INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

A Stepwise Analysis of Aggregated Crowdsourced Labels Describing Multimodal Emotional Behaviors

Alec Burmania and Carlos Busso

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

Abstract

Affect recognition is a difficult problem that most often relies on human annotated data to train automated systems. As humans perceive emotion differently based on personality, cognitive state and past experiences, it is important to collect rankings from multiple individuals to assess the emotional content in corpora, which are later aggregated with rules such as majority vote. With the increased use of crowdsourcing services for perceptual evaluations, collecting large amount of data is now feasible. It becomes important to question the amount of data needed to create well-trained classifiers. How different are the aggregated labels collected from five raters compared to the ones obtained from twenty evaluators? Is it worthwhile to spend resources to increase the number of evaluators beyond those used in conventional/laboratory studies? This study evaluates the consensus labels obtained by incrementally adding new evaluators during perceptual evaluations. Using majority vote over categorical emotional labels, we compare the changes in the aggregated labels starting with one rater, and finishing with 20 raters. The large number of evaluators in a subset of the MSP-IMPROV database and the ability to filter annotators by quality allows us to better understand label aggregation as a function of the number of annotators.

Index Terms: emotion, crowdsourcing, annotation, label aggregation, experimental methods, emotion perception

1. Introduction

Establishing emotional labels for databases is key for the analysis, synthesis and recognition of expressive behaviors. Emotional classifiers rely on the reliability of the labels of the training data. Since the true emotion conveyed on the stimuli is unavailable, these labels are commonly obtained from perceptual evaluations conducted by either many naïve or expert raters. The individual assessments are later aggregated creating consensus labels used as ground truth to describe multimodal emotional behaviors conveyed by the stimuli. This study investigates the consistency of these consensus labels as a function of the number of evaluators.

We have investigated the tradeoff between the number of annotators and the underlying reliability of emotional labels provided by multiple raters [1]. Motivated by the effective reliability concept proposed by Rosenthal for perceptual studies [2], we evaluated difference conditions where categorical emotional labels were aggregated with different numbers of annotations and different level of reliability. We conducted emotion recognition experiments observing differences in classification performance across the chosen conditions. Further analysis demonstrated that the consensus label was not changed by large groups of evaluations, and classification results were not directly related to inter-evaluator agreement. This study made us question the benefit of using more evaluators than what is commonly used in current laboratory-based studies. Is there a benefit in collecting annotations for a given stimuli from more than three or five evaluators? This question is relevant today given the role of crowdsourcing in behavioural signal processing, which offers an affordable, and efficient platform to collect perceptual evaluations from a diverse set of evaluators [3, 4].

This study explores the impact of incrementally adding more evaluators to form the consensus labels. The study uses a set of videos from the publicly available MSP-IMPROV database [5], which was emotionally annotated using crowdsourcing evaluations by over 20 workers. The evaluation includes primary emotions containing happiness, sadness, anger, neutral state, and other. We use majority vote rule to aggregate the labels. This analysis demonstrates that extra evaluations only benefit about 10% of the videos which change labels multiple times. For other cases, the aggregated labels are very stable when we add extra evaluators. This analysis also compares cases where we create the aggregated labels with different number of evaluators (e.g., 3 evaluators versus 20 evaluators). These results show that even for extreme comparisons most of the aggregated labels are consistent. The extra evaluations are effective in certain cases to resolve ties. The findings from this study can guide the design of future evaluations, including selecting the optimal number of evaluators per stimulus.

2. Background

2.1. Relation to Prior Work

Recognizing emotion is a challenging task for both humans and machine learning systems due to the discrepancies between intended and perceived emotions displayed during human interactions [6], especially for ambiguous expressive behaviours [7]. Traditionally ground-truth for emotion recognition is represented by a set of labels provided by experts (i.e. behavioural analysts) or by a large amount of naïve evaluators. The former is expensive and difficult to achieve, while the latter can be time consuming. Crowdsourcing services provide perfect platforms to effectively complete these tasks fast and at a reasonable cost [8]. In fact, several studies have evaluated emotional databases using crowdsourcing services [4, 9, 10]. Given the reduced cost per annotation, it is feasible to collected more annotations per stimulus (e.g., 10) than equivalent studies conducted on laboratory conditions (e.g., 3).

After collecting labels from multiple evaluators, these labels are aggregated to form consensus labels. While there are sophisticated methods for this task, such as the minimax conditional entropy framework [11], a simple method is majority vote. For categorical labels (e.g., anger, happiness and sadness),

This work was funded by NSF CAREER award IIS-1453781.

http://dx.doi.org/10.21437/Interspeech.2017-1278

majority vote selects the class that receives more preferences, where tiebreaks may be necessary. This study uses majority vote due to its simplicity. It aso allows us to make controlled observations about our performance without taking into consideration the effect of quality or quantity on our aggregation metric.

These consensus labels are often used for classification - a main application of this work which is more explicitly defined in Burmania et al. [1]. Bhardwaj et al. [12] suggested in their work that often inter-evaluator agreement does not necessarily correlate with classification performance. They suggest a framework (Anveshan) which solves a similar problem - fusing label data beyond inter-evaluator agreement for the purpose of aggregating a label for natural language processing. Our work can be seen as somewhat of an analog focusing on effective reliability, while moving into the multi-modal domain and providing additional iterative analysis.

An important aspect in these evaluations is to determine the number of evaluators. While previous studies based on laboratory conditions were limited to few evaluators, the increased role of crowdsourcing in perceptual evaluations has allowed researchers to significantly increase the diversity and quantity of annotations. How many annotators are enough to derive reliable labels? Knowing the aggregated method in advance can help answer this question. For example, Zhang et al. [13] proposed an *active learning* (AL) framework, where a stimuli is evaluated until a level of agreement is reached, achieving the same aggregated label from majority vote, but avoiding extra annotations. For example, if five evaluators per stimulus is the target, the evaluation can stop if the first three annotators agree on the label.

The underlying assumption in Zhang et al. [13] illustrates the reduced benefits of adding additional evaluations. Given the flexibility provided by crowdsourcing, it is important to understand the incremental contribution of additional annotators. Understanding this question is fundamental to correctly allocate resources for perceptual evaluations. This study explores this question using a stepwise analysis, where we compare aggregated labels obtained with majority vote after incrementally adding additional raters.

2.2. MSP-IMPROV Database

The analysis relies on the publicly available MSP-IMPROV corpus, an emotional audiovisual database consisting of improvised dyadic interactions between actors [5]. The corpus was originally collected to create conversational sentences conveying the same lexical information across different emotions. This goal was achieved by creating improvisation scenarios carefully designed to (1) lead one of the actor to utter the target sentence, and (2) elicit the target emotions. The target emotions for the corpus were happiness, sadness, anger, and neutrality. The details of the corpus are given in Busso et al. [5].

The MSP-IMPROV corpus is ideal for this analysis since each sentence was perceptually evaluated by at least five annotators using Amazon's Mechanical Turk (Burmania et al. [3] gives the details of the evaluation). Over 1,000 evaluators participated in the study. The evaluation was conducted using an elegant framework which tracked the performance of the workers in real time, stopping the evaluation when the quality of their labels dropped below a given threshold. The evaluation interleaved reference sentences (e.g., gold standard), for which emotional evaluations were previously collected. These videos include 652 recordings, which were used to assess the quality of the labels. Given this protocol, all of the videos from this reference set have more than five evaluations per video. This study only considers this set. The perceptual evaluation contained a multiple choice question asking for the primary emotion among happiness, sadness, anger, neutral state, and other. The class 'other' was included to avoid forcing the evaluators to choose a class that does not represent the perceived emotion. Russell et al. [14] discussed the pitfalls of forced selections in emotional perceptual evaluation in depth. This study only considers the results for the primary emotions.

2.3. Moderate and High Agreement Conditions

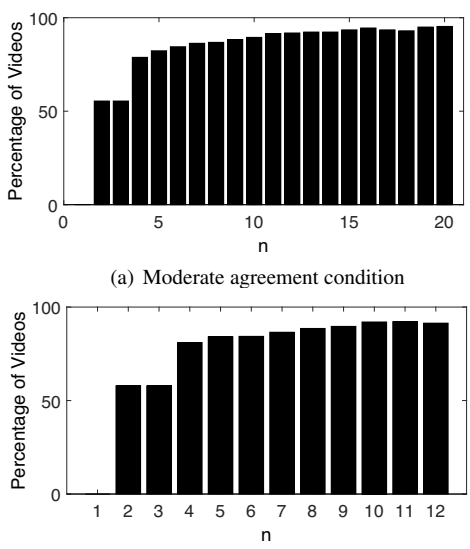
This study includes two conditions: *moderate agreement* and *high agreement*. The moderate agreement condition includes all the annotations collected in the evaluation, achieving a Fleiss' Kappa statistic around $\kappa = 0.4$ (the value for κ depends on the number of evaluators which varies across the videos in the reference set). We consider 599 sentences for this condition, which have more than 20 evaluations per video. The high agreement condition includes a subset of the annotations, selected by post-processing the labels. Burmania et al. [3] used a post-processing step to remove noisy evaluations. If the quality of the worker was above the threshold at a given checkpoint, he/she was allowed to complete 20 additional videos before his/her quality was controlled again. If his/her quality dropped below acceptable values, not only the evaluation was stopped, but also the annotations provided after his/her last successful checkpoint were discarded. This study uses a similar framework to define labels in the high agreement condition, where the threshold for the cosine agreement metric is set at $\Delta\theta_t^s = -25^\circ$ (see Burmania et al. [3]). While this framework discards annotations, we still have roughly 12-15 evaluations per video in this condition. For this set, the Fleiss' Kappa statistic is around $\kappa = 0.45$. We consider 638 sentences for this condition, which have more than 12 evaluations per video. Stricter thresholds increase the quality above $\kappa = 0.55$, but the videos with more than 15 evaluators would be less than 246.

The large amount of evaluations for moderate and high agreement conditions allows us to explore the trends that occur through the labelling process, and allow us to conjecture about best practices for crowdsourced perceptual evaluations.

3. Methodology for Analysis

Starting with one annotator per video, the study incrementally adds new evaluators analyzing the impact on the aggregated labels. As mentioned before, we use the reference set from the MSP-IMPROV corpus. The labels are aggregated using majority vote, where the class with more selections is selected. We consider the primary emotions consisting in a five-class problem: happiness, sadness, anger, neutral state, and other.

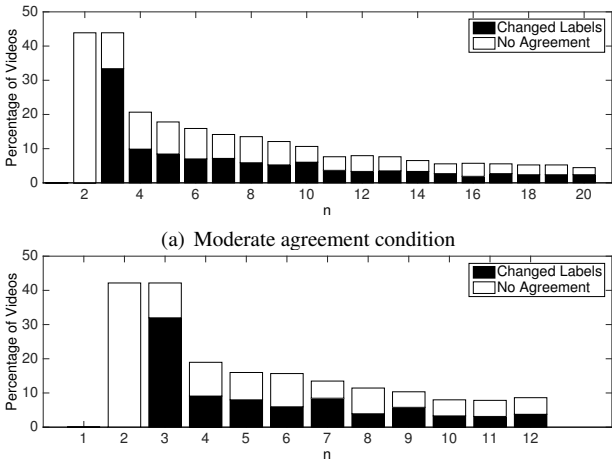
Although the annotations for the MSP-IMPROV database had already been collected, we simulate labels arriving to us one at a time for each video in the database. In this simulation, the order of the annotators correspond to the actual order in which the labels were collected in the database itself. For the moderate and high agreement conditions, we start with one evaluation, corresponding to the first label assigned to each video. The aggregated labels for this initial case are the actual labels. Then, we include an additional evaluator corresponding to the second annotation provided to the videos. This simulation continues, until arriving at the maximum number of annotations per condition ($n_{\max}$). For the moderate agreement condition, $n_{\max}$ is set to 20. For the high agreement condition, $n_{\max}$ is set to 12, since 638 videos have at least 12 annotators after removing the



(a) Moderate agreement condition

(b) High agreement condition

Figure 1: *Percentage of videos with the same aggregated labels before and after adding an additional evaluator.*



(a) Moderate agreement condition

(b) High agreement condition

Figure 2: *Percentage of the videos in which their labels changed as we add one extra evaluator. Black bars correspond to transitions from emotional class to no agreement (EmoA → NA), and black bars correspond to transitions between no agreement to an emotional class (NA → EmoA).*

noisy labels with the post-processing method described in Section 2.3.

This stepwise analysis is proposed to understand the contribution of an additional evaluator. There are four possible scenarios for the aggregated labels:

- EmoA → EmoA: No change on the selected emotional class. The selected emotion is not affected by adding an extra evaluator.
- EmoA → NA: Going from a selected emotional class to *no agreement* (NA). After adding an extra evaluator, there are two or more emotional classes with the same number of selections.
- NA → EmoA: Going from no agreement to a emotional class. The new evaluation resolves a tie.
- NA → NA: Going from no agreement to no agreement. The new evaluation does not resolve a tie.

Notice that with the proposed method, it is not possible to go from one emotion to another emotion in a single step. It has to go through a state of no agreement first.

4. Results of the Analysis

4.1. Stability of Aggregated Labels

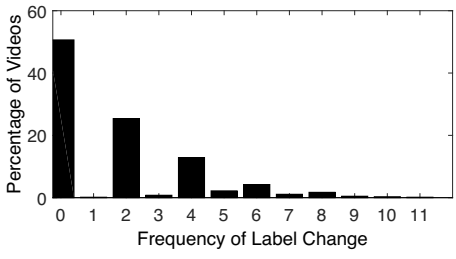
The first part of the analysis considers the stability of the aggregated labels as we incorporate more annotators. We use the annotations provided by $n-1$ annotators to determine the aggregated labels by using majority vote rule. Then, we estimate the aggregated labels after adding the additional annotation (i.e., n raters). Figure 1 shows the percentage of the videos in which their labels remain unchanged with one more rater. This case includes *EmoA→EmoA* and *NA→NA* (Sec. 3). Figure 1 shows that about 60% of the aggregated labels remain the same when $n = 2$ or $n = 3$. We observe this result for moderate and high agreement conditions. After $n = 4$, we observe less than 20% of the labels changed due to the new annotation. After $n = 6$, the changes in aggregated labels are less than 10%. The results are consistent for moderate and high agreement conditions.

In Figure 2, we explore how the aggregated labels of the videos changed as we add new evaluators (e.g., n increases). This analysis includes the cases *EmoA→NA* and *NA→EmoA*, complementing the results shown in Figure 1. White bars represents changes from an emotional class to no agreement (e.g., ties). Black bars represents changes from no agreement to an emotional class (e.g., resolving ties). For example, adding a second annotation which does not agree with the first annotation will create ties (around 40% for both conditions). When $n = 3$, most of these ties are resolved. The number of labels without agreement is below 10% for $n > 6$. The patterns for moderate and high agreement conditions are similar.

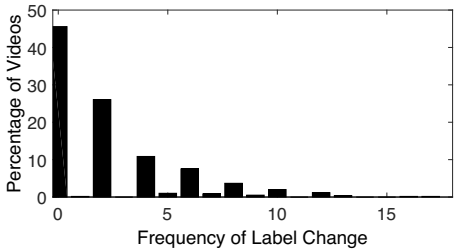
We hypothesize that most of the changes in the aggregated labels correspond to videos with ambiguous emotions. We explore this hypothesis by counting the number of times that each video changed labels as we increase n from 1 to n_{max}. Figure 3 shows the distribution, where the x-axis corresponds to the frequency that a video changed labels. Notice that the high occurrence of even number of transitions indicates that the labels of the videos oscillate from no agreement to emotional categories, but eventually converge to a given emotional class. For example, changing the label three times implies a transition between one emotion to no agreement (*EmoA→NA*) followed by a transition between no agreement to the same or different emotion (*NA→Emo*). A key result from this figure is that between 45% and 50% of the videos do not change labels during the process. The aggregated labels are the labels assigned by the first evaluators. These are videos where evaluators are consistent in assigning the emotional class, reflecting clear emotions on the videos. For both conditions, 75% of the videos change labels less than four times. About 10% of the videos change labels multiple times confirming our hypothesis. The key benefit in collecting extra evaluators per video is to provide better characterization of the range of emotional content conveyed in videos with ambiguous emotions.

4.2. Aggregated Labels with Different Number of Raters

We also compare the aggregated labels obtained by using the typical number of evaluators per stimulus used in previous studies. The purpose of this analysis is to quantify the differences in aggregated labels between different values of n. Notice that this section does not follow the stepwise approach used in Sec-



(a) Moderate agreement condition



(b) High agreement condition

Figure 3: *This figures shows the percentage of the videos in which their aggregated labels changed m times ($m \in \{1, 2, \ldots\}$), as we incrementally add new evaluators from 1 to n_{max}. The x-axis represents the number of changes in the aggregated labels (e.g., m).*

tion 4.1. For the moderate agreement condition, we evaluate $n \in \{3, 5, 9, 20\}$. For the high agreement condition, we evaluate $n \in \{3, 5, 9\}$, since we do not have 20 evaluations per video. In addition to the four scenarios presented in Section 3, this analysis can also have the following transition, which was not possible with the stepwise approach:

- EmoA $\rightarrow$ EmoB: Change from one emotional class to another.

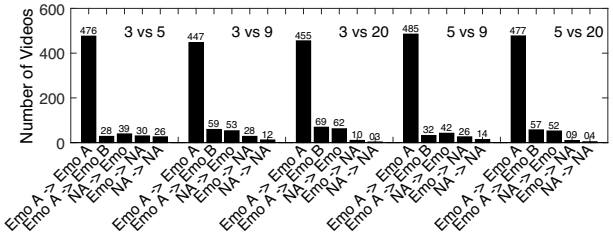
Figure 4 gives the results, presenting pairwise comparison between some of these cases (e.g., $n = 3$ versus $n = 5$). We observe that the labels are very stable even when we compare $n = 3$ versus $n = 20$ (Fig. 4(a)). In this case, only 24% of the videos changed the aggregated labels. Most of the cases correspond to ties resolved by further annotators (10.4%) or changes from one emotion to another (11.5%). This result also indicates that only few videos benefits from having more annotations.

Unlike the stepwise analysis, we observe a difference between moderate and high agreement conditions. In Figure 4(a) (moderate agreement condition), the number of videos that do not change labels is around 468, which represents 78.1% of the videos considered for this condition. In Figure 4(b) (high agreement condition), the number of videos that do not change labels is around 514, which represents 80.6% of the videos considered for this condition. We observe higher consistency between conditions when the inter-evaluator agreement is higher.

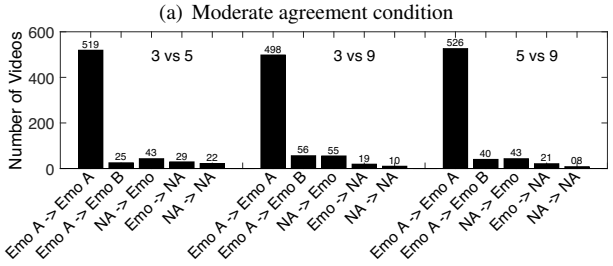
5. Discussion and Conclusions

This study presented a stepwise approach to analyze the incremental contribution of additional annotators for emotional perceptual evaluation. Given the role of crowdsourcing in behavioral signal processing, this analysis can guide the design of future evaluations.

Figure 1 shows that collecting five annotators per video resolves most of the ambiguity. After this point, a reduced number of videos changed labels as we consider additional evaluators. Figure 3 shows that an important percentage of the videos do not change labels at all during the stepwise analysis. Very few videos change labels more than five times. It is interesting to observe this level of stability in the aggregated labels, given the low inter-evaluator agreement associated with emotional perceptual evaluations. We hypothesize that the labels are even more stable for simpler behavioral signal processing evaluations conducted on crowdsourcing services.



(a) Moderate agreement condition



(b) High agreement condition

Figure 4: *Pairwise comparison of aggregated labels obtained by considering different number of evaluators.*

The analysis consistently showed that most of the benefits in adding extra evaluations is for a small set of videos with ambiguous emotions (about 10% in this study). An important question is to assess the impact of these changes in labels on emotion recognition experiments. Our previous study where we analyzed classification experiments at different levels of reliability in the labels suggested that the difference in classification performance will be small [1].

In general, the stepwise analysis shows similar patterns for moderate and high agreement conditions. The effect of higher quality become clearer when we examine the differences in aggregated labels between two distant values of n. The analysis shows that around 80% of the labels for the high agreement condition remain the same even when we compare cases with three and nine evaluators. The extra annotations help to clarify ambiguities on the remaining set, resolving ties and correcting emotional classes. While majority vote is a reasonable rule, it may be interesting to replicate the analysis using more sophisticated aggregated criteria which take into consideration the difficulty of the task and the reliability of the evaluators [11]. We expect that these methods will help to reduce the amount of changes in the labels, suggesting that fewer evaluators may be required.

A limitation of this study is that the absolute results may depend on the given task (corpus, number of emotional categories, reliability of the evaluators). However, we expect that the key observations from this study will generalize to relative performance observations in new domains: most of the agreements are reached with only few annotations, and the benefits of extra labels is to clarify ambiguous samples. These observations have implications on machine learning tasks trained on these labels including classifiers, and opening promising directions on active learning frameworks.

6. References

[1] A. Burmania, M. Abdelwahab, and C. Busso, "Tradeoff between quality and quantity of emotional annotations to characterize expressive behaviors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5190–5194.

[2] 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. Oxford, UK: Oxford University Press, May 2008, pp. 199–234.

[3] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[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] 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. 8, no. 1, pp. 67–80, January-March 2017.

[6] C. Busso and S. Narayanan, "The expression and perception of emotions: Comparing assessments of self versus others," in *Interspeech 2008 - Eurospeech*, Brisbane, Australia, September 2008, pp. 257–260.

[7] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, Amsterdam, The Netherlands, September 2009, pp. 1–8.

[8] 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.

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

[10] 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.

[11] D. Zhou, Q. Liu, J. Platt, and C. Meek, "Aggregating ordinal labels from crowds by minimax conditional entropy," in *International Conference on Machine Learning (ICML 2014)*, Beijing, China, June 2014, pp. 262–270.

[12] V. Bhardwaj, R. Passonneau, A. Salleb-Aouissi, and N. Ide, "Anveshan: A framework for analysis of multiple annotators' labeling behavior," in *Proceedings of the Fourth Linguistic Annotation Workshop (LAW 2010)*, Uppsala, Sweden, July 2010, pp. 47–55.

[13] Y. Zhang, E. Coutinho, Z. Zhang, C. Quan, and B. Schuller, "Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions," in *International conference on Multimodal interaction (ICMI 2015)*, Seattle, WA, USA, November 2015, pp. 275–278.

[14] J. A. Russell, "Forced-choice response format in the study of facial expression," *Motivation and Emotion*, vol. 17, no. 1, pp. 41–51, March 1993.