Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-Expert Transcription

Scott Novotney and **Chris Callison-Burch**
Center for Language and Speech Processing
Johns Hopkins University
snovotne@bbn.com ccb@jhu.edu

Abstract

Deploying an automatic speech recognition system with reasonable performance requires expensive and time-consuming in-domain transcription. Previous work demonstrated that non-professional annotation through Amazon's Mechanical Turk can match professional quality. We use Mechanical Turk to transcribe conversational speech for as little as one thirtieth the cost of professional transcription. The higher disagreement of non-professional transcribers does not have a significant effect on system performance. While previous work demonstrated that redundant transcription can improve data quality, we found that resources are better spent collecting more data. Finally, we describe a quality control method without needing professional transcription.

1 Introduction

Successful speech recognition depends on huge investments in data collection. Even after training on 2000+ hours of transcribed conversational speech, over a billion words of language modeling text, and hand-crafted pronunciation dictionaries, state of the art systems still have an error rate of around 15% for English (Prasad et al., 2005) Transcribing the large volumes of data required for Large Vocabulary Continuous Speech Recognition (LVCSR) of new languages appears prohibitively expensive. Recent work has shown that Amazon's Mechanical Turk[1] can be used to cheaply create data for other natural language processing applications (Snow et al., 2008; Zaidan and Callison-Burch, 2009; McGraw et al., 2009). In this paper we focus on reducing the cost of transcribing conversational telephone speech (CTS) data. Previous measurements of Mechanical Turk stopped at agreement/disagreement with professional annotation. We take the next logical step and measure performance on systems trained with non-professional transcription.

Mechanical Turk is an online labor market where workers (or Turkers) perform simple tasks called Human Intelligence Tasks (HITs) for small amounts of money – frequently as little as $0.01 per HIT. Since HITs can be tasks that are difficult for computers, but easy for humans, they are ideal for natural language processing tasks (Snow et al., 2008). Mechanical Turk has even spawned a business that specializes in manual speech transcription.[2]

Automatic speech recognition (ASR) of conversational speech is an extremely difficult problem. Characteristics like rapid speech, phonetic reductions and speaking style limit the value of non-CTS data, necessitating in-domain transcription. Even a few hours of transcription is sufficient to bootstrap with unsupervised methods like self-training (Lamel et al., 2002). The speech community has built effective downstream solutions for the past twenty years despite imperfect recognition. In topic classification, 90% accuracy is possible on conversational data even with 80%+ word error rate

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

[2] http://castingwords.com/

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 207–215,
Los Angeles, California, June 2010.

(WER) (Gillick et al., 1993). Other successful tasks include information retrieval from speech (Miller et al., 2007) and spoken dialogue processing (Young et al., 2007). Inexpensive transcription would quickly open new languages or domains (like meeting or lecture data) for automatic speech recognition.

In this paper, we make the following points:

- Quality control isn't necessary as a system built with non-professional transcription is only 6% worse for $\frac{1}{30}$ the cost of professional transcription.
- Resources are better spent collecting more data than improving data quality.
- Transcriber skill can be accurately estimated without gold standard data.

2 Related Work

Research into Mechanical Turk by the NLP community has largely focused on comparing the quality of annotations produced by non-expert Turkers against annotations created by experts. Snow et al. (2008) conducted a comprehensive study across a variety of NLP tasks. They showed that high agreement could be reached with gold-standard expert annotation for these tasks through a weighted combination of ten redundant annotations produced by Turkers.

Callison-Burch (2009) showed similar results for machine translation evaluation, and further showed that Turkers could accomplish complex tasks like translating Urdu or creating reading comprehension tests.

McGraw et al. (2009) used Mechanical Turk to improve an English isolated word speech recognizer by having Turkers listen to a word and select from a list of probable words at a cost of $20 per hour of transcription.

Marge et al. (2010) collected transcriptions of verbal instructions to robots with clean speech. By using five duplicate transcriptions, the average transcription disagreement with experts was reduced from 4% to 2%.

Previous efforts at reducing the cost of transcription include the EARS Fisher project (Cieri et al., 2004), which collected 2000+ hours of English CTS data – an order of magnitude more than had previously been transcribed. To speed transcription and lower costs, Kimball et al. (2004) created new transcription guidelines and used automatic segmentation. These improved the speed of transcription from fifty times real time to six times real time, and made it cost effective to transcribe 2000 hours at an average of $150 per hour. Models trained on the faster transcripts exhibited almost no degradation in performance, although discrimanitve training was sensitive to transcription errrors.

3 Experiment Description

3.1 Corpora

We conducted most experiments on a twenty hour subset of the English Switchboard corpus (Godfrey et al., 1992) where two strangers converse about an assigned topic. We used two sets of transcription as our gold standard: high quality transcription from the LDC and those following the Fisher quick transcription guidelines (Kimball et al., 2004) provided by a professional transcription company. All English ASR models were tested with the carefully transcribed three hour Dev04 test set from the NIST HUB5 evaluation.[3] A 75k word lexicon taken from the EARS Fisher training corpus covers the LDC training data and has a test OOV rate of 0.18%.

We also conducted experiments in Korean and collected Hindi and Tamil data from the Callfriend corpora [4]. Participants were given a free long distance phone call to talk with friends or family in their native language, although English frequently appears. Since Callfriend was originally intended for language identification, only the 27 hour Korean portion has been transcribed by the LDC.

3.2 LVCSR System

We used Byblos, a state-of-the-art multi-pass LVCSR system with state-clustered Gaussian tied-mixture acoustic models and modified Kneser-Ney smoothed language models (Prasad et al., 2005). While understanding the system

[3] http://www.itl.nist.gov/iad/mig/tests/ctr/1998/current-plan.html

[4] http://www.ldc.upenn.edu/CallFriend2/

details is not essential for this work, we provide a brief description for completeness.

Recognition begins with cepstral feature extraction using concatenated frames with cepstral mean subtraction and HLDA to reduce the feature dimension space. Vocal track length normalization follows. Decoding then requires three passes: a fast forward pass with coarse one-gaussian-per-phone models and bigram LM followed by a backward pass with triphone models and a trigram LM to generate word confusion lattices. The lattices are rescored with a more powerful quinphone cross-word acoustic model and trigram LM to extract the one best output. These three steps are repeated after unsupervised speaker adaptation with constrained MLLR. Decoding is around ten times real time.

3.3 Transcription Task

Using language-independent speaker activity detection models, we segmented each ten minute conversation into five second utterances, greatly simplifying the transcription task (Roy and Roy, 2009). Utterances were assigned in batches of ten per HIT and played with a simple flash player with a text box for entry. All non-empty HITs were approved and we did not award bonuses except as described in Section 5.1.

3.4 Measuring Annotation Quality

The usefullness of the transcribed data is ultimately measured by how much it benefits a speech recognition system. Factors that inflate disagreement (word error rate) between Turkers and professionals do not necessarily impact system performance. These include typographical mistakes, transcription inconsistencies (like improperly marking hesitations or the many variations of `um`) and spelling variations (`geez` or `jeez` are both valid spellings). Additionally, the gold standard is itself imperfect, with typical estimates of professional disagreement around five percent. Therefore, we judge the quality of Mechanical Turk data by comparing the performance of one LVCSR system trained on Turker annotation and another trained on professional transcriptions of the same dataset.

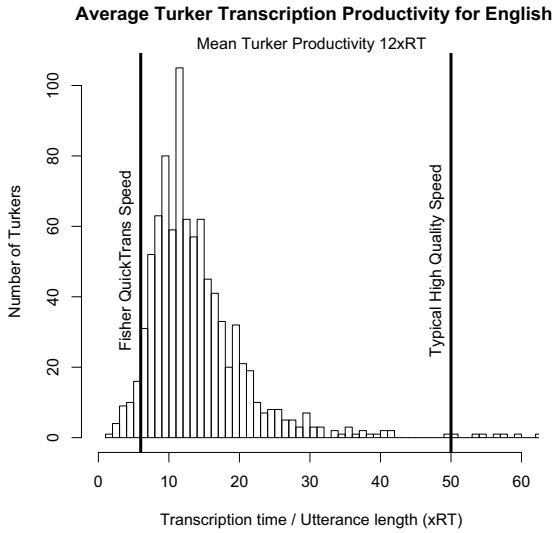

Figure 1: Histogram of per-turker transcription rate for twenty hours of English CTS data. Historical estimates for high quality transcription are 50xRT. The 2004 Fisher transcription effort achieved 6xRT and the average here is 11xRT.

4 Establishing Best Practices with English Switchboard

As an initial test to see how cheaply conversational data could be transcribed, we uploaded one hour of test data from Hub5 Dev04. We first paid $0.20 per HIT ($0.02 per utterance). This test finished quickly, and we measured the average disagreement with professionals at 17%. Next, we reduced payment to $0.10 per HIT and disagreement was again 17%. Finally, we pushed the price down to $0.05 per HIT or $5 per hour of transcription and again disagreement was nearly identical at 18%, although a few Turkers complained about the low pay.

Using this price, we then paid for the full twenty hours to be redundantly transcribed three times. 1089 Turkers participated in the task at an incoming rate of 10 hours of transcription per day. On average, each Turker transcribed 30 utterance (earning 15 cents) at an average professional disagreement of 23%. Transcribing one minute of audio required an average eleven minutes of effort (denoted 11xRT). 63 workers transcribed more than one hundred utterances and one prolific worker transcribed 1223 utterances.

4.1 Comparing Non-Professional to Professional Transcription

Table 1 details the results of different selection methods for redundant transcription. For each method of selection, we build an acoustic and language model and report WER on the heldout test set (transcribed at very high accuracy).

We first randomly selected one of the three transcriptions per utterance (as if the data were only tanscribed once) and repeated this three times with little variance. Selecting utterances randomly by *Turker* performed similarly. Performance of an LVCSR system trained on the non-professional transcription degrades by only 2.5% absolute (6% relative) despite a disagreement of 23%. This is without any quality control besides throwing out empty utterances. The degradation held constant as we swept the amount of training data frome one to twenty hours. Bot the acoustic and language models exhibited the log-linear relationship between WER and the amount of training data. Independent of the amount of training data, the acoustic model degraded by a nearly constant 1.7% and the language model by 0.8%.

To evaluate the benefit of multiple transcriptions, we built two oracle systems. The *Turker oracle* ranks Turkers by the average error rate of their transcribed utterances against the professionals and selects utterances by Turker until the twenty hours is covered (Section 4.3 discusses a fair way to rank Turkers). The *utterance oracle* selects the best of the three different transcriptions per utterance. The best of the three Turkers per utterance wrote the best transcription two thirds of the time.

The utterance oracle only recovered half of the degradation for using non-professional transcription. Cutting the disagreement in half (from 23% to 13%) reduced the WER gap by about half (from 2.5% to 1%). Using the standard system combination algorithm ROVER (Fiscus, 1997) to combine the three transcriptions per utterance only reduced disagreement from 23% to 21%. While previous work benefited from combining multiple annotations, this task shows little benefit.

Transcription	Disagreement with LDC	ASR WER
Random Utterance	23%	42.0%
Random Turker	20%	41.4%
Oracle Utterance	13%	40.9%
Oracle Turker	18%	41.1%
Contractor	$< 5\%$	39.6%
LDC	-	39.5%

Table 1: Quality of Non-Professional Transcription on 20 hours of English Switchboard. Even though disagreement for random selection without quality control has 23% disagreement with professional transcription, an ASR system trained on the data is only 2.5% worse than using LDC transcriptions. The upper bound for quality control (row 3) recovers only 50% of the total loss.

4.2 Combining with External Sources

While in-domain speech transcription is typically the only effective way to improve the acoustic model, out-of-domain transcripts tend to be useful for language models of conversational speech. Broadcast News (BN) transcription is particularly well suited for English Switchboard data as the topics tend to cover news items like terrorism or politics. We built a small one million word language model (to simulate a resource-poor language) and interpolated it with varying amounts of LDC or Mechanical Turk transcriptions. Figure 2 details the results.

4.3 The Value of Quality Control

With a fixed transcription budget, should one even bother with redundant transcription to improve an ASR system? To find out, we transcribed 40 additional hours of Switchboard using Mechanical Turk. Disagreement to the LDC transcriptions was 24%, similar to the initial 20 hours. The two percent degradation of test WER when using Mechanical Turk compared to LDC held up with 40 and 60 hours of training.

Given a fixed budget of 60 hours of transcription, we compared the quality of 20 hours transcribed three times to 60 hours transcribed once. The best we could hope to recover from the three redundant transcriptions is the utterance oracle. Oracle and singly transcribed data had 13% and 24% disagreement with LDC respectively. System performance was 40.9% with 20 hours of

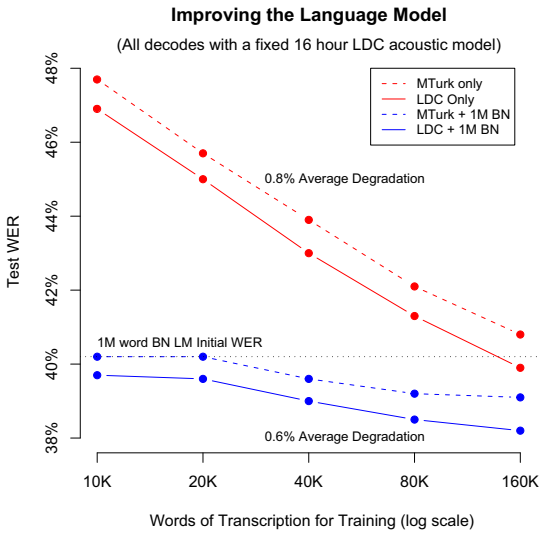

Figure 2: WER with a varied amount of LM training data and a fixed 16hr acoustic model. MTurk transcription degrades WER by 0.8% absolute across LM size. When interpolated with 1M words of broadcast news, this degradation shrinks to 0.6%.

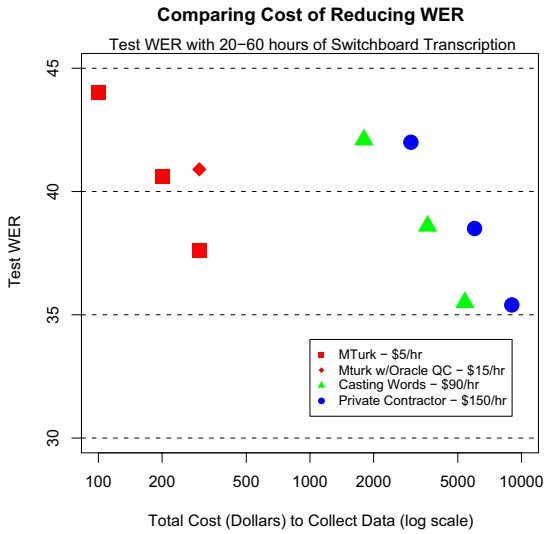

Figure 3: Historical cost estimates are $150 per hour of transcription (blue cirlces). The company Casting Words uses Turkers to transcribe English at $90 per hour which we estimated to be high quality (green triangles). Transcription without quality control on Mechanical Turk (red squares) is drastically cheaper at $5 per hour. With a fixed budget, it is better to transcribe more data at lower quality than to improve quality. Contrast the oracle WER for 20 hours transcribed three times (red diamond) with 60 hours transcribed once (bottom red square).

the former and 37.6% with 60 hours of the latter. Even though perfect selection cuts disagreement in half, three times as much data helps more.

The 2004 Fisher effort averaged a price of $150 per hour of English CTS transcription. The company CastingWords produces high quality (Passy, 2008) English transcription for $90 an hour using Mechanical Turk by a multi-pass process to collect and clean Turker-provided transcripts. While we did not use their service, we assume it is of comparable quality to the private contractor used earlier. The price for LDC transcription is not comparable here since it was intended for more precise linguistic tasks. Extrapolating from Figure 3, the entire 2000 Fisher corpus could be transcribed using Mechanical Turk at the same cost of collecting 60 hours of professional transcription.

5 Collection in Other Languages

To test the feasability of improving low-resource languages, we attempted to collect transcriptions for Korean, Hindi, Tamil CTS data. We built an LVCSR system in Korean since it is the only one with reference LDC transcriptions to use as a test set.

5.1 Korean

Korean is spoken by roughly 78 million speakers world wide and is written in Hangul, a phonetic orthography, although Chinese characters frequently appear in written text. Since Korean has essentially arbitrary spacing (Chong-Woo et al., 2001), we report Phoneme Error Rate (PER) instead of WER, which would be unfairly penalized. Both behave similarly as system performance improves. For comparison, an English WER of 39.5% has a PER of 34.8%.

We uploaded ten hours of audio to be transcribed once, again segmented into short snippets. Transcription was very slow at first and we had to pay $0.20 per HIT to attract workers. We posted a separate HIT to refer Korean transcribers, paying a 25% bonus of the income earned by referrals. This was quite successful as two referred Turkers contributed over 80% of the total transcription (at a cost of $25 per

hour instead of $20). We collected three hours of transcriptions after five weeks, paying eight Turkers $113 at a transcription rate of 10xRT.

Average Turker disagreement to the LDC reference was 17% (computed at the character level). Using these transcripts to train an LVCSR system instead of those provided by LDC degraded PER by 0.8% from 51.3% to 52.1%. For comparison, a system trained on the entire 27 hours of LDC data had 41.2% PER.

Although performance seems poor, it is sufficiently good to bootstrap with acoustic model self-training (Lamel et al., 2002). The language model can be improved by finding 'conversational' web text found with n-gram queries extracted from the three hours of transcripts (Bulyko et al., 2003).

5.2 Hindi and Tamil

As a feasability experiment, we collected one hour of transcription in Hindi and Tamil, paying $20 per hour of transcription. Hindi and Tamil transcription finished in eight days, perhaps due to the high prevalence of Turkers in India (Ipeirotis, 2008). While we did not have any professional reference, Hindi speaking colleagues viewed some of the data and pointed out errors in English transliteration, but overall quality appeared fine. The true test will be to build an LVCSR system and report WER.

6 Quality Control sans Quality Data

Although we have shown that redundantly transcribing an entire corpus gives little gain, there is value in *some* amount of quality control. We could improve system performance by only rejecting Turkers with high disagreement, similar to confidence selection for active learning or unsupervised training (Ma and Schwartz,). But if we are transcribing a truly new domain, there is no gold-standard data to use as reference, so we must estimate disagreement against errorful reference. In this section we provide a practical use for quality control without gold standard reference data.

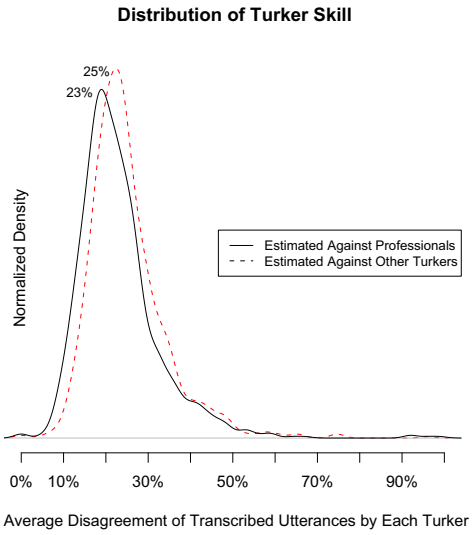

Figure 4: Each Turker was judged against professional and non-professional reference and assigned an overall disagreement. The distribution of Turker disagreement follows a gamma distribution, with a tight cluster of average Turkers and a long-tail of bad Turkers. Estimating with non-professionals (even though the reference is 23% wrong on average) is surprisingly well matched to professional estimate. Turker estimation over-estimated disagreement by only 2%.

6.1 Estimating Turker Skill

Using the twenty hour English transcriptions from Section 4, we computed disagreement for each Turker against the professional transcription for all utterances longer than four words. Note that each utterance was transcribed by three random turkers, so there is not one set of utterances which were transcribed by all turkers. Each Turker transcribed a different, partially overlapping, subset of the data.

For a particular Turker, we estimated the disagreement with other Turkers by using the two other transcripts as reference and taking the average. Figure 4 shows the density estimate of Turker disagreement when calculated against professional and non-professional transcription. On average, the non-professional estimate was 3% off from the professional disagreement.

Given that non-professional disagreement is a good estimate of professional disagreement

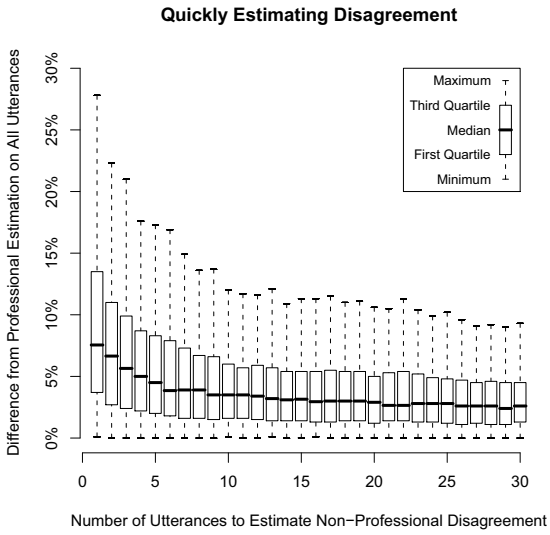

Figure 5: Boxplot of the difference of non-professional disagreement with a fixed number of utterances to professional disagreement over all utterances. While error is expectedly high with one utterance, 50% of the estimates are within 3% of the truth after ten utterances and 75% of the estimates are within 6% after fifteen utterances.

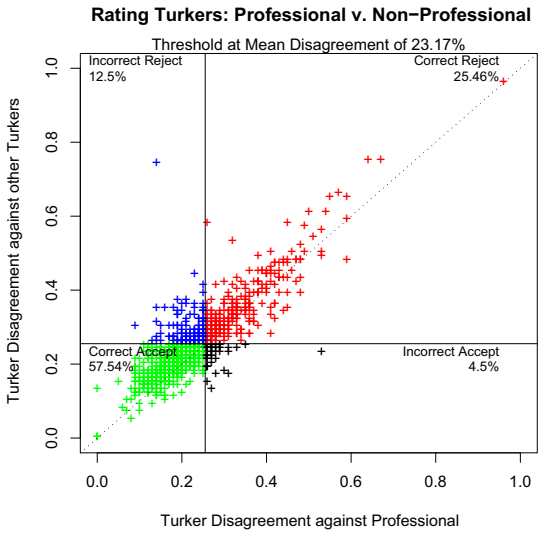

Figure 6: Each Turker is a point with professional (X axis) plotted against non-professional (Y axis) disagreement. The non-professional disagreement correlates surprisingly well with professional disagreement even though the transcripts used as reference are 23% wrong on average. By setting a selection threshold, the space is divided into four quadrants. The bottom left are correctly accepted: both non-professional and professional disagreement are below the threshold. The top left are incorrectly rejected: using their transcripts would have helped, but they don't hurt system performance, just waste money. The top right are correctly rejected for having high disagreement. The bottom right are the troublesome false positives that are included in training but actually may hurt performance. Luckily, the ratio of false negatives to false positives is usually much larger.

over all of a Turker's utterances, we wondered how few needed to be redundantly transcribed by other non-professionals. For each Turker, we started by randomly selecting one utterance and computed the non-professional disagreement. We compared the estimate to the true professional disagreement over all of the utterances and repeatedly sample 20 times. Then we increased the number of utterances used to estimate non-professional disagreement until all utterances by that Turker are selected.

Figure 5 shows a boxplot of the differences of non-professional to professional disagreement on *all* utterances. As few as fifteen utterances need to be redundantly transcribed to accurately estimate three out of four Turkers within 5% of the professional disagreement.

6.2 Finding the Right Turkers

Since we can accurately predict a Turker's skill with as few as fifteen utterances on average, we can rank Turkers by their professional and non-professional disagremeents. By thresholding on disagreement, we can either select good turkers or equivalently reject bad turkers. We can view the ranking as a precision/recall problem to select only the 'good' Turkers below the threshold. Figure 6 plots each Turker where the X axis is the professional disagreement and the Y axis is the non-professional disagreement. Sweeping the disagreement threshold from zero to one generates Figure 7, which reports F-score (the harmonic mean of precision and recall). This section suggests a concrete qualification test by first transcribing 15-30 utterance multiple times to create a gold standard. Using the transcription from the best Turker as reference, approve new Turkers with a WER less than the average WER from the initial set.

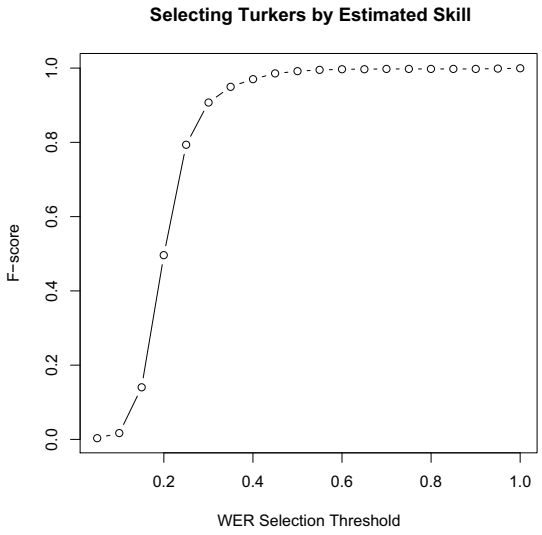

Figure 7: It is difficult to find only good Turkers since the false positives outnumber the few good workers. However, rejecting bad Turkers becomes very easy once past the mean error rate of 23%. It is better to use disagreement estimation to reject poor workers instead of finding good workers.

7 Experience with Mechanical Turk

We initially expected to invest most of our effort in managing Turker transcription. But the vast majority of Turkers completed the effort in good faith with few complaints about pay. Many left positive comments[5] despite the very difficult task. Indeed, the author's own disagreement on a few dozen English utterances were 17.7% and 26.8% despite an honest effort.

Instead, we spent most of our time normalizing the transcriptions for English acoustic model training. Every single misspelling or new word had to be mapped to a pronunciation in order to be used in training. We initially discarded any utterance with an out of vocabulary word, but after losing half of the data, we used a set of simple heuristics to produce pronunciations. Even though there were a few thousand of these errors, they were all singletons and had little effect on performance. Turkers sometimes left comments in the transcription box such as "no audio" or "man1: man2:". These errant transcriptions could be detected by force aligning the transcript with the audio and rejecting any with low scores (Lamel et al., 2000). Extending transcription to thousands of hours will require robust methods to automatically deal with errant transcripts and additionally run the risk of exhausting the available pool of workers.

Finding Korean transcribers required the most creativity. We found success in interacting with the transcribers, providing feedback, encouragement and paying bonuses for referring other workers. Cultivating workers for a new language is definitely a 'hands on' process.

For Hindi and Tamil, Turkers sometimes misinterpreted or ignored instructions and translated into English or transliterated into Roman characters. Additionally, *some* linguistic knowledge is required to classify phonemic categories (like fricative or sonorant) required for acoustic model training.

[5]One Turker left a comment "You don't grow pickles!!" in regards to the misinformed speakers she was transcribing.

8 Conclusion

Unlike previous work which studied the quality of Mechanical Turk annotations alone, we judge its value in terms of the real task: improving system performance. Despite relatively high disagreement with professional transcription, data collected with Mechanical Turk was nearly as effective for training speech models. Since this degradation is so small, redundant annotation to improve quality is not worth the cost. Resources are better spent collecting more transcription. In addition to English, we demonstrated similar trends in Korean and also collected transcripts for Hindi and Tamil. Finally, we proposed an effective procedure to reduce costs by maintaining the quality of the annotator pool without needing high quality annotation.

Acknowledgments

This research was supported by the EuroMatrixPlus project funded by the European Commission by the DARPA GALE program under Contract No. HR0011-06-2-0001, and NSF under grant IIS-0713448 and by BBN Technologies. The views and findings are the authors' alone.

References

Ivan Bulyko, Mari Ostendorf, and A. Stolcke. 2003. Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In *HLT-NAACL*.

Chris Callison-Burch. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazons Mechanical Turk. *EMNLP*.

Seung-Shik Kang Chong-Woo, Chong woo Woo, and Kookmin Univerity. 2001. Automatic segmentation of words using syllable bigram statistics. In *6th Natural Language Processing Pacific Rim Symposium*.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *LREC*.

Jonathan G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover).

L. Gillick, J. Baker, J. Bridle, M. Hunt, Y. Ito, S. Lowe, J. Orloff, B. Peskin, R. Roth, and F. Scattone. 1993. Application of large vocabulary continuous speech recognition to topic and speaker identification using telephone speech. In *ICASSP*.

Jack Godfrey, Edward Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *ICASSP*.

Panos Ipeirotis. 2008. Mechanical turk: The demographics. `http://behind-the-enemy-lines.blogspot.com/2010/03/new-demographics-of-mechanical-turk.html`.

Owen Kimball, Chai-Lin Kao, Teodoro Arvizo, John Makhoul, and Rukmini Iyer. 2004. Quick transcription and automatic segmentation of the fisher conversational telephone speech corpus. In *RT04 Workshop*.

Lori Lamel, Jean luc Gauvain, and Gilles Adda. 2000. Lightly supervised acoustic model training. In *ISCA ITRW ASR2000*.

Lori Lamel, Jean luc Gauvain, and Gilles Adda. 2002. Lightly supervised and unsupervised acoustic model training. *Computer Speech and Language*, 16(1).

Jeff Ma and Rich Schwartz. Unsupervised versus supervised training of acoustic models. In *INTERSPEECH*.

Matthew Marge, Satanjeev Banerjee, and Alexander Rudnicky. 2010. Using the amazon mechanical turk for transcription of spoken language. *ICASSP*, March.

Ian McGraw, Alexander Gruenstein, and Andrew Sutherland. 2009. A self-labeling speech corpus: Collecting spoken words with an online educational game. In *INTERSPEECH*.

D. Miller, M. Kleber, C. Kao, O. Kimball, T. Colthurst, S.A. Lowe, R.M. Schwartz, and H. Gish. 2007. Rapid and Accurate Spoken Term Detection. In *INTERSPEECH*.

Charles Passy. 2008. Turning audio into words on the screen. `http://online.wsj.com/article/SB122351860225518093.html`.

R. Prasad, S. Matsoukas, CL Kao, J.Z. Ma, DX Xu, T. Colthurst, O. Kimball, R. Schwartz, J.L. Gauvain, L. Lamel, et al. 2005. The 2004 BBN/LIMSI 20xRT English conversational telephone speech recognition system. In *INTERSPEECH*.

Brandon Roy and Deb Roy. 2009. Fast transcription of unstructured audio recordings. In *INTERSPEECH*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *EMNLP*.

Steve. Young, Jost. Schatzmann, Karl. Weilhammer, and Hui. Ye. 2007. The hidden information state approach to dialog management. In *ICASSP*.

Omar F. Zaidan and Chris Callison-Burch. 2009. Feasibility of human-in-the-loop minimum error rate training. In *EMNLP*.