Using Crowdsourcing for Labelling Emotional Speech Assets

Alexey Tarasov
*Technological University Dublin*, aleksejs.tarasovs@student.dit.ie

Sarah Jane Delany
*Technological University Dublin*, sarahjane.delany@tudublin.ie

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

Follow this and additional works at: [https://arrow.tudublin.ie/dmccon](https://arrow.tudublin.ie/dmccon)

**Recommended Citation**

This Conference Paper is brought to you for free and open access by the Digital Media Centre at ARROW@TU Dublin. It has been accepted for inclusion in Conference papers by an authorized administrator of ARROW@TU Dublin. For more information, please contact yvonne.desmond@tudublin.ie, arrow.admin@tudublin.ie, brian.widdis@tudublin.ie.

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License.
Using crowdsourcing for labelling emotional speech assets

Alexey Tarasov, Charlie Cullen, Sarah Jane Delany
Digital Media Centre
Dublin Institute of Technology

Abstract

The success of supervised learning approaches for the classification of emotion in speech depends highly on the quality of the training data. The manual annotation of emotion speech assets is the primary way of gathering training data for emotional speech recognition. This position paper proposes the use of crowdsourcing for the rating of emotion speech assets. Recent developments in learning from crowdsourcing offer opportunities to determine accurate ratings for assets which have been annotated by large numbers of non-expert individuals. The challenges involved include identifying good annotators, determining consensus ratings and learning the bias of annotators.

1 Introduction

The automatic recognition of emotion from speech recordings uses supervised machine learning techniques which requires labelled training data in order to operate effectively. The performance of these supervised learning techniques depends on the quality of the training data and therefore on the quality of the labels. For many real life tasks, manual annotation by an expert is the primary way of getting the labels, but it can be an expensive and time-consuming process [20, 23, 4]. In some cases it can be impossible to get the actual label (also known as the ground truth or gold standard) and it is estimated from the subjective opinion of a small number of experts who can often disagree on the labels [14, 29]. It can be argued that emotional expertise does not necessarily correlate with emotional experience [11] suggesting that wider non-expert annotators can provide equally valid labels.

Recently with the availability of crowdsourcing [19] services such as Mechanical Turk\(^1\), reCAPTCHA\(^2\) and Games with a Purpose\(^3\) it has become inexpensive to acquire labels from multiple non-expert annotators. This has led to significant research into learning from crowdsourcing including comparing labels from non-expert annotators with the ground truth [20, 33], analysing consensus versus coverage requirements [7] and investigating methods and techniques for determining the ground truth and learning the bias of annotators [28, 29, 32].

In this position paper we propose the use of crowdsourcing for acquiring emotional labels in the domain of emotion recognition from speech. The rest of the paper is organized as follows—Section 2 presents a review of the techniques that are currently used for labelling emotional speech assets that are to be used as training data. Section 3 discusses crowdsourcing and its main challenges and includes some practical experiences of using crowdsourcing, while Section 4 concludes with a use case for an EmotionML which is appropriate for the usage being presented in this paper.

\footnote{This material is based upon works supported by the Science Foundation Ireland under Grant No. 09-RFP-CMS253}

\(^1\)www.mturk.com
\(^2\)recaptcha.net
\(^3\)www.gwap.com
2 Rating Emotional Assets for Training

Although current research in emotion recognition from speech has used both expert [2, 26, 36] and non-expert [9, 31] annotators for labelling the emotion in speech assets, most of the research does not indicate explicitly what expertise the annotators have.

Various numbers of annotators have been used—from two [2, 13, 36] up to between three to nine [9, 24, 26]. Odd numbers are often used to ensure a majority decision and five has been proposed as a good compromise [6]. There are a few exceptions when larger numbers of annotators have been used [1, 3, 21, 22, 27] but in these cases the annotated assets are not used for training an emotion recognition system so a consensus emotional rating is not required.

The methods used to determine a single rating from the different annotators depend on whether the annotation is categorical or dimensional. Majority voting is regularly used to select a categorical label [36]. With assets rated on a dimensional scale the mean of all annotators’ values for each dimension is usually used [18].

Determining which assets are used as training data is generally based on some measure of annotator agreement. In categorical rating, often a requirement is for a certain proportion of annotators to agree [5, 24] before an asset can be considered as training data. This requirement can be complemented with a strength scale, requiring agreement from a certain proportion of annotators who consider the emotion to be at least of a required strength [12, 30, 31]. However, the most popular measure of agreement between categorical annotators in this area is the $\kappa$-statistic [16]. There is no consensus on what value denotes a high level of agreement, although Fleiss [16] suggests values smaller than 0.4 indicate low agreement, values between 0.4 and 0.7 indicate good agreement and values higher than 0.7 indicate excellent agreement. However, others suggest that values above 0.75 [35] or above 0.80 [10] indicate good agreement and below 0.4 [35] or below 0.67 [10] indicate bad agreement. A significant amount of research in this area report $\kappa$ values between 0.3 and 0.5 [2, 9, 25] which overall does not support strong agreement among annotators, although there are some rare exceptions with very high $\kappa$ values around 0.8 [8, 26]. This high agreement may be explained by the nature of the annotation task undertaken in these cases, requiring the annotator to select from a small number of distinctive categories. For the much more rarely dimensionally rated assets Grimm and Kroschel [17] use the standard deviation of all ratings on a dimension as a measure of agreement between annotators.

There have been limited efforts in the literature into comparing non expert annotations with ‘ground truth’ in emotion recognition in speech. Engberg et al. [15] investigated how well people could recognize emotions in acted speech assets using 20 listeners. Their results revealed that the emotions were identified correctly in 67% of the cases indicating significant room for improvement.

Overall, there is little evidence in the literature of the usage of the recent phenomenon known as crowdsourcing in the labelling of emotional speech assets for use as training data. A limited number of annotators is generally used and, although it is often not often stated in the literature, the expectation is that these annotators are perceived to be experts. Generally, relatively simple statistical techniques are used to determine the actual label for the asset and thresholds on measures of inter-annotator agreement determine the most suitable training data.

3 Crowdsourcing

Crowdsourcing is the use of tasks outsourced to a large group of non-expert individuals [19]. One of the recent applications of crowdsourcing has been to label training data for a wide range of supervised learning application domains. It has been successfully used in machine translation [4], natural language tasks [33], computer aided diagnosis [28, 29], computer vision [32, 34] and sentiment analysis [7, 20].

Practical experiences with crowdsourcing has found that it can offer a fast and effective way to get labels
that are of the same quality as those from experts [33] and usually very cheaply—Ambati et al. [4] report that it is possible to get 20 hours of annotation for only US$45. Many researchers also note that it is possible to get a large number of labels very quickly—Snow et al. [33] obtained 300 ratings from 10 annotators in just 11 minutes using Amazon’s Mechanical Turk.

There are a number of challenges with using crowdsourcing to label speech assets, including (i) how to select which assets are presented for rating, (ii) how to estimate the reliability or bias of the annotators, (iii) how to derive the ground truth or actual rating for the asset and (iv) maintaining the balance between data coverage and data quality.

A recent development in this area is to use active learning to address the first of these concerns, that is selecting the appropriate assets for rating [4, 7, 14]. Active learning proposes techniques for selecting the more informative unlabelled examples to present for labelling. Brew et al. [7] recommend including a clustering-based step which results in identifying a sufficiently diverse set of clusters that represent dominant example types from which to select exemplars for labelling.

Both Donmez et al. [14] and Smyth et al. [32] propose learning approaches for determining the bias of the annotators whereas Brew et al. [7] have found that good annotators are valuable for training and defines good annotators as those that have the highest agreement with the consensus rating. There have also been considerable directions into addressing the challenge of learning the ground truth from multiple possibly noisy labels [28, 29, 32]. Raykar et al.’s [29] approach is to learn a classifier from the multiple annotations using maximum likelihood estimation and estimating the ground truth and the annotator performance is a byproduct of their proposed algorithm. From the point of view of getting annotations, the challenge of balancing the coverage of the assets with the quality of the labels is investigated by Brew et al. [7] who conclude that fewer annotators are needed in domains with high consensus.

A practical issue with using crowdsourcing services such as Mechanical Turk is to analyse the trustworthiness of the users who perform the tasks. Current research shows that the numbers of untrustworthy users is not large, normally a small subset produces most of the invalid input [23]. There is evidence of a number of different techniques used to guard against malicious or lazy users. Some research requires users to show some degree of accuracy on a small test subset [4] while other work uses the percentage of previously accepted submissions from a user in order to determine his or her trustworthiness and motivation [4, 20]. Kittur et al. [23] recommends including explicitly verifiable questions to reduce invalid responses and increase time-on-task.

4 Conclusions

In this position paper we have proposed using crowdsourcing as a mechanism for rating emotional speech assets. The use of crowdsourcing is relatively novel in domains where it can be impossible or too expensive to get the actual label or ground truth and there has been significant research into using machine learning techniques to address the challenges of learning the ground truth labels and learning annotator bias in crowdsourced data. The subjectivity of rating emotional assets offers opportunities for use of these techniques to create datasets of quality labelled speech assets for use in a number of research areas. The area that is of most interest to the authors of this position paper is the use of these quality assets in the classification and prediction of emotion in speech. Below we have included a use case that reflect our requirements of an emotion markup language.

Seán has a set of speech assets extracted from recordings of experiments using mood induction procedures. He wants to get these assets rated on a number of different scales, including activation and evaluation, by a large number of non-expert annotators. He wants to use a micro-task system such as Mechanical Turk to get these ratings. Active learning will be used to select the most appropriate assets to present for labels from the annotators. He will then analyse and evaluate different techniques for identifying good annotators and determining consensus ratings for the assets which will be used as training data for developing prediction systems for emotion recognition.
References


