ELEVATOR ANNOTATOR

Local Crowdsourcing on Audio Annotation for The Netherlands Institute for Sound and Vision

Anggarda Prameswari
Master of Information Sciences
VU University Amsterdam
2589263@student.vu.nl
Supervisor: Victor de Boer
Daily Supervisor: Themistoklis Karavellas

***Abstract* — Crowdsourcing and other human computation techniques have proven their use for collecting large numbers of annotations, including in the domain of cultural heritage. Most of the time, crowdsourcing campaigns are done through online tools. Local crowdsourcing is a variant where annotation activities are based on specific locations related to the task. This study investigates a local crowdsourcing application for audio annotations. We describe a specific use for annotating archival audio content to enrich its metadata. We developed a platform called "Elevator Annotator", to be used on-site. The platform is designed as a standalone Raspberry Pi-powered box which can be placed in an on-site elevator for example. It features a speech recognition software and a button-based UI to communicate with participants. We evaluate the effectiveness of the platform in two different locations and modes of interaction through a local crowdsourcing experiment. The results show that the local crowdsourcing approach is able to achieve annotations with 61% of accuracy, up to 4 annotations per hour. Given that these results were acquired from one elevator, this practice can be a promising method of eliciting annotations from on-site participants.**

Keywords—annotation; audio annotation; crowdsourcing; local crowdsourcing; Raspberry Pi

I. INTRODUCTION

Jeff Howe and Mark Robinson defined crowdsourcing as the act of a company or institution taking a function once performed by employees and outsourcing it to a generally large network of people [3]. This can take both form of collaborative work and sole individuals. This concept sparked the usage of crowdsourcing in different forms. Initially, crowdsourcing paradigms are divided into three categories: mechanized labor, where workers are rewarded financially; games with a purpose, where the task is presented as a game; and altruistic work, which is relied on goodwill [8]. Results from these categories are all depended upon the crowd wisdom and its methods. For example, Threadless[1] is known to practice the concept of crowdsourcing. Threadless is a web-based platform geared towards artist communities that can market their t-shirts there. These designs are crowdsourced through an on-going online competition on the Threadless website [5, 16]. Albeit, crowdsourcing is not only applicable for business and enterprise, but also for cultural heritage institutions. Cultural heritage institutions often dealt with digital archives in order to ensure preservation of their collections. Yet, as technology advanced, the collections are not only important to be archived digitally, but also preserved with enriched content for completion purposes. One way to do this is through annotation, in which human computations can be useful for collecting large quantities of data where a crowd of people helps with the annotation of specific contents [7].

The concepts of using shared information space done by cultural heritage institutions in order to gain engagement among the users and for them to annotate the collections were mostly done using the web technology. However, this practice is not the only solution for annotating the collections. Engaging the crowd in a physical location is another possible way in cultural heritage domain. This approach is called local crowdsourcing. Local crowdsourcing offers the same concept as online crowdsourcing with an addition of exploiting the physical environment as the source and then exposing the crowd to the possible influences from the location [4]. This study evaluates whether the use of local crowdsourcing, that can annotate cultural heritage collections, is an effective option to enrich the content of the cultural heritage's audio collections.

This study primarily focuses on both the audiovisual and broadcasting company archives from Netherlands Institute for Sound and Vision[2] (NISV hereafter). NISV is the largest audiovisual archive in the Netherlands that has a cultural-historical task and functions besides as the public broadcasting company archive is conducted in this study. Being one of the largest audio-visual archives in Europe, NISV collected and preserved a major part of Dutch audio-visual heritage so that it is accessible to as many users as possible. Digitalization is an essential part of this preservation both for efficient long-term management and for making the collection accessible. In this case study, audio collections were used to evaluate the local crowdsourcing approach, and the types of instruments identified in the audio is a part of the annotation.

The identification of musical instrument is an important aspect for information retrieval in musical domain, because such annotation is helpful for database indexing in the recordings. Furthermore, the identification of musical instrument can also be beneficial to classify musical genre, in which instruments were used as its distinctive features [1].

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

[2] https://http://www.beeldengeluid.nl

The ideal place to do the experiments is inside an elevator as it aligns with the goal of this study, which is to investigate whether local crowdsourcing is applicable for eliciting annotation on audio collection. In a cultural heritage institution, such as in NISV, elevators are always running during office hours and they are mostly used in a multi-level building. In this case, the crowd is anyone who used the elevators. This made the approach fell under the paradigms of both game and altruistic work. This study is designed to engage the crowd by approaching them in an interaction similar to playing a game with questions and answers, yet the result still depended on their willingness to participate regardless of any underlying motivation. Thus, an optimal design to effectively engage the crowd to annotate the audio collections and exploring possible influencing factors are the challenges to be overcome in this study.

The aforementioned reasons led to the decisions of naming this study as the Elevator Annotator. The local crowdsourcing approach is implemented in a portable audio annotation platform that is based on a pervasive computing technology concept [17]. The base of this approach is the Raspberry Pi[3]. This study chose the Raspberry Pi because it offers many possible integrations with different hardware and software components. It was fascinating to see whether the different type of institutions produced different results on both quantity and quality of the annotations or not due to the fact that the platform was implemented on-site. This also correlated with the concept of niche sourcing which focused on engaging participants that have knowledge in the domains so that they can perform more complex tasks [2]. Therefore, this study evaluated primarily on the differences on location as well as user input modalities, and their influences on the annotation results.

II. RELATED WORK

This section covers some of the previous researches related to this study. In addition, this study used concepts such as local crowdsourcing and niche sourcing from the past works in order to define its limit and the process of investigation.

A. On Local Crowdsourcing

Local crowdsourcing is a form of hybrid crowdsourcing process which extended the human computation into the physical environment through the use of on-site crowds as workers to do the given collaborative or distributive tasks [4]. Unfortunately, there is only a limited number of research done related to local crowdsourcing, especially with the focus of cultural heritage domains. Some of these researches were found by conducting the crowdsourcing approach with the web technology, while the most relatable research on local crowdsourcing was found in journalism domain. Nevertheless, the concepts of using hybrid traits of physical location and crowd engagements were correlated with this study. This related journalism research in the past has been conducted to investigate the collaborative production of event reports using local workers to collect information in person and remote workers to curate the collected information and generate event reports [4].

The results have revealed opportunities and challenges to extend the online crowdsourcing to the local, physical space. It showed 50% of additional media content on the reports and reasonable content for quality assessment when compared to similar topics from news sources, such as a local newspaper. Biases of the non-expert crowd workers in various locations were some of the challenges found even though they were constrained by the physical space. While being constrained by the physical space, bias of the non-expert crowd workers in different locations were some of the challenges. The results of this research in the past provided a solid background for this study, in which there were possible opportunities for local crowdsourcing to be an option for collecting on-site information. Furthermore, the results provided potential ways to develop an optimal design to overcome the underlying challenges on location constraints and its influences to become an effective audio annotation platform.

B. On Crowdsourcing in Cultural Heritage Domain

In cultural heritage domain, crowdsourcing has several categories based on its purpose. These categories are correction and transcription, contextualization, complementing the collection, classification, co-curation, and crowdfunding [6]. The local crowdsourcing approach of this study fell under the category of classification, where it is defined as gathering descriptive metadata related to objects in collection [6]. Based on the crowd level, this study fell under the contributory type where the public contributes data from the designed platform. A related research in this domain has been conducted as a collaboration of the NISV, KRO Broadcasting [4], and VU University Amsterdam [5] in the form of a video labeling game called Waisda?[6]. The research examined a collaborative way with the public through gaming as a method to annotate television heritage and use the curated vocabularies as a means to integrate tags with professional annotations [10]. The research was successful in getting a large number of matching tags in a 7 months period, as a result of the labeling tags in videos and qualifying them to the databases in NISV [7]. This Waisda? research showed that there are opportunities for studying a variety of innovations in the domains of cultural heritage and time-based metadata like tags can be used by media professionals for accessing specific fragments. These results led the possibilities of this study to explore other methods of annotating the audio collections, by identifying the musical instruments with on-site crowdsourcing.

This study also explored the difference of locations and its effects on the annotation also correlates with the concept of niche sourcing. A related research in the past had investigated the power of experts to optimize the result of human-based computation for certain tasks [2]. An example used in this related research was the prints annotation in the Rijksmuseum[7] that resulted on large quantities of metadata, which were not sufficient in quality. These results revealed challenges on task distribution and quality assurance in niche sourcing. Different crowds with different level of task produced different results,

[3] https://www.raspberrypi.org/
[4] https://www.kro.nl/
[5] https://www.vu.nl/
[6] http://waisda.beeldengeluid.nl/
[7] https://www.rijksmuseum.nl

whereas crowdsourcing by a generic crowd is more ideal for simple task. In this study, the quality of the audio annotation is assessed by its accuracy compared to the pre-annotated annotations, which in this case was done manually by a non-expert. The pre-annotation covered the instrument identification in a generic level (e.g. piano instead of pipe piano) to see if simple tasks could have worked well with the local crowdsourcing approach.

Therefore, these related works in the past showed some opportunities for this study to leverage the crowdsourcing area into the domains of cultural heritage and to take the on-site location as a way to engage the crowd. This study investigates challenges found in the previous works such as location boundaries and its effects on the crowds in order to examine a promising optimal design for a possible effective local crowdsourcing approach in annotating audio collection with an example of the NISV case study.

III. RESEARCH QUESTION

This study explores the application of the principles of local crowdsourcing, which are not only to passively make use the crowd power to do the human computation task, but also to engage the crowd to perform a musical instruments identification present in audio collections as a form of audio annotation. In cultural heritage domain of the NISV case study, preservation and audio content enrichment with the addition of metadata content are important. Then, this extended practice of local crowdsourcing with pervasive computing is evaluated to find out if it is able to deliver an effective result. The evaluation involved two dimensions, which are the different locations and the different user input modalities. The effectiveness of the method and the variations trough annotation accuracy and rate, are evaluated through its design and performance.

Therefore, the following research question is proposed to answer the goal of the study: *What is an effective method for local crowdsourcing metadata gathering for an audio collection?*

Yet in order to answer the main research question, these following sub-questions need to be answered at first:

1. What is an effective technical design for the local crowdsourcing platform?
2. How do the different physical locations of the local crowdsourcing affect the result?
3. How do the different types of user input modalities of the local crowdsourcing affect the result?

IV. ITERATIVE DEVELOPMENT

This section provides detailed explanations of the iterative development that was conducted in the study. The first section discusses the system design and the design decisions made during the process. The second section covers the implementation process of the design, both in terms of hardware and software. Hence, it is implemented as an audio annotation platform, and therefore the local crowdsourcing approach is addressed as the platform from this point.

A. System Design

Several design decisions were made along the process in order to ensure its functionality. The first early decisions related to its physical constraints and its functional designs.

i. **Location**: The elevator was chosen as the location, because it was the ideal place to keep the experiments running in a multi-level building. From the observations made in early stage of this study, people were found to be mostly idle inside the elevators and the time needed to move between floors were long enough to conduct the annotation process. It also showed that on average it required 7 seconds for the process to allow enough time for crowds to participate. Different type of locations was investigated to see if there are any other effects and differences. Therefore, NISV was chosen as it was in the cultural heritage domain in nature and VU University Amsterdam (VU Amsterdam hereafter) was chosen as a comparison due to the different traits that VU Amsterdam had to offer as an academic institution.

ii. **Form**: The platform was designed as a standalone platform, not only because of the pervasive computing concept used in this study, but also because of having the elevator as the location, the platform required to be easily placed and moved. The platform was also placed in a box to provide ease of access to the participants.

iii. **Power Supply and Connectivity**: As it was not possible to have power supply inside the elevator, the platform needed to have reliable energy source to power up. Because it was built on a Raspberry Pi, an external battery pack or a power bank was used to supply the power to the platform. While in terms of network connectivity, the elevators did not have enough coverage for the platform to have an online connection because it did not have a stable connection to the available wireless hotspot in between floors. Thus, the data processing and data storing were designed to be done locally on the Raspberry Pi.

These decisions worked as the frameworks to design a functional platform. An illustration of the functional design of the platform is shown in Figure. 1. The arrows represented interactions among the components and there was no necessary ranks or sequences. The two boxes showed grouped the components based on their interactions.

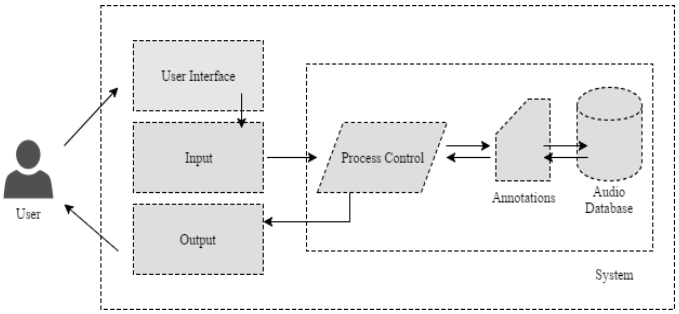

Figure. 1. Functional design

Several design concerns were derived from Figure 1. There were the input/output, process control, and the annotations and database.

i. **Input/Output**: Participants needed to be easily prompted by the platform in order to give their input. Different type of user input modalities were implemented on the user interactions as the audio and buttons input. Difference on the nature of verbal and nonverbal communication with these inputs are interesting to be examined. The type of interactions was done in questions and answers on the willingness of participants to do the annotation and the annotation of the type of instruments heard in the audio. Therefore, to keep the tasks simple for non-specified participants in both location, the expected answers is yes or no. Audio input was possible with an offline speech recognition library called Pocketsphinx[8]. It was chosen because the library was suitable for offline use [12]. Whiles for buttons input, there were two momentary push buttons represented the yes and no answers. For the output, two stereo speakers were utilized to ensure that it was loud enough to hear from the inside elevator.

ii. **Process Control:** All processes had to be done locally on the Raspberry Pi and without an internet connection. The processes also needed to start, run and terminate automatically inside the elevator. Therefore, a motion sensor was applied to trigger the annotation function to start whenever motion was detected and a timeout function was utilized to limit the number of answers attempts to ensure that the platform was kept running the entire time.

iii. **Annotations and Database**: Both annotations and the database were stored locally on the Raspberry Pi. The audio database in this study was stored as a set of audio files from the dataset. These specifications of the audio datasets, as well as the annotation structures, are discussed in details in its own dedicated section. Hence, the annotation was done by identifying a type of instrument heard within the specified time fragment in the audio file that was played over the speakers.

These design decisions were essentials to take into account when building the platform to ensure an optimal design of the platform in both hardware and software. These implementations were discussed further in the next sections.

B. Hardware Implementation

The platform was not only built on a Raspberry Pi as its computing base to connect other hardware components, but also to store and process the annotation. It was important for the platform to utilize both input and output hardware components to support its annotation functionality. A list of hardware components which were used to assemble the platform is provided in the following list.

a. Raspberry Pi 3 model B with 4Gb micro SD card running on Raspbian Jessie[9] [13]
b. PIR motion sensor
c. Mini USB microphone
d. Two momentary push buttons
e. Two 3" speakers 4ohm 3watt
f. 3.5mm (1/8") Stereo audio plug terminal block
g. Stereo 2.1watt class D audio amplifier (TPA2012) [14]
h. Power bank with minimum of 2A output

The annotation function started whenever a motion is detected when the participants entered the elevator. The platform then played the audio over the speakers. The audio input is supported with the USB microphone, while the buttons input supported with the two momentary push buttons, represented in green for the yes answer and in red for the no answer. An illustration of the wiring sketch is shown in Figure 2 that was used as the blueprint to assemble the hardware components to the Raspberry Pi.

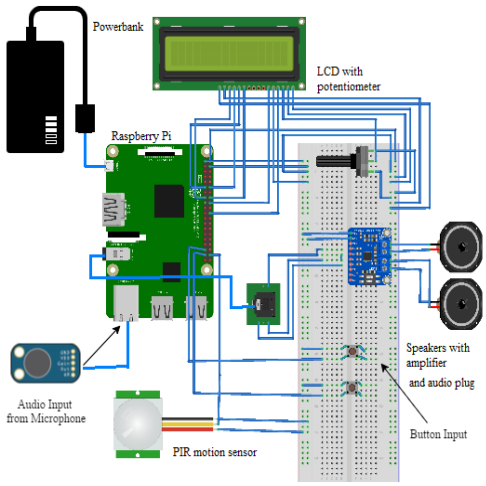

Figure. 2. Wiring sketch

The blue lines in Figure 2 represented the connection among the components, while the arrows showed the two user input modalities used in this platform. A dedicated tutorial and documentation on the assembly process, supporting libraries and components are documented in a Github[10] repository along with its web page. This repository also offered detailed explanations on the source code of the software. The final look of the hardware was implemented in a box and the pictures of the box are shown in Figure 3. The picture also provided an illustration for both types of user interaction. The left side of the picture showed the platform with audio input and the right side of the picture showed the platform with buttons input.

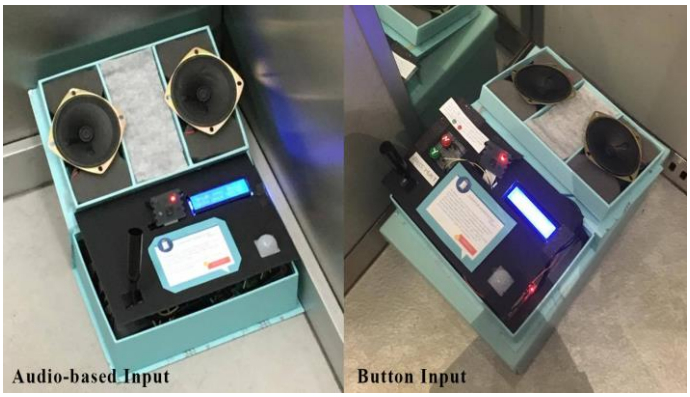

Figure. 3. Implemented platform

[8] https://cmusphinx.github.io/
[9] http://gnutoolchains.com/raspberry/jessie/
[10] https://ajprameswari.github.io/ElevatorAnnotator/

C. *Software Implementation*

The platform was running on a Python script that was implemented locally on the Raspberry Pi. The workflow of how the software was implemented and the interactions among the components are illustrated in Figure 4. The numbers were interpreted as how the software worked in sequences. The workflow is described in details in the following list. It was notable that the speech recognition library and timeout function were defined in the script and they were called whenever the participants were expected to give answers.

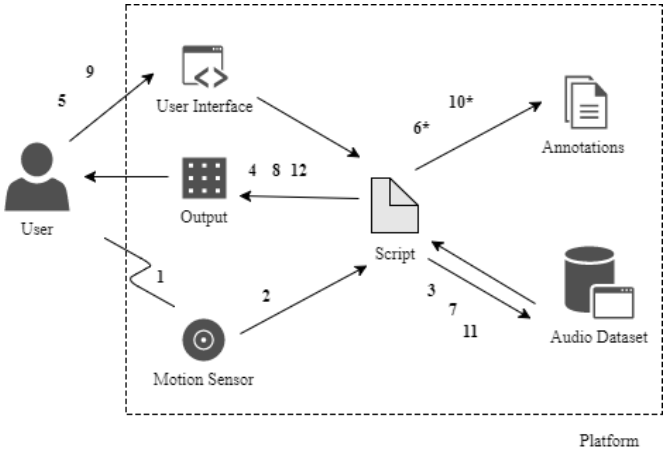

Figure. 4. Software workflow

1. Motion sensor detects any user motion.
2. Detected motion triggers the annotation function in the script to start.
3. The script gets audio files of the song and greetings from the dataset and loads them to the script.
4. The script plays the selected audio files over the output components (speakers) and prompts the user by playing the selected song, continued with the question on participation over the speakers.
5. User gives answers through the user interface (audio and buttons input) and the user interface forwards it to the script.
6. When the user disagrees to participate or the timeout is reached, the script saves the answer to annotations as a record, otherwise continue to next step if the user agrees.
7. The script gets the previously selected audio file of the song and instrument question from the dataset and loads them to the script.
8. The script plays the selected audio files over the speakers and prompts the user with the audio file of the selected song and question on the type of instrument heard.
9. User gives answers through the user interface and the user interface forwards it to the script.
10. The script saves the answers as a record or when the timeout is reached,
11. The script gets the audio file of closing notice to the user from the dataset and loads it to the script.
12. The script plays the closing audio file over the speakers and prompts the user.

The workflow was fundamental to design a working script. A flowchart illustrated both user input modalities is shown in Figure 5. This flowchart showed the logic of how the platform is supposed to work along with the termination conditions on every possible situation.

Figure. 5. Flowchart in audio and buttons input

The flowchart in Figure 5 works as the framework to develop the source code for the script. The annotation function starts every time the sensor detects any motion from the participants. The code snippet below shows how the PIR motion sensor keeps itself waiting to detect any motion from the participants and starts the annotation function. This implementation works for both types of user input modalities.

```
def MOTION(PIR_PIN):
       print("Motion Detected")
       startAnnotation()
def main():
     while True:
       pir.wait_for_motion()
       MOTION(16)
       pir.wait_for_no_motion()
```

In principle, the function starts the annotation by playing the audio file from the dataset randomly. There were in total 10 public domain songs as the dataset with 3 points of starting time in its total duration, which was 25%, 50%, and 75%. The selected audio file was then played for 7 seconds from the starting time in random. In order to manually pre-annotated the dataset for setting up the standards for evaluation, each of the audio was divided into 3 starting positions. The instruments were trumpet, violin, and piano. After the song was played then participants were asked if they wanted to participate or not. The maximum attempts of answer was limited to 3 attempts and with 10 seconds of timeout for buttons input. The numbers were limited to prevent the function to keep waiting for answers. The participants were expected to answer in two forms for two of the platforms.

```
def recogniseMe(mic):
  r = sr.Recognizer()
  try:
     with sr.Microphone(mic)as source:
     r.dynamic_energy_treshold = False
     r.energy_treshold = 4000
     audio = r.listen(source)
     answer=str(r.recognize_sphinx(audio))
     answer=answer.lower()
     return answer
   except sr.UnknownValueError:
       return None
   except sr.RequestError:
       return None
```

The code snipped above showed how the speech recognition library was used in the script. It showed that the microphone was defined as the default index to get input from the participants and energy threshold was set to 4000. Energy threshold was the minimum value for the speech recognition to set itself before it started recognizing phrases. This parameter depended on the noisiness of the surrounding. The lower the value the quieter the surrounding was. The energy value was adjusted several times during and was set up to 4000 due to the higher noise found inside the elevators. However, it was found that the value did not always work as it was expected because the speech recognition kept waiting for a longer time to recognize phrases when there were higher noises. Noises from walls, doors and conversations affected this threshold. A feature to automatically set this threshold was also tried, but it was found that the platform took less time when the value is pre-defined compared to when it was set automatically. When yes or no phrases were recognized by the speech recognition library, it passed the answers to the function, whereas when it failed to recognize them, the number of attempts was incremented. This applied for both participation and instrument answers.

Timeout was also another parameter that was calibrated for both audio and buttons input. The timeout feature from the speech recognition was found to be overruled by the waiting process of speech recognition to set up the energy threshold. While timeout did not work well with the audio input, it worked with the buttons input. The buttons input used a threaded timer to get the timeout to work. The timeout limit was set to 10 seconds to give time for the participants to get to the buttons. When the function reached the timeout in buttons input, the number of attempts for both participation and instrument answer were incremented. Maximum limit was set to 3 times to avoid the function to keep waiting for answers.

D. Dataset and Annotation Structure

The annotations were saved as each line of record within a file. From the some of parameters mentioned earlier, for example: the number of attempts and the type of answers. The structures of each record were determined to provide sufficient data to be evaluated. The structure of each record along with their formats and limitations are provided in Table 1.

TABLE I. ANNOTATION STRUCTURES

Element	Format/Limit
Timestamp	yyyymmdd-hh:mm:ss
Participation answer attempts	Maximum of 3
Participation answer	Yes/No
List of participation answers	False/True
Selected audio file	sound1-10
Start position in percentage	0.25/0.50/0.75
Selected type of instrument	trumpet/violin/piano
Instrument answer attempts	Maximum of 3
Instrument answer	Yes/No
List of instrument answer	False/True
Annotation length	In seconds

All annotation records were stored in a text file with tabular space as its delimiters for each of element. An example of the annotation structures is shown below.

```
20170517-11:53:47 1 YES T sound9 0.25 trumpet 2 YES
FT 38.0137310028
```

From this example, it can be derived that the annotation started on 17 May 2017 at 11:53 and the user was willing to participate in 1 answer attempt. The selected audio to be annotated was audio no. 9 and the starting time fragment of the audio file was from the 25% seconds of its duration. The participant was asked if a trumpet was heard and the participants answered that a trumpet was heard in 2 attempts. The last information from the record was the duration of the annotation from the the beginning of to th moment when the record was saved, and in this case, the example showed approximately 38 seconds.

In summary, all the settings and design decision in both hardware and software implementation were set up in various settings as the result of its iterative development. It is notable that a pilot testing was conducted in the beginning of the the development, to test the functionality of the platform and to calibrate the parameters mentioned in these sections. Adjustments on energy threshold, timeout, and error handling were also set up to improve the platform to perform better in getting annotations from the participants.

V. EXPERIMENT

This section describes how the experiments were, as a continuance from the previous iterative development. The experiments are discussed in several parts. There are dataset, setups, and the experiment results along with and its analysis. It should be noted that the experiments took place in multiple attempts and the pilot testing mentioned in in the previous section was used to improve the performance of the platform.

A. *Dataset*

The dataset that was used in this experiment were acquired from the Europeana Collection[11]. There was a total of 10 random public domain songs. There were no standards on the audio selection, other than the audio needed to have a good audio quality to be played over the speakers. As it was mentioned before, each audio was manually pre-annotated in advance, in order to set the standard for determining the correct or incorrect type of instruments being heard. A complete list of the audio with more information of its pre-annotated answers is provided in the Appendix section A.

B. *Setup*

The experiments took place in two locations, the NISV and VU Amsterdam. In each location, two experiments were carried out for the audio input and buttons input. For each experiment, the standalone platform was placed on the corner side of an elevator and it was left on average of for 4 hours. The platform was placed in a box to ensure its stability on the ground level.

Figure. 6. Setup of the platform in the experiments

Figure 6 showed the how the experiment was set up in one of the experiments.

The number of floors in both locations of NISV and VU Amsterdam was similar. NISV had 7 floors counted from the ground level, whereas VU Amsterdam had 6 floors. The elevator which was used in VU Amsterdam was located in the Science Building. In these experiments, participants were expected to give yes or no answers when asked about their willingness to participate and a type of instrument being from the audio. Albeit, it should be noted that, due to the technical limitation on the error handling and there was a possibility of the annotation to produce an invalid answer which was defined as "not available" (N/A) records, because the speech recognition failed to recognize phrases for audio input and when the timeout limit has exceeded for buttons input. Both failures led to the number of attempts being incremented and invalid records were still stored in this case.

TABLE II. EXPERIMENT RESULTS

		Length (minutes)	Total Recorded Answers	Unidentified Participation Answers	Identified Participation Answers		Instrument Annotation Answers			Accuracy Rate
					No Answers	Yes Answers	Correct Answers	Incorrect Answers	Unidentified Answers	
NISV	Audio	240	55	35	11	9	5	2	2	0.71
	Button	210	81	32	25	24	19	5	0	0.79
VU	Audio	240	67	38	18	11	3	6	2	0.33
	Button	240	61	18	22	21	11	9	1	0.55
Total in location	NISV	450	136	67	36	33	24	7	2	0.77
	VU	480	128	56	40	32	14	15	3	0.48
Total in user input modalities	Audio	480	122	73	29	20	8	8	4	0.50
	Button	450	142	50	47	45	30	14	1	0.68
Total		930	264	123	76	65	38	22	5	0.61

[11] http://www.europeana.eu/portal/

C. Results

From several experiments, the platform had succeeded to acquire a number of records. There were adjustments and calibrations made during the experiments to ensure the functionality of the platform. A complete view of the experiment results is available on Figshare[12] [9]. A summary of the results is shown in Table 2 In general, the acquired data were categorized in both locations of NISV and VU Amsterdam and user input modalities of audio and buttons input. It is notable that the length column in the result table was the approximated total time spent in each experiment

which was done in several attempts.

In total, these experiments took approximately 930 minutes, which were 15.5 hours or equivalent to 15 hours and 30 minutes. This gave the average time spent in each experiment of 232.5 minutes or equivalent to 3 hours and 52 minutes. There were 264 recorded answers which were consisted of 141 identified answers and 123 unidentified answers. The identified answers were the succeeded answers from participation question, while the unidentified answers consisted of those which were recorded due to exceeded attempts and timeouts.

Inclusively there were 65 participants who were willing to participate counted from the total of those answering yes and 76 participants who were not willing to participate. From the instrument annotation questions, there were 38 participants who correctly annotated and 22 who did not. The unidentified instrument answers were the result of the timeout and exceeded answers attempts. As mentioned earlier, the standard to determine the correct or incorrect answers are determined by the set from the manually pre-annotated dataset. Further evaluation of these statistics was given in the next sections.

D. Analysis

This section gives insights into the experimental findings. Performance, the significance of different locations, user input modalities, and crowd participation are evaluated. In order to determine the correctness, the instrument annotation answers were matched with the pre-annotated dataset and the following rules of confusion matrix in the list below were applied. It is notable that correct answers consisted of both True Positives and True Negatives, while incorrect answers consisted of both False Positives and False Negatives.

i. **True Positives** (TP) were cases when the participants predicted that the instrument was present in the audio and it was true.
ii. **True Negatives** (TN) were cases when the participants predicted that the instrument was not present in the audio and it was true that it was not.
iii. **False Positives** (FP) were cases when the participants predicted that the instrument was present in the audio, but it was not true.
iv. **False Negatives** (FN) were cases when the participants predicted that the instrument was not present in the audio, but it actually was.

E. Performance

Based on the aforementioned rules and comparison made to the actual pre-annotated instruments in the audio to the predicted annotation given by the participants, the accuracy rates of the annotation produced by the participants are calculated with the sum of True Positives and True Negatives in both input divided by the total data in the matrices for each dimension:

$$Accuracy\ rate = \frac{TP + TN}{TP + TP + FP + FN} \qquad (1)$$

Based on the location difference, the accuracy rate of annotations that took place in NISV was 0.77 ≃ 77% and the error rate was 1-0.77 = 0.23 ≃ 23%. Whereas, the accuracy rate in VU Amsterdam was 0.48 ≃r 48% with an error rate of 52%. As for difference on user input modalities, the accuracy rate for audio input was 0.50 ≃ 50%, so that its error rate was 50%. While for buttons input in both locations, the accuracy rate reached 0.68 ≃ 68% with an error rate of 32%. From this information, in average the accuracy of the platform was 0.61 or equivalent to 61%. Because both locations and user input modalities revealed the different accuracy rates, it was necessary to test them statistically. The tests were important to evaluate the significant difference on locations and user input modalities to the correctness of annotations.

The first variables were the locations and the annotation correctness. There were two possible annotations acquired in this study, which were the correct and incorrect answers. The correct annotations were the answers that met the condition of True Positives and True Negatives, while the incorrect answers met the conditions of False Positive and False Negatives. Their frequencies in each of locations were counted to see if there were any association. In this case, the appropriate statistical test was the Chi-square test [11], because the variables were categorical and thus they were suitable for the test. The hypotheses for the first test were constructed as:

H_0: There is no association between locations and annotation correctness
H_1: There is an association between locations and annotation correctness

The significance level (p-value) of the test was the standard level 0.05 which indicated a 5% risk The Chi-square test value was calculated as 5.480. With degrees of freedom of 1 and p-value of 0.05, the critical value based on chi-square distribution table [15] was 3.84. Since the test value was more than the table value, then null hypothesis (H_0) is rejected. The p-value for the test has resulted as 0.019 and calculation table is provided in the Appendix section B and C. Therefore, it was concluded that there was an association between location difference and the performance of annotations correctness. It was statistically significant because the p-value test was less than 0.05. With the highest accuracy acquired from the data was taken in NISV, it showed that locations with different traits did have significance difference, where NISV as a cultural heritage institution resulted in higher accuracy compared to VU Amsterdam as an academic institution.

[12] https://figshare.com/

The second variables were the user input modalities and annotation correctness. The same Chi-square statistical variables were used, such as p-value of 0.05 and degree of freedom of 1. For this significance test, the hypotheses were constructed as:

H_0: There is no association between user input modalities and annotation correctness

H_1: There is an association between user input modalities and annotation correctness

Table for the Chi-square test was also provided in the Appendix section B and C. The value of Chi-square test for these variables was 1.670. Because the same critical values from the Chi-square distribution table of 3.84 were used, then the test value is less than the critical value of 3.84. Therefore, the null hypothesis was accepted and that there was no association between the type of user input and annotations correctness. It also meant that these two variables were independent. The p-value of the test was 0.196. Because 0.196 was higher than the earlier p-value of 0.05, this showed that these variables did not correlate to one another.

Therefore, in summary of the applied two statistical significance tests, it indicated that first, different locations had significant associations with the annotation performance, in the sense of its correctness and second, different user input modalities had no significant association with the annotation performance, given it was calculated from the observed data. Whereas, based on the accuracy rates derived from experiment result in Table 2, the average accuracy of the annotations performed by the platform was 61% with the error rate 27%.

F. Crowd Participation

Another parameter to measure the effectiveness of the platform was the crowd participation. It was essential to evaluate whether the local crowdsourcing approach on this platform offered an effective result for acquiring annotations or not, based on the number of people participated in it. In order to do so, information on the numbers of participants was derived from the experiment results in Table 2. First questions given to the participants were whether they wanted to participate or if they did not want to participate and a yes or no answer was expected. From the total of 141 participation answers, there were 65 people who agreed and 76 who disagreed to participate. This gave the participation ratio to a total number of answers in the experiments of 0.46. In contrary, the number of people who did not want to participate was slightly higher with 0.54. While the total time spent for the experiments was approximately 930 minutes or 15.5 hours and given that there were 60 valid and identified annotated instrument answers were, dividing these two values resulted in 3.87 annotations or equivalent to up most 4 annotations per hour. It is notable these numbers were acquired from the result of having the experiments in one physical location at a time.

VI. DISCUSSION

In summary, the main functionality of the platform as a local crowdsourcing approach for audio annotations had proven to succeed acquiring a number of results. Given there were some limitations on the implementation, the results showed that the platform was able to perform as accurate as 61% and was able to acquire 3.87 annotations per hour in one physical location by one platform. With the participation ratio of the people who were willing to annotate of 0.46. The result also showed that while different user input modalities did not have significant association on the platform performance in the sense of its annotation correctness, different locations had significant association. The result of the statistical tests performed also gave further insights on how the numbers can grow in larger population.

The result showed accuracy rate was acquired at the highest in NISV and that it was statistically significant that there was an association in location and the annotation correctness. As mentioned in earlier chapters, since the experiments took place in NISV which is a cultural heritage institution, the participants involved in the experiments were mostly employees who were associated with the domain. This was associated with the niche sourcing principle that crowds of experts produced different level of quality compared to generic crowds [2]. However, further exploration on this matter needed to be done in future work as it was not the primary concerns of this study. Having the standard pre-annotated answers validated by experts on this domain offered different possibilities on the accuracy and reliability of the annotation.

In regards to these rates that were acquired as a result of one platform running in one elevator at a time, there were rooms for improvement in result numbers for future work by double the number of platform used. Different type of speech recognition library, better hardware components, and error handling features have options to be improved as the platform itself was built on a customizable base, such as the Raspberry Pi. Limitations in speech recognition library and its parameters such as energy threshold and timeout limits had influenced the data acquisition. The waiting time spent by the speech recognition library to understand the spoken phrases had caused less data to be acquired, compared to the data acquired from the button based user interactions. Energy threshold which may be adjusted to a different level of noises inside the elevators also have its effect onto the waiting time for its calibration function. Therefore, a different type of speech recognition library is an option to be improved in future work.

There was also an opportunity to locally crowdsource this experiment at different places other than inside an elevator. Different locations with different crowd options offer options for improvement in results, for example at the coffee corners or vending machines where the crowds were. Having the platform to be placed inside an elevator had its own perks, where its constant. In conclusion, this study has shown how local crowdsourcing approach can work when it is integrated with pervasive computing elements such as the use of Raspberry Pi. Ideally, in order to get a larger quantity of data, more platforms can be placed at the same time in several different locations to do the annotation task. The Raspberry Pi and components used in this study were suitable because they were not only low in energy consumption but also affordable and easy to build for different settings.

VII. CONCLUSION

According to the goals of this study which aims to evaluate whether this local crowdsourcing approach offered an effective solution in eliciting audio collection in a cultural heritage domain, with an example of the NISV case study. First, it is important to get a better understanding on an optimal design to build and what other influencing factors are. Then, the two dimensions such as difference on locations and user input modalities are taken into account in the implementation and experiments. From the acquired experiment results and analysis of the experiment, the following research questions can be answered.

1. *What is an effective technical design for the local crowdsourcing platform?*
   The local crowdsourcing platform can be designed as a standalone platform on a Raspberry Pi with supportive input and output components to support the on-site annotation purpose and its process control. The processing and data storing of the platform were developed locally on the Raspberry Pi, using an offline speech recognition for audio inputs and momentary push buttons for buttons input. Despite the location constraints, the platform was able to perform its main function and acquire a number of results. Thus, it is shown that the platform offered an effective design due to its extended portability of being implemented in the external battery powered Raspberry Pi. These features also showed that the platform can be built with low budget and low energy consumption components.
2. *How do the different physical locations of the local crowdsourcing affect the result?*
   Based on the experiment results, the average of annotations collected in NISV is resulted in 77% being accurate, while in VU Amsterdam its accuracy is 50%. From the significance test, it indicated that there were associations between different type of locations and annotation performance on correctness. It showed that the difference is statistically significant.
3. *How do the different types of user input modalities of the local crowdsourcing affect the result?*
   Based on the experiment results, the average of annotations collected with audio input is 50% as accurate, while the buttons input had of 68% accuracy rates. From the significance test, it showed that there was no significant difference and association in between the type of user input modalities and the annotation platform correctness.

From the information and answers given from each of the previous questions, conclusion of the main research question's answer can be derived: *What is an effective method for local crowdsourcing metadata gathering for an audio collection*?

The local crowdsourcing approach, which was designed and implemented in this platform, offered an effective solution for eliciting annotations from on-site participants. It was as accurate as 61% with up to approximately 4 annotations per hour. It showed that there was a significant association between the different locations and its annotation performance. Given the platforms were placed in multiple on-site locations, numbers of annotation had the chance to be increased as it was tested statistically that locations had significant effects to the annotation performance. Therefore, this local crowdsourcing approach, which was combined with pervasive computing components from the platform, showed that the built design was promising for metadata gathering on an audio collection as in NISV case study's.

ACKNOWLEDGMENT

This study would not have been possible without the help and guidance from both of my supervisors, Victor de Boer from VU University Amsterdam, and Themistoklis Karavellas from the Netherlands Institute for Sound and Vision, who have been supporting and sharing their knowledge from the beginning. In addition, many thanks to the Netherlands Institute for Sound and Vision for providing me with opportunities to be a part of this project, along with the access to various dataset from the Europeana Sounds. Lastly, I would like to thank you my sister, Paramita Putri, for helping me to proofread this paper. To this extend, I must express my very profound gratitude.

REFERENCES

[1] McKay, C., & Fujinaga, I. (2005, March). Automatic music classification and the importance of instrument identification. In Proceedings of the Conference on Interdisciplinary Musicology.

[2] De Boer, V., Hildebrand, M., Aroyo, L., De Leenheer, P., Dijkshoorn, C., Tesfa, B., & Schreiber, G. (2012, October). Nichesourcing: Harnessing the Power of Crowds of Experts. In EKAW (pp. 16-20).

[3] Howe, J. (2006). The rise of crowdsourcing. Wired magazine, 14(6), 1-4.

[4] Agapie, E., Teevan, J., & Monroy-Hernández, A. (2015, September). Crowdsourcing in the field: A case study using local crowds for event reporting. In Third AAAI Conference on Human Computation and Crowdsourcing.

[5] Brabham, D. C. (2010). Moving the crowd at Threadless: Motivations for participation in a crowdsourcing application. Information, Communication & Society, 13(8), 1122-1145.

[6] Oomen, J., & Aroyo, L. (2011, June). Crowdsourcing in the cultural heritage domain: opportunities and challenges. In Proceedings of the 5th International Conference on Communities and Technologies (pp. 138-149). ACM.

[7] Oomen, J., Belice Baltussen, L., Limonard, S., van Ees, A., Brinkerink, M., Aroyo, L., ... & Gligorov, R. (2010). Emerging practices in the cultural heritage domain-social tagging of audiovisual heritage.

[8] Sabou, M., Bontcheva, K., Derczynski, L., & Scharl, A. (2014). Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In LREC (pp. 859-866).

[9] Anggarda, P. (2017). Experiment Results. figshare. Retrieved 1 July 2017, from https://doi.org/10.6084/m9.figshare.5106844.v1

[10] beeldengeluid/waisda. (2017). GitHub. Retrieved 1 July 2017, from https://github.com/beeldengeluid/waisda

[11] Chi-squared test. (2017). En.wikipedia.org. Retrieved 1 July 2017, from https://en.wikipedia.org/wiki/Chi-squared_test

[12] cmusphinx/pocketsphinx. (2017). GitHub. Retrieved 1 July 2017, from https://github.com/cmusphinx/pocketsphinx

[13] DebianEdu/Documentation/Jessie - Debian Wiki. (2017). Wiki.debian.org. Retrieved 1 July 2017, from https://wiki.debian.org/DebianEdu/Documentation/Jessie/

[14] Industries, A. (2017). Stereo 2.1W Class D Audio Amplifier - TPA2012 ID: 1552 - $9.95 : Adafruit Industries, Unique & fun DIY electronics and kits. Adafruit.com. Retrieved 1 July 2017, from https://www.adafruit.com/product/1552

[15] Schoonjans, F. (2017). Values of the Chi-squared distribution table. MedCalc. Retrieved 1 July 2017, from https://www.medcalc.org/manual/chi-square-table.php

[16] Threadless. (2017). En.wikipedia.org. Retrieved 1 July 2017, from https://en.wikipedia.org/wiki/Threadless

[17] Ubiquitous computing. (2017). En.wikipedia.org. Retrieved 1 July 2017, from https://en.wikipedia.org/wiki/Ubiquitous_computing

APPENDIX

A. TABLE OF PRE-ANNOTATED DATASET

No	Title	Filename	Classification	Pre-annotated Instrument Answers			Duration
				25%	50%	75%	
1	Oktāvu etīde	sound1	Piano music	Piano	Piano	Piano	1m34s
2	Tautas polka	sound2	Polkas, Folk dancing	Violin, Trumpet	Violin, Trumpet	Violin	2m40s
3	Dienā jaukā	sound3	Popular music	Violin, Trumpet	Trumpet	Violin, Trumpet	2m59s
4	Florentine	sound4	Popular music, Foxtrots	Trumpet	Trumpet	Violin, Trumpet	2m59s
5	Mana dzimtene	sound5	Foxtrots	Violin, Trumpet	Violin, Trumpet	Violin,Trumpet	2m57s
6	Meitenes sirsnina	sound6	Operas	Violin, Piano	Violin, Piano	Violin, Piano	2m8s
7	Kāpēļ tik ilgi vilcinies tu?	sound7	Foxtrots, Jazz,	Violin, Trumpet, Piano	Trumpet, Violin	Trumpet, Piano, Violin	2m46s
8	Dziedu tev	sound8	Popular music	Violin, Piano	Trumpet, Piano,Violin	Violin, Piano	2m53s
9	Serenade iz operas	sound9	Operas, Arranged	Piano	Piano	Piano	1m57s
10	Serenade	sound10	Violin with orchestra	Violin, Piano	Trumpet, Piano, Violin	Trumpet, Piano, Violin	2m31s

B. TABLE OF OBSERVED AND EXPECTED COUNTS FOR CHI-SQUARE TEST

	Observed Counts				Expected Counts		
	Correct	Incorrect	Row Total		Correct	Incorrect	Row Total
Difference on Locations							
NISV	24.00	7.00	31.00	NISV	19.63	11.37	31.00
VU	14.00	15.00	29.00	VU	18.37	10.63	29.00
Column Total	38.00	22.00	60.00	Column Total	38.00	22.00	60.00
Difference on User Input Modalities							
Audio	8.00	8.00	16.00	Audio	10.13	5.87	16.00
Buttons	30.00	14.00	44.00	Buttons	27.87	16.13	44.00
Column Total	38.00	22.00	60.00	Column Total	38.00	22.00	60.00

Expected Counts for each cells = $\frac{Column\ Total \times Row\ Total}{Grand\ Total}$

C. TABLE OF CHI-SQUARE TEST

	Observed (O)	Expected(E)	O - E	(O - E)^2	(O - E)^2 / E
Difference on Locations					
NISV - Correct	24.00	19.63	4.37	19.10	0.97
VU - Correct	14.00	18.37	-4.37	19.10	1.04
NISV - Incorrect	7.00	11.37	-4.37	19.10	1.68
VU - Incorrect	15.00	10.63	4.37	19.10	1.80
				Chi-Square Value	5.49
Difference on User Input Modalities					
Audio - Correct	8.00	10.13	-2.13	4.55	0.45
Buttons - Correct	30.00	27.87	2.13	4.55	0.16
Audio - Incorrect	8.00	5.87	2.13	4.55	0.78
Buttons - Incorrect	14.00	16.13	-2.13	4.55	0.28
				Chi-Square Value	1.67

Chi-Square Value $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$