Guiding Students to Investigate What Google Speech Recognition Knows about Language

David S. Touretzky[1], **Christina Gardner-McCune**[2]

[1] Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213
[2] Department of Computer & Information Science & Engineering, University of Florida, Gainesville, FL 32611
dst@cs.cmu.edu, gmccune@ufl.edu

Abstract

Children of all ages interact with speech recognition systems but are largely unaware of how they work. Teaching K-12 students to investigate how these systems employ phonological, syntactic, semantic, and cultural knowledge to resolve ambiguities in the audio signal can provide them a window on complex AI decision-making and also help them appreciate the richness and complexity of human language. We describe a browser-based tool for exploring the Google Web Speech API and a series of experiments students can engage in to measure what the service knows about language and the types of biases it exhibits. Middle school students taking an introductory AI elective were able to use the tool to explore Google's knowledge of homophones and its ability to exploit context to disambiguate them. Older students could potentially conduct more comprehensive investigations, which we lay out here. This approach to investigating the power and limitations of speech technology through carefully designed experiments can also be applied to other AI application areas, such as face detection, object recognition, machine translation, or question answering.

Introduction

This paper presents an approach to teaching K-12 students about speech recognition and the nature of language. Although most children today have experienced or at least observed speech recognition applications at an early age, few have any idea how this technology works. We present experiments that guide students to examine the Google Web Speech API, investigating its phonological, lexical, syntactic, semantic, and cultural knowledge. As an added benefit, the experiments help develop students' understanding of these linguistic concepts. We present a new open source tool for conducting these experiments, and describe results from a middle school AI class that used it.

Prior Experience with Speech Recognition

Children's first exposure to speech recognition may be via smart speakers such as Amazon Echo, Apple HomePod, or Google's Home and Nest products. These are now found in one third of U.S. homes (Brown 2019; Kinsella 2020). Intelligent voice assistants built into smartphones and tablets,such as Siri, Alexa, Cortana, or Google Assistant, are also likely to be familiar to children.

Students with Android phones likely first encounter speech *transcription* when making voice queries to Google. The query is converted to text and displayed in the search box. The same functionality is available in the Google Chrome browser on laptops and tablets. Transcription occurs in real time, while the user is still speaking, and one can sometimes observe revisions to earlier words in the transcript as more context is supplied. Speech to text functionality is also built in to Google Translate. The web version of this translation service also offers a text to speech option. On phones, Google Translate includes a live translation feature that transcribes an utterance, translates it, displays the translated text alongside the original, and then speaks the translation.

Another place where K-12 students can encounter speech to text software is in AI extensions to children's programming languages. Both MachineLearningForKids (Lane 2022) and Cognimates (Druga 2022) offer speech to text and text to speech add-ins that allow students to incorporate speech recognition and generation capabilities into their Scratch projects. Speech recognition and generation are also available in Snap!, a variant of Scratch, via eCraft2Learn (Kahn et al. 2022), and in MIT App Inventor (MIT App Inventor 2020) and Calypso (Touretzky 2017).

Google offers free access to their Web Speech API that enables applications running in the Chrome browser to incorporate speech input. They provide an online demonstration of the API where again one can see the transcription revised in real time as one is speaking (Google Research 2022). Numerous other online speech to text demonstrations are available, some of which include an option to read back the transcribed text. All these demonstrations share one limitation: they display only the highest ranked transcription, even when lower-ranked hypotheses are available.

The annyang JavaScript package (Ater 2016) provides convenient access to the Google Web Speech API for application developers. Most importantly, it returns multiple hypotheses in rank order when the speech signal is ambiguous. We rely on this feature in the SpeechDemo speech to text educational tool and exercises (Touretzky 2022a).

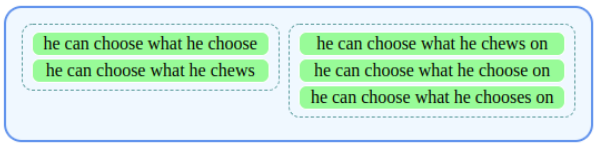


Figure 1: Candidate transcriptions for "He can choose what he chews" and "He can choose what he chews on" produced by the SpeechDemo tool.

The SpeechDemo Tool

SpeechDemo is an online speech to text demonstration that runs in the Chrome browser and uses annyang and the Google Web Speech API. It provides an option for readback of the top-ranked transcription, and supports multiple languages including Spanish, Arabic, and Chinese (Mandarin). It does not display revised hypotheses in real time, but instead waits for a pause in the speech and transcribes the entire utterance. Its most distinctive feature is that it displays multiple transcription hypotheses in rank order, which allows users to see how the speech signal is often ambiguous. The code is open source and has been placed in the public domain (Touretzky 2020). A teacher's guide for SpeechDemo (Touretzky 2022b) has recently been released as part of an AI4K12.org series of activity resource guides for K-12 AI teachers.

Figure 1 shows on the left the transcriptions for the utterance "He can choose what he chews," and on the right "He can choose what he chews on." For the former, the correct transcription is the second-ranked hypothesis, while for the latter it is the top-ranked hypothesis. Note that "what he choose" is not grammatical in Standard American English, but is acceptable in some dialects.

What Students Should Know About Speech

There are three fundamental things students should know about speech recognition. First, they should appreciate that language has multiple levels of structure, including phonological (sounds), lexical (words), grammatical (syntax), semantic (meaning), and cultural (idioms, sayings, famous quotations). Second, they should understand that the speech recognition process begins with a raw waveform and derives the final transcription through a series of transformations, guided at every step by linguistic knowledge. Third, they should be aware that because spoken language is both noisy and ambiguous, effective speech recognition must draw upon knowledge at all these levels to select the best interpretation of the signal from a list of plausible alternatives. This knowledge might be described in terms of discrete constraints and rules, but could also emerge from a complex statistical model.

The AI4K12 Guidelines for Big Idea 1: Perception state that *"The transformation from signal to meaning takes place in stages, with increasingly abstract features and higher level knowledge applied at each stage"* (AI4K12.org 2020). Guideline 1-B-iii introduces the "abstraction pipeline" for language, whose representations run from waveforms to articulatory gestures to sounds to morphemes to words to phrases to sentences. Progress through this pipeline requires phonological, lexical, syntactic, semantic, and cultural knowledge. Familiarity with the pipeline is the second of the three learning goals listed above. In early speech recognition systems these levels of knowledge would have been implemented as separate modules. Modern systems based on deep neural networks use more integrated representations, making it difficult to tease things apart *in silico*. Nonetheless, the distinctions between these types of knowledge are theoretically well-founded and remain important sources of insight into language. In this paper we show how students can investigate knowledge at each of these levels by experiment.

The speech understanding process is further elaborated in the AI4K12 Guidelines for Big Idea 4: Natural Interaction (AI4K12.org 2022). Guideline 4-A-i.K-2 on the structure of language asks students to generate plausible and implausible novel words. Plausibility is based on adherence to phonological rules, e.g., "flurg" is a plausible word in English while "fnurg" is not.

Guideline 4-A-ii.K-2 on the ambiguity of language asks students to give examples of homophones: words that sound the same but have different spellings. Students are also asked to demonstrate how the correct word of a homophone pair can be determined using context. Because K-2 students may not yet be proficient readers, the suggested activities in this grade band are usually designed to be done as "unplugged" activities (Bell and Vahrenhold 2018) rather than on a computer.

Guideline 4-A-iii.3-5 on reasoning about text asks students to experiment with an actual speech to text system to see if it can use context to resolve homophones correctly.

Homophones

Homophones are useful for probing speech recognition systems because they introduce ambiguities that can only be resolved by drawing upon various types of linguistic knowledge, and children are already familiar with them. They appear in guideline 4-A-ii.K-2, which is the youngest of the four grade bands.

The simplest homophone experiment one can do with SpeechDemo is to speak a single word and see the list of candidates produced. Figure 2 shows the results of three such experiments. The rank ordering of alternatives tells us something about Google's biases, which likely reflect the relative frequencies of these words in the training corpus: "break" occurs more frequently than "brake" and "night" more frequently than "knight" or "Knight". The list of results for the spoken word "led" is curious, because "Leed" and "lede" are normally pronounced with a long e as in "bleed", so they should not appear. But the idiom "bury the lede," when pronounced incorrectly as "bury the led," is recognized by Google with "lede" as the top candidate, so perhaps Google has mistakenly coded the short e version of "lede" as a variant pronunciation. These results may change over time as the system is further refined by Google engineers. Also, we found that the web version of Google Translate produces slightly different results than the web speech API.

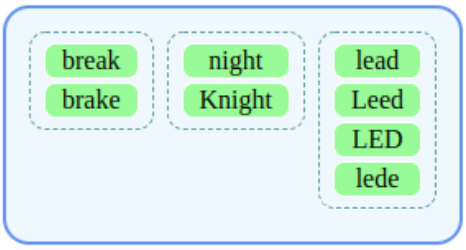


Figure 2: Recognition of homophones in the SpeechDemo tool.

Lexical Knowledge

Google's speech service only returns words in its lexicon; it does not attempt to transcribe non-words phonetically. But it appears to have a massive vocabulary, including archaic and obscure slang terms, acronyms, and many proper nouns. This can sometimes result in surprise phonetic transcriptions. For example, "blick" is not in the Scrabble dictionary (Merriam-Webster Publishing 2022) and is often cited as an example of a phonotactically well-formed non-word in English (Pimentel, Roark, and Cotterell 2020). But it is a proper name (e.g., "BLICK Art Materials"), has an archaic meaning referring to the gleam of gold or silver (OED Online 2018), and has several unsavory slang usages discoverable by a Google search. Likewise the ostensible non-words "flurg" and "plam" have slang meanings known to Google, and can appear in candidate lists returned by the web speech API, although even with careful enunciation it can be difficult to make them the top candidate. Thus, when prompting students to probe the speech API's ability to reject non-words it is important to first verify that those words are in fact unknown to Google.

One can also see semantics-like effects in Google's handling of non-words. One example is "brapes". Google hears "fruit salad with brapes" as "grapes", but "blue velvet brapes" as "drapes".

Google's lexicon also includes extensive knowledge of noun compounds, which can be demonstrated using homophones. Table 1 shows results from testing compounds beginning with break/brake.

Experiments also show that Google can recognize common mispronunciations, automatically correcting "expecially" to "especially' and "excape" to "escape". However, it is still possible to get "expecially' as a less-favored candidate transcription, even though it is widely (but not universally) recognized as a misspelling or mispronunciation, not a dialect difference. A Google search for "expecially" does return some documents, and even two YouTube videos on how to pronounce it. It's possible that a lexicon built by machine learning and not carefully curated may have incorporated a bad entry.

Phonological Knowledge

While the Google Web Speech API's vocabulary is extensive, it does not appear to include words that violate English phonotactic constraints. Thus, although "fnurg", "grimv",

Correct 'brake':	**Correct 'break':**
brake activation	break activity
brake actuator	breakdance / break dance
brake disc	break in
brake failure	break injury
brake handle	breakout / break out
brake job	breakpoint / break point
brake lever	break room
brake light	break schedule
brake line	break time
brake pad	break up
brake pressure	
brake shoes	
brake temperature	
brake wear	
Incorrect: "break." for "break period"	

Table 1: Google's success at disambiguating break/brake in noun compounds demonstrates extensive lexical knowledge. Results shown are the top-ranked candidates.

and "pwego" return results in a Google search, we have been unable to get them to appear in the candidate list shown by SpeechDemo no matter how careful the enunciation. Students may have trouble verbalizing their linguistic intuitions, but they know implicitly that /fn/ and /pw/ are invalid word-initial consonant clusters in English, and /mv/ is an invalid final cluster. Asking them to come up with their own invalid words and test their recognizability is a fun way of learning about phonotactic constraints and supports guideline 4-A-i.K-2.

Another way to probe Google's phonological representations is to see how it maps non-words to words. Both "troo-matized" and "tree-matized" are heard as "traumatized". And "me-nicipal", "my-nicipal", "may-nicipal", "mo-nicipal", and "moo-nicipal" are all heard as "municipal". The uncommon "-matized" and "-icipal" endings in these relatively low frequency words seem to drive the system to discount the first vowel no matter how far away it is from the correct pronunciation. In contrast, changing the initial consonant can be much more disruptive: "tunicipal" may be heard as "tunis Apple", "tuna Sippel", "two miscible", or even "two nipple", but is not corrected to "municipal".

Yet another way to measure the impact of phonological knowledge is to change the language model. The SpeechDemo tool offers recognition and readback in a variety of languages. At least in the US, when a non-English language model is selected the Web Speech API will still recognize English, or a mixture of English and the selected language, and read back the English parts with a noticeable accent. But the English recognition is substantially impaired. One can see this by speaking either the word "Kalamazoo", the sentence "I am going to Kalamazoo", or the sentence "I am going to Kalamazoo, Michigan". With the English language model all three are easily recognized. With the Spanish language model selected, only the last is recognized, and with the Chinese language model none are recognized, although other English sentences with more com-

mon words are recognized successfully. Reproducing these effects themselves can help students appreciate the contribution of phonological knowledge to speech perception.

Syntax and Semantics, Or Statistics?

Context can usually disambiguate homophone usage, and testing this by experiment satisfies guideline 4-A-iii.3-5. The SpeechDemo tool reveals Google's judgments to be less confident than a human's. For example, "Is it their dog?" produces all three candidates (their/there/they're), although "their" is ranked highest. "We walked their dog" produces two candidates, with their/there. But "We fed their dog" produces only the correct candidate.

When people are asked to explain how they disambiguate homophones they typically give syntactic or semantic explanations. For example, they prefer "he knows" but "his nose," and "we choose" but "he chews", because these are the only grammatical hypotheses. The SpeechDemo tool shows that "we choose" is recognized unambiguously by Google, but "he chews" generates two candidates, with the grammatically incorrect (in Standard American English) "he choose" actually ranked higher. So is Google really using grammar to disambiguate homophones? Note that the sequence "he choose" could be correct in some contexts, such as "will he choose".

Google has no trouble with the homophones in "which witch is which", but that well-known quote is likely recognized due to cultural knowledge, discussed below. "Witch" is a noun while "which" can be used as either an adjective or a relative or interrogative pronoun. "Which princess" ("which" used as an adjective in a noun phrase) is recognized as the sole candidate, while "the witch princess" ("witch" used as a noun in a compound noun) produces multiple candidates, but with "witch" correctly ranked higher than "which". Furthermore, for "witch *and* princess" Google correctly ranks the noun "witch" over the pronoun "which", while for "which and why" it correctly does the opposite.

Homophones that share the same part of speech cannot be distinguished grammatically but might be differentiated by meaning. The SpeechDemo tool shows that both "There is no air to breathe" and "There is no heir to the fortune" are recognized unambiguously by Google. One might be tempted to conclude that somewhere inside the Google speech recognition engine lies a semantics module with knowledge about air and breathing, and heirs and fortunes. In classic speech recognition systems this may well have been true. But that is not what is going on here.

Google's language services, including speech recognition and machine translation, are built from deep neural networks that were trained on massive corpora (Wu et al. 2016). These networks function as statistical models that combine phonological, lexical, syntactic, semantic, and cultural knowledge, without representing that knowledge symbolically in the form of phonotactic constraints, grammatical rules, semantic selectional restrictions, or idiom catalogs. But since human speech largely obeys these constraints by definition, a sufficiently detailed statistical model will mostly capture their effect and can even make reasonable predictions for novel utterances. Avoiding discrete levels of linguistic representation and replacing rigid grammatical and semantic categories with more context sensitive statistical correlations has greatly improved the performance of modern speech recognition systems, albeit making them harder to dissect.

Sentence	**Top ranked**
We drove their dog.	their
We drove their dog to New York.	**there**
We flew their dog to New York.	their
We sat their dog for Christmas.	**there**
We called out their dog.	their
We called out their hamster.	**there**
Is it their dog?	their
Is it their hamster?	**there**

Table 2: Inconsistent homonym resolution (highest-ranked candidate) using the Google Web Speech API suggests that it is relying on statistical inference rather than reasoning about meaning.

How then can we guide students to reveal the statistical nature of speech recognition? One way is to look at variability of results across sentences that might be expected from a statistical approximation but would be hard to explain in a system that was using classical linguistic rules. This point might be too technical for younger students to appreciate but should be accessible to upper level high school students. The sentences in Table 2 illustrate this variability in the resolution of there/their.

Cultural Knowledge

The highest level of knowledge contributing to speech recognition is cultural knowledge, which includes things like idioms, sayings, famous quotations, song lyrics, and literature in general. We can test the extent of Google's cultural knowledge by seeing if it is more likely to mis-recognize a word if doing so matches some well-known cultural reference. Consider the John Donne quote "No man is an island." We can show that Google knows this quote by comparing its responses to two variants where "island" is replaced by "eyelid":

- "No man is an eyelid" is mis-transcribed as *island*
- "Nomad is an eyelid" is correctly transcribed as *eyelid*

A similar effect can be observed with the archaic word "ere", which Google generally fails to recognize except in famous quotations. In "able was I ere I saw Elba," a classic palindrome, "ere" is recognized correctly, but in "able were you ere you saw Elba" it is heard as "are". Likewise, "We must away ere break of day" (a line from The Hobbit) is recognized, but the "ere" in "We must obey ere break of day" is transcribed as "are" or "air".

An even stronger effect can be seen with "three blind mice", title of a nursery rhyme. With careful enunciation, "four blind mites" can be recognized correctly, but "three blind mites" is impossible to transcribe; Google insists on correcting it to "three blind mice". Students should be encouraged to search for their own examples to probe

Google's cultural knowledge. This also raises the question of *whose* culture gets to influence speech recognition, as not all English-speaking cultures are likely to be equally represented in the training corpus.

Biases in Speech Recognition

Much has been written about how biases in AI systems can negatively impact members of minority groups. Particular attention has been devoted to facial recognition (Buolamwini and Gebru 2018) and automated decision making systems (Angwin et al. 2016; Corbett-Davies et al. 2016). But recently the topic of bias in speech recognition systems has also started to attract attention. A 2020 study of speech recognition performance across five major vendors showed higher error rates for speakers of African American Vernacular English (AAVE) compared with Standard American English (SAE) (Koenecke et al. 2020; Metz 2020). Given the growing prevalence of speech recognition technology in modern life, difficulty being understood can lead to feelings of "otherness" and not belonging among AAVE speakers (Mengesha et al. 2021).

While we have not attempted a rigorous investigation, an informal test of the Google Web Speech API seems to show that it has no trouble with AAVE syntax, but can be confused by phonological differences with SAE such as dropped consonants. For example, the top-ranked transcription for "He be eatin' wif his cousin" was "He be eaten with his cousin," which sounds similar but has a rather different meaning. The second-ranked candidate was "He be eating with his cousin", and the third was "He be eatin with his cousin". In general the system does not try to preserve dialect; it shows a strong preference for transcribing everything into formal SAE.

One place where this preference dramatically surfaces is in the pronunciation of "ask" or "asked" as "ax" or "axed", respectively. Although this pronunciation is a recognized part of AAVE, it is widely regarded by other English speakers, including some African Americans, as a mispronunciation associated with uneducated persons (McWhorter 2014). Our experiments show that in any grammatical context where "ax" could be interpreted as "ask", Google makes that substitution no matter how careful and deliberate the enunciation. This is an experiment that can easily be done in a K-12 classroom. It can lead to a nuanced discussion of how speech recognition systems should handle dialect. Users may have different preferences in different social contexts.

Middle School Experiment

Subjects

Here we report results from a nine-week middle school AI elective, "Living and Working with Artificial Intelligence," that includes a module on speech recognition. The module covers the audio abstraction pipeline from waveforms to spectrograms to phonemes and beyond. The observations come from two sections of the course taught simultaneously by the same teacher. Each section had roughly 25 students, all in grade 8, which comprises 13 to 14 year olds. One class period was devoted to the speech recognition module, during which students engaged in three activities. The first involved visualizing waveforms, the second involved spectrograms, and the third looked at recognition of homophones. Students worked in groups of 2 (or in a few cases 3) to complete a worksheet on homophone disambiguation while running the SpeechDemo tool on their Chromebooks.

The Worksheet

The worksheet used in this activity was designed by the teacher, who was participating in a year-long professional development program on how to teach AI to middle school students. Teachers were empowered to take resources we provided to them and create developmentally appropriate materials for their students. A sample worksheet is shown in Figure 3.

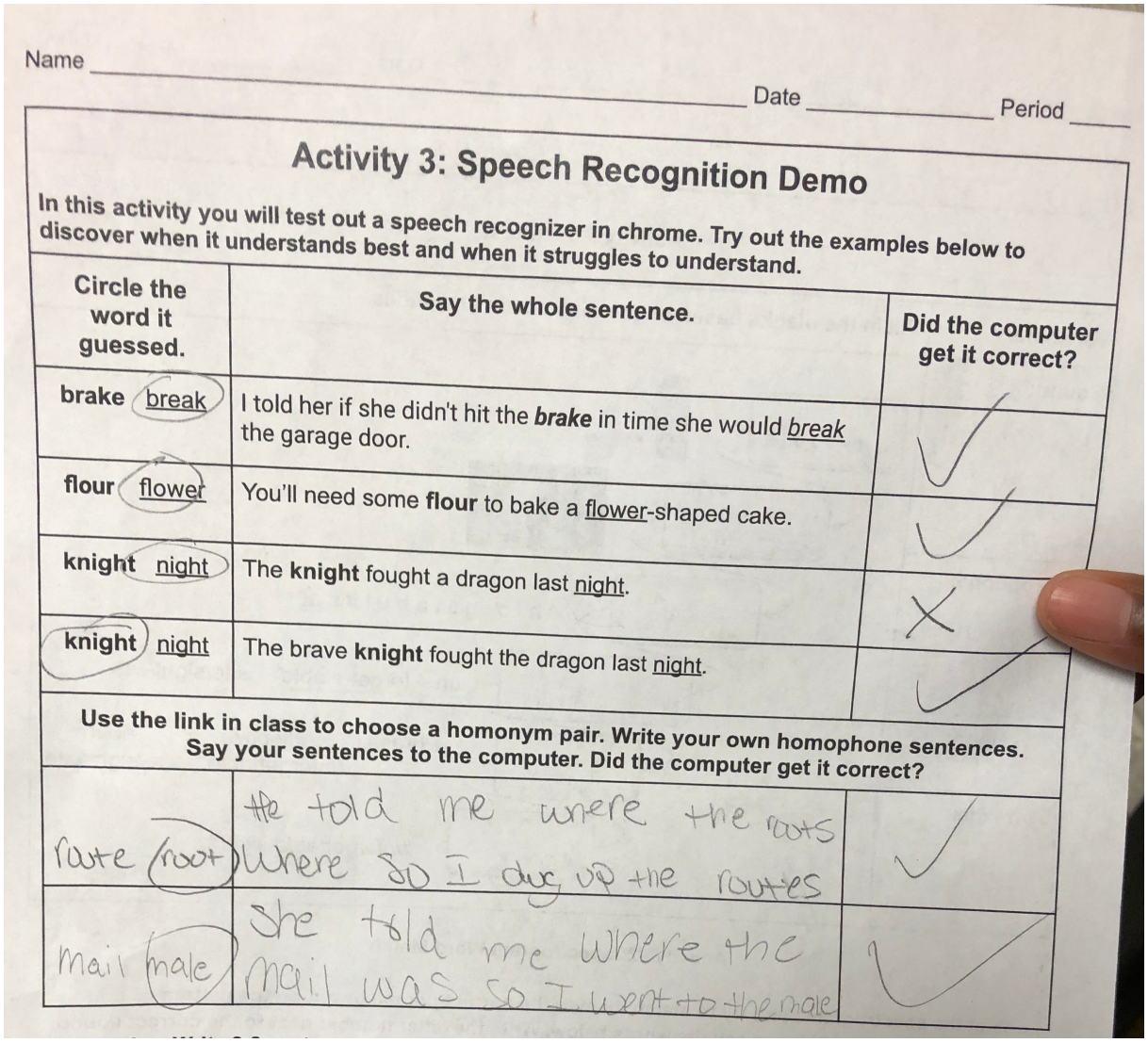
The worksheet was divided into two parts. In the top half students tested the speech recognition system's ability to disambiguate homophones by using the words and sentences provided. First they spoke a word in isolation, e.g., "break," and circled which of the two spellings break/brake the system picked as its first choice. To emphasize that the spellings were different, the first was shown in boldface and the second was underlined. Next the students spoke the associated test sentence, e.g., "I told her if she didn't hit the **brake** in time she would break the garage door." They then examined the top-ranked transcription candidate to see if both instances of the homophone were resolved correctly. They recorded the result in the third column by writing "yes" or drawing a check mark if it got both right, or writing "no" or drawing an X if it did not. Besides break/brake, the other homophone pairs used were flour/flower and night/knight. The latter pair was used twice; one test sentence began with "the knight" and the other with "the brave night". (Adding "brave" substantially improves the results.)

The second part of the worksheet asked students to make up two test sentences of their own using a homophone pair. They were provided links to resources with example homophones to choose from. They wrote the homophone pairs in the first column and the sentences in the second column. Students spoke their sentences to the SpeechDemo tool and recorded in the third column whether the words were transcribed correctly.

Implementation Issues

Having nearly a dozen teams of 2-3 students simultaneously speaking to their Chromebooks made for a noisy activity. The teacher anticipated this and physically dispersed the teams as much as possible, even putting two teams out in the hall. She also advised the students to lean in to their machines so the microphone could pick up their voice better. If extraneous sounds interfered with recognition they could repeat the trial as necessary. These measures proved sufficient.

Unfortunately, the mechanics of the worksheet did not work out as expected this first time around. In a retrospective interview several months after the activity, after we'd had time to review the collected data, the teacher remarked that her instructions for use of the worksheet were not as clear as they needed to be, and that the students didn't always follow



Name _____ Date _____ Period ____

Activity 3: Speech Recognition Demo

In this activity you will test out a speech recognizer in chrome. Try out the examples below to discover when it understands best and when it struggles to understand.

Circle the word it guessed.	Say the whole sentence.	Did the computer get it correct?
brake break	I told her if she didn't hit the ***brake*** in time she would *break* the garage door.	✓
flour flower	You'll need some **flour** to bake a flower-shaped cake.	✓
knight night	The **knight** fought a dragon last night.	✗
knight night	The brave **knight** fought the dragon last night.	✓

Use the link in class to choose a homonym pair. Write your own homophone sentences. Say your sentences to the computer. Did the computer get it correct?

route (root)	He told me where the roots where so I dug up the routes	✓
mail (male)	She told me where the mail was so I went to the male	✓

Figure 3: Student worksheet.

directions. Had they followed the procedure correctly, the circled word in the first column would be an indication of Google's biases about homonyms, and most likely would indicate relative frequency. Unfortunately, the data in the first column proved unreliable because some students went back and circled spellings that the system used correctly when transcribing the test sentence in the second column. Others circled the spellings that it got wrong. So it was not uncommon to see both words circled.

The second part of the worksheet also had issues. Of the two links provided for homophone resources, only the first contained true homophones such as male/mail. The second contained homonyms: words that have the same spelling but different meanings, such as "saw," which could be either a hand tool or the past tense of "see". Students who chose words from the latter list would not be able to test the system's disambiguation abilities because both meanings use the same spelling. Fortunately most students used true homophones, in some cases generating their own instead of choosing from the list of suggestions, and were therefore able to check spellings.

Results

Despite the issues described above, the teacher felt that the activity was very successful. Students were able to test which spellings the system favored, even if the record of those results was corrupted on the worksheet. They also understood the need to resolve ambiguity and could identify cases where the speech recognition system fell short, as in "The **knight** fought a dragon last night."

Students particularly enjoyed making up their own homophone-pair sentences in an attempt to "fool Google". And they understood that Google's only hope of resolving the ambiguities was by attending to the context in which

each word occurred. In evidence of this, in another part of the worksheet where students were asked to explain "how computers understand what we say," several cited "context clues", a phrase the teacher had used during the lesson. The reason more students didn't answer this way is that the day's activities also included examining waveforms and spectrograms; most responses focused on those.

How much did students retain from this activity? In a course review session 9 days later, students were shown four sentences and asked which one would likely cause trouble for a computer to understand. The sentences were:

1. He likes to sing with his friends on Thursdays.
2. Let's sit close to the front of class so that we can hear.
3. Although broccoli is yummy, she prefers spinach.
4. He said the sea was so clear that he could see the fish swimming.

Roughly 88% correctly chose the sentence with a homophone pair (sea/see), and they were able to explain why this made the sentence difficult.

We are making improvements to the activity worksheet and plan to use it in additional middle school classrooms this year.

Discussion

We've presented a lengthy series of experiments that probe the various types of knowledge the Google Web Speech API employs to correctly transcribe English. Some of these experiments are suitable for young children, while others are more appropriate for high school students. Our initial results with middle school students show that they are able to perform experiments with homophones.

We found that middle school students enjoy experimenting with AI systems to test the system's understanding and see if they can "fool" it. Engaging in the activities set out in this paper is also a way for them to learn about language itself—the first of our three learning goals—outside of traditional ELA (English Language Arts) instruction. And thinking about all the phrases Google can recognize can help them grasp the enormity of training sets used by neural language models.

For older students, the SpeechDemo tool's ability to show multiple candidate transcriptions, not just the top-ranked one, is important for appreciating the probabilistic nature of the speech recognition process. Thinking about the way the system weighs different sources of evidence when ranking transcription candidates gives students a window on complex AI decision making. Does Google truly "understand" the speech it transcribes, or is it merely applying a hugely complex statistical model? Its occasional lapses, as in Table 2, seriously undermine the illusion that it "knows" what we're talking about. Similar arguments have been made against other large language models constructed via machine learning (Marcus and Davis 2020).

One limitation of the SpeechDemo tool is its reliance on the current implementation of the Google Speech API. As Google continues to improve its speech recognition service, some experiments that reveal limitations of the service may yield different results; perfect reproducibility cannot be guaranteed in perpetuity. But experiments that demonstrate correct inference should continue to work as expected.

Our approach to teaching AI by "dissection" could also be applied to other AI technologies, such as face detection, object recognition, language translation, or question answering. The approach begins with selecting an AI system to examine and a theoretical framework for understanding its domain. In the case of speech recognition the theoretical framework is linguistic analysis. The framework posits certain representations or abilities the system must have, e.g., an understanding of phonotactic constraints, syntax rules, semantic relations, and so on. One then designs inputs to probe the system's behavior for evidence that it embodies these representations or abilities, as we have done here. In this way, one can gain some insight into the operation of the system without having to examine the actual algorithms or data structures, which are too complex for K-12 students to grapple with.

Acknowledgments

This work was funded in part by National Science Foundation awards DRL-2049029 and DRL-2048502.

References

AI4K12.org. 2020. Big Idea 1: Perception. https://ai4k12.org/big-idea-1-overview/. Draft version of May 18, 2020.

AI4K12.org. 2022. Big Idea 4: Natural Interaction. https://ai4k12.org/big-idea-4-overview/. Draft version of March 14, 2022.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed: 2020-09-10.

Ater, T. 2016. annyang! SpeechRecognition that just works. https://www.talater.com/annyang/. Accessed: 2022-09-02.

Bell, T. C.; and Vahrenhold, J. 2018. CS Unplugged - How Is It Used, and Does It Work? In *Adventures Between Lower Bounds and Higher Altitudes*, 497–521. Springer.

Brown, B. 2019. Alexa and Google Home smart speakers bring A.I. to nearly one in three U.S. homes. https://www.digitaltrends.com/home/alexa-and-google-home-smart-speakers-bring-ai-to-one-in-three-us-homes/. Accessed: 2022-09-04.

Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Friedler, S. A.; and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 77–91. PMLR.

Corbett-Davies, S.; Pierson, E.; Feller, A.; and Goel, S. 2016. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *The Washington Post*. October 17, 2016.

Druga, S. 2022. Cognimates. https://cognimates.me. Accessed: 2022-09-10.

Google Research. 2022. Web Speech API Demonstration. https://www.google.com/intl/en/chrome/demos/speech.html. Accessed: 2022-09-04.

Kahn, K.; Naveen, N.; Prasad, R.; and Veera, G. 2022. AI Snap! blocks for speech input and output, computer vision, word embeddings, and neural net creation, training, and use. https://ecraft2learn.github.io/ai/publications/EAAI-2022.pdf. Accessed: 2022-09-10.

Kinsella, B. 2020. Nearly 90 million U.S. adults have smart speakers, adoption now exceeds one-third of consumers. https://voicebot.ai/2020/04/28/nearly-90-million-u-s-adults-have-smart-speakers-adoption-now-exceeds-one-third-of-consumers/. Accessed: 2022-09-10.

Koenecke, A.; Nam, A.; Lake, E.; Nudell, J.; Quartey, M.; Mengesha, Z.; Toups, C.; Rickford, J. R.; Jurafsky, D.; and Goel, S. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14).

Lane, D. 2022. Machine Learning for Kids. https://machinelearningforkids.co.uk. Accessed: 2022-09-10.

Marcus, G.; and Davis, E. 2020. GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review*. August 22, 2020.

McWhorter, J. 2014. The 'ax' versus 'ask' question. *The Los Angeles Times*. January 19, 2014.

Mengesha, Z.; Heldreth, C.; Lahav, M.; Sublewski, J.; and Tuennerman, E. 2021. "I don't think these devices are very culturally sensitive." — Impact of automated speech recognition errors on African Americans. *Frontiers in Artificial Intelligence*, 4:725911.

Merriam-Webster Publishing. 2022. Scrabble Word Finder. https://scrabble.merriam.com/. Accessed: 2022-09-05.

Metz, C. 2020. There is a racial divide in speech-recognition systems, researchers say. *The New York Times*. March 23, 2020.

MIT App Inventor. 2020. MIT App Inventor Component Reference: Media Components. http://ai2.appinventor.mit.edu/reference/components/media.html. Accessed: 2022-09-05.

OED Online. 2018. "blick, n.". Accessed: 2022-09-09.

Pimentel, T.; Roark, B.; and Cotterell, R. 2020. Phonotactic Complexity and Its Trade-offs. *Transactions of the Association for Computational Linguistics*, 8: 1–18.

Touretzky, D. S. 2017. Computational thinking and mental models: From Kodu to Calypso. In *2017 IEEE Blocks and Beyond Workshop (B&B)*, 71–78.

Touretzky, D. S. 2020. SpeechDemo GitHub Repository. https://github.com/touretzkyds/SpeechDemo. Accessed 2022-11-27.

Touretzky, D. S. 2022a. Speech Recognition Demo. https://www.cs.cmu.edu/~dst/SpeechDemo. Accessed 2022-11-27.

Touretzky, D. S. 2022b. SpeechDemo Activity Guide. https://ai4k12.org/wp-content/uploads/2022/11/SpeechDemo-Activity-Guide-4.pdf. Accessed 2022-11-27.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; Klingner, J.; Shah, A.; Johnson, M.; Liu, X.; Kaiser, L.; Gouws, S.; Kato, Y.; Kudo, T.; Kazawa, H.; Stevens, K.; Kurian, G.; Patil, N.; Wang, W.; Young, C.; Smith, J.; Riesa, J.; Rudnick, A.; Vinyals, O.; Corrado, G.; Hughes, M.; and Dean, J. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. https://arxiv.org/abs/1609.08144.