Large-scale Data Collection and Analysis via a Gamified Intelligent Crowdsourcing Platform

Simone Hantke[1,2] Tobias Olenyi[1] Christoph Hausner[3]
Tobias Appel[3] Björn Schuller[1,4]

[1] ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

[2] Machine Intelligence &Signal Processing Group, Technische Universität München, München, Germany

[3] audEERING GmbH, Gilching, Germany

[4] Group on Language, Audio & Music (GLAM), Department of Computing, Imperial College, London SW7 2AZ, UK

1 Introduction

The present emerging trend for innovative artificial intelligence applications and deep learning technologies is unbroken, leading to a tremendous need for large-scale labelled training data to adequately train newly developed systems and their underlying machine learning models.

In particular, for audio classification, training data is required to come from large pools of speakers in order for models to generalise well. Current technologies bring the opportunity to collect masses of new data via the Internet, making use of ubiquitous embedded microphones in laptop PCs, tablets and smartphone devices. This technological progress enables collection of speech data under real-life conditions (e.g., different microphone types, devices, background noises, reverberations, to name but a few) of a large number of speakers with different geographic origins, languages, dialects, cultural backgrounds, age groups, and many more differences. Speech samples collected in-the-wild may also contain various types of environmental noises such as crowd noises at events, traffic noises, and other city noises. This makes them ideally suited for research areas dealing with noise-cancellation or source-separation, e.g., modern speech recognition tasks.

Unfortunately, this mass of data is generally unstructured, lacks reliable labels and high-quality annotation procedures are costly, time-consuming, and tedious work[1–5]. What seems like the answer to these needs of big data has come in the form of a technique called "crowdsourcing"[6–8]. Recently, numerous scientific research projects have turned their backs on only collecting annotations in a controlled laboratory setting by groups of experts and started to recruit annotators by outsourcing the labelling tasks to an open, unspecific, and mostly untrained group of individuals on the web. Therefore, crowdsourcing can be harnessed for lots of different application areas and offers immediate access to a wide and diverse range of individuals with different backgrounds, knowledge, and skills, everywhere and at any time.

Hence, crowdsourcing has emerged as a collaborative approach highly applicable to the area of language and speech processing[4, 9–12], offering a fast and effective way to gather a large amount of labels[4, 13, 14] that are of the same quality as those determined by small groups of experts[4, 10, 15, 16] but at lower costs[1, 4].

In this context, we developed the online crowdsourcing-based gamified data collection, annotation and analysis platform iHEARu-PLAY[1][17–19] and its integrated novel web-based speech classification tool voice analysis application (VoiLA)[2][18] to encourage players to provide large-scale labelled speech data in an efficient manner on a voluntary basis while playing a game and supporting science.

1.1 Related work

For our proposed approach, we consider two categories of related work: 1) the usage of gamified crowdsourcing platforms for tasks related to data collection and 2) present automatic emotion classification systems:

1) Gamified crowdsourcing for data collection

Crowdsourcing individuals can be categorised into two groups[20]: firstly, people who are motivated by money (extrinsic incentives) and secondly people who dedicate their leisure time to something they feel passionate about (intrinsic incentives). The first group will likely be found at some typical crowdsourcing marketplace but the second group are the people who can be recruited for scientific purposes. In this regard, a novel strategy named games with a purpose (GWAP) was introduced, offering a potential solution to motivate individuals to voluntarily help in scientific research projects by providing a positive, gamified environment instead of financial rewards[21, 22].

GWAP have a short history, the first one – called the ESP Game – being designed and introduced in 2006. This GWAP deals with the annotation of images and opened new grounds in human computation[22]. Directly after, the music-related GWAP Tagatune[23] was introduced, focusing on retrieving new music annotations via a multi-player game. Apart from von Ahn's projects, a variety of different game mechanisms have also been introduced. In Peekaboom[24], players are awarded points for identifying objects within an image; Matchin[25] presents players with two photos and awards points when agreeing on which one is more appealing; and City Lights[26] is a music metadata validation approach which makes use of gamification elements to make the validation process more enjoyable. Similarly, Curator[27] was introduced as a class of GWAP where players create collections and are awarded points based on collections that match. In addition to that, the GWAP Wordrobe[28] consists of a large set of multiple-choice questions on word senses, where players need to answer these questions to receive points according to the agreement with fellow players. Moreover, ARTigo[29] attracts people who are interested in art as it supplies artworks with tags while presenting them to two opposing players, giving points for the tags identical to the ones entered by the co-player. The creators of *ARTigo* experimented with handing out financial rewards to the players and showed that only very few players competed in fact for the money[30]. Considerably higher motivational factors include altruism due to the scientific background, or reputation thanks to highscore lists[30].

[1]https://ihearu-play.eu

[2]https://ihearu-play.eu/voila

2) Automatic emotion classification systems

To date, there have been numerous efforts to develop automatic speech (emotion) classification systems[31–34] and potential applications such as service robot interactions[35–37], call-centre monitoring[38], health monitoring[39], smart homes[40, 41], and driver assistance systems[42–44] that benefit from this technology. In this context, a range of different applications for automatic emotion classification from speech have been introduced in the literature, such as the open speech & music interpretation by large-space extraction (openSMILE) toolkit[45], EmoVoice[46], and the web-based interactive speech emotion (WISE) classification system[47]. These frameworks are mostly standalone software packages with a focus on audio recording, file import, feature extraction, and emotion classification.

Unlike openSMILE and EmoVoice, VoiLA can conventionally be run on any PC or smartphone device without setting it up or installing any software. WISE is more similar in that it is also web-based and it does provide automatic emotion classification. However, VoiLA also provides different states and traits like gender, interest, emotion, and arousal/valence, and is directly connected to the web-based crowdsourcing game iHEARu-PLAY[17] for gathering annotations and recordings for new datasets in an efficient manner.

1.2 Contributions of this work

In this contribution, we describe an alternative crowdsourcing method – our online crowdsourcing-based, gamified, intelligent annotation platform iHEARu-PLAY[14, 17] – which motivates people (non-professional annotators) by giving them a playful environment where they can have fun and at them same time voluntarily help scientific research projects by recording and annotating data.

Furthermore, we herein propose our interactive, speech analysis framework VoiLA[18]. Its main aim is to obtain training data and to allow the people who helped annotate data within iHEARu-PLAY, and anyone else, to test and evaluate the trained system. Once a speech recording is uploaded to the server, the system classifies the speaker states and traits arousal, dominance, valence, gender, and 24 different kinds of emotions using a model trained on previously labelled training instances. Players unsatisfied with the classification results are able to submit corrections of their results. The corrected labels are saved and will be used in the future to enhance the emotion classification system. In this regard, VoiLA is directly connected to iHEARu-PLAY, which makes it unique, as – to the best of the authors' knowledge – no other purely web-based emotion analysis tool exists aiming at the crowdsourcing-based collection of annotations and recordings for new datasets.

2 Gamified crowdsourcing-based data collection

Despite their successes, all the GWAP described earlier have been designed with a specific aim and target a single-modal labelling task. Furthermore, these GWAP can only be applied for their developed purpose and due to its method of implementation, they cannot be easily adapted to other labelling or data collection procedures. In this regard, the authors proposed the crowdsourcing platform iHEARu-PLAY[17]. This platform is accessible on any standard PC or smartphone and offers audio, video and image labelling for a diverse range of annotation tasks. It also features audio-visual or just audio data collections and analysis, taking into account a range of novel annotator trustability-based machine learning algorithms to reduce the amount of manual annotation work[14, 15, 48].

In detail, iHEARu-PLAY was realised with the free and open-source Python web framework Django[49], using a free HTML 5 theme which ensured compatibility with all current browsers on standard PCs and mobile devices. Therefore, data recordings and annotations are easily collectable at any time and anywhere as long as audio can be played back and/or a microphone is available.

iHEARu-PLAY's primary intended use is the collection and annotation of audio datasets. It is, however, modality-independent, i.e., images and videos can also be collected and annotated via the platform. iHEARu-PLAY offers a wide range of multi-task annotation options, including discrete (single-choice and multiple-choice), discrete numeric, continuous numeric, continuous numeric 2D, time-continuous numeric, self-assessment manikin, pairwise comparison, and free-input labels.

An overview of the different components in iHEARu-PLAY is given in Fig. 1. Shown are the data collection options, the pre-processing and intelligent audio analysis components, the integrated machine learning component, and the annotator trustability score calculation[14, 15]:

Data collection. New data can be collected either simply with the recording feature within iHEARu-PLAY or by making use of VoiLA, which will be described in more detail in the following sections.

Pre-processing. Recorded speech of each player automatically runs through a pre-processing step, ensuring a good audio quality by applying voice and event activity detection and a volume normalisation of the recordings.

Intelligent audio analysis. For the annotation part, data owners and researchers upload their audio data to iHEARu-PLAY which then automatically runs through

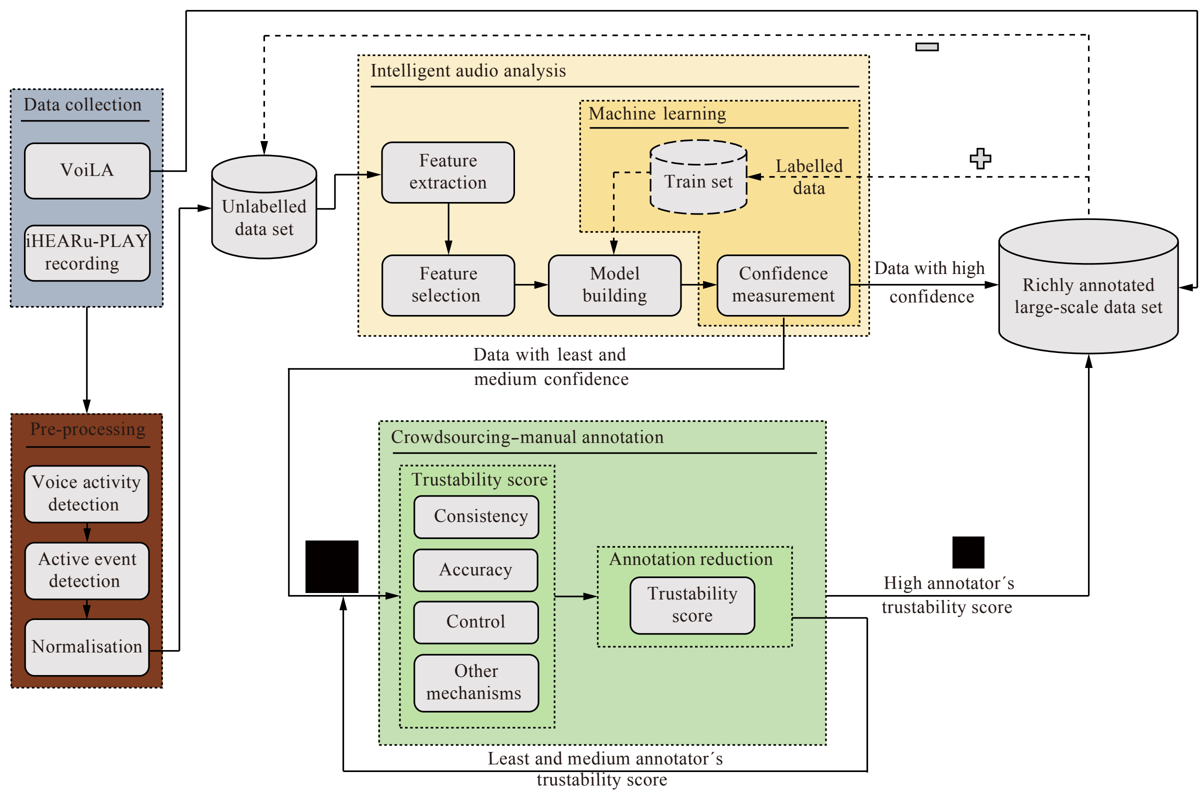

Fig. 1 iHEARu-PLAY's interaction between the intelligent audio analysis, the active learning, and the data quality management components, including the annotator trustability calculation and the annotation reduction components; extended from [14, 15] the intelligent audio analysis (IAA) component. After having chosen one of several available feature sets (e.g., IS09[50] with 384 features or IS16 ComParE with 6 373 features[51], the acoustic features are automatically extracted by using the integrated openSMILE toolkit[45]. Then, a classifier is trained with the small amount of pre-labelled training data on iHEARu-PLAY and the results are transferred to the trustability-based machine learning component.

Machine learning. A range of selectable machine learning algorithms create a sorted list from the highest confidence on the instances to the lowest and extract a subset of instances based on the prediction confidence values[15]. This low and medium confidence subset is then removed from the unlabelled data and automatically passed on for manual labelling while taking the annotator's trustability calculation into account.

Trustability score and annotation reduction. Low quality annotations can result in training the model using incorrectly labelled data which may lead to a reduced accuracy of the trained classifier. Therefore, one of the goals of iHEARu-PLAY is to obtain annotations from non-expert annotators that are qualitatively close to gold standard annotations created by experts. In this context, several data quality mechanisms are applied such as pre-time quality checks and tracking the annotator's behaviour[15].

For more information on the outlined components, the reader is referred to [14, 15]. In summary, iHEARu-PLAY has many advantages over conventional crowd-sourcing platforms and is unique in that it provides volunteers a game-like environment to record and annotate speech[19], i.e., work is presented to players in an interesting and accessible way by incorporating elements that are typically found only in games.

In this context, just as humans differ from each other, player types can greatly differ as well[52]. Not every player reacts or experiences game design elements in the same way, which makes it difficult to anticipate how well certain design decisions will be received by players[53]. People can have different interests and preferences such as enjoying narratives of the game or playing to compete with others, whereas others may find competition or getting points irrelevant; instead, they might enjoy socialising with others or interacting with the world. Therefore, within the literature[54], people are categorised into four different types of players: achievers, explorers, socialisers, and killers (Table 1). It should be noted though, that these player types are theoretical extremes of players and their behaviours. In practice, players mostly have the characteristic of all player types[52]. However, only one or two playing styles and behaviours are predominant.

Therefore, to include as many players as possible, iHEARu-PLAY provides a specially designed gamification concept and collectable points for each annotation or recording handed in by the player depending on different mechanisms[19]. In this context, the platform takes into account the interests of the different player types and utilises a combination of points, leaderboards, badges, and a social platform. These gamification elements are also the most widely used, since if applied right, they are known to be powerful, practical, relevant[55] and potentially able to turn the mundane labelling work into a more enjoyable and motivating task. For more details on iHEARu-PLAY's gamification concept, the reader is referred to the work presented in [19].

3 Voice analysis application

iHEARu-PLAY features a web-based interface with an audio-visual or just audio recording feature, where players are asked to record and upload their speech directly from within their browser. These speech prompts can be individual sentences, a short story, describing an image, or include special tasks, e.g., recording a prompt while whispering or acting out different emotions. Making the mundane performance of speech recording tasks more appealing – besides the inclusion of different gamification elements[19] – VoiLA encourages players to take an active role in improving the system by providing their (labelled) speech data[18].

3.1 Architecture

VoiLA comprises a unique and novel mix of different components. The relationship and information flow between the components is illustrated in Fig. 2. Speech recorded by players on VoiLA is first uploaded to the VoiLA server and then forwarded to classification. Internally, the system employs openSMILE[45] to extract an extended and enhanced version of the GeMAPS feature set[56]. A manually selected subset of these features is

Table 1 General overview of four different possible player types and their typical characteristics[54]. Note that these types are valid in general and not specifically within iHEARu-PLAY

Player type	Characteristics
Achievers	Enjoy completing challenges → Collecting achievements, points or items and levelling up
Explorers	Have a thirst for adventure → Exploring the game and strive to discover new features
Socialisers	Like to stay in the social circle → Socialising or interacting with other players
Killers	Look for open battles Using the virtual construct to cause distress on other players

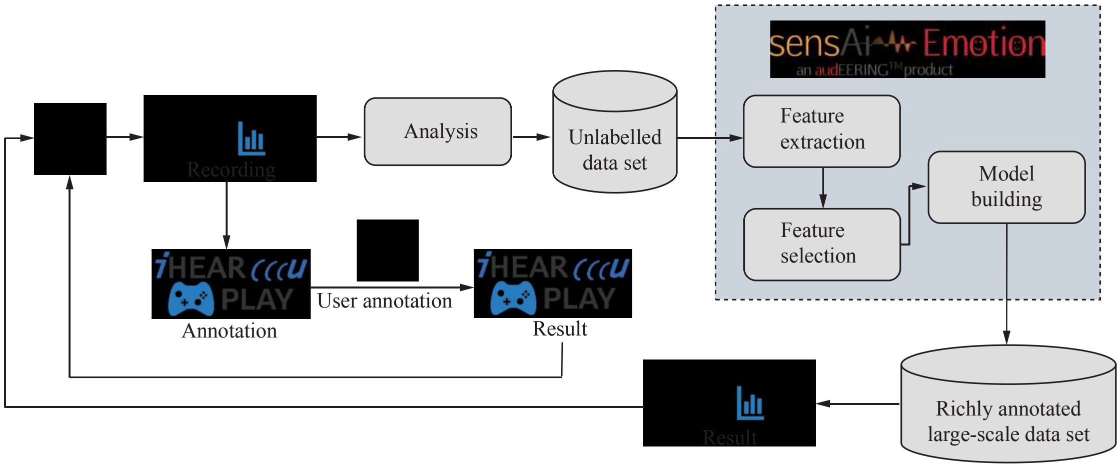

Fig. 2 Schematic overview of the integration of the different VoiLA components into the iHEARu-PLAY crowdsourcing platform for annotation, training and classification of speech

then used in hand-crafted linear and non-linear regression models for emotion prediction. Upon completion of the analysis, the results are sent back to the VoiLA server for generation of the report page.

Players are provided with the unique ability to correct the system-proposed labels and to suggest an alternative label if they do not agree with the automatic analysis result. This new label will be used later to adapt the models and improve the accuracy and robustness of the system over time (Fig. 1). In this way, the integration within iHEARu-PLAY will serve as a scalable way to improve the accuracy by retraining the classifier on more training data and adding new classification capabilities in the future.

All communication with the servers is handled via encrypted HTTPS connections. Through a clearly defined web API interface which only permits the uploading of audio and the retrieval of results, it is ensured that uploaded audio cannot be accessed by any means from the outside. Our web API internally uses queuing mechanisms and auto-scaling to adaptively respond to changes in load. Therefore, VoiLA is able to scale to hundreds of concurrent players.

3.2 Player interface

VoiLA's brower-based interface has been deliberately kept minimalistic in order to make usage as simple as possible, while still providing players with detailed information about their voice. Through the single click of a button, visitors initiate the recording process (Fig. 3). To generate voice content, players are encouraged to accomplish one of currently three kinds of tasks: reading a text, acting emotions/situations or describing an image/story. However, since the linguistic content of the sample is not leveraged in the classification system, players are also free to improvise.

Through an integrated voice activity detection component, the recording stops automatically when the player stops speaking. Alternatively, the player may end the recording by another click on the recording button. At this point, their recorded voice sample is uploaded to the server where it is analysed. After the analysis has finished, results are retrieved from the server and presented to the player (Fig. 4). If desired, players can correct the

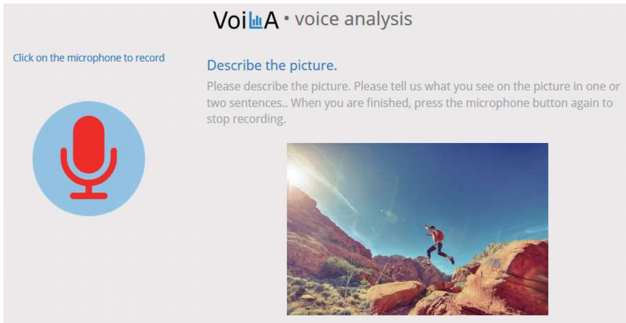

Fig. 3 Exemplary recording page of VoiLA as it is shown to players while recording their voice. The red microphone shows the player that the recording has started. After having described the presented picture in own words, the speech sample will be sent to the server for analysis. Color versions of the figures in this paper are available online

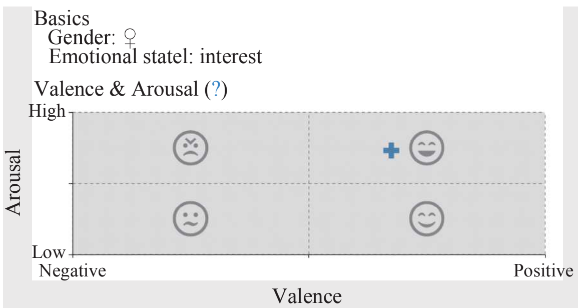

Fig. 4 Top part of VoiLA's results page as it is shown to players after the analysis of their uploaded voice sample is finished. From top to bottom: basics as gender and dominant emotional state, followed by the chart plotting the mood of the player in the 2D arousal/valence space

results and thereby provide labelled audio data, which can be used as new training data to improve the classifier later on (Fig. 1). This makes VoiLA unique, as it is – to the best of the authors knowledge – the only tool allowing players to alter analysis results and thereby improve the classification process in-line. A player can repeat the analysis as many times as they desire. Registered players have the additional option to review the results of their previous recordings.

As the system is still evolving, the analysis is currently restricted to five aspects: arousal, valence, dominance, emotion classes and gender. Arousal and valence are reported both as continuous values between –1 and 1 and plotted in a scatter chart, while dominance is given as a percentage. The emotions are also represented by percentage values for 24 categories: affection, anger, boredom, contentment, depression, disgust, enthusiasm, excitement, fear, frustration, happiness, interest, irritation, joy, nervousness, panic, passion, pride, relaxation, sadness, satisfaction, stress, tension, and worry. Even more categories are planned to be added in the future, including emotional states such as admiration, amusement, confusion, disappointment, impressed, loving, serenity, and surprise.

4 Evaluation

An evaluation study was conducted to assess the effectiveness of the current system, to determine what could be improved, and to identify the needs and wishes of the players for new features. We evaluated the iHEARu-PLAY platform and VoiLA to answer the following questions:

- How do players feel about the design and content of the platform?
- What is the usability of the current prototype? What are possible usability improvements?
- How interesting are the different recording tasks?
- How well are the current features accepted?
- What would players like to see added to the platform experience?
- What do players dislike about the platform and how can these issues be improved?

To ensure that all data necessary to answer these questions could be collected, an evaluation survey[3] was tailored specifically to iHEARu-PLAY. In addition, the system usability scale (SUS) by Brooke[57] was included to evaluate the usability of the platform in a comparable manner.

Over the course of two months, 157 players participated in the online survey describing their iHEARu-PLAY experience. Out of the 157 overall annotators, 131 gave us their complete metadata (Table 2). Among these participants were 72 male and 59 female volunteers. Altogether, we reached a variety of ages from 18 to 57 years (mean: 31.3, standard-deviation: 9.6). A large majority of participants were students (78.2%), followed by people employed for wages (18.7%) and self-employed participants (3.1%). Many players had a high school degree or equivalent (51.7%) as their highest academic degree, followed by a bachelor's degree (18.3%), a master's degree (12.3%) and other qualifications (8.4%). Only a few people went to a college without degree (3.4%) or had no qualification (5.9%).

Concerning the usability of iHEARu-PLAY, evaluation of the collected data shows that the platform reaches a 87.9% SUS usability-score (Table 3). According to Bagor[58] who divided this scale into categories, this suggests that iHEARu-PLAY has an excellent, bordering

[3]The questionnaire was created and hosted on the online-platform
SoSci Survey (https://www.soscisurvey.de).

Table 2 Statistics of the participants of the user evaluation
Note: Out of 157 overall annotators, only 131 gave their complete metadata

Gender
72 female / 59 male
Age
Range = 18–57 years old; Medium = 31.3 years old; Standard deviations (SD) = 9.6 years old
Occupation
Students (78.2%), employed for wages (18.7%), self-employed (3.1%)
Education
High school degree or equivalent (51.7%), bachelor's degree (18.3%), master's degree (12.3%), other qualifications (8.4%), college without degree (3.4%), no qualification (5.9%)

Table 3 Results of the evaluation survey. Overall results are displayed as star ratings (intervals incrementing in 20% steps), followed by absolute numbers (0–100%) and standard deviations

Topic	Rating	%	SD
General			
Content		86.6	1.8
Design		83.5	2.3
Usability		87.9	1.6
Fun		68.1	2.9
Interest		79.8	1.7
Tasks			
Annotation		81.5	2.1
Recording			
Acting		84.6	1.4
Image		85.2	1.6
Text		76.1	2.3
Results			
Acceptance		72.4	1.9
Alteration		68.3	1.4
Presentation		73.6	0.9

on best imaginable, usability. iHEARu-PLAY's content (86.6%) and design (83.5%) was rated similar positive, followed by rating the platform as interesting (79.8%) and fun to use (68.1%). Independent of this obtained results, we are aware that there is still space for further improvement like optimising the usability of our mobile version.

The acceptance rates of our annotation and recording features with its different tasks were measured individually and answered on a five-point Likert scale. While the annotation feature achieved an acceptance rate of 81.5%, the recording feature image task has the highest acceptance rate (85.2%), followed by the game/acting task (84.6%) and the text task (76.1%). This leads to the conclusion that players generally prefer visual or interactive tasks over the less demanding text task.

The analysed results of VoiLA show an acceptance rate of 72.4%, while its presentation was rated with 73.6% and the alteration with 68.3% by the players. The obtained evaluation results are summarised in Fig. 5.

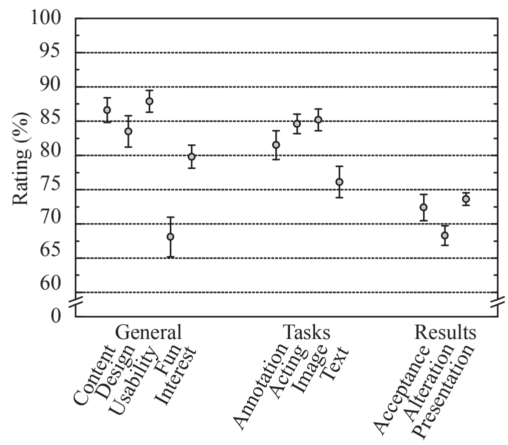

Fig. 5 Results of the evaluation survey on iHEARu-PLAY and VoiLA

To gather insights on the opinions of players, and to receive more detailed feedback and feature requests, we encouraged participants of our survey to submit free-text comments where they could explain the choices they made in the survey and could request features or emphasise positive aspects. Among other things, participants mostly reported on the VoiLA feature. A representative example was a blurriness of the classifier near the edges of emotion classes, i.e., incorrectly classified emotions close together – e.g., irritation and anger. This issue will be addressed in a future release of VoiLA, where we will publish an improved classification system based on the label corrections that players can already perform today. Another common player request was the introduction of more diverse recording tasks, allowing the recording procedure to be even more interesting and fun. This feature is already under development and is implemented by introducing an additional task where players are able to play a small game while recording their speech.

Overall, the system predominantly received positive feedback, stating that iHEARu-PLAY and VoiLA were easy to use and that it was interesting to see an automatic analysis demonstrated on the own voice. Additionally, the analysis increased interest in the science behind voice analysis and the willingness to participate in improving the system and therefore performing annotation tasks. This collected feedback allows the conclusion that iHEARu-PLAY is broadly accepted among players.

5 Conclusions and outlook

Within this contribution, the browser-based crowdsourcing platform iHEARu-PLAY and its web-based speech classification tool VoiLA were introduced, with VoiLA following a unique approach by leveraging iHEARu-PLAY for speech annotation to obtain required training data.

In detail, VoiLA encourages people who helped annotate data – and anyone else – to try and evaluate the trained system by having their own voice analysed. It allows visitors to record and upload their voice directly from a website in their browser. On the backend, the uploaded speech data is run through a classification pipeline using a set of pre-trained models that target different kinds of speaker states and traits like gender, dominance, 24 kinds of emotions, arousal, and valence. The gathered analysis results are then sent back to the player and visualized in the browser, giving players unique objective insights into how their voice sounds.

Finally, an extensive player assessment and evaluation of the first-of-its-kind proposed platform and the introduced methods was performed. The player evaluation survey showed that the proposed system has an excellent, bordering on best imaginable, usability and the task system proposed for voice recording is accepted well by the players. Additional player comments indicated that some enhancements could be made in terms of accuracy of the emotion classification.

In the future, it is planned to integrate the concept of transfer learning, allowing for the adaptation of existing models to an unseen topic. Hence, the goal is to maximise the knowledge transfer from an existing task and obtain new knowledge relevant to a new task. This adaptive learning strategy could also allow for continuous improvement of the models of VoiLA. In addition, we will further improve the classifiers by retraining them with already collected and annotated player data within VoiLA. In this context, a future idea is to give players the possibility to train their own classifier which in turn would help to improve the overall system. From a player's point of view, the performed recording and annotation tasks are handled in a gamified way and could be seen as a way of feeding their own "tamagotchi"™ (i.e., the classifier), which can only grow with good care by performing annotation or recording tasks on a daily basis.

Other potential additions to VoiLA include giving players the possibility to have their voice analysed not only by machine learning but by human annotators, as well. We see our platform iHEARu-PLAY as an ideal platform to collect these manual labels and plan a tighter integration with VoiLA. Additionally, we are currently integrating the player feedback from the conducted evaluation survey.

Finally, a long-term goal is to develop and integrate a classifier which is capable of presenting the results to the player in real-time while they are speaking. Therefore, VoiLA has the potential to popularise the science behind voice analysis and the annotation process of iHEARu-PLAY.

Acknowledgements

This work was supported by the European Community's Seventh Framework Programme (No. 338164) (ERC Starting Grant iHEARu). We thank audEERING for providing sensAI and all iHEARu-PLAY players for taking part in our evaluation.

References

[1] V. Ambati, S. Vogel, J. Carbonell. Active learning and crowd-sourcing for machine translation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Association for Computational Linguistics, Valletta, Malta, 2010.

[2] V. C. Raykar, S. P. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, L. Moy. Learning from crowds. *Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.

[3] A. Kittur, E. H. Chi, B. Suh. Crowdsourcing for usability: Using micro-task markets for rapid, remote, and low-cost user measurements. In *Proceedings of ACM Conference on Human Factors in Computing Systems*, ACM, Florence, Italy, pp. 1–4, 2008.

[4] A. Tarasov, S. J. Delany, C. Cullen. Using crowdsourcing for labelling emotional speech assets. In *Proceedings of W3C Workshop on Emotion Markup Language*, Telecom ParisTech, Paris, France, 2010. DOI: 10.21427/D7RS4G.

[5] B. Settles. Active Learning Literature Survey, Computer Sciences Technical Report 1648, University of Wisconsin-Madison, Madison, USA, 2009.

[6] J. Howe. The rise of crowdsourcing. *Wired Magazine*, vol. 14, no. 6, pp. 1–4, 2006.

[7] M. Eskénazi, G. A. Levow, H. Meng, G. Parent, D. Suendermann. *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*, Chichester, UK: Wiley, 2013.

[8] X. J. Niu, S. F. Qin, J. Vines, R. Wong, H. Lu. Key crowdsourcing technologies for product design and development. *International Journal of Automation and Computing*, vol. 16, no. 1, pp. 1–15, 2019. DOI: 10.1007/s11633-018-1138-7.

[9] A. Burmania, S. Parthasarathy, C. Busso. Increasing the reliability of crowdsourcing evaluations using online quality assessment. *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, 2016. DOI: 10.1109/TAFFC.2015.2493525.

[10] S. Hantke, E. Marchi, B. Schuller. Introducing the weighted trustability evaluator for crowdsourcing exemplified by speaker likability classification. In *Proceedings of the 10th Language Resources and Evaluation Conference*, Association for Computational Linguistics, Portorož, Slovenia, pp. 2156–2161, 2016.

[11] O. F. Zaidan, C. Callison-Burch. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, ACM, Portland, USA, pp. 1220–1229, 2011.

[12] R. R. Morris, D. McDuff. Crowdsourcing techniques for affective computing. *The Oxford Handbook of Affective Computing*, R. A. Calvo, S. D'Mello, J. Gratch, A. Kappas, Eds., Oxford, UK: Oxford University Press, pp. 384–394, 2015.

[13] P. Y. Hsueh, P. Melville, V. Sindhwani. Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, ACM, Boulder, USA, pp. 27–35, 2009.

[14] S. Hantke, Z. X. Zhang, B. Schuller. Towards intelligent crowdsourcing for audio data annotation: Integrating active learning in the real world. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, ISCA, Stockholm, Sweden, pp. 3951–3955, 2017.

[15] S. Hantke, A. Abstreiter, N. Cummins, B. Schuller. Trustability-based dynamic active learning for crowdsourced labelling of emotional audio data. *IEEE Access*, vol. 6, pp. 42142–42155, 2018. DOI: 10.1109/ACCESS.2018.2858931.

[16] R. Snow, B. O'Connor, D. Jurafsky, A. Y. Ng. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACM, Honolulu, USA, pp. 254–263, 2008.

[17] S. Hantke, F. Eyben, T. Appel, B. Schuller. iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing. In *Proceedings of International Conference on Affective Computing and Intelligent Interaction*, IEEE, Xi'an, China, pp. 891–897, 2015. DOI: 10.1109/ACII.2015.7344680.

[18] S. Hantke, T. Olenyi, C. Hausner, B. Schuller. VoiLA: An online intelligent speech analysis and collection platform. In *Proceedings of the 1st Asian Conference on Affective Computing and Intelligent Interaction*, IEEE, Beijing, China, pp. 1–5, 2018. DOI: 10.1109/ACIIAsia.2018.8470383.

[19] S. Hantke, T. Appel, B. Schuller. The inclusion of gamification solutions to enhance user enjoyment on crowdsourcing platforms. In *Proceedings of the 1st Asian Conference on Affective Computing and Intelligent Interaction*, IEEE, Beijing, China, pp. 1–6, 2018. DOI: 10.1109/ACIIAsia.2018.8470330.

[20] J. Howe. *Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business*, New York, USA: Crown Business, 2009.

[21] B. M. Good, A. I. Su. Games with a scientific purpose. *Genome Biology*, vol. 12, no. 12, pp. 135, 2011. DOI: 10.1186/gb-2011-12-12-135.

[22] L. von Ahn. Games with a purpose. *Computer*, vol. 39, no. 6, pp. 92–94, 2006. DOI: 10.1109/MC.2006.196.

[23] E. L. Law, L. von Ahn, R. B. Dannenberg, M. Crawford. TagATune: A game for music and sound annotation. In *Proceedings of International Conference on Music Inform-*

ation Retrieval, Vienna, Austria, pp. 361–364, 2007.

[24] L. von Ahn, R. R. Liu, M. Blum. Peekaboom: A game for locating objects in images. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, ACM, Montréal, Canada, pp. 55–64, 2006. DOI: 10.1145/1124772.1124782.

[25] S. Hacker, L. von Ahn. Matchin: Eliciting user preferences with an online game. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, ACM, Boston, USA, pp. 1207–1216, 2009. DOI: 10.1145/1518701.1518882.

[26] P. Dulačka, J. Šimko, M. Bieliková. Validation of music metadata via game with a purpose. In *Proceedings of the 8th International Conference on Semantic Systems*, ACM, Graz, Austria, pp. 177–180, 2012. DOI: 10.1145/2362499.2362526.

[27] G. Walsh, J. Golbeck. Curator: A game with a purpose for collection recommendation. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, ACM, Atlanta, USA, pp. 2079–2082, 2010. DOI: 10.1145/1753326.1753643.

[28] N. J. Venhuizen, V. Basile, K. Evang, J. Bos. Gamification for word sense labeling. In *Proceedings of the 10th International Conference on Computational Semantics*, Association for Computational Linguistics, Potsdam, Germany, pp. 397–403, 2013.

[29] C. Wieser, F. Bry, A. Berárd, R. Lagrange. ARTigo: Building an artwork search engine with games and higher-order latent semantic analysis. In *Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing*, AAAI, Palm Springs, USA, pp. 15–20, 2013.

[30] C. Wieser. Building a Semantic Search Engine with Games and Crowdsourcing, Ph. D. dissertation, Ludwig-Maximilians-Universität, München, Germany, 2014.

[31] V. Sethu, E. Ambikairajah, J. Epps. Empirical mode decomposition based weighted frequency feature for speech-based emotion classification. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, IEEE, Las Vegas, USA, pp. 5017–5020, 2008. DOI: 10.1109/ICASSP.2008.4518785.

[32] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, B. Schuller. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Prague, Czech Republic, pp. 5688–5691, 2011. DOI: 10.1109/ICASSP.2011.5947651.

[33] C. Busso, S. Lee, S. Narayanan. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 582–596, 2009. DOI: 10.1109/TASL.2008.2009578.

[34] D. Bitouk, R. Verma, A. Nenkova. Class-level spectral features for emotion recognition. *Speech Communication*, vol. 52, no. 7–8, pp. 613–625, 2010. DOI: 10.1016/j.specom.2010.02.010.

[35] J. S. Park, J. H. Kim, Y. H. Oh. Feature vector classification based speech emotion recognition for service robots. *IEEE Transactions on Consumer Electronics*, vol. 55, no. 3, pp. 1590–1596, 2009. DOI: 10.1109/TCE.2009.5278031.

[36] P. Rani, C. C. Liu, N. Sarkar, E. Vanman. An empirical study of machine learning techniques for affect recognition in human-robot interaction. *Pattern Analysis and Applications*, vol. 9, no. 1, pp. 58–69, 2006. DOI: 10.1007/s10044-006-0025-y.

[37] T. M. Wang, Y. Tao, H. Liu. Current researches and future development trend of intelligent robot: A review. *International Journal of Automation and Computing*, vol. 15, no. 5, pp. 525–546, 2018. DOI: 10.1007/s11633-018-1115-1.

[38] L. Vidrascu, L. Devillers. Detection of real-life emotions in call centers. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, ISCA, Lisbon, Portugal, pp. 1841–1844, 2005.

[39] Z. J. Yao, J. Bi, Y. X. Chen. Applying deep learning to individual and community health monitoring data: A survey. *International Journal of Automation and Computing*, vol. 15, no. 6, pp. 643–655, 2018. DOI: 10.1007/s11633-018-1136-9.

[40] B. Lecouteux, M. Vacher, F. Portet. Distant speech recognition in a smart home: Comparison of several multisource ASRs in realistic conditions. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, ISCA, Florence, Italy, pp. 2273–2276, 2011.

[41] A. Fleury, N. Noury, M. Vacher, H. Glasson, J. F. Seri. Sound and speech detection and classification in a health smart home. In *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Vancouver, Canada, pp. 4644–4647, 2008. DOI: 10.1109/IEMBS.2008.4650248.

[42] F. Eyben, M. Wöllmer, T. Poitschke, B. Schuller, C. Blaschke, B. Färber, N. Nguyen-Thien. Emotion on the road—necessity, acceptance, and feasibility of affective computing in the car. *Advances in Human-computer Interaction*, vol. 2010, Article number 263593, 2010. DOI: 10.1155/2010/263593.

[43] A. Tawari, M. Trivedi. Speech based emotion classification framework for driver assistance system. In *Proceedings of IEEE Intelligent Vehicles Symposium*, San Diego, USA, pp. 174–178, 2010. DOI: 10.1109/IVS.2010.5547956.

[44] C. M. Jones, I. M. Jonsson. Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses. In *Proceedings of the 17th Australia Conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future*, ACM, Canberra, Australia, pp. 1–10, 2005.

[45] F. Eyben, F. Weninger, F. Gross, B. Schuller. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, ACM, Barcelona, Spain, pp. 835–838, 2013. DOI: 10.1145/2502081.2502224.

[46] T. Vogt, E. André, N. Bee. EmoVoice—a framework for online recognition of emotions from voice. In *Proceedings of the 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-based Systems*, Springer, Kloster Irsee, Germany, pp. 188–199, 2008. DOI: 10.1007/978-3-540-69369-7˙21.

[47] S. E. Eskimez, M. Sturge-Apple, Z. Y. Duan, W. Heinzelman. WISE: Web-based interactive speech emotion classification. In *Proceedings of the 4th Workshop on Sentiment Analysis Where AI Meets Psychology*, IJCAI, New York, USA, pp. 2–7, 2016.

[48] S. Hantke, C. Stemp, B. Schuller. Annotator trustability-based cooperative learning solutions for intelligent audio analysis. In *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, ISCA, Hyderabad, India, 2018. DOI: 10.21437/Interspeech.2018-1019.

[49] Django. Computer Software, [Online], Available: https://djangoproject.com, 2019.

[50] B. Schuller, S. Steidl, A. Batliner. The interspeech 2009 emotion challenge. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association*, ISCA, Brighton, UK, pp. 312–315, 2009.

[51] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Bur-

goon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini. The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, ISCA, San Francisco, USA, pp. 2001–2005, 2016. DOI: 10.21437/Interspeech.2016-129.

[52] R. Bartle. Hearts, clubs, diamonds, spades: Players who suit MUDs. *Journal of MUD Research*, vol. 1, no. 1, pp. 19, 1996.

[53] M. Meder, B. J. Jain. *The Gamification Design Problem*, [Online], Available: https://arxiv.org/abs/1407.0843, 2014.

[54] Y. W. Xu. Literature Review on Web Application Gamification and Analytics, CSDL Technical Report 11–05, University of Hawaii, Hawaii, USA, 2011.

[55] K. Werbach, D. Hunter. *For the Win: How Game Thinking can Revolutionize Your Business*, Philadelphia, USA: Wharton Digital Press, 2012.

[56] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, K. P. Truong. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016. DOI: 10.1109/TAFFC.2015.2457417.

[57] J. Brooke. SUS-A quick and dirty usability scale. *Usability Evaluation in Industry*, P. W. Jordan, B. Thomas, B. Weerdmeester, I. L. McClelland, Eds., London, UK: CRC Press, pp. 4–7, 1996.

[58] A. Bangor, P. Kortum, J. Miller. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, vol. 4, no. 3, pp. 114–123, 2009.

pervised by Simone Hantke, he explored different approaches to vocal emotion analysis based on feature mapping and he developed the initial version of VoiLA. In addition, he integrated the emotion analysis capabilities of audEERING's sensAI into the newly-developed tool.

E-mail: tobias.olenyi@tum.de