Crowd Translator: On Building Localized Speech Recognizers through Micropayments

Jonathan Ledlie
Nokia Research
Cambridge, MA, USA
jonathan.ledlie@nokia.com

Billy Odero
Nokia Research
Nairobi, Kenya
billy.odero@nokia.com

Einat Minkov
Nokia Research
Cambridge, MA, USA
einat.minkov@nokia.com

Imre Kiss
Nokia Research
Cambridge, MA, USA
imre.1.kiss@nokia.com

Joseph Polifroni
Nokia Research
Cambridge, MA, USA
joseph.polifroni@nokia.com

ABSTRACT

We present a method to expand the number of languages covered by simple speech recognizers. Enabling speech recognition in users' primary languages greatly extends the types of mobile-phone-based applications available to people in developing regions. We describe how we expand language corpora through user-supplied speech contributions, how we quickly evaluate each contribution, and how we pay contributors for their work.

Categories and Subject Descriptors

C.2.4 [**Computer Communication Networks**]: Distributed Systems—*Distributed Applications*; H.5.3 [**Information Interfaces and Presentation**]: Groups and Organization Interfaces—*Collaborative Computing*

General Terms

Algorithms, Experiments, Human Factors, Measurement

Keywords

Speech Recognition, Crowd-Sourcing, Self-Verification

1. INTRODUCTION

Speech is a key modality for mobile devices and applications, particularly for low-literate users in developing regions. Even for simple tasks, speech-based interaction requires a speech recognizer trained in the target phrases in the user's language. Because creating a speech recognizer for a new language is an expensive and time-consuming task, recognizers exist for only the most popular languages. Many of the users who would most benefit from a speech-based interface are often forced to speak in a secondary language, such as English, or use an alternative modality. A fundamental component of any speech interface is reliable data with which to build acoustic and language models. A critical prerequisite to deploying speech interfaces in the developing world is these data.

In order to ensure reliable and robust performance in a speaker-independent task, large amounts of spoken data must be collected from as broad a range of people as possible. We demonstrate a system, Crowd Translator (CX), that gathers speech data from people through their mobile phones to create a high-quality speech recognizer. After we automatically validate each set of contributions, we pay into the contributor's mobile phone bank account.

CX aims to make it easy and cheap to develop simple speech recognizers in many more languages than are currently available. This paper describes how we acquire speech data in new languages and how we automatically verify and pay the people who contribute to each language's corpus of recognized phrases.

This paper makes the following contributions:

- We show how user-generated content can be subject to self-verification, quickly classifying it as useful or not.
- We demonstrate a prototype that automatically screens out invalid user data, suggesting that CX can be used to create simple speech recognizers for local languages at significantly lower cost than previous methods.

1.1 How Crowd Translator Works

We start with a target corpus of text phrases chosen as likely to conform to a set of simple telephone-based applications. A single person with an unmarked accent, our "voice talent", records each phrase to be used as an audio prompt. This same voice talent also reads a short introduction that instructs people on how to use the system. When subjects call the system, they hear the instructions, which ask them to repeat each audio prompt after they hear it.

Each contributor, or *worker*, is recruited from among native speakers of the target language. Immediately after each worker's contribution, or *session*, CX automatically determines if the recordings the worker provided are valid. If the session is valid, the worker is sent a payment, either to a mobile bank account or to a mobile minutes account, depending on what is available in the particular country. After sufficiently many users have contributed many sessions for the

(a) Make Canonical Recordings

English	Swahili	Gold Std. Utterance
car	gari	"gari"
boat	mashua	"mashua"
plane	ndege	"ndege"
...	...	"..."

(b) Gather User Input

Prompt	User$_1$ Utterance
$gari_g$	$gari_1$
$ndege_g$	$ndege_1$
$mashua_g$	$mashua_1$
$\dots_g$	$\dots_1$
$mashua_g$	$mashua'_1$
$\dots_g$	$\dots_1$
$gari_g$	$gari'_1$

(c) Verify Input

Intra-session Agreement?
$gari_1 \approx gari'_1$
$mashua_1 \approx mashua'_1$

(d) Expand Corpus

Word	Utterance
car	$gari_g$
car	$\underline{gari_1}$
car	$\underline{gari'_1}$
car	$gari_2$
...	...
boat	$mashua_g$
boat	$\underline{mashua_1}$
...	...

Figure 1: Crowd Translator Overview: (a) In a lab setting, a single native speaker translates target utterances and records canonical, "gold standard" phrases, *e.g.,* "gari" (gari$_g$). (b) A worker, who speaks the target language natively, calls the Crowd Translator phone number and provides input, mimicking each prompt. Each set of utterances provided by one worker in one phone call is called a *session.* (c) CX quickly determines the validity of the session using a new technique, testing for *intra-session agreement.* The worker is paid within a few seconds of completing the task if it contains a sufficient fraction of valid input. (d) Lastly, we add the user's input to the overall speech corpus for the target language. Slower automatic post-processing further filters the corpus.

same corpus, a corpus is created to be used to train a speech recognizer. Note that we are not building a phoneme-aware model of each language. Instead, we are collecting many samples of each utterance, which can directly be used for statistical matching. Figure 1 illustrates this process.

2. BACKGROUND

2.1 Crowdsourcing Data

Amazon's Mechanical Turk is a "marketplace for work" [1]. Through its web interface, users can submit short tasks they want other people to complete. These tasks are typically ones that humans are good at but machines are not: for example, determining whether a video is suitable for children. Mechanical Turk takes care of allocating tasks and routing payments to workers, but leaves determining the validity of each worker's results to the user who submitted the task. Crowd Translator could be thought of as a specialized instance of Mechanical Turk's marketplace, one which builds in verification. While currently focused on collecting and automatically verifying speech data, CX could expand to include other speech-driven tasks, such as sentence simplification and annotation of spoken utterances.

Like Mechanical Turk, txteagle is an outsourcing marketplace [7]. Unlike Mechanical Turk, workers are drawn from developing regions and sent tasks via SMS. Current tasks focus on text translation, although image recognition tasks may be supported in the future (using MMS).

The success of txteagle's trials, in Kenya, Rwanda, Nigeria, and the Dominican Republic, have varied widely depending on levels of trust. In Nigeria, for example, where SMS-based scams are common, workers have been hesitant to use txteagle, because they are unsure they will really be paid on successfully completing a task. As we expand CX beyond a prototype, we will likely encounter similar trust issues.

While txteagle and Crowd Translator employ some similar techniques to determine worker and result credibility, txteagle does not use sessions to establish user credibility. In addition, users are paid per individual task, not per set of tasks. Because establishing the correctness of a task may occur hours or days after a worker has provided input, trust may be greater in a system like CX that provides immediate feedback.

2.2 Data Annotation

Machine learning researchers quickly realized that Mechanical Turk was a new resource for collecting large quantities of human-labeled data. An evaluation of a machine learning algorithm typically requires training and testing data, where both sets of data have been labeled, or annotated, by a human. For example, after humans assign categories (labels) to large numbers of videos, researchers can test their algorithms to see if this association can be made automatically.

Machine learning researchers also quickly realized that labels generated with Mechanical Turk were different from "expert" annotations. Snow *et al.* asked Mechanical Turk's workers to assign emotions to news headlines — for example, assigning "surprised" to the phrase "Outcry at North Korea Nuclear Test" [19] — and found the results were noisy. Snow *et al.* and Kitter *et al.* found that Turk's human annotators sometimes cheat or make mistakes, producing incorrect results when compared to an expert [11]. To correct for these errors, researchers used spot checks, cross-validation, and worker filtering: (a) *spot checks* compare the purported result against a known truth, or "gold standard;" (b) *cross-validation*, also called inter-annotator agreement, compares the results from different users for the same task [18] and (c) *worker filtering* assigns tasks to people who have performed the same types of tasks well in the past [6]. Combined, these approaches facilitate *verification*, the process by which utterances are vetted and transcriptions authenticated. Interestingly, these methods for verifying human input are essentially the same as those developed a decade earlier to catch cheaters in volunteer computing projects, such as SETI@Home [2].

Without human intervention, these techniques greatly improve the reliability of the labeled data, sometimes leading to expert level performance [18]. However, they come at the cost of redundancy, reducing throughput and increasing cost per task.

Another powerful technique is to measure individual consistency. Kruijff-Korbayová *et al.* examined *intra*-annotator agreement, where the same person was presented with the same question two or more times [12]. If the person gave the same answer, he or she was considered more reliable; if not, less so. The results were verified by a human expert, an especially tedious task with speech data.

Crowd Translator uses a new type of intra-annotator agreement to measure input validity. Instead of spreading redundant queries over months (as in [12]), we ask workers the same question within the same session. This is feasible because our questions (voice prompts) are very brief relative to the length of the session. With multiple data points gathered in the same short period, we can estimate the validity of the user input immediately after each session (or even during the session). This allows us to provide the worker with immediate feedback on his or her work. This technique of prompting users with a small number of redundant, very short tasks appears to generalize to verifying other forms of user-provided data.

2.3 Speech Data Collection

Large-scale telephone speech data collection has been used for more than a decade to bootstrap development of new domains [4]. Broad-based recruitment efforts have been effective in targeting large numbers of subjects in a short period of time [5, 9]. These efforts involved non-spontaneous speech, with subjects either reading from prepared prompt sheets or responding to scripted or situational prompts.

Non-spontaneous data collection makes verification easier, but can have an unnatural priming effect. A human verifier, or *annotator*, can be given the anticipated transcription which, in the majority of cases, is what the user has actually said. Annotators need some level of training, and typically both hardware and software is required for playing, displaying, and manipulating annotations: they add greatly to the cost of collection.

In contrast to these highly manual approaches, Crowd Translator provides a method for collecting and verifying large amounts of speech data in "low resource" languages, *i.e.,* languages for which there are few, if any, existing corpora, either written or spoken. Other research has examined the issue of speech-to-speech translation in low resource languages [10, 15]. In these cases, some existing data were available, as well as resources for transcribing parts of it. This differs from CX in that we are concerned with the rapid development of simple applications to be used in cases where neither corpora nor resources for annotation are available.

A different approach to large-scale telephone speech data collection is GOOG411 [8]. GOOG411 is a public telephone directory service, similar to 411 in the U.S. In contrast to most directory services, the user speaks the name of the desired listing — no human operator is ever involved — and Google pays for the call, not the user. While Google loses money on this service directly, the recorded utterances, in aggregate, improve speech recognition for its other applications and services. Both GOOG411 and Crowd Translator pay users for their time, although GOOG411 does so indirectly.

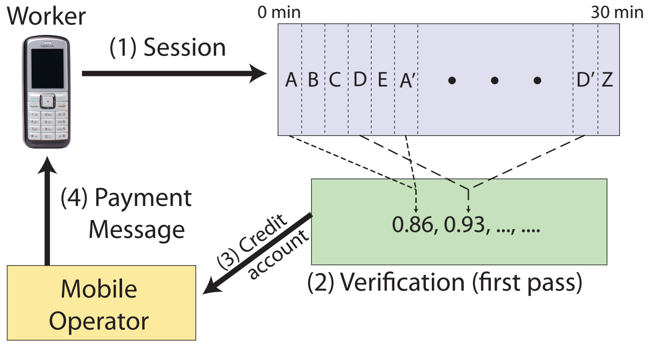

Figure 2: During a session, each worker is presented with a small number of redundant utterances *e.g., A* and A'. After each session, the verification process can quickly assign an overall validity to the session. If the session passes verification, the user's mobile payment account is credited, by sending a signed message to the mobile operator. Within a few seconds, the user sees his or her account has been credited, increasing trust in the system.

GOOG411 collects spontaneous speech; users can ask for any listing. This type of speech often more closely matches normal speech than mimicked, non-spontaneous speech, but it is much harder to annotate because there is no anticipated transcription. Because GOOG411 is single-shot, there is little opportunity to automatically decide if a user has provided useful input for a speech recognizer. Thus, screening out poor input may be harder. In addition, with GOOG411, it is more difficult to infer if the result presented to the user is correct: even if the user connects to the top result, this may not have been what the user originally asked for.

Our phrase-based recognizer is in contrast to the phoneme-based recognizers that have come to dominate the field in the past fifteen years. Phoneme-based recognizers are, in many respects, superior to non-phoneme ones. Because they match only subcomponents of words, they are easier to expand to new words, they can be easier to tune to new dialects, and they use less memory. Unfortunately, they are extremely expensive to build: so expensive that only 28 languages are available from the market leader [16]. Phrase-based recognizers have been shown to provide very accurate results ($\approx$ 100%) for speaker-independent matching for small vocabularies [13]. Interestingly, because this approach was dropped in favor of phoneme-based recognizers, it is an open problem to examine how this approach scales to hundreds or thousands of phrases. But because our target applications have only a few dozen possibilities at each user prompt and because acquiring new phrases is inexpensive, full-phrase recognition appears sufficient.

3. VERIFYING USER INPUT

Crowd Translator uses a two pass model to construct a speech corpus. The first pass, illustrated in Figure 2, is a verification step whose goal is to determine, once a working session is over, whether the worker is likely to have supplied a sufficient fraction of valid input and should be paid. Invalid inputs correspond to words or phrases that are different than requested. This may occur due to workers' attempts to receive an award without performing the task as instructed, as well as due to issues of line quality and workers' misinterpretation of the word spoken. We are interested in a model

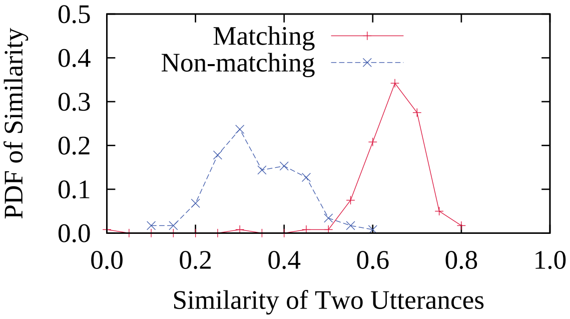

Figure 3: When workers provide different content than is expected, these invalid inputs will match the "disagreement" distribution, not the "agreement" one. The figure shows two distributions: "agreement" is the similar metric when users have said the same word twice; "disagreement" is when they have said different words. This same method of testing for the similarity of a few user-supplied values can be applied to other contexts where users supply the bulk of the system's data. The data come from five people whose inputs were verified and then artificially scrambled to create mismatches.

where the workers are rewarded according to their effort. A good model should discard low quality sessions, but it should allow moderately noisy inputs to establish worker trust.

The goal of a second pass is to further eliminate irrelevant samples from the set of approved sessions, such that the quality of the speech recognizer trained on the accepted samples is high. The focus of this paper is on approaches to the first pass validation, where we leave the theoretical and empirical study of the second pass for future work.

We next describe two approaches to evaluating session validity: (a) intra-session agreement and (b) comparing against gold standard samples. The results of an empirical evaluation of the two approaches are given in the following section.

3.1 Intra-Session Agreement

We suggest intra-session agreement as a method to rapidly determine the likely validity of a session of user data. The intra-session agreement measure is aimed at discovering whether a user was consistent throughout the session. It is assumed that if the user did not or partially followed the instructions given to him, overall consistency would be low.

The three steps to testing intra-session agreement are:

1. Make a small fraction of the user's queries redundant.
2. Measure the similarity between each query pair.
3. The distribution of the output similarity scores is compared against a known distribution of truly similar pairs of the same type of data.

We evaluate step (2) as the similarity between utterance pairs using an acoustic similarity measure. However, this model is general: different similarity measures can be used according to the subject domain. For example, in a sentiment labeling scenario, the user-supplied labels could be compared for semantic similarity ("surprised" and "shocked"). In an indoor localization scenario, it could be used to test the quality of user-supplied radio-frequency fingerprints [20].

Figure 3 shows the distribution we used for step (3). The comparison can be performed simply by contrasting the median of the output distribution against a threshold, set according to the reference distribution. An alterinative, more robust, method is to apply a standard statistical test, such as the Student's t-test, which states whether it is likely that the output sample was generated from the reference distribution.

The outcome of the intra-session agreement test is that the user-supplied content is determined to be likely valid or likely invalid. This approach does not track low-quality or erroneous samples as long as the user is consistent. In particular, if a user has malicious intentions, this type of test is relatively easy to falsify, but this problem can be overcome by combining intra-session agreement testing with other approaches.

In addition to determining whether a specific session was valid, we consider a user's credibility score as in Sarmenta [17]. We let a user's credibility vary over time with an exponentially-weighted moving average, such that the user's recent historical information can be integrated into each session evaluation. We then send a session on to the second pass when the overall credibility score is above a threshold. This allows users to be paid even when they have provided invalid data in the past, but for CX to remain cautious about this user.

3.2 Gold Standard Comparisons

Another approach is to compare a sample of user annotations with gold standard data, generated by experts. In our framework, a sample of the utterances provided in the session may be compared against gold standard pronunciation. This forms a tighter control over the session's quality, as beyond consistency, the produced utterances are evaluated directly against a reference acoustic signal. However, we conjecture that this approach may be suboptimal for the speech domain, as it makes a strong assumption that there is high similarity between the user's pronunciation and the gold standard. This assumption may be false if the worker uses a particular dialect that is not represented by the gold standard. While it is in our interest to discard irrelevant samples, valid outliers due to dialects are important for system coverage. Snow *et al.* and others have also based tests on a comparison to a gold standard [19].

As we build up utterances that come from users with high credibility, we can construct an acoustical model of each word that is based on multiple inputs, of both experts and credible users. Such model will form a broader standard that can be used as with the previous approach, but yielding higher coverage. To make this validation step quick, we would only pick a small random sample of the user's utterances. If both this test and the intra-session agreement test succeed, the session will be accepted. We plan to implement this combined approach in the future, once the size of our user base scales up.

4. PROTOTYPE

We implemented a prototype of Crowd Translator and used it to collect a small sample of English and Swahili utterances in Kenya. Our target corpus of a few hundred words uses the vocabulary of a speech-based classifieds application we are deploying in East Africa [14]. We recruited fifteen local workers overall, where each working session was comprised of 110 prompts. For evaluation purposes, half of the

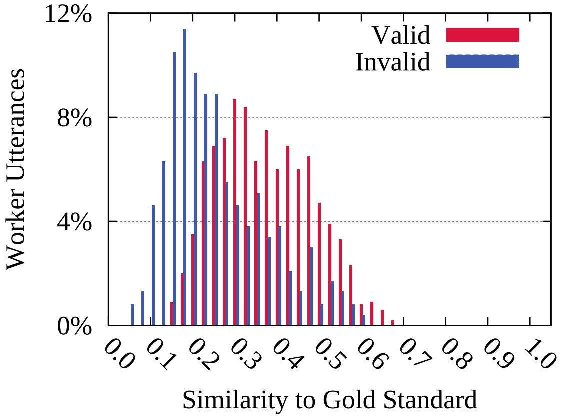

Figure 4: Screening out invalid data is difficult if only comparing user-supplied input to a known gold standard. When we manually labeled data as valid or invalid and compared it to the gold standard utterance, we found that much of the invalid data had a high similarity score. Using a threshold to determine input validity using this method seems unlikely to give solid results.

prompts in each session are duplicates; in the experiments, however, as well as in a real deployment, only a small number of duplicates will be used.

Because of low levels of literacy among our target population, we do not use prompt lists, even if they could be sent via SMS. Our subjects hear audio prompts and are asked to repeat what they hear. While we realize the priming effect this may have on the data we collect, we made the choice to use audio prompts based on the average literacy level of our intended subjects and the nature of the data being collected (i.e., the words represent abstract concepts such as "repeat" or "next" and cannot be rendered through pictograms).

Workers used mobile phones to call into the prototype, which was built using Asterisk [3]. While a programmable API to credit accounts exists, we have not yet linked our back-end to it. (This, however, does not appear to be a significant technical hurdle.)

We used the data collected using the prototype to evaluate intra-session agreement and learned that several aspects of CX need to be altered before the next phase of the project. Overall, based on manual validation, the data contained 1229 valid utterances and 421 invalid ones. Invalid data occurred because users misheard the prompts, did not know when to speak (before or after a beep), or simply said nothing. A proposed alteration of the prototype is therefore adding a training phase to improve user success rates.

4.1 Experiments

Our main goal in the experiments is to evaluate automatic methods for validating user inputs. We first evaluate the common approach of comparing the input utterances against gold standard samples. Figure 4 shows that there may be little acoustic resemblance between two valid samples of the same word. This confirms our conjecture that gold standard samples may not account for the variety of personal pronunciations and accents in the speech domain (see Section 3). While an expanded method — where the sample is compared to a set comprising of the gold standard as well as other presumed valid samples — may improve validation results, comparing only to the gold standard is not recommended for bootstrapping a corpus in these settings.

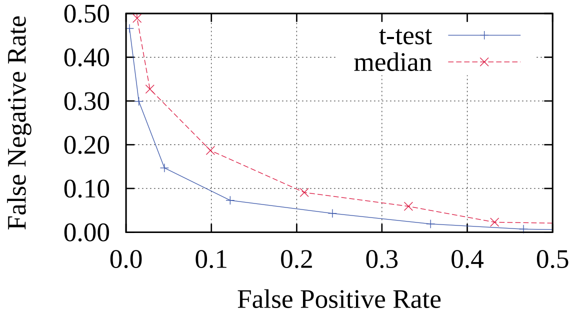

Figure 5: Comparing to a known distribution of valid utterances using a simple statistical test, such as the t-test, gives moderately better performance over a simple median threshold. In particular, we do not want high false negative rates because this would lessen trust in the system. The figure shows the false positive and false negative rates as we vary the median threshold and the t-test distributions sampled from, respectively.

The second measure that we evaluate in our experiments is the intra-session agreement. To examine intra-session agreement, we used our acoustical analyzer to compute a similarity score for 16 duplicate pairs per every session. To estimate a session's validity, we sampled from this distribution and either (a) compared the sample's median to a threshold or (b) determined whether the sample matched the known valid intra-session distribution from Figure 3.

Figure 5 displays a receiver operating curve (ROC), showing the false positive and false negative rates as we vary the acceptance threshold (for the median comparison) and the ratio of valid-to-invalid distributions (for the t-test comparison). In the figure, a curve closer to the origins is preferable. That is, Figure 5 confirms that a statistical comparison of the sample distributions performs better than considering the sample's median.

Several considerations may guide the selection of the specific threshold or valid-to-invalid ratio to be applied. In particular, we consider trust as a key element of Crowd Translator. We would like to ensure that people who have supplied it with primarily valid data will be rewarded. A lack of trust will prevent the users from recommending Crowd Translator to others. Thus, we would prefer a low false negative rate (*i.e.,* have only few high-quality sessions rejected) to a low false positive one (*i.e.,* allowing lower accuracy among the set of accepted sessions), when automatically determining session validity. After letting some invalid data through the first pass, we can clean it more carefully using slower, more precise techniques, lowering the false positive rate with the second pass.

By varying the valid-to-invalid ratio, we can model the effect of a user whose utterances are mostly but not exclusively valid. In our experiments, we assumed a session was valid if 80% of its utterances were valid. We compare the result of our testing mechanism against the correct decisions, based on the manual evaluation of the samples. Overall, we observed that, by choosing a moderately lenient threshold, the automatic evaluation can yield a low ($< 5\%$) false negative rate per session, while limiting the ratio of invalid sessions accepted in the first pass to between $25 - 35\%$. This result should satisfy the human factor; a second pass that further filters noisy data may be required for training high-quality speech recognizers.

4.2 Discussion

In addition to including a training phase, we believe that three other changes should improve our prototype. First, instead of a single voice model, we should have at least one male and one female model. While this may not affect intra-session agreement, we found that it does improve second pass verification in preliminary tests. Broadening the selection of voice talents further — including non-standard accents, for example — should also reduce the priming effects on the collected corpus. Second, once a "starter" corpus of valid utterances for a phrase has been collected, new utterances of this phrase can be cross-validated against them. This is broader than simply comparing against the gold standard. Like intra-session agreement, this test can be done before paying the user. By rejecting sessions where a significant fraction of the utterances fail this test, we can avoid accepting sessions where, for example, the user repeats exactly the same utterance in response to each prompt. Third, some volunteers were extremely reticent to have themselves recorded and required lengthy reassurance that no one outside of our research group would listen to their recordings ("I don't like the sound of my voice" said an early participant from Tanzania). In addition to developing trust on payment, participants must develop trust that their recordings will be kept private.

5. CONCLUSION

We presented a method that allows to expand the number of languages covered by speech recognizers, enabling new applications in developing regions. We focused on a method for automatically validating sessions of user-generated content, a new form of intra-annotator agreement. Self-validating content is useful in contexts where large corpora of human-generated data are required. Instead of having a small number of people painstakingly generate corpora, we showed how many people could be used to build them.

A promising direction to pursue in terms of user verification and data quality is to apply learning techniques such as clustering to find trends in the data. Suppose that inter-user similarity is evaluated, using duplicate samples across multiple users; given user similarity scores, users may be clustered into cohesive groups (for example, using methods such as agglomerative clustering). It is reasonable that the groups formed will represent different dialects. The association of a worker to a particular dialect profile is very valuable, as it may enable control over the distribution of samples selected to train the speech recognizer. In addition, users that are not found to be tightly related to one of the groups formed in clustering will be considered low-quality (noisy) workers.

In the future, we aim to deploy Crowd Translator in several countries in East Africa. We hope to have created recognizers for at least five new languages in the next year and make these recognizers available for phone-accessible and on-device applications.

Acknowledgements

We thank: Elishibah Msengeti for being our voice model; Nokia Research Center Africa in Nairobi, Kenya for testing the CX prototype; and Prof. Tavneet Kaur Suri for orchestrating the data collection process in the field.

6. REFERENCES

[1] Amazon Mechanical Turk. `http://mturk.com`.

[2] D. P. Anderson, J. Cobb, E. Korpela, M. Lebofsky, and D. Werthimer. Seti@home. *Communications of the ACM*, 45(11):56–61, 2002.

[3] Asterisk. `http://asterisk.org`.

[4] J. Bernstein, K. Taussig, et al. MACROPHONE: An American English Telephone Speech Corpus for the Polyphone Project. In *ICASSP*, Apr. 1994.

[5] R. Cole, M. Fanty, et al. Telephone speech corpus development at CSLU. In *ICSLP*, 1994.

[6] P. Donmez, J. G. Carbonell, and J. Schneider. Efficiently Learning the Accuracy of Labeling Sources for Selective Sampling. In *KDD*, 2009.

[7] N. Eagle. txteagle: Mobile Crowdsourcing. In *HCII*, July 2009.

[8] GOOG-411. `http://www.google.com/goog411`.

[9] E. Hurley, J. Polifroni, and J. Glass. Telephone data collection using the world wide web. In *ICSLP*, 1996.

[10] A. Kathol, K. Precoda, D. Vergyri, W. Wang, and S. Riehemann. Speech Translation for Low-Resource Languages: The Case of Pashto. In *Interspeech*, 2005.

[11] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. In *CHI*, Apr. 2008.

[12] I. Kruijff-Korbayová, K. Chvátalová, and O. Postolache. Annotation Guidelines for Czech-English Word Alignment. In *LREC*, 2006.

[13] K. Laurila and P. Haavisto. Name dialing: How useful is it? In *ICASSP*, 2000.

[14] J. Ledlie, N. Eagle, M. Tierney, M. Adler, H. Hansen, and J. Hicks. Mosoko: a Mobile Marketplace for Developing Regions. In *DIS*, Feb. 2008.

[15] S. Narayanan et al. Speech Recognition Engineering Issues in Speech to Speech Translation System Design for Low Resource Languages and Domains. In *ICASSP*, 2006.

[16] Nuance: OpenSpeech Recognizer. `http://nuance.com`.

[17] L. Sarmenta. Sabotage-Tolerance Mechanisms for Volunteer Computing Systems. In *CCGRID*, May 2001.

[18] V. S. Sheng, F. Provost, et al. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *KDD*, Aug. 2008.

[19] 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 *EMNLP*, Oct. 2008.

[20] S. Teller, J. Battat, B. Charrow, D. Curtis, R. Ryan, J. Ledlie, and J. Hicks. Organic Indoor Location Discovery. Tech. Report CSAIL TR-2008-075, MIT, Dec. 2008.