CROWD'S PERFORMANCE ON TEMPORAL ACTIVITY DETECTION OF MUSICAL INSTRUMENTS IN POLYPHONIC MUSIC

Ioannis Petros Samiotis
Delft University of Technology
i.p.samiotis@tudelft.nl

Christoph Lofi
Delft University of Technology
c.lofi@tudelft.nl

Alessandro Bozzon
Delft University of Technology
a.bozzon@tudelft.nl

ABSTRACT

Musical instrument recognition enables applications such as instrument-based music search and audio manipulation, which are highly sought-after processes in everyday music consumption and production. Despite continuous progresses, advances in automatic musical instrument recognition is hindered by the lack of large, diverse and publicly available annotated datasets. As studies have shown, there is potential to scale up music data annotation processes through crowdsourcing. However, it is still unclear the extent to which untrained crowdworkers can effectively detect when a musical instrument is active in an audio excerpt. In this study, we explore the performance of non-experts on online crowdsourcing platforms, to detect temporal activity of instruments on audio extracts of selected genres. We study the factors that can affect their performance, while we also analyse user characteristics that could predict their performance. Our results bring further insights into the general crowd's capabilities to detect instruments.

1. INTRODUCTION

Studies of the last decade have shown the success of data-driven algorithms to tackle complex classification tasks. Such algorithms require large annotated datasets to train and capture the nuances of multi-faceted problems, with crowdsourcing being successfully utilized to scale annotation processes to meet the ever higher demands [1–3]. While works such as [4] and [5] show that crowdsourcing can be a viable and powerful tool to distinguish and annotate music audio, it still remains underutilised as a tool in the domain, primarily due to the complexity of the annotation tasks [6] which are believed to demand extensive domain knowledge and training – arguably, musical elements such as tempo, chords and timbre can be demanding for an untrained human annotator to detect.

With this study, we aim at providing more evidence that complex music audio annotation tasks can be performed on crowdsourcing platforms. We focus on the task of musical instrument activity detection, and investigate non-experts' capability to recognise their activity and annotate the times in which they perform. Our study builds upon the findings of [4] where users were able to detect if an instrument was present in an audio excerpt or not. We extend this detection task to also cover the exact time-frames of instrument activity. This is a type of task where experts are commonly employed [7] to annotate data, due to several challenges such as multiple instruments playing simultaneously [8,9], or instruments of the same family exhibiting similar timbre [10, 11].

More specifically, we explore and analyse the capabilities of crowd workers to effectively detect temporal aspects of musical instrument activity in polyphonic audio (with focus on trio ensembles). We seek answer to the following questions:

- `RQ1`: To what extent non-experts can detect the onset and offset of a musical instrument's activity on polyphonic audio?
- `RQ2`: How their self-assessed perceptual abilities and musical knowledge relate to their performance?

Our study takes place on Prolific [1] . The audio excerpts were chosen from three different genres (namely *classical*, *jazz* and *rock*) to understand if different instruments and rhythms can affect the performance of crowd workers. We also utilize a set of pre-established and evaluated questionnaires to retrieve user attributes, that can potentially relate to their performance. We employ the "Musical Training" and "Perceptual Abilities" categories from Goldsmith's Music Sophistication Index (GMSI) [12], a questionnaire specifically designed to capture an individual's ability to engage with music. These specific categories were found previously to most significantly predict the workers' musical perceptual abilities [13].

Our results show that non-experts can demonstrate good perception of musical instruments' temporal activity for the chosen audio excerpts. Their self-assessed perceptual abilities reflect reasonably well their actual perception skill. These results open possibilities of further future studies on instrument activity annotation, and provide a positive outlook for systems relying on such annotations.

[1] `https://www.prolific.co`

2. RELATED WORK

The work in OpenMic 2018 [4] is one of the first attempts to annotate instrument presence for instrument recognition at scale, employing 2,500 unique annotators from CrowdFlower [2], using excerpts from Free Music Archive [3] and the AudioSet [14]. The researchers followed specific task design approaches to assist the crowd workers in their task, which they adapted after an initial study. The annotation process was limited to binary annotations, indicating the presence or absence of a musical instrument in an audio excerpt. Showcasing that crowd workers are able to provide strongly-labeled data, e.g. with temporal annotation, as in our study, can enable new opportunities for instrument activity detection and source separation.

Even though the study in [15] is not based on music audio, it demonstrates the crowd's ability to annotate temporal aspects of audio events. Our interface design is inspired by this study, as the crowd workers had to draw bounding boxes on spectrogram visualisations of audio excerpts. The sounds were synthesized using Scaper [16], for a greater control over *max-polyphony* and *gini-polyphony* (amount of sound overlap).

Our study is also motivated by recent findings regarding crowd workers music perception abilities [13]. Users of crowdsourcing platforms were shown to possess considerable skills to detect music aspects such as tempo and melody.

To the best of our knowledge, the current literature lacks works that study the performance of crowd workers on temporal activity detection of musical instruments in relationship with worker demographic or musical properties, which is the goal of this work.

3. EXPERIMENTAL DESIGN

We designed our experiment to study and understand if users on crowdsourcing platforms can perceive the temporal activity of a musical instrument in audio excerpts. We aim to focus on realistic use cases, thus testing the workers' capacity to perceive instruments in audio excerpts that are performed, recorded, mixed and mastered professionally. Therefore, we used existing recordings instead of synthesized audio which would have been less representative of real-life scenarios, but could have given us higher control on the musical aspects of the audio and instrumentation. To that end, we carefully selected the audio excerpts to control, as much as possible, musical aspects such as timbre and performance.

We employed previously established and evaluated questionnaires, to learn about workers' (a) "Perceptual Abilities" and "Music Training" through Goldsmith's Musical Sophistication Questionnaire (GMSI); (b) cognitive load through NASA's Task Load IndeX (NASA-TLX) survey [4]; (c) equipment quality [17] and (d) outside noise [18].

[2] https://visit.figure-eight.com/People-Powered-Data-Enrichment_T

[3] https://freemusicarchive.org

[4] https://humansystems.arc.nasa.gov/groups/tlx/

The task workflow started with simple demographic questions, followed by the GMSI questionnaire. The user was then introduced with the main task to annotate audio excerpts. The study concluded with a post-task survey regarding their cognitive load, equipment and a general feedback entry.

3.1 Selected Audio Excerpts

For the main annotation task, we made use of audio excerpts from trio ensembles of three major genres, *classical*, *jazz* and *rock*. We used audio excerpts of these particular three genres due to their wide discrepancy in instrumentation and rhythm. Even though in some occasions the instruments used in each genre can showcase timbre similarities (like double bass and bass guitar), in other cases the timbre can differ wildly (electric guitar compared to cello). To the best of our knowledge, there is no previous baseline of the crowd workers' perception of polyphonic music, so we decided to control for the maximum number of instruments that would play simultaneously in an excerpt, by selecting recordings of trio ensembles for each genre. Each audio excerpt had a length of 10 seconds, as used also in similar studies [4, 15]. The authors annotated the instrument activity per audio excerpt, which was later used to evaluate the crowd's annotations.

For the classical music excerpts, we made use of a specific type of a trio ensemble, namely *piano*, *clarinet* and *cello*. On the selected music clip, we selected an excerpt where both *clarinet* and *cello* have prominent parts, while *piano* is mostly following in the background. For our jazz excerpt, we used of the more standardized trio ensemble of *piano*, *double bass* and *drums*, where *double bass* and *drums* keep the rhythm and *piano* is performed in small melodic bursts. Lastly, for the category of rock, we made use of a music excerpt from "power trio" bands, which most frequently consist of *electric guitar*, *bass guitar* and *drums*. It follows the same performance pattern with the jazz excerpt on the *bass guitar* and *drums*, while the *electric guitar* enters near the middle of the excerpt with a sustained, distorted power chord.

We hypothesise that bass instruments will be more difficult to annotate in these genres, as bass-related sounds are more often "pushed back" during the mixing stage for such types of music. The different genres were selected to lessen the impact of possible enculturation bias. We believe that if only one genre was selected, participants who would be more familiar with it, would find it easier to spot the activity of instruments prominent in the genre. With the selected genres, we cover a variety of rhythms, instrumentations and performative aspects, which could impose a challenge to non-experts.

3.2 Task and Interface Design

To assess the music expertise of the crowd we employed parts of GMSI, namely: "Music Training" and "Perceptual Abilities". The choice of the categories was based on a study on music perception skills of crowd workers [13],

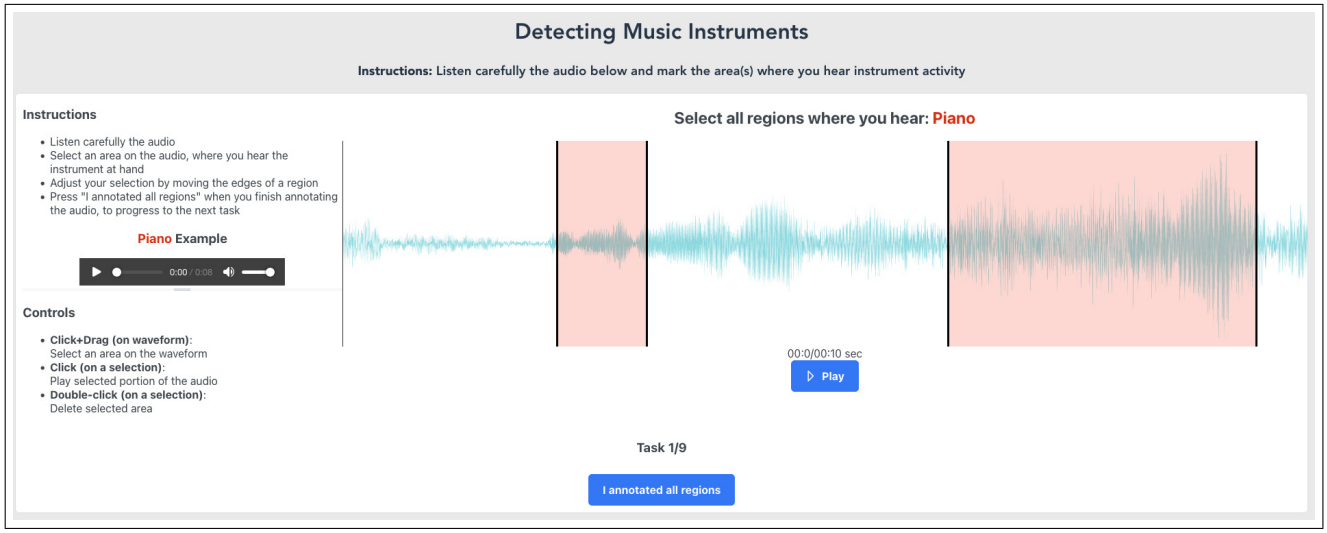

Figure 1: Main audio annotation task

where results in these two categories were found to most significantly predict their auditory capabilities.

The questions of both GMSI categories were aggregated to one questionnaire, with one attention question placed in-between the questionnaire's items. The users also had the ability to use a "Back" button to return to a previous question and alter their answer. We used the complete set of questions on both "Music Training" and "Perceptual Abilities", after consulting the online GSMI "configurator" [5] .

The users were greeted with an "Instructions" message before the main annotation task, which described the steps to complete each microtask and a warning regarding the volume (as seen in Figure 2). The main audio annotation task (see Figure 1) consisted of four main parts: (a) audio waveform and controls (center-right), (b) instructions and instrument example (upper left), (c) description of controls and (d) submission button with a simple progress indication. The instrument to be identified, was indicated on both (a) and (b) in red, to draw the attention of the users.

Based on the findings during the OpenMic 2018 work [4], the crowd workers were found to struggle to detect multiple instruments at once. To that end, we followed their task design of annotating one instrument at a time; we presented the participants with the audio excerpt and requested to annotate the regions where a chosen single instrument, was active during the recording.

The worker would be presented with an audio excerpt and was instructed to detect the activity of one of the instruments present in the excerpt. The same procedure would follow for each of the instruments per audio excerpt, presented in a random order across genres (e.g. piano from classical music excerpt, followed by the electric guitar from rock music excerpt).

In the audio annotation interface, the users could play and pause the audio excerpt while also draw bounding boxes on the audio waveform. The regions drawn on the waveform were adjustable on both ends and the user could

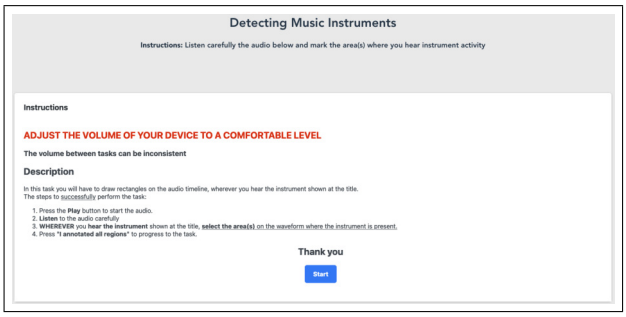

Figure 2: Task instructions and warning

easily dismiss them with a double-click. A single-click on a region would play only the selected part of the audio excerpt. A crowd worker could only progress to the next excerpt if they had drew at least one bounding box on the waveform.

For the design of the interface, we utilized `wavesurfer.js` [6] to draw the waveform and used the `regions` package to enable the bounding boxes interaction. Our choice of these tools was based on previous studies on audio annotation that utilized them successfully [15, 19].

Finally, as mentioned in [4], crowd workers could experience high cognitive load during instrument detection tasks, ultimately affecting their psyche. It was important for us to capture such a phenomenon, so we included the NASA-TLX questionnaire and a free text input to accommodate their feedback towards the study.

3.3 Evaluation methods

Our task design is based around one audio excerpt per genre (10 seconds), where maximum three instruments can play simultaneously. As described before, per task, a worker had to draw the regions where they detect the activity of the selected musical instrument.

[5] `https://shiny.gold-msi.org/gmsiconfigurator/`

[6] `https://wavesurfer-js.org`

To evaluate their performance, we followed the same methods established in [15, 20] and in the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [21]. We segmented each excerpt into 100ms-long frames which had binary values, depending on the presence or absence of the selected instrument. A frame is considered active when there is an overlap between the annotation region and any portion of the time interval of the frame. We believe that the frame's resolution of 100ms can help us to adequately assess the extent of crowd workers' precision when annotating the temporal activity of an instrument. Based on the ground truth values, we later calculated *Accuracy*, *Precision* and *Recall* of the workers' annotations. To evaluate their performance, we followed the same methods established in [15, 20] and in the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [21]. We segmented each excerpt ($N = 3$) into 100ms-long frames which had binary values, depending on the presence or absence of the selected instrument. A frame is considered active when there is an overlap between the annotation region and any portion of the time interval of the frame. We believe that the frame's resolution of 100ms can help us to adequately assess the extent of crowd workers' capabilities to detect the temporal activity of an instrument. Based on the ground truth values, we later calculated *Accuracy*, *Precision* and *Recall* of the workers' annotations.

4. RESULTS

The study took place on Prolific, employing 28 crowd workers. We used the built-in prescreening filters of Prolific, setting criteria for fluency in English – for instructions' comprehension and higher chance of affinity to western music – and minimum task approval rate to 90% – to maximise the chances for good-quality work. The reward was set to 4.5 GBP (5.62 USD) which was classified as "*Good*" by the platform. We preserved the results of the 14 workers (see their demographics on Table 1) that successfully passed the attention question. Filtering the results based on the attention questions.

	Variables	**Statistics**
Gender, n	Female	10
	Male	17
	Prefer not to say	1
Age (years)	Range	18-55
Occupation	Full-time	12
	Part-time	5
	Unemployed	11
Education	Associate degree	2
	Bachelor's degree	12
	High school/HED	4
	Master's degree	4
	Some college, no diploma	3
	Technical/trade/vocational training	3

Table 1: Participant demographics

4.1 Demographics and Equipment

The workers used mostly earphones, headphones and laptop speakers, while three reported using dedicated speakers. Most workers (15) reported the quality of their equipment as "Excellent", with the majority (22) reporting "Imperceptible" impairment. Finally, the majority (15) reported that conducted the study in near silence conditions, while one reported performing the tasks in an environment with high noise levels.

4.2 Detecting Musical Instruments

The crowd workers showed high performance detecting most instrument activities on all three audio clips (`RQ1`). Studying the results per genre, we see on Table 2 that "Clarinet" was the most easily identifiable instrument. In the given audio excerpt, "Clarinet" had a prominent and distinct timbre, compared to the rest of the instruments. This might have helped annotators to detect its activity correctly. "Piano" on the other hand was more difficult to detect its temporal activity, as it accompanied the rest of the instruments with a softer tone.

	Accuracy	Precision	Recall
Piano	70.6%	91.5%	66.5%
Clarinet	**84.5%**	**95.8%**	**82.9%**
Cello	62.6%	95.5%	59.6%

Table 2: Accuracy, Precision, Recall and F-score on Classical audio excerpt (the highest scores per metric are in bold)

"Cello" though appears to be the hardest instrument to detect in the audio excerpt, as both accuracy and recall are near 60%. The high precision combined with low accuracy, could indicate that most workers mistook the activity of another instrument, with that of a cello. The results are surprising, as "Cello" was equally prominent as the "Clarinet", playing at a lower register than the rest of the instruments.

In the case of "Jazz" we find the "Drums" to be the most recognizable instrument, while "Double Bass" yielded better results than "Cello" in the "Classical" excerpt (see Table 3. Recordings of "Double Bass" in jazz can vary from barely noticeable to accentuated, depending on the recording setting or the part of the song (being more prominent during solo performance). Despite being the prominent instrument alongside "Drums" for a large portion of the excerpt, the workers still had trouble identifying the regions where it was active.

	Accuracy	Precision	Recall
Piano	81.8%	70.9%	**87.7%**
Double Bass	64%	**100%**	64%
Drums	**84.4%**	**100%**	84.4%

Table 3: Accuracy, Precision, Recall and F-score on Jazz audio excerpt (the highest scores per metric are in bold)

It is very interesting to highlight how the performance on "Piano" which is present in both "Classical" and "Jazz" music clips, changes greatly between the two samples. A possible explanation could be on the rather more prominent role it plays in piano jazz trios, where in most cases carries the melodic part of a composition (which would explain also the high recall score). In this specific example, we see that on average the crowd workers accurately selected the small rhythmic bursts of piano play, although not as precisely. This shows that they could definitely detect its activity correctly, but could not indicate precisely its onset and offset regions.

	Accuracy	Precision	Recall
Electric Guitar	**91.7%**	96.5%	**91.6%**
Bass Guitar	82.4%	**100%**	82.4%
Drums	73%	**100%**	73%

Table 4: Accuracy, Precision, Recall and F-score on Rock audio excerpt (the highest scores per metric are in bold)

The participants performed better on average, in the "Rock" excerpt. We speculate that the sounds of "Electric Guitar" and "Bass Guitar" are more familiar to the demographics of the participating workers, who scored quite highly on accuracy and recall, on both instruments.

The sustained power chord of the "Electric Guitar" was easy to identify and correctly annotate its onset and offset. On the other hand, despite "Drums" and "Bass Guitar" being present during the entirety of the audio excerpt, crowd workers found "Drums" more difficult to recognize correctly, despite the results in the jazz excerpt. Difference in "Drums" between the two excerpts, show higher use of the snare drum in the jazz excerpt, while in the rock, the use of lower tone tom drums was more prominent.

4.3 Self-assessed Music Characteristics and Performance

On Table 5 we see the self-assessed "Perceptual Abilities" and self-reported "Musical Training" of the participants. The low "Musical Training" is consistent with the results of [13] but pretty low when compared to the participant pool of [12] (scoring near the bottom 30% of the population in the original study).

	Range	Median	Standard Deviation(1σ)
Perceptual Abilities	29-63	47.5	8.19
Musical Training	7-41	18.5	9.04

Table 5: Range, Median, Mean and Standard Deviation of Perceptual Abilities and Musical Training

The self-assessed "Perceptual Abilities" are also low compared to the sample of [12] but considerably higher than in [13]. The results in our study certainly showcase adequate perceptual skills, in regards with the task at hand.

We study the connection of their musical properties to their performance from a more qualitative perspective, due to the size of our participant pool. Their self-assessed "Perceptual Abilities" show that the users felt quite confident on the degree they can detect musical traits on sound, despite their lack of expertise as shown by their "Musical Training" average score (Table 5).

Comparing their assessment to their actual performance we further see that their "Musical Training" is not indicative of their capability to detect temporal activity of musical instruments. Their median score as shown on the table, is close to the low 25th percentile of the results in the original GMSI study [12], showing a general low formal musical training. While formal training could certainly be beneficial for such tasks, people are still exposed to different musical instruments through casually enjoying music, especially as it is widely and easily accessible through streaming services. We also believe that the task design with the inclusion of an audio example of a given instrument, assisted the workers in their task to identify instruments.

4.4 Cognitive Load and Feedback

The results on the NASA-TLX questionnaire, show that from the total of 14 crowd workers, 10 found the task's difficulty average, while 9 were very confident on their performance. All of the participants reported average to low mental and physical demand, with mental load being higher than the physical. 10 workers experienced very low temporal demand, with most finishing the study in near 10 minutes. The results though show that the workers' self-assessed performance varied greatly between individuals, with scores from "Very Low" to "Very High".

Finally, crowd workers expressed their opinions on the study through a free form text area. Through their feedback we found that they greatly enjoyed the study through comments such as: "*Study was very well thought out. Nothing else to add.*", "*It was fun, I would love to take part similar studies again*" and "*the study was interesting and I am finding the piano very interesting instrument after this study*". Some even gave their insights for future improvements in comments such as: "*Put more instruments in there*" and "*it was ok but i propose next time the sounds be played slowly for us to easily identify. thank you*".

5. DISCUSSION

Non-experts exhibited high precision with a rather high recall on most instruments, especially on the "Jazz" and "Rock" audio clips. Despite their low expertise as indicated through the "Musical Training" attribute, the results show that they were capable of perceiving the temporal activity of instruments. These abilities are in line with the findings from [13] but also people's innate understanding of music, as shown in studies [22–24].

The high precision scores combined with lower accuracy and recall scores though, could indicate that the participants underestimated the activity of the instruments in

the excerpts. This means that the users although detected correctly segments of an instrument's activity, they weren't able to identify the totality of temporal activity for the given instrument. By selecting more, smaller and precise regions, one would select only the most prominent "True Positive" frames in an excerpt, but fail to select all of them, as is apparent on the cases of "Cello" and "Double Bass". Additionally, in our evaluation, we used a quite short and strict frame resolution which could potentially affect their recall scores. However, further studies are needed with variable frame resolution to test its suitability for this type of annotation task.

While it is inevitable to experience issues of sampling bias when executing crowdsourcing studies (i.e. participants will always be a smaller set of the userbase, which by itself is highly specific and smaller than the general public), we justify the differences with [12] based on the form of incentive from the side of participants, to perform the study. In our case the incentive was strictly monetary, therefore we employed participants who could be less enthusiastic about music, compared to [12]. When comparing to [13] though, while the results are consistent regarding "Musical Training", the results on "Perceptual Abilities" were higher in our case, despite the use of the same crowdsourcing platform. Of course, the landscape of crowdsourcing platforms is constantly changing, but it could be a nice indication of adequately, musically perceptive crowd workers.

Finally, we believe that our interface design with the inclusion of short examples of the musical instruments on each task, must have assisted the crowd workers during annotation. We encourage further experimentation on interface design, to explore effective ways to assist workers during their audio annotation task.

Limitations. Being an exploratory study, we acknowledge that the number of participating crowd workers is lower than in traditional crowdsourcing studies. Nonetheless, we believe that the rigorous set up and the in-depth qualitative analysis of the obtained results allow us to provide valuable and robust insights, which could be used to design and deploy larger-scale studies in the future.

The music excerpts we used in our study focus on popular genres of music. As such, despite the diverse demographics of Prolific, the participants in our study were expected to be familiar with the instruments in our excerpts. We strongly encourage future studies to experiment with instruments of different traditions, as we believe that similar techniques could yield equally promising results for those instruments.

6. CONCLUSION

Our study focuses on exploring the ability of non-experts to identify the temporal activity of musical instruments in audio excerpts of western music. This is an important task during dataset production for instrument recognition, as it can provide strongly-labeled annotations which enable event detection classification tasks. Results show that untrained crowd workers can successfully detect the activity of instruments like *clarinet* and *electric guitar*, one at a time, given an example of the instrument. The overall cognitive load that workers experienced was average, while most of them expressed their enjoyment of the tasks through free-form feedback. The positive outcomes of this work encourage conducting further studies on the topic, with focus on a larger participant pool and a more extensive evaluation dataset that includes additional genres, instruments, and identification complexities.

7. REFERENCES

[1] S. Nowak and S. Rüger, "How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation," in *Proceedings of the international conference on Multimedia information retrieval*, 2010, pp. 557–566.

[2] T. Yan, V. Kumar, and D. Ganesan, "Crowdsearch: exploiting crowds for accurate real-time image search on mobile phones," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*, 2010, pp. 77–90.

[3] N. Sawant, J. Li, and J. Z. Wang, "Automatic image semantic interpretation using social action and tagging data," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 213–246, 2011.

[4] E. Humphrey, S. Durand, and B. McFee, "Openmic-2018: An open data-set for multiple instrument recognition." in *ISMIR*, 2018, pp. 438–444.

[5] H. Schreiber and M. Müller, "A crowdsourced experiment for tempo estimation of electronic dance music." in *ISMIR*, 2018, pp. 409–415.

[6] S. Grollmisch, E. Cano, C. Kehling, and M. Taenzer, "Analyzing the potential of pre-trained embeddings for audio classification tasks," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 790–794.

[7] J. J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera, "A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals." in *ISMIR*. Citeseer, 2012, pp. 559–564.

[8] Y. Han, J. Kim, and K. Lee, "Deep convolutional neural networks for predominant instrument recognition in polyphonic music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 208–221, 2016.

[9] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–15, 2006.

[10] D. L. Wessel, "Timbre space as a musical control structure," *Computer music journal*, pp. 45–52, 1979.

[11] J. Pons, O. Slizovskaia, R. Gong, E. Gómez, and X. Serra, "Timbre analysis of music audio signals with convolutional neural networks," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 2744–2748.

[12] D. Müllensiefen, B. Gingras, J. Musil, and L. Stewart, "The musicality of non-musicians: an index for assessing musical sophistication in the general population," *PloS one*, vol. 9, no. 2, p. e89642, 2014.

[13] I. P. Samiotis, S. Qiu, C. Lofi, J. Yang, U. Gadiraju, and A. Bozzon, "Exploring the music perception skills of crowd workers," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 9, 2021, pp. 108–119.

[14] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[15] M. Cartwright, A. Seals, J. Salamon, A. Williams, S. Mikloska, D. MacConnell, E. Law, J. P. Bello, and O. Nov, "Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, pp. 1–21, 2017.

[16] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 344–348.

[17] I. Recommendation, "General methods for the subjective assessment of sound quality," *ITU-R BS*, pp. 1284–1, 2003.

[18] E. F. Beach, W. Williams, and M. Gilliver, "The objective-subjective assessment of noise: Young adults can estimate loudness of events and lifestyle noise," *International journal of audiology*, vol. 51, no. 6, pp. 444–449, 2012.

[19] M. Marolt, C. Bohak, A. Kavčič, and M. Pesek, "Automatic segmentation of ethnomusicological field recordings," *Applied Sciences*, vol. 9, no. 3, p. 439, 2019.

[20] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[21] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[22] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, no. 1, pp. 1–36, 1985.

[23] S. Koelsch, K. Schulze, D. Sammler, T. Fritz, K. Müller, and O. Gruber, "Functional architecture of verbal and tonal working memory: an fmri study," *Human brain mapping*, vol. 30, no. 3, pp. 859–873, 2009.

[24] B. Gingras, H. Honing, I. Peretz, L. J. Trainor, and S. E. Fisher, "Defining the biological bases of individual differences in musicality," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, no. 1664, p. 20140092, 2015.