Using crowdsourcing to provide prosodic annotations for non-native speech

Keelan Evanini and Klaus Zechner

Educational Testing Service
Princeton, New Jersey, USA
KEvanini@ets.org, KZechner@ets.org

Abstract

We present the results of an experiment in which 2 expert and 11 naive annotators provided prosodic annotations for stress and boundary tones on a corpus of spontaneous speech produced by non-native speakers of English. The results show that agreement rates were higher for boundary tones than for stress. In addition, a crowdsourcing approach was implemented to combine the naive annotations to increase accuracy. The crowdsourcing approach was able to match expert agreement for stress (62.1%) with 3 naive annotators, and come within 7.2% of expert agreement for boundary tones (82.4%) with 11 naive annotators. This experiment also demonstrates that noticeable improvements in naive annotations can be obtained with a small amount of additional training.

Index Terms: crowdsourcing, prosodic annotation, stress, boundary tone

1. Introduction

The recent availability of cheap and fast labor through the Amazon Mechanical Turk interface has led to several experiments that have used naive annotators to perform linguistic tasks that have traditionally been performed by trained experts. In order to obtain more accurate annotations, some researchers combine multiple naive annotations for the same item in a crowdsourcing approach. In many cases, these results have matched expert performance, even for complex tasks such as transcription of non-native speech [1], grammatical error detection [2], textual entailment [3], and many others.

In this study, we consider the task of annotating utterance-level stress (prominence) and boundary tones in spontaneous speech produced by non-native speakers of English. Since this task is quite difficult compared to most other linguistic annotation tasks that use crowdsourcing with naive annotators, we partnered with an off-shore outsourcing company for the task instead of using Amazon Mechanical Turk. This enabled us to have more control over the demographics of the workers as well as provide them with more detailed feedback.

The performance of naive annotators on the task of annotating prominence and boundary tones has been studied before ([4], [5], and [6]). However, these studies have not provided comparisons between the naive annotations and a set of expert annotations. In addition, no previous studies have used a crowdsourcing approach in an attempt to improve the quality of naive annotations for these tasks. This study builds on previous research by addressing these two questions.

2. Data

The speech data used in this experiment consist of responses to an English proficiency assessment for non-native speakers. The responses are either 45 or 60 seconds in duration and contain spontaneous speech in response to open-ended questions. Table 1 summarizes the characteristics of this data set.

# of responses	50
# of speakers	33
# of words	5641
Audio duration	45.1 minutes

Table 1: Characteristics of the data set

3. Methodology

3.1. Expert Annotation

Two native-speaker linguists with training in phonetics and phonology each provided independent annotations for the set of 50 responses. For the stress annotations, they were asked to label words which received utterance-level stress. For the boundary annotations, they were asked to label words preceding a strong prosodic juncture (corresponding to a ToBI break index of 4) with one of two boundary tone labels: -L% (falling) or -H% (rising). For both tasks, the annotators viewed the waveform and spectrogram for the response in Praat along with word and phoneme boundaries produced by forced alignment. They were able to listen to the audio stimulus as many times as was necessary to provide their annotations.

3.2. Naive Annotation

A team of 11 annotators was arranged through a contract with an off-shore outsourcing company. All annotators were highly proficient non-native speakers of English and had obtained university-level educations. They were provided with brief written guidelines about the annotations tasks. For the stress task, they were instructed to mark words that "sound like they receive the most stress from the speaker." For the boundary task, they were instructed to mark the "final word of an intonational phrase," and were told that intonational phrases usually coincide with clauses or sentences. They were told to provide boundary tone labels of -L% and -H% to correspond to perceived falling and rising intonation, respectively. For both tasks, they were provided with a few additional guidelines; for example, that stressed words are often louder and longer in duration than unstressed words and that clause boundaries are typically followed by a pause. In addition to the annotation guidelines, they were provided with a set of 3 responses containing gold standard annotations (N=411) that were produced by having a third annotator adjudicate the annotations from the two experts. In the annotation task, they were presented with an orthographic transcription of each response in a spreadsheet with one word per

row. They were able to listen to the response as many times as was necessary to provide their annotations.

After receiving the training material, the naive annotators provided annotations for a set of 3 calibration responses (N=340) to ensure that they had understood the annotation guidelines. The performance of all 11 annotators on this set was deemed acceptable to let them participate in the experiment. The annotators were paid $0.80 per audio minute of annotation (including both the stress and boundary tone tasks), and the annotators completed the annotations in approximately 10 times real time.

3.3. Crowdsourcing

To evaluate the effectiveness of crowdsourcing for the annotation tasks, a voting procedure was implemented to combine the naive annotations. For each $n \in \{1, ..., 11\}$, all unique combinations of n naive annotators were selected, resulting in $\binom{11}{n}$ sets of annotations. For each set, the n annotations were used to produce a crowdsourced annotation for each word by majority vote (ties were broken by random choice). Then, the agreement rates between each of the $\binom{11}{n}$ crowdsourced annotations and the expert annotations were calculated, and the average κ for each n was determined to represent the estimated performance when n naive annotators are used.

3.4. Re-annotation

After the first set of annotations from the 11 naive annotators was obtained, a second round of training and re-annotation of the same 50 responses was conducted in an attempt to see whether this would lead to an improvement in performance. In the second round of training, the naive annotators were supplied with an additional set of 3 gold standard annotations (N=357), and then asked to re-annotate the 50 responses, using their initial annotations as a starting point. No explicit feedback was provided about the 50 responses, and the guidelines were not changed. The improvement in agreement rates between the naive and expert annotators after this second round of training demonstrate the effect of additional training on naive annotators.

4. Results

4.1. Frequency of Annotations

Table 2 provides the frequencies of stress and boundary tone annotations for the two expert annotators. As the table shows, Expert2 labeled a larger number of words (12% more) as stressed than Expert1. Both expert annotators labeled approximately 6% - 7% of the words with a falling boundary tone, and labeled very few words with a rising boundary tone.

Annotator	Stress		Boundary Tone		
	None	Stressed	None	-L%	-H%
Expert1	0.575	0.425	0.936	0.057	0.007
Expert2	0.455	0.545	0.930	0.068	0.002

Table 2: Frequencies of expert annotations

Table 3 provides descriptive statistics (mean, minimum, maximum, and standard deviation) for the frequencies of annotations from the 11 naive annotators. As the table shows, all of the naive annotators labeled more words as stressed than Expert1 (the minimum frequency of stress labels among the naive annotators was 0.487, which was greater than Expert1's frequency of 0.425). Expert2's frequency of 0.545 for stressed words falls directly in the middle of the distribution for the naive annotators (5 naive annotators had a smaller frequency, and 6 had a greater frequency).

Task	Label	Mean	Min.	Max.	S.D.
Stress	Stressed	0.557	0.487	0.636	0.046
Boundary	-L%	0.048	0.036	0.065	0.009
	-H%	0.011	0.004	0.018	0.005

Table 3: Descriptive statistics for frequencies of naive annotations (N=11)

The naive annotators, in general, provided a slightly smaller number of falling boundary tones labels and a slightly greater number of rising boundary tone labels than the experts, although the overall frequencies of boundary tone labels between the two sets of annotators were roughly equivalent.

In order to understand the behavior of the annotators in more detail, the words in the data set were labeled as *function words* (conjunctions, determiners, prepositions, and pronouns), *content words* (all other parts of speech), and *disfluencies* (filled pauses and partial words). Figure 1 shows the frequencies of stress annotations on function and content words for all 13 annotators. As the figure shows, both expert annotators had lower frequencies of stress annotations on content words than the 11 naive annotators. This suggests that the naive annotators were relying somewhat too heavily on the transcriptions in their decisions about stress annotations and not enough on the audio stimuli. For the function words, on the other hand, the two expert annotators showed markedly different behavior: Expert1 labeled 10.8% of function words as bearing stress (lower than all 11 naive annotators), whereas Expert2 labeled 30.6% (greater than all but 2 of the naive annotators). Expert2's behavior was also somewhat anomalous with regard to disfluencies: Expert2 labeled 19.7% of the disfluencies as bearing stress, whereas Expert1 and the 11 naive annotators all had rates around only 1% for the disfluencies (the guidelines for the naive annotators stated that disfluencies are generally not prominent).

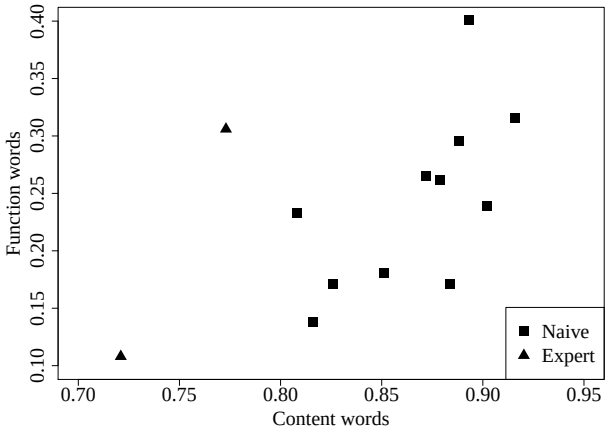

Figure 1: Frequency of stress annotations by lexical type

The annotations for boundary tones patterned similarly with regard to the different lexical types for all annotators: annotation frequencies ranged from 8%-12% for content words and 1% to 2% for function words for both the expert and naive an-

notators (tone boundary annotations on disfluencies were negligible for all annotators).

As Tables 2 and 3 show, the frequency of rising boundary tone annotations (-H%) is quite low in both groups. Due to this fact, subsequent analyses merge the labels -H% and -L% and report on the agreement about the presence or absence of a boundary tone, regardless of its type.

4.2. Agreement

In addition to the annotation frequencies presented in the previous section, it is also necessary to measure inter-rater agreement, in order to measure the similarity of the annotators' behaviors. Previous studies have reported agreement rates either among expert annotators or among naive annotators. As a comparison, Table 4 presents those values for this study. The expert agreement rate was calculated using Cohen's κ (for pairwise agreement) and the naive agreement rate was calculated using Fleiss' κ (for multi-way agreement).

Annotators	Stress	Boundary Tone
Naive (N=11)	0.464	0.787
Expert (N=2)	0.612	0.824

Table 4: κ values for expert and naive annotators

The agreement rates achieved by both groups of annotators are comparable with those obtained in other studies for both naive ([4], [5], and [6]) and expert ([4], [7]) annotations (although the specific tasks in each study were slightly different). Table 4 also shows that both naive and expert annotators show higher levels of agreement for boundary tones than for stress. This result is also consistent with the findings from [4], [5], [6], and [7].

While the inter-rater agreement among the naive annotators can show how consistently they behave, it does not necessarily indicate how useful the naive annotations are (for example, the naive annotators could have a very high agreement rate simply due to the fact that they all mis-interpreted the task in the same way). In order to investigate how well the naive annotators actually performed the tasks, it is necessary to compare their annotations with the expert annotations. For this purpose, pairwise κ values were calculated for each of the 11 naive annotators with each of the two experts. Table 5 summarizes these results for both the stress and boundary tone annotations.

Task	Expert	Mean	Min.	Max.	S.D.
Stress	Expert1	0.618	0.524	0.675	0.048
	Expert2	0.505	0.427	0.526	0.044
Boundary	Expert1	0.728	0.645	0.790	0.050
	Expert2	0.680	0.608	0.724	0.036

Table 5: Descriptive statistics for pairwise κ values between all naive annotators (N=11) and both experts

Table 5 shows that the naive annotators consistently agreed more closely with Expert1 than with Expert2. This suggests that Expert2's annotations may not have followed the guidelines as closely as Expert1's (Expert2's anomalous behavior with regard to stress annotations was already mentioned in Section 4.1).

A comparison between the naive-expert agreement values in Table 5 and the expert-expert agreement values in Table 4 shows that many of the naive annotators attained the expert-expert agreement rate on the stress task when Expert1 is used as the gold standard. The expert-expert agreement rate for stress annotations is 0.612, and the mean naive-expert agreement rate is 0.618 (6 out of the 11 naive annotators had an agreement rate greater than 0.612 with Expert1). For the boundary tone annotations, however, all of the naive annotators fall short of the expert-expert agreement level of 0.824: the mean naive-expert κ on this task is 0.728, and the highest value for a single naive annotator is 0.790 when Expert1 is used as the gold standard.

4.3. Crowdsourcing results

This section presents the results of using crowdsourcing with the naive annotators according to the methodology described in Section 3.3. Figure 2 presents the results for the stress annotation task. The figure shows how the average κ values increase monotonically as the number of naive annotators is increased from 1 to 11. The figure plots the average κ values for comparisons with gold standard annotations from both Expert1 and Expert2, and the dotted line shows the expert-expert agreement rate of 0.612. When Expert1 is used as the gold standard, the average κ values increase from 0.609 ($n = 1$) to 0.679 ($n = 11$), and the performance of the naive annotators surpasses the expert-expert agreement at $n = 3$.

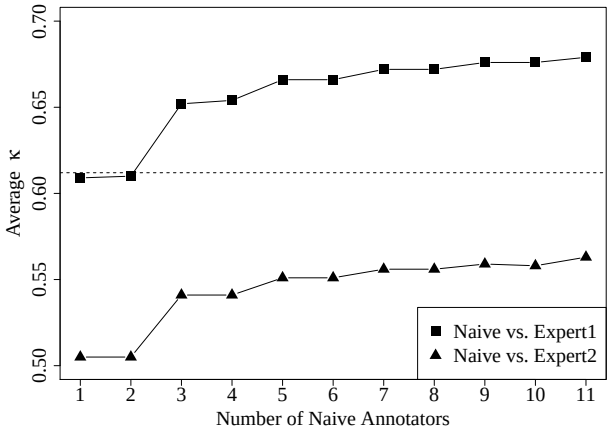

Figure 2: Average κ values for crowdsourcing results in the stress annotation task

Figure 3 presents the crowdsourcing results for the boundary tone annotation task. This figure again shows a monotonic improvement in the average κ value as the number of naive annotators was increased. In the case of the boundary tones, however, the performance of the crowdsourced annotations does not reach the expert agreement level of 0.824, and reaches a maximum average agreement of 0.765 when 11 naive annotators are compared to Expert 1 (a relative difference of 7.2%). Furthermore, the rate of improvement for the boundary tone task is nearly flat between $n = 3$ and $n = 11$, whereas it continued to show substantial increases for the stress annotation task. This fact suggests that the behavior of the 11 naive annotators is more uniform in the boundary tone annotation task; thus, the addition of more naive annotations does not provide as much as a benefit as it did for the stress annotation task.

For both tasks, Figures 2 and 3 show that the largest single increase in performance is obtained when n is increased from 2 to 3. This result has been found in several other tasks involving annotations of linguistic phenomena, such as preposition error detection [2] and textual entailment [3]

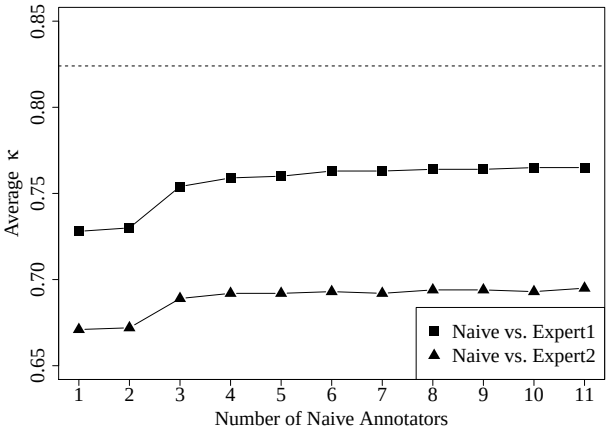

Figure 3: Average κ values for crowdsourcing results in the boundary annotation task

4.4. Effect of Additional Training

As described in Section 3.4, the annotators re-annotated the 50 responses after reviewing a second set of 3 gold standard annotations. Figure 4 shows the changes in the pairwise κ values (compared to both experts) obtained by each naive annotator for each task after this second pass. The figure shows that nearly all of the κ values improved, and that the only negative changes were small in magnitude (-1% or less). Some individual annotators showed large improvements after the second round of training; for example, the first annotator in Figure 4 showed an improvement of 0.146 in the stress annotation task (vs. Expert1) and the third annotator in Figure 4 showed an improvement of 0.121 in the boundary tone annotation task (vs. Expert 1).

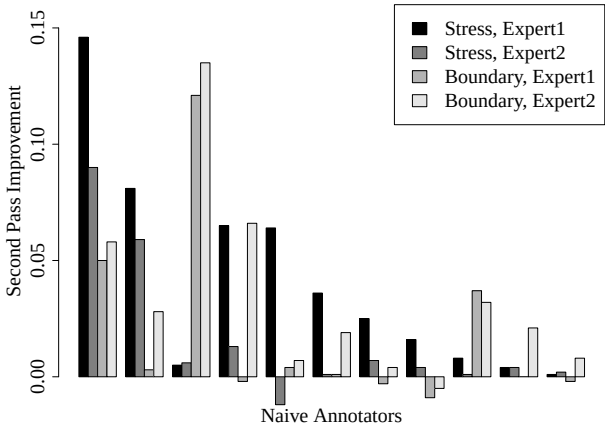

Figure 4: Improvment of naive annotations after second pass

Finally, Table 6 presents a summary of the pairwise κ values for all naive annotators after the second annotation pass. A comparison with the results after the first pass in Table 5 shows that the overall performance improved for both tasks. The mean κ values improved by 3.2% for stress and 1.6% for boundary tones compared to Expert1's annotations, and by 2.3% for stress and 2.5% for boundary tones compared to Expert2.

Task	Expert	Mean	Min.	Max.	S.D.
Stress	Expert1	0.650	0.598	0.695	0.032
	Expert2	0.521	0.435	0.581	0.035
Boundary	Expert1	0.751	0.655	0.787	0.035
	Expert2	0.705	0.615	0.737	0.033

Table 6: Descriptive statistics for pairwise κ values between all naive annotators (N=11) and both experts, second pass

5. Conclusions

This study has shown that using naive annotators is a reasonable and cost-effective way of obtaining prosodic annotations of stress and boundary tones for non-native speech. Despite the difficulty of the task, some of the individual naive annotators were able to achieve agreement rates that matched the expert-expert agreement for the stress annotations. When a crowdsourcing approach was used for this task, we showed that only 3 naive annotators were needed to match the level of agreement attained by the two experts. For the boundary tone annotation task, the naive annotators still fell a little short of the expert-expert agreement even when all 11 annotations were used in a crowdsourcing approach. However, the ultimate test of the usefulness of the naive annotations is whether they can be used to produce meaningful results in a supervised classification framework (similar to the study described in [8] which used crowdsourcing of naive annotators in a regression system to rate the proficiency of non-native speakers).

Finally, this study also demonstrated that meaningful improvements in the performance of naive annotators could be achieved by a small amount of additional training. This suggests that future studies employing naive annotators for complex tasks might benefit from using a service that allows a greater amount of interaction with the annotators than Amazon Mechanical Turk; in future work, we plan to explicitly compare the results from these two approaches.

6. References

[1] M. Marge, S. Banerjee, and A. I. Rudnicky, "Using the amazon mechanical turk for transcription of spoken lanugage," in *Proc. ICASSP*, 2010.

[2] J. Tetreault, E. Filatova, and M. Chodorow, "Rethinking grammatical error annotation and evaluation with the amazon mechanical turk," in *Proc. NAACL HTL Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, 2010.

[3] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks," in *Proc. EMNLP*, 2008.

[4] C. J. Mayo, "Prosodic transcription of glasgow english: an evaluation study of glatobi," Master's thesis, University of Edinburgh, 1996.

[5] J. Buhmann, J. Caspers, V. J. van Heuven, H. Hoekstra, J.-P. Martens, and M. Swerts, "Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the spoken dutch corpus," in *Proc. LREC*, 2002.

[6] Y. Mo, J. Cole, and E.-K. Lee, "Naive listeners' prominence and boundary perception," in *Proc. Speech Prosody*, 2008.

[7] T.-J. Yoon, S. Chavarría, J. Cole, and M. Hasegawa-Johnson, "Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI," in *Proc. Interspeech*, 2004.

[8] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "How many labellers? modelling inter-labeller agreement and system performance for the automatic assessment of non-native prosody," in *Proc. ISCA Workshop on Speech and Language Technology in Education*, 2010.