CROWDSOURCING EMOTIONAL SPEECH

Jennifer Smith[⋆] *Andreas Tsiartas*[⋆] *Valerie Wagner*[⋆]
Elizabeth Shriberg[†1] *Nikoletta Bassiou*[⋆]

[⋆]SRI International, USA
[†]Ellipsis Health, USA

ABSTRACT

We describe the methodology for the collection and annotation of a large corpus of emotional speech data through crowdsourcing. The corpus offers 187 hours of data from 2,965 subjects. Data includes non-emotional recordings from each subject as well as recordings for five emotions: angry, happy-low-arousal, happy-high-arousal, neutral, and sad. The data consist of spontaneous speech elicited from subjects via a web-based tool. Subjects used their own personal recording equipment, resulting in a data set that contains variation in room acoustics, microphone, etc. This offers the advantage of matching the type of variation one would expect to see when exposing speech technology in the wild in a web-based environment. The annotation scheme covers the quality of emotion expressed through the tone of voice and what was said, along with common audio-quality issues. We discuss lessons learned in the process of the creation of this corpus.

***Index Terms*—** Emotional speech, speech corpora, spontaneous speech, crowdsourcing, annotation

1. INTRODUCTION

In recent years, the increase in the number and scope of speech-based human-machine interfaces has fueled a growing interest in the recognition, modeling, and generation of emotion in speech. Accurate classification of emotion in speech is integral to many speech technology applications, such as affective computing, remote health monitoring, and speech synthesis. All of these applications depend on emotional speech data for training and development.

The collection of emotional data presents several challenges. The most crucial of these challenges is the fact that emotion is a complex phenomenon that is not easily defined [1]. Psychology offers multiple theories to help define and map the emotional space, but each of these has their own strengths and weaknesses [1, 2]. In addition, context and cultural factors can play a role in the way that emotion is expressed in speech. For example, anger may be expressed differently when speaking with a family member than when speaking with a coworker. Also, emotions often co-occur resulting in mixed emotions. For example, a wedding may elicit simultaneous joy and sadness in the same individual. The complexity of emotion presents many challenges to the design of collection and annotation for emotional speech data. In this collection, we target "full-blown" emotion in five different categories that span the arousal/valence space: angry, happy-low-arousal, happy-high-arousal, neutral, and sad.

Although recent research has shown that classification using multimodal data sets outperforms experiments using unimodal data [3], many applications preclude the use of video or physiological data. For example, speech alone is more appropriate for monitoring the state of children in the classroom due to privacy concerns [4]. In this work, the collection focuses on speech alone.

Researchers have been collecting emotional speech data for decades. Many collection designs have relied on professional actors reading from scripts or memorizing passages [5, 6, 7, 8, 9]. To get closer to natural data, some researchers have tried eliciting the desired emotion through an interaction designed to induce emotion in the participant (for example, [5, 10, 11, 12]). Collecting at scale using this method is very expensive because participants may vary in when and to what degree they experience the desired emotion. Furthermore, if the interaction requires a researcher's active participation or the development of a bot, this adds extra costs. Others have moved toward more natural emotion by using emotional recall and story telling (for example, [13]). We describe a corpus of data collected more like the latter, with the added advantage of an increase in the number of subjects by an order of magnitude.

We were able to collect and annotate such a large data set by implementing crowdsourcing methods. The proliferation of crowdsourcing platforms has inspired numerous researchers to investigate online annotation for natural language processing tasks, as reviewed in [14]. Suggestions for best practices in pay, handling errors, interface design, and task design have been offered [15, 16, 17, 14, 18, 19, 20]. In this paper, we describe a method for eliciting short "full blown" emotional utterances from thousands of speakers at a low cost. We also describe strategies for inexpensive and fast annotation and show that annotations increase the discriminability of the emotions.

2. DATA COLLECTION

Data was collected via a web-based crowdsourcing tool, enabling the rapid collection of data from a large number of subjects. All consent and personally identifying information is handled by the web service and data was transmitted to us already anonymized. Recordings were collected in batches of ten recordings: eight emotion-recall free speech recordings, one read passage, and one non-emotional free speech recording. The latter prompted the subject to discuss the date, time, and weather in their normal voice. The read passage was as follows:

> "Bridges, tunnels, and ferries are the most common methods of river crossing. The eastern coast is a place for pure pleasure and excitement."

The read speech and non-emotional free speech provide a baseline for comparison of each subject's emotion-recall recordings and their normal speaking voice. All emotion-recall recordings in an individual batch targeted the same emotion to avoid the additional cognitive load of emotion switching. Batches of recordings were collected

[1] Contribution made during tenure at SRI International.

978-1-5386-4658-8/18/$31.00 ©2018 IEEE

for five different emotions: angry, happy-low-arousal, happy-high-arousal, neutral, and sad. Subjects were free to record batches of data for anywhere from one to five of the five emotions.

Subjects were prompted to use past emotional experience as the basis for expressing emotion in the emotion-recall free speech recordings. Specifically, they were asked to recall an event from their past in which they experienced the emotion of interest. They were asked to imagine that the experience was happening in the present moment and to express the "full blown" emotion in one or two sentences, as if talking to a trusted friend or family member. The subjects improvised their words rather than reading from or memorizing a script. The prompt is inspired by the concept of "Emotional Memory," developed by the influential theater practitioner Konstantin Stanislavski. The expectation is that one can evoke true emotion in the present moment by recalling and role playing with past emotional experiences.

Subjects were also prompted to remember how the emotional experience manifested in the body. They were asked to recreate the embodiment of the emotion while recording. For example, subjects may smile while speaking about a happy event, frown while speaking about a time they felt angry, or slump their shoulders while recalling something that made them sad. Research on the embodiment of emotion has shown that subjects report that when they embody emotion through emotion-specific postures, they feel the associated emotion [21]. Many of the subjects reported that they experienced the elicted emotion while recording.

Further, some bodily expressions of emotion have been shown to affect the speech signal. For example, smiling changes the acoustics and prosody of speech [22, 23], and smiling can be identified by humans listening to speech without visual information [24, 25, 26]. Other bodily expressions of emotion may also plausibly affect articulatory positioning and other mechanisms of speech. Thus, subjects were also instructed to embody the emotion.

Lastly, subjects were required to listen to a series of three short audio examples of the targeted emotion. We found that providing audio examples was the most efficient way to communicate the desired emotion and was especially important for helping subjects understand the difference between happy-low-arousal and happy-high-arousal, for example. We tested this by collecting data using identical prompts, with and without audio examples. We found that without audio examples, subjects were less expressive in tone of voice and more frequently off-task. To avoid guiding subjects towards an overly narrow version of each emotion, we carefully chose three examples for each emotion, with substantial variation of emotional expression.

3. DESCRIPTIVE STATISTICS

The emotional speech corpus consists of 187 hours of recorded speech. A total of 110,068 audio recordings were collected. The corpus consists of 86,904 (79%) emotional recordings and 23,164 (21%) non-emotional recordings. Data was collected from 2,965 subjects. Before recording their speech, subjects answered a survey in which they reported their age group, gender, device type, and microphone type. Subjects ranged in age, with the 98.2% of subjects falling between the ages of 18 and 65 (Figure 1). Slightly more females participated than males, with 57.1% reporting their gender as female, and 42.6% as male. The other 0.3% declined to report their gender.

Unlike laboratory speech collection set-ups, subjects recorded themselves remotely, using their own personal recording equipment. This resulted in variation in room acoustics, microphone, etc., that matches the type of variation one would expect to see when exposing speech technology in the wild in a web-based environment. Roughly 60% of subjects recorded themselves using the built-in microphone from a laptop computer (Figure 2). Only 24.1% of subjects used a close-talking microphone from a headset to record their speech.

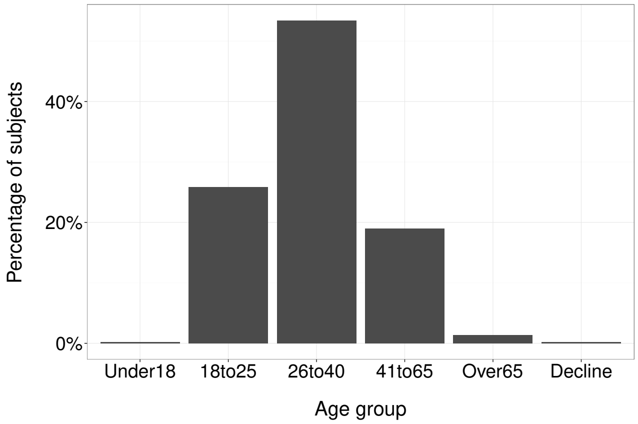

Fig. 1. The distribution of the age of subjects.

Data collection was implemented via a web-based collection tool. Since subjects had to access the collection via the web and recorded themselves independently, we can assume that the majority of subjects probably have basic computer literacy. The web-based design also enabled collection of data from subjects across the United States, with a variety in American English accents. The majority of subjects are native English speakers.

Data was collected for five emotions: angry, sad, neutral, and happy-low-arousal and happy-high-arousal. Subjects were free to contribute data for anywhere from one to all five of the five emotions collected. The majority of subjects (39.7%) contributed data for only one of the five emotions collected, while 16.0% contributed data for five of the five emotions. We collected roughly equal numbers of samples for each emotion.

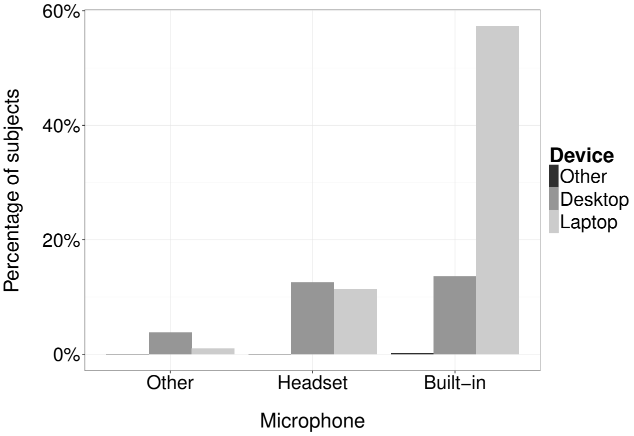

Fig. 2. The distribution of the type of device and microphone used by subjects.

We targeted "full blown" emotion, but the expression of emotion in the voice is not easily sustained over long periods of speech. Thus, we designed the collection to elicit short utterances. The majority (98%) of audio recordings are less than 60 seconds and the median length of a recording is 7 seconds. The distribution of the length of audio recordings is shown as a density plot estimation in Figure 3.

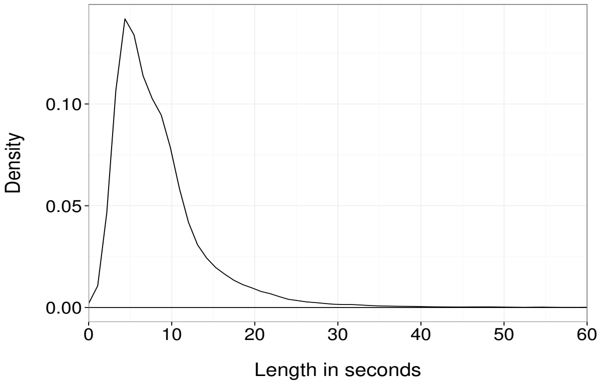

Fig. 3. The distribution of the length of audio recordings.

4. ANNOTATION

Annotation, like data collection, was also designed for implementation via a web-based tool. Web-based crowdsourcing enables the annotation of a large corpus to be divided among many annotators, making it faster and cheaper than employing a small group of highly trained annotators. One of the trade-offs for speed is that extensively training and vetting annotators is more difficult. Our initial approach to this challenge was using the web-based application to vet annotators via a paid training session and a test. Unfortunately, after completing the training and passing the test, a large proportion of the vetted annotators chose not to return to annotate the data.

We found that a more efficient approach was to allow anyone to annotate the data after completing a brief training session and to deal with discrepancies in annotation skills in post-processing. To account for differences in annotation skills, we used two rounds of majority voting. The first round of majority voting was applied at the level of the individual audio file. The annotators each annotated fifty randomly selected audio files. The audio files were selected randomly to ensure that each batch of audio files recorded by a single subject was not annotated by the same group of annotators. Each audio file was annotated by at least three annotators and majority voting was used to determine the final annotation. If there was no agreement among annotators, the file was flagged for discarding.

The second round of majority voting was applied at the level of the batch of recordings from a single subject for a single emotion. Recordings were collected in batches, and each batch consisted of eight samples for one emotion from one subject. We noticed that most subjects gave consistent performance across all eight samples. For example, if there were problems with background noise or distortion, those were generally present in most, if not all, of the eight files. Emotive ability was also fairly consistent across the batch. We took advantage of this consistency by annotating the first, fourth, and seventh of the eight audio recordings, and using majority voting to map those annotations to the rest of the files. These two rounds of majority voting reduced the likelihood that the annotations from the less skilled annotators would end up influencing the final annotations.

The data was annotated for both the quality of emotion in the tone of the voice and the emotion expressed in the content of the words spoken. We also asked annotators to listen for common audio-quality issues. Specifically, annotators were asked to report voices or other noises in the background and any distortion or skipping in the audio. Lastly, we asked annotators to mark files in which a subject expresses emotion that sounds particularly authentic as “very realistic.”

Annotators completed a training session before annotating files. The training session offered a concise explanation of the annotation scheme, with audio examples to illustrate common issues. The most complex concept that we needed to impart in training is that a subject’s tone of voice may not match the content of their words. For example, a subject may speak with a neutral voice, but say “I am so angry. I can’t believe he didn’t show up again.” If we simply ask an annotator “what emotion do you hear in this recording?,” there would be two right answers: neutral and angry. To control for this issue, we asked the annotator to separately annotate the content of the words spoken and the tone of voice. To help illustrate what we meant, we required the annotators to listen to audio examples in which the tone of voice and content of the words did not match. To help define what we meant by tone of voice, we explained it this way:

> “We are going to ask you to judge the emotion in the *tone* of a person’s voice while ignoring the *words* they say. To do this, imagine that you are listening to the person through a wall, so that you can’t understand what they’re saying, but you can hear the emotion in their voice.”

In addition to asking annotators to report on which emotion they heard in the tone of voice, we asked them to annotate the quality of the tone of voice. Annotators could choose between the following options for the tone of voice:

- Emotion is very faint (i.e. it almost sounds like normal talking, but there’s a hint of emotion)
- Emotion is very exaggerated or the person is goofing around
- Emotion sounds very realistic, great acting!
- None of the above

We wanted to annotate and filter out “faint” data in order to increase the discriminiability of the classes. The “exaggerated” label targeted recordings for which subjects either over acted or did not follow instructions.

We annotated a subset of the data to investigate the quality of the crowdsourced data and to determine whether or not annotating the data before building a classification system would improve performance. The annotated subset consists of 5,168 recordings, approximately 5% of the total 110,068 recordings. Annotators found that 98% of the subset have no audio-quality issues, like background noises, distortion, or skipping. We further subsetted the data by filtering out “faint” or “exaggerated” data. We included only recordings in which subjects express the elicited emotion in both the tone of voice and the content of what was said, leaving 46% of the data. When we apply the additional constraint that data should sound “very realistic,” we are left with 29% of the data.

Interestingly, the percentage of recordings from each emotion that passed the strictest quality test varied greatly. The pass rates for angry, sad, happy-high-arousal, happy-low-arousal, and neutral were 50.5%, 27.0%, 38.2%, 22.7%, and 4.2%, respectively. As seen in Figure 4, if we ease the requirement that data sound “very realistic”, the percentage of data that passes increases about 8-10% for each of the classes except for neutral. For the neutral data, removing the “very realistic” requirement results in a pass rate of 60.0% rather than 4.2%. It appears that subjects had the hardest time speaking naturally when asked to speak without emotion.

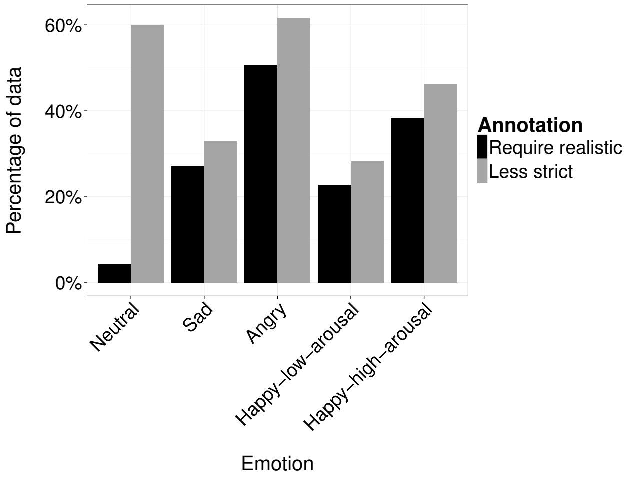

Fig. 4. Comparison of the percentage of data for each emotion that meets two sets of annotations requirements. Black bars represent the most strict quality control, including the requirement that data sound "very realistic". The gray bars represent quality control made less strict by removing the "very realistic" requirement.

To automatically evaluate the value of annotating the data using our scheme, we trained classifiers using unannotated and annotated data sets, and compared the results. The two classifiers were built using the SenSay$^{\text{TM}}$ platform [27]. The platform uses state-of-the-art machine learning approaches and performs real-time feature extraction and classification. The features used capture spectral, prosodic, articulatory, auditory, discourse, and fluency characteristics, as well as features designed specifically for robustness to noise and reverberation. For these experiments, we employed a single-layer DNN model because the subsets were too small for a larger network.

One classifier was trained using a class-balanced subset of the annotated data that met the strictest quality constraints and the other was trained using a balanced random sample of the same size. Both were tested on an unseen class-balanced set drawn from the annotated data that met the strictest quality constraints. We excluded "neutral" data because of the lack of samples after annotation. The four-way classifier trained on the annotated set produced a 5.3% absolute and 13.3% relative improvement in accuracy over the classifier trained using the unannotated set and the self-ratings as gold-standard. This result indicates that the annotation scheme is useful in increasing the discriminability of the four classes.

5. DISCUSSION AND LESSONS LEARNED

Web-based crowdsourcing is a fast and inexpensive way to collect and annotate large amounts of audio data from a multitude of subjects. As investigated in [16], optimizing the size and clarity of the task is crucial for maximizing quality of data collection and annotation. We tested various task designs by collecting data from 15-20 subjects and utilizing expert in-house annotators. We found that subjects produced the most realistic "full blown" emotion when we requested multiple short utterances. Specifically, we found that ten recordings was the maximum number of recordings for which we received the most consistent quality of emotion. When we requested more than ten recordings, we observed less consistent quality in emotion and complaints from subjects. We also achieved a boost in quality by collecting emotions in separate batches, thus avoiding the additional cognitive load of switching between different emotions. Lastly, requiring subjects to listen to examples of emotional speech before recording themselves also increased the quality of emotion in the data. We were careful to provide not one but three varied examples of the elicited emotion in order to avoid eliciting an overly narrow range of emotional expressions.

We were not successful in recruiting annotators to train and annotate data over a longer term using a crowdsourcing interface. We found that annotators often completed training without returning to annotate data and the time and financial commitment was not returned. Our final annotation method allowed anyone to annotate data and accounted for errors by employing two rounds of majority voting. Others have also suggested that annotation crowdsourcing methods should 'embrace' error [15, 28]. Classifiers trained using the data that met our strictest criteria showed that the annotation scheme was effective in improving the differentiation of the emotional classes.

The results of the annotation suggest that emotional speech data collected via this crowdsourcing method will have mostly (98%) high audio-quality, and that approximately one third of the data will be perceptually realistic. However, special considerations may be necessary when collecting "neutral" data as subjects had the hardest time speaking naturally when asked to speak without emotion.

6. SUMMARY AND FUTURE DIRECTIONS

In this paper, we presented a methodology for collecting a large corpus of emotional speech data. The 187 hours of recordings were collected and annotated via a web-based crowdsourcing tool. The corpus offers data from a much larger number of subjects than the majority of existing emotional speech corpora. Speech was collected for five emotions: angry, happy-low-arousal, happy-high-arousal, neutral, and sad. Non-emotional utterances were also collected for each recording session. Emotion was elicited by instructing subjects to recall a time in their life when they experienced the target emotion and to recreate the embodiment of the emotion while recording. The subjects then described why they felt that emotion in their own words.

Recordings were made using the subjects' personal recording equipment, thus the data contains the kind of variation that we would expect to see when exposing speech technology in a web-based environment without control of the equipment. Specifically, the data contains variation from the microphone, room acoustics, etc.

An annotation scheme was developed and tested on a subset of recordings. The annotation scheme targets several quality-control measures: the quality of emotion in the tone of voice, the quality of emotion in the choice of words, and common audio-quality issues. The scheme was tested on a subset of data and shows that annotation improves the ability to distinguish among the five classes.

The next step is to annotate the rest of the data. We expect that the annotation process will yield similar results and the subset of recordings that meet the strictest quality control criteria will be about 29% of the original 110,068 recordings, or approximately 32,000 recordings. We plan to use this corpus in order to train models for our SenSay$^{\text{TM}}$ emotion recognition platform. But this type of data can be useful for other research efforts, for example to study emotion variation impact on speaker identification tasks.

7. REFERENCES

[1] Roddy Cowie and Randolph R Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1, pp. 5–32, 2003.

[2] Stefan Steidl, *Automatic classification of emotion related user states in spontaneous children's speech*, Logos Verlag, 2009.

[3] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[4] Jennifer Smith, Harry Bratt, Colleen Richey, Nikoletta Bassiou, Elizabeth Shriberg, Andreas Tsiartas, Cynthia DAngelo, and Nonye Alozie, "Spoken interaction modeling for automatic assessment of collaborative learning," in *Speech Prosody*, 2016, pp. 277–281.

[5] Ellen Douglas-Cowie, Roddy Cowie, and Marc Schröder, "A new emotion database: Considerations, sources and scope," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[6] Dimitrios Ververidis and Constantine Kotropoulos, "A state of the art review on emotional speech databases," in *Proceedings of 1st Richmedia Conference*, 2003, pp. 109–119.

[7] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas, "The enterface05 audio-visual emotion database," in *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*. IEEE, 2006, pp. 8–8.

[8] 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.

[9] 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.

[10] Anton Batliner, Christian Hacker, Stefan Steidl, Elmar Nöth, Shona D'Arcy, Martin J Russell, and Michael Wong, "'You stupid tin box'- Children interacting with the AIBO robot: A cross-linguistic emotional speech corpus," in *LREC*, 2004.

[11] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[12] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[13] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss, "A database of German emotional speech," in *Interspeech*, 2005, vol. 5, pp. 1517–1520.

[14] Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl, "Corpus annotation through crowdsourcing: Towards best practice guidelines.," in *LREC*, 2014, pp. 859–866.

[15] Ranjay A Krishna, Kenji Hata, Stephanie Chen, Joshua Kravitz, David A Shamma, Li Fei-Fei, and Michael S Bernstein, "Embracing error to enable rapid crowdsourcing," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 3167–3179.

[16] Ujwal Gadiraju, Patrick Siehndel, Besnik Fetahu, and Ricardo Kawase, "Breaking bad: Understanding behavior of crowd workers in categorization microtasks," in *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 2015, pp. 33–38.

[17] Leib Litman, Jonathan Robinson, and Cheskie Rosenzweig, "The relationship between motivation, monetary compensation, and data quality among us-and india-based workers on mechanical turk," *Behavior Research Methods*, vol. 47, no. 2, pp. 519–528, 2015.

[18] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng, "Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 254–263.

[19] R Morris and Daniel McDuff, "Crowdsourcing techniques for affective computing," in *The Oxford Handbook of Affective Computing*, pp. 384–394. Oxford Univ. Press, 2014.

[20] Alexey Tarasov, Sarah Jane Delany, and Charlie Cullen, "Using crowdsourcing for labelling emotional speech assets," 2010.

[21] Paula M Niedenthal, "Embodying emotion," *Science*, vol. 316, no. 5827, pp. 1002–1005, 2007.

[22] Véronique Aubergé and Marie Cathiard, "Can we hear the prosody of smile?," *Speech Communication*, vol. 40, no. 1, pp. 87–97, 2003.

[23] John J Ohala et al., "An ethological perspective on common cross-language utilization of F0 of voice," *Phonetica*, vol. 41, no. 1, pp. 1–16, 1984.

[24] Vivien C Tartter, "Happy talk: Perceptual and acoustic effects of smiling on speech," *Attention, Perception, & Psychophysics*, vol. 27, no. 1, pp. 24–27, 1980.

[25] Sascha Fagel, "Effects of smiling on articulation: Lips, larynx and acoustics," *Development of Multimodal Interfaces: Active Listening and Synchrony*, pp. 294–303, 2010.

[26] Amy Drahota, Alan Costall, and Vasudevi Reddy, "The vocal communication of different kinds of smile," *Speech Communication*, vol. 50, no. 4, pp. 278–287, 2008.

[27] A Tsiartas, C Albright, N Bassiou, M Frandsen, I Miller, E Shriberg, J Smith, L Voss, and V Wagner, "Sensay analytics^{TM}: A real-time speaker-state platform," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 6582–6483.

[28] Emily Jamison and Iryna Gurevych, "Noise or additional information? Leveraging crowdsource annotation item agreement for natural language tasks.," in *EMNLP*, 2015, pp. 291–297.