserts

There were 87 cities (7.6% of 1,139 cities) for which there were no events in the LocalifyMusicEvents-USA-2019 dataset. The three "music desert" cities with the largest populations were Renton, WA (pop. 101,747), Deltona, FL (pop. 92,752), and Newton, MA (pop. 88,411). When we examine all three using simple Google web searches, it is clear that there are music events taking place in these cities, as they are listed on their corresponding local websites [9] but not found when we scraped event information from our two data sources (BandsInTown, Facebook). This suggests that our dataset is incomplete due to the imperfect nature of our scraping procedure and our limited set of data sources. We will further discuss these limitations in Section 5.1.

4. CORRELATION WITH SOCIOECONOMIC INDICATORS

In this section, we explore correlations between LMER and the 28 different socioeconomic indicators using statistical testing. If we assume that LMER is a rough indicator of the strength of a local music scene, we can use it to study how music scenes are related to other aspects of our society. We have grouped the 28 socioeconomic indicators into six categories: Transportation, Population, Economics, Age, and Education & Race. The 28 indicators we sampled, as well as the sources they were collected from, are shown in Table 2.

We use a Bonferroni correction when determining statistical significance since we are conducting multiple hypothesis statistical tests [32]. Usually, in significance testing, we perform one statistical test, and obtain the correlations r and p-values p; an indicator is deemed significant if $p < \alpha$, where α is the p-value threshold. The Bonferroni correction, however, deems an indicator significant if $p < \alpha/I$, where I is the number of statistical tests (i.e., one per indicator). In our experiment, there are $I = 28$ indicators and we set α to 0.05; thus, a correlation is significant if the p-value is less than $\alpha/I =$ 1.8e-3.

Table 4 shows the correlation coefficients (or r-values) and the probability of observing the data assuming no correlation (p-value) for each of these indicators grouped in their categories, with the p-values (p) ranked from smallest (most significant) to largest (least significant) for each category.

[8] https://seatgeek.com/

[9] Renton, WA: https://rentondowntown.com/happenings/summer-concert-series/, Deltona, FL: https://www.deltonafl.gov/parks-recreation-department/events/22814, Newton, MA: https://patch.com/massachusetts/newton/newton-porchfest-2019-what-know

4.1 Transportation

As shown in Table 4, we find that there is a strong positive correlation between LMER and the first four transportation-based indicators (percentage of people who bike, take public transit, walk, or all three together). Good public transportation reduces the overall cost of attending a show [20] and enables more people to drink alcohol (more safely) at night, which is very often associated with live music events. This finding supports the claim by Terrill et al. [11] that supportive urban infrastructure is an important factor for a strong local music scene. We also considered the mean travel time to work, but did not find it to have a statically significant correlation.

4.2 Population

We observe that population density as well as one-year and ten-year population growth are all positively correlated with LMER. Van der Hoeven and Hitters [19] argue that "density and diversity provide the critical mass of participants that alternative [music] scenes need to thrive." When people are clustered together, it is easier to engage and interact with one another. This idea also relates to population growth; as people with diverse backgrounds immigrate to the city, there is an increase in opportunities for musicians to influence one another in novel ways [18].

4.3 Economics

As we discussed in Section 2.2, many researchers have suggested that having a strong local music scene is good for the local economy [6, 8, 9, 11]. This is consistent with our findings that employment rate, housing cost, per capita income, and median property values are all positively correlated with LMER. We found it interesting that the owner-occupied housing rate is negatively correlated which suggests that there are more live music events when there is a higher proportion of renters relative to homeowners. This is consistent with the observation that home affordability has dropped in the United States, making it harder for young people to own their homes [33], and as we observe in the next subsection, having a larger percentage of younger adults is positively correlated with LMER.

Concerning poverty, Harrison [25] examines how “music projects develop the skills, education levels, incomes, or occupational possibilities of participants living in material poverty, which in turn can enhance their socio-economic status.” Accordingly, we might expect a lower poverty rate where there was more live music but we did not find a statistically significant correlation between poverty rate and LMER. We assume therefore that the dynamic between poverty and music is too complex to be measured in a simple quantitative analysis.

4.4 Age

Our results show that cities with a high proportion of young people between the ages of 18 and 29 tend to have a large live music event rate. Conversely, cities with a large population of people under 18 are negatively correlated with LMER. This may be because many music venues (i.e., bars) tend to have 18-and-up and 21-and-up policies due to laws related to serving alcohol. These two results, along with the fact that neither median age nor the percentage of the population under 30 is correlated with LMER, suggest that cities with many young adults (college students, young professionals, aspiring young artists) are places where we expect to have many live music events. Conversely, cities with a relatively large percentage of families are less likely to have a high rate of music events. As mentioned in the previous subsection, age is also positively correlated to home ownership, as decreasing home affordability in the United States has caused fewer young people to own a home; this entails a negative correlation between owner-occupied housing rate and LMER, as higher housing ownership rates means fewer young people, an indicator of a lower LMER.

4.5 Education

Terrill et al. [11] suggests that "cities such as Toronto, Adelaide, Austin, and Berlin point to their large student populations as helpful factors in generating engaged audiences." Our results support this in that we find a positive correlation between LMER and cities with a high proportion of people with undergraduate or graduate degrees. This is also reflected by the fact that many of the top Music Havens in Table 3 are college towns as was discussed in Section 3.1.

4.6 Race

We did not find a significant correlation between our racial diversity index and the LMER. This might be surprising considering the associations suggested between healthy music scenes and demographic diversity [34]. In that, we only explore one indicator that explicitly relates to race, a more thorough analysis is required to explore the complex relationship between race and the strength of music scenes.

5. DISCUSSION

In this paper, we explore how using the live music event rate (LMER) is straightforward to estimate, easy to interpret, and correlated with a large and diverse set of socio-economic indicators. We found a strong positive correlation between LMER and the percentage of people who take public transit, walk, or bike. Education is also strongly correlated with LMER. We also observed that many economic indicators (e.g., employment rate, per capita income, median property value) are also correlated with LMER. Finally, there is a high live music event rate when there is a high proportion of late teens and individuals in their twenties. We did not find a significant correlation between LMER and our race diversity index. This is not to say that no relationship exists, but rather our simplistic analysis did not reveal a statistically significant correlation.

Indicator	**r**	**p**
Transportation		
Bicycle	**0.39**	**5.1e-43**
Public Transit+Bicycle+Walkability	**0.34**	**1.1e-32**
Public Transit	**0.26**	**1.8e-18**
Walkability	**0.24**	**7.7e-17**
Mean Travel Time	0.04	1.5e-01
Population		
Population Density	**0.21**	**1.5e-12**
10 Year Population Growth	**0.18**	**9.8e-10**
Migration Rate Since Previous Year	**0.13**	**1.7e-05**
Economics		
Employment Rate	**0.26**	**4.7e-19**
Owner-Occupied Housing Unit Rate	**-0.21**	**3.3e-13**
Median Owner Cost w/ Mortgage	**0.19**	**2.4e-10**
Median Owner Cost w/o Mortgage	**0.17**	**2.9e-09**
Median Property Value	**0.17**	**1.8e-08**
Per Capita Income	**0.16**	**3.7e-08**
Median Gross Rent	**0.15**	**2.7e-07**
Median Household Income	0.06	3.1e-02
Poverty Rate	-0.00	8.2e-01
Age		
Percent Under 18	**-0.24**	**7.6e-16**
Percent 20-29	**0.17**	**2.0e-08**
Percent 18-29	**0.14**	**1.8e-06**
Age Diversity Index	**0.11**	**1.1e-04**
Percent 10-29	**0.10**	**4.5e-04**
Median Age	-0.06	4.1e-02
Percent Under 30	0.02	4.1e-01
Education		
Bachelor Or Higher	**0.27**	**6.6e-20**
Postgrad Degree	**0.24**	**8.1e-16**
High School Or Higher	**0.12**	**6.6e-05**
Race		
Race Diversity Index	-0.07	1.9e-02

Table 4: Table showing the correlation coefficient (r) and p-values (p) for all 28 socioeconomic indicators across all 1,139 cities. Statistically significant indicators are in **bold** font.

5.1 Limitations

As mentioned in Section 3.2, the music events dataset was not collected for this research, but rather for our Localify.org [10] music events recommendation application. As a result, the scraping was limited in the number of data sources (Facebook & BandsInTown), collected in an ad-hoc manner (snowball sampling), only covers one year of data (2019), and only looks at one country (United States). It is probable that many active artists, venues, and events were not found using our music event data collection process.

To increase our coverage and add redundancy, we have since added additional scrapers (e.g., SongKick, Google, Eventbrite) but the music event data we have scraped in 2020 and 2021 is problematic since such a high percentage of scheduled music events were canceled due to COVID-19. However, even with a large number of scrapers, our approach necessarily ignores music events that do not have a digital footprint. This includes many underground (private house parties, DIY shows), music in religious spaces (e.g., churches), and impromptu performances (e.g., busking). Our future work will be to explore how our approach to scraping music event information might be incomplete and/or biased by taking a detailed census of music events in individual cities using more principled ethnographic techniques (e.g., observation, interviews, historical records).

As described in Section 3, the 28 socioeconomic indicators for each city were collected using three different web sources, rather than one unified source. A more systematic approach is to use one unified source for our data collection, such as ESRI datasets [11] . For certain indicators addressed in Section 2 and Table 1, such as mental/physical health, cultural heritage, and local government regulations on music, we could not find any available quantitative data sources to measure them.

5.2 Future Work

Both popular culture and academic research tend to focus on "music havens" like Nashville, TN and Seattle, WA. These are cities that are known to have a great local music scene and they reap economic, social, and cultural rewards because of it. We, by contrast, are especially interested in identifying cities with underdeveloped music scenes. We plan to study these “music deserts"’ and explore how various strategies could be used to help them strengthen their local music scenes. For example, Terrill et al. [11] outlines a number of potential strategies which include developing music-friendly government policies, creating music-focused offices and advisory boards, investing in audience development, and creating music tourism plans. We argue that cities with low local music event rates (LMER) may be good initial candidates for future research involving the study of music deserts.

6. REPRODUCIBILITY

To make our research both fully reproducible and transparent, our full LocalifyMusicEvents-USA-2019 dataset and the associated data processing code (in the form of Jupyter Notebooks) can be found at `https://github.com/JimiLab/LocalifyMusicEventData`.

7. ACKNOWLEDGEMENTS

Doug Turnbull, Tim Clerico, John Hunter, and Emmett Barry all contributed to the Localify.org music event scraper code. This research was supported by NSF grant IIS-1901330/1901168 and NEH grant HAA-280975-21. Finally, we would like to thank the anonymous ISMIR reviewers for their highly constructive feedback.

8. REFERENCES

[1] S. Baker, R. Nowak, P. Long, J. Collins, and Z. Cantillon, “Community Well-Being, Post-Industrial Music

[10] https://localify.org/

[11] https://www.esri.com/en-us/arcgis/products/arcgis-open-data

Cities and the Turn to Popular Music Heritage," in *Music Cities*. Springer, 2020, pp. 43–61.

[2] D. Wheatley and C. Bickerton, "Subjective Well-Being and Engagement in Arts, Culture and Sport," *Journal of Cultural Economics*, vol. 41, no. 1, pp. 23–45, 2017.

[3] T. M. G. (Firm), *The Austin Music Census: A Data-Driven Assessment of Austin's Commercial Music Economy*. Titan Music Group, LLC, 2015. [Online]. Available: https://books.google.com/books?id=fStOjwEACAAJ

[4] A. J. Baker, "Algorithms to Assess Music Cities: Case Study — Melbourne as a Music Capital," *SAGE Open*, vol. 7, no. 1, pp. 1–12, 2017.

[5] S. Frith, M. Brennan, M. Cloonan, and E. Webster, "'Analysing Live Music in the UK' Findings One Year into a Three-Year Research Project," *IASPM Journal*, vol. 1, no. 1, pp. 1–30, 2010.

[6] F. Holt, "The Economy of Live Music in the Digital Age," *European Journal of Cultural Studies*, vol. 13, no. 2, pp. 243–261, 2010.

[7] D. Siegel and F. Bovenkerk, *Crime and Music*. Springer, 2021.

[8] "Economic Contribution of the Venue-Based Live Music Industry in Australia," *Australasian Performing Right Association*, vol. 20, 2011.

[9] S. Hallam, I. Cross, and M. Thaut, *Oxford Handbook of Music Psychology*. Oxford University Press, 2011.

[10] A. J. Ferrer, "The Effect of Live Music on Decreasing Anxiety in Patients Undergoing Chemotherapy Treatment," *Journal of Music Therapy*, vol. 44, no. 3, pp. 242–255, 2007.

[11] A. Terrill, D. Hogarth, A. Clement, P. Assistant, and R. Francis, *The Mastering of a Music City: Key Elements, Effective Strategies and Why It's Worth Pursuing*. Music Canada, 2015.

[12] E. Webster, M. Brennan, A. Behr, M. Cloonan, and J. Ansell, "Valuing Live Music: The UK Live Music Census 2017 Report," 2018.

[13] A. Williams and A. Lal, "Correlations between Crime Rates in US Cities, and the Popularity of Rap and Hip-Hop Music," 2015.

[14] S. Oakes and G. Warnaby, "Conceptualizing the Management and Consumption of Live Music in Urban Space," *Marketing Theory*, vol. 11, no. 4, pp. 405–418, 2011.

[15] C. Pinkney and S. Robinson-Edwards, "Gangs, Music and the Mediatisation of Crime: Expressions, Violations and Validations," *Safer Communities*, 2018.

[16] S. Brown and U. Volgsten, *Music and Manipulation: On the Social Uses and Social Control of Music*. Berghahn Books, 2005.

[17] H. Shobe and D. Banis, "Music Regions and Mental Maps: Teaching Cultural Geography," *Journal of Geography*, vol. 109, no. 2, pp. 87–96, 2010.

[18] C. Gibson, "Rural Transformation and Cultural Industries: Popular Music on the New South Wales Far North Coast," *Australian Geographical Studies*, vol. 40, no. 3, pp. 337–356, 2002.

[19] A. Van der Hoeven and E. Hitters, "The Social and Cultural Values of Live Music: Sustaining Urban Live Music Ecologies," *Cities*, vol. 90, pp. 263–271, 2019.

[20] P. E. Earl, "Simon's Travel Theorem and the Demand for Live Music," *Journal of Economic Psychology*, vol. 22, no. 3, pp. 335–358, 2001.

[21] M. Brennan, "Live Music History," *The SAGE Handbook of Popular Music*, pp. 207–222, 2015.

[22] K. Swanwick, *Music, Mind and Education*. Routledge, 2003.

[23] M. Davyd, "GRASSROOTS MUSIC VENUES (GMVs)–DEFINITION," 2016.

[24] D. A. Economics, "The Economic, Social and Cultural Contribution of Venue-based Live Music in Victoria," *Melbourne: Arts Victoria*, 2011.

[25] K. Harrison, "Music, Health, and Socio-Economic Status: A Perspective on Urban Poverty in Canada," *Yearbook for Traditional Music*, vol. 45, p. 58–73, 2013.

[26] E. Klinenberg, *Palaces for the People: How Social Infrastructure Can Help Fight Inequality, Polarization, and the Decline of Civic Life*. Crown, 2018.

[27] R. D. Putnam, "Bowling Alone: America's Declining Social Capital," in *Culture and Politics*. Springer, 2000, pp. 223–234.

[28] A. McGraw, "Mapping Sonic and Affective Geographies in Richmond, Virginia," *Sonic Identity at the Margins*, p. 19, 2022.

[29] L. K. Richerme, "Equity via Relations of Equality: Bridging the Classroom-Society Divide," *International Journal of Music Education*, vol. 39, no. 4, pp. 492–503, 2021.

[30] L. A. Goodman, "Snowball Sampling," *The Annals of Mathematical Statistics*, pp. 148–170, 1961.

[31] "Which U.S. Cities Get the Most Concerts?" https://seatgeek.com/tba/music/which-u-s-cities-get-the-most-concerts/.

[32] C. Y. Wijaya, "Multiple Hypothesis Testing Correction for Data Scientist," Oct 2021. [Online]. Available: https://towardsdatascience.com/multiple-hypothesis-testing-correction-for-data-scientist-46d3a3d1611d

[33] G. Paz-Pardo, “This is Why Owning a Home is More Difficult than Ever for Young People,” Jan 2022. [Online]. Available: https://www.weforum.org/agenda/2022/01/younger-generations-homeownership-housing-market-wealth-inequality/

[34] C. Ballico and A. Watson, *Music Cities: Evaluating a Global Cultural Policy Concept*. Springer, 2020.

LEARNING MULTI-LEVEL REPRESENTATIONS FOR HIERARCHICAL MUSIC STRUCTURE ANALYSIS

Morgan Buisson[1] **Brian McFee**[2,3] **Slim Essid**[1] **Hélène C. Crayencour**[4]

[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2] Music and Audio Research Laboratory, New York University, USA
[3] Center of Data Science, New York University, USA
[4] L2S, CNRS-Univ.Paris-Sud-CentraleSupélec, France

ABSTRACT

Recent work in music structure analysis has shown the potential of deep features to highlight the underlying structure of music audio signals. Despite promising results achieved by such representations, dealing with the inherent hierarchical aspect of music structure remains a challenging problem. Because different levels of segmentation can be considered as equally valid, specifically designed representations should be optimized to improve hierarchical structure analysis. In this work, unsupervised learning of such representations using a contrastive approach operating at different time-scales is explored. The proposed system is evaluated on flat and multi-level music segmentation. By leveraging both time and the hierarchical organization of music structure, we show that the obtained deep embeddings can encode meaningful patterns and improve segmentation at various levels of granularity.

1. INTRODUCTION

Common approaches for music structure analysis can usually be broken down into two main steps: segmentation and structural grouping [1]. The segmentation task aims at determining the boundary locations between consecutive musical sections while the grouping step consists in assigning labels to each of the retrieved segments based on certain musical similarities. Traditional algorithms for structural segmentation use different hand-crafted features [2] and their combinations to detect abrupt changes of particular musical characteristics or repetitions of certain patterns throughout the song. However, recent progress in deep learning has given rise to new systems automatically producing more robust representations which manage to combine several acoustic characteristics to enhance the recognition of musical sections [3–5]. While these representations have consistently improved downstream segmentation methods, their performance is mostly evaluated on *flat* structural annotations and metrics. However, musical structure naturally exhibits a hierarchical organization where a variety of cues can trigger boundaries between segments of different length [6], depending on the time scale at which they are observed [1]. At the lowest temporal level, short segments might only last a few measures. Coarser annotation levels are generally composed of longer segments, grouping various shorter fragments into larger musically meaningful units (ex: chorus, verse ...). This nested organization of musical events at different levels holds crucial information about music structure [7]. While the original task of music structure analysis is commonly performed at a pre-defined level of granularity (i.e. *flat* segmentation), the problem of *hierarchical* structural analysis consists in predicting a set of segmentation candidates called hierarchy, ordered by their amount of detail (from the coarsest to the most refined level). Recent efforts have been made to compile datasets with multi-level structural annotations [2, 8], which greatly facilitates the study of musical structure in a hierarchical manner. Although a few methods have been proposed for such task, the role of hierarchy in music structure has never been explicitly considered while building better-suited representations of audio music signals prior to segmentation.

1.1 Our contributions

In this work, we propose a deep unsupervised hierarchical metric learning approach for music structure analysis. We show that leveraging both time information and the hierarchical structure of music can help building efficient representations for music segmentation at different levels without requiring any supervision from structural annotations. We demonstrate the effectiveness of these representations for both flat and multi-level segmentation and show that they can accommodate structural annotations of varying styles and levels.

1.2 Related work

The method proposed here builds upon recent work in music structure analysis devoted to finding efficient representations using deep learning methods to improve already existing downstream algorithms. The work by McCallum [3] proposes an unsupervised method to learn deep features

Attribution: F. Author, S. Author, and T. Author, “Learning Multi-Level Representations for Hierarchical Music Structure Analysis”, in *Proc. of the 23rd Int. Society for Music Information Retrieval Conf.,* Bengaluru, India, 2022.

using a triplet-based approach. It relies on the assumption that frames temporally close to each other are more likely to belong to the same musical section than those separated by a certain amount of time. Therefore, triplets are sampled in such a way that the temporal distance between the anchor point and the positive example is smaller than the distance separating the anchor from the negative example. Wang et al. [4] adopt a similar approach by using structural annotations in a supervised fashion to mine informative sets of frames.

One of the main challenges in estimating the structure of a musical piece is to account for the different temporal levels at which it can be decomposed. Up to now, only a few approaches have been proposed for the task of multi-level segmentation. McFee and Ellis [9] use spectral clustering to decompose an enhanced self-similarity matrix and produce segmentations at different temporal levels. This approach is later improved by Tralie and McFee [10] where the input self-similarity matrix is obtained by combining different features using Similarity Network Fusion. Salamon et al. [5] further extend this method by employing two types of deep embeddings along with CQT features. They capture local timbral patterns with few shot-learning and long-term similarities with disentangled deep metric learning [11]. While these works demonstrate the advantage of combining multiple representations of a same signal to extract meaningful structural patterns, our approach shows instead that these can be directly encoded into the representations using time proximity and the hierarchy of music structure.

2. HIERARCHICAL REPRESENTATIONS

The method introduced here constructs deep representations which allow for structural segmentations at various time-scales. To facilitate the decomposition of a song at different levels, these representations should provide strong discriminative capabilities for time frames belonging to different musical sections and separated by a large amount of time. Conversely, they should be more homogeneous for frames belonging to the same section and happening within a short time interval. As section lengths might vary from one annotator to another due to the ambiguity of the task [1], the aforementioned constraint is imposed at different temporal scales. Additionally, most datasets for music structure analysis come with only one level of annotations, which motivates us to learn such representations in an unsupervised fashion, taking advantage of large quantities of unlabelled data. A base convolutional neural network is used to output embeddings which are divided into multiple sub-regions. Each of them is optimized independently using specific triplets of frames efficiently sampled to encode the temporal structure of the song at different levels. We show that each level of the final representations can model frames proximity with its own amount of granularity.

2.1 Sampling

The objective of the sampling method introduced by McCallum [3] is to build triplets of frames where the anchor and the positive example belong to the same musical section, while the anchor and the negative example are labelled differently. The method proposed here can be viewed as its multi-level extension. A hierarchy is defined as a set of L levels of structural segmentations ordered from the coarsest to the most refined. For each level $\ell \in \{0; \dots; L-1\}$ in the hierarchy, triplets of beat indices are sampled using a specific set of parameters $\delta = \{\delta^{\ell}_{p,min}, \delta^{\ell}_{p,max}, \delta^{\ell}_{n,min}, \delta^{\ell}_{n,max}\}$. Intuitively, they rule how "close" or "far away" from the anchor the positive and negative examples will be sampled throughout the song. More specifically, for a given anchor beat index i_a, positive and negative examples respectively located at beat indices i_p and i_n are uniformly sampled from the interval I^{ℓ}_p defined by $\delta^{\ell}_{p,min}$ and $\delta^{\ell}_{p,max}$ and I^{ℓ}_n specified by $\delta^{\ell}_{n,min}$ and $\delta^{\ell}_{n,max}$. An example for an arbitrary level ℓ is shown in Figure 1. The δ parameters actually define a notion of temporal distance d_{ℓ} between frames at a given level of the hierarchy (analogous to a notion of dissimilarity). Therefore, the set of triplets T_{ℓ} at level ℓ can be expressed as:

$$T_{\ell} = \{(i_a, i_p, i_n; l) \mid d_{\ell}(x_a, x_p) < d_{\ell}(x_a, x_n)\} \qquad (1)$$

where x_k is the input feature patch observed at beat index i_k, $i^{\ell}_p \sim \mathcal{U}(I^{\ell}_p)$ and $i^{\ell}_n \sim \mathcal{U}(I^{\ell}_n)$.

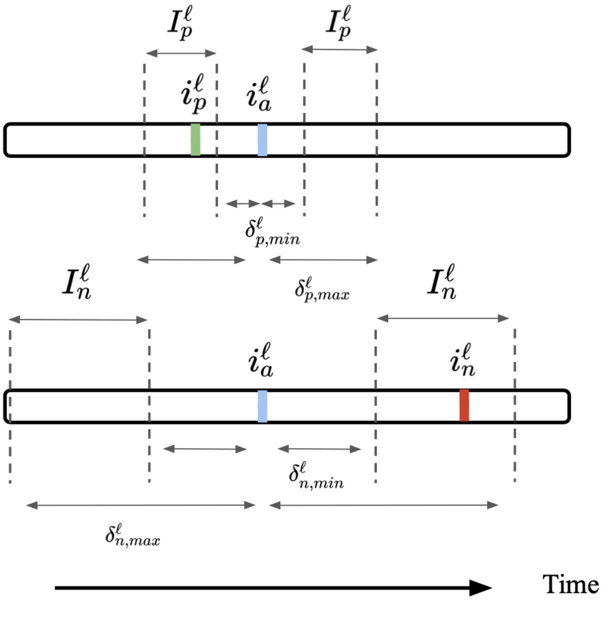

Figure 1. Initial triplet sampling method at level ℓ.

In order for the learned hierarchy levels to remain consistent with one another, monotonicity is encouraged by modifying the initial triplet mining technique. In addition to the time constraint imposed on triplets of the same level, their probability of being sampled is restricted from one level in the hierarchy to the next. For a randomly sampled anchor index at level $\ell = 0$, a complete triplet (i^a_0, i^p_0, i^n_0) is built only using time proximity (*i.e.* δ parameters). Then