Evaluation of crowdsourcing transcriptions for African languages

Hadrien Gelas[1,2], Solomon Teferra Abate[2], Laurent Besacier[2], François Pellegrino[1]
[1]Laboratoire Dynamique Du Langage, CNRS - Université de Lyon, France
[2]Laboratoire Informatique de Grenoble, CNRS - Université Joseph Fourier Grenoble 1, France
{hadrien.gelas, francois.pellegrino}@univ-lyon2.fr, {solomon.abate, laurent.besacier}@imag.fr

Abstract

We evaluate the quality of speech transcriptions acquired by crowdsourcing to develop ASR acoustic models (AM) for under-resourced languages. We have developed AMs using reference (REF) transcriptions and transcriptions from crowdsourcing (TRK) for Swahili and Amharic. While the Amharic transcription was much slower than that of Swahili to complete, the speech recognition systems developed using REF and TRK transcriptions have almost similar (40.1 vs 39.6 for Amharic and 38.0 vs 38.5 for Swahili) word recognition error rate. Moreover, the character level disagreement rates between REF and TRK are only 3.3% and 6.1% for Amharic and Swahili, respectively. We conclude that it is possible to acquire quality transcriptions from the crowd for under-resourced languages using Amazon's Mechanical Turk. Recognizing such a great potential of it, we recommend some legal and ethical issues to consider.

1 Foreword

This paper deals with the use of Amazon's Mechanical Turk (MTurk) which is a subject of controversy among researchers for obvious legal and ethical issues. The goal of this paper is to evaluate the quality of the data produced via crowdsourcing and not to produce a mass of data for a low price (in this experiment, we have actually re-transcribed speech data for which we already had transcriptions). Ethical issues on working with MTurk are discussed in the last section of this paper where guidelines of "good conduct" are proposed.

2 Introduction

Speech transcriptions are required for any research in speech recognition. However, the time and cost of manual speech transcription make difficult the collection of transcribed speech in all languages of the world.

Amazon's Mechanical Turk (MTurk) is an online market place for work. It aims at outsourcing difficult or impossible tasks for computers called "*Human Intelligence Tasks*" (HITs) to willing human workers ("*turkers*") around the Web. Taking use of this "crowd" brings two important benefits against traditional solutions (employees or contractors): repetitive, time consuming and/or costly tasks can be completed quickly for low payment.

Recently MTurk has been investigated as a great potential to reduce the cost of manual speech transcription. MTurk has been previously used by others to transcribe speech. For example, (Gruenstein et al., 2009; McGraw et al., 2009) report near-expert accuracy by using MTurk to correct the output of an automatic speech recognizer. (Marge et al., 2010b) combined multiple MTurk transcriptions to produce merged transcriptions that approached the accuracy of expert transcribers.

Most of the studies conducted on the use of MTurk for speech transcription take English as their subject of study which is one of the well resourced languages. The studies on English, including (Snow et al., 2008; McGraw et al., 2009), showed that MTurk can be used to cheaply create data for natural language processing applications. However, MTurk is not yet widely studied as a means to acquire useful data for under-resourced languages except a research conducted recently (Novotney and Callison-Burch, 2010) on Korean, Hindi and Tamil. On the other hand, there is a growing research interest towards speech and language processing for under-resourced and African languages. Specific workshops in this domain are appearing such as SLTU (Spoken

Languages Technologies for Under-resourced languages[1]) and AfLaT (African Language Technology[2]). Moreover, (Barnard et al., 2010a; Barnard et al., 2010b) highlighted interests using Automatic Speech Recognition for information access in Sub-Saharan Africa, with a focus on South-Africa.

In this paper we investigate the usability of MTurk for speech transcription to develop Automatic Speech Recognition (ASR) for two under-resourced African languages without combining transcription outputs. In Section 3, we review some of the works conducted on the use of MTurk for speech transcription. We then describe our experimental setups including the subject languages in Section 4. Section 5 presents the result of the experiment. Discussions and conclusions are presented in Section 6.

3 Related work

We find a lot of work on the use of MTurk in creating speech and language data (Marge et al., 2010b; Lane et al., 2010; Evanini et al., 2010; Callison-Burch and Dredze, 2010). It shows the increasing interests of the research community in the use of MTurk for various NLP domains such as collecting speech corpora as in (McGraw et al., 2010; Lane et al., 2010) and for speech transcription as in (Novotney and Callison-Burch, 2010; Evanini et al., 2010; Marge et al., 2010a)

Among the works, (Novotney and Callison-Burch, 2010) is the most related one to our study. The study investigated the effectiveness of MTurk transcription for training speech models and the quality of MTurk transcription is assessed by comparing the performance of one LVCSR system trained on Turker annotation and another trained on professional transcriptions of the same data set. The authors pointed out that average Turker disagreement to the LDC reference for Korean was 17% (computed at the character level giving Phone Error Rate-PER) and using these transcripts to train an LVCSR system instead of those provided by LDC decreased PER only by 0.8% from 51.3% to 52.1%. The system trained on the entire 27 hours of LDC Korean data obtained 41.2% PER.

Based on these findings, it is concluded that since performance degradation is so small, redundant annotation to improve quality does not worth the cost. Resources are better spent collecting more transcription.

4 Experiment Description

4.1 Languages

Amharic is a member of the Ethio-Semitic languages, which belong to the Semitic branch of the Afroasiatic super family. It is related to Hebrew, Arabic, and Syrian. According to the 1998 census, it is spoken by over 17 million people as a first language and by over 5 million as a second language throughout different regions of Ethiopia. The language is also spoken in other countries such as Egypt, Israel and the United States. Amharic has its own writing system which is syllabary. It is possible to transcribe Amharic speech using either isolated phoneme symbols or concatenated CV (Consonant Vowel) syllabary symbols.

Swahili is a Bantu language often used as a vehicular language in a wide area of East Africa. It is not only the national language of Kenya and Tanzania but also spoken in different parts of Democratic Republic of Congo, Mozambique, Somalia, Uganda, Rwanda and Burundi. Most estimations give over 50 million speakers (with only less than 5 million native speakers). Structurally, Swahili is often considered as an agglutinative language (Marten, 2006). Even if non-total, it has typical Bantu features, such as noun class and agreement systems and complex verbal morphology. It was written with an Arabic-based orthography before it adopted the Roman script (standardized since 1930).

4.2 Corpora

Both Amharic and Swahili audio corpora were collected following the same protocol. Texts were first extracted from news websites and then segmented by sentence. Recordings were made by native speakers reading sentence by sentence with the possibility to re-record anytime they considered having mispronounced. The whole Amharic speech corpus (Abate et al., 2005) contains 20 hours of training speech collected from 100 speakers who read a total of 10850 sentences (28666 tokens). Still in its first steps of development, Swahili corpus corresponds to 3 hours and a half read by 5 speakers (3 male and 2 female). The sentences read by speakers were used as our gold standards to compare with the transcriptions obtained by MTurk. So the transcribed data were already available for control. We recall that the goal

[1] www.mica.edu.vn/sltu-2010/

[2] aflat.org/

of this paper is to evaluate the quality of crowdsoursing tools to obtain good enough transcriptions for resource scarce languages.

4.3 Transcription Task

For our transcription task, we selected from the Swahili corpus all (1183 files) the audio files between 3 and 7 seconds (mean length 4.8 sec and total one hour and a half). The same number of files were selected from the Amharic corpus (mean length 5.9 sec). These files were published (a HIT for a file) on MTurk with a payment rate of USD 0.05 per HIT. To avoid inept Turkers, HIT descriptions and instructions were given in the respective languages (Amharic and Swahili). For the Amharic transcription to be in Unicode encoding, we have given the address of an online Unicode based Amharic virtual keyboard[3] (Swahili transcriptions need no requirement).

5 Results

5.1 Analysis of the Turkers work

On such a small amount of sentences we chose to do the approval process manually via the MTurk web interface. Table 1 shows proportion of approved and rejected HITs for both languages. The higher rate of rejected HITs for Amharic can be explained by the much longer time the task was available for Turkers. We rejected HITs containing empty transcriptions, copy of instructions and descriptions from our HITs, non-sense text and HITs which were made by people who were trying to transcribe without any knowledge of the language. Doing this approval process manually can be considered as time consuming on a large amount of data. However, it was out of the scope of this paper to consider automated filtering/rejecting methods (this is part of future works). With the help of Mturk web interface directly allowing to reject or approve all works made by turkers known to do correct or incorrect work, this approval process took us only a few minutes each day (approximately 15min). Table 2 shows rejected HIT details.

	♯ workers	
	AMH	SWH
APP	12	3
REJ	171	31
TOT	177[4]	34

	♯ HITs	
	AMH	SWH
APP	589 (54.49%)	1183 (82.50%)
REJ	492 (45.51%)	250 (17.43%)
TOT	1081	1434

Table 1: Submitted HITs approval

Content of Rejected HITs	Percentage	
	Swahili	Amharic
Empty	92.86	60.57
Non-sense	3.17	20.33
Copy from instructions	1.98	5.70
Trying without knowledge	1.98	13.40

Table 2: Content of Rejected HITs

Figure 1 shows the detailed completion rate per day for both languages. Among the 1183 sentences requested, Amharic has reached 54% of approved HITs in 73 days. On the other hand, Swahili was completed after 12 days showing a real variety of work rate among different languages.

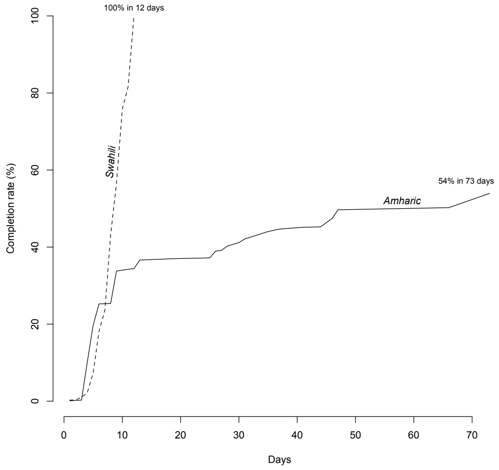

Figure 1: Completion rate per-day

One hypothesis for such a difference could simply be the effective population having access to MTurk. A recent survey (Ipeirotis, 2010) shows that 47% of the turkers were from the United States, 34% from India and the last 19% were divided among 66 non-detailed other countries. However, against this eventual demographic reason, we learn from U.S.ENGLISH[5], that Swahili speakers are less numerous than Amharic speakers in the United States (36690 Swahili speakers against 82070 Amharic speakers).

[3] www.lexilogos.com/keyboard/amharic.htm

[4] This is the number of all the Turkers who submitted one or more Amharic HITs. It is not, therefore, the sum of the number of rejected and approved Turkers because there are Turkers who submitted some rejected HITs and some approved ones

[5] www.usefoundation.org/view/29

Moreover, Table 1 shows that numbers of workers doing coherent work was higher for Amharic than Swahili (12 and 3, respectively). Thus, a more likely reason would be the input burden for Amharic using the external virtual keyboard and copy/paste from another web page. The difficulty to do this while at the same time manage and listen to the audio file may have complicated the task and discouraged Turkers.

Nevertheless, HITs transcription productivity (Figure 2) indicates similar mean Turker productivities (15 and 17xRT for Amharic and Swahili, respectively). Obvious false values brought by some bias in *working time* indicated in MTurk results were removed (lower than 4xRT). Comparing with values in (Novotney and Callison-Burch, 2010), it is much less than historical high quality transcription rate (50xRT), but slightly more than MTurk transcriptions of English (estimated at 12xRT).

5.2 Evaluation of Turkers transcriptions quality

To evaluate Turkers transcriptions (TRK) quality, we computed accuracy against our reference transcriptions (REF). As both Amharic and Swahili are morphologically rich languages, we found relevant to calculate error rate at word-level (WER), syllable-level (SER) and character-level (CER). Besides, real usefulness of such transcriptions must be evaluated in an ASR system (detailed in 5.4). Indeed, some misspellings, differences of segmentation (which can be really frequent in morphologically rich languages) will not necessarily impact system performance but will still inflate WER (Novotney and Callison-Burch, 2010). The CER is less affected and, therefore, it reflects the transcription quality more than the WER. Our reference transcriptions are the sentences read during corpora recordings and they may also have some disagreements with the audio files due to reading errors and are imperfect.

Table 3 presents ER for each language depending on the computed level accuracy[6]. As expected, WER is pretty high (16.0% for Amharic and 27.7% for Swahili) while CER is low enough to approach disagreement among expert transcribers. The word level disagreement for a none agglutinative language ranges 2-4% WER (NIST, web). The gap between WER and SER can be a good indication of the weight of different segmentation errors due to the rich morphology.

	Amharic		
Level	♯ Snt	♯ Unit	ER
Word	584	4988	16.0
Syllable	584	21148	4.8
Character	584	42422	3.3

	Swahili		
Level	♯ Snt	♯ Unit	ER
Word	1179	10998	27.7
Syllable	1179	31233	10.8
Character	1179	63171	6.1

Table 3: Error Rate (ER) of Turkers transcriptions

The low results for Swahili are clarified by giving per-Turker ER. Among the three Turkers who completed approved HITs, two have really similar disagreement with REF, 19.8% and 20.3% WER, 3.8% and 4.6% CER. The last Turker has a 28.5% WER and 6.3% CER but was the most productive and performed 90.2% of the HITs. By looking more closely to error analysis, it is possible to strongly suggest that this Turker is a second-language speaker with no difficulty to listen and transcribe but with some difference in writing to the reference transcription (see details in 5.3).

5.3 Error analysis

Table 4 shows most frequent confusion pairs for Swahili between REF transcriptions and TRK transcriptions. Most of the errors can be grouped into five categories that can also be found in Amharic.

Frq	REF	TKR
15	serikali	serekali
13	kuwa	kwa
12	rais	raisi
11	hao	hawa
11	maiti	maiiti
9	ndio	ndiyo
7	mkazi	mkasi
6	nini	kwanini
6	sababu	kwasababu
6	suala	swala
6	ufisadi	ofisadi
5	dhidi	didi
5	fainali	finali
5	jaji	jadgi

Table 4: Most frequent confusion pairs for Swahili.

[6] Five of the approved Amharic transcriptions and four of the Swahili ones were found to be not usable and were disregarded

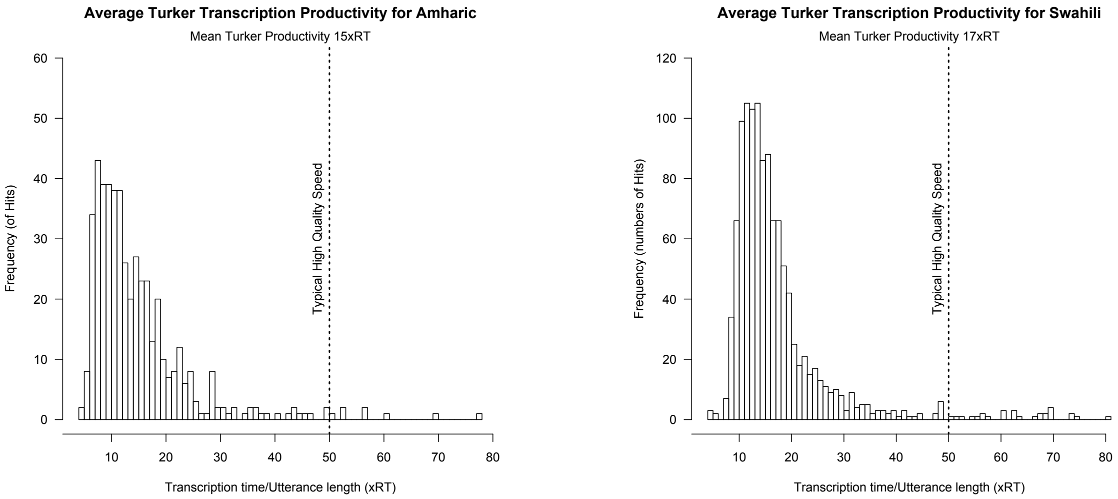

Figure 2: Histogram of HITs transcription productivity

- Wrong morphological segmentations: see words *nini, sababu*, both preceded by *kwa* in REF.
- Common spelling variations of words such as *serikali* and *rais* (sometimes even found in newspapers article); and misspellings due to English influence in loanwords like *fainali* and *jaji* (meaning final and judge).
- Misspellings based on pronunciation (see words *kuwa, ndio, suala*) and due to personal orthographic convention that can be seen in words *maiti, mkazi, ufisadi, dhidi*.

Errors in the last category were all made by the same Turker (the most productive one but having a high WER). Their frequency and regularity are the bases of our strong assumptions to consider this Turker as a second-language speaker. To illustrate this on the phoneme level, the phoneme [z] (voiced alveolar fricative always transcribed 'z' in Swahili) between vowels was always transcribed with an 's' as it is in other languages (like French or German). Similarly, phonemes [θ] and [ð] (dental fricatives transcribed 'th' and 'dh' in Swahili) were never recognized and may not be part of his consonant system.

5.4 Performance in Automatic Speech Recognition (ASR)

Considering the lack of data for Swahili, we used a very preliminary system. Based on a text corpus collected from 7 news websites (over 10 millions words), we built a statistical 3-gram language model using the SRI[7] language model toolkit. Then, to generate a pronunciation dictionary, we extracted 64k more frequent words from the text corpus and automatically created pronunciations taking benefit of the regularity of the grapheme to phoneme conversion in Swahili. For Amharic, we have used the 65k vocabulary and the 3-gram language model that are developed by (Tachbelie et al., 2010).

We used SphinxTrain[8] toolkit from Sphinx project for building Hidden Markov Models based acoustic models (AMs) for both languages. We trained context independent acoustic models of 36 and 40 phones for Swahili and Amharic, respectively. With the respective speech corpora used in the MTurk transcription task, we trained two (for each language) different AMs, one with REF transcriptions and the other using TRK transcriptions.

We computed WER using test sets which contain 82 and 359 utterances for Swahili and for Amharic, respectively. Table 5 presents the WER for both languages.

Languages	ASR	♯ Snt	♯ Wrd	WER
Swahili	REF	82	1380	38.0
	TRK	82	1380	38.5
Amharic	REF	359	4097	40.1
	TRK	359	4097	39.6

Table 5: Performance of ASRs developed using REF and TRK transcriptions

Results indicate nearly similar performances for both languages with a slightly higher WER for the one based on TRK transcriptions (+0.5%) for Swahili and on the opposite direction for Amharic (-0.5%). This suggests, therefore, that non-expert transcriptions using crowdsourcing can be accu-

[7] www.speech.sri.com/projects/srilm/

[8] cmusphinx.sourceforge.net/

rate enough for ASR. Moreover, not only for major languages such as English, languages from developing countries can also be considered. It also highlights the fact that even if most of the transcriptions are made by second-language speakers, it will not particularly affect ASR performances.

6 Discussion and Conclusions

In this study, we have investigated the usability of Amazon's Mechanical Turk speech transcription for the development of acoustic models for two under-resourced African languages. The results of our study shows that we can acquire transcription of audio data with similar quality to a text that can be used to prepare a read speech corpus. However, all languages are not equal in completion rate. The two languages of this study clearly had a lower completion rate than English. And even among the languages of this study, Amharic's task was not completed totally in a period of 73 days.

Thus, MTurk is proved to be a really interesting and efficient tool for NLP domains and some recommended practices were already proposed in (Callison-Burch and Dredze, 2010), mainly on how to be productive with MTurk. However, the use of this powerful tool also happens to be controversial among the research community for legal and ethical issues[9]. As in many fields of research, one should be careful on the manner the data are collected or the experiments are led to prevent any legal or ethical controversies. Indeed, it is often adopted that some charter or agreement need to be signed for any experiments or data collection; which is most of the time totally omitted by the requesters/turkers relationship in MTurk. In order to keep a research close to the highest ethical standards and attenuate these drawbacks, we propose a few guidelines of good conduct while using MTurk for research:

- Systematically explain "who we are", "what we are doing" and "why" in HITs descriptions (as done traditionally for data collection);
- Make the data obtained available for free to the community;
- Set a reasonable payment so that the hourly rate is decent;
- Filter turkers by country of residence to avoid those who consider MTurk as their major source of funding.

[9] http://workshops.elda.org/lislr2010/sites/lislr2010/IMG/pdf/W2-AddaMariani-Presentation.pdf

References

S. T. Abate, W. Menzel, and B. Tafila. 2005. An amharic speech corpus for large vocabulary continuous speech recognition. In *Interspeech*.

E. Barnard, M. Davel, and G. van Huyssteen. 2010a. Speech technology for information access: a south african case study. In *AAAI Symposium on Artificial Intelligence*.

E. Barnard, J. Schalkwyk, C. van Heerden, and P. Moreno. 2010b. Voice search for development. In *Interspeech*.

C. Callison-Burch and M. Dredze. 2010. Creating speech and language data with amazon's mechanical turk. In *NAACL HLT 2010 Workshop*.

K. Evanini, D. Higgins, and K. Zechner. 2010. Using amazon mechanical turk for transcription of non-native speech. In *NAACL HLT 2010 Workshop*.

A. Gruenstein, I. McGraw, and A. Sutherland. 2009. A self-transcribing speech corpus: collecting continuous speech with an online educational game. In *SLaTE Workshop*.

P. Ipeirotis. 2010. Demographics of mechanical turk. *CeDER-10–01 working paper*. New York University.

I. Lane, M. Eck, K. Rottmann, and A. Waibel. 2010. Tools for collecting speech corpora via mechanical-turk. In *NAACL HLT 2010 Workshop*.

M. Marge, S. Banerjee, and A. Rudnicky. 2010a. Using the amazon mechanical turk for transcription of spoken language. In *ICASSP*.

M. Marge, S. Banerjee, and A. Rudnicky. 2010b. Using the amazon mechanical turk to transcribe and annotate meeting speech for extractive summarization. In *NAACL HLT 2010 Workshop*.

L. Marten. 2006. Swahili. In Keith Brown, editor, *The Encyclopedia of Languages and Linguistics, 2nd ed.*, volume 12, pages 304–308. Oxford: Elsevier.

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

I. McGraw, C. Lee, L. Hetherington, S. Seneff, and J. Glass. 2010. Collecting voices from the cloud. In *LREC'10*.

The nist rich transcription evaluation project. *http://www.itl.nist.gov/iad/mig/tests/rt*.

S. Novotney and C. Callison-Burch. 2010. Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *NAACL HLT 2010*.

R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *EMNLP'08*.

M. Y. Tachbelie, S. T. Abate, and W. Menzel. 2010. Morpheme-based automatic speech recognition for a morphologically rich language - amharic. In *SLTU'10*.