Speech Recognition for Early Literacy

Anna Utgoff

October 10, 2019

Executive Summary

Helping all children read fluently by third grade is an incredibly important public policy goal. 4th graders who aren't reading on grade level are at a disadvantage for the rest of their lives. Children need practice reading aloud to become fluent readers, but many don't get the practice they need. Speech recognition for children could help solve this problem by allowing educational apps to "listen" to children read and give them helpful feedback. This isn't happening yet because speech recognition for children is not yet accurate enough to be widely useable. Public and private investment could change that by accelerating the development of speech recognition for children. This proposal argues for an investment in public data sets and "common task"-style challenges with metrics for evaluating research. This approach has proven successful in adult speech recognition, where DARPA's investment in early speech laid the groundwork for technological advances that eventually led to widespread adoption of consumer speech recognition like Siri and Alexa.

Proposal

Introduction to the problem

The opportunity to learn to read well is perhaps the most important we can give a child. The 60% of children who aren't reading on grade level by the end of 3rd grade are unlikely to catch up, and will find themselves at a disadvantage for the rest of their lives. They will be unable to absorb 50% of their school curriculum[1], less likely to graduate from high school[2], and more likely to end up in prison[3].

In order to reach the critical milestone of reading fluently by third grade, students need practice. And not just any practice will do: encouraging students to practice reading independently doesn't measurably impact their reading skills[4]. To grow as readers, students need practice reading aloud to someone who can listen and provide feedback.[5]

Unfortunately, many students are starved for this kind of practice. Less than half of students get the 15 minutes a day of reading time that researchers recommend[6]. Teachers

---

[1] Fiester, L. (2010, Jan 1). "Early Warning! Why Reading by the End of Third Grade Matters." p9
[2] The Campaign for Grade Level Reading. "THIRD GRADE READING SUCCESS MATTERS." Retrieved from http://earlychildhoodfunders.org/pdf/GLR_Brochuretoprint_3-1.pdf
[3] Hudson, J. (2012, Jul 2). "An Urban Myth That Should Be True." *The Atlantic.*
[4] National Reading Panel. (2000, April.) "TEACHING CHILDREN TO READ: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction."
[5] Ambruster, B, et al. *Putting Reading First:* The Teacher's Guide to the National Reading Panel findings.
[6] "The magic of 15 minutes: Reading practice and reading growth" https://www.renaissance.com/2018/01/23/blog-magic-15-minutes-reading-practice-reading-growth/

resort to techniques like "round-robin reading", where each child gets about a minute of read aloud practice time.[7] Busy parents struggle to fit in reading time: only 55% of families read with their children daily, and less than half of those families ask children to practice reading aloud.[8]

Reading experts have thought for years that speech recognition could help solve this problem. If computers can process human speech, why not have them listen to a child practice reading and provide helpful feedback? As early as 2000, the National Reading Panel identified speech recognition as a promising area for research.[9] Venerated reading researcher Marilyn Adams experimented with speech recognition technology in the early 2000s and concluded, "In our study of grade 2–5 classrooms, those using [speech recognition] showed remarkable growth in fluency."[10]

Despite this promise, speech recognition is not yet widely used to teach reading. As it turns out, there are technical challenges to making speech recognition work well for kids, which we will explore below. These problems be solved, so that we speech recognition technology can help to close the reading achievement gap.

Review of past attempts to solve the problem

Speech recognition technology has made incredible advances in the past decade. Scientists have developed deep neural networks that transcribe human speech more accurately than people can. Multiple companies[11] are attempting to apply this technology to early literacy and yet none have succeeded at scale. Why?

Speech recognition for kids is a hard problem, and technology designed for adults is only partially suited to solve it. First and foremost, kids' voices are harder for computers to understand. They're just learning to speak, and they often have difficulty enunciating certain sounds and letters. Their voices are higher pitched, which makes them harder signals to interpret.

Second, the goals of speech recognition for kids are different than those of adult speech recognition. Siri, for example, wants to give you the benefit of the doubt. If you misspeak or mispronounce a word, she will try to ignore your mistake and guess what you meant to say. When teaching small children how to read, on the other hand, mispronunciations,

---

[7] Shanahan, T. (2019, July 30). "Is Round Robin Reading Really That Bad?" https://www.readingrockets.org/blogs/shanahan-literacy/round-robin-reading-really-bad

[8] "Kids & Family Reading Report: The Rise Of Read-Aloud," 7th Edition. (2019) scholastic.com/readingreport

[9] National Reading Panel, 2000

[10] Adams, M. J. (2006) "The promise of automatic speech recognition for fostering literacy growth in children and adults."

[11] See, for example: Amira Learning, Lalilo, and Soapbox Labs

hesitations, and false starts are key pieces of information. Siri has been trained to ignore mistakes, but children's speech recognition must be trained to detect them.

Both of these obstacles are surmountable, with enough data. Speech recognition algorithms, like all machine learning systems, are pattern recognition systems. Data scientists "train" algorithms by feeding them many different recordings and then telling them how humans transcribed those recordings. Eventually, the algorithms learn patterns, and can use those patterns to transcribe new recordings. The more training data the algorithms get, the more accurate they become.

Which brings us to our third challenge: lack of data. There is no large corpus of recordings of children reading, because privacy regulations make it difficult to collect & share that data. Speech recognition for children is being deprived of oxygen because there's such scant data for researchers to use for training algorithms.

Entrepreneurs and data scientists working on this problem agree: they need access to more data. According to Mark Liberman, executive director of the Linguistic Data Consortium, "Speech recognition for kids hasn't come as far as ASR for adults, in part because it's hard to get good data."[12] Dr. Patti Price, cofounder of Soliloquy and researcher on the TBALL project[13] (both of which used speech recognition to teach reading), explains "The data needs to be specific to the application: you need recordings of children reading to develop speech recognition for early readers."[14] Benjamin Abdi, cofounder of literacy speech recognition company Lalilo, believes "Data is the number one thing we need."[15]

The history of adult speech recognition shows that public datasets can spur enormous innovation. Alexa and Siri exist because of a new approach to government funding adopted in the 1980s, the "common task" methodology. The idea was simple: science moves faster in the sunlight. The best way to accelerate speech recognition, the Defense Advanced Research Projects Agency (DARPA) concluded, was to provide scientists with both large shared data sets and clear metrics for success[16]. Access to data meant that researchers could train algorithms without spending lots of money on collecting data. Sharing data sets and metrics meant that researchers could directly compare their results. Before common tasks, it was easy to claim "100% accuracy" on some small data set. Common tasks solved that problem: if your algorithm achieved a 94% Word Error Rate on the Switchboard corpus

---

[12] Private email correspondence, September 14 2019
[13] 2005, "Tball Data Collection: The Making of a Young Children's Speech Corpus," A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Andersen, S. Narayanan, and A. Alwan, Proc. Eurospeech, Portugal. http://www.pprice.com/papers/tball_data_coll_final.pdf
[14] Private correspondence, September 26 2019
[15] Private email correspondence, September 16 2019
[16] 2017. "Finding a voice." *The Economist.*
https://www.economist.com/node/21710907/sites/all/modules/custom/ec_essay

(the most widely used data set), it was crystal clear whether that represented an advance over previous research.

This approach proved enormously effective. Since the founding of the Linguistic Data Consortium (the organization that maintains the shared data sets), common task metrics such as the "Word Error Rate" have steadily declined, and major commercial systems continue to use these datasets and metrics[17].

Introduction of your idea

I propose a three stage investment that uses the common task model to accelerate speech recognition to the point where it can be widely and effectively used in literacy instruction: Data collection, research, and application.

Phase 1: Data Collection

To collect a large, privacy-compliant, public dataset of children reading, we will donate free online reading tutoring to participating families. This will allow us to collect recordings of children reading aloud, both uninterrupted and with corrections made by trained educators. As an added benefit, this project would make hundreds of hours of tutoring available to families, most of whom could not otherwise afford it.

In order to comply with privacy standards:

- Parents will be fully informed of the intended uses for their children's voices and asked to provide verifiable consent
- The data will be encrypted and anonymized (see below for details)
- Parents will have the right to request their children's voices be deleted from the corpus

We will recruit participants across demographic groups: regions, primary languages, socioeconomic status, and reading ability. Any algorithm trained on this corpus will work well for all demographic groups, including the children who most need supplementary reading support.

Each tutoring sessions will be transcribed and annotated to make it usable training data for machine learning algorithms. Some participating families will be provided with eye-tracking peripheral devices[18], to enrich the data set with information about which words children are looking at as they read. We will also collect demographic data and reading

---

[17] Paul, S. (2017, March 20.) "Voice Is the Next Big Platform, Unless You Have an Accent." *Wired Magazine.* https://www.wired.com/2017/03/voice-is-the-next-big-platform-unless-you-have-an-accent/

[18] The advent of consumer eye tracking devices for use in video games makes this possible to do at scale, see for example https://www.tobii.com/. Eye movements during reading have been shown to be predictive of reading disabilities, so this data would make the corpus more useful to scientists researching dyslexia and related disorders.

level progress data for the children. The data collection effort should be done in partnership with school districts, who will be able to provide additional data such as state test scores and the classroom reading instruction methods used.

To ensure privacy, the data set would be scrubbed to remove names and any other personally identifiable information about the participants[19]. Each child's data will be tagged with an ID number, so that a single child can be tracked anonymously over time. This ID will also allow parents to request removal of their child's data from the corpus.

The corpora will be hosted and managed by the Linguistic Data Consortium (LDC), which already hosts corpora for adult speech. The data will be encrypted, and the LDC will require researchers using the dataset sign confidentiality agreements and complete the same background checks required of adults who work with children and children's data.

Stage 2: Research

After the data sets are collected, the next step will be to establish common task-style benchmark metrics. The two most important metrics for these data sets are:

- **Accurate detection of disfluencies: % of miscues[20] detected - % false alarms[21]**
  whether the algorithm successfully detected mistakes made while reading.
- **Reading level accuracy: % confidence with which the algorithm can determine a child's reading level based on a sample of their reading**
  whether an algorithm could confidently guess a child's reading level as determined by an educator using a standard assessment.

The availability of data & common tasks will incentivize research. We will accelerate progress further by offering research funding to scientists working on this problem. Prize money should also be offered to teams that are able to meet or exceed goals on the common tasks, on the condition that the teams open source the winning algorithms, and all competitors agree to publish the details of their research. This will make some of the best children's speech recognition algorithms widely available, and allow researchers to learn from each other's efforts.

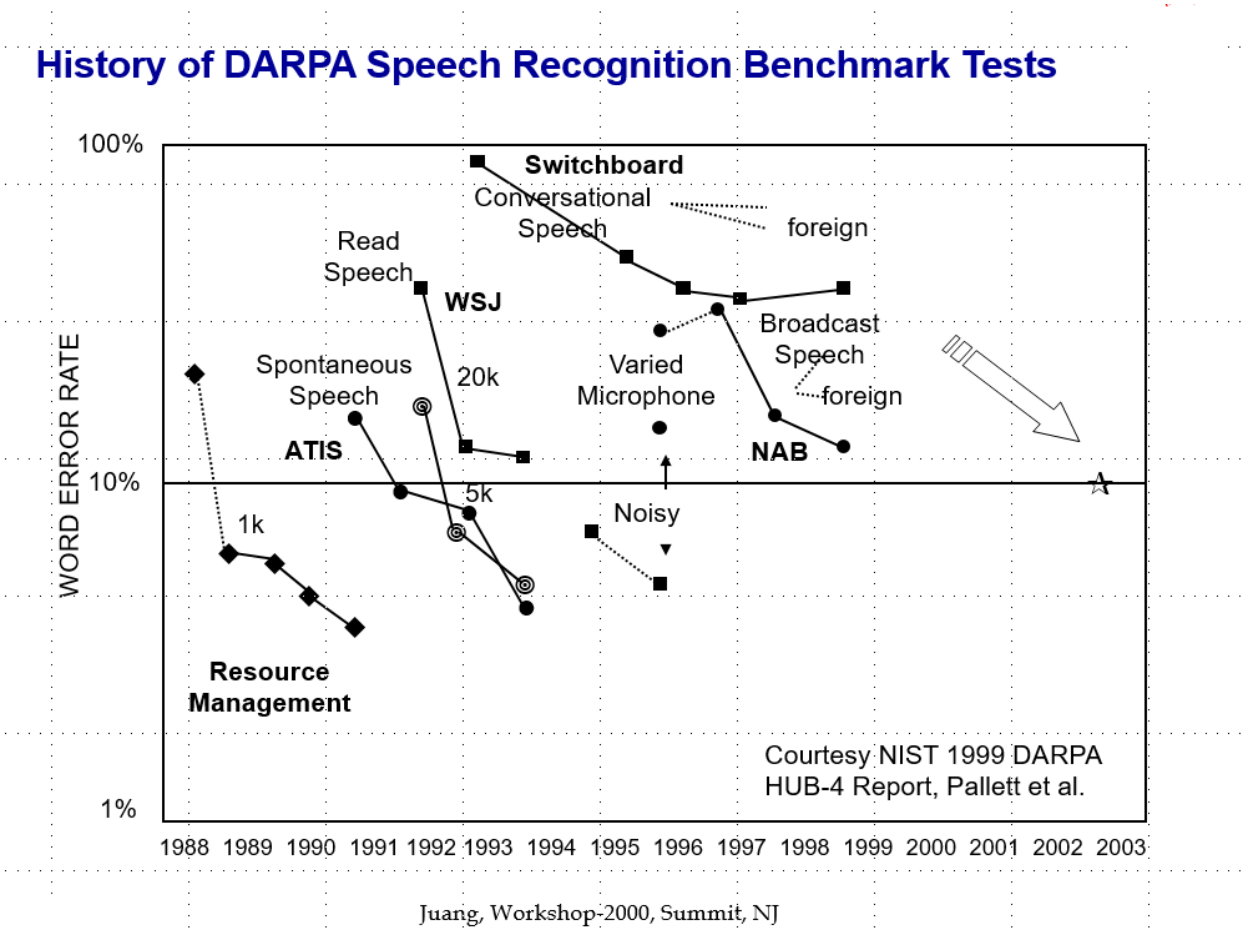
---

[19] In the case of small classes (only one or two children within a given demographic designation (e.g., age 7, grade 2, 2022, native language Lao, reading level grade 1.5, California), data would be invented to shield identification.

[20] "Disfluency" is the word speech recognition researchers use to describe a mistake in reading; "Miscue" is the word used by educators. They are used interchangeably throughout this proposal.

[21] Subtracting the false alarm rate (% of words inaccurately labeled as disfluencies) would ensure that the algorithms did not attempt to succeed by overcorrecting (e.g. labeling every word as a miscue).

As progress is made, new and more challenging common tasks will be established: speech with background noise, speech from younger children, etc. Here again we are following the DARPA model. As the chart below shows, DARPA pushed forward work in speech recognition by continually establishing new and more challenging benchmark tests. Each new step forward in speech recognition will open up new problems previously thought unsolvable.



Stage 3: Applications

The research conducted in stage 2 will be an enormous boon to educational software. Adding accurate speech recognition to a children's product will suddenly be feasible even for small companies. App developers will be able to experiment with ways that ASR can improve students' learning. Just as the ease of developing with Alexa has led to the release of over 80,000 Alexa skills, the availability of open source algorithms will enable the development of many children's speech recognition apps.

The reading level algorithms developed for the second common task may prove even more valuable than disfluency detection. Right now the only way for app developers to know what works is to conduct expensive randomized control trials, or to build their own

assessments and hope they're valid measures of learning. When phase 2 produces an algorithm that can estimate student reading levels, educational app developers will suddenly have a valid way to assess children's reading skills. That will make it easy to bake formative assessment into the user experience, simply by asking children to read a few sentences during game play. This data would help app developers to track student growth, and experiment rapidly with ways to maximize that growth.

When app developers can easily measure student reading progress, education funders can ask for that data and use it to make funding decisions. Foundations and government programs such as the SBIR that are looking for effective programs to support will be able to invest with more confidence using this data.

Philanthropic funding should also support the development of a free reading level assessment app. Elementary teachers spend hours assessing their student's reading level via "running records", which are so time consuming they're usually only administered a few times a year. The algorithms developed in phase 2 could power "automated running records" to supplement or even replace manual assessments. Funders should make this tool free and universally available. Users of this automated assessment tool could be asked to voluntarily donate their data, so that new voice recordings continue to flow into the speech corpus.

Implementation Description

The program will be supervised by a committee of experts in reading, speech recognition, and privacy. The committee would be supported by a small permanent staff. This supervisory body would be responsible for:

- Defining the specifications for data to be collected, issuing calls for proposals to collect it, and awarding contracts
- Setting metrics for success for common tasks, and determining when sufficient progress had been made to merit new, more challenging tasks
- Reviewing proposals for speech recognition research funding and awarding funds
- Establishing prizes for meeting common task goals, evaluating prize submissions, and awarding prize money
- Ensuring all funded research and data collection complied with privacy regulations and human subject research regulations

A process for incorporating feedback and evaluation to improve program delivery and outcomes

The strength of the common task methodology is that evaluation is baked into the approach: participating in the initiative requires participants to evaluate their work using a

set of clearly established metrics. The supervisory body described above would report regularly on progress on established common tasks.

In addition, the supervisory body would release annual reports on the degree to which the funded research was meeting outcome goals other than technical improvements, such as the degree to which:

- Data collected reflected the demographics of the United States K-12 school population
- Funded research was resulting in applied and widely available technological advancements
- Education applications identified as having "proven results" via reading level algorithms proved to be so when assessed via more rigorous methods like large-scale randomized control trials

Conclusion

Voice recognition technology for children has enormous promise as an educational tool. Before that promise is realized, the technical challenges described above must be addressed. Those technical challenges are far from easy. But one must remember that decades ago, the challenge of using voice recognition to transcribe a conversation or search a database seemed insurmountably difficult. Public investment made those things feasible, and turned adult voice recognition into a commercially viable and widely used tool. Public investment can likewise make voice recognition for children educationally viable. And when that happens, millions more children will be reading fluently by 3rd grade, and have access to all the opportunities that reading unlocks.