Lost in Digital Translation

This article first appeared in *Orange County Lawyer* magazine in November 2011, Vol. 53 No. 11 (page 30).

by Early Langley

A leading voice-to-text provider who is attempting to gain a toehold in the legal arena touts the technology as "almost 100%" accurate. My son Andy, a UC Berkeley EECS senior—electrical engineering computer science—happened to have a professor with data indicating a 50% word error rate from voice to text on TV. In addition, the most recent National Institute of Standards and Technology evaluations show the best word error rate posted for multimicrophone speech recognition in a conference room was about 40%. I decided to investigate.

A couple of referrals led to a world-renowned expert in automatic speech recognition and understanding, Nelson Morgan, Ph.D. My son and I paid him a visit at his International Computer Science Institute headquarters in Berkeley, California, where he serves as director.

About 12 years ago, Morgan began an experiment within the four corners of the room we were sitting. He recorded his research meetings, and the recordings became the "American corpus" or data for his research in voice recognition. He placed microphones in front of each speaker. What he found is what we court reporters find: sound varies; it carries; and it gets lost in a room due to multiple speakers speaking at once; people speaking with heavy accents; people speaking indistinctly; air conditioning; shuffling papers; reverberation; tapping; doors opening and closing. His word error rate then is what he predicts now: 20% to 50% in meetings.

Just how does digital V2T work? "It's elementary, my dear Watson." Hardly.

Professor Morgan: "You compute representations of what frequencies have an energy from low to high. You think of that as a pattern which you're going to identify with a particular kind of speech sound." Researchers went to "probabilistic representations with statistical models which you have trained up on a whole bunch of other data which you hope is going to be similar to the data that you're going to get. There's the rub, by the way. The rub is that it won't be. Unless you're in the same room, with the same microphone in the same position with no external background noise, it won't be exactly the same."

Professor Morgan recounted that AT&T's first V2T rollout of interactive voice recognition was about 12, 15 years ago in answer to an old question: "Do you accept the charges for this call?" The "yes" recognizer depended upon a 200–word vocabulary for the one single word "yes."

If you extrapolate that out into a real-world scenario of millions of words pronounced thousands of different ways, you arrive at numbers that become exponentially difficult to translate automatically.

Speed and accuracy are hampered because the technology can't catch up with the demands this puts on it. You wouldn't think so because technology is advancing so quickly. But multispeaker voice to text with accuracy of context is one of the most elusive and sought-after pieces of research in artificial intelligence. In fact, it is hard to obtain grants for what many feel is a 100–year problem. Professor Morgan explained, "Some problems are just 100–year problems and speech recognition is one of them."

Let's visit your depo on calendar today. You arrive five minutes early and scan the depo room. Four microphones are available at the table. You state your appearance. You must lay claim to a microphone in order for the realtime transcript to reflect your client's appearance. And if microphones are shared, voices are mixed up. A monitor swears the witness. Realtime feed begins. There is no distinction between the question and answer. Words appear in brief. There is simultaneous speaker confusion and external noise. Numbers are confusing, and words are unintelligible. Readback of your question and answer is unintelligible.

Welcome to a digital voice-to-text (V2T) deposition. Translate that into your VIP appearance in front of the budget-strapped, V2T or digitally-recorded (DR) court, sans live Certified Shorthand Reporter and you may be in for a shock.

At your deposition, the reason you get only colloquy is because speech recognition, though it tries to understand context, cannot distinguish the difference between a question and answer; it cannot capitalize or punctuate; nor can it distinguish nuances in numbers such as, "The columns across are in a sequence of 12, 15, 18, 20," and "The columns across are in a sequence of 12151820." Homophones are problematic.

The V2T company's control of the deposition is critical to achieve a minimum word error rate—and expensive. The user must bear the cost of the company's package, which includes a four-track digital recorder using a laptop running Windows 7, i3 or higher, a USB mixer, lapels, table microphones, headphone and speaker, a Web camera (taking the place of the legal videographer) and cables; a "speech transcript," the most costly component, and a digital player that comes with earphones and foot pedal for the monitor. Costs not factored into the equation are an update package; back-up systems in the event of software or hardware failure; and an uncertified monitor taking copious notes. Last but not least, the cost of the human transcriber who must painstakingly play back and replay to get the right word, the right speaker, the right sequence of speakers, the right punctuation with meaning, and the right Qs and As, all to produce a transcript that is uncertified and unusable in court.

How does it work? The technology behind the speech engine, simply put, translates voice into words based upon algorithmic patterns. It then builds a "profile" on you mined from audio and written transcripts of your case and transcripts from other subject-specific litigation, and Internet documents.

Profiles on you take some time, and the first several go-rounds may be pretty rocky. Women's voices, it's been reported, are particularly hard to translate, as difficult as heavy accents. But take heart, by the fourth or fifth day your word error rate might go down—as long as the same people are talking about the same subject, using the same microphone, with the same equipment, in the same room. Of course, that doesn't count the newly-arrived, unprofiled deponent. And that's who you're most concerned with anyway, isn't it? The machine doesn't know any different.

Back to the transcript mine. What if your litigation were being mined in other cases? Where does the information come from and where does it reside? Who could get access to it? Is there a right to mine this sort of information, particularly with security breaches reported regularly? With identity theft looming over us, your work product and your client's trade secret or strategy information, personal information, such as HIPAA and bank records, risk exposure to your adversary and to the public.

Enter the Intersection of Ethics and the Digital World of Discovery

If you think this is far-fetched, think again. To save dollars, administrators are throwing a valuable person in California courtrooms under the bus: the live Certified Shorthand Reporter. It is incumbent upon litigators and consumers to be aware of the risks they face.

Imagine being in a DR courtroom with no reporter and given a CD to review and interpret for the next day's testimony. It may be six to seven hours' worth of testimony to review overnight in preparation for the next day's key witness or witnesses. Found someone else's highly sensitive information on the same CD? Found an attorney/client conversation recorded? Can't find what you're looking for in a hurry? Having problems locating exhibits? There's no keyword indexing. There's no word search capability. There's no condensed transcript. Day after day. How would this affect your practice, your time, your client's interest, and the integrity of the record?

While the closure of courts will impact public access to justice, long lines, delays in unlawful detainers, divorce and civil cases, for those courts still open, it need not take with it the integrity, accuracy, reliability, and neutrality of the Certified Shorthand Reporter. The risk of mistrials and retrials from garbled transcripts and lost recordings is a hefty price for your client to pay.

Next time you either notice a deposition or appear as a party at trial, make sure that you ask for a live Certified Shorthand Reporter to protect your client's rights. If you don't, your case may turn on testimony that's lost in translation.

Early Langley, B.A., CSR, RMR, is a Phi Beta Kappa graduate of the University of California Berkeley, President-Elect of the California Court Reporters Association, member of the Ethics First Task Force of the National Court Reporters Association, and Senior Staff Reporter for Aiken Welch Court Reporters. She can be reached at early.langley@cal-ccra.org.