TRADEOFF BETWEEN QUALITY AND QUANTITY OF EMOTIONAL ANNOTATIONS TO CHARACTERIZE EXPRESSIVE BEHAVIORS

Alec Burmania, Mohammed Abdelwahab, and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory, Department of Electrical Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
axb124530@utdallas.edu, mxa129730@utdallas.edu, busso@utdallas.edu

ABSTRACT

Emotional descriptors collected from perceptual evaluations are important in the study of emotions. Many studies on emotion recognition depend on these labels to train classifiers. The reliability of the emotion descriptors vary with the number and quality of the raters. Conducting perceptual evaluations used to be an expensive and time demanding task, resulting in emotional databases with poor labels annotated by few raters. Nowadays, crowdsourcing services have simplified the process, reducing the cost, facilitating more evaluations per stimuli. The key challenge in using crowdsourcing for perceptual evaluation is the quality which significantly varies across workers. Is it better to have multiple annotations with lower inter-evaluator agreement or to have few annotations with higher inter-evaluator agreement? This study explores this tradeoff between quality and quantity in emotional annotations to characterize expressive behaviors. The analysis relies on emotional labels from the MSP-IMPROV database, where each video was evaluated by over 20 workers. We discuss the theoretical concept of effective reliability to address this problem. We demonstrate that a reduced set of labels with higher inter-evaluator agreement can provide similar classification performance than unfiltered set of labels from multiple workers. We discuss best practices to collecting annotations for emotion recognition tasks using crowdsourcing.

Index Terms— inter-evaluator agreement, crowdsourcing, emotion recognition

1. INTRODUCTION

Emotional labels are important in the study of emotion. For example, supervised machine learning frameworks in emotion recognition require training data, usually in the form of labels from perceptual evaluations [1]. These labels are typically collected using hired evaluators or volunteers. The quality of the work that these participants produce is crucial to obtaining useful classifiers for emotion recognition systems. The number of annotators who participate in the evaluation is also a factor that affects the overall agreement within the labels and, therefore, the classifiers. Ideally, a corpus would have a large amount of annotations with near-perfect agreement, but this is not usually the case. The videos in most of the emotional databases are evaluated by few raters, due to the cost associated with conducting perceptual evaluations [1].

Recently, studies have relied on crowdsourcing services to annotate emotional databases [2–5], offering an interesting alternative where multiple annotators can be recruited. The annotations can be purchased for a fraction of the cost of traditional evaluations, in a short amount of time, and with minimal effort [6]. Crowdsourcing allows researchers to collect large amounts of annotations from a diverse pool of workers [7]. The challenge with crowdsourcing services is the quality in the labels. The reliability varies across evaluators, so the quality in the labels may not be as high as the quality obtained by expert in controlled laboratory. Is it better to have multiple annotations with lower inter-evaluator agreement or to have few annotations with higher inter-evaluator agreement? Is it worthwhile to recruit 20 evaluators per video instead of just 5? It is important to construct a framework for assessing the quality and quantity of labels required to produce desired results for a given cost.

This study evaluates the tradeoff between quality and quantity of emotional annotations. The study relies on a subset of 648 videos from the MSP-IMPROV corpus [8], which are evaluated by 28 raters, on average. By removing less reliable annotations, we can significantly increase the inter-evaluator agreement from $\kappa = 0.42$ to $\kappa = 0.57$ (Fleiss' Kappa statistic). We discuss the theoretical concept of effective reliability [9] to compare different tradeoffs between quality and quantity in the perceptual evaluations. This metric combines qualitative and quantitative measures, fitting the scope of this study. Finally, we evaluate the performance of emotion classifiers trained with labels derived from four different conditions, where we vary the quality and quantity of the annotations. We discuss classification performance in terms of effective reliability and experimental cost. The findings in this paper can aid in experimental design when using crowdsourcing to derive emotional labels.

2. BACKGROUND

Emotional databases are commonly annotated by few evaluators per sample. Examples include [number of evaluators listed after the corpus' name] the IEMOCAP - 3 [10], AVIC - 4 [11], FAU AIBO - 5 [12], and CCD - 4 [13] databases. Recently, crowdsourcing services have allowed researchers to increase the number of evaluations per sample. An example is the CREMA-D database annotated by 9.8 raters, on average, using crowdsourcing [3].

2.1. Crowdsourcing and Quality

Amazon Mechanical Turk (MTurk) [14] is a well-known platform for crowdsourcing tasks, which allows a researcher (requester) to seek help on tasks (HITs) from subjects (Turkers). We refer to Turkers as workers, using a more general terminology. MTurk has been used for the annotation of video, speech, and emotion annotations [2, 4, 15]. A problem with crowdsourcing is the quality due to spam from bots, and workers who are not honestly interested in completing their task, seeking only the payment. Many studies have focused on how to detect these forms of cheating [16]. For example, Buchholz et al. [17] described a quantitative analysis of workers who blatantly cheat on tasks for payment (e.g., completing the task without watching the videos required for the evaluation). MTurk provides some built in methods for prescreening workers including location, as well as qualitative metrics such as number of tasks completed and percentage of tasks approved by requesters. Eickhoff and Vries [16] analyzed these filters, concluding that some (especially location filters)

This work was funded by NSF (IIS-1217104, IIS-1453781).

are more helpful at preventing spam, while others (such as approval ratings) may not prevent spam.

2.2. Increasing the Quality in Crowdsourcing

Pre-filters, real-time filters, and post-filters are all used in crowdsourcing evaluations. Usage of each type of filter varies by the type of task and preference of the requester. Parent and Eskenazi [18] describe different usage scenarios for filters with respect to MTurk including pre-filters provided by the service. They noted that real-time systems are uncommon, though do exist [19, 20]. That study also noted that post-filtering is the most common approach in the tasks they observed. Post-filtering usually consists of removing labels that have been already collected. This can be implemented via majority vote or filtering by inter-evaluator agreement. The use of majority vote can even be employed as a basis for rejecting or accepting workers annotations [21]. Marge et al. [22] proposed to ask workers to evaluate or improve the evaluations from other workers [22]. They implemented a multi-stage corrective process, where workers are asked to investigate and correct disagreements between workers in the first stage of speech transcription, leading to a significant increase in quality compared with expert raters.

2.3. Effective Reliability: Quality versus Quantity

The amount of evaluations and quality of evaluations to create a "good" classifier is not inherently obvious, as data sets have different levels of difficulty in the annotation (samples with ambiguous emotions versus prototypical emotions). What qualifies as a "good" classifier is entirely relative to other classifiers for that specific task. The annotation cost can usually be seen as a function of desired quality from a reliability standpoint (qualifications, higher payment, more training) and/or quantity standpoint (amount of annotations desired). Evaluating large corpora can quickly become expensive so it is important to find the right tradeoff between quantity and quality.

Rosenthal et al. [9] presented an interesting analysis to evaluate the effect of rater's quality and quantity in term of effective reliability, which we adopt in this study. They started with a reliability metric about inter-evaluator agreement such as Spearman's ρ, phi coefficient, or Fleiss' kappa. This value, denoted as r, is then applied to the Spearman-Brown equation as shown in Equation 1, where n is the number of evaluators.

$$R_{SB} = \frac{nr}{1+(n-1)r} \tag{1}$$

The effective reliability, R_{SB} can be used as a metric for interpreting and comparing the goodness of labels derived by multiple raters who produce annotations with a given inter-evaluator agreement. For example, the effective reliability metric can be used to determine the number of evaluators that are required to replace the annotations from few experts. Table 1 provides a similar table to the one provided by Rosenthal et al. [9], where we choose a subset of the quantitative and qualitative values which are relevant to this study. We use Fleiss' kappa statistics. The reliabilities in bold refer to data points which will be compared in the speech emotion classification evaluation described in Section 6. The table shows that the same effective reliability can be achieved from two entirely different cases on the chart. [n=10, r=.42] and [n=5, r=.6] have an effective reliability of 88 (e.g., 10 less reliable workers are as effective as 5 experts). We wish to study the concept of effective reliability in the evaluation of emotions, including emotion classification. Is it worthwhile to remove samples with lower inter-evaluator agreement with post-filters just to justify a higher Fleiss' kappa statistics?

Table 1. Effective reliability in terms of number of workers, and reliability.

	Mean reliability (r)						
n raters	0.42	0.45	0.48	0.51	0.54	0.57	0.60
5	**78**	80	82	84	85	**87**	88
10	88	89	90	91	92	93	94
15	92	**92**	93	94	95	95	96
20	**94**	94	95	95	96	96	97

3. THE MSP-IMPROV DATABASE

The MSP-IMPROV [8] database is a multimodal corpus of dyadic interactions recorded from twelve actors from the University of Texas at Dallas. The corpus provides conversational renditions of sentences with fixed lexicon content, conveying different emotions. The elicitation scheme consists of designing carefully selected hypothetical scenarios that two actors improvise. The context leads one of the actor to utter a target sentence. By changing the scenarios, we achieve renditions of the same sentences portraying different emotions (e.g., "*How can I not?*" - scenario for happiness: receiving a job offer; scenario for anger: reacting to a lazy friend who suggests that you do not have to attend classes). This corpus includes 648 samples containing these target sentences (*Improvised - Target* set). The target emotions are happiness, anger, and sadness, plus neutral state. We recorded the entire dyadic improvisations that led to the target sentences. We also recorded the interactions between actors during the breaks. In total, the corpus has 8,438 speaking turns. This study only considers the 648 target sentences, since the emotions in this subset was perceptually evaluated by more workers. Busso et al. [8] presents further details of this corpus.

4. ONLINE QUALITY ASSESSMENT OF WORKERS

The perceptual evaluation for this corpus involves a multi-task evaluation per HIT over the videos of the MSP-IMPROV corpus. We implement an approach that assess in real-time the performance of the workers. A comprehensive description of this perceptual evaluation can be found in Burmania et al. [20]. This section describes the aspects that are relevant for this study. A reference set is pre-evaluated and used as gold standard. Each video in the reference set is originally evaluated by five workers. A random set of 5 reference sentences are interleaved after every 20 sentences from the rest of the corpus (5 reference, 20 new, 5 reference, and so on). After annotating 30 videos (20 new, plus 10 reference sentences), we evaluate whether the new labels from the worker increase or decrease the inter-evaluator agreement over the reference set. We measure quality using the angular similarity metric described in Section 4.1. If the inter-evaluator agreement increases, the worker can evaluate 20 new videos plus 5 reference videos. This process continues until the inter-evaluator agreement decreases (e.g., due to fatigue or lack of interest), the worker quits, or the evaluation concludes after 105 videos. Since the *Improvised - Target* set is the most important sentences for this corpus, we use these 648 sentences as s reference set. These sentences are evaluated on average by 28 workers due to the overhead associated with assessing the quality of the workers during the evaluation. While the questionnaire includes multiple questions (e.g., secondary emotions, dimension attributes for activation, valence, and dominance), we only assessed the performance of the workers in selecting the primary emotions that best describes the video: anger, sadness, happiness, neutrality and other.

4.1. Angular Similarity

To evaluate the performance of workers in selecting primary emotion over five videos we use the angular similarity. The metric takes values between 0° and 90°, and provides an angular representation of the difference between the labels over this 5-class problem. Using the labels pre-collected for the reference set (5 annotations per video), we estimate θ^s_{ref} for each video in the reference set as follows (see Eq. 2). First, we map the annotations for a video into a 5D vector, where the dimensions correspond to the emotional classes (e.g., [anger, happiness, sadness, neutral, others]). For example, if a video is evaluated as "happiness" by three workers, and "sadness" by two workers, its corresponding vector is [0,3,2,0,0]. We defined $\vec{\mathbf{v}}^s_{(i)}$ as the vector formed by considering all the annotations from the N workers except the ith worker. We also define the elementary vector $\hat{\mathbf{v}}^s_{\mathbf{i}}$ in the direction of the emotion provided by the ith worker (e.g., [1 0 0 0 0] for "anger"). Then, we estimate the angle between $\vec{\mathbf{v}}^s_{(i)}$ and $\hat{\mathbf{v}}^s_{\mathbf{i}}$. We repeat the process for all the annotations, deriving the average angle θ^s_{ref} for video s. If all the evaluations are consistent (e.g., all the annotations are "happiness"), the value for θ^s_{ref} is 0° (easy case). The worse case is for θ^s_{ref} equals to 90°, when all the workers disagree on the labels (difficult case).

$$\theta^s_{ref} = \frac{1}{N}\sum_{i=1}^{N} \arccos\left(\frac{< \vec{\mathbf{v}}^s_{(i)}, \hat{\mathbf{v}}^s_{\mathbf{i}} >}{||\vec{\mathbf{v}}^s_{(i)}|| \cdot ||\hat{\mathbf{v}}^s_{\mathbf{i}}||}\right) \qquad (2)$$

For a new worker t, we estimate the angle θ^s_t between his/her evaluation $\hat{\mathbf{v}}^{\mathbf{s}}_{\mathbf{t}}$ and the vector $\vec{\mathbf{v}}^s$ formed by considering all annotations from the N workers for video s (Eq. 3). Finally, we subtract this angle to the reference angle, obtaining $\Delta\theta^s_t$. A positive value indicates that the inter-evaluator agreement increases when the labels from worker t are included. A negative value for $\Delta\theta^s_t$ indicates a drop in inter-evaluator agreement.

$$\theta^s_t = \arccos\left(\frac{< \vec{\mathbf{v}}^s, \hat{\mathbf{v}}^{\mathbf{s}}_{\mathbf{t}} >}{||\vec{\mathbf{v}}^s|| \cdot ||\hat{\mathbf{v}}^{\mathbf{s}}_{\mathbf{t}}||}\right) \qquad (3)$$

$$\Delta\theta^s_t = \theta^s_{ref} - \theta^s_t \qquad (4)$$

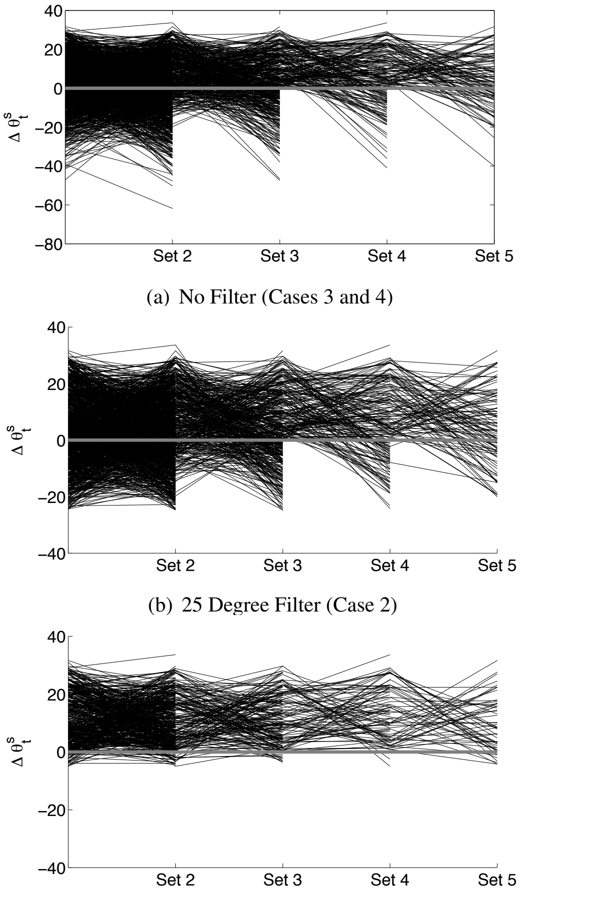
The benefits of this metric are (1) it is sensitive to minority labels, which are not penalized as much as labels never selected by the N workers in the reference set, (2) it directly measures increases and decreases in inter-evaluator agreement, (3) it captures the differences in agreement, so it is invariant to the complexity in evaluating different videos, and (4) it can be robustly estimated over few videos.

5. POST-FILTER APPROACH: QUALITY AND QUANTITY

Figure 1(a) illustrates the approach used for the perceptual evaluation. Each line represents the performance of an individual worker. At each evaluation point (Set 2 - Set 5), we measure $\Delta\theta^s_t$. We stop the evaluation when this value is negative. The gray line at $\Delta\theta^s_t$=0 illustrates this threshold.

Since this evaluation is conducted every 25 videos, workers can drop their performance between evaluation points. For these cases, we can either keep or reject the labels provided after the previous successful evaluation point. We implement a post-filtering approach, where we compare the value for $\Delta\theta^s_t$ for rejected workers to a second threshold $t_\triangle$. If the value is below this threshold, we ignore all the labels assigned by this worker after the previous successful evaluation point (e.g., we assume that his/her performance really dropped during the evaluation of the last 25 videos). By changing the value of this threshold, we can achieve different qualities, removing less reliable annotations. Figure 1(b) shows the case when the threshold is set to $t_\triangle = -25°$. All the evaluations below this threshold are ignored. Figure 1(c) shows the evaluations that survive $t_\triangle = -5°$, which is a more strict threshold, removing many of the evaluations.

The threshold $t_\triangle$ can really impact the quality and quantity of the labels. Table 2 reports the Fleiss' Kappa statistic for different values of $t_\triangle$ over the first N workers, where $N \in [5, 10, 15, 20, 25]$. For example, with no post-filter approach (e.g., $t_\triangle = -90°$; last row in Table 2), there are 381 sentences with 25 (or more) workers. The inter-evaluator agreement for the first 25 workers is $\kappa = 0.409$. We can achieve Fleiss' Kappa statistic as high as κ=0.572 with strict threshold ($t_\triangle = -5°$; first row in Table 2). However, the number of videos with more than 15 or more labels decreases to 246 (only 38% of the set). The table highlights four cases that we further study:



(a) No Filter (Cases 3 and 4)

(b) 25 Degree Filter (Case 2)

(c) 5 Degree Filter (Case 1)

Fig. 1. Maps of workers performance through the evaluation task given in degrees. Each filter increases the angular similarity between raters by eliminating evaluations below a set threshold.

Case 1 ($t_\triangle$=−5°; N=5; κ = 0.572): This case uses the most restrictive filter and uses a small number of worker (5). This case trades off cost and quantity for high inter-evaluator agreement.

Case 2 ($t_\triangle$=−25°; N=15; κ = 0.450): This case uses a moderate filter and uses a decently large number of workers (15). This case balances quality and quantity.

Case 3 ($t_\triangle$=−90°; N=5; κ = 0.422): This case uses no filter and uses a small number of workers (5). This case represents a task that aims to complete a minimal amount of evaluations with the filtering present in the crowdsourcing task (pre-filtering, CAPTCHA [23]).

Case 4 ($t_\triangle$=−90°; N=20; κ = 0.419): This case uses no filter and uses a large sample size (20). This case trades off quality for a high sample size without filtering as in Case 3.

The ranking of these cases in terms of effective reliability is: Case 4 (94), Case 2 (92), Case 1 (87) and Case 3 (78) – See Table 1.

Table 2. Fleiss' Kappa statistic achieved under different quality threshold and number workers. Highlighted are four cases of interest.

$t_\triangle$ [°]	5 Workers		10 Workers		15 Workers		20 Workers		25 Workers	
	# sent.	κ	# sent.	κ	# sent.	κ	# sent.	κ	# sent.	κ
-5	**638**	**0.572**	525	0.558	246	0.515	52	0.488	0	–
-10	643	0.532	615	0.522	466	0.501	207	0.459	26	0.455
-15	648	0.501	643	0.495	570	0.483	351	0.443	112	0.402
-20	648	0.469	648	0.471	619	0.463	510	0.451	182	0.414
-25	648	0.452	648	0.450	**643**	**0.450**	561	0.440	247	0.416
-30	648	0.438	648	0.433	648	0.436	609	0.431	298	0.410
-35	648	0.425	648	0.433	648	0.426	619	0.424	346	0.403
-40	648	0.420	648	0.427	648	0.425	629	0.423	356	0.402
-90	**648**	**0.422**	648	0.419	648	0.422	**629**	**0.419**	381	0.409

Table 3. Differences in the labels between cases when considering only the common sentences (514 videos)

	Case 1	Case 2	Case 3	Case 4
Case 1	–	26	40	32
Case 2		–	32	10
Case 3			–	36

6. IMPLICATIONS IN EMOTION CLASSIFICATION

This section evaluates whether the effective reliabilities translate into classification performance. Using the four cases described in Section 5, we derive labels for the sentences using majority vote. Some of the turns may not appear in every case, since they may not have the required amount of evaluations after the filter, or may not have a majority vote consensus (and are thus not considered in the classification evaluation). For each classification problem, we use *OpenSmile* [24] to extract the Interspeech 2013 standard feature set. We then perform feature selection in two steps. In the first step, we employ Correlation Attribute Evaluation to select features that correlate well with the class labels, reducing the feature set from 6373 to 1000 features. The second step further reduces the number of features used to 50, by selecting the features that optimize the classifier's accuracy, using *floating forward feature selection* (FFFS). We use *LIBSVM* [25] to train and test our SVM classifier with RBF kernel. We ensure that the data used for training and testing was balanced using random under sampling of the abundant classes. We employ a six-fold *leave-one-speaker-out* (LOSO) cross-validation. To account for the variations introduced by the random under sampling, we ran the experiment 4 times, resulting in a total of 24 folds. In addition to accuracy, we estimate the average precision and recall rates across emotional label. Their average values are used to estimate the F-score.

Before discussing the classification results, it is interesting to evaluate the difference between the actual labels assigned with majority vote to each case. We consider 514 sentences, which are common to the four cases. Table 3 shows the number of sentences that have different labels between cases. All the differences are less or equal to 40 (7.8% of sentences). Most of the labels are consistent across cases, so we do not expect significant differences in the classification results. The most different case is between case 1 and case 3. Since both cases have 5 annotations per video, the difference is only due to the increased quality in case 1.

The top half of Table 4 reports the average classification results across all the folds. The results from Table 4 show some interesting trends. The highest difference is between cases 1 and 2. The trend is consistent with effective reliability, indicating that extra evaluations in case 2 overcame its lower inter-evaluator agreement. We speculate that some of the differences in performance are due to the inclusion or exclusion of ambiguous turns that reach agreement with majority vote for certain cases. Therefore, we repeat this experiment using only common turns between each of the cases to keep the training and testing content consistent. The bottom half of Table 4 shows these results. While the differences are marginal, the trends are consistent with the effective reliability (cases 4 and 2 are slightly better than cases 1 and 3).

Table 4. Speech emotion classification results for the four cases of interest (*Acc.*=Accuracy; *Pre.*=Precision; *Rec.*=Recall).

			All Turns				
	$t_\triangle$	N	# Turns	Acc. [%]	Pre. [%]	Rec. [%]	F-score [%]
Case 1	-5°	5	605	45.8	44.7	45.8	45.3
Case 2	-25°	15	615	48.4	47.9	48.4	48.1
Case 3	-90°	5	575	47.9	47.4	47.9	47.7
Case 4	-90°	20	614	46.9	45.9	46.9	46.4
			Only common turns across conditions				
	$t_\triangle$	N	# Turns	Acc. [%]	Pre. [%]	Rec. [%]	F-score [%]
Case 1	-5°	5	514	47.4	46.5	47.4	47.0
Case 2	-25°	15	514	48.2	47.4	48.2	47.8
Case 3	-90°	5	514	47.1	46.6	47.1	46.8
Case 4	-90°	20	514	47.9	47.2	47.9	47.5

7. CONCLUSIONS

This study explored the tradeoff between quality and quantity in emotional annotations collected with perceptual evaluations. We leveraged the concept of effective reliability to understand optimal configurations to achieve a desire quality from (unreliable) raters. The analysis and classification results provide valuable insights to guide future perceptual evaluations.

An interesting result is that very few sentences change labels when the number of workers increases from 5 to 15 (or 20). To annotate emotional databases for emotion classification, five annotations per video may be enough. We notice that when we use common videos in the classification evaluation, the results are closer to what we expected from the analysis on effective reliability. It would be interesting to collect additional annotations to increase the number samples under each case. This can allow us to better understand the effects of filtering and majority vote on turns with more ambiguous emotional content. These are the videos that are more likely to change labels across cases, increasing the difference in classification performance. It may also be interesting to use different corpora for each case. This would diversify the labels and content for each classifier, allowing us to observe greater differences in the results. Notice that this framework can also be applied to other emotional metrics such as activation, dominance and valence.

8. REFERENCES

[1] C. Busso, M. Bulut, and S.S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds., pp. 110–127. Oxford University Press, New York, NY, USA, November 2013.

[2] A. Tarasov, S. Delany, and C. Cullen, "Using crowdsourcing for labelling emotional speech assets," in *W3C workshop on Emotion ML*, Paris, France, October 2010.

[3] H. Cao, D.G. Cooper, M.K. Keutmann, R.C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, 2014.

[4] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Interspeech 2014*, Singapore, September 2014, pp. 238–242.

[5] E. Mower Provost, Y. Shangguan, and C. Busso, "UMEME: University of Michigan emotional McGurk effect data set," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 395–409, October-December 2015.

[6] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng, "Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks," in *Conference on empirical methods in natural language processing (EMNLP 2008)*, Honolulu, HI, USA, October 2008, pp. 254–263.

[7] J. Ross, L. Irani, M.S. Silberman, A. Zaldivar, and B. Tomlinson, "Who are the crowdworkers?: shifting demographics in Mechanical Turk," in *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, April 2010, CHI EA '10, pp. 2863–2872.

[8] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. To appear, 2015.

[9] R. Rosenthal, "Conducting judgment studies: Some methodological issues," in *The new handbook of methods in nonverbal behavior research*, J. Harrigan, R. Rosenthal, and K. R. Scherer, Eds., pp. 199–234. Oxford University Press, Oxford, UK, May 2008.

[10] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[11] B. Schuller, R. Müeller, B. Höernler, A. Höethker, H. Konosu, and G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," in *9th international conference on Multimodal interfaces (ICMI 2007)*, Nagoya, Aichi, Japan, November 2007, pp. 30–37.

[12] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*, Ph.D. thesis, Universität Erlangen-Nürnberg, Erlangen, Germany, January 2009.

[13] C.M. Lee and S.S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, March 2005.

[14] "Amazon Mechanical Turk," https://www.mturk.com, 2014, Retrieved July 29st, 2014.

[15] L. Nguyen-Dinh, C. Waldburger, G. Troster, and D. Roggen, "Tagging human activities in video by crowdsourcing," in *ACM International Conference on Multimedia Retrieval (ICMR 2013)*, Dallas, TX, USA, April 2013, pp. 263–270.

[16] C. Eickhoff and A.P. de Vries, "Increasing cheat robustness of crowdsourcing tasks," *Information Retrieval*, vol. 16, no. 2, pp. 121–137, April 2013.

[17] S. Buchholz and J. Latorre, "Crowdsourcing preference tests, and how to detect cheating," in *Interspeech 2011*, Florence, Italy, August 2011, pp. 3053–3056.

[18] G. Parent and M. Eskenazi, "Speaking to the crowd: Looking at past achievements in using crowdsourcing for speech and predicting future challenges," in *Interspeech 2011*, Florence, Italy, August 2011, pp. 3037–3040.

[19] Q. Xu, Q. Huang, and Y. Yao, "Online crowdsourcing subjective image quality assessment," in *ACM international conference on Multimedia (ACMMM2012)*, Nara, Japan, October-November 2012, pp. 359–368.

[20] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. To Appear, 2015.

[21] M. Hirth, T. Hoßfeld, and P. Tran-Gia, "Cost-optimal validation mechanisms and cheat-detection for crowdsourcing platforms," in *International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS 2011)*, Seoul, Korea, June-July 2011, pp. 316–321.

[22] M. Marge, S. Banerjee, and A. Rudnicky, "Using the Amazon Mechanical Turk to transcribe and annotate meeting speech for extractive summarization," in *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, California, June 2010, pp. 99–107.

[23] L. von Ahn, M. Blum, N.J. Hopper, and J. Langford, "CAPTCHA: Using hard AI problems for security," in *Advances in Cryptology - EUROCRYPT 2003*, E. Biham, Ed., vol. 2656 of *Lecture Notes in Computer Science*, pp. 294–311. Springer Berlin Heidelberg, Warsaw, Poland, May 2003.

[24] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[25] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27:1–27, April 2011.