MuSE-ING ON THE IMPACT OF UTTERANCE ORDERING ON CROWDSOURCED EMOTION ANNOTATIONS

Mimansa Jaiswal*, Zakaria Aldeneh*, Cristian-Paul Bara*, Yuanhang Luo*, Mihai Burzo†,
Rada Mihalcea, Emily Mower Provost*
*University of Michigan–Ann Arbor
†University of Michigan–Flint

ABSTRACT

Emotion recognition algorithms rely on data annotated with high quality labels. However, emotion expression and perception are inherently subjective. There is generally not a single annotation that can be unambiguously declared "correct." As a result, annotations are colored by the manner in which they were collected. In this paper, we conduct crowdsourcing experiments to investigate this impact on both the annotations themselves and on the performance of these algorithms. We focus on one critical question: the effect of context. We present a new emotion dataset, Multimodal Stressed Emotion (MuSE), and annotate the dataset using two conditions: randomized, in which annotators are presented with clips in random order, and contextualized, in which annotators are presented with clips in order. We find that contextual labeling schemes result in annotations that are more similar to a speaker's own self-reported labels and that labels generated from randomized schemes are most easily predictable by automated systems.

Index Terms— emotion, crowdsourcing, annotation, emotion perception, classifier performance

1. INTRODUCTION

Emotion technologies, both recognition and synthesis, are heavily dependent on having reliably annotated emotional data, annotations that describe the observed emotional display. The hope is often that these annotations capture the speaker's true underlying state. Yet, in practice, this true *felt sense* emotion is unknown, and researchers must resort to manual labeling of data. The hope is that these manual labels are sufficiently "correct" to enable the training and evaluation of emotion technologies. One method of ensuring quality labels has been to require the participation of expert raters. However, it can be both expensive and time consuming to hire expert raters. More recently, researchers have embraced crowdsourcing services (e.g., *Amazon Mechanical Turk*) to efficiently collect annotations from non-expert workers in a cost-effective and timely manner [1]. Once collected, annotations from non-expert workers are aggregated to form ground-truth labels that are used for training and evaluating automated systems. However, the method through which these annotations are collected can profoundly impact the behavior of the annotators. In this paper, we study how the setup of a crowdsourcing task can influence both the collected emotion labels as well as the performance of classifiers trained using these labels.

The effective use of crowdsourcing for collecting reliable emotion labels has been an active research topic. Burmania et al. investigated the trade-off between the number of annotators and underlying reliability of the annotations [2]. Other work has looked at quality-control techniques to improve the reliability of annotations. For example, Soleymani et al. used qualification tests to filter out spammers and retain high-quality annotators [1]. Burmania et al. investigated the use of gold-standard samples to monitor annotators' reliability and fatigue [3].

However, variability also results from context, relevant past information that provides cues as to how to interpret an emotional display. Context, such as tone, words, expressions can affect how individuals perceive emotion [4]. Context is also implicitly included in the labeling schemes of many of the most common emotion datasets (e.g., IEMOCAP [5] and MSP-Improv [6]) because annotators rate each utterance (or time period) in order. That means that annotators are influenced by information that they recently observed [7]. However, emotion recognition systems are often trained over single utterances [8–11], leading to a mismatch in the information available to annotators and to classification systems.

In this work, we study the difference between annotations obtained for audio clips when emotional displays are presented to annotators with context and when presented randomly. In both cases, annotators are affected by the emotion displays that they have recently observed [12, 13]. However, only in the contextual presentation there is also a cohesive story. We investigate the following research questions:

- Q1: Is there a significant difference between annotations obtained from random and contextual presentations?

- Q2: Are annotations obtained from contextual presentations more similar to a speaker's own self-reported labels than those from random presentations?

- Q3: Is there a significant difference between the inter-rater agreements obtained from random and contextual presentations?

- Q4: How does the performance of an emotion recognition system, operating on single utterances, vary given annotations obtained from random and contextual presentations?

- Q5: How does the performance gain of an emotion recognition system operating across multiple utterances vary given different amounts of context (defined as number of prior utterances) and labels obtained from random and contextual presentations?

This paper is organized as follows. First, we introduce the dataset and explain the design, collection and post-processing procedures. Then, we present an analysis of the dataset and the collected corpus labels. We then present the results of a state-of-the-art speech-based emotion classification system trained on the random presentation vs. the contextual presentation labels. The findings from this work will provide insight into performance implications of emotion recognition system given mismatches between the amount of context provided to the annotators generating the labels and the ultimate classification system.

2. DATASET

2.1. Data Collection

We introduce the Multimodal Stressed Emotion (MuSE) dataset, designed to understand how stress and emotion interplay in spoken communication. The dataset consists of fifty-five recordings from twenty-eight participants, each recorded under two conditions, stressed and not-stressed (one subject participated in only the stressed condition). The stress condition was recorded during the final exam period at the University of Michigan, the not-stressed condition was recorded after exams concluded. The emotion component was generated through video stimuli, sampled from [14] and [15] and through emotionally evocative monologue topics [16]. The data used in this study are a subset of the corpus. Table 1 shows the questions used to evoke emotions, which fall under following sections: (a) icebreaker; (b) non-neutral (c) non-neutral; (d) non-neutral; (e) ending. The non-neutral sections (b), (c) and (d) were presented in random order using prompts from each of the categories: positive, negative, and intensity. In each case, one question was used in the stress recording and the other was used in the non-stress recording. Each participant was asked to rate his/her emotions after the completion of each section using the scales of activation (calm vs. excited) and valence (positive vs. negative). We refer to these annotations as the *self-report annotations*.

2.2. Data Preprocessing

The monologues in each section are divided into utterances. However, since the monologues are spontaneous, often there is not a clear sentence boundary. We create utterances by identifying prosodic or linguistic boundaries in spontaneous speech as defined by [17]: (a) a clear sentence boundary (full stop or exclamation); (b) a change of context after filler words, or revision of sentence; (c) an extended pause (i.e., a silence greater than three seconds); or (d) filler or example words instead of a full stop.

The dataset contains 2,648 utterances with a mean duration of 12.44 ± 6.72 seconds (Table 2). The mean length of stressed utterances (11.73 ± 5.77 seconds) is significantly different from that of the non-stressed utterances (13.30 ± 6.73 seconds).

We perform data selection, excluding utterances that are shorter than 3-seconds and longer than 35-seconds (2.8% of the original data). This is because short segments may not have enough information to capture emotion, and longer segments can have variable emotion. This results in 2,574 utterances.

2.3. Crowdsourcing

We posted our experiments as Human Intelligence Tasks (HITs) on *Amazon Mechanical Turk*. HITs were defined as sets of utterances in either the contextual or random presentation condition. In each condition, workers were presented with a single utterances and were asked to annotate the activation and valence values of that utterance using Self Assessment Manikins [18]. Once completed, the worker was presented with a new HIT and could not go back to revise a previous estimate of emotion. This annotation strategy is different than the one deployed in [19],where the workers could go back and re-evaluate utterances.

In the randomized experiment, each HIT is an utterance from any section, by any speaker, from any session and all HITs appear in random order. So, a worker might see the first HIT as *Utterance 10 from Section 3 of Subject 4's stressed recording* and see the second HIT as *Utterance 1 from Section 5 of Subject 10's non-stressed recording*. This setup ensured that the workers couldn't condition to any speaker's specific style or contextual information.

Table 1. Emotion elicitation questions.

Icebreaker
1. Given the choice of anyone in the world, whom would you want as a dinner guest?
2. Would you like to be famous? In what way?

Positive
1. For what in your life do you feel most grateful?
2. What is the greatest accomplishment of your life?

Negative
1. If you could change anything about the way you were raised, what would it be?
2. Share an embarrassing moment in your life.

Intensity
1. If you were to die this evening with no opportunity to communicate with anyone, what would you most regret not having told someone?
2. Your house, containing everything you own, catches fire. After saving your loved ones and pets, you have time to safely make a final dash to save any one item. What would it be? Why?

Ending
1. If you were able to live to the age of 90 and retain either the mind or body of a 30-year old for the last 60 years of your life, which would you choose?
2. If you could wake up tomorrow having gained one quality or ability, what would it be?

In the contextual experiment, we posted each HIT as a collection of ordered utterances from a section of a particular subject's recording. Because each section's question was designed to elicit a particular emotion, we still posted the HITs in a random order over sections from all subjects. This prevented workers from conditioning to the speaking style of an individual participant. For example, a worker might see the first HIT as *Utterance 1...N from Section 3 of Subject 4's stressed recording* and see the second HIT as *Utterance 1...M from Section 5 of Subject 10's non-stressed recording* where *N, M* are the number of utterances in those sections respectively.

We recruited from a population of workers in the United States who are native English speakers, to reduce the impact of cultural variability. We ensured that each worker had $> 98\%$ approval rating and number of HITs approved as > 500. We ensured that all workers understood the meaning of activation and valence using a qualification task that asked workers to rank emotion content. The workers were asked to select, given two clips, which clip had the higher valence and which had the higher activation. The options were chosen from a set including: (1) a speaker in low activation, high valence state and (2) a speaker in high activation, low valence state.

We assigned each HIT to eight workers. All HIT workers were paid a minimum wage ($9.25/hr), pro-rated to the minute. We removed and re-posted assignments where the worker completed the assignment in time shorter than the audio length. The ground-truth for each utterance was formed by taking the average of the eight annotations.

3. EXPERIMENTAL SETUP

Acoustic Features. We extract acoustic features using OpenSmile [20] with the eGeMAPS configuration [21]. The eGeMAPS

Table 2. Data summary (R:random, C:context, F:female, M:male).

Monologue Subset	
Mean num of utterances/monologue	9.69 ± 2.55
Mean duration of utterances	12.44 ± 6.72 seconds
Total num of utterances	2,648
Selected num of utterances	2,574
Gender distribution	19 (M) and 9 (F)
Total annotated speech duration	~ 10 hours

Crowdsourced Data	
Num of workers	160 (R) and 72 (C)
Blocked Workers	8
Mean activation	3.62 ± 0.91 (R) 3.69 ± 0.81 (C)
Mean valence	5.26 ± 0.95 (R) 5.37 ± 1.00 (C)

feature set consists of 88 utterance-level statistics over the low-level descriptors of frequency, energy, spectral, and cepstral parameters. We perform speaker-level z-normalization on all features.

Static Network Setup (Hypothesis 4). We train and evaluate four Deep Neural Networks (DNN) models: {random, contextual} × {valence, activation}. In all cases, we predict the continuous annotation using regression. For each network setup, we follow a five-fold evaluation scheme and report the average RMSE across the folds. For each test-fold, we use the previous fold for hyper-parameter selection and early stopping. The hyper-parameters include: number of layers $\{2, 3, 4\}$ and layer width $\{64, 128, 256\}$. We use ReLU activation and train the networks with MSE loss using Adam optimizer.

Dynamic Network Setup (Hypothesis 5). We use Gated Recurrent Unit networks (GRU). The hyper-parameters are: number of layers $\{1, 2\}$ and layer width $\{64, 128, 256\}$. We pass the GRU output of the last time step through a regression layer to get the final outputs. We train the networks with MSE loss using Adam optimizer.

Network Training. We train our networks for a maximum of 100 epochs and monitor the validation loss after each epoch. We stop the training if the validation loss does not improve for 15 consecutive epochs. We revert the network's weights to those that achieved the lowest validation loss during training. Finally we train each network five times and average the predictions to reduce variance due to random initialization.

4. RESULTS AND ANALYSIS

4.1. Question 1

Hypothesis: *Human annotations collected through randomized labeling are significantly different from those collected through contextualized labeling.* Prior work has shown context effects emotion perception [7], even when observers are explicitly asked not to take it under consideration [22, 23]. Hence, we believe that context provided by previous utterances would lead to a change in perception of a particular utterance. Tables 3 and 4 (sets of significantly different means are bolded (t-test, $p < 0.01$)) show the mean activation and valence, for the random and contextualized labeling schemes, grouped by condition and question, respectively. Table 3 shows that, for non-stress conditions, the mean of the activation ratings obtained through contextual labeling is significantly higher than that obtained through random labeling. The table also shows that, for both stress and non-stress conditions, the valence means obtained through contextual labeling are significantly higher than those obtained through random labeling. Table 4 shows that, although the mean valence and activation values were consistently different for the labelling schemes across all emotion elicitation techniques, the differences

Table 3. Mean activation and valence values obtained from the two crowdsourcing labeling schemes (random and context) grouped by speaker condition (stress and non-stress).

	Activation		Valence	
	Random	Context	Random	Context
Stress	3.63	3.59	**5.27**	**5.36**
Non-Stress	**3.61**	**3.79**	**5.26**	**5.39**

Table 4. Mean activation and valence values obtained from the two crowdsourcing labeling schemes (random and context) grouped by emotion elicitation question.

	Activation		Valence	
	Random	Context	Random	Context
Icebreaker	3.55	3.60	**5.41**	**5.61**
Positive	3.64	3.71	5.11	5.13
Negative	**3.57**	**3.67**	**5.40**	**5.55**
Intensity	**3.64**	**3.74**	**5.17**	**5.31**
Ending	3.69	3.71	**5.23**	**5.29**

were significant in some elicitation techniques and not in others.

4.2. Question 2

Hypothesis: *Annotations of outside observers are more similar to self-annotations in the contextual case, compared to the randomized case.* Path models [24] suggest that subjective voice variation, from the established mental baseline accounts for much of the variance in emotion inference. Hence, emotion inference is aided with more cues about the speech patterns that are more readily provided through context. Figure 1 shows the absolute differences between the mean crowdsourced labels (valence and activation, each for random and contextual schemes) and self-reported scores as a function of utterance position. The figure shows that contextual labels have consistently lower absolute differences, compared to self-reported labels, than the random labels. A paired t-test shows that these differences between the contextual and random labels are significant ($p < 0.01$) for both valence and activation.

Our results suggest that crowdsourced emotion labels collected with access to contextual information are closer to self-reported emotion labels. Our results further suggest that these differences are consistent across recording conditions (Table 5) and emotion elicitation questions (Table 6, sets of significantly different means are bolded, t-test, $p < 0.01$).

4.3. Question 3

Hypothesis: *Individual annotators differ in annotation similarity in the contextual presentations, compared to the randomized presentation.* Joseph et al. in [25] show that while insufficient context results in noisy and uncertain annotations, an overabundance of context may cause the context to outweigh other signals and lead to lower agreement. Further, contextual information biases different people differently on both temporal and intensity metrics [26, 27]. Our results highlight the impact of context: the agreement is significantly higher in the case of labels obtained from the randomized presentations, compared to the contextualized presentations: (1.55 vs. 1.62) for activation and (1.07 vs. 1.14) for valence. This trend holds true for all experimental design setups i.e. {random, contextual} × {valence, activation} and {random, contextual} × {icebreaker, positive, negative, intensity and ending}. As shown in Tables 3 and 4, the labels obtained in both cases are significantly different due to context-based conditioning. However, the conditioning may not impact the labels consistently across all workers, which may lead to lower inter-

Table 5. Mean difference between the self-reported activation and valence ratings from the two labeling schemes (random and context) grouped by speaker condition (stress and non-stress).

	Activation		Valence	
	Random	Context	Random	Context
Stress	**2.03**	**1.96**	**1.20**	**1.14**
Non-Stress	**1.82**	**1.67**	**1.20**	**1.12**

Table 6. Mean difference between the self-reported activation and valence ratings from the two labeling schemes (random and context) grouped by emotion elicitation question.

	Activation		Valence	
	Random	Context	Random	Context
Icebreaker	1.81	1.80	**0.97**	**0.85**
Positive	**1.89**	**1.74**	1.14	1.11
Negative	**1.96**	**1.76**	**1.18**	**1.07**
Intensity	**2.19**	**2.08**	**1.49**	**1.44**
Ending	**1.81**	**1.73**	1.23	1.28

annotator agreement values. This suggests that it may be beneficial to consider the distribution of annotations as ground-truth, rather than averaging labels, which presumes that the impact of conditioning is consistent across all workers [28].

4.4. Question 4

Hypothesis: *A static classifier will perform better when trained and evaluated using labels annotated with a randomized presentation, compared to a contextualized presentation.* Prior studies have shown that it is easier to classify data with less target variation [29] and matched classifier input, which in our case is labels obtained from the random labelling presentation (the classifier processes single utterances at a time, no context).

We test this hypothesis by training and evaluating classifiers for the four possible setups: $\{random, contextual\}$ x $\{valence, activation\}$. The classifier is described in Section 3. We find that the RMSEs are lower for the contextual labels in the case of activation (0.91 vs. 1.00) while the errors are lower for the random labels in the case of valence (1.13 vs. 1.20). Using a paired t-test, we find that the differences in errors are significant in the case of valence but not activation. These findings suggest that classification performance is impacted by the labelling methodology, but that this effect may depend on emotion dimension.
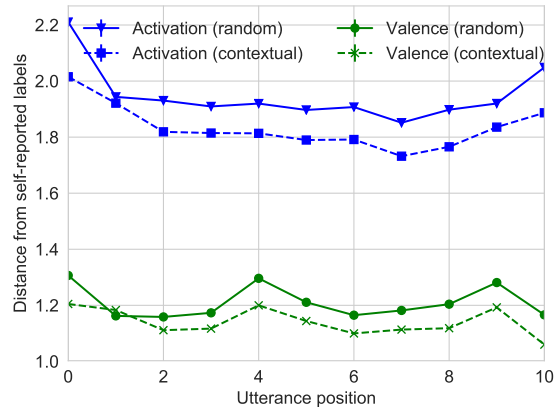
Prior work has demonstrated the importance of considering long-term context when predicting valence (the same effect has not been shown in activation) [30]. The contextual annotations provided the annotators with this information, but the classifier could not take advantage of this effect. This mismatch may have contributed to the relatively lowered performance of the valence classifier, compared to the activation classifier.

4.5. Question 5

Hypothesis: *We anticipate that systems trained on contextualized labels will see greater increases in performance as the amount of provided context increases.* This finding would support results in the literature regarding the ordinal nature of emotion perception [7] and previous works in emotion recognition that have demonstrated that context can influence the performance of emotion classifiers [30].

The classifier is described in Section 3. We test this hypothesis by using the contextual annotations in one classifier and the non-contextual (random) annotations for the other classifier. We select a subset of utterances in each section that have at least five consecutive utterances before them (59% of the original data). The initial

Fig. 1. Mean difference between the self-reported activation and valence ratings and the random and contextual presentations.

Table 7. Relative improvement in RMSE (%) obtained for each additional previous utterance, comparing random and contextual labels.

	Activation		Valence	
Past steps	Random	Context	Random	Context
0	-	-	-	-
1	$+1.96\%$	$+1.24\%$	$+0.85\%$	$+3.32\%$
2	$+2.28\%$	$+2.93\%$	$+5.23\%$	$+7.63\%$
3	$+3.36\%$	$+8.72\%$	$+6.08\%$	$+8.43\%$
4	$+4.41\%$	$+10.5\%$	$+8.23\%$	$+8.36\%$

classifier is trained without temporal context (but with the contextualized labels). We incrementally increase the number of past utterances (from zero to five). We run this for every task combination and report the results in Table 7.

Table 7 shows the performance gains after incrementally adding the past utterance, relative to the baseline performance. The addition of past utterances improves the performance over baseline for all setups. Where using contextual labels, however, the performance gains are generally higher than the gains obtained after using random labels. Our results suggest that it is necessary to consider the mismatch the amount of context provided to the annotators generating the labels and the ultimate classification system.

5. CONCLUSION

In this work we showed that the amount of context provided to annotators when assigning emotion labels affects both the annotations themselves and the performance of classifiers using these annotations. We also studied the implications of a mismatch between annotation context and classifier context on classifier performance. For future work, we will analyze the effect of context given multimodal information and the differences in perception of emotion expression in stress vs. non-stressed situations.

References

[1] Mohammad Soleymani and Martha Larson, "Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus," *SIGIR-Workshops*, 2010.

[2] Alec Burmania, Mohammed Abdelwahab, and Carlos Busso, "Tradeoff between quality and quantity of emotional annotations to characterize expressive behaviors," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5190–5194.

[3] Alec Burmania, Srinivas Parthasarathy, and Carlos Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, 2016.

[4] Debi Laplante and Nalini Ambady, "On how things are said: Voice tone, voice intensity, verbal content, and perceptions of politeness," *Journal of Language and Social Psychology*, vol. 22, no. 4, pp. 434–441, 2003.

[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[6] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily 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, 2017.

[7] Georgios N Yannakakis, Roddy Cowie, and Carlos Busso, "The ordinal nature of emotions," in *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*. IEEE, 2017, pp. 248–255.

[8] Zakaria Aldeneh and Emily Mower Provost, "Using regional saliency for speech emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2741–2745.

[9] Mohammed Abdelwahab and Carlos Busso, "Incremental adaptation using active learning for acoustic emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5160–5164.

[10] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2227–2231.

[11] Mousmita Sarma, Pegah Ghahremani, Daniel Povey, Nagendra Kumar Goel, Kandarpa Kumar Sarma, and Najim Dehak, "Emotion identification from raw speech signals using dnns," *Proc. Interspeech 2018*, pp. 3097–3101, 2018.

[12] Emilie Qiao-Tasserit, Maria Garcia Quesada, Lia Antico, Daphne Bavelier, Patrik Vuilleumier, and Swann Pichon, "Transient emotional events and individual affective traits affect emotion recognition in a perceptual decision-making task," *PloS one*, vol. 12, no. 2, pp. e0171375, 2017.

[13] James A Russell, "Emotion recognition: Is it universal?," 2017.

[14] James J Gross and Robert W Levenson, "Emotion elicitation using films," *Cognition & emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[15] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," *Cognition and Emotion*, vol. 24, no. 7, pp. 1153–1172, 2010.

[16] Arthur Aron, Edward Melinat, Elaine N Aron, Robert Darrin Vallone, and Renee J Bator, "The experimental generation of interpersonal closeness: A procedure and some preliminary findings," *Personality and Social Psychology Bulletin*, vol. 23, no. 4, pp. 363–377, 1997.

[17] Jáchym Kolář, *Automatic Segmentation of Speech into Sentence-like Units*, Ph.D. thesis, University of West Bohemia in Pilsen, 2008.

[18] Margaret M Bradley and Peter J Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[19] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al., "Emotionlines: An emotion corpus of multi-party conversations," *arXiv preprint arXiv:1802.08379*, 2018.

[20] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[21] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al., "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.

[22] Nhi Ngo and Derek M Isaacowitz, "Use of context in emotion perception: The role of top-down control, cue type, and perceivers age.," *Emotion*, vol. 15, no. 3, pp. 292, 2015.

[23] Richard T Cauldwell, "Where did the anger go? the role of context in interpreting emotion in speech," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[24] Tanja Bänziger, Georg Hosoya, and Klaus R Scherer, "Path models of vocal emotion communication," *PloS one*, vol. 10, no. 9, pp. e0136675, 2015.

[25] Kenneth Joseph, Lisa Friedland, William Hobbs, Oren Tsur, and David Lazer, "Constance: Modeling annotation contexts to improve stance classification," *arXiv preprint arXiv:1708.06309*, 2017.

[26] Leaf Van Boven, Katherine White, and Michaela Huber, "Immediacy bias in emotion perception: Current emotions seem more intense than previous emotions.," *Journal of Experimental Psychology: General*, vol. 138, no. 3, pp. 368, 2009.

[27] W Richard Walker and John J Skowronski, "The fading affect bias: But what the hell is it for?," *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, vol. 23, no. 8, pp. 1122–1136, 2009.

[28] Biqiao Zhang, Georg Essl, and Emily Mower Provost, "Predicting the distribution of emotion perception: capturing inter-rater variability," in *International Conference on Multimodal Interaction*, 2017, pp. 51–59.

[29] Tongliang Liu and Dacheng Tao, "Classification with noisy labels by importance reweighting," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 38, no. 3, pp. 447–461, 2016.

[30] Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, Melvin McInnis, and Emily Mower Provost, "Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition," *Proc. Interspeech*, 2017.