# Speaking to the Crowd: looking at past achievements in using crowdsourcing for speech and predicting future challenges

*Gabriel Parent, Maxine Eskenazi*

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania

gparent@cs.cmu.edu, max@cs.cmu.edu

## Abstract

This paper examines the literature on the use of crowdsourcing for speech-related tasks: speech acquisition, transcription and annotation as well as the assessment of speech technology. 29 papers were found, representing, 37 different experiments, which were annotated and analyzed to find trends in the field. The paper focuses on the different techniques used for quality control and the variety of sources of "crowds". Finally, we propose several challenges for the future of crowdsourcing for speech processing.

**Index Terms**: crowdsourcing, human computation, speech acquisition, transcription

## 1. Introduction

In the past few years a new trend has swept through the natural language processing community. Based on the idea of the wisdom of the crowd [1], crowdsourcing is the act of providing a way for non-experts to complete a task that would normally be reserved for experts. The majority of the recent crowdsourcing work has been done via the web, since it allows access to a large population. While some platforms have been designed specifically to allow communication between requesters and workers (e.g., Amazon Mechanical Turk (AMT)), crowdsourcing can also be achieved through games, and other platforms where work is done by volunteers (e.g., Wikipedia).

In speech processing, there are three tasks that seem to be ideal beneficiaries of crowdsourcing. The first, spoken data acquisition, benefits from a larger amount of realistic data that can be collected, with uncontrolled noise level and microphone distance. Another task is transcription/annotation/labeling, providing notation of the linguistic and/or extralinguistic content. Finally, crowdsourcing can assess a technique or a system faster and at lower cost than previously. The quality of the result, as will be seen below, is dependent on the care taken in **task construction** and in **quality assurance**.

This paper reviews the speech crowdsourcing literature to gain a perspective on present achievements and future directions. One goal is to explain a variety of aspects of this tool for those who have not yet used it, providing information about what to do and pointers to where they can get more detailed information. Another goal is to help readers judge a paper, determining whether the data obtained using this technique is of good quality and the scientific results can be relied upon. It is too early to say precisely how this tool will affect the field of speech. The results of its use are just now being put into use. We expect that that answer will come in a few years when some studies are replicated and others are built upon.

Section 2 will review the 29 papers individually, Section 3 will provide a quantitative analysis of trends, types of quality control and sources of crowds and Section 4 will discuss some present issues.

## 2. Background

Many publications have addressed the use of crowdsourcing for language-related tasks. If we limit ourselves to papers that describe at least one experiment using crowdsourcing for speech processing, we find 29 publications through early 2011: theses, journals, books, conference and workshop articles. These papers use the crowd to either acquire or label speech, to assess a speech technology or to conduct a listening study [27].

### 2.1. Speech Acquisition

There had been a bottleneck in real time acquisition, playback and archiving of speech over the web which was solved by the use of Flash and other players which can be embedded, for example, in an AMT "human intelligence task" (HIT). Although there is no way to control microphone type or distance, or the ambient noise, speech gathered in this way is representative of the signal that a speech recognizer processes in a web or phone application. In order to acquire speech in a new language, starting with only one speaker and no screen to read from, [2] has many speakers hear and repeat that person's utterances. Since cellphones can collect speech that is read from the screen, [3] gathers training data from speakers reading words from a mobile phone screen. To acquire speech for a targeted domain, [4] have the speaker dictate a restaurant review of her own invention. [5] used AMT to collect read speech (Figure 1) and had workers complete a predefined dialog scenario, thus obtaining more varied and realistic speech data. [32] created a photo annotation HIT where turkers record a spoken description of photos which, in conjunction with a transcription task, is used to grow a spoken language interface. Others have created interactive games. This often means that they are using their own interface, not AMT. To record speech from non-native speakers while keeping them interested in the task, [6] and [7] use a language learning game. [8] also uses online games to get material for human-robot dialog research. In order to obtain speech for new synthetic voice models, [9] use a quiz game with speakers reading text from the screen.

### 2.2. Speech labeling

Speech, or extralinguistic content, can be annotated using crowdsourcing. At present, more of the literature deals with transcribing speech than with annotation. This trend may change in the near future with increased interest in detecting sentiment and other non-linguistic information. [10] uses an automatic transcription scheme to process very large amounts of speech data from their call center systems. [11] has workers transcribe a large amount of spoken dialog turns from real callers. [12] has workers transcribe meeting data, [13] has them transcribe human-robot dialog speech. Similarly, [26] asks multiple workers to transcribe speech, and use the audio for acoustic model adaptation. [31] integrates processing steps in a transcription task in AMT in order to filter out bad transcripts, and to evaluate word level confidence to provide feedback to

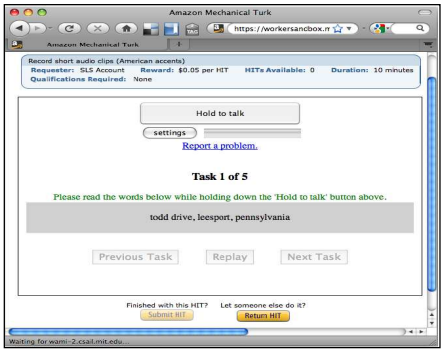


Figure 1 - Example of an AMT task for speech acquisition [5]


Figure 2 - Example of a Facebook transcription game [14]

workers. [14] features a Facebook game where players earn points by transcribing audio (Figure 2). [15] (for Korean, Hindi, Tamil),[16] (for Spanish) and [29] (for Swahili and Amharic) obtain transcriptions of speech for low resource languages. [17] has workers transcribe non-native speech. [24] presents the implementation of a crowdsourcing platform allowing volunteers to correct transcriptions of audio. Finally, [18] looks at audio reCAPTCHAs, acquiring transcriptions of audio streams. For annotation, some authors study the identification of accents. One paper carries out emotion studies. [14]'s Facebook game collects non-native speech and then asks workers to identify the accent in each utterance. [19] asks listeners to identify accents, and they research the influence of the listeners' backgrounds on their annotations. In [25], naïve annotators are asked to provide prosodic annotation of non-native speech. [6] and [11] also use crowdsourcing to label whether a text snippet corresponds to the transcription of a specific utterance, thus accomplishing a task similar to automatic speech recognition (ASR) output validation.

### 2.3. Assessment of speech technology

Only four papers to date use crowdsourcing for assessment. [20] has workers listen to a spoken dialog and then asks them to complete a questionnaire on dialog success. [28] also uses crowdsourcing for dialog system assessment, and concludes that results provided by AMT and Cambridge in-house evaluation are indistinguishable. [21] compared AMT workers and students at the University of Edinburgh on the task of evaluating speech synthesis. [30] report on lessons learned from running 127 crowdsourced speech synthesis preference tests, and provides insight on how to detect cheaters. These pioneering studies lead the way for others to increasingly turn to crowdsourcing to assess the value of a given type of system has from the user's point of view.

## 3. Analysis of the literature

In order to better understand how crowdsourcing has been used for speech processing, we separated each paper into the individual studies it described and attached a type of task to each study. In the 29 papers, there were 37 individual studies (e.g., a paper can describe both an acquisition and a transcription study ([11]). The following section shows some overall statistics and discusses the trends we observed.

### 3.1. Trends in the studies

We see a growing interest in crowdsourcing for speech processing. There were only 4 publications in 2009, the number increased to 14 in 2010. In early 2011, we found 10 papers, 9 of which were submitted at Interspeech11. The present paper was submitted too early to include 2011 ICASSP, ACL, and other venues.

The majority of the 37 indexed studies were on speech labeling and transcription (59%), with speech acquisition being the second most frequent topic (27%). To our knowledge, only 5 studies (14%) have harnessed crowdsourcing for assessment of speech technology. Most of the studies (57%) used the AMT microtask market. While this platform provides access to a large quantity of workers, it can also become relatively expensive, especially for large amounts of data. 7 of the 37 studies (19%) involved a game from which researchers could obtain the players' judgments for free. Other sources of workers include volunteers (14%) and other crowdsourcing platforms, including in-house (11%). Section 3.3 discusses the different sources of workers.

Analysis of the literature shows homogeneity in the geographical source of papers: of the 29 papers, 22 are from the United States. Six are from Europe. There are only two from Asia. This raises questions about future adoption of crowdsourcing outside the US. One cause has been that most studies used the AMT platform (57%), which makes creating a task easier since it not only procures the workers and pays them, but also provides the framework of the interface for the HITS. However, as of the writing of this paper, AMT "Requesters" of HITs must be from the United States. Moreover, in order to be paid in cash, workers need to have a bank account in the US or in India. The alternative for other workers is to receive payment in Amazon gift card credit. Thus, at present, monetary and legal issues may be hampering the adoption of crowdsourcing outside of the US. There are also issues concerning the ethics of the scheme of low payments, social coverage and taxes that are not yet resolved ([22], [29]). If sites outside the US and India cannot easily perform crowdsourcing, then, with the large databases of English speech that are increasingly used by US researchers, it becomes difficult for groups interested in other languages or working on the language of their home country, to have a publication based on relatively smaller amounts of data compete in international venues.

### 3.2. Quality control

Quality control involves a set of measures that is used to oversee and positively influence the quality of the data obtained. It can be present at different stages in the crowdsourcing process. Pre-qualification requirements such as good work history (e.g., how successful a worker was in the past), native language, and even success on a pre-qualification task can be required before the participant is given access to a task. Online filtering can be used during the task to evaluate the quality of the workers' production (e.g., like the gold-unit concept in CrowdFlower [23], see below for more detail). Finally, quality control can be carried out after a worker has submitted all of her work. Table 1 shows the types of quality control in the studies we found. A first surprising observation from Table 1 is that 38% of the experiments did **not** apply **any** quality control. In some cases, this can be explained by the fact that the experiment was a "proof of concept". However, evidence from other papers seems to indicate that quality control is essential in order to obtain reliable data.

Table 1 - Trends in quality control

| Quality Control | # of exp. |
|---|---|
| Before the task | 8 (22%) |
| During the task | 1 (3%) |
| After the task | 18 (47%) |
| None | 14 (38%) |

In the papers that did compare the performance of the crowd to an expert baseline, although the crowd sometimes approached the level of the experts, it never surpassed it. This is true for a classification task (in [5] and [11], where experts show a slightly higher kappa), a transcription task (in [5], [7] and [17] experts have slightly lower word-error rate) or an acquisition task (in [3], experts produce more valid speech utterances). One way to reduce the gap between the crowd and the experts is through quality control. Thus quality control should be central to the design of every task.

The two most popular quality control mechanisms are inter-worker and gold-standard. The main advantage of the former is that an extra set of human interventions is not required, the idea being that most workers are assumed to be correct, and thus the aggregation of workers' judgments provides a good estimation of the true value, which can in turn be used to identify poor workers or cheaters. This unsupervised approach to quality control was, for example, used in [6] and [11] through a voting scheme on a labeling task and in [7], [11], [12], [13], [15], [17] and [26] using string merging algorithms such as ROVER. Inter-worker agreement can be considered to be a sufficient measure, and can be used to measure the quality of individual performance. The biggest issue here is that the majority in the crowd could be wrong, in which case not only is the final solution wrong, but good workers may be penalized. For this reason, various studies have either built new gold-standard datasets ([2], [20], etc.), or reused existing labeled datasets (e.g., from LDC, such as in [15] and [16]). The quality control mechanism introduces an extra cost, but ensures that the workers can be assessed on a valid basis (although even gold-standard datasets can contain errors and ambiguous labels). Finally, although intra-worker ([2]) and peer-reviewed ([20], [31]) control haven't been used extensively, authors note that these mechanisms seem to increase the quality of the data obtained.

While performing quality control after the task is completed is the most frequent option, it seems that waiting this long to filter workers could cause loss of time and money. If inadequate workers can be detected earlier in the process, these losses could be avoided. Only 22% of the studies report doing this by using filters (e.g., AMT success rate ([12] and others) or country of origin ([21] and others)). This low number is surprising given that these mechanisms are natively supported by AMT and can be easily gathered by others. We also haven't found any speech studies where quality control is applied while the workers are performing the task. An example of this type of quality control is the "gold-unit" in CrowdFlower: a requester defines gold-units which are inserted in the task that the workers complete. A worker receives feedback about whether she is completing the task properly (thus reinforcing her understanding of the task). This online quality control also shows the participants that quality matters and that they are being monitored. It may be that the prevalent use of AMT, where it is not easy to automatically insert gold-unit-type checks, is the reason that this has not yet been adopted by the speech community. This kind of feedback could be useful for any type of speech processing task.

A large proportion of the experiments (47%) investigated the use of quality control after the data has been collected (post quality control). These quality control mechanisms can be further divided into 4 categories: intra-worker, where the variability of the work submitted by an individual worker is monitored for discrepancy (e.g., asking the same question twice to the same worker), *inter-worker*, where the variability of the work submitted by multiple workers is monitored for discrepancy (e.g., with a voting scheme, where, for example, all workers agree but one), *gold-standard,* where there are known answers for some of the tasks that the workers complete, and *peer-verified,* where quality control is carried out in a separate crowdsourced task. Table 2 shows the frequency of these types of quality control.

Table 2 - Type of post quality control

| Type of post quality control | # of exp. |
|---|---|
| Intra-worker | 5 (14%) |
| Inter-worker | 9 (22%) |
| Gold-standard | 10 (27%) |
| Peer-reviewed | 3 (8%) |

A frequent question related to quality on a microtask market is whether the payment has an effect on the quality of the work. 5 experiments reported investigating the effect of wage on quality ([5],[13],[15]), and none found a significant difference among the different wage levels. However, [13] noted a significant difference between the latency of a task paying \$0.005 (62 hrs) and the same task paying \$0.05 (13 hrs).

### 3.3. Source of crowds

The experiments used many different sources of crowds. As mentioned previously, AMT has been the most popular platform, but other studies have successfully used games and even volunteers. While the diversity of the tasks does not afford enough data of any one type to enable us to compare the quality of the different sources of crowds, we do observe a difference in the distribution of the number of distinct individual workers on a task on the AMT platform and on the game platforms. (Figure 3)

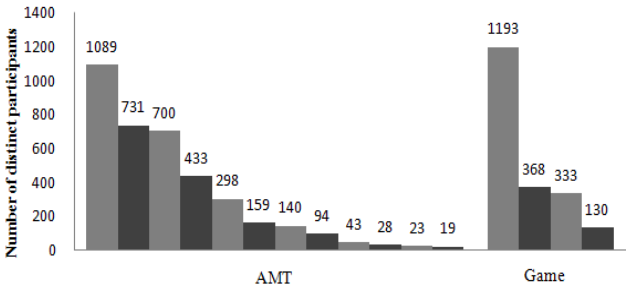


Figure 3 - Distribution of distinct individual workers with AMT and games

Although it would seem that reaching a larger number of workers is easier with AMT because of the financial incentive, the data shows that games may provide a larger variety of workers. This becomes an important issue in a task where a variety of opinions is necessary such as speech acquisition, where a wide sampling of speech affords the construction of more reliable acoustic models.

Another worker issue is native language. Several papers used the AMT country filter to either allow only US workers ([19], [21], etc.) or disallow them ([16] on the Spanish task). In previous work ([11]) we also used the AMT country filter, but found that it was ill-adapted for native language detection for

two reasons: many US residents aren't native speakers of English, and many workers on AMT are registered as US workers, but aren't working from there (as given by the IP addresses we collected). We believe that most crowdsourced speech processing tasks would benefit from having a more effective native language filter. One option would be to assess the participant's native language in a pre-test. While this would decrease throughput, it would improve overall quality.

## 4. Discussion

Crowdsourcing is becoming useful to the speech community. However, there are two issues that may determine whether this resource is actually adopted as a valid tool. One is the quality of the results. A second is the problem of being able to crowdsource for any language.

Anyone can devise a method to crowdsource speech data. Thus, when the results are in, anyone can draw conclusions on that data. In order for crowdsourcing to be accepted by the scientific community, it is essential that each task be accompanied by a valid quality control method. The method should be explained in the publication. Just like methods of assessing the validity of a new algorithm, this will enable others to replicate the work and judge its validity. In general, we suggest that readers judge a paper on the presence of some quality assessment, on the solidity of that assessment and on the quality of their data according to this assessment.

Another issue that we believe to affect the quality of the data produced by crowdsourcing is cognitive load. A task may be simple, such as writing down the word that was heard or repeating what was heard [2]. Or it may be more complex such as writing down the sentence that was pronounced with non-linguistic sounds such as coughing being annotated at the same time as the linguistic content [13]. When creating a task, we believe that researchers should analyze whether it involves some complex set of decisions or conditions that can be divided into separate simpler tasks. In the above example, there should be two passes to annotate the sentence with one pass for the linguistic content and another for the non-linguistic sounds. We believe that eventually the speech community, perhaps in combination with other communities, will devise a flexible measure of the cognitive load in microtasks. This will involve an analysis of the pieces of individual information that are requested in each task as well as of the items and layout used to present the task on the screen. For now we suggest the readers examine the clarity and ease of the task that was given to the workers.

Again, even though a few of the publications we reference deal with languages other than English, there is a growing gap in the speech research community that has been caused by crowdsourcing. There is one group of those who speak English or work on English and also have a US bank account and then there are all the others. While there has been some division in the past as to the amount of data that commercial enterprises dispose of as opposed to academia, now there is a division between the amounts of data that the first group has at their disposal compared to the second group. It is highly desirable that groups of researchers from many countries (this would be a highly expensive enterprise for one country alone to undertake) come together to create a means to crowdsource all other languages.

## 5. Acknowledgements

This work was funded by NSF grant IIS0914927. The opinions expressed in this paper do not necessarily reflect those of NSF.

## 6. References

[1] Surowiecki, J. "The Wisdom of Crowds", Doubleday, 2004
[2] Ledie, J., Odero, B., Minkov, E., Kiss, I., Polifroni, J., "Crowd translator: on building localized speech recognizers through micropayments", in ACM SIGOPS Operating Systems Review, vol 43, 4, New York, NY, 2010.
[3] Lane, I., Waibel, A., Eck, M. and Rottmann, K. "Tools for Collecting Speech Corpora via Mechanical-Turk", Proc. NAACL Workshop: Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, 2010.
[4] Polifroni, J. Kiss, I., Seneff, S., "Speech for Content Creation," Proc. SiMPE, Lisbon, Portugal, 2010.
[5] McGraw, I., Lee, C., Hetherington, L., Seneff, S., Glass, J., "Collecting Voices from the Cloud," Proceedings of LREC, Malta, 2010.
[6] McGraw, I., Gruenstein, A., Sutherland, A., "A Self-Labeling Speech Corpus: Collecting Spoken Words with an Online Educational Game," Proceedings Interspeech, Brighton, UK, 2009.
[7] Gruenstein, A., McGraw, I., Sutherland, A., "A Self-Transcribing Speech Corpus: Collecting Continuous Speech with an Online Educational Game", Proceedings of SLaTE 2009, Brighton, UK, 2009.
[8] Chernova, S., Orkin, J., Breazeal, C., "Crowdsourcing HRI through online multiplayer games", Proc. Dialog with Robots: AAAI fall symposium, 2010.
[9] Freitas, J., Calado, A., Braga, D., Silva, P., Dias, M., "Crowdsourcing platform for large-scale speech data collection", Proc. FALA, Vigo, 2010.
[10] Suendermann, D., Liscombe, J., Pieraccini, R., "How to Drink from a Fire Hose: One Person Can Annoscribe 693 Thousand Utterances in One Month". Proc. SIGDIAL 2010, Tokyo, Japan, 2010.
[11] Parent, G. & Eskenazi, M., "Toward better crowdsourced transcription: transcription of a year of the Let's Go Bus information System data". Proc. IEEE SLT, Berkeley, 2010.
[12] Marge, M., Banerjee, S., Rudnicky, A., "Using the Amazon Mechanical Turk to Transcribe and Annotate Meeting Speech for Extractive Summarization". Proc. NAACL Workshop: Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, 2010.
[13] Marge, M., Banerjee, S., Rudnicky, A., "Using the Amazon Mechanical Turk for Transcription of Spoken Language". Proc. ICASSP, Dallas, 2010.
[14] Akasaka, R., "Foreign Accented Speech Transcription and Accent Recognition Using a Game-based Approach", Masters Thesis, Swarthmore Department of Linguistics, 2009.
[15] Novotney, S., Callison-Burch, C., "Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-Expert Transcription". Proc. NAACL, 2010.
[16] Audhkhasi, K., Georgiou, P., Narayanan, S. "Accurate Transcription of Broadcast News Speech Using Multiple Noisy Transcribers and Unsupervised Reliability Metrics", Proc. ICASSP, Prague, 2011.
[17] Evanini, K., Higgins, D., Zechner, K., "Using Amazon Mechanical Turk for transcription of non-native speech", Proc. NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010.
[18] Schlaikjer, A., "A dual-use speech CAPTCHA: Aiding visually impaired web users while providing transcriptions of Audio Streams", LTI-CMU Technical Report 07-014, 2007.
[19] Kunath, S.A., and Weinberger, S.H., "The wisdom of the crowd's ear: speech accent rating and annotation with Amazon Mechanical Turk", Proc. NAACL HLT Workshop: Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, 2010.
[20] Yang, Z., Li, B., Zhu,Y., King, I., Levow, G. and Meng , H., "Collection of User Judgments on Spoken Dialog System with Crowdsourcing", Proc. SLT 2010.
[21] Wolters, M., Isaac, K., Renals, S. "Evaluating Speech Synthesis Intelligibility using Amazon Mechanical Turk", In Proc. 7th Speech Synthesis Workshop (SSW7), 2010.
[22] Adda, G., Mariani, J., "Language resources & Amazon Mechanical Turk ethical, legal and other issues", In Proc. LISLR2010, LREC2010.
[23] CrowdFlower, http://www.crowdflower.com
[24] Goto, M., Ogata, J., "PodCastle: Recent Advances of a Spoken Document Retrieval Service Improved by Anonymous User Contributions", Proc Interspeech2011, Florence, 2011.
[25] Evanini, K., Zechner, K., "Using crowdsourcing to provide prosodic annotations for non-native speech", Proc. Interspeech2011, Florence, 2011.
[26] Audhkasi, K., Georgiou, P., Narayanan, S., "Reliability-weighted acoustic model adaptation using crowd sourced transcriptions", Proc. Interspeech2011, Florence, 2011.
[27] Cooke, M., Barker, J., Lecumberri, ML, Wasilewski, K., "Crowdsourcing for word recognition in noise", Proc, Interspeech2011, Florence, 2011.
[28] Jurcicek, F., Keizer, S., Gasic, M., Mairesse, F., Thomson, B., Yu, K., Young, S., "Real User evaluation of spoken dialogue systems using Amazon Mechanical Turk", Proc, Interspeech2011, Florence, 2011.
[29] Gelas, H., Abate, ST, Besacier, L., Pellegrino, F., "Quality assessment of crowdsourcing transcriptions for African languages", Proc Interspeech2011, Florence, 2011.
[30] Bucholz, S., Latorre, J., "Crowdsourcing preference tests and how to detect cheating", Proc. Interspeech2011, Florence, 2011.
[31] Lee, C., Glass, J., "A Transcription Task for Crowdsourcing with Automatic Quality Control", Proc. Interspeech2011, Florence, 2011.
[32] McGraw, I., Glass, J., Seneff, S., "Growing a Spoken Language Interface on Amazon Mechanical Turk", Proc. Interspeech2011, Florence, 2011.