ANALYZING QUALITY OF CROWD-SOURCED SPEECH TRANSCRIPTIONS OF NOISY AUDIO FOR ACOUSTIC MODEL ADAPTATION

Kartik Audhkhasi, Panayiotis G. Georgiou and Shrikanth S. Narayanan
Signal Analysis and Interpretation Lab (SAIL)
University of Southern California, Los Angeles, CA
audhkhas@usc.edu, {georgiou,shri}@sipi.usc.edu

ABSTRACT

The accuracy of crowd-sourced speech transcriptions varies depending on a variety of factors. This paper studies the impact of one such factor, namely, the quality of audio. We employed a speech database with babble noise at three SNR levels (clean, 2 dB and -2 dB) and asked workers on Amazon Mechanical Turk to transcribe it. Two interesting observations emerge. First, as expected, the quality of transcripts combined by word frequency based ROVER decreases with decreasing SNR. Further, we demonstrate that the use of some unsupervised reliability scores can improve the transcription quality, with increasing benefits at lower SNR. Second, we do not observe a significant drop in the performance of acoustic models adapted with increasing transcription noise. This highlights the surprising robustness of crowd-sourced transcripts for acoustic model adaptation.

Index Terms— Crowd-sourcing, speech transcription, automatic speech recognition

1. INTRODUCTION

Clean and accurate transcription of speech is crucial for the development of spoken language processing applications such as speech recognition and speech-to-speech translation systems. However, high quality professional speech transcription is expensive and time consuming. This is mainly because such transcriptions are done by well-trained experts who are acquainted with the potential variabilities in audio (such as speaking rate, pronunciation, accent) and the transcription conventions of the domain. Due to these limitations, there has been extensive interest recently in utilizing crowd-sourcing services like Amazon Mechanical Turk (MTurk) and Crowd Flower[1] for speech transcription. MTurk is a web-service which allows *workers* from all over the world to perform some *Human Intelligence Tasks* (HITs) required by *requesters*. The requester designs a web template form through which the workers submit their work, and pays a pre-specified amount of money for every satisfactory HIT. Crowd Flower is similar to MTurk except that it incorporates an intermediate layer of quality control in the work flow.

The natural language processing (NLP) community has explored crowd-sourcing for data collection and system evaluation. Snow et al. [1] demonstrate strong correlation between non-expert and expert annotations in five text processing tasks – affect recognition, word similarity, textual entailment recognition, temporal event recognition and word sense disambiguation. Callison-Burch shows that the MTurk can be used to evaluate machine translation quality [2]. Lambert et al. [3] describe the annotation of a linguistic plausibility database, where sentences generated from an N-gram language model (LM) are rated by workers in MTurk. Some further examples of the use of crowd-sourcing in NLP include paraphrasing for English-Arabic machine translation [4], building parallel corpora for machine translation systems [5] and elicitation of Wikipedia articles [6].

There have also been many works within the speech recognition and processing domain regarding the use of crowd-sourcing. Marge et al. [7] describe transcription of audio consisting of route instructions of robots using MTurk. They use word frequency based ROVER [8] for combining the multiple noisy transcriptions and achieve a reduction in transcription error rate with respect to a gold standard. Novotney and Callison-Burch [9] demonstrate that an automatic speech recognition (ASR) system trained on MTurk transcriptions achieves similar performance as one trained on expert transcriptions, but with only a fraction of the cost. Roy et al. [10] conclude that the force alignment score from an ASR system is significantly correlated with the inter-annotator agreement, and thus can be used for predicting transcription difficulty. A special session in InterSpeech-2011 focussed on the use of crowd-sourcing in speech processing [11]. Papers presented at the session ranged over a wide-variety of applications such as speech perception [12], transcription quality control [13, 14], prosodic annotations [15] and spoken document retrieval [16].

In our previous work, we have shown that the use of some unsupervised reliability metrics for MTurk transcripts can help reduce the transcription error rate and improve acoustic model performance for ASR [17, 18]. However, the audio transcribed in those works was noise-free, which resulted in very high inter-transcriber agreement. This paper attempts to understand the nature of non-expert transcriptions for noisy audio. For our study, we use 2000 audio clips from the 1997 English Broadcast News Speech corpus (HUB4) that we artificially corrupt with additive babble noise at 3 levels from the NOISEX database. We observe a significant degradation in transcription error rate with respect to LDC transcriptions with increasing noise. ROVER combination using unsupervised reliability scores performs better than word frequency based ROVER, and the benefit increases with decreasing SNR. We also use the resulting fused transcripts for acoustic model adaptation and observe insignificant drop in performance with increasing noise, which indicates that an increase in transcription error rate does not necessarily impact the adapted acoustic model performance.

This paper is organized as follows - the next section discusses the database preparation and its transcription in MTurk. We present some statistics collected from the MTurk transcriptions in subsection 2.1. Section 3 discusses the unsupervised transcription reliability scores used for fusing transcripts in ROVER. Experiments on transcription accuracy and acoustic model adaptation are discussed in section 4. We conclude the paper in section 5 and present some directions for future work.

[1] mturk.com, crowdflower.com

2. AUDIO TRANSCRIPTION IN MECHANICAL TURK

We select 2000 audio clips from the 1997 English Broadcast News Speech corpus (HUB4), having an average duration of 3.3 seconds. Babble noise from the NOISEX corpus was added to each of these clips at three levels of SNR – clean (no noise added), 2 dB and -2 dB. A simple HTML webpage was prepared for MTurk with the following instructions:

"The task is to hear a set of English broadcast audio files and transcribe the voice of the main speaker. Some clips may contain background babble noise. Please follow the instructions given below. Failure to do so would result in rejection of the HITs."

This was followed by instructions about transcribing non-alphabetical characters, numbers and special symbols, acronyms, unintelligible words, fragments/partial words and non-speech events. These instructions have been omitted from this paper due to space constraints. In addition, the workers were asked to answer the following questions for every HIT:

- *Audio quality: 5 (excellent), 4 (good), 3 (fair), 2 (poor), 1 (bad).*
- *Difficulty of transcribing: 3 (high), 2 (moderate), 1 (low).*
- *Confidence of transcription: 3 (high), 2 (moderate), 1 (low).*

Answers to these questions will be later used as predictors of transcription quality. To get an idea of the worker demographics, we also asked them to fill the following one-time questions:

- *Native Language: English, Other.*
- *Previous audio transcription experience: Yes/No.*

A total of 18000 HITs (2000 audio clips, 3 noise levels, 3 HITs per clip) were offered at a price of $0.03/HIT. This price was fixed based on our previous experience with audio transcription in MTurk. The HITs were reviewed for quality every day. In our prior experience with audio transcription in MTurk, we have observed that most transcription errors are merely convention differences and minor spelling mistakes. Hence, our quality check simply consisted of verifying the completed HITs for non-compliance with the provided instructions. An attempt was made to keep the text cleaning to a minimum, since we wanted to capture the inherent noise in the non-expert transcriptions. The next section presents an analysis of various transcription statistics across the three noise levels.

2.1. Analysis of Transcription Statistics

Table 1 shows the distribution of audio quality scores for three noise scenarios. We observe that the worker scores indeed reflect the change in audio quality. Next, we analyse the responses to the other two per-clip questions – *difficulty* of transcribing and *confidence* of transcription, as shown in Table 2. While the distributions shift to high difficulty and low confidence scores with increasing noise, the mode remains at low difficulty and high confidence. This is in consonance with our observations in [18] – the workers are overly optimistic about the quality of transcription.

Next, we consider the correlation coefficients between the various scores for the three noise conditions in Table 3. As expected, we observe strong correlation between the three scores. Interestingly, the correlation coefficients increase with a decrease in SNR. One possible reason for this is a reduction in the heavy skew of the three score distributions with an increase in noise level. In other words, at an SNR of -2 dB, the variation in scores is much more pronounced as compared to the case of clean audio.

	Audio Quality				
Score	1	2	3	4	5
Clean	0	2	5	23	**70**
2 dB	1	5	22	**47**	25
-2 dB	5	14	**37**	33	11

Table 1. Distribution (in %) of audio quality scores given by MTurk workers for three noise scenarios.

	Difficulty of transcribing			Confidence of transcription		
Score	1	2	3	1	2	3
Clean	**84**	14	2	1	8	**91**
2 dB	**66**	29	5	4	17	**79**
-2 dB	**48**	37	15	11	23	**66**

Table 2. Distribution (in %) of difficulty and confidence of transcription scores given by MTurk workers for three noise scenarios.

The next section discusses a few unsupervised transcription reliability scores (similar to our work in [17, 18]) which can be used in ROVER for combining the multiple noisy transcripts corresponding to a given audio file.

3. UNSUPERVISED TRANSCRIPTION RELIABILITY SCORES

Given multiple noisy transcripts for an audio clip, one can use word frequency based ROVER [8] for combination. As noted in previous works, such a process gives an appreciable reduction in the transcription error rate. However, such an approach assumes that the transcripts being combined are equally reliable. Furthermore, in a practical scenario, a reference transcript is not available for computing this reliability. Thus our approach consists of using worker responses to the three per-clip questions and normalized scores from a generic acoustic and language model to compute the overall reliability. These scores are unsupervised in the sense that the true reference transcript is not used. The word frequency based ROVER transcript is used as a proxy for the true reference, even for tuning of the hyperparameters.

Let S be the number of audio clips in the training database. Let t_{sw_i} be the transcription of clip s by a worker w_i, $i \in \{1, 2, ..., N\}$ where N is the number of (unique) workers transcribing each clip. Let r_s^0 denote the transcription obtained by word frequency based ROVER combination of $\{t_{sw_1}, ..., t_{sw_N}\}$. We assume r_s^0 to be a proxy for the gold standard transcription of s and compute the following unsupervised reliability metrics for t_{sw_i}:

1. *Per-clip question based scores*: The difficulty rating is subtracted from 1, all scores are normalized to [0,1], and then added to yield the total question-based confidence score. Let this score be denoted by $a_{\mathrm{q}}(s, i)$.

Correlation coefficient	(Quality, Difficulty)	(Quality, Confidence)	(Difficulty, Confidence)
Clean	-0.52	0.42	-0.50
2 dB	-0.57	0.49	-0.61
-2 dB	-0.59	0.53	-0.65

Table 3. Correlation coefficients between the scores from the three per-HIT questions for the three noise scenarios.

2. *Average worker HIT acceptance rate*: This is the fraction of a worker's overall HITs which have been accepted throughout his/her history in MTurk, and is denoted by $a_{\text{ar}}(w_i)$.

3. *Normalized force alignment score*: Given a baseline acoustic model, we force align r_s^0 and t_{sw_i} $\forall i \in \{1, ..., N\}$ for each audio clip s. Let the force alignment log-likelihoods be $L(r_s^0)$ and $L(t_{sw_i})$ respectively. The ratio of $L(r_s^0)$ to $L(t_{sw_i})$ can be taken as a reliability score since the force alignment score for a transcript is expected to increase and come closer to the score for r_s^0 as its reliability increases. We denote this score by $a_{\text{fa}}(s, i)$.

4. *Normalized LM score*: Similar to the force alignment score, we can compute the log-likelihoods of r_s^0 and t_{sw_i} using a baseline LM and take its ratio as an indicator of the quality of the MTurk transcript. This score is denoted by $a_{\text{lm}}(s, i)$.

5. *Word error rate*: Assuming r_s^0 to be a proxy for the true reference transcript, the word error rate between r_s^0 and t_{sw_i} (denoted by $a_{\text{wer}}(s, i)$) is another simple reliability score.

To reduce the number of hyperparameters in the model, we compute the overall reliability score as follows:

$$a(s,i) = \frac{1}{7}\Big\{\beta \sum_{k\in\{\text{q,ar}\}} a_k(s,i) + (1-\beta) \sum_{k\in\{\text{fa,lm,wer}\}} a_k(s,i)\Big\}$$

where $\beta \in [0, 1]$ is a hyper-parameter to be tuned. Note that we normalize by 7 and not 5 since $a_{\text{q}}(s, i) \in [0, 3]$.

4. EXPERIMENTS AND RESULTS

We first present experiments on the estimation the true LDC transcripts using the multiple noisy MTurk transcripts.

4.1. Estimation of True Reference LDC Transcript

We trained the baseline acoustic models (AMs) in Sphinx-III [19] using the Wall Street Journal (WSJ) corpus and databases from the DARPA TRANSTAC program. Gaussian mixture models with 32 components were used in these models. The number of tied states was set to 3000. The LM used for computing the LM-based reliability scores was trained on a large text corpus which included WSJ, DARPA databases and text collected from the web by data mining. 20 % of the 2000 sentences put up on MTurk for transcription were set aside for tuning the two hyper-parameters – β and α (weight given to word frequency while combining it with the total reliability score in ROVER). It was observed that tuning these hyper-parameters on the true LDC transcripts or the word frequency based ROVER transcripts of the held-out set gave the same performance for all cases.

Table 4 shows the word and sentence error rates (WER and SER) in estimating the true LDC transcripts for a variety of cases. The following observations can be made:

1. As has been previously reported in many papers, simple word frequency based ROVER gives an appreciable decrease in transcription error rate.

2. Incorporation of the unsupervised reliability scores decreases the error rates further.

3. The benefit gained by incorporating reliability scores increases with increasing noise. Its relative improvement over word frequency based ROVER increases from 2% to 6.7% (WER) and from 2 % to 2.9 % (SER) from the clean to -2 dB noise case. While this increase in performance benefit may not be large, it does indicate the usefulness of the reliability scores for combining very noisy transcripts. One intuition behind this trend is that in case of extremely noisy transcripts, using solely word frequencies in each confusion bin for ROVER is not sufficient. Highly frequent words in a bin may be erroneous, and thus the incorporation of some form of reliability score in the fusion process is expected to provide benefit.

	No ROVER		Word frequency based ROVER		Reliability weighted ROVER	
	WER	SER	WER	SER	WER	SER
Clean	10.2	53.3	8.2 *-19.6*	46.4 *-13.0*	**8.0** *-21.6*	**45.3** *-15.0*
2 dB	16.4	68.8	13.3 *-18.9*	63.1 *-8.3*	**12.7** *-22.6*	**61.3** *-10.9*
-2 dB	26.9	79.6	23.3 *-13.4*	76.5 *-3.9*	**21.5** *-20.1*	**74.2** *-6.8*

Table 4. Variation of transcription error rate in estimating the true LDC transcripts with noise. The "No ROVER" case used all the MTurk transcripts without fusion. All numbers are in percent, and the numbers in italics represent the percent decrease in error rate relative to the case without ROVER.

Based on the transcription error rates for word frequency based ROVER, it is natural to expect that acoustic models adapted using these transcripts and corresponding to clean audio will be significantly poorer than ones adapted on the LDC transcripts. The next subsection discusses experiments and results in this direction.

4.2. Acoustic Model Adaptation using MLLR

We next considered adaptation of the baseline acoustic models using maximum likelihood linear regression (MLLR) [20]. For these experiments, the background LM was mixed with a LM trained on the text from the HUB4 corpus excluding the set of 2000 sentences put up on MTurk and an evaluation set of 1000 files. All out of vocabulary (OOV) words in the MTurk transcript were added to the dictionary after passing through a grapheme-to-phoneme converter [21].

The baseline (unadapted) acoustic models gave a WER and SER of 42.8 % and 86.5 % respectively. On the other hand, acoustic models trained on 50 hours of audio from the HUB4 corpus gave a WER of 28.8 % and SER of 74.7 %. These two error rates provide a lower bound on the expected recognition performance after MLLR. Table 5 shows the word and sentence error rates of various acoustic models. It is interesting to note that the performance of the adapted models does not vary significantly even with the use of noisy transcripts (fused using word frequency based ROVER) for adaptation. In case of 1 regression class, MLLR applies a single transformation matrix and bias vector to the means of all Gaussians. This matrix is estimated using frames pooled together from all phonemes, irrespective of context. Due to the substantial averaging, transcription errors are not expected to affect the acoustic model drastically.

In case of 47 regression classes (one class for all triphones having a particular basephone at the center), one has to estimate a different regression matrix for each subset of Gaussians. The little variation in performance can now be explained by looking at the transcription error rate in terms of phonemes. The LDC reference and word frequency-based ROVER transcripts were converted to phoneme sequences using a simple dictionary lookup. It was observed that the phoneme error rate with respect to LDC was 4.5 %, 8.9 % and 18.3 % for clean, 2 dB and -2 dB noise cases respectively.

	HUB4 models		Unadapted WSJ models	
	WER	SER	WER	SER
	28.8	74.7	42.8	86.5
	WSJ Adapted on LDC		WSJ adapted on clean MTurk word frequency based ROVER	
	WER	SER	WER	SER
1 class	40.0	84.7	39.9	84.4
47 classes	37.5	83.6	37.7	83.5
	WSJ adapted on 2 dB MTurk word frequency based ROVER		WSJ adapted on -2 dB MTurk word frequency based ROVER	
	WER	SER	WER	SER
1 class	40.0	84.5	39.7	84.5
47 classes	37.7	83.3	38.1	83.1

Table 5. Variation of WER and SER before/after MLLR adaptation with 1 and 47 regression classes. All numbers are in percent.

Hence, the number of misclassified frames during force alignment is expected to be overshadowed by the overwhelming majority of the correctly assigned ones. This, and the averaging of parameters (considering that each MLLR transformation matrix is on an average estimated using approximately 10000 frames) can be one cause of the negligible performance degradation after adaptation.

5. CONCLUSION AND FUTURE WORK

This paper presented an analysis of the quality of transcripts obtained from Amazon Mechanical Turk for audio with varying levels of noise. It was observed that the use of unsupervised reliability scores benefits the ROVER combination, as compared to a combination using only word frequency information in each confusion bin. Moreover, this benefit increases with decrease in SNR, indicating that the proposed unsupervised reliability scores may be extremely useful in transcription of databases with corrupted audio. We also presented results on acoustic model adaptation using the noisy crowd-sourced transcripts and found out that the performance of adapted models is relatively robust to transcription noise. This can be attributed to the relatively low level of phoneme errors which took place during the transcription process.

Future work will involve crowd-sourcing experiments with databases with natural noise and distortions, exploration of more unsupervised transcription reliability scores and finding ways of estimating reliability scores at word rather than utterance level. Another direction of future work is on devising an active learning framework for transcribing audio on MTurk, with a closer coupling between the demands of the system being trained (e.g. ASR) and the kinds of HITs offered on MTurk.

6. ACKNOWLEDGEMENT

This work was supported by the NSF, DARPA and Army.

7. REFERENCES

[1] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks," in *Proc. EMNLP*. ACM, 2008, vol. 1, pp. 254–263.

[2] C. Callison-Burch, "Fast, cheap and creative: Evaluating translation quality using Amazon's Mechanical Turk," in *Proc. EMNLP*. ACM, 2009, vol. 1, pp. 286–295.

[3] B. Lambert, R. Singh, and B. Raj, "Creating linguistic plausibility dataset with non-expert annotators," in *Proc. InterSpeech*, 2010, pp. 1906–1909.

[4] M. Denkowski, H. Al-Haj, and A. Lavie, "Turker-assisted paraphrasing for English-Arabic machine translation," in *Proc. HLT*. NAACL, 2010, pp. 66–70.

[5] V. Ambati and S. Vogel, "Can crowds build parallel corpora for machine translation systems?," in *Proc. HLT*. NAACL, 2010, pp. 62–65.

[6] S. Novotney and C. Callison-Burch, "Crowdsourced accessibility: Elicitation of Wikipedia articles," in *Proc. NAACL-HLT*, 2010.

[7] M. Marge, S. Banerjee, and A. I. Rudnicky, "Using the Amazon Mechanical Turk for transcription of spoken language," in *Proc. ICASSP*, 2010.

[8] J. Fiscus, "A post processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proc. ASRU*. IEEE, 1997, pp. 347–354.

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

[10] B. C. Roy, S. Vasoughi, and D. Roy, "Automatic estimation of transcription accuracy and difficulty," in *Proc. InterSpeech*. ISCA, 2010, pp. 1902–1905.

[11] G. Parent and M. Eskenazi, "Speaking to the crowd: looking at past achievements in using crowdsourcing for speech and predicting future challenges," in *Proc. InterSpeech*, 2011.

[12] M. Cooke, J. Barker, M. L. G. Lecumberri, and K. Wasilewski, "Crowdsourcing for word recognition in noise," in *Proc. InterSpeech*, 2011.

[13] C-Y. Lee and J. Glass, "A transcription task for crowdsourcing with automatic quality control," in *Proc. InterSpeech*, 2011.

[14] S. Buchholz and J. Latorre, "Crowdsourcing preference tests, and how to detect cheating," in *Proc. InterSpeech*, 2011.

[15] K. Evanini and K. Zechner, "Using crowdsourcing to provide prosodic annotations for non-native speech," in *Proc. InterSpeech*, 2011.

[16] M. Goto and J. Ogata, "Podcastle: Recent advances of a spoken document retrieval service improved by anonymous user contributions," in *Proc. InterSpeech*, 2011.

[17] K. Audhkhasi, P. G. Georgiou, and S. S. Narayanan, "Accurate transcription of broadcast news speech using multiple noisy transcribers and unsupervised reliability metrics," in *Proc. ICASSP*, 2011.

[18] K. Audhkhasi, P. G. Georgiou, and S. S. Narayanan, "Reliability-weighted acoustic model adaptation using crowdsourced transcriptions," in *Proc. InterSpeech*, 2011.

[19] Carnegie Mellon University, "Sphinx-3," *Pittsburgh, Pennsylvania*.

[20] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer speech and language*, vol. 9, no. 2, pp. 171, 1995.

[21] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.