Reliability-Weighted Acoustic Model Adaptation Using Crowd Sourced Transcriptions

Kartik Audhkhasi, Panayiotis G. Georgiou and Shrikanth S. Narayanan

Signal Analysis and Interpretation Lab (SAIL)
Electrical Engineering Department
University of Southern California, Los Angeles, CA 90089-2564, USA
audhkhas@usc.edu, {georgiou, shri}@sipi.usc.edu

Abstract

This paper focuses on adaptation of acoustic models using speech transcribed by multiple noisy experts. A simple approach involves combining multiple transcripts using word frequency based Recognizer Output Voting Error Reduction (ROVER) followed by adaptation using the combined transcripts. But this assumes that the transcripts being combined are equally reliable. To overcome this assumption, we use two sets of scores to estimate this reliability. The first set is based on answers to some questions given by the transcribers. The second set is derived in an unsupervised way using the word frequency based ROVER transcripts and baseline acoustic models. The overall confidence is a convex combination of these scores and is used to perform a confidence weighted fusion. We adapt the baseline acoustic models using these combined transcripts. Recognition results for a Mexican Spanish ASR system show an absolute improvement of 0.5% in word error rate and 0.9% in sentence error rate.

Index Terms: Speech transcription, evaluator reliability, crowd sourcing

1. Introduction

Training of nearly all *spoken language* (SL) systems is done in a supervised manner. For example, training of an *Automatic Speech Recognition* (ASR) system requires the presence of a transcribed speech corpus, and a conventional *Statistical Machine Translation* (SMT) system requires parallel text in the source and target languages. Thus, accurate transcription of corpora is an essential ingredient in the development of SL systems. Typically this is done by well-trained transcribers who are not only acquainted with potential variabilities in the audio (such as due to accent, dialect and speaking rate), but are also knowledgeable in the pronunciation and orthographic conventions of the language. We argue that this expert transcription has two potential drawbacks. First, it is expensive and time consuming. This severely limits its use when the data has to be transcribed quickly within strict budget constraints. Second, the transcription conventions followed by the experts may not generalize to the data at hand. For example in [1], we observed that the LDC and Mechanical Turk (MTurk) transcriptions of Mexican Spanish broadcast news contained notational differences like “ex presidente” vs “expresidente” (both meaning “former president”) and “adónde” vs “donde” (both meaning “where”).

Crowd sourcing services like Amazon Mechanical Turk (MTurk)[1] promise ways to address both of the above drawbacks. MTurk is a web-service which allows *workers* from all over the world to perform some *Human Intelligence Tasks* (HITs) assigned by *requesters*. The requester makes a web interface through which the workers can complete the HITs. The workers are paid a pre-specified amount of money for each completed HIT, pending approval by the requester. MTurk received considerable attention initially from the natural language processing (NLP) community. Snow et al. [2] used MTurk for getting annotations for five NLP tasks – affect recognition, word similarity, recognizing textual entailment, temporal event recognition and word sense disambiguation. They demonstrated that non-expert annotations were as good as the gold standard annotations in terms of correlation and overall system performance. Callison-Burch [3] used MTurk for evaluating machine translation quality. He observed that the worker evaluations have a high agreement with the gold standard evaluations, and are a better indicator of translation quality than the often used BLEU score. Lambert et al. [4] created a linguistic plausability database by generating sentences from an N-gram language model (LM) and getting them annotated by workers in MTurk. A workshop in NAACL HLT-2010 focussed on annotation of speech and language data using MTurk, subject to a maximum budget of $100. Examples included paraphrasing for English-Arabic machine translation [5], building parallel corpora for machine translation systems [6] and elicitation of Wikipedia articles [7].

The ASR community has also used MTurk. Marge et al. [8] discuss transcription of audio consisting of route instructions of robots using MTurk. They demonstrate that a combination of multiple non-expert transcriptions using word frequency based ROVER [9] drastically reduces the WER of the combined transcripts with respect to a gold standard. Novotney and Callison-Burch [10] demonstrate that an ASR system trained with transcripts from MTurk gives similar performance as one trained on gold standard transcripts, at nearly one-thirtieth the cost. They also present a quality control scheme where disagreement between multiple transcribers is used as an estimate of their skill. Transcribers with skill less than a threshold are removed from ASR training. Recently, Roy et al. [11] have shown that the force alignment score from an ASR system is reasonably correlated with inter-transcriber agreement.

This work is an extension of [1], where we presented some simple unsupervised metrics of transcription reliability. Combination of multiple noisy transcripts using these reliability scores was shown to bring the combined transcript closer to an available gold standard, as compared to a combination without reliability information. It is important to note that the final metric of interest is the ASR word error rate (WER) – a direct measure of

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

overall system performance. Hence, this paper focuses on adaptation of acoustic models (AMs) using the reliability weighted combination of transcripts. Also in addition to the unsupervised reliability metrics proposed in [1], we use the worker's responses to some questions (asked during the transcription process) for obtaining an overall reliability score. MLLR adaptation using reliability-weighted ROVER combination gives an improvement in WER of the ASR system as compared to an unweighted ROVER combination. More importantly, the insights gained from this work are expected to motivate better ways of training an ASR system using crowd-sourced transcripts.

This paper is organized as follows – Section 2 describes the database, its pre-processing and the setup on MTurk. Section 3 presents an analysis on the responses of transcribers to various questions and other statistics (like time taken to complete a HIT). Section 4 describes the method used to obtain a transcription reliability score. Section 5 presents the experimental setup and the results of decoding a held-out test set using the adapted AMs. We conclude the paper in Section 6 with a summary of the observations and some directions for future work.

2. Mexican Spanish Audio Database

The audio database used in this work was collected as part of a NSF project on "An Integrated Approach to Creating Enriched Speech Translation Systems". The aim of this project is to develop a speech-to-speech (S2S) translation system for a doctor-patient interaction scenario, where the patient and doctor are proficient in different languages (Mexican Spanish and English respectively). The focus is on enriching the overall system by using various contextual features which are otherwise ignored in a conventional S2S system. This includes developing methods to detect and transfer speaker emotions, disfluencies etc. through the system, and using these cues to tightly integrate the otherwise pipelined system architecture.

As part of the initial thrust in this project, two small scale data collections were organized. Senior medical students and native Spanish actors (also called standardized patients) were given a set of scenarios involving an interaction between a doctor and a patient. Many sessions also had a bilingual interpreter. Each session was roughly 20 minutes long, and involved spontaneous conversation between the doctor, patient and interpreter. The audio was recorded using close-talking microphones.

A conventional S2S system consists of a sequence of an ASR system, SMT system and a text-to-speech (TTS) system. Thus, one of the initial tasks was to develop a Mexican Spanish ASR system using the collected audio data. Since the collected audio is in contiguous chunks of approximately 20 minutes each, an important pre-processing step was voice activity detection (VAD) and segmentation of the audio into small clips. The algorithm proposed in [12] was used for doing this. Clips with duration more than 1 second were selected for transcription on MTurk, as clips with smaller duration would be tough to transcribe and sometimes corresponded to VAD errors. This approximately gave 11 hours of audio (15911 clips) in the two languages combined. The next subsection describes the transcription setup.

2.1. Audio Transcription Setup in MTurk

A simple HTML page was created for transcription of the 15911 audio clips in MTurk. The instructions for annotation were as follows:

The task is to hear a set of Mexican-Spanish/English audio files and transcribe whatever you hear (please do not translate). The average length of a clip is 2-3 seconds. Please follow the instructions given below. Failure to do so would result in rejection of the HITs.

Further detailed instructions were related to the use of proper Mexican Spanish characters and words, transcription of disfluencies and unintelligible words etc. The following four questions were also asked for every HIT:

1. *Background noise/music: Mark yes/no.*
2. *Difficulty of transcribing: Rate on a scale of 1-3, 1 being very easy and 3 being extremely difficult.*
3. *Confidence of transcription: Rate on a scale of 1-3, 1 being not confident at all and 3 being totally confident.*
4. *Language of the clip: Mark Spanish/English/Mixed.*

In addition, the workers were also required to respond to the following one-time questions:

1. *Native language: Mark Spanish/English/Other.*
2. *Previous audio transcription experience: Mark yes/no.*

Three HITs were assigned for each audio clip and the reward for a HIT was set to $0.03 (based on our prior experience with audio transcription in MTurk). Only workers with at least 95% approval rate (defined as the percentage of the worker's HITs which were accepted) were allowed to accept the HITs. Regular review of the completed HITs was done, and all HITs with one or more unanswered fields were rejected. The overall rejection rate was approximately 1.64%. No manual cleaning of the transcripts was done. The entire batch of 15911 clips (47733 HITs) was completed in 2 weeks. Based on the labels provided by the workers, 9190 clips were in Mexican Spanish, corresponding to approximately 6 hours of audio. The next section discusses some statistics obtained from this process.

3. Analysis of Transcription Statistics

The distribution of answers to the first three per-clip questions is shown in Table 1. Interestingly, the workers thought that the clip was extremely difficult to transcribe for only 1.8% of the HITs. The same observation holds for confidence in transcription, where workers had low confidence in only 1.8% of the HITs. For an overwhelming majority of the HITs (more than 80%), the workers were extremely confident of their responses.

Table 1: *Relative frequencies of worker responses to the first three per-clip questions.*

Response →	1	2	3
(1) Background Music/Noise	0.137 (yes)	0.863 (no)	-
(2) Difficulty in transcribing	0.826	0.156	0.018
(3) Confidence in transcribing	0.018	0.130	0.852

Table 2 shows the correlation coefficient between the responses to the first three per-clip questions. We can observe that the correlation coefficient between the response to question 2 (difficulty in transcribing) and 3 (confidence in transcription) is large and negative. This is intuitive since we expect workers to be less confident about HITs that are difficult to transcribe. What is interesting is the small correlation between the presence of background noise/music and the response to the remaining two questions. It is natural to expect that the presence of background noise makes transcription difficult. However, the small

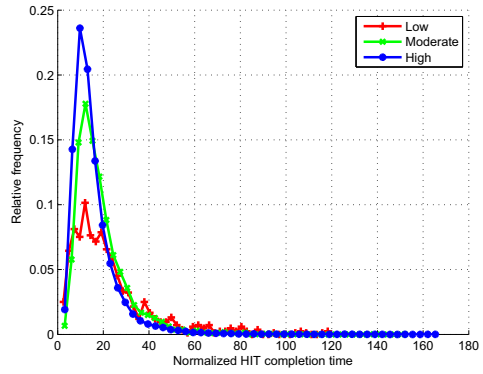

Figure 1: *Histograms of normalized HIT completion duration (completion duration divided by clip duration) for three levels of confidence of transcription.*

correlation values can be explained on the basis of the workers being over-confident of their responses most of the time.

Table 2: *Correlation coefficients between responses to the three per-clip questions.*

Question number	1	2	3
1	1.00	-0.09	0.17
2	-	1.00	-0.59
3	-	-	1.00

In addition to responses for all the requested questions, MTurk also records the time taken for the completion of each HIT. The correlation coefficient between the duration of the clip and time taken to complete the HIT was only 0.44. We hoped that this value would be much higher than what was observed, since it is intuitive to expect longer clips to take more time to complete. However, it is likely that the workers heard a clip multiple times (depending on the *true* difficulty of transcription) before transcribing.

Figure 1 shows the histogram of the normalized HIT completion duration (completion duration divided by clip duration) for responses to the question on confidence of transcription. We can conclude that in general, clips with high worker confidence invariably had a smaller normalized completion time than other clips. A similar trend was observed in the histograms for the responses to the other two questions - clips with no background noise and low difficulty in general had a lower normalized completion time.

4. A Transcription Reliability Score

We used two sets of features for obtaining a transcription reliability score. The first set is based on the worker responses to the first three per-clip questions. For the first question, absence of background noise was assigned a score of 1 and its presence was assigned 0. Scores for the question on transcription difficulty were assigned three levels - 0 for high, 1/3 for moderate and 1 for low difficulty. Scores for the question on confidence of transcription were also similarly assigned - 0 for low, 1/3 for moderate and 1 for high confidence. This score assignment is ad-hoc and better ways to generate numeric scores in $[0, 1]$ from Likert scale ratings can be further investigated in future work. We also incorporated the overall HIT acceptance rate of a worker as a feature in this set.

The second set of features was derived in an unsupervised manner using a baseline acoustic model and word frequency based ROVER transcripts, similar to our previous work [1]. 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 unavailable gold standard transcription of s and compute the following unsupervised reliability metrics for t_{sw_i}:

1. Normalized force alignment score: Given the baseline acoustic models, for each audio clip s, we force align r_s^0 and t_{sw_i} $\forall i \in \{1, ..., N\}$. Let the force alignment scores (log likelihoods) for r_s^0 and t_{sw_i} be $L(r_s^0)$ and $L(t_{sw_i})$. Intuitively, the force alignment score for a transcript is expected to increase and come closer to the force alignment score for r_s^0 as its reliability increases. Thus, the normalized force alignment score defined below is an indicator of the reliability of t_{sw_i}:

$$a_1(s, i) = \frac{L(r_s^0)}{L(t_{sw_i})} \tag{1}$$

2. WER with respect to ROVER transcript: Another simple transcription reliability metric is the WER between r_s^0 and t_{sw_i}, denoted by $a_2(s, i)$.

Once the two feature sets have been computed, we find the worker-wise means of each feature over the entire training set and append them to original sets. This is done to compensate (or smooth) any abnormal scores for each transcription with the global averages. To avoid a large number of hyper parameters to tune, we compute the individual mean of the two sets of scores (each set including the worker-wise means). Let the mean of the question based scores for worker i and clip s be $a_q(s, i)$, and the mean of the set of unsupervised features be $a_u(s, i)$. The overall transcription reliability score for t_{sw_i} is computed to be a convex combination of these two mean scores:

$$a(s, i) = \beta a_q(s, i) + (1 - \beta) a_u(s, i) \tag{2}$$

where $\beta \in [0, 1]$ is a hyper parameter to be tuned.

5. Experimental Setup and Results

Baseline continuous observation, 32 mixture component, triphone acoustic models with 3000 tied states were trained in Sphinx-3 [13] on 26 hours of Mexican Spanish audio from the HUB4-NE broadcast news corpus [14]. A trigram language model was trained in SRILM [15] using the LDC transcripts corresponding to the AM training data. When tested on a held-out set from HUB4-NE corpus, the WER of the ASR was 32%.

The 9190 Mexican Spanish audio clips transcribed in MTurk were split into random training, test and development sets of 6390, 1400 and 1400 clips each. We first decided to test the performance of the baseline acoustic models on the MTurk test set. Since the MTurk transcripts contain many out of vocabulary (OOV) words, possibly due to spelling errors, we trained a trigram LM using the transcripts of the MTurk training set without any ROVER combination. This LM was then interpolated with broadcast news LM with a weight of 0.8 being given to the former. Since the gold-standard transcriptions are not

available for the MTurk corpus, we combined the 3 transcripts for each test clip using word frequency based ROVER. The resulting transcripts were used as a proxy for the gold standard, and compared with the output of the ASR system for getting the WER in all experiments. The broadcast news AM and interpolated LM (MTurk + broadcast news) gave a WER of 40.7% on the test set.

To set the baseline for adaptation, we used the word frequency based ROVER transcripts of the training set and the corresponding audio to adapt the broadcast news acoustic models using one iteration of maximum likelihood linear regression (MLLR) in Sphinx. All tied states corresponding to the same basephone were assigned a single regression class. Thus we used 27 regression classes. Since the MTurk transcripts contained many OOV words, a simple rule-based letter-to-sound rule converter for Mexican Spanish was used to generate pronunciations for OOVs, which were then added to the dictionary. The WER of the resulting ASR system on the test set was 36.5% when using the word frequency based ROVER transcripts as reference. This is an improvement of 4.2% absolute as compared to the unadapted acoustic models. We also tried to use a single regression class (i.e. a global regression matrix) while performing adaptation. The adapted AMs gave a WER of 37% on the test set. As expected, the improvement was less than that obtained in the case of 27 regression classes.

To understand the benefit of fusing the three transcripts for each clip, we randomly selected one transcript per clip in the training set. Baseline AMs adapted on the resulting set gave a WER of 43.1% – poorer than even the baseline unadapted AM performance. This highlights the benefit of fusing multiple noisy transcripts.

Next, we tested the effect of introducing the previously discussed transcription reliability scores in ROVER. After the reliability scores have been computed as discussed in Section 4, two hyper-parameters have to be tuned. The first one is β - the weight given to the question based average reliability score. The second is $\alpha \in [0, 1]$ - the weight given to the word frequency in each confusion bin in the `avgconf` method in SCTK's [16] implementation of ROVER. A full grid search with step size of 0.1 would have been computationally too expensive. Thus we first optimized α keeping β fixed at 0.6, approximately corresponding to equal weight being given to the two average scores. Then, β was optimized keeping α at its optimal value. The best α and β values based on the WER on the development set were found out to be 0.8 and 0.1. The corresponding adapted AMs gave a WER of 36%, and absolute improvement of 0.5% over the AMs adapted using word frequency based ROVER. The sentence error rate (SER) reduced from 72.6% to 71.7% – an improvement of 0.9%.

6. Conclusion and Future Work

This paper extended our previous work in [1] by evaluating the effect of incorporating transcription reliability information on the WER of an ASR system with adapted acoustic models. When compared with an adapted system using word frequency based ROVER, we obtained an improvement of 0.5% (absolute) in WER. While this improvement may not be large, it does show the merit of incorporating the proposed transcription reliability scores into the ROVER based combination process. We also found out that the scores based on worker responses to some per-clip questions were less beneficial in improving the WER (as evident by the optimal β of 0.1) when compared with the unsupervised scores. This may be attributed to the observation that most of the time, the workers are overly confident about their transcriptions.

There are many directions of future work in this domain. Finding more accurate ways to estimate transcription quality is an obvious direction. It would also be interesting to find a better way to evaluate the performance of adapted ASR system without using a gold standard reference, which is unavailable. One intuitive solution is to get the test set transcribed by trained experts and treat that as the gold standard reference. However, there is no way to guarantee that the transcription convention used by the experts is general enough to accommodate the variability observed in the perfectly correct non-expert transcripts. Asking the experts to correct the spelling errors in the non-expert transcripts is another solution.

7. Acknowledgement

This work was supported by DARPA and NSF.

8. References

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

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

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

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

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

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

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

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

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

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

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

[12] P. K. Ghosh, A. Tsiartas, P. G. Georgiou, and S. S. Narayanan, "Robust voice activity detection in stereo recording with crosstalk," in *Proc. InterSpeech*, 2010, pp. 3098–3101.

[13] Carnegie Mellon University, "Sphinx-3," *Pittsburgh, Pennysylvania*.

[14] HUB4-NE, "1997 Spanish Broadcast News Speech," *Linguistic Data Consortium*, 1998.

[15] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. ICSLP*, 2002, pp. 901–904.

[16] SCTK Version 1.2c, "NIST Speech Recognition Scoring Toolkit," *NIST*, 2008.