Every Voice: A Proposal for a National Initiative in Children's Speech Recognition

Jill Fain Lehman

It has been scarcely more than 50 years since the main theoretical and practical underpinnings of continuous speech recognition were established. In that time the technology has left the research lab and entered the commercial mainstream, with products and services that produce tens of billions of dollars in revenue each year. Speech-enabled applications are a growth sector and language-based interactions, which sit on top of robust recognizers, are being deployed across a wide spectrum of domains. Just as important, the availability of free, albeit limited, access to some of the best speech recognition technology through APIs and SDKs has made it possible for researchers, start-ups, and small businesses to go further than they could have done without having that capability at hand.

Circumscribing this substantive progress, however, is a fact that typically does not get trumpeted: recognition depends on data, and almost without exception recognizers are trained on adult speech. That they do not work well for children, particularly young children, has been demonstrated empirically and repeatedly. Researchers do not all agree about how to solve the many problems in children's speech that make it more difficult than adult recognition, but they do all agree that the problems cannot be solved without children's speech data. A lot of it, at all ages.

We are at an inflection point. At this moment there are companies large and small that have begun to notice the area of children's language interaction as a potentially lucrative one. They, too, understand that data is key, and although the question of how to collect the data within existing constraints is slowing them down, it will not do so indefinitely. If it is left to the commercial sector to extend the technological benefits of speech recognition to children's voices, that data will be proprietary and its uses tied to the narrow interests of the businesses that collect it.

There is an alternative path and this is the moment to take it. A national initiative to collect and transcribe children's speech would start everyone – public and private, research and enterprise – on a more even playing field. We could guarantee that the data is acquired ethically. We could guarantee that children's voices all across

Jill Fain Lehman
Carnegie Mellon University, Pittsburgh, PA e-mail: jfl@andrew.cmu.edu

the country, irrespective of location, community and personal access to technology, become part of a national database. We could leapfrog a decade or more of narrowly focused, slowly and unevenly-accessible functionality by making the resource freely available to everybody – jump starting new areas of research into educational and therapeutic applications, as well as start ups and new technologies. Finally, we could demonstrate the viability of such an effort to other countries which, by virtue of their own infrastructure, social institutions, and language communities, would necessarily differ in implementation details without differing in intent.

A national effort, sponsored by a governmental scientific agency, and implemented by existing trusted institutions at the local level can succeed in producing the quantity and representativeness of data necessary. A proposal to ensure inclusion, anonymity and privacy protection, and breadth of access to the data is presented below and rests firmly on the belief that as a national initiative, the creation of this resource should involve, in one way or another, as much of the population as is willing to contribute. To that end, the particular mechanisms and institutions mentioned may be replaceable, but the principles behind them are not:

- To promote both inclusion and privacy, collection should allow data to be uploaded only from accounts registered by schools and public libraries, so that geographic area is known but the IP addresses of individual's phones or computers are not. Additional demographic data should include only gender and month/year of birth as identifying information so that the combination of elements cannot be linked to an individual.
- To make sure no specialized hardware is required, data should be collected via a downloadable app that uses existing microphones (phone, tablet, computer).
- Because the distribution of sounds in children's language is much more variable as a function of both age and language task than in adults', the protocol should involve multiple language activities to capture listen-and-repeat, conversational, and storytelling data, with versions that are adjusted to be age appropriate.
- To further promote involvement across the population, the elicitation protocol should be designed so that it can be run by almost anybody with little to no training, enabling data collection by teachers and librarians as well as parents and student volunteers – for example, scout troops, youth groups, etc. – at those locations.
- Guidelines for word-level transcription should be straightforward enough that the transcription itself can be done through crowdsourcing. There is now enough expertise in how to design redundant crowdsourced tasks to overcome most error. The use of crowdsourcing to transcribe the speech at a basic level presents a temporary employment opportunity for a large, distributed portion of the population and can usefully be thought of as an effort to maintain the country's information infrastructure in the same way that bridge and road work maintains the country's physical infrastructure.
- The curated and transcribed data should become part of the Library of Congress's electronic collection as a national resource available to every citizen, researcher, entrepreneur and company.

A national initiative works only to the extent that the country as a whole both wants to participate and can participate, independent of economic class, political belief, race, gender, and immigration status. To succeed requires a clear message of celebration: we are a country made up of different voices and we work and build for our children so that in the future *every* voice can be heard.