Introducing the Weighted Trustability Evaluator for Crowdsourcing Exemplified by Speaker Likability Classification

Simone Hantke[1,2], Erik Marchi[1,3], and Björn Schuller[1,4]
[1] Chair of Complex & Intelligent Systems, University of Passau, Germany
[2] Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[3] audEERING UG, Gilching, Germany
[4] Department of Computing, Imperial College London, UK
Email: simone.hantke@uni-passau.de

Abstract

Crowdsourcing is an arising collaborative approach applicable among many other applications to the area of language and speech processing. In fact, the use of crowdsourcing was already applied in the field of speech processing with promising results. However, only few studies investigated the use of crowdsourcing in computational paralinguistics. In this contribution, we propose a novel evaluator for crowdsourced-based ratings termed Weighted Trustability Evaluator (WTE) which is computed from the rater-dependent consistency over the test questions. We further investigate the reliability of crowdsourced annotations as compared to the ones obtained with traditional labelling procedures, such as constrained listening experiments in laboratories or in controlled environments. This comparison includes an in-depth analysis of obtainable classification performances. The experiments were conducted on the Speaker Likability Database (SLD) already used in the INTERSPEECH Challenge 2012, and the results lend further weight to the assumption that crowdsourcing can be applied as a reliable annotation source for computational paralinguistics given a sufficient number of raters and suited measurements of their reliability.

Keywords: Crowdsourcing, Speech Corpus Annotation, Computational Paralinguistics, Speech Classification

1. Introduction

Computational paralinguistics deals with the computer-based analysis and synthesis of paralinguistic phenomena. As opposed to many related phenomena, such as speakers states and traits, the term involves related fields such as social signal processing (Vinciarelli et al., 2012) and affective computing (Picard, 2000). A suitable definition can be potentially given in an 'ex-negativo' fashion: it comprises everything that is not dealt within phonetics or linguistics (Schuller and Batliner, 2014).

The success of supervised machine learning techniques for paralinguistic tasks depends highly on the quality of the existing labelled training data and therefore on the quality of the labels. The manual annotation of data by an expert is the primary way of gathering labelled training data for speech recognition. This way of getting the labels can be very time-consuming and expensive (Tarasov et al., 2010; Ambati et al., 2010; Hsueh et al., 2009; Kittur et al., 2008). Furthermore, the labels have to be estimated from the subjective opinion of a small number of experts who can often disagree on the labels (Tarasov et al., 2010; Donmez et al., 2009; Raykar et al., 2010).

This paper proposes an alternative way to gather annotated data for paralinguistic tasks with the use of non-professional annotators via crowdsourcing. Recently, with the access to crowdsourcing services such as Mechanical Turk[1] and CrowdFlower[2], it has become easier to get labels from multiple non-expert annotators.

Laboratory studies allow us, as researchers, to control many variables during the experiments. However, there are considerable differences in the environment between crowdsourcing workers and the traditional laboratory subjects. Therefore, trusting the work of the participant is only one of the many differences between crowdsourcing and laboratory environments. For instance, a crowdsourcing worker might not be completely concentrated on the task as he might be in a controlled laboratory environment.

However, it has been shown that the application of crowdsourcing can offer a fast and effective way to get labels (Tarasov et al., 2010; Hsueh et al., 2009) that are of the same quality as those from experts (Tarasov et al., 2010; Snow et al., 2008) at lower costs (Tarasov et al., 2010; Ambati et al., 2010).

1.1. Related Work

The idea of crowdsourcing was already applied in earlier works for several tasks. One of the earlier crowdsourcing studies in the field of language and speech processing was carried out to create speech and language data (Callison-Burch and Dredze, 2010), and for transcription of non-native speech (Evanini et al., 2010), or spoken language (Marge et al., 2010). Zaidan and Callison-Burch (2011) evaluated the quality of translations gathered from non-professionals and tried to increase the quality with special applied mechanisms. A comparison of crowdsourced data and data annotated via a university laboratory was performed in (Smucker and Jethani, 2011) by measuring the participants' judging behaviours and their relevance. Hsueh et al. (2009) proposed a study on quality management of crowdsourced data and examined the quality of the annotation data from expert annotators in a research lab and non-expert annotators from the internet applying the three criteria: noise level, sentiment ambiguity, and lexical uncertainty. Subsequently, in order to increase the quality

[1] https://www.mturk.com
[2] http://www.crowdflower.com

management of crowdsourced results, an algorithm which generates a scalar score representing the inherent quality of each worker was implemented by Ipeirotis et al. (2010).

1.2. Contribution of this Work

We suggest a novel method for the use of crowdsourced ratings and introduce the novel Weighted Trustability Evaluator (WTE). The WTE is computed from the rater-dependent accuracy over the given test questions to be annotated and is based on the weights of the ratings derived from the trustability of the rater. In this contribution, we further investigate the reliability of crowdsourced annotations as compared to the ones obtained in the 'traditional' labelling experiments. This comparison includes an in-depth analysis of obtainable classification performances. Finally, we conclude that crowdsourcing can be applied as a reliable annotation source for computational paralinguistics tasks given a sufficient number of raters and suited measurements of their reliability. Besides, to our best knowledge, only few studies investigated the use of crowdsourcing in the area of computational paralinguistics.

The paper is structured as follows: first, a description of the database used during experiments is given (Section 2). Then, Section 3 reports the way of gathering the labelled data via subjects in traditional laboratory experiments and via non-professionals with the use of crowdsourcing. In Section 4 we define the experimental tasks, the acoustic features, the set-up and the evaluation procedures. Results are presented in Section 5 before concluding the paper in Section 6.

2. The Speaker Likability Database (SLD)

For our experiments, we used the Speaker Likability Database (SLD) which is provided by Burkhardt et al. (2011) and was used in the INTERSPEECH Challenge 2012 (Schuller et al., 2012) and subsequently evaluated in (Eyben et al., 2013). The SLD is a subset of the German Agender Database (Burkhardt et al., 2010), which was originally recorded to study automatic age and gender recognition from telephone speech. The speech is recorded over fixed and mobile telephone lines at a sample rate of 8 kHz. The database contains 18 utterance types taken from a set listed in detail in (Burkhardt et al., 2010). For the SLD an age and gender balanced set of 800 speakers is selected. For each speaker, the longest sentence consisting of a command embedded in a free sentence is used, in order to keep the effort for judging the data by many listeners as low as possible. The SLD serves to evaluate features and algorithms for the detection of speaker age, gender and the average subjective likability of the speaker's voice by others. It is given with distinct definition of training, development, and test partitions, incorporating speaker independence, as needed in most real-life settings.

3. Gathering the Annotations

3.1. Laboratory Experiments

For our experiments with labelled data from traditional laboratory procedures, we used the annotated data of the SLD provided by (Burkhardt et al., 2011). Age, Gender and likability ratings of the data were established by presenting the stimuli to 32 participants (17 male, 15 female, aged 20–42 years, mean=28.6, standard deviation=5.4). To control the effects of gender and age group on the ratings, the stimuli were presented in six blocks with a single gender/age group. To mitigate the effects of fatigue or boredom, each of the 32 participants rated only three out of the six blocks in randomised order with a short break between each block. The order of stimuli within each block was randomised for each participant as well. Furthermore, the participants were instructed to rate the stimuli according to their likability, without taking into account sentence content or transmission quality. The rating of the likability was done on a seven-point Likert scale. All participants were paid for their service. (Burkhardt et al., 2011; Schuller et al., 2012).

To establish a consensus from the individual likability ratings (16 per instance), the Evaluator Weighted Estimator (EWE) (Grimm and Kroschel, 2005) was used. The EWE is a weighted mean, with weights corresponding to the reliability of each rater, which is the cross-correlation of her / his rating with the mean rating (over all raters). Hence, the EWE is – slightly adapted to our needs – defined as

$$\hat{x}_n^{EWE} = \frac{1}{\sum_{k=1}^{K} r_k} \sum_{k=1}^{K} r_k \hat{x}_{n,k}, \tag{1}$$

where r_k is the reliability of the k-th rater. For each rater, the cross-correlation is computed only on the block of stimuli s(he) rated. In general, the raters exhibit varying reliability ranging from a cross-correlation of .057 to .697.

3.2. Crowdsourcing

We examine the idea of creating further annotations for the SLD data via the crowdsourcing service CrowdFlower[2]. For our experiments we hired non-professional raters, and asked them to annotate the audio data for speaker gender and likability with the same set-up for better reproducibility.

In this study, we used the CrowdFlower interface and prepared a front-end using the CrowdFlower Markup Language (CML) and custom JavaScript. Furthermore, we used the interface to calibrate our task for changing parameters such as the amount of time required to complete a rating task, and the desired accuracy level to derive the payment.

However, crowdsourcing is prone to spammers trying to get paid without performing the task. Raters are given a low wage, and they are working in their own uncontrolled environments. For those reasons in order to ensure a high quality of the crowdsourced ratings, we adopted CrowdFlowers' quality control system of test question tasks, by pairing data items with correct responses. Due to the fact that the original dataset already has professional reference annotations, it allowed us to objectively and quantitatively compare the quality of our gathered non-professional ratings with respect to the existing ratings from the lab. We assumed that the performance obtained with laboratory participants should be considered as a 'ground truth' reference. Within the annotation procedure, the test questions are automatically mixed into the stream of regular tasks. If the accuracy – defined as trustability – of a worker, measured

Table 1: *Top ten origin countries of the raters.*

Rank #	Country	# Raters
1.	Romania	1.637
2.	India	1.114
3.	United Kingdom	1.026
4.	Spain	943
5.	Turkey	888
6.	Italy	824
7.	Serbia	760
8.	Bulgaria	745
9.	Russia	713
10.	Ukraine	624

by the number of test questions passed over the total number of test questions seen by the rater, drops below 90 %, we decided to not accept the worker for any further tasks. CrowdFlower claims that error rates are reduced by a factor of two when test questions are used[2]. All raters of the crowdsourced labels had a trustability over 90 % and thus passed our qualification test.

The selected raters labelled the audio data from different countries all over the world. We provide the details on the raters distribution per country in Table 1, showing the ten most frequent countries.

For each audio file we received 20 ratings, for a total of 16 000 ratings from different raters. Since we do not have a crowdsourced-based rating for each file coming from the same subject, the EWE is not applicable. Therefore, we established the Weighted Trustability Evaluator (WTE), based on the weights of the ratings derived from the trustability of the rater. The WTE is a weighted mean over all raters for an utterance. The weights are computed from the trustability of a rater, which is derived from the number of correct test answers over the total number of test questions. The WTE is defined as:

$$\hat{x}_n^{WTE} = \frac{1}{\sum_{k=1}^{N} t_k} \sum_{k=1}^{N} t_k \, x_{n,k} \quad (2)$$

$$\text{with} \quad t_k = \frac{P_k}{M}, \quad (3)$$

where t_k is the trustability of the k-th rater, computed from the P correct answers over M questions.

In the INTERSPEECH 2012 Speaker Trait Challenge (Schuller et al., 2012) the EWE was discretised into the two classes 'likable' (L) and 'non-likable' (NL). While the given original annotation provides likability in multiple levels, for our experiments the EWE rating was discretised into the three classes 'likable' (L), 'neutral' (N) and 'non-likable' (NL) based on the EWE rating of all stimuli in the SLD (Schuller et al., 2012; Burkhardt et al., 2011; Eyben et al., 2013). For our experiments, the WTE was also discretised into the three classes 'likable' (L), 'neutral' (N) and 'non-likable' (NL). The respective detailed partition into a training, development, and test set can be found in Table 2. Since both EWE and WTE ratings were discretised into these three classes, we can observe the distribution of EWE and WTE ratings in two histograms depicted in Figure 1. The histograms show the normalised EWE and WTE with a range [-1,1] and as in (Grimm and Kroschel, 2005), we used a fixed threshold equal to 0 for the EWE and WTE to obtain the three discrete classes. It seems that, the WTE ratings have a mean value closer to 0 and a lower standard variation than the EWE ratings.

Table 2: *Partitioning of the SLD into training, development, and test set for the gender task (top), the 2-class likability task (middle), and the 3-class likability task (bottom). Number of instances for the laboratory and crowdsourcing labels are given. (F: female / M: male, L: likable / N: neutral / NL: non-likable).*

	Laboratory			Crowdsourcing		
SLD #	**Train**	**Devel**	**Test**	**Train**	**Devel**	**Test**
F	199	89	115	197	90	114
M	195	89	113	197	88	114
L	189	92	119	348	128	156
NL	205	86	109	46	50	72
L	91	35	54	53	82	99
N	174	75	93	276	81	114
NL	129	68	81	65	15	15
Σ	394	178	228	394	178	228

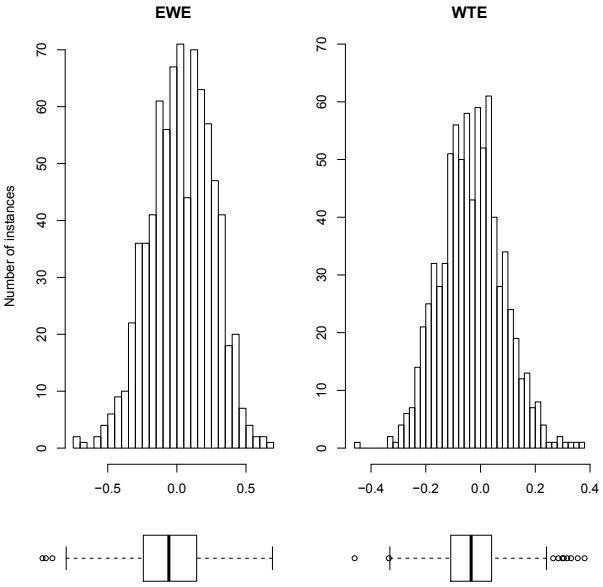

Figure 1: Normalised EWE and WTE histograms with a range [-1,1] and a fixed threshold equal to 0 to obtain the three discrete classes. The WTE ratings have a mean value closer to 0 and a lower standard variation than the EWE rating.

4. Experiments

In the following, we validate the effectiveness of our suggested novel approach by describing first the experimental tasks, then the feature sets and experimental setup, and finally our evaluation and analysis criteria.

Table 3: *Unweighted Average Recall (UAR) and Weighted Average Recall (WAR) for the 2-class and 3-class* Likability *tasks. Shown are the best performances obtained with traditional ratings (EWE), and crowdsourced ratings (WTE) using SVM with linear kernel. Cross-label results are also shown by training with EWE labels and testing with WTE ratings, and vice-versa. C: complexity parameter, optimised on the development set. Average results are computed over different complexities and different WTE and EWE thresholds.*

Task	**Chance**	*EWE*		*WTE*		*EWE vs WTE*		*WTE vs EWE*	
	Level	C	**UAR** (WAR)	C	**UAR** (WAR)	C	**UAR** (WAR)	C	**UAR** (WAR)
Likability 2-class									
{L, NL}	50.0	10^{-2}	59.5 (59.6)	10^{-3}	**60.2** (64.5)	10^{-1}	57.4 (57.5)	10^{-2}	**59.2** (60.1)
Likability 3-class									
{L, N, NL}	33.3	10^{-2}	45.7 (46.5)	10^{-2}	**58.6** (57.0)	10^{-3}	35.3 (36.7)	10^{-3}	**45.5** (43.8)
Average	33.3	-	41.6 (43.6)	-	**48.1** (52.3)	-	33.5 (36.7)	-	**40.5** (41.8)

4.1. Experimental Tasks

Three tasks were evaluated: gender, 2-class likability, and 3-class likability. The gender task concerns the classfication of male, and female subjects. In the original INTERSPEECH 2012 Speaker Trait Challenge (Schuller et al., 2012) condition the likability task aims to distinguish between likable and non-likable vocal expressions. Subsequently, in addition to this the likability task was performed as 3-class task covering the recognition of likable, neutral, and non-likable vocal expressions. A first evaluation of the effectiveness of the crowdsourced ratings is carried out on the gender task. Then, we evaluate crowdsourcing on the likability tasks.

4.2. Acoustic Features

For better reproducibility the acoustic feature set used in our experiments corresponds to the feature set of the INTERSPEECH 2012 Speaker Trait Challenge (Schuller et al., 2012). We use the open-source openSMILE feature extractor (Eyben et al., 2010) to 'brute-force' a high-dimensional feature set by applying statistical functionals to frame-wise low-level descriptors (LLDs), which comprise energy, spectral, and voicing related LLDs. Regarding functionals, we aim at a compromise between a broad variety of functionals, including mean, min, max, and moments. Altogether, the INTERSPEECH 2012 Speaker Trait Challenge feature set contains 6 125 features.

4.3. Setup

Also, for better reproducibility we applied the same set-up used in the INTERSPEECH 2012 Speaker Trait Challenge (Schuller et al., 2012). Adopting the Weka toolkit (Hall et al., 2009), Support Vector Machines (SVMs) with linear kernel were trained with the Sequential Minimal Optimization (SMO) algorithm. SVMs have been chosen as classifier since they are a well known standard method for computational paralinguistics. SVMs are discriminative classifiers which do not require large amounts of training data, making them especially suited for our task. The SVM training has been made at different complexity constant values $C \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. The 3-class classification problem is handled by constructing exhaustive pairwise one-vs-one SVMs.

Table 4: *Unweighted Average Recall (UAR) and Weighted Average Recall (WAR) for the 2-class* gender *task. Shown are the best performances obtained with traditional ratings (EWE), and crowdsourced ratings (WTE) using SVM with linear kernel. Cross-label results are also shown by training with EWE labels and testing with WTE ratings, and vice-versa. C: complexity parameter, optimised on the development set.*

Ratings	C	**UAR** (WAR)	AUC
EWE	10^{-2}	96.1 (96.1)	98.9
WTE	10^{-1}	95.6 (95.6)	97.8
EWE vs WTE	10^{-2}	95.6 (95.6)	97.8
WTE vs EWE	10^{-1}	96.0 (96.1)	98.0

4.4. Evaluation

As evaluation measure, we stick with unweighted average recall (UAR) as used in (Schuller et al., 2011). In the given case of two classes ('X' and 'NX'), it is calculated as (Recall(X)+Recall(NX))/2, in the given case of the three classes ('X', 'MX' and 'NX'), it is calculated as (Recall(X)+Recall(MX)+Recall(NX))/3, i. e., the number of instances per class is ignored by intention. The motivation to consider unweighted average recall rather than weighted average recall (WAR) is that it is also meaningful for highly unbalanced distributions of instances among classes, and for more than two classes. In the case of equal distribution, UAR and WAR naturally resemble each other. The evaluation is performed on the test set, where we retrain the models using the training and development set.

5. Results

5.1. Gender

Using the gathered labels, we investigated the efficacy of our crowdsourcing framework in providing reliable ratings. We first evaluated the gender task which is trivial, but still it is worth mentioning as an initial proof-of-concept.

Therefore, we compared performances achieved using the labels obtained from the laboratory and crowdsourcing annotators. Table 4 shows the results obtained with laboratory ratings (EWE), and crowdsourced ratings (WTE). We further performed cross-label experiments in order to com-

pare the EWE-based and WTE-based systems under mismatched conditions.
EWE and WTE show similar performances up to 96.1 % and 95.6 % UAR, respectively. As an additional metric, we also provided results in terms of area under the curve (AUC) on which we achieve performances up to 98.9 % and 97.8 % respectively for EWE and WTE. Subsequently, we achieved similar performances also in the cross-label experiments providing a further indication that our crowdsourced labels can reliably be used in this task.

5.2. Likability

Table 3 shows the best performances for the 2-class task obtained with traditional ratings (EWE) up to 59.5 % UAR. For crowdsourced ratings (WTE) results up to 60.2 % UAR were archived. In addition, we show cross-label results by training with EWE labels and testing with WTE ratings, and vice-versa. We can observe the best results for EWE testing with WTE labels of up to 57.4 % UAR, and WTE testing with EWE labels of up to 59.2 % UAR, respectively. Besides the comparison with state-of-the-art performances, we also applied an alternative 3-class discretisation by introducing the neutral class. More specifically, the 3-class task was computed averaging different complexities and different WTE and EWE thresholds. The best result was obtained with the thresholds between -0.14 and 0.14 for EWE and -0.17 and 0.08 for WTE.
The second row of Table 3 shows the performances for the 3-class task, up to 45.7 % UAR (41.6 % UAR on average) for EWE. The WTE archived results up to 58.6 % UAR (48.1 % UAR on average), showing an improved performance against EWE. The cross-label results show best results for EWE testing with WTE labels of up to 35.3 % UAR (33.5 % UAR on average), and WTE testing with EWE labels of up to 45.5 % UAR (40.5 % UAR on average), respectively.
For all performances obtained, the 3-class task provided better performance compared to the 2-class task. Comparing the cross-label results for the 3-class task and the 2-class task, we observe further improvement using the WTE instead of EWE. The proposed WTE seems to be more robust under cross-label evaluation, with a significant absolute improvement (one-tailed z-test (Smucker et al., 2007), $p<0.01$) of 12.9 % UAR against the traditional ratings coming from different raters.

6. Conclusions

We proposed a novel evaluator for crowdsourced-based ratings termed Weighted Trustability Evaluator (WTE) which is computed from the rater-dependent consistency over the test questions. We further investigated the reliability of crowdsourced annotations as compared to the ones obtained from different annotators with traditional labelling procedures. This comparison showed a significant improvement in performances against the traditional labels of up to 12.9 % absolute improvement. Additionally, cross-label experiments showed, that WTE-based labels seem to perform more robustly. The experiments were conducted on the Speaker Likability Database (SLD) already used in the INTERSPEECH Challenge 2012. The caveat has to be made that this is a pilot study, with a limited number of samples per class; the results will be reviewed and verified with larger databases and crowdsourced ratings collected in the future. However, the results lend further weight to the assumption that crowdsourcing can be applied as a reliable annotation source for computational paralinguistics given a sufficient number of raters and suited measurements of their reliability.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreements No. 338164 (ERC Starting Grant iHEARu) and the European Union's Horizon 2020 Programme through the Research and Innovation Action No. 688835 (DE-ENIGMA) and No. 644632 (MixedEmotions).

Ambati, V., Vogel, S., and Carbonell, J. (2010). Active learning and crowd-sourcing for machine translation. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages x pages – no pagination, Valletta, Malta, may. European Language Resources Association (ELRA).

Burkhardt, F., Eckert, M., Johannsen, W., and Stegmann, J. (2010). A database of age and gender annotated telephone speech. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, page no pagination, Valletta, Malta, may. European Language Resources Association (ELRA).

Burkhardt, F., Schuller, B. W., Weiss, B., and Weninger, F. (2011). "would you buy a car from me?" - on the likability of telephone voices. In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, pages 1557–1560. ISCA.

Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 1–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Donmez, P., Carbonell, J. G., and Schneider, J. (2009). Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 259–268, New York, NY, USA. ACM.

Evanini, K., Higgins, D., and Zechner, K. (2010). Using amazon mechanical turk for transcription of non-native speech. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 53–56, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 1459–1462, New York, NY, USA. ACM.

Eyben, F., Weninger, F., Marchi, E., and Schuller, B. (2013). Likability of human voices: A feature analysis and a neural network regression approach to automatic likability estimation. In *Proceedings 14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services, WIA2MIS 2013*, Paris, France, July. IEEE, IEEE.

Grimm, M. and Kroschel, K. (2005). Evaluation of natural emotions using self assessment manikins. In *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, pages 381–385, San Juan, November. IEEE.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Hsueh, P.-Y., Melville, P., and Sindhwani, V. (2009). 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*, HLT '09, pages 27–35, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ipeirotis, P. G., Provost, F., and Wang, J. (2010). Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 64–67, New York, NY, USA. ACM.

Kittur, A., Chi, E., and Suh, B. (2008). Crowdsourcing for usability: Using micro-task markets for rapid, remote, and low-cost user measurements. In *Proceedings of CHI 2008*, no pagination.

Marge, M., Banerjee, S., and Rudnicky, A. I. (2010). Using the amazon mechanical turk for transcription of spoken language. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5270–5273, Dallas, TX, March. IEEE.

Picard, R. W. (2000). *Affective computing*, volume Reprint. MIT press.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. (2010). Learning from crowds. *J. Mach. Learn. Res.*, 11:1297–1322, August.

Schuller, B. and Batliner, A. (2014). *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley, November.

Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (2011). Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge. *Speech Communication, Special Issue on Sensing Emotion and Affect - Facing Realism in Speech Processing*, 53(9/10):1062–1087, November/December.

Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., and Weiss, B. (2012). The INTERSPEECH 2012 Speaker Trait Challenge. In *Proceedings INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, Portland, OR, September. ISCA, ISCA.

Smucker, M. D. and Jethani, C. P. (2011). The crowd vs. the lab: A comparison of crowd-sourced and university laboratory participant behavior. In *Proceedings of the SIGIR 2011 Workshop on crowdsourcing for information retrieval*, page no pagination, Beijing, China, July.

Smucker, M., Allan, J., and Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on information and knowledge management*, pages 623–632. ACM.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). 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*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tarasov, A., Delaney, S. J., and Cullen, C. (2010). Using crowdsourcing for labelling emotional speech assets. *W3C workshop on Emotion ML, Paris, France*, page no pagination, October.

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., and Schroeder, M. (2012). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *Affective Computing, IEEE Transactions on*, 3(1):69–87.

Zaidan, O. F. and Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1220–1229. Association for Computational Linguistics.