2012-05-26

A Crowdsourcing Approach to Labelling a Mood Induced Speech Corpus

John Snel
Technological University Dublin, john.snel@student.dit.ie

Alexey Tarasov
Technological University Dublin, tarasovsaleksejs@gmail.com

Charlie Cullen
Technological University Dublin, charlie.cullen@tudublin.ie

See next page for additional authors

Funder: Science Foundation Ireland

A Crowdsourcing Approach to Labelling a Mood Induced Speech Corpus

John Snel*, Alexey Tarasov†, Charlie Cullen*, Sarah Jane Delany†

*Digital Media Centre, Dublin Institute of Technology
Aungier St, Dublin 2, Ireland
john.snel@student.dit.ie, charlie.cullen@dmc.dit.ie

†School of Computing, Dublin Institute of Technology
Kevin St, Dublin 8, Ireland
aleksejs.tarasovs@student.dit.ie, sarahjane.delany@dit.ie

Abstract

This paper demonstrates the use of crowdsourcing to accumulate ratings from naïve listeners as a means to provide labels for a naturalistic emotional speech dataset. In order to do so, listening tasks are performed with a rating tool, which is delivered via the web. The rating requirements are based on the classical dimensions, activation and evaluation, presented to the participant as two discretised 5-point scales. Great emphasis is placed on the participant's overall understanding of the task, and on the ease-of-use of the tool so that labelling accuracy is reinforced. The accumulation process is ongoing with a goal to supply the research community with a publicly available speech corpus.

1. Intro

As part of building a naturalistic speech corpora, annotators are required to label and index emotional episodes associated with the acquired speech. In most cases, rather small numbers of "expert" labellers are asked to participate in listening tasks; the assignment of gathering large numbers of annotators is rarely a principal research objective. Moreover, most research does not indicate explicitly what expertise the annotators have. Expert listeners are usually researchers who are part of the wider field of emotional research.

Emotion is an important aspect of communication between *all* humans. The method used to accumulate ratings in this paper is through the use of crowdsourcing, which has been suggested by Tarasov et al. (2010). It diverges from others as we focus on large-scale listening groups not depicted as "expert" annotators—we suggest equal validity between an *expert* and a *non-expert* annotator's emotional judgement. That is to say, we aim to accumulate judgment ratings from a broader sample population that are not necessarily familiar with emotion theory.

The paper is structured as follows. Section 2 describes the related work in crowdsourcing and emotional speech labelling. Aims of our research are stated in section 3, and section 4 covers the methods used for creating the rating framework. The preliminary results are covered in section 5, and section 6 concludes the report.

2. Related work

In this section, a brief outline is given of related work in the area of accumulating labellers for corpora, and the labelling methods that have previously been used.

2.1. Crowdsourcing

Crowdsourcing is the use of tasks outsourced to a large group of non-expert individuals (Howe, 2008). Typically, a large number of tasks are distributed across a population of raters, and there from the results of several task solutions are combined. In the context of labelling corpora, each asset is presented to several raters and labelled separately by each individual. The final label for the asset is some combination of these labels; take majority voting for example (Brew et al., 2010). Crowdsourcing has recently been used for the task of getting labels for different corpora in numerous domains such as machine translation (Ambati et al., 2010), computer vision (Smyth et al., 1995; Sorokin and Forsyth, 2008), and sentiment analysis (Brew et al., 2010; Hsueh et al., 2009). Crowdsourcing is a fast way to accumulate labels; for instance, the work of Snow et al. (2008) received 151 ratings per hour, while Sorokin and Forsyth (2008) reported a speed of 300 ratings per hour. Nevertheless, with sufficient number of raters the quality of labels remains high and comparable to that of experts (Ambati et al., 2010; Snow et al., 2008; Sorokin and Forsyth, 2008).

Support for using crowdsourcing with regard to rating emotional speech is shown in the work of Cowie and Cornelius (2003). According to them, it can be argued that emotional expertise does not necessarily correlate with emotional experience, suggesting that the wider, non-expert population can provide labels that are equally valid to those of experts, who are primarily used to perform rating of emotional speech assets in state-of-the-art research.

2.2. Labelling naturalistic emotional speech

An early example of work that highlights the complexities in labelling naturalistic emotional speech is on the Leeds-Reading database (Roach et al., 1998). Emotional annotation came to four levels. The first level used freely chosen everyday emotion labels; the second specified the strength of the emotion, together with a sign to indicate valence; and, the third and fourth described emotional episodes based on the individual's appraisal of the event. Understandably, they specified that the number of categories associated with an in-depth qualitative coding strategy will amount to smaller occurrences in each category.

The development of the Belfast Naturalistic database

(Douglas-Cowie et al., 2000) followed from the Leeds-Reading experience. Their focus was to develop a quantitative description. They developed "trace" techniques to evaluate, quantitatively, emotion as it changes over time along underlying affect dimensions—positive to negative and active to passive. They argued that quantitative measurement using the Feeltrace tool better estimated real consensus compared to categorical labels, because of the inclusion of similarity—rather than only identical—measures. As somewhat unexpected, dimensional ratings showed less individual differences compared to categorical ratings, showing closer agreement on the evaluation dimension. For the rating task, however, they acquired three trained raters to use the tool; therefore, for this study, which excludes the need for comprehensive training inappropriate for crowdsourcing (large-scale, non-expert listening groups), the methods are adapted to meet the relevant requirements.

A comprehensive labelling schema for the JST/CREST Expressive Speech Corpus (Campbell, 2006) also included a version of the Feeltrace tool—and noted that labellers understood the meaning and validity of the two dimensions. Further, they proposed three levels for labelling: state of speaker, style of speaker, and physical aspects of the voice. This comprehensive schema is data-driven and appeared to be necessary when listening to speech in context and over long segments. For example, they familiarised themselves with the speakers mannerism when labelling someones speech over a five-year period. Such a comprehensive scheme, however, is not suitable for short segments of speech found in this particular study's speech dataset.

The study by Grimm, Kroschel and Narayanan (Grimm and Kroschel, 2008) used a three-dimensional model—valence, activation, and dominance. Interestingly, they discretised the continuous dimensional scales into 5 classes.

3. Aims

The focus of this paper, as part of an ongoing corpus building project, is to provide labels based on how naïve listeners judge conveyed emotional dimensions (i.e. effect-type orientation (Cowie and Cornelius, 2003)), for speech extracted from a previously constructed naturalistic, mood induced, emotional speech dataset (Cullen et al., 2008). Listeners are asked to rate on two scales that represent the activation-evaluation space.

Considering there is no absolute "ground truth" in emotion labels, and given that an individual's impression of emotion in speech is subjective in nature, it is suggested here that the use of crowdsourcing is a convenient method for determining more robust consensual ratings.

To collect ratings from large-scale listening groups, the listening tasks are performed through an online listening tool. The tool has its focus on user-centred design (UCD), developed and tested keeping in mind ease-of-use, ensure adequate understanding for each scale, and encourage participation by minimising the requirements of personal details. Moreover, the tool aims to be suited for repeated use to accumulate continual ratings from all participants.

4. Methods

This section describes the methods used to obtain the speech data, the framework chosen to label it, the available tool for the labellers, and the validation of tool design.

4.1. Data acquisition

The designated naturalistic emotional speech corpus for labelling is constructed based on Mood Inducing Procedures (MIPs) (Gerrards-Hesse et al., 1994). With inevitable restrictions in obtaining truly natural material while at the same time isolating the desired speech signal from unwanted noise, MIPs provide for a convenient trade-off. In this dataset, the inducing methods were performed on participants in a controlled environment with soundproof isolation booths. The build of the corpora (Cullen et al., 2008) investigated 3 different experiments incorporating the MIP 4 group (Success/Failure and Social Interaction MIP) and the MIP 3 group (Gift MIP). It considered several critical factors. Amongst these were: authenticity of emotional content, demand effects[1], ethical issues, and audio quality. The speech clips have been extracted from 8 different MIP sessions, and a total of 160 speech clips were chosen from 16 different speakers (7m/9f).

4.2. Labelling framework

To avoid the issues with subjective category labels, the labelling framework used in this paper is the dimensional approach as it appears to be more suited for cross-studies in a wider context (Eyben et al., 2003). Our method is comparable to the Feeltrace tool (Cowie et al., 2000), as mentioned above, mainly because of the number and type of dimensions used. We employ two-dimensions: activation and evaluation. Our method differentiates from the Feeltrace tool in two major ways.

First, our method is renouncing time-continuous evaluation, i.e. trace labelling (see also the work by Grimm and Kroschel (2005)), and instead provides annotation for utterances of discrete periods of time (termed as *quantised* labelling (Cowie et al., 2011)). The speech utterances rated are of short length (~5 seconds), and we are assuming that within the speech segment no changes in emotion occur, and are thus kept constant (Busso et al., 2008). For this study, prioritising large-scale rating via crowdsourcing is at odds with trace labelling that necessitates trained labellers.

Second, participants are presented with two discretised scales (colour-coded) rather than a continuous circular—or square—representation of the evaluation/activation space.

4.3. Design of web-based tool

To assist crowdsourcing, the rating tool is delivered via the Internet. The objective of the tool[2] is to have a simple but clean interface to make it easy for participants to understand and use. The participant's understanding about each rating scale is given considerable importance. The tool includes a detailed instructions page about how and what to annotate. As a more straightforward representation of the

[1] Demand effects are those possibilities of the subject guessing the purpose of the procedure and hence act the desired emotion.

[2] The online tool can be found at http://dmcx.dit.ie/emovere

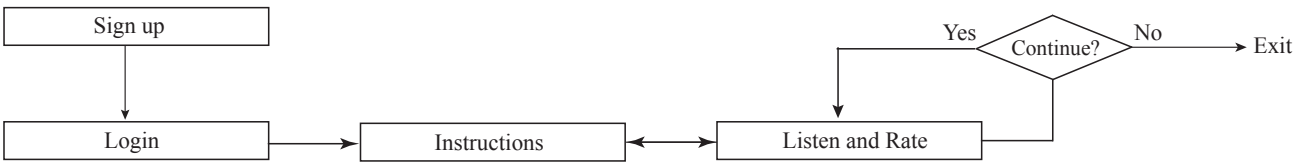

Figure 1: Flowchart of the presented web pages to the participant

circumplex model, whether circular or square, each dimension, activation and evaluation, is presented as two scales. For each scale, the participant is provided with a definition and an accompanied example. The design of the site (see Figure 1) ensures that the instructions are presented prior to the listening task, although the participants can refer back to the instructions at any stage during the task.

The participants are required to create a login account; and to prevent the impression of a daunting task and encourage participation, minimal details are required. However, mandatory information on first language and hearing impairment is required.

The listening task is presented as 3 successive steps i.e. listen to the speech clip and rate accordingly on both scales (see Figure 2). Each clip is only rated once by each participant. To avoid order effects, speech clips are randomised; and, to avoid fatigue and boredom effects participants are presented with just 6 speech clips before given the option to exit. Participants are given the option to skip a speech clip if they feel they cannot rate it by choosing "Do not rate". To prevent participants from continually doing this, it is required to fully listen to—or at least until the audio player has reached the end of the speech clip—before rating is activated. If a participant chooses "Do not rate" for 3 consecutive speech clips, they are notified and asked if they want to exit. A total of 160 speech clips are available for each participant to rate, and each clip can be replayed as many times as the participant wants. Participant details and rating information has been kept in two separate databases.

4.4. Preliminary survey (design validation)

Prior to implementation, we surveyed 7 non-expert (in emotional judgment) individuals to assess their understanding of the instructions using a multi-choice questionnaire. We ensured they were able to set up an account, and complete the task without difficulties. Participants were from a technical (college staff and other researchers) and non-technical (first year journalism students) background. The procedure for this was as follows:

1. Read instructions.
2. Answer questions about the definitions of both *evaluation* and *activation*.
3. Rate assets.
4. Assessment on workload.

For the activation question, 6 were correct and 1 incorrect; similarly, for the evaluation question, 6 were correct and 1

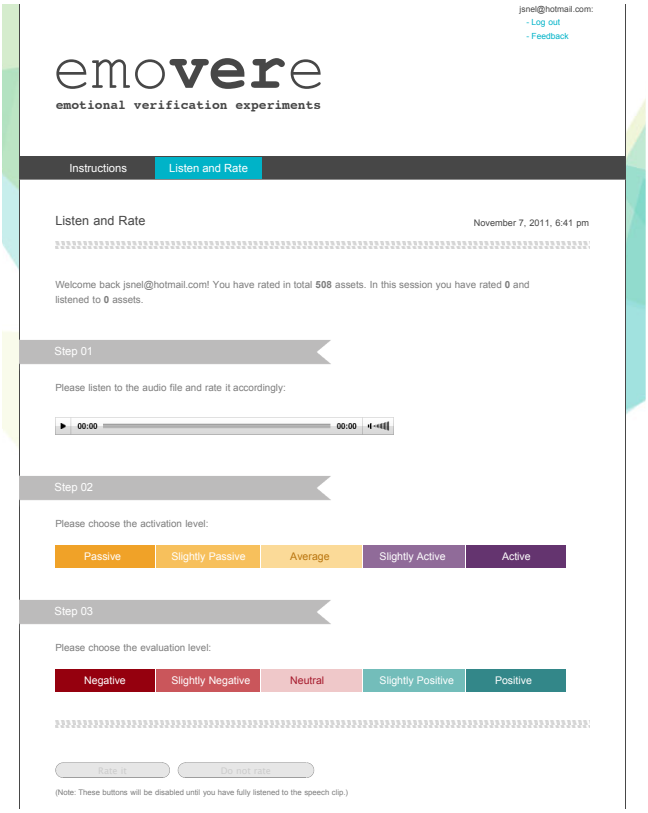

Figure 2: Online listening task

incorrect (see Table 1). It should be noted that the incorrect answers were from the same participant. The participant didn't follow the order of the above procedure. Instead, participant read instructions, rated assets, and then answered the questions on evaluation and activation. From this, it was concluded there was a sufficient amount of understanding among the raters for the instructions of both scales.

	Correct	Incorrect
Activation	6	1
Evaluation	6	1

Table 1: No. of correct and incorrect answers given for the multiple choice questions on activation and evaluation

A survey based on the NASA TLX (Hart and Staveland, 1988)—a subjective workload assessment tool—assessed the cognitive load on mental demands; temporal demands; and uncertainty, irritation, and stress (effort) while using

the online ratings tool. Overall, we concluded the cognitive demands were in adequate conditions (see Table 2).

Demand	VL	L	N	H	VH
Mental	1	1	3	2	0
Temporal	0	3	4	0	0
Effort	3	1	2	1	0

Table 2: Subjective workload assessment, VL=Very low, L=Low, N=Normal, H=High, VH=Very high

Participants were asked on the amount of assets that they would rate on a daily basis. 4 participants chose to keep it at 3 assets per day and 3 chose to increase the number. We concluded that participants should be presented with 3–7 assets at a time to prevent *boredom* and/or *fatigue* effects.
Besides querying cognitive load, participants gave free-response feedback on any other information they felt gave difficulties. Accordingly, technical issues within reason—such as browser issues and password restrictions—were addressed.
A brief summary of some interesting remarks from the free-respone feedback from the different participants is given as follows:

- Evaluation would be easier as binary.
- The definition of activation is easier to understand in terms of the dynamics of emotion.
- Scale for authenticity/genuineness could be introduced.
- There is a need for a baseline speech clip to compare others against.
- It was necessary to listen to some clips several times to hear the tone of voice, rather than the linguistic content.
- Others noted they assessed the clips along the scales according to the linguistic content.
- One participant said the speech clips were "weird".

5. Discussion

Since July 2011, we have received 1243 activation-evaluation pairs of ratings, which is 7.77 ratings per asset in average. The distribution of ratings for activation is shown in Figure 3, and for evaluation is shown in Figure 4. In total, 71 people have been registered as raters. Unfortunately, the majority have rated $<$20 assets, with a select few who provided labels for the whole corpus. The proportion of "Do not rate" ratings is only 3%, which shows that raters are rarely confused by the recordings. The evaluation dimension exhibits the same trend as would have been expected—it contains a large number of neutral ratings, gradually decreasing towards positive or negative classes. However, the corpus seems to have a relatively big number of active, non-neutral assets. One of the explanations can be the nature of the task faced by participants that forced them to act fast. In any case, it indicates that the MIP procedures used were successful in inducing non-neutral emotions.
We calculated the standard deviation (SD) for the ratings of each asset and used the mean value as a measure of rater agreement. The mean SD for the *activation* scale was 20% proportional to the width of the scale. Likewise, the *evaluation* scale came to 21%; that is to say, the participants are deviating from the average label by one class. The mean SD for this corpus was compared with the mean SD for the VAM corpus, which also used 5 discrete classes. The degree of agreement is comparable for both studies—VAM corpus is 14% for *activation* and 18% for *evaluation*.

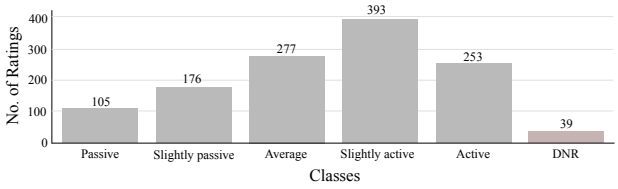

Figure 3: The number of ratings for the *activation* scale in the overall speech dataset, DNR=Do not rate.

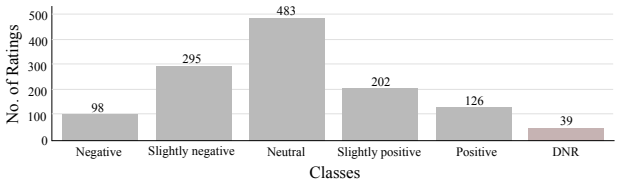

Figure 4: The number of ratings for the *evaluation* scale in the overall speech dataset, DNR=Do not rate.

6. Conclusions

One of our aims was to have participants engage with the tool on a daily basis, and rate six at a time to avoid fatigue and boredom effects that may cause spurious labelling. In spite of several reminders, it was difficult to achieve consistent daily rating from individual participants. However, the process of getting labels is on going. As an alternative, we are considering using crowdsourcing platforms such as Amazon Mechanical Turk[3] in addition to volunteer raters.
The release of all sets of ratings will be in the near future, including the single target label for each asset, obtained by aggregating the ratings submitted by raters. All rated assets will be freely available to the research community, with downloadable versions updated as ratings accumulate. With that, analysis on ratings will also be published. Finally, the corpus' speech dataset will be extended using other emotion eliciting methods, all in the same recording environment.

7. Acknowledgements

This work was supported by the Science Foundation Ireland under Grant No. 09-RFP-CMS253. Authors would like to

[3] http://www.mturk.com

thank Anna Deegan for the help with the implementation of the tool. We also express gratitudes to all raters, who participated in the research.

8. References

V. Ambati, S. Vogel, and J. Carbonell. 2010. Active Learning and Crowd-Sourcing for Machine Translation. In *Procs of LREC*, pages 2169–2174.

A. Brew, D. Greene, and P. Cunningham. 2010. Using Crowdsourcing and Active Learning to Track Sentiment in Online Media. In *Procs of PAIS*, pages 1–11.

C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan. 2008. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Language Resources and Evaluation*, 42(4):335–359.

N. Campbell. 2006. A language-resources approach to emotion: corpora for the analysis of expressive speech. In *The Workshop Programme Corpora for Research on Emotion and Affect Tuesday 23 rd May 2006*, page 1.

R. Cowie and R.R. Cornelius. 2003. Describing the Emotional States that Are Expressed in Speech. *Speech Communication*, 40(1-2):5–32.

R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schroder. 2000. 'FEELTRACE': An Instrument for Recording Perceived Emotion in Real Time. In *Procs of ITRW on Speech and Emotion*, pages 19–24.

R. Cowie, C. Cox, J. C. Martin, A. Batliner, D. K. J. Heylen, and K. Karpouzis. 2011. Issues in data labelling. In R. Cowie, C. Pelachaud, and P. Petta, editors, *Emotion-Oriented Systems. The Humaine Handbook*, Cognitive Technologies, pages 213–241.

C. Cullen, S. Kousidis, and J. McAuley. 2008. Emotional Speech Corpus Construction, Annotation and Distribution. In *Procs of LREC*.

E. Douglas-Cowie, R. Cowie, and M. Schroder. 2000. A new emotion database: considerations, sources and scope. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Citeseer.

Florian Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl. 2003. Cross-Corpus Classification of Realistic EmotionsSome Pilot Experiments. In *The Workshop Programme*, page 77.

A. Gerrards-Hesse, K. Spies, and F.W. Hesse. 1994. Experimental Inductions of Emotional States and Their Effectiveness: A Review. *British Journal of Psychology*, 85(1):55–78.

M. Grimm and K Kroschel. 2005. Evaluation of Natural Emotions Using Self Assessment Manikins. In *Procs of IEEE ASRU*, pages 381–385.

M. Grimm and K Kroschel. 2008. The Vera am Mittag German Audio-Visual Emotional Speech Database. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 865–868, Hannover, Germany.

S.G. Hart and L.E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human mental workload*, 1:139–183.

J. Howe. 2008. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Business.

P.Y. Hsueh, P. Melville, and V. Sindhwani. 2009. Data Quality from Crowdsourcing: a Study of Annotation Selection Criteria. In *Procs of ALNLP*, pages 27–35.

P. Roach, Richard Stibbard, Jane Osborne, Simon Arnfield, and Jane Setter. 1998. Transcription of Prosodic and Paralinguistic Features of Emotional Speech. *Journal of the International Phonetic Association*, 28(1-2):83–94.

P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi. 1995. Inferring Ground Truth from Subjective lLabelling of Venus Images. *Advances in Neural Information Processing Systems*, 7:1085–1092.

R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. 2008. Cheap and Fast—But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Procs of EMNLP*, pages 254–263.

A. Sorokin and D. Forsyth. 2008. Utility Data Annotation with Amazon Mechanical Turk. In *Procs of IEEE CVPR*, pages 1–8.

A. Tarasov, S.J. Delany, and Charlie Cullen. 2010. Using Crowdsourcing for Labelling Emotional Speech Assets. In *Procs of W3C workshop on Emotion ML*.