TOWARD BETTER CROWDSOURCED TRANSCRIPTION: TRANSCRIPTION OF A YEAR OF THE LET'S GO BUS INFORMATION SYSTEM DATA

Gabriel Parent, Maxine Eskenazi

Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA USA
gparent@cs.cmu.edu, max@cs.cmu.edu

ABSTRACT

Transcription is typically a long and expensive process. In the last year, crowdsourcing through Amazon Mechanical Turk (MTurk) has emerged as a way to transcribe large amounts of speech. This paper presents a two-stage approach for the use of MTurk to transcribe one year of Let's Go Bus Information System data, corresponding to 156.74 hours (257,658 short utterances). This data was made available for the Spoken Dialog Challenge 2010 [1][1]. While others have used a one stage approach, asking workers to label, for example, words and noises in the same pass, the present approach is closer to what expert transcribers do, dividing one complicated task into several less complicated ones with the goal of obtaining a higher quality transcript. The two stage approach shows better results in terms of agreement with experts and the quality of acoustic modeling. When "gold-standard" quality control is used, the quality of the transcripts comes close to NIST published expert agreement, although the cost doubles.

Index Terms— Crowdsourcing, speech recognition, spoken dialog systems, speech data transcription

1. INTRODUCTION

The creation of the Fisher Corpus (2000 hours) was motivated by the need for a greater volume of conversational telephone speech that could help bring automatic speech recognizers (ASR) to a new level. In order to reduce the cost of transcriptions, the Quick Transcription (QTr) specification was used, where the transcribers aren't asked to provide detailed transcriptions such as punctuation or background noise. The new guidelines, in conjunction with automatic segmentation, enable transcription to be carried out in 6 times real time, bringing the average cost to $150/hour of speech.

Even at this rate, as the amount of data increases, it becomes prohibitively expensive to obtain high quality transcripts. An alternative is to use a large amount of unlabeled data to boost recognizer accuracy. In unsupervised acoustic model training [2], unlabelled data was recognized using a seed model, and the well-recognized audio (as judged by a confidence model) was reused as training data to train a new model. Multiple iterations can be carried out. [3] use this unsupervised approach on broadcast data, additionally aligning the seed model results with available closed captions.

More recently, crowdsourcing has been used to obtain speech labels at a fraction of the cost of traditional methods. Gruenstein et al [4] used an educational game in parallel with MTurk to collect orthographically transcribed speech with near-expert accuracy. [5], [6] and [7] have also investigated the use of MTurk to transcribe speech. They all show high levels of agreement with their experts' annotations. While the cost varies from one study to another (between $5/hour for [7] and $20/hour for [4]), the average cost is still one order of magnitude less than the traditional cost of in-house or contractor transcription.

This paper presents a two stage approach to transcription. The goal is to increase the quality of the transcripts while decreasing their cost. In the first stage, the MTurk workers see the best ASR hypothesis and are asked whether that transcription exactly matches what they hear in the corresponding audio file. In the second stage, only the portion that did not exactly match the ASR output is transcribed. This paper also addresses quality control using a "gold-standard" (GS). We present both a static and a dynamic method of using the GS to improve the quality of the transcription.

2. BACKGROUND

The Let's Go Bus Information System is a spoken dialog system used nightly by the Port Authority of Allegheny County. Callers obtain bus schedule and planning information. The system "went live" in March of 2005. The year of data described in this paper is from October 2008 to September 2009 and contains a total of 18,070 dialogs. The dialogs were segmented in real time to produce 257,658 user turns, totaling 156.74 hours (average utterance length is 2.2 seconds). Table 1 shows a sample dialog.

[1] The data is available through the Dialog Research Center (DialRC, http://www.dialrc.org/). Contact the authors.

System:	WELCOME TO CMU LET'S GO BUS INFORMATION SYSTEM. WHAT CAN I DO FORYOU?
User:	*I'd like to go to Squirrel Hill.*
System:	GOING TO SQUIRREL HILL. IS THIS CORRECT?
User:	*Yeah.*
System:	WHERE ARE YOU LEAVING FROM?
User:	*I'm leaving from CMU.*
System:	LEAVING CMU. AT WHAT TIME DO YOU WANT TO LEAVE?
User:	*Now.*
System:	LET ME LOOK THAT UP FOR YOU. THE NEXT 61C LEAVES FORBES AND CRAIG AT […]

Table 1. *Example of a dialog in Let's Go SDS*

The quality of the transcriptions obtained using crowdsourcing in previous studies varies considerably. In [5], the workers provided transcripts that had an average 4.96% word error rate (WER) compared to an in-house annotated GS. The authors noted that the subset of non-native speech shows a higher level of disagreement with the GS. Crowdsourced transcription of non-native speech is investigated in more detail by [6]: they obtain an average WER of 15.5% on three different datasets. Finally, using a similar interface (i.e., plain text input box), [7] gives a WER of 23% between the workers' transcripts and the Switchboard corpus expert annotations. This variance may be due to the difference in the type of application (e.g., size of the vocabulary).

The above WERs were calculated using a transcript provided by only one worker per utterance. One could also ask multiple workers to transcribe the same audio utterance and aggregate the results. ROVER [8] is well suited for this task since it was first designed to combine the results from multiple ASRs. ROVER aligns transcriptions into word transition networks where null transitions are allowed. It then selects the most frequent word in every aligned set as the output word. [5] and [6] used ROVER for the transcripts from 5 workers and got a WER reduction of 2.8% and 3.9% respectively (relative improvement of 56% and 25%). [7] seems to have benefitted less from aggregating results, since applying ROVER on 3 workers' transcripts only brought the WER down from 23% to 21% (7% relative improvement). [6] tried two other aggregation methods: the first one involved finding the longest common subsequence; the second one found the optimal path through a word lattice built with multiple transcriptions. While on average ROVER performs the best, the lattice approach seems to perform better than ROVER on spontaneous speech.

Merging the work from multiple MTurk workers therefore provides better quality transcripts. However, [7] shows that for acoustic modeling, it is better to obtain more lower quality data from one worker than less higher quality data from multiple workers. They trained one ASR using 20 hours of "high-quality" crowdsourced data (transcribed by 3 different workers and then aggregated it using ROVER) and another recognizer with 60 hours of data (transcribed by only one worker). The second ASR outperformed the first by 3.3% WER (37.6% and 40.9% respectively, for a relative improvement of 8%).

Despite this result, for research where better transcriptions are essential, a better quality control mechanism must be used. For example, if there is little available data, or if the transcribed data is difficult or expensive to obtain (e.g. speech-impaired, children's speech or speech from a small group of individuals in a very remote area), WER improvement comes with the extra expense of quality control. Applications where the end product is the transcription itself (e.g., closed captioning, dictation) also need to produce high quality transcripts. As mentioned above, using techniques that merge multiple transcripts is one way to improve the quality. Another is unsupervised quality control [7], where the workers' skills are evaluated by comparing their transcripts with the transcripts of other workers. [7] showed that it is better to use a disagreement estimation to reject bad workers rather than to find the right threshold to accept only good workers. [9] investigated the use of a GS to improve the classification of noisy data: this work has lead to an open source tool [10] which uses the EM algorithm "to infer the 'true' results and the underlying quality of the workers". This approach cannot be naively transferred to speech transcription since the latter is not a classification task. This paper will evaluate the use of a two-stage approach that uses a GS to improve the quality of transcriptions obtained through MTurk.

3. TWO-STAGE APPROACH

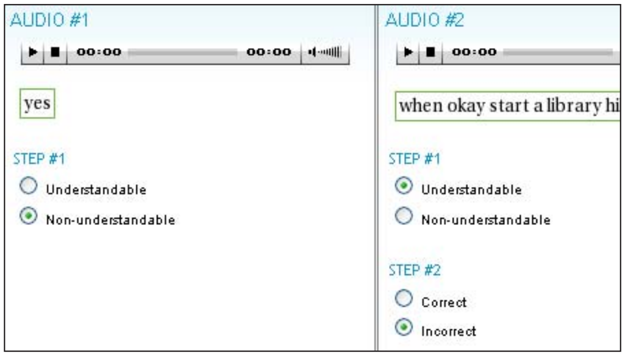
If not properly designed, transcription tasks can be very cognitively demanding. [11] presents a simple resource model that describes how humans allocate cognitive resources when doing problem solving while listening to audio. The transcriber first needs to allocate short-term memory resources for speech input, and then uses the remaining resources for problem solving: providing the annotation. If the annotation task involves too many subtasks (e.g., provide the orthography of the words, detect false starts, detect and classify different noises, etc.), short-term memory becomes saturated and the worker either provides low quality annotation or has to listen to the audio two or more times. Also, the worse the quality of the audio recording, the more resources need to be allocated for speech input, thus leaving fewer resources for problem solving. One way to circumvent potential problems is to provide shorter recordings to prevent short-term memory saturation.

The goal of this two-stage approach is to separate the annotation process into different subtasks in order to avoid resource saturation and obtain better quality transcripts. In the first stage, we provide the workers with 10 audio recordings per task (called Human Intelligence Task or HIT) and, for each recording, the best corresponding recognizer hypothesis. For each recording, they are asked to indicate if it is *understandable* or *non-understandable.* The instructions define understandable as an utterance where each and every word is clearly intelligible. If they select *understandable,* the interface asks them to identify whether the transcript (ASR output) is *correct* or *incorrect* (Figure 1). C*orrect* means that the transcript corresponds word by word to what was said in the audio. If this is not the case, the utterance is marked as *incorrect*. Every utterance is thus classified in one of the 3 categories: *understandable and correct* (UC), *understandable and incorrect* (UI) and *non-understandable* (NU). Table 2 presents an example of the pairs that the workers would hear and see for each category.

Category	Workers hear:	Workers see:
UC	*I need the next bus*	I NEED THE NEXT BUS
UI	*Yes, I do*	NO I DON'T
NU	Only noise, or unintelligible speech	LEAVING FROM FORBES AND MURRAY

Table 2. *Sample utterance text and audio pairs and their classification*

Since the first stage does not require the workers to provide actual transcriptions, they have more resources available to do a good job of classifying the utterances, and accomplishing this more quickly. Thus this costs less than a complete transcription HIT. This classification task also identifies the recordings that the ASR did not recognize correctly which are completely intelligible. Those utterances are sent to the second stage where the workers are asked to type in each and every word they hear. Since the workers do not have to worry about background noise and false starts and mumbles, they have more resources available to provide good orthographic transcriptions. This leaves the *non-understandable* portion of the data unlabeled (found to be 17%, see details in Section 5). Since most of the utterances in this portion of the data contain noises and false starts, it is very useful for training more specialized models. This data annotation will be done in a third stage, where workers will use an interface that enables them to mark different kinds of noises, etc.


AUDIO #1

yes

STEP #1

○ Understandable

● Non-understandable

AUDIO #2

when okay start a library hi

STEP #1

● Understandable

○ Non-understandable

STEP #2

○ Correct

● Incorrect

Figure 1. *The first stage interface*

4. QUALITY CONTROL

4.1. Gold-standard

One "gold-standard" (GS) utterance was inserted for every 9 unknown utterances in each 10-utterance HIT of both stages. For the first stage, 3 different experts labeled 2540 utterances as being UC, UI or NU. The 2119 utterances for which the 3 experts all agreed on the same label constitute the GS for the first stage. For the second stage, 2 experts transcribed 890 utterances. The transcriptions were normalized in order to account for different possible orthographies (e.g., number, time, places). The two experts had exact agreement on 696 utterances; this set was selected as the GS for the second stage. Workers can be evaluated using Kappa in the first stage and word error rate in the second stage.

4.2. Distribution of good workers

The Kappas of each of the 991 workers who submitted **at least four HITs** in the first stage were computed and, while the majority (57.3%) of workers display a high Kappa (> 0.70), there is still a non-negligible portion of the workers with a Kappa below 0.70. The better workers also submitted more HITs than the "bad workers" (Kappa < 0.7): 66.7% of the HITs were submitted by workers with a Kappa of over 0.7. Figure 2 shows the distribution of workers' Kappas and of the HITs based on the Kappa of the worker who submitted them.

This phenomenon is even more pronounced in the second stage where workers were asked to transcribe rather than just to classify. While only 60% of the workers had a WER smaller than 15%, these workers submitted 81.1% of all the HITs (Figure 3). We could hypothesize that while there is still a non-negligible proportion of workers who provide bad results, they tend to work less and thus affect the overall quality less. However, in both stages, there is still a large amount of data (33.3% and 18.9%) that is provided by workers whose submission quality is questionable and thus there is room for improvement by means of quality control.

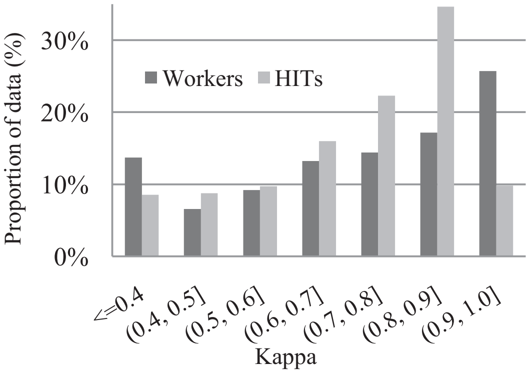


Figure 2. *Quality and workload of the workers, first stage*

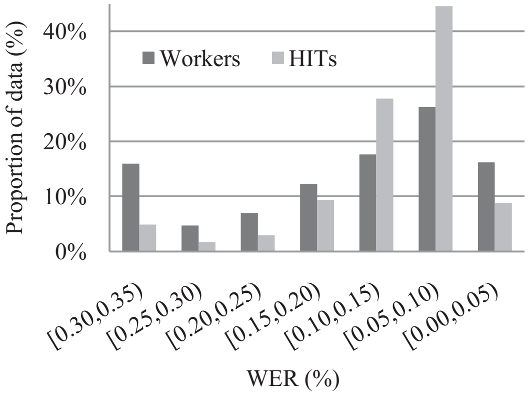


Figure 3. *Quality and workload of the worker, second stage*

4.3. Self-confidence as a quality control

Instead of inserting a GS in every HIT, we might be able to avoid the extra expense simply by asking how confident the workers are in each of their answers. To investigate this possibility, the workers were asked to check a checkbox containing the judgment "I'm not 100% confident of my answer" for any transcript they weren't sure of (without being penalized monetarily). The box was checked for 6% of the transcripts, which had an average WER of 0.619. The other 94% of the transcripts had an average WER of 0.086, thus indicating that self-confidence could be used as a first filter to improve the quality of the data. The comparison between workers' WER on the GS and self-confidence ratio (how many times they indicated they weren't sure, out of the total utterances transcribed) shows a correlation of 0.37, which is not sufficient to warrant the use of this filter with a high degree of confidence to eliminate workers.

4.4. Gold-Standard quality control

We implemented two quality control mechanisms for the first and second stage of the speech transcription task. Unlike [7] who used unsupervised quality control, we investigated the use of inserting one gold-standard utterance in each HIT. The GS dataset was divided in two, one was kept as the development set, used to evaluate the workers, and the other was used for testing purposes.

The first mechanism that was tested is simple **static rejection**, where the worker is below some threshold. (By rejection, we mean not including that worker in the final data set[2]). One approach is to have a conservative threshold, keeping only the workers with a score higher than the median score (*Perc. 50*), and another is to have a more relaxed threshold, where the 25th percentile (*Perc. 25*) score is used as the lower bound. Obviously this approach has the drawback of lowering the quantity of data (see exact quantities in Section 5).

In order to compensate for this fact, we tried a **dynamic rejection** approach where, when a HIT is rejected because the worker's score is below the threshold, that HIT is resubmitted to MTurk and put through the two stage process once more. The tradeoff is that extra submissions to MTurk make this mechanism more expensive.

5. RESULTS

This section first presents results for the first stage classification task, and then provides results for the complete two stage transcription. Table 3 provides an overview of the data processed on MTurk.

	First-stage	Second-stage
Unknown utterances	257,658	73,643
GS utterances	25,766	7,364
Total utterances	283,424	81,007
Utterances/HIT	10	10
$ per HIT	0.05	0.10
Throughput	6160 utts. /hour	3903utts. /hour

Table 3. *Summary of the HITs*

5.1. First-stage: classifying the ASR output

As mentioned in Section 4.1, three experts classified 2540 utterances and unanimously agreed on 2119 of them. By comparing the experts amongst each other, we obtain an **interannotator agreement value (ITA),** which represents how much the annotators agreed with one another. In the case of a classification task, the ITA is obtained by averaging the Kappa calculated between each pair of experts. For example, in the case of the simple classification of whether an utterance was *understandable* or *non-understandable,* expert1 (E1) and expert2 (E2) have a reciprocal Kappa of 0.73, while E1 and E3 have a Kappa of 0.77 and E2 and E3 have 0.71. By averaging these three Kappas, we obtain the ITA of 0.74. A crowd of good workers is expected to agree with the experts with an average agreement close to the ITA.

To evaluate whether the crowd can achieve a comparable ITA, we aggregated, with majority vote, three workers' answers to the same 2540 utterances completed by the experts. The rightmost column of Table 4 presents the ITA for the classification of an utterance as *UC, UI* or *NU*, and also shows the result for the two sub-classification tasks (understandable vs. non-understandable and correct vs. incorrect). Columns E1, E2 and E3 in Table 4 present the agreement between the crowd and the three experts (and the next column presents the average agreement) for the same three classification tasks.

[2] All workers were paid for the tasks with only few exceptions, such as one worker who transcribed 6890 utterances as "not clear", and workers who had a Kappa below 0.35 on the first stage. Workers with a Kappa above 0.9 were awarded bonuses.

	E1	E2	E3	Avg. 3 experts	ITA
Understandable vs. Non-Und.	0.70	0.74	0.68	**0.71**	**0.74**
Correct vs. Incorrect	0.90	0.88	0.86	**0.88**	**0.92**
Classification in *UC/UI/NU*	0.72	0.74	0.70	**0.72**	**0.76**

Table 4. *ITA and crowd-experts agreement*

There is strong agreement amongst experts on overall utterance classification (0.76), but there is even stronger agreement on whether the ASR output should be classified as correct or incorrect (0.92). On the three classification tasks, the crowd does not achieve an ITA that is as high as that of the experts. However, the Kappas are all higher than 0.70 and so the crowd's data can be considered to be fairly reliable.

Past studies [12] showed that adding people to the crowd could increase the agreement with experts. To verify if this was the case in this study, 11 workers were asked to complete the first stage for the GS utterances. Figure 4 shows the relationship between the size of the crowd and the quality of the classification. Once again, majority vote was used and ties were resolved randomly. Adding more than 3 workers to the crowd yields a maximum Kappa improvement of 0.02, and thus does not seem to be worth the additional expense.

Finally, out of the 257,658 utterances that were processed, 139,185 (**54%**) were labeled as *understandable correct*, 73,643 as *understandable incorrect* (**29%**) and 44,829 (**17%**) as *non-understandable*.

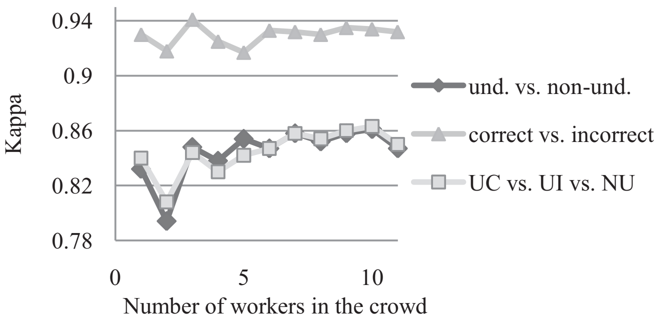


Figure 4. *Crowd size effect*

5.2. Overall quality of the transcribed corpus

NIST reports that the average disagreement between 2 expert transcribers is between 2 and 4% WER [13]. As described in Section 2, transcriptions obtained through crowdsourcing in previous studies have been reported to be between 5% and 25% WER, depending on the application. For the sake of comparison, workers were paid to transcribe the test GS utterances using a one-stage interface similar to that used in [5] [7]. Table 2 presents the WER of the test GS (Section 4.4) for the two stage approach and for the two quality control mechanisms described in Section 4.4. Sphinx3 [14] was used to build acoustic models for each condition (13th-order MFCC with 1st-order and 2nd-order delta MFCC, semi-continuous HMMs, 256 Gaussians, 4000 senones). These models were tested on a held-out dataset as an additional way to compare the quality of the transcripts obtained.

	Thresh. (Perc.)	Hours	$/Hour	Expert WER	ASR WER
1-stage	N/A	60	17.4	13.7%	64.6%
2-stage no qual. control	N/A	60	14.5	8.1%	62.3%
2-stage **static rejection**	25	47.4	14.5	6.6%	65.1%
	50	25.1	14.5	3.1%	67.5%
2-stage **dynamic rejection**	25	60	16.8	6.9%	62.4%
	50	60	28.9	5.4%	62.1%

Table 5. *Results of 1-stage vs. 2-stage and quality control*

For dynamic rejection with a threshold set at the 25th percentile of the workers' scores, it took an average of 1.3 HIT requests for the first stage and 1.1 for the second stage. When the threshold was set to percentile 50, the average number of HITs for the first stage was 2.5 and 1.5 for the second.

6. DISCUSSION

If we first evaluate the quality of the different approaches by considering the disagreement with the experts, the two stage approach presented here has a 5.6% lower WER than an approach where all of the annotation task is completed in one stage (a relative improvement of 41%). It is also 17% cheaper since a large proportion of the data (54%) was correctly recognized by the ASR and did not have to go to the second stage. When quality control is applied, the disagreement with experts goes below 8% and approaches NIST inter-expert measures. As expected, the transcripts obtained when using a threshold that accepts many workers are of lower quality than if the threshold only accepts the best workers. However, in the static approach, a tighter filter keeps a smaller portion of the data (41% if only the results from the best 50% of the workers are accepted). In order to keep the quantity of data constant, the expense of resubmitting the task to MTurk is necessary. The dynamic approach with a threshold set at the median score costs almost the double of the version without quality control, and provides 33% relative improvement in term of WER. Thus, better quality and cheaper data could be obtained by finding the best threshold for a specific task.

Considering how well the acoustic models perform based on the data on which they are trained, there is a small but not significant difference between the approaches. The WER of the model built using the data from the two stage approach (62.3%) is slightly lower than its counterpart, built using the one stage approach (64.6%). The two dynamic approaches also seem to provide better training data, with a

WER around 62%. The inherent complexity of the dataset explains the low accuracy. The signal properties and quality vary from one call to another and many utterances are directed toward another person rather than toward the system. Also, since the system is being used at night with a wide variety of callers, there are many out-of-domain utterances that the language model does not expect. For example, a caller could yell: "I smoke crack rocks" which isn't likely to be correctly parsed by the grammar, and for which words are not in the dictionary.

However, both *static* rejection approaches perform considerably worse than the other approaches. This lends credence the idea that the quality of the acoustic models seems to be more heavily influenced by the quantity of data used in the training phase more than by its quality. This corroborates the conclusion of [7]. Consequently, if the ultimate goal is to obtain acoustic modeling, and a very large amount of data is available, it is better to obtain more data using a two stage approach **without quality control** since it is the cheapest. The two quality control mechanisms do not seem to provide a significant improvement in term of ASR WER. A more detailed analysis of the kind of errors each ASR made might provide more insight into the difference between the various approaches. Also, the more HITs with a GS that are completed by the workers, the better we can estimate their skills. The quality control mechanisms presented here would thus probably give better results when used on larger dataset.

The cost of the different options presented in Table 5 varies from $14.50/hour to $28.90/hour of speech. It is an order of magnitude cheaper than traditional transcription, but could probably be even lower using a seed model of better quality, or by providing a more efficient interface. One could also try to bring down the raw payment, paying 0.25 cents per utterance instead of 0.5 cents. However, it is important to keep in mind that even if the task is optimized and the worker can accomplish the transcription at 5 times real time, paying $5 per hour of speech comes down to paying a worker $1 per hour.

6. CONCLUSION

This paper has presented a crowdsourcing approach to speech transcription that improves the quality of the transcripts by dividing the task in subtasks, thus reducing cognitive load. The approach has better agreement with experts' transcriptions than an alternative approach where the entire annotation task is completed in one HIT. The results even approach the NIST measures for expert agreement. Even though we found that "good" workers submit more HITs than "bad" ones, quality control has been shown to yield improved results. A dynamic mechanism where low-confidence HITs are resubmitted to MTurk until a "good" worker completes them has been shown to provide the best quality transcripts, but is twice as expensive as not using quality control.

In the future, a third stage will be developed where workers will be asked to annotate audio recordings with noise markers. This will be done in parallel with an automatic pipeline that will send all the audio data collected in a day for Let's Go to MTurk for transcription. Active learning could then be achieved by retraining the ASR with the new data, thus improving accuracy and lowering the cost of MTurk by reducing the quantity of audio needed to go through the second stage.

7. ACKNOWLEDGMENTS

This work was funded by NSF grant IIS0914927. The opinions expressed in this paper do not necessarily reflect those of NSF. We would like to thank Adam Skory and Susi Burger for their help in transcription expertise. Many thanks to David Huggins-Daines for his support with the CMU Sphinx Speech Recognition Toolkit training and decoding.

8. REFERENCES

[1] Black, A., and Eskenazi, M. "The Spoken Dialogue Challenge" In *Proceedings of SIGDIAL 2009*.

[2] Yu, K., Gales, M., Wang, L. and Woodland, P. "Unsupervised training and directed manual transcription in LVCSR". *Speech Communication*. Volume 52, p. 652-663, 2010.

[3] Lamel, L., Gauvain, JL. And Adda, G. "Lightly supervised and unsupervised acoustic model training". *Computer Speech & Languag*e. Volume 16, p. 115-129. 2002.

[4] Gruenstein, A., McGraw, I. and Sutherland, A. "A self-transcribing speech corpus: collecting continuous speech with an online educational game", In *Proc. of the Speech and Language Technology in Education (SLaTE) Workshop*. 2009.

[5] Marge, M., Banerjee, S. and Rudnicky, A. "Using the Amazon Mechanical Turk for transcription of spoken language". *IEEE-ICASSP*. 2010.

[6] Evanini, K., Higgins, D. and Zechner, K. "Using Amazon Mechanical Turk for transcription of non-native speech". In *Proceedings of the NAACL workshop on creating speech and language data with Amazon's Mechanical Turk*. 2010.

[7] Novotney, S. and Callison-Burch, C. "Cheap, fast and good enough: automatic speech recognition with non-expert transcription". In *Proceedings of NAACL*. 2010.

[8] Fiscus, J. "A post-processing system to yield reduced error rates: recognizer output voting error reduction (ROVER)". In *Proceedings of IEEE ASRU workshop*. 1997.

[9] Sheng, VS., Provost, F. and Ipeirotis, PG. "Get another label? Improving data quality and data mining using multiple, noisy labelers." In *Proceedings of 14th ACM SIGKDD*. 2008

[10] get-another-label. http://code.google.com/p/get-another-label/

[11] Schneiderman, B. "The limits of speech recognition". *Communications of the ACM.* Volume 43. 2000.

[12] Snow, R., O'Connor, B., Jurafsky, D. and Ng, AY. "Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks". In *Proc. of EMNLP-08*. 2008.

[13] NIST, RTE Project. http://www.itl.nist.gov/iad/mig/tests/rt/

[14] Placeway, P., Chen, S., Eskenazi, M., Jain, U., Parikh, V., Raj, B., Ravishankar, M., Rosenfeld, R., Seymore, K., Siegler, M., Stern R. and Thayer, E. "The 1996 Hub-4 Sphinx-3 System". In *DARPA Speech Recognition Workshop*. 1996.