Combining Collaborative and Content Filtering in a Recommendation System for a Web-based DAW

Jason Smith
Georgia Tech Center for Music Technology
840 McMillan Street
Atlanta, Georgia USA
jsmith775@gatech.edu

Mikhail Jacob
Georgia Tech Center for Interactive Computing
85 5th Street NW
Atlanta, Georgia USA
mikhail.jacob@gatech.edu

Jason Freeman
Georgia Tech Center for Music Technology
840 McMillan Street
Atlanta, Georgia USA
jason.freeman@gatech.edu

Brian Magerko
Georgia Tech School of Literature, Media, and Communication
686 Cherry Street
Atlanta, Georgia USA
magerko@gatech.edu

Tom Mcklin
The Findings Group
2646 Woodridge Drive
Decatur, Georgia USA
tom@thefindingsgroup.org

ABSTRACT

EarSketch is a web-based audio production and education platform that uses an online coding environment and the Web Audio API to teach introductory programming and music production to students. One of the main challenges of implementing an educational online music production platform is providing users with a variety of foundational audio loops to use in order to foster creative personal expression. EarSketch aims to achieve this through the inclusion of a sound browser for users to navigate and select sounds to use as part of their compositions.

This paper describes the implementation and evaluation of a hybrid recommendation engine, combining collaborative and content filtering, designed to guide users through the sound browser and promote diversity in student compositions. The paper also presents a preliminary analysis of the impact of different recommendation strategies on user sound selection, and how the application of recommendation strategies can inform the design of EarSketch and other web-based DAWs.

CCS Concepts

•**Applied computing** → **Sound and music computing;**

Keywords

music, recommendation systems

Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution**: owner/author(s).

Web Audio Conference WAC-2019, December 4–6, 2019, Trondheim, Norway.

© 2019 Copyright held by the owner/author(s).

1. INTRODUCTION

EarSketch [10] is an online platform that integrates Python and JavaScript coding environments with a Web Audio API-based digital audio workstation (DAW). It helps to teach students coding and music production through the manipulation of audio loops from a large sound library [11].

EarSketch engages diverse student populations in computing through a curriculum and learning environment that is authentic in both the computing domain (i.e. industry-standard programming languages) and the music domain (i.e. interface and API designs resembling digital audio workstations and the inclusion of audio loops created by professional artists) [3]. Previous research has found student perceptions of *authenticity* to be correlated with attitudes towards computing and intention to persist in the field, which EarSketch is designed to reinforce through its focus on pervasive music production paradigms, programming languages, and musical styles [13].

A critical component of this authentic learning environment is the ability of students to find personally relevant, expressive loops that fit musically with the compositions that they are creating through code. EarSketch includes a sound library of nearly 4000 audio loops across a variety of popular music genres, created for the platform by sound designer Richard Devine and hip-hop engineer and DJ Young Guru. These sounds form the building blocks of student compositions, and are designed to promote a wide range of creative expression in styles that are personally meaningful to students.

An analysis of 20,000 non-tutorial user scripts, both from experienced and novice users, collected in spring 2018 [16] reveals a significant under-utilization of a majority of the sound library. As seen in Figure 1, fewer than 200 sounds were used in over 1% of scripts. Under 20 sounds were used in over 10% of scripts. We hypothesize that this relative lack of usage of most sounds in the library stems from the difficulties users face in searching for and finding sounds in the browser interface, which has been limited to simple filters and text search. In fact, over half of the sounds most

often used in scripts closely align with those found in sample tutorial scripts.

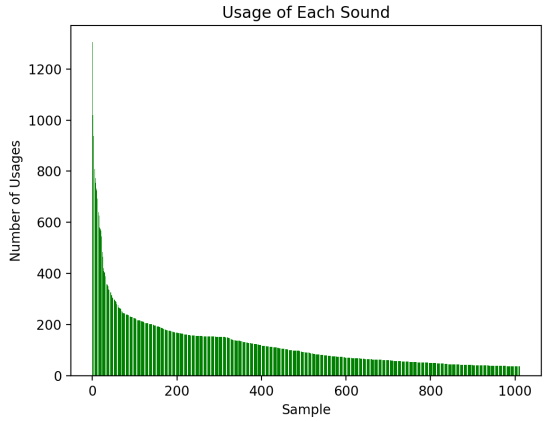

Figure 1: The 1,000 most commonly used EarSketch sounds in 20,000 user scripts [16].

We explored the implementation of a hybrid recommendation system [16] in order to a) address the cause of this usage problem, b) improve the diversity and coverage in the representation of the sound library by user scripts, and c) to better facilitate student creativity, potentially leading to higher perceptions of authenticity of EarSketch as a creative environment.

This project emphasizes the inclusion of an intelligent recommendation system in a web-based music production software. The guiding principle behind this that when a program gathers statistical data from its user base's creative decisions, it can generate its own decisions that are representative of its users. With a large sound library and user base, this allows for intelligent recommendations that approximate the collective design choices of users. Such recommendations may be of particular importance in web-based music production systems, which are typically targeted to novice users who may desire more guidance in locating and selecting sounds from a massive library of choices.

Common approaches to evaluating hybridized recommendation systems focus on the domain of listening preferences for songs. These examples use previous listening history and ratings [19] or social media [18] to collect user preferences. The domain of full songs differentiates these methods from our approach, which gathers preferences for short loops used in combination, and for composition instead of listening.

An example of ongoing research into feature analysis of shorter audio loop is Groove Explorer [2]. This drum loop visualization tool measures similarity between rhythmic vectors and evaluates performance using genre labels. Our recommendation system does not currently use these labels, but the inclusion of textual metadata may be used to improve future performance (see future work).

In this article, we present our research on a recommendation system for discovering new sounds for use in EarSketch, informed by a user-centered design study and proposed recommendation engine described in [16]. The main contributions discussed are:

- The implementation of a hybrid recommendation system using collaborative filtering of previous user scripts and analysis of audio similarity between sounds in the sound library.
- The redesign of the EarSketch sound browser to accommodate recommendations, and to improve user experience in navigating options.
- An evaluation methodology and preliminary results from two studies designed to assess the relative performance of four variations of the recommendation algorithm.

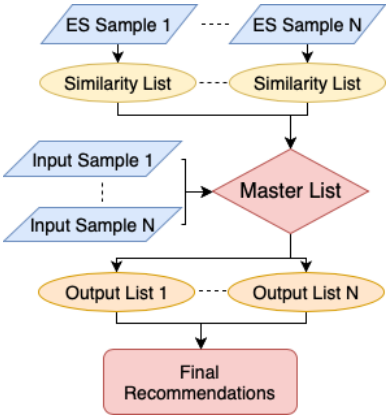

Figure 2: Program flow of recommendation generation

2. IMPLEMENTATION OF THE RECOMMENDATION SYSTEM

The base model of the hybrid recommendation system (Figure 2) is comprised of six steps.

1. Each sound in EarSketch is used as an input sound during a separate iteration of the recommendation process.
2. The system generates a *co-usage list* (representing the most commonly used sounds with the input) for each EarSketch sound, with usage statistics generated from an analysis of a collection of 20,000 user scripts.
3. The system generates a *similarity list* for each input sound, using content-based filtering of audio features as well as the co-usage data to compare it to every other sound in EarSketch and generate a recommendation score.
4. The system then combines each *similarity list* into a *master list* and uploads it to EarSketch.
5. When a user script is active in EarSketch, all of its component sounds are indexed in the *master list*.
6. The system combines the *similarity lists* for each input sound in order to generate weighted recommendation scores, and displays the sounds with the highest scores to the user.

This model is used, as explained in [16], to provide recommendations based on acoustic similarity as well as the habits of users without collecting personally identifiable data in conformance with EarSketch's privacy policy.

2.1 Collaborative Filtering

The *co-usage list* is generated from a collection of previous user scripts [16]. Every sound in each script containing the input has its score increased, so the sounds that most commonly appear with the input have the highest score.

This system is intended to regularly update with new user scripts, as improved sound representation in user scripts due to the addition of the recommendation system will increase the diversity of coverage scores. The system will increase in computational time as more scripts are added, but this will not be an issue due to it being an offline process. Scripts will be removed from the system over time, with a maximum set of scripts to be defined by future comparative studies.

2.2 Content-based Filtering

The system employs precomputed feature distances to quickly calculate recommendations for each sound, using each sound as input to create the *co-usage list* and *similarity list*, which are used to form the *master list*. The distance between every combination of EarSketch sounds is recorded for two common audio features, STFT and MFCC [7]. The features are calculated as *fingerprints* representing short sections of the samples, generated from an extended version of Kyle McDonald's AudioNotebooks code [12].

2.2.1 STFT

Comparison of Short-time Fourier Transform vectors, representing spectral features of each sound on a frame-by-frame basis, assesses temporally-based similarity. [8]. This time-series data is used in dimensionality reduction and clustering techniques to group and visualize sounds [5].

In Figure 3, STFT fingerprints are represented in a two-dimensional space using t-SNE visualization techniques [9] as inspired by the Infinite Drum Machine [17], with colors representing genre labels.

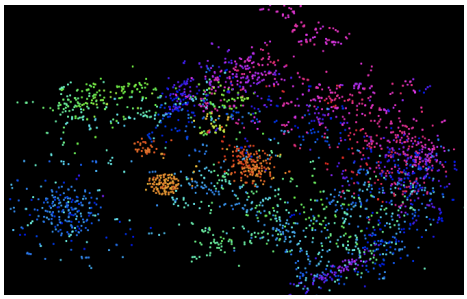

Figure 3: Short-time Fourier Transform used to cluster and represent EarSketch sounds in 2D space.

2.2.2 MFCC

Mel-frequecncy Cepstrum Coefficients represent sounds in the domain of frequency distribution. They are created by decorrelating the spectral information between frames and can be used in temporally-independent timbral speech analysis and genre recognition [8].

2.3 Limiting Factor

The recommendation system includes a limiting factor to save computational resources, both when generating the *master list* and when generating recommendations in real time. The *similarity list* for each sound stores only the 50 highest recommendation scores, and uses only the highest 10 values of the *co-usage list* to generate the scores.

2.4 Combined Recommendation Scores

The recommendation scores for each (*input*) and *output* sound pair, split into component scores labeled R_A R_B and R_C which are added to form the final score R, are generated as follows:

$$R_A = D_{co-usage}(hc, input) + D_{STFT}(output, hc) + D_{MFCC}(output, hc) \quad (1)$$

The base model calculates the feature distances between samples with the highest co-usage scores to generate R_A, which can be interpreted as similarity to a sound *hc* that is highly co-used with the input.

$$R_B = D_{co-usage}(output, input) \quad (2)$$

R_B is co-usage between the recommended sound and the input found in the collection of stored user scripts.

$$R_C = D_{STFT}(output, input) + D_{MFCC}(sound, input) \quad (3)$$

R_C is the feature similarity between the recommended sound and the input, calculated using precomputed MFCC and STFT distances.

$$[R_A, R_B, R_C] = [R_{A1} + ... + R_{AN}, R_{B1} + ... + R_{BN}, R_{C1} + ... + R_{CN}]/\sqrt{N} \quad (4)$$

With multiple sounds in a user script, the system accommodates for multiple inputs. It adds the component scores for each input and divides by the square root of the number of inputs (N) when generating recommendations. This is in order to balance - the scores recommended highly by multiple input sounds receive a score increase which cannot be achieved by using a mean of multiple scores, but without the hard increase of pure summation that would negate single recommendations.

$$R = R_A + C_U * R_B + S * R_C \quad (5)$$

The model adds the three component scores together to generate the final recommendation score R. Tuneable paramaters C_U and S can be set to -1 or 1 to maximize or minimize co-usage and similarity respectively. Minimization is chosen over neglect to allow for more parameter combinations that make use of the negative, such as intentionally choosing the lowest co-usage and the highest similarity or vice versa. These combinations generate the following labeled recommendation categories: *highest co-usage, highest similarity:* "Others Like You Use These Sounds", *lowest co-usage, highest similarity:* "Sounds That Fit Your Script", *highest co-usage, lowest similarity:* "Discover Different Kinds of Sounds", and *lowest co-usage, lowest similarity:* "Are You Feeling Lucky?".

2.5 Sound Browser Redesign

The EarSketch sound browser (Figure 4) has been redesigned to display recommendations as they are generated.

The addition of openable and closeable sound folders allows users to see a larger amount of sound types in immediate succession and to explore the options available to them through navigating the list. Recommendation folders appear at the beginning of the list when they are generated, with the type of recommendation indicated by a highlighted label.

The sound browser updates with new recommendations when a change is detected in the user script, such as when a user types in a new sound name, pastes from the sound browser, saves a script, or switches tabs. The sound names found in the active script are entered as the input, and if none are detected, a user's previous scripts are used to avoid the 'cold-start' problem [18] faced by usage-based systems.

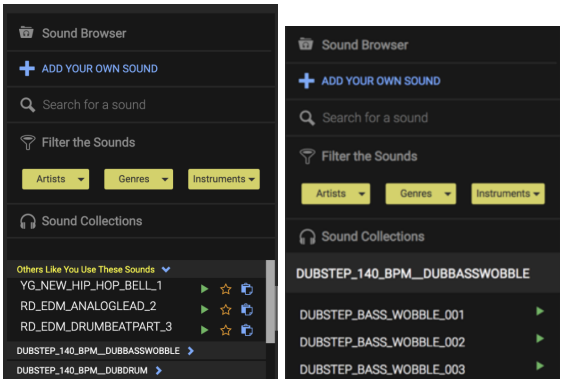

Figure 4: Sound Browser redesign (left) featuring open and closed folders and recommendation labels, with original Sound Browser design (right).

This design arose from a user study [16] that discovered that users wanted to find sounds based on categorical recommendations, highlighting the similarity or difference to their current work. Users also expected a degree of novelty or serendipity [1], which are emphasized by the recommendation labels.

3. EVALUATION

The recommendation engine and updated user interface were implemented and deployed to EarSketch users in May 2019. Our preliminary evaluation of the system seeks to understand the utility of the sound recommendations across the four categories ('Others Like You Use These Sounds', 'Sounds That Fit Your Script', 'Discover Different Kinds of Sounds', and 'Are You Feeling Lucky?') that differentially weight co-usage and similarity. The following research questions motivated our evaluation design: 1) Which recommendation category do a general pool of adult subjects prefer for completing a given musical fragment? 2) Which recommendation category do EarSketch users prefer while composing music in a web-based DAW?

This section presents two initial studies addressing these questions, with the intent that the findings will guide iterative design of the EarSketch recommendation system and potentially inform the design of sound recommendation systems in other web-based DAWs as well.

3.1 General Subject Pool Study: Ranking Recommendation Categories

An initial study was conducted based on the research question: which recommendation category do a general pool of adult subjects prefer for completing a given musical fragment? The study was conducted outside the context of a web-based DAW in order to understand relative user preferences between the four recommendation categories among a large number of subjects from a general population. Rather than actual EarSketch users, study participants consisted of adults from a general subject pool with a range of experiences in various musical activities (see Table 1). Subject experience levels with musical activities were divided into categories of never, rare (once in their lifetime or a few times a year), and regular (a few times a month, week, or day). Participants (N = 919 subjects in total) were recruited using the Amazon Mechanical Turk crowd-worker platform [1] and were offered $0.25 to participate in the study.

Table 1: General Subject Musical Experience

Activity	Never	Rare	Regular
Listening To Music	2	8	908
Playing Music	436	266	215
Writing Music	634	170	113
Amateur/Professional DJing	768	94	57
Read Musical Notation	593	182	141
Music Production Software	650	175	91

The study was designed with a within subjects repeated measures configuration. Each participant was asked to first listen to an audio track of a partial composition representing an in-progress musical composition on the EarSketch DAW. They were then asked to listen to four options for recommended audio loops, each obtained from a different type of recommendation. Recommendations were generated for each example before the test was published, and each subject was given the same four recommendations. They were asked to rank the options from one (best fit) to four (worst fit) in terms of how well they thought the suggested audio loop fit with the partial composition. Ties in ranks and missing ranks were grounds for rejecting participant responses as they were explicitly not permitted in the task.

The rankings obtained from subjects was considered to be ordinal data with one independent variable and four levels (1 to 4). The hypothesis (H_1) for the study stated that significant differences were present in the rankings between the four recommendation types. The null hypothesis (H_0) stated that there were no significant differences between the rankings for each recommendation option. The options were labeled as A - "Sounds That Fit Your Script", B - "Others Like You Use These Sounds", C - "Discover Different Kinds of Sounds", and D - "Are You Feeling Lucky?" for convenience of analysis.

The individual rank distributions for each option were first run through a Shapiro-Wilk test for normality [15] to see if they could be analyzed using parametric tests. All four distributions of responses were classified as non-normal. Therefore, the Friedman's rank sum (omnibus) test [4] was used to test whether there were significant differences between

[1] https://www.mturk.com

the ordinal ratings data for individual types of recommendations. The results of the Friedman's test allowed us to reject H_0 at a significance level $\alpha < 0.05$, showing significant differences between the distributions with $p = 2.7 \times 10^{-20}$.

Table 2: Pairwise Nemenyi Post-hoc Tests

	A	B	C
B	$\mathbf{9.70 \times 10^{-12}}$	-	-
C	$\mathbf{2.61 \times 10^{-14}}$	2.00×10^{-1}	-
D	$\mathbf{1.06 \times 10^{-3}}$	$\mathbf{4.91 \times 10^{-3}}$	$\mathbf{7.16 \times 10^{-7}}$

Pairwise Nemenyi post-hoc tests [14] were conducted on the data after finding significance with the Friedman's omnibus test. The results are shown in table 2 with significant differences highlighted in bold, showing that all pairs except for types B vs. C are statistically different. Median rankings, mean rankings, and standard deviations for all four distributions can be seen in table 3 with lower values being better and the best rankings highlighted in bold. The relative rankings and significant difference results placed the recommendation types ordered as $B, C > D > A$. However, a computation of effect size using the Kendall's W test of concordance [6] resulted in values of $W = 0.03$. This showed that our study was overpowered and the statistical significance observed was likely due to the large sample size.

Table 3: Recommendation Category Rankings

Recommendation Category	Median	Mean	Standard Deviation
A	3	2.80	1.22
B	**2**	**2.37**	**1.11**
C	**2**	**2.25**	**1.04**
D	3	2.57	1.02

Our results from studying a general subject pool of adult users with varying experience in musical activities did not reveal a clear preference between categories of recommendations. However, the study was conducted on a population unlike the target audience for EarSketch. Additionally, the context for ranking the different recommendation categories was far removed from the context of music composition in a web-based DAW. Therefore, a second study was conducted using aggregated data from actual EarSketch users composing music within EarSketch, in order to understand if significant differences would be observed in the relative usage of the different recommendation categories.

3.2 EarSketch Users Study: Comparing Relative Usage of Recommendation Categories

Over a period of 14 days in June 2019, we collected aggregate data from 21,368 EarSketch users. Users were randomly assigned to a recommendation type such that they only saw one category of recommendations in the sound browser (not all four). Users received single random category assignments, so they would see the same category on all sessions during the trial period. For each of the four recommendation categories, a web-based analytics engine collected aggregate data on a) the total number of recommendations displayed to users, b) the number of recommendations previewed by users (suggesting that they saw the recommendation and were intrigued by the sound's name), and c) the number of recommendations pasted by users into their code (suggesting that they found the recommendation to be a good fit for their project and proceeded to experiment with it, potentially using it in their final code).

Calculation of the number of recommended sounds pasted into user code relative to the number of recommendations previewed functioned as a simple comparison of the four recommendation types. A greater percentage of previewed sounds that are pasted into code may suggest that users find the recommendations better suited for their projects.

Due to the aggregate nature of this data collection, this system is limited in its inability to recognize sounds that have been pasted without being previewed. The general assumption made is that a user is first previewing a sound before pasting it into their script, but this assumption cannot be validated without collecting personal account data.

Table 4: EarSketch User Study Results

Recommendation Category	Pastes	Previews	Percentage
1:"Others Like You Use These Sounds"	173	1043	16.587%
2:"Sounds That Fit Your Script"	225	1448	15.539%
3:"Discover Different Kinds of Sounds"	214	1033	20.716%
4:"Are You Feeling Lucky?"	96	767	12.516%

Category 3 is found to be significantly higher than 2 and 4, using a general proportion test with a pairwise post-hoc test [14]. However, 3 is not significantly higher than 1, and 1 and 2 are not significantly higher than 4. The two significantly less used categories were the two with minimized co-usage scores, suggesting that showing the user intentionally rare combinations of sounds does not lead to higher usage of these sounds. EarSketch users showed a significant increase in usage for sounds in the 'Discover Different Kinds of Sounds' category above all other categories, unlike the general subject pool.

The difference in population of EarSketch's student body and adult subjects, as well as the difference in context between general EarSketch usage and selecting sounds for the completion of an example script, can explain the statistical variation between tests. The general subjects' task may have led them towards sounds that were more acoustically similar to the ones in the example, while EarSketch users free to create were more inclined to try sounds presented as intentionally different to what they were using.

The labeling and presentation of the categories is a possible cause of this difference, in that the 'Discover Different Kinds of Sounds' category appeals most to the desire for novel recommendations from users. Users in past interviews suggested that recommendations of sounds similar to previously used sounds was of lesser importance to them [16]. The statistical variation between subject groups reflects the importance of evaluating recommendations in context. The different tasks created different reasons for users to consult sound recommendations, and other web-based systems should tailor their evaluations to user behavior.

4. FUTURE WORK

As the studies in this section compared usage of the four categories in completing scripts without user input on recommendation quality, future explicit comparison between styles is necessary. A possible format for this evaluation is having participants rank sample recommendations for active scripts in a blind procedure in terms of relevance, novelty, diversity, and serendipity, in order to determine the strengths and weaknesses of each category beyond usage statistics.

Once enough time has passed for data collection, coverage analysis using a new set of post-recommendation user scripts will be conducted. The changes in the usage of each sound compared to Figure 1 will be used to evaluate the long-term effectiveness of the recommendation engine. Additionally, paste-to-preview percentages (as found in Table 4) for non-recommended sounds can provide a baseline for comparison.

Other changes to the recommendation system can be made to better adhere to user interests and evaluate performance. Metadata can be used to improve recommendation relevance or provide more contextual recommendations. Finally, the recommendation engine will be re-seeded with a larger collection of user scripts. The effects that re-seeding using scripts built with recommendations has on usage statistics will be evaluated as the system improves.

5. CONCLUSIONS

As our sound recommendation system has been integrated with EarSketch, a web-based music production system, we can evaluate performance in terms of user trends and studies designed to compare categorical recommendations. Performance goals defined by user study and performance analysis have concrete influences on design choices, such as the integration of recommendation systems into EarSketch and other web-based DAWs.

6. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation Award No. 1814083. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Additional thanks to Nashawn Cherry for assistance in creating precomputed feature distance vectors for the EarSketch sound library. EarSketch is available online at *https://earsketch.gatech.edu*.

7. REFERENCES

[1] C. C. Aggarwal et al. *Recommender systems*. Springer International Publishing, 2016.

[2] F. Bruford, M. Barthet, S. McDonald, and M. Sandler. Groove explorer: An intelligent visual interface for drum loop library navigation. In *Joint Proceedings of the ACM IUI 2019 Workshops*, 2019.

[3] J. Freeman, B. Magerko, T. McKlin, M. Reilly, J. Permar, C. Summers, and E. Fruchter. Engaging underrepresented groups in high school introductory computing through computational remixing with earsketch. In *Proceedings of the 45th ACM technical symposium on Computer science education*, pages 85–90. ACM, 2014.

[4] M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.

[5] S. Haskey, B. Blackwell, and D. Pretty. Clustering of periodic multichannel timeseries data with application to plasma fluctuations. *Computer Physics Communications*, 185(6):1669–1680, 2014.

[6] M. G. Kendall and B. B. Smith. The problem of m rankings. *Annals of mathematical statistics*, 1939.

[7] A. Lerch. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. Wiley-IEEE Press, 1st edition, 2012.

[8] T. Li, M. Ogihara, and Q. Li. A comparative study on content-based music genre classification. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 282–289. ACM, 2003.

[9] L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[10] B. Magerko, J. Freeman, T. Mcklin, M. Reilly, E. Livingston, S. Mccoid, and A. Crews-Brown. Earsketch: A steam-based approach for underrepresented populations in high school computer science education. *ACM Transactions on Computing Education (TOCE)*, 16(4):14, 2016.

[11] A. Mahadevan, J. Freeman, B. Magerko, and J. C. Martinez. Earsketch: Teaching computational music remixing in an online web audio based learning environment. In *Web Audio Conference*, 2015.

[12] K. McDonald. Audio notebooks. https://github.com/kylemcdonald/AudioNotebooks, 2016.

[13] T. McKlin, B. Magerko, T. Lee, D. Wanzer, D. Edwards, and J. Freeman. Authenticity and personal creativity: How earsketch affects student persistence. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, pages 987–992. ACM, 2018.

[14] P. Nemenyi. Distribution-free multiple comparisons (doctoral dissertation, princeton university, 1963). *Dissertation Abstracts International*, 25(2):1233, 1963.

[15] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[16] J. Smith, M. J. Dillon Weeks, J. Freeman, and B. Magerko. Towards a hybrid recommendation system for a sound library. In *Joint Proceedings of the ACM IUI 2019 Workshops*, 2019.

[17] M. Tan and K. McDonald. Infinite drum machine. https://experiments.withgoogle.com/ai/drum-machine/, 2017.

[18] A. Vall and G. Widmer. Machine learning approaches to hybrid music recommender systems. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 639–642. Springer, 2018.

[19] D. Wu. Music personalized recommendation system based on hybrid filtration. In *2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, pages 430–433. IEEE, 2019.