A Self-Labeling Speech Corpus:
Collecting Spoken Words with an Online Educational Game

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

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

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

Abstract

We explore a new approach to collecting and transcribing speech data by using online educational games. One such game, Voice Race, elicited over 55,000 utterances over a 22 day period, representing 18.7 hours of speech. Voice Race was designed such that the transcripts for a significant subset of utterances can be automatically inferred using the contextual constraints of the game. Game context can also be used to simplify transcription to a multiple choice task, which can be performed by non-experts. We found that one third of the speech collected with Voice Race could be automatically transcribed with over 98% accuracy; and that an additional 49% could be labeled cheaply by Amazon Mechanical Turk workers. We demonstrate the utility of the self-labeled speech in an acoustic model adaptation task, which resulted in a reduction in the Voice Race utterance error rate. The collected utterances cover a wide variety of vocabulary, and should be useful across a range of research.

Index Terms: speech recognition, data collection, games with a purpose, self-labeled data

1. Introduction

Collecting and labeling spoken natural language can be time consuming and expensive. Recordings are typically made of subjects performing a task that elicits speech, which must then be transcribed. We present a new approach to speech data collection via online educational speech games. We describe Voice Race, a game which provides a fun way to review flashcards to aid in memorizing vocabulary words, scientific terms, mathematical concepts, and so forth. While each instance of the game is a small-vocabulary recognition task, in aggregate the collected utterances cover a large vocabulary, and should be useful in a variety of speech tasks. Over a 22 day trial deployment on Quizlet.com, a website for creating, sharing, and studying flashcards, Voice Race elicited over 55,000 utterances, constituting 18.7 hours of speech data.

Because each utterance is collected in the course of playing the game, a combination of recognition N-best lists and game context can be used to automatically infer the transcripts for a significant subset of the utterances. Intuitively, when the top recognition hypothesis is known to be a correct answer, this is a strong indication that it is accurate. Using such constraints, 34% of the collected utterances were *automatically* transcribed with near perfect accuracy. For the remaining utterances, game context can also be used to simplify the task of human transcription to one of choosing among several alternative transcripts on a short list. Such a simple task is easy to complete with no training, so we explored using Amazon Mechanical Turk[1] (AMT) for transcription. The transcripts produced by AMT workers agree well with those of two experts.

The approach to collecting spoken language data just outlined is quite powerful in a number of key ways. First, unlike other online games which produce labeled data (*e.g.* [1]), Voice Race is not only fun, but educational as well – providing a tangible benefit to its players. From the perspective of data collection, this also means that it is easy to recruit a large number of willing subjects, which gives rise to a diversity of ages, genders, fluency, accents, noise conditions, microphones, and so forth. Second, its design allows for the automatic transcription of a significant subset of the collected data, and for cheap transcription of the vast majority of the remainder. This means that an arbitrary amount of transcribed utterances may be collected over time at no, or slowly increasing, cost. Third, using the web to elicit speech data is a practical and cheap way to reach a huge number of users. While the first two authors have previously experimented with collecting speech data via the web [2, 3, 4], the results discussed here demonstrate the feasibility of such methods on a significantly larger scale.

2. Related Work

Voice Race was inspired, in part, by the success of so-called "games with a purpose" (GWAPs), as pioneered in [1]. Such games typically provide casual online entertainment for two players, with the covert purpose of harnessing "human computation" to produce labeled data. GWAPs have previously been shown to provide useful data for a speech recognition task: People Watcher [5] elicits alternative phrasings of proper nouns, which are used to improve accuracy in a directory assistance application. It does not, however, produce or label actual speech.

Voice Race is different from GWAPs in a few key respects. First, it is a single-player game. Whereas typical GWAPs rely on the agreement of two humans to obtain labels, Voice Race instead uses artificial intelligence. Contextual constraints and small vocabulary speech recognition are paired to bootstrap the collection of a labeled corpus, which covers a larger vocabulary and a variety of noise conditions. Second, GWAPs label existing data, whereas Voice Race both elicits new speech data, and automatically transcribes much of it. Thus, while Voice Race cannot label arbitrary speech data, it can *continuously* provide new, transcribed speech without any supervision. Third, unlike GWAPs which offer only diversion, Voice Race directly benefits its players by helping them to learn.

Amazon Mechanical Turk (AMT) has been shown to be useful in a number of natural language processing tasks. Notably, [6] show that by aggregating labels of non-expert AMT workers, it is possible to obtain annotations of expert, or near-

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

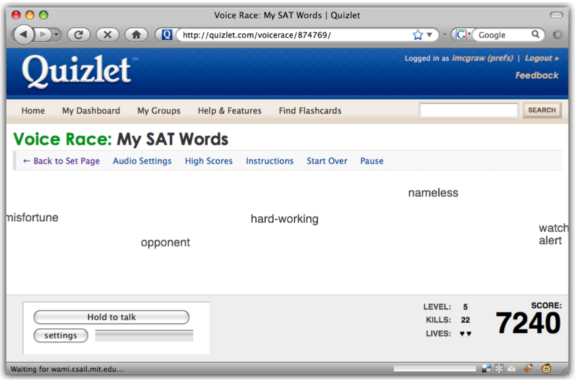


Figure 1: *The Voice Race game with vocabulary flashcards. As a definition moves from left to right, the player must say the corresponding vocabulary word before it flies off the screen. Each such "hit" earns points and makes the definition disappear.*

expert, quality in several NLP tasks. We are unaware, however, of previous evaluations of using AMT for speech transcription.

3. The Voice Race Game

Voice Race is an online game which can be played by anyone with a web browser and a microphone. It is available as part of Quizlet.com, a website where users can create, share, and study virtual flashcards. Like real flashcards, each virtual card has two sides: typically one is used for a *term* – a word or short phrase – and the other for its *definition*. The site is typically used by students to study sets of vocabulary words, science concepts, mathematical facts, and so forth. Quizlet currently has over 420,000 registered users, who have contributed over 875,000 such sets, comprised of more than 24 million flashcards.

In Voice Race, shown in Figure 1, definitions from a set of flashcards move across the screen from left to right. Players must say its matching term before a definition flies off the screen. Each such "hit" earns points and makes the definition disappear. If a definition is never hit, then the player is shown the correct answer and prompted to repeat it aloud. As the game progresses, the definitions move more quickly, and appear more frequently.

Voice Race was built using the WAMI Toolkit for developing web-accessible multimodal interfaces [2]. Specifically, it uses the publicly available WAMI Javascript API,[2] which can be used to integrate speech recognition capabilities available via MIT servers with any web site. A simple context-free grammar is constructed on-demand each time the game is played, in which any term in the set may appear any number of times in an utterance. The SUMMIT speech recognizer [7] is used with a dictionary containing 145,773 words, and an automatic pronunciation module to generate pronunciations for other words.

4. Self-Labeled Data

Beyond providing educational value, Voice Race serves as a vehicle for collecting significant amounts of speech data. Since it is available on a well-trafficked website with a motivated user base, it is well positioned to elicit a significant number of utterances from many users using different microphones and in varied noise conditions. Moreover, the collected data is especially valuable because each utterance occurs in a context where the correct answer (or answers) is known. This information, when combined with the recognition results, can be used to automatically infer the transcript for a significant subset of the utterances, and greatly limit the set of likely transcripts for the rest. The subsets of interest are as follows:

Hit In Voice Race, a "hit" occurs when the top speech recognition hypothesis contains the correct term associated with a definition visible on the screen. Players typically aim for the right-most definition, so such "hits" are likely to be the most reliable indicators of an accurate recognition hypothesis.

Miss A "miss" occurs when the user has spoken, but a hit has not been detected. There is no way of knowing if a miss is due to a human error or a speech recognition error. However, when misses are due to recognition errors, the correct transcript for the user's utterance is likely to be one of the correct answers. As such, when considered in aggregate, misses may be useful for automatically identifying difficult terms to recognize.

Prompted Hit/Miss The taxonomy above applies to most Voice Race utterances. Voice Race also provides an additional category of labeled data: when a definition flies off the screen without being "hit", players are shown the correct answer and prompted to read it aloud. As such, when players are cooperative, the transcript of their utterances should be known in advance. Moreover, when these utterances are run through the same small-vocabulary recognizer used for the game, they can again be classified as "hits" – which indicate a high likelihood that the player faithfully repeated the prompt – or as "misses".

5. Simplified Transcription

The contextual game constraints identified in the previous section are useful for automatically transcribing a significant portion of the data – the "hits" in particular. In addition, the same constraints may be used to greatly decrease the human effort required to transcribe the remainder. For each utterance, the transcript is likely to be one of the correct answers, to appear on the N-best list, or both. This means that the task of human transcription for most utterances can be reduced to one of choosing a transcript from a short list of choices drawn from these two sources. Given that it requires no expertise or knowledge of the task domain to listen to a short audio clip and choose a transcript from a list, we designed a transcription task which could tap the large pool of Amazon Mechanical Turk (AMT) workers.

We designed the AMT transcription task such that workers listen to a Voice Race utterance and then choose from one of four likely transcripts. They can also choose "None of these"' or "Not Speech". The likely transcripts were drawn in order from the following sources, until four unique candidate transcripts were obtained:

1. The prompted term, if the user was asked to repeat aloud
2. The top two distinct terms in the recognition N-best list
3. The terms associated with the two right-most definitions
4. Any remaining terms on the N-best list
5. Random terms from the flashcard set

After transcribers select a transcript, they can optionally label two additional attributes. *Cut-off* indicates that the speech was cut off – this happens occasionally because players release the space bar, which they must hold while speaking, before they finish. Future iterations of the game will likely correct for this by recording slightly past the release of the key. Transcribers may also select *Almost* if the utterance was understandable, but

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

Games Played	4184	Mean Words per Utt.	1.54
Utterances	55,152	Total Distinct Phrases	26,542
Total Hours of Audio	18.7	Mean Category Size	53.6

Table 1: *Properties of Voice Race data collected over 22 days.*

5-way agreement	69.2%	Majority "None of these"	12.9%
4-way agreement	18.0%	Majority "cut-off"	12.1%
3-way agreement	9.8%	Majority "almost"	7.2%

Table 2: *Agreement obtained for transcripts and attributes of 10,000 utterances, each labeled by 5 AMT workers.*

Game Context:	miss	hit	prompted-miss	prompted-hit
% Correct:	13.9	86.4	12.7	97.5
% of Total Data:	43.7	43.8	8.9	3.6

Hit Context:	4-hit	3-hit	2-hit	1-hit
% Correct:	41.3	69.4	81.7	**98.5**
% of Hit Data:	1.8	3.4	9.0	**69.4**

Table 3: *% of AMT-labeled data and recognition accuracy grouped by game context. Hits are further broken down in terms of the position of the item on the screen at the time the hit occurred. Statistics for the four right-most positions are shown.*

contained hesitations, extra syllables, mispronunciations, and so forth.

6. Data Analysis

Voice Race was made available on Quizlet.com for a 22 day trial period. No announcements or advertisements were made; the two games were simply added to the list of activities available to study each (English) flashcard set. Nonetheless, as Table 1 shows, a total of 55,152 utterances were collected, containing 18.7 hours of speech.

6.1. Transcription

10,000 utterances representing 173 minutes of audio were drawn from 778 Voice Race sessions and then submitted for transcription to Amazon Mechanical Turk (AMT). Within 16 hours, each utterance had been labeled by 5 different AMT workers using the simplified transcription task discussed in the previous section, at a cost of $275.

Table 2 shows agreement statistics for the workers. A majority agreed on one of the transcript choices for 97% of the utterances, agreeing on "None of these" only 13% of the time. Thus, the simple forced choice among 4 likely candidates (and "no speech") yielded transcripts for 84% of the utterances.

To judge the accuracy of the produced labels, the first two authors each labeled 1,000 randomly drawn utterances. Their transcript choices showed a high level of agreement, with a Cohen's Kappa score of 0.89. Each of their label sets agreed well with the majority labels produced by the AMT workers, as measured by Kappa scores of 0.85 and 0.83.

Using the AMT majority labels as a reference transcription, the utterance-level recognition accuracy on the set of 10,000 Voice Race utterances was found to be 53.2% . While accuracy is low, it's important to note that the task is a difficult one. The two authors noted while transcribing that (1) the vast majority of the utterances seemed to be from teenagers, (2) there was often significant background noise from televisions, music, or classrooms full of talking students, and (3) many microphones produced muffled or clipped audio. While these problems lead to imperfect speech recognition accuracy, they also lead to a richer, more interesting corpus. Moreover, usage levels suggest that accuracy was high enough for many successful games. In the next section, we show that, despite relatively poor recognition performance overall, it is nonetheless possible to use game context to automatically obtain *near-perfect* transcriptions on a significant subset of the data.

6.2. Automatic Transcription

Because each utterance occurs in the course of playing Voice Race, we hypothesized that it should be possible to identify a subset of the data for which transcripts can be inferred automatically with high accuracy. In this section, we evaluate this hypothesis using as reference the transcripts agreed upon by a majority of AMT workers. We explore the utility of *hits*, *misses*, *prompted-hits* and *prompted-misses*. Table 3 shows the amount of speech data collected in each category out of the 10,000 AMT-labeled utterances.

Over 4,000 of the 10,000 utterances were *hits*, and the recognition accuracy on this data is 86.4%. In addition, *prompted-hits* yield an accuracy of 97.5%, meaning that they yield nearly perfectly transcribed data. Unfortunately, they represent less than 5% of the data.

Using game-context to filter data for accurately labeled utterances can be taken further in the case of a *hit*. Students are most likely to aim for the right-most label on the Voice Race screen. It stands to reason then, that *hits* of definitions which are not the right-most one are more likely to be due to a misrecognition. We call a hit that occurred while the item was in the *nth* position on the screen (from right-to-left) an *n-hit*. Recognition accuracies for $n = 1 \ldots 4$ are presented in Table 3.

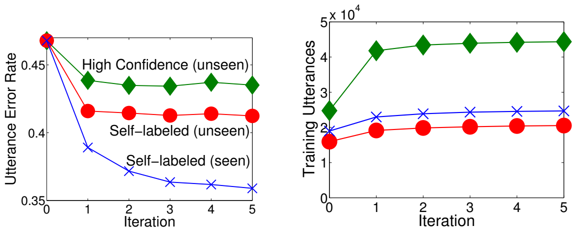
It is exciting to note that *1-hits* constitute 30.4% of the total AMT-labeled data, and are recognized with 98.5% accuracy. Of all 55,152 utterances collected, 18,699 – representing 5.8 hours of audio – are self-labeled in this fashion.

6.3. Self-Supervised Acoustic Model Adaptation

One common use of transcribed speech data is to perform acoustic model adaptation. While typically this requires human transcription, we explore using the automatically transcribed utterances to adapt the telephone acoustic models used by Voice Race in a fully automatic fashion. We performed iterative MAP adaptation using "hits", which resulted in a decrease in utterance error rate from 46.8% to 41.2%, as calculated using the 10,000 utterances transcribed by AMT workers.

In the first iteration, the original acoustic models are adapted using models trained on all collected self-labeled utterances from the 45,152 utterances without human labels, treating the 10,000 AMT-labeled utterances as "unseen" data. Then, all 45,152 utterances are re-recognized and the "hits" are recomputed, yielding a larger pool of self-labeled training data, which is then used in the following iteration. As a baseline, the same technique was used to iteratively select utterances which received high recognition confidence scores.

As Figure 2 shows, using less than half of the data, the training set of self-labeled utterances achieves an additional 2.2% absolute improvement in error rate over the iteratively calculated high-confidence utterances. Figure 2 also shows the results of iteratively selecting from all 55,152 utterances (without using the AMT-labels), in effect, treating the 10,000 utterance test set as "seen" data. Iteratively selecting high-confidence training utterances from *all* the data achieves error rates simi-



(a) Utterance Error Rate

(b) Training Set Size

Figure 2: *Iterative acoustic model adaptation, trained using: (1) Iteratively calculated high-confidence utterances, excluding the 10,000 AMT-transcribed test set (i.e. the test data is unseen), (2) An iteratively calculated set of self-labeled utterances (unseen). (3) An iteratively calculated set of self-labeled utterances, including test utterances (seen).*

lar to those found when selecting self-labeled utterances from the original 45,152 utterances, and is omitted from the graph for clarity. Iteratively selecting *self-labeled* utterances from all of the data, however, improves performance significantly, even across iterations. The *iterative* gains are likely due to the fact that the self-adaptation set now includes utterances gathered from the same session, meaning that the speaker, acoustic environment, and vocabulary are the same. This hints at the potential for games like Voice Race to improve in a personalized, fully automatic, online fashion.

Furthermore, since the improved acoustic models are used to re-label existing data, we can assess the quality of these iteratively adapted automatic labels. Using the models iteratively trained on all 55,152 utterances we were able to automatically label 44.8% of the data while maintaining an accuracy of 96.8% on the subset of labels corresponding to the AMT-transcribed data.

7. Self-Labeled Continuous Speech

Voice Race is a fun game, and an excellent source of automatically and/or cheaply transcribed data; however, the elicited utterances typically contain only a single word or short phrase. To explore collecting labeled continuous speech, we have developed another flashcard game, Voice Scatter, which elicits significantly longer utterances. In Voice Scatter, players match terms and their definitions, which are scattered randomly across the screen. The spoken term and its definition are highlighted and then collide. If correctly paired, they “explode” and disappear from the screen. By including arbitrary definitions, Voice Scatter elicits much longer utterances than Voice Race, and, again, each utterance is associated with a particular game context.

Over the same 22-day period, we collected 30,938 Voice Scatter utterances, constituting 16.8 hours of speech. In [8], we show that similar automatic labeling methods, in combination with recognition confidence scores, yield near human-level transcription accuracy.

8. Conclusion

We have demonstrated a new approach to collecting and transcribing speech data, through the use of online educational games. We presented Voice Race, a flashcard study game that elicits speech, and showed how game context can be used to automatically transcribe much of the data, and to simplify the transcription task for most of the rest. Moreover, we demonstrated the utility of the automatically transcribed utterances in a self-supervised acoustic model adaptation task, which showed an improvement in accuracy without any human transcription required. With techniques like this, the accuracy of games like Voice Race can be improved over time automatically, while also yielding transcribed data which may be useful for other tasks.

There are a number of open questions in speech research for which the ever-growing Voice Race corpus might provide a useful experimental platform. Because players can play with any set of flashcards, words for which no pronunciation exists in the dictionary will routinely occur. Voice Race is a small vocabulary task, so automatic pronunciation rules may function “well enough” for game play, while simultaneously labeling data from which new pronunciations may be learned.

Beyond the corpus presented here, the automatic labeling and transcription methodologies we’ve developed should be applicable to collecting a variety of data. An educational game for learning English, for example, might produce a large corpus of automatically labeled, non-native speech. Alternatively, a children’s math game might yield a child-digit corpus.

The WAMI toolkit [2], which was used to add speech capabilities to Voice Race, provides an easy framework for researchers to make games and other speech applications available via the web. The web makes it simple and cheap to collect large amounts of data, so if even only a small portion of data from an application can be automatically and/or cheaply transcribed, with enough usage a large transcribed corpus may still be produced. By partnering with existing websites, as was done in this paper, it is possible for researchers to collect significant amounts of data from motivated users.

9. Acknowledgments

Jim Glass and Stephanie Seneff provided many valuable discussions. Tara Sainath helped us with acoustic model adaptation. This research is sponsored in part by the T-Party Project, a joint research program between MIT and Quanta Computer Inc., Taiwan.

10. References

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

[2] 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.

[3] A. Gruenstein and S. Seneff, “Releasing a multimodal dialogue system into the wild: User support mechanisms,” in *Proc. of the 8th SIGdial Workshop on Discourse and Dialogue*, 2007, pp. 111–119.

[4] I. McGraw and S. Seneff, “Speech-enabled card games for language learners,” in *Proc. of the 23rd AAAI Conference on Artificial Intelligence*, 2008.

[5] 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.

[6] 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.

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

[8] A. Gruenstein, I. McGraw, and A. Sutherland, “A self-transcribing speech corpus: collecting continuous speech with an online educational game,” Submitted to *the Speech and Language Technology in Education (SLaTE) Workshop*, 2009.