愛媛大学教育学部紀要　第66巻　153~163　2019

Differences between Computer and Human Speech Recognition

-A Brief Look at Dictation on MacOS/iOS-

David R. BOGDAN（英語教育講座）

(Accepted 9/2/2019)

In this study, the author (DRB) is interested in the use of the speech recognition (Dictation) function in Apple's MacOS (and eventually iOS). He recently conducted a very quick machine versus human test which pitted the machine against several non-native EFL/ESL college students and himself. He discusses the results of this experiment in light of how computer speech recognition in dictation might be used as a teaching/learning aid for ELLs (English Language Learners).

This paper gives a very brief introduction to computer speech recognition and then goes on to discuss the test and its results. The test itself was run with only three (four, if you include the author) participants and therefore should not be considered definitive in any statistical sense, but rather should be thought of as a jumping-off point for further investigation.

MOTIVATION

Several years ago, the author, DRB, was advising an EFL/ESL graduate student working on his master's degree who was having difficulty in deciding on a thesis research topic. DRB suggested that he might look into the use of the built-in speech-to-text (STT) capabilities of MacOS and iOS as a dictation tool in EFL/ESL instruction and/or learning.

This is something that DRB had been interested in for sometime because he was using the built-in speech-to-text capability in iOS for his daily foreign language practice with Duolingo (a very popular language-learning platform). Specifically, he would use the STT function whenever Duolingo called for typed input, except in cases where there was no iOS keyboard available for a particular language and, therefore, no dictation capability.

The graduate student in question ended up deciding to investigate how to use speech-to-text in MacOS. He created specific tasks for students to complete which were supposed to help them to improve their pronunciation in English and to gain confidence in their speaking ability. One thing he kept running into during his testing, however, was the speech-to-text feature not picking up what the students were saying in English (or what they were trying to say).

DRB was concerned that this might have a detrimental effect on the students by actually causing them to lose confidence in their pronunciation ability. This concern led him to suggest that the device was probably less capable of dealing with the same amount of variation in pronunciation that an actual person was, and that, even if it (the computer) did not understand some things, a human being might do better. In light of this, he stressed that students should not be overly disheartened when the machine did not understand them perfectly, but should rather use that as a stimulus to try and work harder on their pronunciation.

SPEECH RECOGNITION

To understand a stream of speech sounds, one needs to recognize the phonemes and how they are strung together into morphemes, words, and longer utterances, There is a tremendous level of variation in pronunciation among different speakers and even with the same speaker. Because we are human, we do not pronounce the same thing exactly the same every time.

When we are processing an utterance, we need to be able to deal with that variance, and often we have to guess at what was produced based on our knowledge of what sequences or sounds are likely or even possible. Our internal grammar provides us with these rules.

In the case of utterances by non-native speakers, we often have to deal with even greater variation due to L1 interference or perhaps certain phonemes or sequences of phonemes not existing in their native grammar. In many cases, we are not accustomed to this degree of variation, which can lead to failures in comprehension.

We also have to deal with ambiguity. For example, in the three utterances below:

(1) I was surprised, too.
I was surprised to hear he was coming.
I was surprised two times during the speech.

we see the same pronunciation in three homophones: *too*, *to*, and *two*. By the linguistic context, however, we can guess which one it will be. A machine also has to be able to deal with this phenomenon when processing natural language.

For speech-to-text to work with any degree of accuracy on a computer, it has to be able to emulate a human speaker's ability to handle pronunciation variation, ambiguity, and other foibles of human speech.

Speech recognition (sometimes referred to as voice recognition) allows computers to translate spoken words into text, and "modern processes involve the use of two key methods: acoustic modeling and language modeling". (Thompson). The people at Microsoft Research (2004a) also point out that "speech recognition engines usually require two basic components in order to recognize speech." The first component is the acoustic model, which is "created by taking audio recordings of speech and their transcriptions and then compiling them into statistical representations of the sounds for words."

The language model, on the other hand, gives the probabilities of sequences of words. To quote from Microsoft Research (2004b) again, "language models help a speech recognizer figure out how likely a word sequence is, independent of the acoustics." This provides the computer with clues as to which words to choose when it encounters ambiguity in the acoustic data which has been processed by the acoustic model.

THE TEST

A mini experiment was run in August, 2019, following an event during summer vacation after which DRB asked some students to remain and, if they were willing, to take part in a small test. As this qualifies as low-risk linguistics research and was also completely a spur-of-the-moment exercise, consent from the participants was received orally, both before and after the experiment.

Prior to running the experiment, the author explained the goals and methodology of the test, that the participants would be recorded while participating and how the recordings would be used for this experiment, and that the recordings may be used in further research. They were all fine with the conditions laid out and kindly volunteered to participate. In the follow-up discussion, they also gave their consent to use and discuss the data they had provided. In addition, the participants are anonymized to protect their identities in this written report, excluding that of the author himself.

The participants

The three university students were all majoring in teaching EFL/ESL. One was a senior, one was a junior, and the third one was a sophomore, and they were all native speakers of Japanese. One was male, while the other two were female.

The author is a native speaker of American English with mainly Midland American dialectal characteristics (with some Inland North influences thrown in), although, with his extended experience abroad and exposure to foreign language and non-native pronunciation of English, some changes have no doubt occurred in his English.

Hereafter, the participants will be referred to only as *LL1*, *LL2*, *LL3*, and *NS*, with *LL* indicating the three respective language learner students and the *NS* representing the one native speaker of English, i.e., the author himself.

The overall environment

The experiment took place in the ICT (Mac) lab in the Faculty of Education. The author has used the room for his ICT classes over a long period of time and has had problems with the acoustic conditions because some of his classes involved creating video (including audio) teaching/learning materials. Located in that room is a large box that handles the LAN connections for the entire floor which has a very loud fan for cooling off the equipment. There is a containment box around this equipment which should, in theory, dampen the sound from the exhaust fan, but this has never been the case, and the acoustic conditions have always been, to put it bluntly, atrocious.

To make matters worse, just a month prior to the experiment, the equipment in this box was "upgraded". Not only did the fan become louder, but the door to the "sound-dampening" box no longer closed all the way, resulting in even worse sound pollution. The author's office is soundproofed to a certain extent, but it is not large enough to accommodate all the test participants at one time, so the ICT lab was used.

The lab is equipped with 16 fairly new iMacs (2017), and, while their internal microphones are quite decent, they do tend to pick up background noise. Therefore, an external microphone was used, which did seem to work somewhat better than the internal microphone.

The informant dictating the text was seated before one of the iMacs which was facing away from the previously-mentioned LAN equipment container and also away from the "listening" participants who were to transcribe the utterances. While the view of the speaker was not completely blocked, having the iMac between the speaker and the listeners effectively prevented them from being able to see the face of the person speaking. (When talking about acoustic/auditory conditions, it should be noted that DRB was just recovering from middle ear infections in both ears, which meant his hearing ability was somewhat impaired.)

The language material used

The author provided the participants with three different EFL/ESL textbooks—*Power On English Communication I* (Asami, 2019), *All Aboard! English Communication I* (Kiyota, 2019), and *Prominence English Communication I* (Tanabe, 2019)—all published in Japan and used for teaching sophomores English at Japanese high schools.

The original plan had called for the students to randomly choose three sentences, of varying lengths, from the textbook they were provided with, write down those sentences while making sure none of the other participants (including DRB) saw them, and then use them for the dictation exercise. As the students were doing this, however, DRB quickly realized that it would take too long to run the experiment and would also make things unnecessarily complicated, and he revised his instructions to where each student was asked to find only two sentences: one somewhat shorter, and the other, of a more respectable length.

They could select any sentence from the reading passages or dialogs in the textbooks with the following proviso. They were asked to avoid sentences that contained foreign personal names and place names because this might throw off the computer with items that were not in its lexicon database. After they had chosen and written down the sentences that they were going to use, DRB also chose two sentences that he would use when it was his turn to dictate. He selected his sentences from one of the textbooks used by the students, taking care that he chose from pages other than those that the student had used. Again, because this was a dictation exercise, none of the participants, including DRB, were to know in advance what the sentences of any of the others were.

The equipment and software

As mentioned above, a 2017 iMac was used to conduct the speech-to-text dictation exercise discussed here. Also as noted above, an external microphone was used rather than the built-in internal microphone on the iMac. The microphone used was the Fifine K669 Podcast Condenser Microphone. This is an inexpensive USB microphone which records in a cardioid polar pattern.

Only two applications—both included on the Mac—played a role in the test; TextEdit was used in conjunction with the dictation function to produce the written text, and QuickTime Player simultaneously made audio recordings of the speech production. The iMac had been set for enhanced dictation during a regular class in the year prior to the experiment. DRB had not himself used that particular Mac for dictation, nor had any of the three students, which meant that, if the dictation function somehow uses some sort of learning to adjust to a particular person's speech and pronunciation, it should not have occurred on this particular machine.

The experiment itself

First, one of the students acted as the speech production informant by just sitting in front of the Mac and reading their first sentence into the microphone. They then waited while the other three participants, including DRB for the three times he acted as a listener, attempted to write down what they had heard. When the listeners had all raised their hands indicating that they were finished, the speaker would then read the sentence a second time. Following this, they proceeded on to their second sentence, again repeating it a second time. This activity was conducted by the three remaining participants in turn, finally concluding with DRB acting as the speech informant.

The procedure was actually quite simple, although, with the first student, DRB found himself running back and forth between the utterances being produced in order to check whether everything was recording and transcribing properly (making sure, of course, that he did not see what the student had written down or the actual transcription produced by the dictation function).

He also checked to see that the dictation remained switched on, being worried because the students had had no experience in using that function. Also, quite foolishly in retrospect, he had initially thought to make a separate recording for each utterance. Subsequently, however, he realized that it would be simpler just to keep QuickTime running throughout the entirety of each participant's turn and then do the splicing and dicing at his leisure during post-editing.

Following this, all of the participants gathered around the computer to find out what the actual sentences were and how their transcriptions compared to that of the machine's transcriptions and then to discuss their impressions.

The data

In (2) – (5), we see the original sentences used for the dictation exercise in the order in which they occurred.

(2) LL1-1: Some animals sleep longer than others do.
LL1-2: Some ocean animals, such as whales, have to swim to the surface for air.

(3) LL2-1: What course do you take at your school?
LL2-2: During their school life, the students often face the lives and deaths of their animals.

(4) LL3-1: I had no interest in trying the phone in our new house.
LL3-2: When I was very young, my family had one of the first telephones in our neighborhood.

(5) NS1: By deep frying noodles, they become hard and dry.
NS2: The angels celebrate the birth of [Jesus] Christ by singing and playing musical instruments.

Again, the *LL* represents each of the three language-learning students, while *NS* is the native speaker of English. Also, as mentioned earlier, the first and second sentences for each informant differ in length, with the second one being longer and more complicated.

RESULTS

In this section, we give the resulting transcription attempts for the data seen above, beginning with the computer's version, followed by those of the human listeners. The underlined blank sections indicate places where the listeners knew something had been said, but could not figure it out well enough to even make a stab at it.

The first line in (6) – (9), respectively, indicates which of the human participants is acting as the speaker. Following this, the data for each speaker is divided into two sections, reflecting the short and long sentences. In each section, we first see what the speaker tries to dictate in italics, followed by the attempts at transcription by all of the remaining participants, including the computer. The computer's attempts come first, then DRBs (except for in the last session, where he was acting as the speaking informant), followed by those of the listening student language learners.

(6) LL1-1
Some animals sleep longer than others do. (Original)

C-a:: I must say wrong about us do Sam I'm saying
C-b:: Somebody Mars Street on the other of us do
NS-a: Some animals _______ longer than others do.
NS-b: Some animals live longer than others do.
LL2a: Some animals sleep longer than other species.
LL2b: Some animals live longer than others, too.
LL3a: Some animals _______ sleep.
LL3b: Some animals sleep no longer than s_______.

LL1-2
Some ocean animals, such as whales, have to swim to the surface for air. (Original)

C-a:: Some astronomers such as bears have to seem to have faith for air
C-b:: Ocean such as bears have to seem to have phase for air
NS-a: Some _______ animals_______
NS-b: Some sea animals, such as whales, have to swim to the surface to breathe.
LL2a: Some ocean animals, such as whales have to swim _______
LL2b: Some ocean animals such as whales have to swim to the surface for the air.
LL3a: Some animals _______ sleep.
LL3b: Some animals sleep no longer than s_______.

(7) LL2-1
What course do you take at your school? (Original)

C-a: What colors do you take at your school
C-b: What course do you take at your school
NS-a: What course do you take at your school [courses?]
NS-b: What course do you take at your school
LL1a: What _______ do you take at a school.
LL1b: What course do you take at your school.
LL3a: What color do you take at the school?
LL3b: What course do you take at your school?

LL2-2
During their school life, the students often face the lives and deaths of their animals. (Original)

C-a: During their school life the students often freeze dog lives and death of their I'm else
C-b: During their school life the students often face lives and this is of their animals
NS-a: During their school life, the children often _______ the life and death of their animals.
NS-b: During their school life, the students often face the life and death of their animals
LL1a: During the _______ the students _______ of their animals.
LL1b: During their school life, the students _______ lives of their animals.
LL3a: During their school lifes, _______ their animals.
LL3b: During their school life, the students often live(s) their animals

(8) LL3-1
I had no interest in trying the phone in our new house. (Original)

C-a: I have no interest in trying the phone eat our new house
C-b: I have no interest in trying to cell phone eight our new house
NS-a: I have no interest in _______ into our new house

NS-b: I have no interest in trying _______ in our new house
LL1a: I have no entrance _______ phone _______ in our new house
LL1b: I have no entrance _______ phone _______ on our new house
LL2a: I have no entrance _______ new phone
LL2b: I have no entrance and trying toward the new phone in our new house

LL3-2
When I was very young, my family had one of the first telephones in our neighborhood. (Original)
C-a Bananas very young I find a hot while I'm at the foster funds eat our neighborhood
C-b When I last Verianc my family hot while I was in foster phones in our neighborhoods
NS-a: When I was very young, our family had one of the very first telephones in our neighborhood.
NS-b: When I was very young, my family had one of the very first telephones in our neighborhood.
LL1a: When I was young, _______ family _______ in our neighborhood.
LL1b: When I was very young, my family had one of the first telephones in our
LL2a: When I was very young, my parents _______ in our neighborhood.
LL2b: When I was very young, my family had one of the first phone in our neighborhood.

(9) NS-1
By deep frying noodles, they become hard and dry. (Original)
C-a Jo buy the frying noodles they became hard and dry
C-b Buy deep frying noodles they became hard and dry
LL1a: By the flying, _______ became hard to dry.
LL1b: By the flying, noodles became hard to dry.
LL2a: By deep frying noodles, _______ became _______
LL2b: By deep frying noodles, they became hard and dry.
LL3a: By deep flying _______
LL3-b: By deep flying noodles, they became hard and dry.

NS-2
The angels celebrate the birth of (Jesus) Christ by singing and playing musical instruments.(Original)

C-a: The Angels celebrate the birth of Jesus Christ by singing and playing musical Yzerman
C-b: The Angels celebrate the birth of Christ by singing and playing musical instruments
LL1a: The engles _______ some _______ by singing and playing musical instruments.
LL1b: The engles _______ some of birds to cry by singing and playing musical _______
LL2a: The angels celebrates Jesus by singing and playing musical instruments.
LL2b: The angels celebrates the birth of Christ by singing and playing musical instruments.
LL3a: The engles celebrate the _______ music instruments.
LL3b: The engles celebrate the birds of price by singing and playing music instruments.

SOME OBSERVATIONS

This is a very small database involving an extremely limited number of participants, which precludes drawing far-reaching conclusions. However, some interesting observations are possible.

In looking at the results in the section above, it appears that the native speaker beat out the computer in five out of the six non-native utterances, at least once he heard the sentence a second time. The exception was for the shorter sentence in (8), in which for some reason (perhaps the middle ear infection was affecting his hearing), he could not pick out the word *phone* in either of the utterances and left that portion blank.

All of the other participants, including the computer, did pick up on the word *phone*. However, the computer did come up with some rather confusing collocations by using *eat* and *eight* in the sentence. DRB wondered whether the informant might have mistakenly substituted *at* for *in* in his utterances. Upon listening to the audio file, however, he ascertained that the word used was, in fact, *in*, although with a pronunciation that was perhaps slightly different from that of a "typical" native speaker.

On the other hand, the computer seemed to fare somewhat better when competing with the non-native ELLs in understanding DRBs utterances, seen in (9). The apparent mystery of the *Jo* at the beginning of the first transcription of the shorter sentence can possibly be explained by listening to the corresponding audio file.

In his instructions to the students immediately prior to uttering that sentence, DRB says *Yarimasu yo* "I'm going to start", and, in the confusion between what is Japanese and what is English, the *yo* may have been added to the actual sentence as *Jo*. The computer has to distinguish between what are and are not legitimate speech sounds (phones or phonemes) particular to the language it has been set to.

Otherwise, the two iterations of the shorter sentence were processed nearly flawlessly, except for the use of the homophone *buy* for *by (*and the *the* added after the *Jo)*. Early in the paper, we saw that the computer needs to deal with such ambiguities through a knowledge of collocational possibilities and restrictions. Obviously, it failed in both attempts here. A human is unlikely to make the same mistake owing to their internalized grammar, and the non-native speaker participants were aware of this infraction of grammar rules, for none of them mistook *by* for *buy* in this context. On the other hand, two of the three NNSs (non-native speakers) used *flying* instead of *frying*, stemming no doubt from the lack of a distinction between *l* and *r* in their L1, Japanese.

One would hope that an equivalent to the language model used in speech recognition for computers (if it could be applied to humans) would forestall this mistake occurring, but Japanese speakers will still occasionally make similar errors. DRB has seen it many times throughout his teaching career in Japan and can even remember how surprised he once was to see a large permanent sign on the wall of a supermarket informing shoppers that they were at the *Flesh Meat* section. All learners of any foreign language seem bound to make common errors owing to L1 interference: the author being no exception. The computer, without such L1 interference to impede it, did not commit the same error in this instance.

There were also other areas where the difference between native and non-native grammar came up. For instance, in (9), we see problems with number agreement with *the angels celebrates,* and, in (8), we see a similar problem with number in *one of the first phone*. These are mistakes that you would not expect to see coming from a typical native speaker; owing to their internalized grammar (the Language Model, for computers) they would normally add the plural *s* even if they did not hear it. There was also the plural *lifes* in (7): a typical non-native speaker error.

In discussing this in the after-action review, the reaction was that it should be treated as a very good learning experience, so this is actually evidence of how using the dictation function for teaching/learning can be quite useful.

In talking about number, also in (7), you can see where DRB had put in *courses?* in his first rendering of LL2's shorter sentence. This was because, contrary to what his own ears were telling him, he felt that the plural would be more logical, as in *What courses do you take at your school?* However, the second rendering convinced him that his ears were not, in fact, deceiving him.

It was interesting that, in (9), the computer came up with *Yzerman* instead of *instruments* in the first rendering of DRB's production of the longer sentence. Perhaps it is a hockey fan, for an Internet search produces numerous hits for Steve Yzerman, a Canadian-American ice hockey former player and executive.

A review of the audio files forces the author to admit here that he made an embarrassing mistake in his reading of the longer sentence. The first time, he correctly said *Jesus Christ*, while for the second, he had left out the word *Jesus*. Both the computer and LL2 picked up on that, although the latter had left off the word *Christ* in their rendering of the first reading. While this episode was embarrassing for DRB, he is thankful he had the foresight to record the audio files because they came in very handy in clearing up discrepancies in the transcriptions.

An interesting pattern—if one can accept that a pattern may be formed with this limited data set—can be seen in how humans and machines differ in speech recognition. The human participants, the author included, often left blank the sections of the speech stream they did not understand, while the computer just spewed out some sort of text, no matter how crazy. This may have something to do with memory span; the computer does not have the limitations that humans do and can go ahead and process what it has "heard." Of course, another factor may be that, in many cases, the human participants were consciously aware that they would get a second crack at transcribing the sentence, and so it seems to be a perfectly reasonable listening strategy to leave such portions blank. This would not, of course, account for the use of blank areas in the second rendering of sentences.

CONCLUSION

Again, any far-reaching pronouncements are impossible here, but a cursory look at the data suggests (or at least does not contradict) that one can infer that computers process spoken language differently than humans. Interestingly, these differences differ in different ways when comparing the computer with native speakers and then with non-native speakers. Of course, between the two groups, there are bound to be differences.

Future investigation is warranted because the speech-to-text capabilities keep improving at an astonishing rate, and we may find even an average home/school computer passing the Turing test sometime in the not-so-distant future with respect to speech recognition. This suggests that this function shows great promise in becoming a very viable aid in foreign language teaching/learning.

ACKNOWLEDGEMENTS

I would like to thank the three undergraduate students for volunteering to participate in this off-the-cuff experiment. They were fantastic!

Also, I would like to express my appreciation to my colleague at the university, R. Vergin, who has always been willing to take a look at my fractured English. However, I would like to stress, under no uncertain terms, that responsibility for any errors is mine only.

Thanks are also owed to another of my colleagues, D. Tatematsu, who kindly loaned me the three English textbooks which were used in the experiment.

REFERENCES

Asami, Michiaki. (2019). *Power On English Communication I*. Tokyo: Tokyo Shoseki.

Kiyota, Yoichi. (2019). *All Aboard! English Communication I*. Tokyo: Tokyo Shoseki.

Microsoft Research. (2004a). "Acoustic Modeling". [Online] https://www.microsoft.com/en-us/research/project/acoustic-modeling/.

Microsoft Research. (2004b). "Language Modeling for Speech Recognition". [Online] https://www.microsoft.com/en-us/research/project/language-modeling-for-speech-recognition/.

Tanabe, Masami. (2019). *Prominence English Communication I*. Tokyo: Tokyo Shoseki.

Thompson, Tyler, Kieran Harrigan, Matthew Wong, and Patrick Liao (Team 13). "Voice and Image Recognition AI" Past student presentation listed in the syllabus for the Fall 2019 *CSE352 Artificial Intelligence* course at Stony Brook University. [Online] https://www3.cs.stonybrook.edu/~cse352/T13talk.pdf.