Evaluating Acoustic Features from Web Audio Recordings A Crowdsourcing Survey on Background Noise Characteristics

Rafael Zequeira Jiménez, Babak Naderi, Sebastian Möller
Quality and Usability Lab, Technische Universität Berlin, Germany
rafael.zequeira@tu-berlin.de | babak.naderi@tu-berlin.de | sebastian.moeller@tu-berlin.de

Abstract

Crowdsourcing permits to reach a large pool of users for gathering and annotating data in an efficient and cost effective manner. In a crowdsourcing context, users employ their own hardware to execute the tasks from the comfort of their environment. However, there is too little information about the users' background noise and environment characteristics, which is mandatory to judge the validity of the data being collected. Specially, in speech quality assessment and audio related tasks. There have been some attempts to investigate the conditions in which users from English speaking countries, India and Asia conduct crowdsourcing tasks, yet any effort have been made regarding the German crowd-workers. To address this issue, a study in a German-based crowdsourcing platform has been conducted. Users answered questions regarding the environment in which they normally execute crowdsourcing tasks. Additionally, while conducting the survey, audio and visual data was collected to validate the submitted answers. This paper reports about the environment conditions in which crowdsourcing tasks are being executed. And we evaluate whether the wavelet time scattering and MFCC features derived from the collected background noise files, are any good to predict the workers environment background noise.

Introduction

Crowdsourcing (CS) has established as a mechanism to distribute and accomplish tasks over the Internet. In a CS paradigm, small tasks that normally require human intelligence for being resolved, are outsource to anonymous individuals over the Internet. The users (also called workers or crowd-workers) can carry out those tasks from their computer in exchange of a monetary compensation. This approach is being adopted in multiple domains as a fast and low cost way to collect human input for data acquisition and labeling [1]. Experiments conventionally executed in a laboratory setup can now be addressed to a wider and diverse audience.

Still, conceptual and technical challenges arise due to the remote test settings, and the fact that crowd-workers work without supervision. Some of the issues arising from CS studies have been already addressed in the literature to some degree, e.g. the reliability of user ratings [2], task length [3], workers performance [4] and influence of environmental conditions [5].

Often, users in CS do not follow the given instructions to execute a certain task. This is especially critical when it comes to human-centered studies that investigate the user experience of multimedia content. For instance, crowd-workers might not be able to judge properly about the quality of audio files if they conduct the task from a place where they could be potentially distracted. Or they might not be able to assess properly the quality or impairments in speech files if they conduct the test in a place with a loud background noise. All of these situations would lead to poor experimental results. To avoid such a waste of resources to some extent, we first need to understand about the environment characteristics in which workers typically execute CS tasks. And also about the common sources of background noise they experience when doing CS jobs.

In this work, we report on the results of a survey about the environment characteristics of workers from German speaking countries. A CS study has been conducted with participants recruited through a German based CS platform. We collected audio and visual data, which we used to contrast with the workers responses. Furthermore, we investigate if the MFCC (Mel Frequency Cepstral Coefficients) and the wavelet time scattering features derived from the collected background noise files, are any good to infer the workers environment background noise. We expect our findings to help the Quality of Experience (QoE) research community to design human centered studies in CS proactively, taking into account the workers' environments characteristics so reliable results can be gathered.

Study Overview

The goal of the crowdsourcing study was to gather information about the environment characteristics of crowd-workers from German speaking countries. For this purpose we used clickworker which is a German-based CS platform that reported to have 1.5 million users worldwide (15% from Germany and 25% from other European countries [1]). Then, we addressed our study to crowd-workers from Germany, Austria and Belgium. This platform have been employed successfully in a number of user studies that required German participants [3, 6, 7], therefore a good fit to our experimental needs.

The CS study had three different phases, i.e. "Audio recording setup", "Environment video recording", "Envi-

[1] https://www.clickworker.com/about-us/clickworker-crowd

ronment Questionnaire". We designed and implemented a HTML JavaScript based framework to administer the test to the workers and to collect the data. 325 workers participated in total in the study. A number of them stopped the study at different phases and unfortunately, we did not collect the same amount of data at each of the stages of our experiment.

Setup for Audio Recording

This phase was a tutorial to instruct and guide the workers to disable some of the noise reduction options that are normally enable by default in Windows and macOS computers. This was important to collect accurate audio recordings of the users environmental scene. Otherwise, the recordings would be too corrupt to extract any useful information from them. Additionally, we request the participants to take and upload a screenshot of their configuration, so they could prove to have the proper setup for the recording.

248 workers in total submitted a screenshot (183 from Windows). We manually analyzed the files and found that 54% of the workers configured their computers according to our instructions, 23.4% failed and 22.6% were not able to set the requested configuration (e.g. the options to select were not available on their computers).

Environment Video Recording

In this second phase, workers were requested to take and submit a short video of their current environment. To preserve privacy, we pointed out that they should avoid recording other people or documents with sensitive information, nor any element that would permit to identify their identity. These videos were manually analyzed to verify that they current environment corresponded with the one they reported in the questionnaire.

We collected a total of 230 videos from different users and recognized a house room in 82.17% of the recordings (mostly living room, bedroom, workroom or kitchen). An office space was identified in 3.48% of the videos and in some others, workers failed to do a proper recording or they submitted a video different from what we were asking. More details are presented in [8].

Environment Questionnaire

The third and last phase of the study was a questionnaire in which workers answered questions regarding the environment characteristics of the place in which they normally conduct CS tasks, and also about the environment in which they were at that moment.

213 crowd-workers responded to the environment questionnaire. Most of them were conducting the study from home (86.9%) or from work 7.5%. 93.4% of the workers reported that they normally execute CS tasks from home and, alone in a room while doing so (79.6%). Overall, around 60% of the workers spend between one and three hours per week doing CS tasks. And a number of them expressed to take part in other activities while doing CS jobs, e.g. listening to music (28.4%), other activities in computer (18.3%), executing other CS tasks (16.3%), watching TV (15.1%). See [8] for more details.

The last two elements of the questionnaire were two open ended questions that we collected in a text field. This information we analyzed with the "IBM® SPSS® Text Analytics for Surveys". This program uses Natural Language Processing (NLP) to code the input text into terms and it creates categories.

The first open ended question was to report which sounds they could hear at that moment. Our analysis revealed that 21.6% of the 213 workers could hear car or street noises, 18.8% the TV, 14.6% other people talking, and 13.6% declared that it was quiet (see Table 1.A). The second open ended question asked the workers to report what sources of noise normally distract them when they execute jobs at clickworker or other CS platforms. We found that 33.3% of the workers get distracted by other people, 28.6% by phone calls and 17.8% reported to work from a quiet place. More details in Table 1.B.

Tabelle 1: Categories found in the open ended answers from the 213 workers that reported about the noises or sounds they could hear while conducting our experiment (Table 1.A), and the sources of distractions they face when executing CS tasks (Table 1.B).

Category	%
car	21.6
TV	18.8
people	14.6
quiet	13.6
computer	13.6
music	7.0
bird	5.2
radio	3.8
baby	3.3
keyboard	2.8
mouse click	2.3
mowing	1.4
dog	0.9

TABLE 1.A

Category	%
people	33.3
phone calls	28.6
quiet	17.8
smartphone	6.6
baby	5.2
TV	5.2
music	2.8
social media	2.8
pets	1.4
WhatsApp	1.4
chat	0.9
autos	0.5

TABLE 1.B

Analysis of Background Noise Recordings

When workers answered the first item of the environment questionnaire, a JavaScript code embedded within the HTML permitted to record the workers' environment background noise during 15 seconds. We collected a total of 131 background noise files from different users.

These audio files were labeled manually according to the sound they contained and to whether they carried any information at all. Then, we evaluate if from these files we could extract useful features that would allow us to train a classifier to automatically identify the background noise. There is a big diversity in computers, type of microphones and softwares, specially in the Windows eco-

system, which make it difficult for collecting useful audio environmental recordings via web browsers. Most of the limitations of current web standards have been slightly compensated with the introduction of the "Web Audio API-[9] which enables extended access to the host audio interface, including sample manipulation of audio streams and control of the channel configuration. However, a lot of work still have to be done.

The number of background noise files collected was low compared to the data gathered in other phases. We believe that some users did not granted access to their microphone when requested, or that hardware or browsers issues happened on their system which prevented the recording from happening.

Table 2 present the labels that were manually assigned to the environment background files. In 43.08% of the recordings we were not able to identify any noise due to the recording being corrupted, and in 23.85% the background noise was low or not present, therefore the label "quiet". Additionally, we detected in some recordings noises from the kitchen, bird sounds, mobile ring tones, ventilation systems, water flowing and a radio. However, those were detected in only one recording in most cases and thus, not included in the subsequent analysis.

Features Evaluation

Furthermore, we evaluate whether the features derived from wavelet time scattering and from MFCC (Mel Frequency Cepstral Coefficients) extracted from the environment noise recordings can be used to make predictions about the workers' background noise conditions.

In wavelet scattering, the information is propagated through a series of wavelet transforms, nonlinearities, and averaging to produce low-variance representations of the data, which are then used as inputs to a classifier. This approach has been used successfully in other audio applications such as music genre classification, and achieved state-of-the-art performance [10].

The parameters needed to be specified in a wavelet time scattering framework are the duration of the time invariance, the number of wavelet filter banks, and the number of wavelets per octave. We computed the wavelet scattering features for the collected background noise files and resulted in a matrix of $2541 - by - 504$. Each row of the resulting feature set is one scattering time window across the 504 paths in the scattering transform of each audio signal. For each file, we have 21 of such time windows. Then, the number of rows is equal to the number of background files (121) multiplied by the number of scattering windows per example (21).

We considered the labels that were manually assigned to the recording (see Table 2). MatLab was employed to tried out all of the most common classification algorithms. Since we had 504 features, we applied principal component analysis (PCA), and parameters were optimized in a 5-fold cross-validation. After training only five components were kept by PCA, which was enough to explain 95% variation on the data. An Ensemble classifier with a preset of subspace KNN achieved the best accuracy of 84.1%, Figure 1 show its confusion matrix. Table 3 present the four classifier with the highest accuracy and the one with the lowest.

Tabelle 2: Labels assigned to the audio recordings.

Background Noise	No. of Files	Percentage
not defined (NA)	56	43.08%
quiet	31	23.85%
TV	18	13.85%
electric device	11	8.46%
music	8	6.15%
street noise	4	3.08%
people talking	3	2.31%

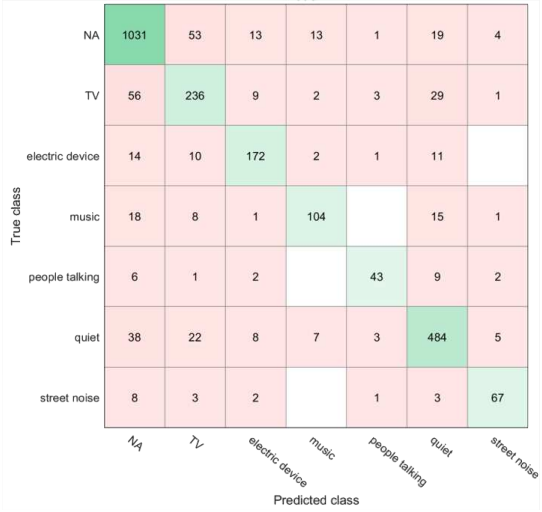

Abbildung 1: Confusion matrix of the Ensemble (preset: subspace KNN) classifier with the highest accuracy on the wavelet time scattering feature set.

Additionally, MFCC features were extracted from the background audio recordings to train the aforementioned classifiers and to evaluate their performance. MFCC has been widely used in speech and music applications, and its adoption is mainly due to the stability of the coefficients against signal deformations. Recently, MFCC has been employed successfully for environmental sound classification [11], and for audio event recognition [12].

The first 13 MFCC coefficients were computed using a 30 ms windows with 75% overlap. The resulting feature matrix had a dimension of $257850 - by - 13$. Again, a 5-fold cross-validation was set and no PCA was applied this time. The best accuracy was achieved by an Ensemble Bagged Trees classifier (82.8%). Figure 2 show the confusion matrix.

Conclusion

This work presents preliminary results from a study that investigated the environment characteristics of crowdworkers from German speaking countries. A survey was conducted in a German based CS platform and audio and visual data was collected from the workers environmental

Tabelle 3: Four best classifiers (and worst) trained with features derived from wavelet time scattering.

Classifier	Accuracy (%)
Ensemble (Preset: Subspace KNN)	84.1
Ensemble (Preset: Bagged Trees)	83.4
KNN (Preset: Weighted KNN)	82.8
KNN (Preset: Fine KNN)	82.5
SVM (Coarse Gaussian)	44.6

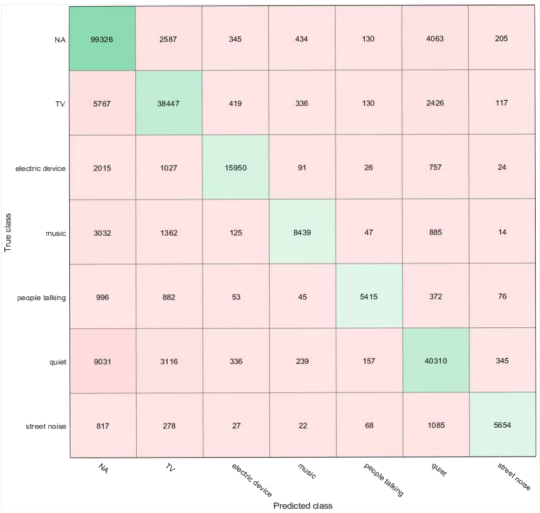

Abbildung 2: Confusion matrix of the Ensemble Bagged Trees classifier on the MFCC feature set.

scene. We evaluate whether the features derived from wavelet time scattering and from MFCC extracted from the environment audio files, that were recorded via the web browsers, are useful to make predictions about the workers' background noise. More details about the findings from the environmental survey are reported in [8].

We found that some background recording files were corrupted despite the computers being properly configured. Nevertheless, the wavelet time scattering and MFCC features permitted to train an Ensemble classifier that predicts the workers' background noise with an accuracy above 80%. Our results suggest that, despite the diversity in hardware and software, audio web recordings could be used to infer information about the background noise characteristics of users in CS.

Literatur

[1] Sebastian Egger-Lampl, Judith Redi, Tobias Hoßfeld, Matthias Hirth, Sebastian Möller, Babak Naderi, Christian Keimel, and Dietmar Saupe, "Crowdsourcing Quality of Experience Experiments," in *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*, Daniel Archambault, Helen Purchase, and Tobias Hoßfeld, Eds., Cham, 2017, pp. 154–190, Springer International Publishing.

[2] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger, "SOS: The MOS is not Enough!," in *Third International Workshop on Quality of Multimedia Experience (QoMEX)*, 2011, pp. 131–136.

[3] Rafael Zequeira Jiménez, Laura Fernández Gallardo, and Sebastian Möller, "Influence of Number of Stimuli for Subjective Speech Quality Assessment in Crowdsourcing," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, may 2018, pp. 1–6.

[4] Rafael Zequeira Jiménez, Anna Llagostera, Babak Naderi, Sebastian Möller, and Jens Berger, "Modeling Worker Performance Based on Intra-rater Reliability in Crowdsourcing, A Case Study of Speech Quality Assessment," in *accepted for: 11th International Conference on Quality of Multimedia Experience (QoMEX)*, 2019.

[5] Babak Naderi, Sebastian Möller, and Gabriel Mittag, "Speech Quality Assessment in Crowdsourcing: Influence of Environmental Noise," in *44. Deutsche Jahrestagung für Akustik (DAGA)*, Alte Jakobstraße 88, 10179 Berlin, 2018, pp. 229–302, Deutsche Gesellschaft für Akustik DEGA e.V.

[6] Babak Naderi, Sebastian Möller, and Rafael Zequeira Jiménez, "Evaluation of the Draft of P.CROWD Recommendation," ITU-T Contribution SG12-C.204, International Telecommunication Union, CH-Geneva, may 2018.

[7] Rafael Zequeira Jiménez, Anna Llagostera, Babak Naderi, Sebastian Möller, and Jens Berger, "Intra- and Inter-rater Agreement in a Subjective Speech Quality Assessment Task in Crowdsourcing," in *accepted for: Companion Proceedings of the 2019 World Wide Web Conference*. 2019, WWW '19, International World Wide Web Conferences Steering Committee.

[8] Rafael Zequeira Jiménez, Babak Naderi, and Sebastian Möller, "Background Environment Characteristics of Crowd-Workers from German Speaking Countries. A Survey on User Environment Characteristics," in *submitted to: 11th International Conference on Quality of Multimedia Experience (QoMEX)*, 2019.

[9] Paul Adenot and Raymond Toy, "Web Audio API, W3C Candidate Recommendation," 2018.

[10] Xu Chen and Peter J Ramadge, "Music Genre Classification Using Multiscale Scattering and Sparse Representations," in *47th Annual Conference on Information Sciences and Systems (CISS)*, 2013, pp. 1–6.

[11] Dharmesh M Agrawal, Hardik B Sailor, Meet H Soni, and Hemant A Patil, "Novel TEO-based Gammatone Features for Environmental Sound Classification," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1809–1813.

[12] Anurag Kumar and Bhiksha Raj, "Features and Kernels for Audio Event Recognition," *arXiv preprint arXiv:1607.05765*, 2016.