FOR DOCTORS AND JOURNALISTS, IT'S BEEN AI'S HOLY GRAIL

TO BE ABLE TO RECORD AN INTERACTION AND LET A MACHINE TRANSCRIBE IT HAS LONG BEEN THE DREAM. IT'S FINALLY HERE—ALMOST.

DAVID HECHLER

> *have spent most of my career as a journalist. The information a reporter gathers mostly boils down to documents and interviews. And many interviews are conducted on deadline, so you're writing or typing as fast as you can—hoping what you're getting is accurate. But sometimes for a feature article, or one that requires more depth, you want to record the conversation to capture it all. But that means you have to spend a lot of time transcribing. Or at least you did until recently.*

For many journalists the Holy Grail has long been speech transcription software that could save us the time and drudgery of transcribing recorded interviews ourselves. And I, for one, was happy to celebrate when it appeared that advances in artificial intelligence had finally produced speech recognition software that worked. But like so many "saviors" that promise a whole new way of life, it has not delivered all it had promised. At least not yet.

I will show you just a few examples of what I mean in this article, which was transcribed by the program I use—and required hours to fix. We've **[highlighted in brackets]** *the transcript's errors to make them easy to see.*

But I'm not here to kvetch. I'm happy to have it. It's just that when an expert on this subject joined our company, I couldn't resist asking some questions. I wanted to understand the challenges software developers face in creating the kind of product I want. That's why I asked Jay Wilpon, one of the true pioneers of speech recognition technology and now TAG's AI expert (see the article by him in these pages), to sit for this interview. I was hoping he'd tell me that perfection is just around the corner.

David Hechler: Jay, have you ever used software to transcribe an interview or a speech?

Jay Wiipon: Yes. I use it all the time for first drafts of articles and to dictate text messages. The accuracy is imperfect. It's good enough to be able to use to get my thoughts down quickly. And I like talking better than typing.

Hechler: Have you done work professionally related to this field?

Wiipon: Yeah, I started working on speech recognition before there was speech recognition—back in the mid 1970s. In 1977, I joined AT&T Bell Labs and helped create a lot of the technology that led to the modern dictation software that you're using today.

Hechler: I noted that the company Dragon, which I'm going to refer to, was founded in 1975, although Dragon NaturallySpeaking didn't come out until the '80s. So we're talking about a similar time frame. When I became aware of this technology, I remember that doctors often recorded their patient notes on Dictaphones. And then they gave their tapes to an assistant to transcribe them. I'm sure

those doctors were eager to be able to avoid paying assistants to type their notes. They were probably some of the early people eager for the technology we're talking about. And then, as a journalist, I was constantly looking for a product that would obviate my need to sit down after a recorded interview and type away. And then have to play back, and rewind, and listen to it over and over—to save me all that time and spare me that drudgery. And Dragon NaturallySpeaking was the first one that got good reviews, that sounded like maybe this is possible. But it was, in its earliest form, designed for one speaker, to dictate kind of like the doctors did. And that person dictating had to devote quite a bit of time to train the software to recognize his voice, his inflections, his accent. So the individual had to do a lot of work just to have it be somewhat accurate. It was clear to me that this wasn't what I was looking for. It wasn't for an interview, because I would have had to train it myself. And then it wouldn't have been trained for the person I was talking to. And since that person was changing all the time, it didn't seem like it was ever going to work for me. From what I read it did get better over time, but it wasn't close enough for me to decide to buy it. Because that's what I would had to do: spend several hundred dollars and then try it out. Does this match your recollection of the slow progress on the way to getting technology that could transcribe interviews?

Wiipon: You gave a good summary of the challenges that were in place almost 40 years ago. The first dictation product probably came out in the early 80s. Yes, Dragon also came out around that time. There were a number of companies that were offering dictation software. And one of the variables you actually missed. For the early dictations, you… had… to… speak… one… word… at… a… time. Not only did you have to train it to understand you, but you had to speak in that unnatural way.

Hechler: Yes, I remember that now. It seems that there was mutual training going on. There were a lot of ways in which the technology trained the human to behave a certain way. What you just said is going to be hard for me to transcribe because the software isn't going to know how to add those pauses you inserted a minute ago. I'll have to add dots between the words.

Wiipon: The point I was making was that being able to recognize what before was sort of isolated speech, one word at a time, to be able to move to continuous speech was a scientific achievement. And it took a a decade once we could recognize isolated speech reasonably well, to be able to recognize continuous speech very well. The second variable, you've got to look at at the computing capabilities. You had kilobytes of memory, or megabytes maybe back in the '80s. Even if you could train something, even if you had the data, the machines couldn't physically handle it in real time. So everything always started very small: 10 digits, alphabet, the dictation piece you talked about with doctors, especially radiologists. They were one of the first users of dictation engines. And it may sound like that was hard because they were saying a lot of polysyllabic words. But radiologists have a very short vocabulary. Most of it is boilerplate. They say one or two words and it generates a paragraph. So actually, it was very useful for them to increase their efficiency.

YOU DON'T WANT TO LEAVE A DEATH PENALTY CASE TRANSCRIPT IN THE "HANDS" (QUOTE, UNQUOTE) OF A MACHINE. AGREED?

Additionally, things like Dragon—you said something about hundreds of dollars. When Dragon came out in the '80s, it was around $10,000. Who else was going to be able afford that besides doctors? So it wasn't a cheap piece of software. A lot of research went into it. Research that the enterprise private sector did, a lot that the government sector did. Dragon was partially funded by DARPA. And DARPA allowed people to form their own companies. And the Bakers [Dragon's founders] did that with Dragon. What changed the equation on the price was IBM coming out with a dictation product in the '80s for a few bucks. Then in 1999, IBM released its ViaVoice program for $200. And so literally overnight Dragon had to drop their price to be competitive. Dragon wasn't long for this world. There were a bunch of mergers in the speech industry. The company eventually merged into Lernout & Hauspie and then ScanSoft, which eventually merged into Nuance.

Hechler: So it was only within the past 10 years that this has changed substantially, at least in my experience. You've now got Otter.ai, more recently Google Live Transcribe, Microsoft Word Dictate transcribes, and most recently OpenAI Whisper has joined the party. I haven't tried them all. But they seem to be much better. You agree?

Wiipon: They're infinitely better for a few reasons. One, 20 years ago the vocabulary sizes were still very small. Hundreds of words, maybe thousands of unique words. And you had to train it, and you had to know the accents. Now, speech engines can pretty much recognize any word that you could say, and it's easy to add new vocabulary. You can also speak normally. You can speak in a crowded environment and it will still pick it out based on newer algorithms and signal processing. It's like night and day. It's like the caveman inventing the wheel versus having a Tesla. It's completely usable, except for things that are critical to my life. I'm not going to use it to do something that could affect the safety of my family, for example.

Hechler: That's a perfect segue to my next question. To add context, I don't think we've ever seen machines relied on to transcribe testimony in court. Humans are required for that work—they're called court reporters—even where courtroom testimony is recorded electronically. I expect that will continue for a long time, because accuracy is so important. You don't want to leave a death penalty case transcript in the "hands" (quote, unquote) of a machine. Agreed?

Wiipon: I think it's going to be that way for a while. It may be the case that the machine can transcribe it better. But right now it's not allowed. The law is the human sitting there with their transcriber machines. I think that will change over time. It's an obvious evolution.

Hechler: Where accuracy is *not* so important, transcriptions have crept into the culture. When you get a voicemail on your iPhone, you don't have to listen to it. Usually, there's a transcript you can read instead. Zoom meetings like the one we're on now can be recorded and transcribed simultaneously. So what do you think about these developments? And are there other examples you want to throw in?

Wiipon: Those things will continue to happen. Another place, to continue the media theme, is television shows, movies. The accessibility acts [under the **ADA**] requires closed captions. The funny thing is that up until maybe 10 years ago, things like YouTube didn't require that because they weren't broadcast over the waves. But now you see that if you go on YouTube, you can get transcriptions in pretty much real time of anything you see there, because Google can put their speech engine on there, collect the data, train it, and build great models. So, from an accessibility point of view, for people that are disabled—have a problem hearing—it's a boon to be able to apply this technology to make their world much better. It's happened pretty much organically. Because once technology gets good enough, people will use it.

But there's another aspect of this technology we should discuss. Even in the early days of dictation, the companies you mentioned—Dragon, IBM—the thing that distinguished their products more often than not wasn't their accuracy. It was the user interface that allowed users to easily dictate and easily edit on the fly, go back, change a word. If they weren't sure of a word, the recognizer might highlight it with an underline that you could go back to. So a lot of what made these companies differentiate themselves was the user experience they provided to enable people to use a flawed technology.

Hechler: Let me turn to my own experience using this software. For my work, it is essential that the transcript I get of an interview like the one we're doing right now is accurate, because we're going to publish it. That means I have to rigorously review the text against the recording. I'm always disappointed by the results. My only consolation, and it's substantial, is that I remember how much work it was when I had to transcribe it all myself. There have been times I have spelled out words or acronyms in the recording, and the program still got them wrong. Here's an example. It's not so much spelling it out as using an acronym. When I refer to a company's information technology department as the IT department, IT is too often lowercased. It can be pretty confusing when someone says something like, "I called IT and asked what the problem was." And then it comes out, "I called it and asked what the problem was." **[In the example above, the only time "IT" was capitalized in the transcript was the first time I used it.]** Have you run into these kinds of problems over the years in your work?

Wiipon: Yeah, this gets to language. Speech recognizers are there to listen and transcribe. They don't know about punctuation. They don't know about accents. They don't know about pauses. They don't know about the things that you're talking about, like acronyms. They have to learn those things. Up until a few years ago, you didn't get punctuation when you dictated. You just got a long string of words. If you want to send a text message and use dictation—even now, I say something and I add "period" or "comma," because it doesn't know where to put any kind of punctuation. The difference between dictation, which is just recognizing the words without meaning, and more advanced forms, which I believe the modern dictation engines are moving towards, is to understand a little bit of the meaning.

Hechler: We'll come back to grammar. But first I want to say that I recognize jargon is likely to confuse the program. But I expected it to improve over time, and it hasn't. I interview people about chief information security officers regularly. Really **[the program rendered this word as "Billy"]** all the time. The acronym is [spelling it out] CISO. And I pronounce it *seaso*. But sometimes I will actually spell it out. It doesn't even get all those letters right when I spell it out. So it seems clear to me that there is no

machine learning going on with the product I'm using. My subscription plan allows me to teach the program five words or phrases per conversation. That's it. I presume it won't remember them the next time. It's very frustrating. Isn't there a better way for the company to handle this? Can't they build in some form of machine learning without requiring payments and my having to train it? This takes me back to Dragon NaturallySpeaking.

Wiipon: Well, it's always the case that things advance forward slowly. Once things become good and ubiquitous, then you make it part of the product. If there's things that everybody's got, you make them available, and you differentiate yourself with other things. I don't want to second guess what people's business models are for generating revenue, but there's a reason that you have a bunch of software that's open source. And there's a reason that there are tools which allow companies to be able to customize AI, in this case speech recognition, so that there's a business. But over time, it does improve.

Hechler: Yes. That's the reason I'm using this software now. It is so far better than anything that used to be available. I want to be clear that once I saw what it could do, I didn't hesitate to start using it. That had never happened before. And I'm not unhappy that I did. It has and does save me an enormous amount of time. It's just that I really expected that there would be improvements—either from what the company was able to offer, or that somehow it would learn from me, from my work.

IT'S NOT BUILT TO WRITE. [TRANSCRIPT: "IT'S NOT BUILT TO RIGHT."] IT'S BUILT TO RECOGNIZE THE WORDS YOU SAID. THINGS LIKE PUNCTUATION ARE PART OF MEANING.

Wiipon: How do you know it's not? It definitely is. [This came out: "I know it's not. It definitely is."] Take Siri. Let's say you've got five Jays in your contact list. And the first time you say, "Call, Jay" it may say, "There's five Jays, which one do you want?" After you've done it for a while, you say "Call Jay" and it knows which one you want and it dials it. So clearly the machines and the technology and the services are learning. They may take a while before they home in on the exact thing you want. But they need to get enough examples of what you're trying to say and what you mean.

Hechler: But it doesn't absorb changes I make. It doesn't even know I'm making changes, as far as I can tell. If it actually absorbed what I'm doing to the transcript before I export it into Word, it would learn. But that hasn't happened. So there's no feedback loop at all. How do I know that? Because it makes the same mistakes over and over and over. That's why it's been such a disappointment to me: it hasn't learned, as far as I can tell, anything.

Wiipon: Well, in that case you try other software. You should experiment.

Hechler: I will. In the meantime, let's return to grammar. I'll just say It's atrocious. Sentences go on and on and on. Commas appear with alarming regularity, often where [the program changed "where" to "when"] they have no business being in the first place. What's the problem? Why doesn't the software know how to write?

Wiipon: That's pretty funny. It's not built to write. [Transcript: "It's not built to right."] It's built to recognize the words you said. Things like punctuation are part of meaning. They're not part of syntax of a language. They're part of the semantics of the language. You put punctuation in and by and large you intend a particular meaning or something to happen. And, yeah, it's something that they

are learning. Like I said, a decade ago you wouldn't get punctuations. Now, at least Microsoft puts punctuation in, and most of the ones that I've tried will put punctuation in. If I'm dictating while I'm on the phone, I'm in the car responding to a text message, I'll still put it in myself [by saying "comma, period," etc.], just to make sure it's there. But that is coming along.

Hechler: The program often hears the wrong word. Sometimes it seems way off **[like rendering "really" as "Billy."]** Other times, it's plausible, but in context it's clearly wrong. But the program doesn't understand context.

Wiipon: Yeah. Context, again. When you think of syntax, syntax doesn't know, by and large, the context. Context is usually a long-distance relationship across sentences, across paragraphs, or across the meaning of the document. You can see in the generative AI space, with ChatGPT, if you ask it to produce things, it's making good context relationships throughout a document. So in the generative case, it's working really well. Would you expect that to happen in transcription? Probably. I'm sure there is some of that in there already. But speech adds another level of complication. Generating text is a lot easier than first recognizing something that someone says and then trying to make it fit into the context and meaning that somebody wants. It's a lot harder to be able to do that.

Hechler: What do you expect will be the future of speech transcription? And how long do you think it will take for it to become really reliable? By which I mean, getting the right words and getting some sort of grammatical structure that is closer to what I would do if I were doing the transcript myself? Maybe 10 years?

Wiipon: It will constantly improve, but I don't think anything will be perfect in 10 years. Too many things to deal with. I am not sure where the advancements would be. I mean people consider ChatGPT an advancement. Not dictation. I think that's showing you what large language models can do. So the language part the machines are going to handle really well. The speech part is still a difficult thing. Interestingly there's not as many researchers working on the speech side as there are on the language side.

Hechler: Does the language side include grammar?

Wiipon: Yeah. When I say speech I mean for signal processing. You know, sounds come out of your mouth, they go into your ear, they go into a microphone, which goes into a computer. How are those analyzed, and used in such a way that a machine can recognize or distinguish one sound from another? That's still a science. And it used to be very sexy. It hasn't been as sexy as language in the past 25 years.

Hechler: Well, we will see. When I get the transcript of this conversation, I'm going to look for weird anomalies, to show readers what I'm talking about. I'll show them some sentences that might be really hard to figure out. Or maybe it'll surprise me. Maybe it understands that it's under special scrutiny today. And it will do its best work ever. But I suspect there will be a number of passages that will have us saying, "Huh??"

(Reprinted from the TAG Security Annual, 2nd Quarter 2024.)