VoiLA: An Online Intelligent Speech Analysis and Collection Platform

Simone Hantke[1,2]
[1]Machine Intelligence & Signal Processing Group,
Technische Universität München, Germany
Email: simone.hantke@tum.de

Tobias Olenyi[2]
[2]ZD.B Chair of Embedded Intelligence
for Health Care and Wellbeing,
University of Augsburg, Germany

Christoph Hausner[3]
[3]audEERING GmbH,
Gilching, Germany

Björn Schuller[2,4]
[4]GLAM – Group on Language, Audio, and Music,
Imperial College London, UK

***Abstract*—In this study, we propose the novel *Voi*ce Ana*L*ysis *A*pplication "VoiLA", a free web-based speech classification tool designed to educate users about state-of-the-art speech analysis paradigms. Further, the platform encourages users to take an active role in improving the service by providing labelled speech data via an associated web interface. VoiLA allows users to record and upload voice samples directly from their browser. The speech data is then analysed in a state-of-the-art classification pipeline, using a set of pre-trained models which target a range of speaker states and traits such as gender, valence, arousal, dominance, and 24 different discrete emotions. The analysis results are then visualised in the browser, giving users a unique insight into how their voice sounds. We assess the effectiveness of VoiLA via a series of user evaluations which indicate that it is fun and easy to use and routinely provides accurate and informative voice analysis.**

***Index Terms*—human computation; speech analysis; crowdsourcing; survey.**

I. Introduction

There have been numerous efforts to develop automatic speech (emotion) classification systems [1]–[4] and potential applications such as service robot interactions [5], [6], call-centre monitoring [7], smart homes [8], [9], and driver assistance systems [10]–[12] that benefit from this technology. The success of these supervised machine learning techniques highly depends on the amount and quality of labelled training data required to create robust classification models. Many (emotional speech) datasets are small with respect to the number of subjects who participated in the recording, often resulting in poor generalisations of systems to new speakers. Current state-of-the-art technologies allow for the gathering of vast amounts of speech data from the web [13], [14], yet analysing this content is challenging. This fact, among others, prevents many interesting large-scale investigations.

Recently, crowdsourcing has emerged as a collaborative approach highly applicable to the area of language and speech processing. Crowdsourcing offers a fast and effective way to gather a large amount of labels [15]–[17] that are of the same quality as those determined by small groups of experts [15], [18], [19] but at lower costs [15], [20]. Our online crowdsourcing-based, gamified annotation platform iHEARu-PLAY [1] [21], [22] and its integrated novel feature "VoiLA"[2], the web-based speech classification tool proposed herein, encourage users to provide labelled speech data on a voluntary basis while playing a game and supporting science.

A. Related Work

A range of different applications for automatic emotion classification from speech have been introduced in the relevant literature, such as the open-Source Media Interpretation by Large feature-space Extraction (openSMILE) toolkit [23], EmoVoice [24], and the Web-based Interactive Speech Emotion (WISE) classification system [25]. These frameworks are standalone software packages with a focus on audio recording, audio file import, feature extraction, and emotion classification.

openSMILE [23], is a cross-platform classification toolkit that includes libraries for feature extraction. Pre-trained models and scripts to train custom models are available from the related openEAR toolkit [26]. openSMILE supports real-time processing and is able to extract more than 500k features. Other modules exist to allow external classifiers and libraries such as LibSVM [27] to be integrated and used for classification. openSMILE also supports a variety of data formats from popular machine learning frameworks such as CURRENNT [28], the Hidden Markov Model Toolkit (HTK) [29], WEKA [30], and scikit-learn for Python [31].

EmoVoice [24] allows the user to create a personal speech-based emotion recogniser and can track the affective state of the user in real-time. Each user records their own speech corpus to train the system which can then be used for real-time emotion classification for the same user. EmoVoice has been employed in several practical applications including robot-human and virtual agent-human interactions.

WISE [25] features a web-based interface that allows users to upload speech data and automatically classify emotion using pre-trained models. Users can suggest alternative labels if the

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

978-1-5386-5311-1/18/$31.00 ©2018 IEEE

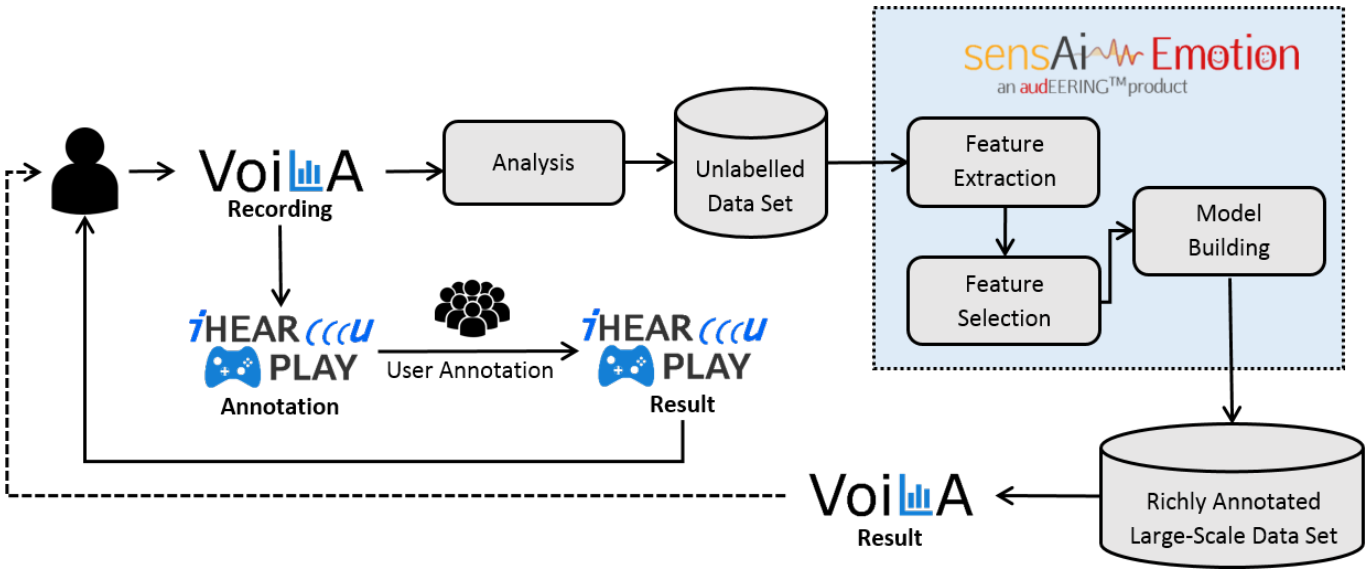

Fig. 1. Schematic overview of the integration of the different components into a common framework for annotation, training and classification of speech.

result of the classifier is deemed unfitting. Updated labels are then fed back into the system to retrain the models.

Unlike openSMILE and EmoVoice, VoiLA can be used conveniently without installing any software. WISE is similar in that it is also web-based and provides automatic emotion classification. VoiLA, however, 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 [21] for gathering annotations and recordings for new datasets.

B. Contributions of this Paper

We herein propose the interactive speech analysis framework VoiLA which adopts a unique approach by leveraging the public crowdsourcing annotation platform iHEARu-PLAY. The aim is to obtain training data and to allow the people who helped annotate the data, and anyone else, to test and evaluate the trained system. VoiLA features a web-based interface that allows users to record and upload their speech directly from within their browser. Once a speech sample is uploaded, the system classifies the speaker states and traits arousal, dominance, valence, gender, and 24 different kinds of emotions using a model derived from previously labelled training samples.

II. Architecture and User Interface

VoiLA comprises a unique and novel mix of different components. The relationship and information flow between the components are illustrated in Figure 1. Speech recorded by users on VoiLA is first uploaded to the VoiLA server and then forwarded to classification. Internally, the system employs openSMILE to extract an extended and enhanced version of the GeMAPS feature set [32]. A manually selected subset of the features is then used in hand-crafted linear and non-linear regression models for emotion prediction. After the analysis is finished, the results are sent back to the VoiLA server for generation of the report page.

Users 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. 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.

A. The Crowdsourcing Platform

Training data designed to further improve the classifier behind VoiLA is obtained through iHEARu-PLAY, our crowdsourcing platform specialized on acquisition of labelled audio data [21]. The platform is unique in that it provides volunteers a game-like environment to record and annotate speech [22], i.e., work is presented to players in an interesting and accessible way by incorporating elements that are typically found only in games [22].

B. VoiLA

VoiLA's brower-based interface has been deliberately kept minimalistic in order to make usage as simple as possible, while still providing users with detailed information about their voice. Through the single click of a button, visitors initiate the recording process (cf. Figure 2). To generate voice content, users are encouraged to accomplish one of currently three kinds of tasks – a text task, where the user is prompted to read out a text loud, an image task where the user is asked to describe an image, or a game task where the user is allowed to act out emotions. However, since the linguistic content of the

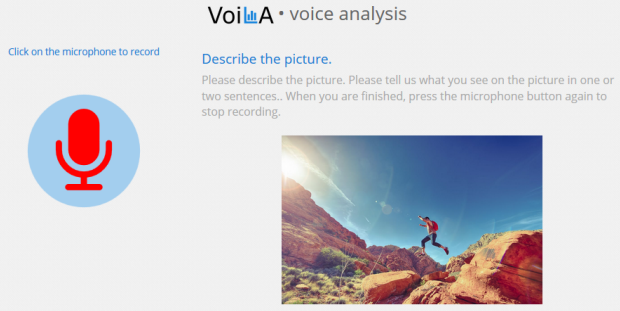

Fig. 2. The recording page of VoiLA as it is shown to users while recording their voice. The red microphone shows the user that recording has started. After having described the picture in own words, the speech sample will be sent to the server for analysis.

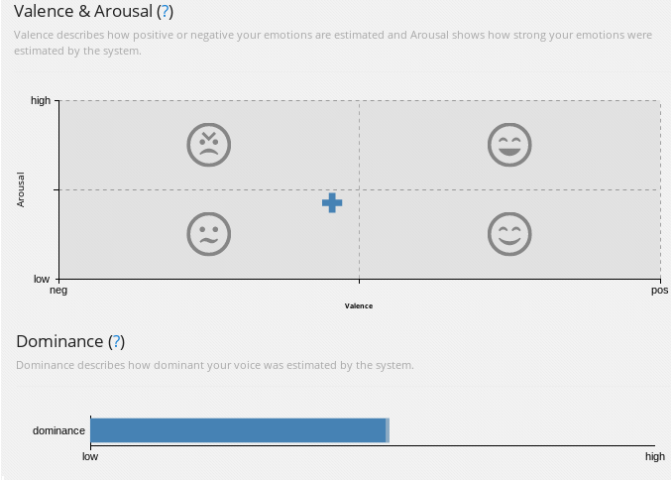

Fig. 3. Top part of VoiLA's results page as it is shown to users after the analysis of their uploaded voice sample is finished. From top to bottom: chart plotting the mood of the user in the 2D arousal/valence space and the bar chart indicating the dominance of the voice.

utterance is not leveraged in the sensAI classification system, users are free to improvise.

Through an integrated Voice Activity Detection component, the recording stops automatically when the user stops speaking. Alternatively, the user 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 user (cf. Figure 3). If desired, users can correct the results and thereby provide labelled audio data, which can be used to improve the classifier later on. This makes VoiLA unique, as it is – to the best of the authors knowledge – the only tool allowing users to alter analysis results and thereby improve the classification process in-line. A user can repeat the analysis as many times as they desire. Registered users have the additional option to review the results of their previous recordings. More information on how the system works and how users can contribute to science by annotating data through iHEARu-PLAY can also be found on the website [3].

As the system is still evolving, the analysis is currently restricted to five aspects: arousal, dominance, emotions, gender, and valence. Arousal and valence are modelled as continuous values between -1 and 1, while dominance is given as percentage. The emotions are 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. However, 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.

III. User 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 users for new features. We evaluated the prototype to answer the following questions:

- What is the usability of the current prototype? What are possible usability improvements?
- How well are the current features accepted?
- What would users like to see added to the analysis experience?
- What do players dislike about it and how can these issues be improved?

To ensure that all data necessary to answer these questions could be collected, the evaluation survey was tailored specifically to VoiLA. In addition, the System Usability Scale (SUS) by Brooke [33] was included to evaluate the usability of VoiLA in a comparable manner. The questionnaire was created and hosted on the online-platform SoSci Survey[4].

Over the course of 22 days, 26 users had their voice analysed on VoiLA and afterwards participated in our online survey, describing their experience. Among these participants were 16 male and 10 female volunteers. Altogether, we reached a variety of ages (cf. Figure 4), from 16 to 53 years (mean: 23.6, standard-deviation: 8.69). A large majority of participants were students (88.5 %), followed by people employed for wages (7.7 %) and self-employed participants (3.8 %). Many users had a high school degree or an equivalent (53.8 %) as their highest academic degree, followed by a bachelor's degree (23.1 %), a master's degree (7.7 %) and other qualifications (7.7 %). Only few people went to a college without degree (3.8 %) or had no qualification (3.8 %).

Concerning the usability of VoiLA, evaluation of the collected data shows that VoiLA reaches a 78.13 % SUS usability-score (cf. Table 1). According to Bagor et al. [34] who divided this scale into categories, this indicates that VoiLA has a good, bordering on excellent, usability. As VoiLA is still

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

[4]https://www.soscisurvey.de

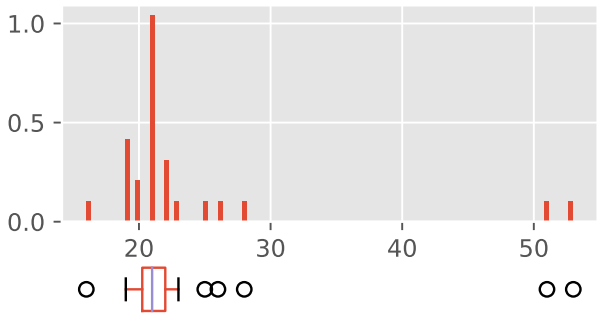

Fig. 4. Age distribution of users participating in the evaluation survey. The x-axis displays the age in years, the y-axis the score of participation.

evolving, we are positive about reaching excellent usability in a future release. The acceptance rate of different tasks was measured individually and answered on a five-point likert scale. From these three questions, an acceptance rate of the task was calculated (cf. Table 1). The image task has the highest acceptance rate (74.58 %), followed by the game task (73.33 %) and the text task (67.56 %). This leads to the conclusion that users generally prefer visual or interactive tasks over the less demanding text task.

To gather insights on the opinion that users have on VoiLA, 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 emphasize positive aspects of the system. Among other things, participants reported a blurriness of the classifier near the edges of emotion classes, i.e. wrongly 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 users can already perform. Another common user request was the introduction of a delay before the recording, allowing users to read and think before the recording starts. This feature is already under development and is implemented by introducing an additional page where users are able to familiarize themselves with the given task before starting the recording.

Overall, the system predominantly received positive feedback, stating that VoiLA is easy to use and it is interesting to see an automatic analysis demonstrated on the own voice. Additionally, the analysis increased the interest in the science behind voice analysis and the willingness to participate in improving the system. This collected feedback allows the conclusion, that VoiLA is broadly accepted among users.

TABLE I
RESULTS OF THE EVALUATION SURVEY. OVERALL RESULTS ARE DISPLAYED AS STAR RATINGS (INTERVALS INCREMENTING IN 20% STEPS), FOLLOWED BY ABSOLUTE NUMBERS.

	Rating	%
Tasks		
Acting	★★★★★	74.58
Image	★★★★★	73.33
Text	★★★★½	67.56
Results		
Acceptance	★★★½☆	61.67
Alteration	★★★★☆	67.78
Presentation	★★★½☆	62.86
General		
Usability	★★★★★	78.13
Fun	★★★½☆	60.00
Interesting	★★★★☆	69.33

IV. Conclusion and Outlook

This paper introduced the novel web-based speech classification tool VoiLA which follows a unique approach by leveraging the browser-based crowdsourcing game iHEARu-PLAY for speech annotation to obtain required training data. It encourages people who helped annotate data – and anyone else – to try and evaluate the trained system by having their own voice analysed. VoiLA 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 user and visualized in the browser, giving users unique objective insights into how their voice sounds.

An extensive user evaluation survey showed that the proposed system has a good, bordering on excellent, usability and the task system proposed for voice recording is accepted well. Additional user comments indicated that some enhancements could be made in terms of accuracy of the emotion classification.

In the future, we will improve the classifiers by retraining them with already collected and annotated user data within VoiLA. Further future additions to VoiLA include giving users 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 user feedback from the conducted evaluation survey. Another long-term goal is to develop and integrate a classifier, which is capable of presenting the results to the user in real-time while they are speaking. Therefore, VoiLA has the potential to popularize the science behind voice analysis and the annotation process of iHEARu-PLAY.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement No. 338164 (ERC Starting Grant iHEARu). We thank audEERING for providing sensAI and all VoiLA users for user for taking part in our evaluation.

References

[1] V. Sethu, E. Ambikairajah, and J. Epps, "Empirical mode decomposition based weighted frequency feature for speech-based emotion classification," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. Las Vegas, USA: IEEE, April 2008, pp. 5017–5020.

[2] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep Neural Networks for Acoustic Emotion Recognition: Raising the Benchmarks," in *Proceedings of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing*. Prague, Czech Republic: IEEE, May 2011, pp. 5688–5691.

[3] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 582–596, 2009.

[4] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech Communication*, vol. 52, pp. 613–625, 2010.

[5] J.-S. Park, J.-H. Kim, and Y.-H. Oh, "Feature vector classification based speech emotion recognition for service robots," *IEEE Transactions on Consumer Electronics*, vol. 55, pp. 1590–1596, 2009.

[6] P. Rani, C. Liu, N. Sarkar, and E. Vanman, "An empirical study of machine learning techniques for affect recognition in human–robot interaction," *Pattern Analysis and Applications*, vol. 9, pp. 58–69, 2006.

[7] L. Vidrascu and L. Devillers, "Detection of real-life emotions in call centers," in *Proceedings of the 6th Annual Conference of the International Speech Communication Association*. Lisboa, Portugal: ISCA, September 2005, pp. 1841–1844.

[8] B. Lecouteux, M. Vacher, and 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*. Florence, Italy: ISCA, August 2011, pp. 2273–2276.

[9] A. Fleury, N. Noury, M. Vacher, and J.-F. Seri, "Sound and Speech Detection and Classification in a Health Smart Home," in *Proceedings of the 30th Annual International IEEE EMBS Conference*. Vancouver, Canada: IEEE, August 2008, pp. 4644–4647.

[10] F. Eyben, M. Wöllmer, T. Poitschke, B. Schuller, C. Blaschke, B. Färber, and N. Nguyen-Thien, "Emotion on the Road – Necessity, Acceptance, and Feasibility of Affective Computing in the Car," *Advances in Human Computer Interaction, Special Issue on Emotion-Aware Natural Interaction*, vol. 2010, p. 17, 2010.

[11] A. Tawari and M. Trivedi, "Speech based emotion classification framework for driver assistance system," in *Proceedings of the Intelligent Vehicles Symposium (IV)*. San Diego, USA: IEEE, June 2010, pp. 174–178.

[12] C. M. Jones and 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*. Narrabundah, Australia: CHISIG, June 2005, pp. 1–10.

[13] S. Amiriparian, S. Pugachevskiy, N. Cummins, S. Hantke, J. Pohjalainen, G. Keren, and B. Schuller, "CAST a database: Rapid targeted large-scale big data acquisition via small-world modelling of social media platforms," in *Proc. of the Conference on Affective Computing and Intelligent Interaction*. San Antonio, USA: IEEE, October 2017, pp. 340–345.

[14] S. Amiriparian, M. Schmitt, S. Hantke, V. Pandit, and B. Schuller, "Humans Inside: Cooperative Big Multimedia Data Mining," in *Innovations in Big Data Mining and Embedded Knowledge: Domestic and Social Context Challenges*, ser. Intelligent Systems Reference Library (ISRL), A. Esposito, A. M. Esposito, and L. C. Jain, Eds. Springer, 2018, 25 pages, invited contribution, to appear.

[15] A. Tarasov, S. J. Delaney, and C. Cullen, "Using crowdsourcing for labelling emotional speech assets," in *W3C Workshop on Emotion Markup Language*, Paris, France, October 2010, p. no pagination.

[16] P.-Y. Hsueh, P. Melville, and V. Sindhwani, "Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria," in *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, Stroudsburg, PA, USA, June 2009, pp. 27–35.

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

[18] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and Fast—but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, October 2008, pp. 254–263.

[19] S. Hantke, E. Marchi, and B. Schuller, "Introducing the Weighted Trustability Evaluator for Crowdsourcing Exemplified by Speaker Likability Classification," in *Proceedings of the 10th Language Resources and Evaluation Conference*. Portoroz, Slovenia: ELRA, May 2016, pp. 2156–2161.

[20] V. Ambati, S. Vogel, and J. Carbonell, "Active Learning and Crowd-Sourcing for Machine Translation," in *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Valletta, Malta, May 2010, p. no pagination.

[21] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing," in *Proceedings of the 1st International Workshop on Automatic Sentiment Analysis in the Wild held in Conjunction with the 6th Biannual Conference on Affective Computing and Intelligent Interaction*. Xi'an, P. R. China: IEEE, September 2015, pp. 891–897.

[22] S. Hantke, T. Appel, and 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 (ACII Asia)*. Beijing, P. R. China: AAAC, May 2018, 6 pages, to appear.

[23] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, Barcelona, Spain, October 2013, pp. 835–838.

[24] T. Vogt, E. André, and 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*. Kloster Irsee, Germany: Springer Berlin Heidelberg, June 2008, pp. 188–199.

[25] S. E. Eskimez, M. Sturge-Apple, Z. Duan, and W. Heinzelman, "WISE: Web-based Interactive Speech Emotion Classification," in *Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology held in conjunction with the 25th International Joint Conference on Artificial Intelligence*, New York City, USA, July 2016, pp. 2–7.

[26] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit," in *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. Amsterdam, the Netherlands: IEEE, September 2009, pp. 576–581.

[27] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, pp. 1–27, 2011.

[28] F. Weninger, J. Bergmann, and B. Schuller, "Introducing CURRENNT: the Munich Open-Source CUDA RecurREnt Neural Network Toolkit," *Journal of Machine Learning Research*, vol. 16, pp. 547–551, 2015.

[29] S. J. Young and S. Young, *The HTK hidden Markov model toolkit: Design and philosophy*. University of Cambridge, Department of Engineering, 1993.

[30] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, pp. 10–18, 2009.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[32] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andr, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and 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.

[33] J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability Evaluation in Industry*, vol. 189, pp. 4–7, 1996.

[34] A. Bangor, P. Kortum, and J. Miller, "Determining what individual sus scores mean: Adding an adjective rating scale," *Journal of Usability Studies*, vol. 4, pp. 114–123, 2009.