# Crowd-sourcing for difficult transcription of speech

Jason D. Williams, I. Dan Melamed, Tirso Alonso, Barbara Hollister, and Jay Wilpon

*Shannon Laboratory, AT&T Labs - Research, Florham Park, NJ, USA*

jdw@research.att.com

***Abstract*—Crowd-sourcing is a promising method for fast and cheap transcription of large volumes of speech data. However, this method cannot achieve the accuracy of expert transcribers on speech that is difficult to transcribe. Faced with such speech data, we developed three new methods of crowd-sourcing, which allow explicit trade-offs among precision, recall, and cost. The methods are: incremental redundancy, treating ASR as a transcriber, and using a regression model to predict transcription reliability. Even though the accuracy of individual crowd-workers is only 55% on our data, our best method achieves 90% accuracy on 93% of the utterances, using only 1.3 crowd-worker transcriptions per utterance on average. When forced to transcribe all utterances, our best method matches the accuracy of previous crowd-sourcing methods using only one third as many transcriptions. We also study the effects of various task design factors on transcription latency and accuracy, some of which have not been reported before.**

## I. Introduction and background

Modern speech recognition engines are typically trained on hundreds or thousands of hours of manually transcribed speech. Historically, transcription has been an expensive and slow process done by expert transcribers. Such experts typically require at least 6 hours of work per hour of speech. In the U.S.A., the resulting cost is $90–$150 per hour of speech [1], [2], [3], [4]. The slow pace and high cost of transcription are major obstacles to improving speech recognition technology. In addition, it is often difficult to find enough expert transcribers for large transcription projects, especially when the volume of work fluctuates.

Recently, *crowd-sourcing* has emerged as a promising method for inexpensive, fast, large-scale, on-demand transcription of speech data [2], [5], [6], [7], [8], [9]. In crowd-sourcing, a large task is divided into many small tasks. These small tasks are then distributed to a large pool of workers through a co-ordinating web service. The workers work concurrently, greatly speeding up the completion of the original large task. The supply/demand economics of such work allow each small task to be done for as little as $0.01, which can reduce the overall cost by several orders of magnitude.

The key challenge in transcription by crowd-sourcing is quality control. Crowd-workers typically produce lower quality transcriptions than experts. Some try to get paid without even doing the work, e.g. by submitting garbage [10]. Previous research – which studied composed monologues [9], conversations [2], meetings [8], and bus timetable utterances [5] – found that crowd-workers differed from experts in 5% [9] to 23% [2] of the transcribed words.

One approach to quality control is to have several workers transcribe each utterance, and then to choose the most frequent transcription [2], [5] or to combine the transcriptions into one composite transcription [9]. Another approach is to use automatic speech recognition (*ASR*). For example, crowd-workers can edit ASR output [7], or they can decide whether ASR output is accurate [5]. In previous work, such methods achieved almost expert quality transcription for as little as $5 per hour of speech [2], while enjoying the high availability and speed of crowd-sourcing.

In this paper, we address the challenges posed by a more difficult transcription task — business name queries from a publicly accessible telephone directory service. Such utterances are difficult to transcribe because there is no conversational context, the vocabulary is very large, and location-specific knowledge is often required. We found that crowd-workers disagreed with experts by 37% or more — a relative increase of 60% over the worst word error rate (WER) reported in previous studies [2]. Using the majority vote of 7 workers reduced the WER to 17%, which is still higher than the WER of *individual* crowd-workers on some other kinds of data, and also 7 times more expensive. Lower transcription quality might be acceptable for some applications. For example, studies have shown that useful acoustic models and language models can be built from transcriptions with a WER as high as 23% [2]. For other uses of transcriptions, such as measuring the accuracy of a deployed ASR system, expert-quality is necessary.

We propose three new techniques for improving transcription by crowd-sourcing. Our techniques can produce a transcription for every utterance, but they can also estimate the relative reliability of different transcriptions. Thus, they offer system builders a way to trade off between transcription precision, transcription recall, and cost. Depending on the use case, the remaining utterances can be re-transcribed by more expensive experts, used with special handling, or discarded. The first technique requests transcriptions one at a time until a desired number of matching transcriptions is obtained, saving the cost of many unnecessary transcriptions. The second technique uses ASR output in a manner than is simple to implement and less susceptible to cheating than previous methods. The third technique uses a regression model to estimate the reliability of the most frequent transcription for each utterance, so that additional transcriptions can be requested only when this reliability is below a threshold. For a given level of precision and recall, each technique yields an

incremental reduction in the average number of transcriptions required, and thereby also the average cost per utterance.

Section II of this paper describes the data we used and explains the transcription collection procedure. Section III discusses factors that affected the latency and accuracy of individual transcriptions, including several factors that have not been studied before. Sections IV-VI then present a baseline and our three new methods for reducing transcription cost.

## II. Data and experimental design

Our speech corpus consisted of 900 telephone-quality audio snippets from a deployed directory assistance service. Each snippet was a response to a prompt in English asking for a business name. Professional transcribers (the *experts*) transcribed the corpus. 12.8% of their transcriptions were empty because the snippets contained only background noise or background speech. Snippets that contained speech directed at the system had a mean of 2.63 words.

We used Amazon's Mechanical Turk (*MTurk*) at `mturk.com` as a crowd-sourcing platform. MTurk tasks are called "Human Intelligence Tasks" (*HITs*).

The independent variables in our experiments were:

- price per utterance: \$0.002, \$0.004, or \$0.01;
- HIT size: 5, 10, or 15 utterances per HIT; and
- time of day when HITs were submitted to MTurk: 12AM or 12PM GMT.

The effects of price have been studied before [7], [9], but we are not aware of any studies of the effects of task size and request time on the accuracy and latency of crowd-sourced transcriptions. The 3 independent variables yielded 18 experimental conditions. For each condition, 5 HITs were created, and 7 copies of each HIT were submitted to MTurk, for a total of 630 HITs. Workers were barred from transcribing a given utterance more than once. Our HITs were open to any worker: no skill qualifications were required.

To encourage workers to participate, the worker GUI was designed to be simple. It consisted of brief instructions such as "Type exactly what the main speaker says", a few examples of correct and incorrect transcriptions, links to audio files, a text box for transcribing each audio file, and an extra text box for comments. The instructions were much shorter and simpler than the standard guidelines used by our experts. For example, the experts' guidelines say to mark background speech with a special tag instead of transcribing it.

## III. Baseline results

### A. Worker response

126 different workers worked on the 630 HITs. The most productive worker did 48 HITs, or 7.6% of all HITs. 53 crowd-workers did only one HIT. Much of the crowd's work was done by a small number of loyal workers. This finding can inform strategies of scaling up crowd-sourcing to a large volume of transcriptions.

Table I shows the types of comments that the workers wrote in the comment box.

TABLE I
Crowd-worker comments, grouped by type.

| Count | Comment type |
|---|---|
| 61 | Indication that an utterance was hard to understand |
| 19 | Positive comment about the task/interface |
| 6 | Questions about how to handle profanity |
| 4 | Questions about the operation of the GUI |
| 3 | Questions about how to handle other languages |
| 1 | Questions about how to handle capitalization |
| 1 | Questions about how to handle multiple speakers |
| 1 | Questions about how to handle non-speech noises |
| 1 | Indicating a previous HIT had an error |
| 1 | Questions about payment |

### B. Latency

Figure 1 shows that higher prices yielded lower latencies for the first 90% of HITs, consistent with prior studies [7], [9]. The variances of the latencies were much higher among the last 10% of HITs, so we cannot compare their means with confidence. We conjecture that the higher variances in the last 10% were due to the fact that workers can sort HITs by their time of arrival. Workers are less likely to see older HITs, so both the mean and the variance of their completion time rises, sometimes dramatically. Overall, the average latency per transcription at \$0.01 was significantly faster than at cheaper prices ($p < 0.0001$, pair-wise Mann-Whitney). We were unable to show that other differences were significant.

Figure 2 reports a new result: the effect of HIT size on latency. Again, we found the differences to be reliable only for the first 90% of HITs completed, but the overall trend is clear: smaller HITs reliably yield lower latencies. We conjecture that this trend is due to the way that MTurk pays workers. Workers can be paid only for whole HITs, not for fractions of HITs, and HIT requesters can choose to pay workers or not. Therefore, from a worker's point of view, larger HITs carry a risk of not getting paid for a larger amount of work. Fewer workers are willing to take larger risks, so it takes longer for a larger

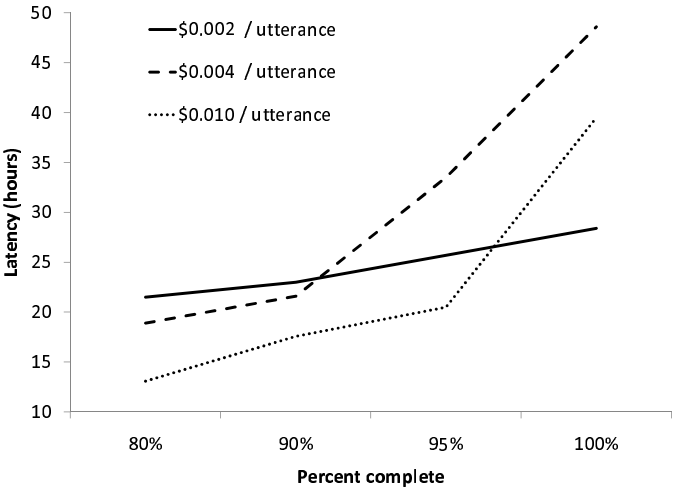



Fig. 1. Mean latency for different prices per utterance. Higher prices yield lower latencies for the first 90% of HITs.

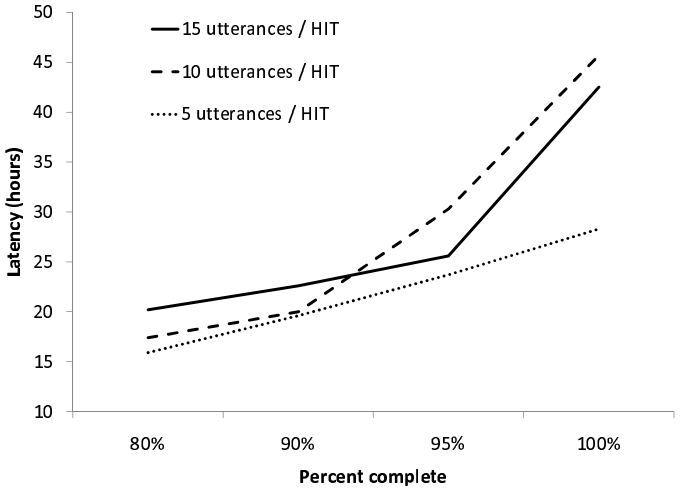



Fig. 2. Mean latency for different HIT sizes, i.e. number of utterances per HIT. Smaller HITs yield lower latencies for the first 90% of HITs.

HIT to be accepted by a worker. Overall, the average latency per transcription with 5 utterances/HIT was significantly faster than larger-sized HITs ($p \leq 0.015$, pair-wise Mann-Whitney). We were unable to show that other differences were significant.

We also examined the combined effect of price per utterance and HIT size. Each variable was largely independent of the other, with one glaring outlier. HITs priced at \$0.01 (i.e. 5 utterances per HIT at \$0.002 per utterance) had much higher latency — 25.5 hours for 80% completion. This outlier might be attributed to the vast majority of HITs on MTurk being priced at \$0.01. Workers can sort the available HITs by price, so HITs priced above \$0.01 are appealing to more workers.

We found no correlation between request time and latency. Crowd-workers were available around the clock.

### C. Accuracy

We measured the accuracy of the crowd in terms of exact match with the experts' reference transcriptions, ignoring differences in whitespace and capitalization. Since our experts did not transcribe the 12.8% of utterances with no foreground speech, we excluded these utterances from this analysis.

Refer to Table II. The mean whole-utterance error rate (*UER*) was 44.6% — nearly half of the transcriptions differed from their references. Past studies reported WERs between 5% and 23%. In contrast, the WER on our data was 36.8%. The WER drops by only 20.2% to 16.6% for the most frequent of 7 transcriptions, still 3 times higher than individual transcriptions in some prior studies.

TABLE II
MEAN ACCURACY OF CROWD-SOURCED TRANSCRIPTIONS

| | UER | WER |
|---|---|---|
| all transcriptions | 44.6% | 36.8% |
| most frequent transcriptions | 21.9% | 16.6% |
| ROVER over all transcriptions | 22.8% | 15.6% |
| ASR (1-best) | 34.1% | 27.7% |

ROVER [11] is an algorithm for combining multiple transcriptions by word-level voting. ROVER has been shown to lower the WER of the crowd, in some cases substantially [9], [8], in others only marginally [2]. Here ROVER increases utterance error rate slightly, and decreases WER slightly. With business names, we are primarily interested in UER, so we did not pursue ROVER further.

We next compared crowd-workers to ASR. We built a statistical language model from a large separate set of expert transcriptions, and combined it with a generic acoustic model using AT&T's WATSON ASR system [12]. On this task, WATSON was more accurate than individual crowd-workers, but less accurate than 7 crowd workers.

Prior work [9] reported the counter-intuitive trend that higher prices lead to *more* errors. Table III suggests the same trend in our data. The error rate at the cheapest price was significantly lower than at higher prices ($p \leq 0.014$, pair-wise Mann-Whitney applied to utterances with ties broken randomly).

TABLE III
PRICE PER UTTERANCE VS. UTTERANCE ERROR RATE

| Price per utterance | % UER |
|---|---|
| \$0.01 | 50.8% |
| \$0.004 | 47.0% |
| \$0.002 | 42.7% |

Utterances requested at midnight GMT had a UER of 50.9%, but those requested at noon had a significantly lower UER of 42.8% ($p \leq 0.0006$, Mann-Whitney Test applied to utterances with ties broken randomly). We conjecture that the noon HITs were more often done by workers in North America, who were more likely to be familiar with the business names in the data, whereas the midnight HITs were more often done by workers elsewhere. We can't be sure, because MTurk does not reveal worker location. Another possibility is that more workers did the midnight hits at night, when they were more tired. We found no correlation between HIT size and accuracy.

The rest of the paper describes better methods for crowdsourcing. Since one of our methods employs machine learning, we randomly divided the data into training and test sets of 450 utterances each. We included the utterances that the experts labeled as background speech, since such labels would not be available to real-world applications of crowd-sourcing. For this 12.8% of utterances, we used the most frequent crowd-worker transcription as the reference. All results from here on were measured on the test set.

We computed two different measures of accuracy. *Lexical accuracy* was measured in terms of exact match with the reference, ignoring variations in whitespace and case. *Phonetic accuracy* was measured the same way, but also ignoring the following kinds of variation:

- dropped dashes: "uhaul" vs. "u-haul"
- dropped apostrophes: "johns" vs. "john's"
- homophonic differences: "john" vs. "jon"

- dropped fillers: “vermont vs. “vermont uh”
- dropped partial words: “hotel” vs. “ho- hotel”
- differences attributable to genuinely unclear audio, as determined by another expert: “united airline” vs. “united airlines”

The most useful measure depends on how the transcriptions are used. Lexical accuracy is more appropriate for measuring the accuracy of a deployed system. Phonetic accuracy is more relevant for training acoustic models, where phonetically identical transcriptions will yield the same model and a garbage model usually absorbs word fragments.

To improve the accuracy of crowd-sourced transcription, previous studies collected multiple transcriptions for each utterance, and then used the most frequent one. We applied this approach to our data, using from 1 to 7 transcriptions for each utterance, and breaking ties in favor of transcriptions completed earlier. The resulting lexical and phonetic accuracies are shown as the dashed line with squares in Figure 3. These baseline curves are identical in the upper and lower panels, since the baseline method produced a transcription for every utterance, so its recall was always 100%. Lexical accuracy on the test set rose from 53.1% for a single worker to 83.6% for 7. Phonetic accuracy rose from 73.6% to 97.1%. The differences between the two measures highlight the difficulty of transcribing business names correctly. Expert transcribers diligently look up correct spellings, but crowd-workers often chose a similar-sounding variant. With either measure, the accuracy of the crowd is substantially below 100% even with 7 transcriptions per utterance. That is why we pursued methods that allow trade-offs between cost, precision, and recall.

## IV. Incremental redundancy

In the baseline method above, a fixed number of transcriptions was requested for every utterance. In practice, it is wasteful to request as many transcriptions for easy utterances (on which most crowd-workers agree) as for difficult ones. In addition, the baseline method cannot estimate which utterances are likely to be transcribed correctly and which are not. The method of *incremental redundancy* addresses both of these concerns. The idea is to stop requesting transcriptions after $N$ transcriptions are obtained, as before, but also to stop if $K$ matching transcriptions are obtained, for a fixed value of $K < N$. If a certain transcription is repeated $K$ times, then that transcription is deemed *reliable*. If $N$ transcriptions are obtained that do not contain a matching set of size $K$, the majority transcription is deemed *unreliable*. To our knowledge, this approach has not been studied before.

Figure 3 uses the dotted line with circles to show precision and recall for various values of $K$. The circles do not line up with the x-axis tic marks because the mean number of utterances requested for a given $K$ was always a non-integer greater than $K$. The upper plots are based on reliable utterances only; the lower plots use all utterances. At $K = 2$, the method of incremental redundancy deemed 93% of utterances reliable, achieving the same accuracy as the baseline using only two thirds as many transcriptions. At higher values of $K$, this method still deemed more than half the utterances reliable. The reliable subset was transcribed more accurately than the baseline, and at a slightly lower cost. When all transcriptions are considered, the method achieves very similar accuracy to the baseline but at half the cost.

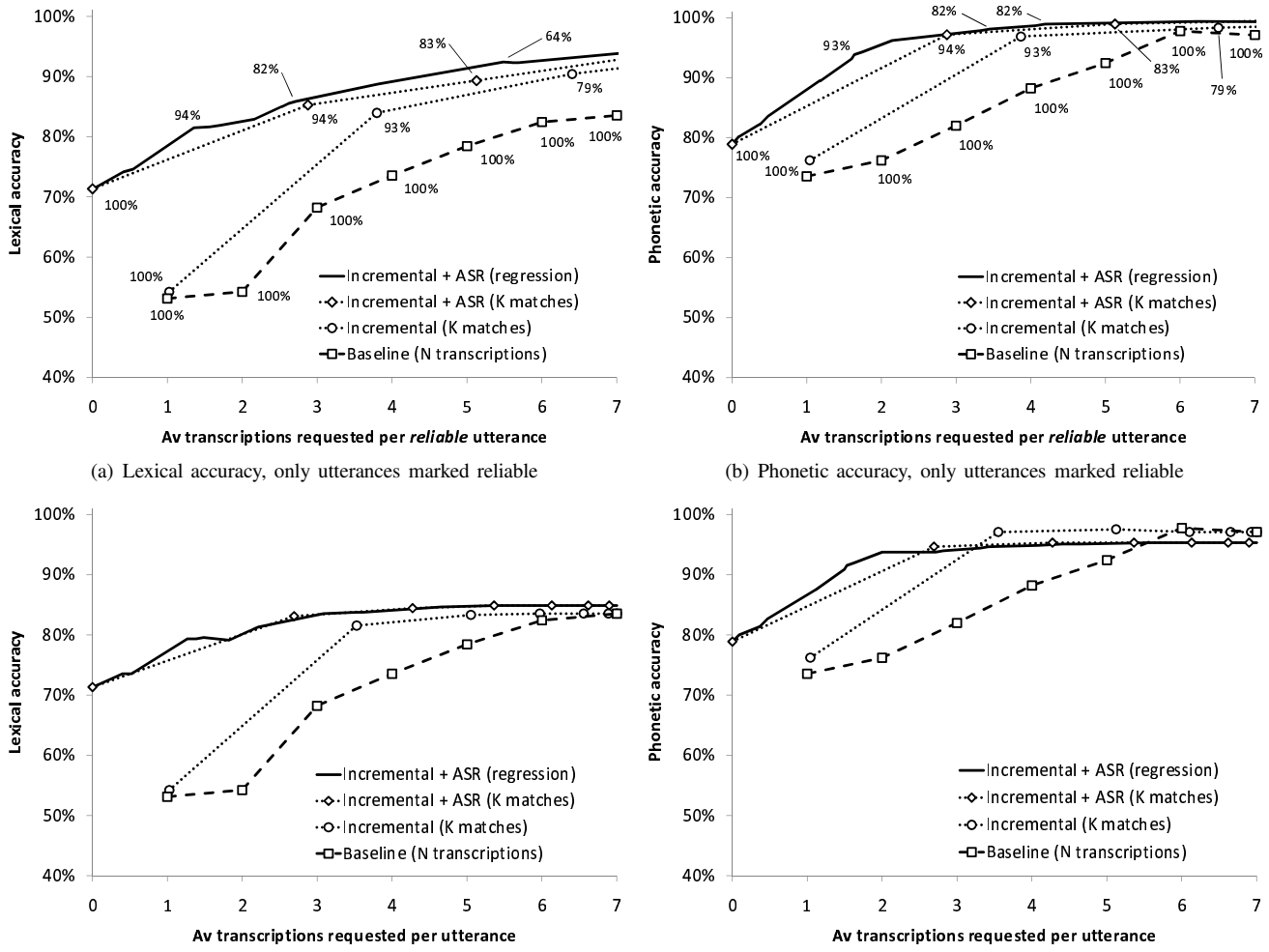
## V. Using ASR with crowd-sourcing

In large-scale transcription settings, there is often enough previously transcribed data available to build an ASR system. ASR is another source of transcriptions. ASR output often contains errors, but the marginal cost to gather each ASR transcription is negligible.

Prior work has suggested two methods for using ASR in transcription by crowd-sourcing. Both of these approaches are susceptible to cheating, which degrades transcription quality. One method is to ask crowd-workers to edit ASR output [7]. Since workers aim to minimize the time they spend on each HIT, in this approach there is a bias towards making fewer changes than necessary to the ASR output. The researchers who proposed this method found that 24% of workers edited no words in more than 10% of their HITs. The researchers decided to discard *all* of the HITs from those crowd-workers, to reduce the risk that erroneous ASR transcriptions will be deemed reliable. However, some of the ASR transcriptions were probably correct, so some useful information was discarded. The other method uses a two-stage approach: first, crowd-workers decide whether ASR output is accurate; then (later) crowd-workers are asked to transcribe only the utterances which were deemed inaccurate [5]. This method also runs the risk that some erroneous ASR transcriptions will be deemed reliable, because workers have an incentive to classify utterances without examining them.

We propose a method that is not susceptible to cheating, because it does not show any ASR output to the crowd. Instead, we treat ASR as just another worker, who works for free, is always available, and completes work almost instantly. Not surprisingly, we always treat ASR output as the *first* worker. As a bonus, this worker’s results come with a confidence score.

The effects of using a single ASR hypothesis in this manner are shown as the dotted line with diamonds in Figure 3. In general, using ASR substantially reduces the number of manual transcriptions required to achieve a given level of accuracy. In particular, many easy utterances can be deemed reliably transcribed after just one transcription from the crowd. Thus, phonetic accuracy similar to the baseline can be achieved at less than half the cost. We also tried using 100 best hypotheses from ASR, but did not see a substantial difference. Workers tended to make different kinds of errors than ASR, such as mis-spellings or invented phonetic transcriptions, which didn’t appear in the 100-best list. There were a few additional matches, but they were less likely to be correct, negating their benefit.




(a) Lexical accuracy, only utterances marked reliable

(b) Phonetic accuracy, only utterances marked reliable

(c) Lexical accuracy, all utterances

(d) Phonetic accuracy, all utterances

Fig. 3. Mean transcription accuracy (y axis) as a function of the mean number of transcriptions requested (x axis). The lower two plots are based on always accepting the most frequent transcription. Thus the recall in the lower plots is always 100%. In the upper two plots, only transcriptions deemed *reliable* are considered, so the x axis shows the number of transcriptions requested per *reliable* transcription generated. The percentages inside the upper plots show recall, i.e. the fraction of transcribed utterances which our methods deem reliable. In all plots, the squares for the baseline method show $N = 1 \ldots 7$, where $N$ is the number of transcriptions requested. The circles and diamonds for the incremental methods represent different values of $K$, the number of matches required.

## VI. Predicting reliability with a regression

When we incorporated ASR into the transcription process, we were curious whether the ASR confidence score could be useful: if the confidence was very high, perhaps we could avoid requesting any transcriptions from the crowd. More generally, there were several unused features of the crowd-sourced transcriptions which could help to predict their accuracy. For example, the audio player logged the number of times the play button was pressed. Figure 4 shows that this feature was a strong predictor of accuracy. Some lexical features also seemed relevant: transcribing plurals was often problematic whereas labeling silence was comparatively easy.

We performed regression on various features of the transcription process, to predict the likelihood of correctness of the majority transcription at any given point in the process. If the probability of correctness was below a threshold, then another transcription was requested. In other words, we were still using incremental redundancy, but the decision about when to stop used more information than just the constant $K$. In prior work, transcription accuracy has been predicted using purely acoustic features [13]. Our contribution is to use a wider variety of features, to combine them in a regression model, and to show end-to-end utility in a transcription process.

We built two regression models — one each for lexical and phonetic accuracy. As in the previous section, the first transcription came from ASR and subsequent transcriptions came from MTurk. The regression models used these features:

- ASR confidence score;
- number of transcriptions requested so far;
- size of majority/plurality set;
- number of times the play button was pressed;
- number of words in the majority transcription;
- whether majority transcription is marked as silence; and
- whether majority transcription includes a plural.

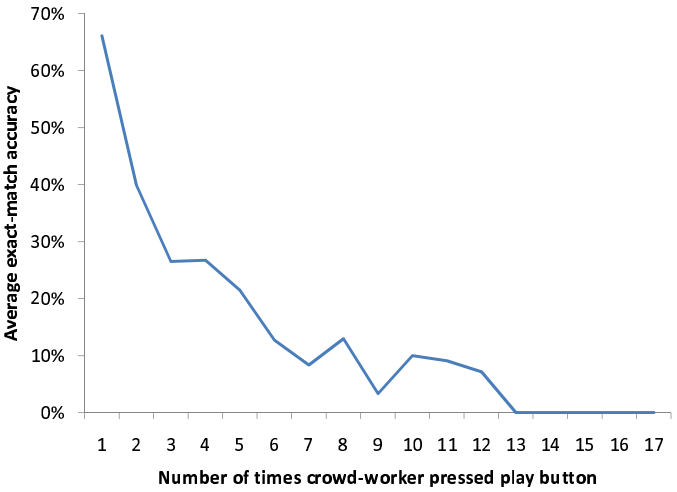



Fig. 4. Number of times worker pressed play button vs. lexical accuracy.

The models were trained on the training set using an alternating decision tree (LADTree) as implemented in the WEKA toolkit [14], [15]. Results are shown as the solid line in Figure 3. The line was created by using 1000 evenly spaced probability thresholds between 0 and 1.

For phonetic accuracy of reliable transcriptions only, precision and recall are similar to the previous two methods, but the average number of transcriptions required is reduced by approximately 1. For lexical accuracy, there is an improvement in all dimensions: precision and recall improve *and* the required number of transcriptions declines. The average accuracy of a single crowd-worker was only 55%, but our regression method transcribed 93% of the utterances with 90% phonetic accuracy, using only 1.3 MTurk transcriptions per utterance. The method achieved 96% phonetic accuracy on the same 93% of utterances, using only 2.0 MTurk transcriptions per utterance. At 100% recall (the lower panels in the figure), there was only a slight improvement over using ASR without regression, for both accuracy measures.

## VII. Conclusions and Future Work

This paper studied crowd-sourcing for difficult transcription of speech. We first described how various factors affected the accuracy and latency of crowd-sourced transcriptions. Larger tasks increased latency, suggesting it is better to divide work up into smaller chunks. However, the chunks shouldn't be too small: tasks priced too low took a long time to complete. Inexplicably, transcription accuracy varied by time of day.

We then proposed three methods for using MTurk for difficult transcriptions more effectively. Since crowd-workers showed relatively low accuracy on this task, and since different uses of the data have different requirements for transcription accuracy, we developed methods that explicitly expose the trade-offs between precision, recall, and cost. We first proposed gathering transcriptions one at a time until $K$ matches are obtained. We then proposed treating ASR output as the first crowd-worker. Finally, we proposed a way to use regression to estimate the probability of correctness of crowd-sourced transcriptions. Using this probability to decide whether to request more transcriptions maintained or improved precision and recall while further lowering the number of transcriptions required. When forced to produce a transcription for every utterance, these methods yield the same accuracy as baseline methods using less than half the transcriptions, and therefore half the expense. When configured to maximize precision, these methods yield transcriptions that are more accurate than the baseline for a known 80-90% of utterances, again at about half the cost of the baseline.

In future work, we plan to extend the regression method to use the reliability of individual crowd-workers, based on their agreement rates on previous HITs. Prior work has shown that crowd-worker agreement correlates highly with accuracy [2]. We also plan to re-evaluate all of our methods in terms of the accuracy and cost of ASR systems built from the transcriptions that our methods produce. Lastly, we hope to find ways to use utterances that are deemed unreliable.

## References

[1] C. Passy, "Turning audio into words on the screen," *Wall Street Journal*, 2008. [Online]. Available: http://online.wsj.com/article/SB122351860225518093.html

[2] S. Novotney and C. Callison-Burch, "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription," in *Proc NAACL HLT, Los Angeles, California*, 2010.

[3] O. Kimball, C.-L. Kao, T. Arvizo, J. Makhoul, and R. Iyer, "Quick transcription and automatic segmentation of the Fisher conversational telephone speech corpus," in *RT04 Workshop*, 2004.

[4] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: a resource for the next generations of speech-to-text," in *LREC*, 2004.

[5] G. Parent and M. Eskenazi, "Toward better crowdsourced transcription: Transcription of a year of the Let's Go bus information system data," in *Proc SLT, Berkeley, CA*, 2010.

[6] M. Wald, "Crowdsourcing correction of speech recognition captioning errors," in *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*. ACM, 2011.

[7] I. McGraw, C. Y. Lee, L. Hetherington, and J. Glass, "Collecting voices from the cloud," in *Proc International Conference on Language Resources and Evaluation (LREC), Malta*, 2010.

[8] M. Marge, S. Banerjee, and A. I. Rudnicky, "Using the Amazon Mechanical Turk to transcribe and annotate meeting speech for extractive summarization," in *Proc NAACL HLT Workshop on Creating Speech and Language Data with Amazons Mechanical Turk, Los Angeles, California*, 2010.

[9] M. Marge, S. Banerjee, and A. Rudnicky, "Using the Amazon Mechanical Turk for transcription of spoken language," in *Proc Intl Conf on Acoustics, Speech, and Signal Processing (ICASSP), Dallas, USA*, 2010.

[10] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on Amazon Mechanical Turk," in *Proc ACM SIGKDD Workshop on Human Computation (HCOMP)*, 2010.

[11] J. G. Fiscus, "A post-processing system to yield redcued word error rates: Recognizer output voting error reduction (ROVER)," in *Proc IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Santa Barbara, USA*, 1997.

[12] V. Goffin, C. Allauzen, E. Bocchieri, D. Hakkani-Tur, A. Ljolje, S. Parthasarathy, M. Rahim, G. Riccardi, and M. Saraclar, "The AT&T Watson speech recognizer," in *Proc ICASSP, Philadelphia*, 2005.

[13] B. C. Roy, S. Vosoughi, and D. Roy, "Automatic estimation of transcription accuracy and difficulty," in *Proc INTERSPEECH, Makuhari, Japan*, 2010.

[14] G. Holmes, B. Pfahringer, R. Kirkby, E. Frank, and M. Hall, "Multiclass alternating decision trees," in *Machine Learning: ECML 2002*, 2002.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.