A Self-Transcribing Speech Corpus: Collecting Continuous Speech with an Online Educational Game

Alexander Gruenstein[1]*, Ian McGraw*[1]*, Andrew Sutherland*[1,2]

[1]MIT Computer Science and Artificial Intelligence Lab, Cambridge, MA, USA
[2]Quizlet.com, Albany, CA, USA

alexgru@mit.edu, imcgraw@mit.edu, asuth@mit.edu

Abstract

We describe a novel approach to collecting orthographically transcribed continuous speech data through the use of an online educational game called Voice Scatter, in which players study flashcards by using speech to match terms with their definitions. We analyze a corpus of 30,938 utterances, totaling 27.63 hours of speech, collected during the first 22 days that Voice Scatter was publicly available. Though each individual game covers only a small vocabulary, in aggregate speech recognition hypotheses in the corpus contain 21,758 distinct words. We show that Amazon Mechanical Turk can be used to orthographically transcribe utterances in the corpus quickly and cheaply, with near-expert accuracy. Moreover, we present a filtering technique that automatically identifies a sub-corpus of 39% of the data for which recognition hypotheses can be considered human-quality transcripts. We demonstrate the usefulness of such self-transcribed data for acoustic model adaptation.

1. Introduction

In this paper, we present a new approach to collecting and orthographically transcribing a significant amount of continuous speech via an online educational game called Voice Scatter. Voice Scatter uses speech recognition to provide a fun way for users to review flashcards by speaking aloud terms and their definitions. We have recently made the game publicly available on the website Quizlet.com, and have collected 30,938 utterances, constituting 27.63 hours of speech, over a 22 day period. Each individual game uses only eight flashcards, and thus speech recognition can be performed using a narrow-domain, strict grammar. However, an estimated 1,193 speakers played the game with 1,275 distinct flashcard sets, so recognition hypotheses in the corpus cover 21,758 distinct words.

Speech recognition errors do, of course, occur. However, in this paper we explore filtering techniques to identify high quality recognition hypotheses. The best technique pairs confidence scores from narrow-domain speech recognition with information from the game context about whether a hypothesis represents a correct answer. In this way, we automatically identify a sub-corpus of 39% of the data for which recognition hypotheses can be taken to be human-quality orthographic transcripts. We establish human agreement levels, and obtain manual transcripts of a 10,000 utterance development set, by "crowdsourcing" the transcription task via Amazon Mechanical Turk[1]. When compared to a 1,000 utterance subset transcribed by experts, the crowdsourced transcripts show near expert-level agreement.

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

Voice Scatter exemplifies a paradigm of collecting and (automatically) transcribing significant amounts of speech via games that have four key properties. First, a game should be easy for a large number of users to access, *e.g.* via the Web or telephone. Second, the game must be attractive to users in its own right, ideally providing some kind of benefit to its users. Third, speech recognition for game play should require only a small vocabulary, narrow-domain language model; yet, a variety of datasets should be available so that a diverse vocabulary is covered in aggregate. Fourth, while not required, contextual constraints such as whether the recognition hypothesis makes sense in context, and whether it is a "good move", may often be helpful to filter data. Games with these four properties are extremely valuable, as they both benefit their players and have the potential to provide automatically transcribed speech data.

2. Related Work

In [1], we present a related online speech flashcard game called Voice Race, which elicits isolated spoken words that can be automatically transcribed using game constraints. Voice Scatter extends this technique to continuous speech, and introduces confidence scores as an additional filter.

Voice Scatter is similar to so-called "games with a purpose" (GWAPs) – introduced in [2] – in that while it is ostensibly just a game, it also has the covert purpose of using "human computation" to label data. We are aware of one GWAP which has been applied to a speech recognition task, People Watcher [3]. People Watcher elicits typed alternative proper noun phrasings, which proved useful in a speech directory assistance application. Voice Scatter is different in that it yields orthographically transcribed, continuous speech data.

Moreover, Voice Scatter differs significantly in design from typical GWAPs. Whereas GWAPs are two player games in which existing data is labeled by relying on agreement between two human players, Voice Scatter is a single player game in which new data (utterances) are both elicited and labeled (orthographically transcribed). Labeling is performed via narrow-domain speech recognition, and labeled data is winnowed using confidence scores and game constraints. Voice Scatter also differs in that it is educational, benefiting its players.

The use of an educational website to transcribe data was explored in [4], in which a task intended to help students learn a foreign language was deployed via a prototype website, and used by 24 students to label 10 sentences. Again, unlike Voice Scatter, the task was intended to label an existing corpus, rather than elicit a new one.

Finally, we are unaware of previous evaluations of Amazon Mechanical Turk (AMT) for transcribing continuous speech.

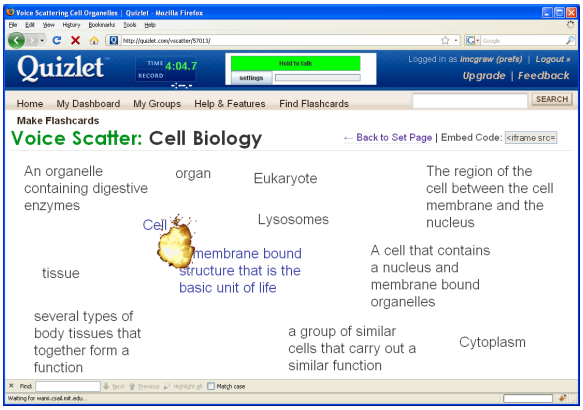

Figure 1: *Screenshot of Voice Scatter.*

However, AMT has been shown to be useful in a number of other natural language labeling tasks; [5] provides an overview of this work, and demonstrates that AMT may be used to produce high quality annotations.

3. The Voice Scatter Game

The Voice Scatter game was introduced to the popular flashcard website Quizlet.com as one of the available study activities. Quizlet users can make and share sets of virtual flashcards, which each contain a *term* on one side and a *definition* on the other. Quizlet boasts 420,000 registered users who have created over 875,000 sets of flashcards, which altogether contain more than 24 million individual flashcards. Although we did not advertise, and we restricted the sets accessible from the game to those which appeared to contain only English words, the level of existing traffic to the site allowed us to collect a significant amount of speech data in a relatively short period of time.

A screenshot of Voice Scatter is shown in Figure 1. Players first choose (or create) a set of flashcards to study. Then, up to eight terms and definitions are "scattered" randomly across the screen. Using a microphone and a web browser, players speak short commands to connect each term to its definition: e.g. "*match cell to a membrane bound structure that is the basic unit of life.*" Players hold the space bar, or click an on-screen hold-to-talk button, while speaking.

When a term is correctly paired with its definition (a "hit"), they come together in a fiery explosion, and then disappear from the screen, as shown in Figure 1. When they are incorrectly paired (a "miss"), they collide and then bounce off of each other. A timer counts upward at the top of the screen, encouraging (though not requiring) players to set a speed record for the flashcard set.

To incorporate speech recognition capabilities into Voice Scatter, the publicly available WAMI Javascript API [2] was used, which is part of the WAMI Toolkit [6]. With it, any web developer can easily make use of MIT's SUMMIT speech recognizer [7] via a web service. The following simple grammar is used as the speech recognizer's language model:

```
[match] <TERM> [to|with|and|equals] <DEF>
[match] <DEF>  [to|with|and|equals] <TERM>
```

where the brackets indicate optionality, and TERM and DEF are any of the terms or definitions on the screen as the game begins.

[2] http://wami.csail.mit.edu

Games Played	4,267	Distinct Words Recognized	21,758
Utterances	30,938	Total Number of "Hits"	10,355
Hours of Audio	27.63	Recognized Words per "Hit"	8.327
Distinct Speakers	1,193†	Distinct Flashcard Sets	1,275

Table 1: *Properties of Voice Scatter data collected over 22 days. † Distinct speakers estimated as one speaker per IP address.*

match robust to strong and vigorous
local area network lan
match silk road with an ancient trade route between china and europe
anything that makes an organism different from others variation
match newtons first law of motion to an object at rest tends to stay at rest and a moving object tends to keep moving in a straight line until it is affected by a force
match what does friar lawrence point out to get romeo to see that life isnt so bad juliet is alive and still his wife tybalt wanted to kill romeo but romeo killed him instead the prince could have condemned him to death but he banished him instead

Table 2: *Example transcripts drawn from the corpus.*

4. Corpus Overview

Voice Scatter elicits utterances containing spontaneous continuous speech; however, because terms and definitions are visible on the screen, utterances – especially long ones – sometimes have the feel of being read aloud. While there is no specific requirement that players read the terms and definitions verbatim, there is a strong incentive to do so to avoid speech recognition errors. In addition, some (but certainly not all) players speak quickly because of the timer displayed during game play.

Table 1 gives a quantitative summary of the collected data. However, the type and variety of the data can be immediately understood by examining the sample transcripts shown in Table 2. As is shown, even though each individual Voice Scatter game is restricted to a small vocabulary, in aggregate there is a large and varied vocabulary. Moreover, by examining a random sample of utterances, we noted that almost all speakers appeared to be teenagers, and that utterances were recorded both in quiet and noisy environments. Noise typically came from televisions, music, computer noise, and people talking in the background. Finally, since players are trying to master unfamiliar material, some words are mispronounced. We observed one player, for example, who consistently mispronounced vocabulary words like "proliferate", "unanimity", and "steadfast".

5. Crowdsourced Transcription

Amazon Mechanical Turk (AMT) is a service where anyone can create web-based tasks and pay anonymous *workers* to complete them. We used AMT to orthographically transcribe 10,000 Voice Scatter utterances drawn from 463 random users (as determined by IP address), which totaled 11.57 hours of speech. Workers were given 10 utterances per page to transcribe. A text box for transcription was initialized with the speech recognizer's top hypothesis, and workers were asked to edit it to reflect the words actually spoken. To guide the transcriber, each utterance was accompanied by a list of terms and definitions from the game associated with that utterance. Each utterance was transcribed by three different workers, yielding 30,000 transcripts created by 130 workers for a total cost of $330.

Since we have 3 transcripts for each utterance, we must combine them somehow to form a gold-standard *AMT-*

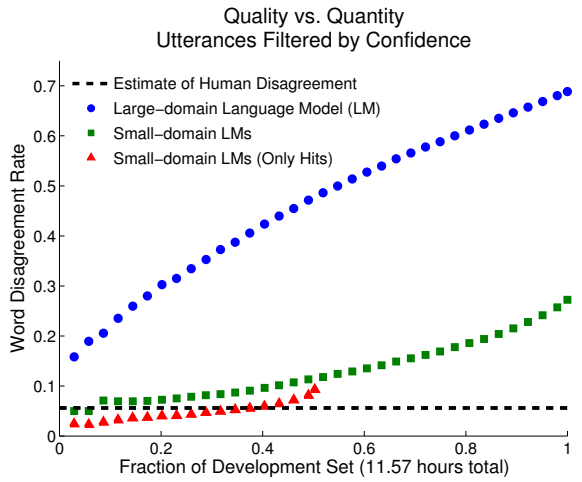

Figure 2: *Cumulative Word Disagreement Rate (WDR) for recognition hypotheses produced using either a large domain trigram or many small-domain grammars on the 10,000 utterance AMT-transcribed set. Cumulative subsets are created by incrementally adding hypotheses ordered by confidence score. An estimate of human WDR, calculated using the 1,000 utterance expert-transcribed subset, is shown for comparison.*

transcript. We chose the majority transcript if there was exact agreement by at least two of the workers, and selected a transcript at random if all three workers disagreed. There was majority agreement on 86.7% of utterances.

To assess the reliability of transcripts obtained in this manner, the first two authors each performed the same transcription task on a 1,000-utterance subset of the AMT-transcribed data. Inter-transcriber "Word Disagreement Rate" (WDR) was computed, given N transcripts from two transcribers A and B, as follows:

$$WDR = \left(\frac{\sum_{i=1}^{N} Sub_i + Del_i + Ins_i}{\sum_{i=1}^{N} \frac{1}{2}(length_{i,A} + length_{i,B})} \right)$$

WDR is simply a symmetric version of Word Error Rate, as the denominator is the sum of the average length of each pair of compared transcripts.

The inter-expert WDR was 4.69%. The WDRs between the AMT-transcripts and the first and second authors were 5.55% and 5.67% respectively. Thus, it seems reasonable to treat the AMT-transcripts as a near-expert reference orthography. In addition, the average WDR produced by pairing the three sets of transcripts produced by AMT workers was 12.3%, indicating that obtaining multiple transcripts of each utterance is helpful when using AMT to obtain a reference.

6. Filtering for Accurate Hypotheses

Because Voice Scatter players often read terms and definitions verbatim, a significant portion of the utterances ought to be recognized with no, or very few, errors. In this section, we explore the usefulness of three sources of information in identifying this subset of utterances, with our goal being to identify a subset of the data which can be automatically transcribed with human-like accuracy. First, we consider the utility of speech recognition confidence scores, which provide a measure of uncertainty based on acoustic and lexical features. Second, we look at information from the game context associated with each utterance. In particular, speech recognition hypotheses which produce "hits", where a term is correctly matched to its definition, are unlikely to occur by chance. Third, we explore the importance of using a small vocabulary, strict grammar during recognition by comparing our results to those produced by a trigram trained on all flashcards appearing in the corpus.

Figure 2 explores the usefulness of each of these factors in identifying high-quality subsets of the data. The curves shown are produced from three experiments performed on the 10,000 utterance AMT-transcribed development set. First, we ordered the set of hypotheses logged from game play based on their confidence scores, as produced by the module described in [8]. We then drew utterances from the set in order from high to low confidence, and measured their cumulative Word Disagreement Rate (WDR) to produce the curve indicated with green squares. Second, we performed the same experiment, using only the 4,574 utterances which were identified as "hits" according to their recognition hypotheses. This produced the curve of red triangles. Third, to explore the effect of vocabulary and language model size, we trained a trigram on all flashcard terms and definitions which appeared in the corpus. Using this n-gram as the language model, we re-recognized each utterance to produce a new hypothesis and confidence score. We then drew hypotheses from these results in order of confidence score, to create the curve of blue circles. Finally, the dotted line shows the average WDR between the AMT-transcripts and each expert on the 1,000 utterance expert-transcribed subset. It represents an expectation of human transcription agreement on the set.

First and foremost, it is clear from Figure 2 that the small-domain nature of our recognition tasks is essential. The n-gram language model had an overall WDR of 68.8% when compared to the AMT-transcripts on all 10,000 utterances, whereas the narrow domain LMs achieved a WDR of 27.2%. Moreover, using only confidence scores, it is possible to select a subset containing 15% of the original data with a near-human WDR of 7.0%. Finally, by considering only "hits", it is possible to select a subset containing 39% of the data at a human-like WDR of 5.6% by discarding just 78 minutes of low-confidence "hits". Indeed, ignoring confidence scores altogether, and simply choosing all "hits", yields 50.2% of the data at a WDR of 9.3%. It is worth noting, however, that on these filtered subsets, human transcripts are still likely to be better. For example, the average WDR between experts and the AMT-transcripts on the 511 expert-transcribed "hits" was only 3.67%.

6.1. Self-Supervised Acoustic Model Adaptation

Orthographically transcribed speech corpora are useful for many tasks. Here we explore using the self-transcribed Voice Scatter sub-corpora in the common task of acoustic model adaptation. We adapt the original acoustic model, used by both Voice Scatter and a related flashcard game called Voice Race [1]. We explore how the quantity and quality of orthographically transcribed *Voice Scatter* data influences the effectiveness of the adapted acoustic model on the *Voice Race* recognition task.

We drew self-transcribed utterances from the 16.05 hours of data that was *not* transcribed by AMT workers, so that we can analyze the usefulness of this transcribed data as a development set. Utterances and their self-"transcripts" were accumulated in one hour increments using each of the three filtering methods described above. After each new hour of data was added to the set, acoustic model MAP adaptation was performed using forced alignments of the self-transcripts. Each adapted acous-

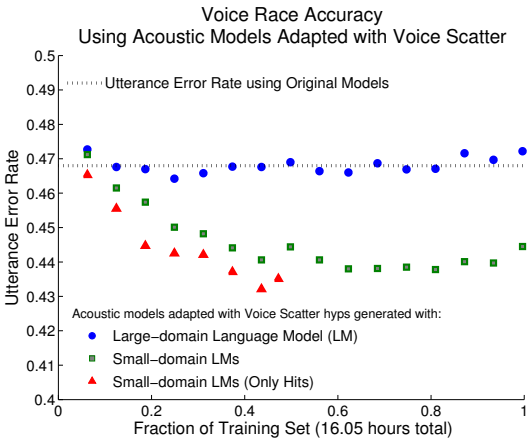

Figure 3: *Voice Race utterance error rate using an acoustic model trained with incrementally more self-transcribed Voice Scatter utterances (sorted by confidence). The self-transcripts are generated using the original acoustic model via: a "Large-domain" n-gram, the small-domain grammars used in the online system, and the "hits" found in hypotheses generated from these small-domain grammars.*

tic model was then used by the speech recognizer to produce hypotheses for 10,000 AMT-labeled utterances collected from Voice Race.

Figure 3 shows the utterance error rate (used here because Voice Race utterances are typically isolated words) found on the the AMT-labeled data using successively larger sets of Voice Scatter utterances filtered via the three methods. First, it is clear that using errorful hypotheses produced by the n-gram language model does not result in an improvement in utterance error rate, regardless of the amount of training data used. Second, using high-confidence hypotheses of utterances recognized with a small-domain language model achieves significant gains, and appears to reach a local minimum when between 60% and 80% of the Voice Scatter data is used. Third, when only "hits" are used, error rates fall faster, and achieve a better local minimum, even though less than half as much total data is available.

Finally, by comparing Figures 2 and 3, we can see that the manually transcribed utterances serve as a useful development set, both to select a filtering method and set a confidence threshold at which to consider data self-transcribed. According to the development set, selecting the high-confidence "hit" data that comprises roughly 39% of the total corpus should yield a human-like WDR. Choosing a training set from utterances based on this operating point would achieve an utterance error rate in Voice Race quite close to the best local minimum shown in Figure 3. Moreover, in the absence of a development set, a 7.8% relative reduction in utterance error rate would have been attained simply by using all of the "hit" data.

7. Conclusions

We presented Voice Scatter, an online educational game that uses speech recognition constrained by many, small-domain language models to collect a rich variety of automatically orthographically transcribed continuous speech utterances. Using game constraints and confidence scores to filter for high-accuracy recognition hypotheses, we show that automatically identified subsets of our data perform well as training corpora for an acoustic model adaptation task. This paper also serves as a compelling case-study of the power of making speech applications available via the World Wide Web. Here, we make use of the open-source WAMI Toolkit [6].

It is not difficult to imagine a wide variety of games, educational or otherwise, which fit the model exemplified by Voice Scatter. Unlike traditional GWAPs, which at times require somewhat contrived game-constraints to produce a label, small-domain speech recognition games may naturally fit into existing web sites that already have large user-bases. Educational games are particularly compelling, because they offer a situation in which players may be satisfied to choose among a small set of answers, the correct one of which is known to the computer. Such small domains both make accurate speech recognition feasible, and provide the opportunity to identify subsets of self-transcribed utterances.

In the future, it may be interesting to explore games which harness additional "human computation". Suppose, for instance, that an English language-learning game was created where learners performed small-domain tasks in English. In a small domain, non-native speech may still be accurately recognized. Moreover, utterances recorded from the game could be made available to a teacher, who would provide feedback to the student by correcting pronunciation errors. A by-product of this fun and educational activity would be a teacher-labeled corpus of non-native speech, which could be used to research algorithms that automatically detect pronunciation errors.

We believe that there are a range of online applications that could benefit by incorporating speech technology. Moreover, the speech research community can benefit from large amounts of cheaply collected, self-transcribed data.

8. Acknowledgments

This research is funded in part by the T-Party project, a joint research program between MIT and Quanta Computer Inc., Taiwan. We are grateful to Stephanie Seneff and Jim Glass for many useful discussions on the ideas presented here. Mitch Peabody and Tara Sainath also provided helpful scripts and advice.

9. References

[1] I. McGraw, A. Gruenstein, and A. Sutherland, "A self-labeling speech corpus: Collecting spoken words with an online educational game," September 2009, *Submitted to INTERSPEECH, draft available at http://wami.csail.mit.edu/papers/Interspeech2009.pdf*.

[2] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. of SIGCHI Conference on Human Factors in Computing Systems*, 2006.

[3] T. Paek, Y.-C. Ju, and C. Meek, "People watcher: A game for eliciting human-transcribed data for automated directory assistance," in *Proc. of INTERSPEECH*, 2007.

[4] J. Cai, J. Feldmar, Y. Laprie, and J.-P. Haton, "Transcribing southern min speech corpora with a web-based language learning system," in *Proc. of International Workshop on Spoken Language Technologies for Under-resourced languages (SLTU)*, May 2008.

[5] R. Snow, B. O'Conner, D. Jurafsky, and A. Y. Ng, "Cheap and fast — but is it good? evaluating non-expert annotations for natural language tasks," in *Proc. of EMNLP*, Oct 2008, pp. 254–263.

[6] A. Gruenstein, I. McGraw, and I. Badr, "The WAMI toolkit for developing, deploying, and evaluating web-accessible multimodal interfaces," in *Proc. of ICMI*, October 2008.

[7] J. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, vol. 17, pp. 137–152, 2003.

[8] T. J. Hazen, S. Seneff, and J. Polifroni, "Recognition confidence scoring and its use in speech understanding systems," *Computer Speech and Language*, vol. 16, pp. 49–67, 2002.