"Transcription versus Speech Recognition"

By Andy Braverman, President of Apptec Corporation… a developer of *next generation* speech processing products.

When anyone mentions "Transcription versus Speech Recognition"… it sounds like it will be the next Heavyweight Boxing Match on the marquee at Madison Square Garden.

But in fact, Speech Recognition does not work without Transcription!

Before I describe why this is, please allow me to provide a little background information first.

Transcription is a process whereby a skilled typist types a document while listening to the audio dictation of it.

Speech Recognition (hereafter referred to as "SR") is a process whereby a computer program performs that same task. But instead of "listening" to the audio dictation as the human transcriptionist does, SR extracts mathematical (frequency and amplitude) characteristics for each spoken word, and then chooses the word from its vocabulary list that most closely matches those characteristics.

In fact, most advanced SR systems do their best-fit word matching not simply by looking at one word at a time, but by looking at the words that surround each word… this is called "context modeling or language modeling". And to further enhance their accuracy, SR systems usually limit their vocabulary list to the words that are expected to be spoken in a particular application (like radiology versus pathology in a medical application, or like criminal versus civil law in a legal application)… this pared-down vocabulary list is called their "lexicon". And one more thing that SR systems do to further enhance their accuracy is to use a "voice model", that is, to take into consideration an individual's unique pronunciation.

Long ago, a few decades ago in fact, SR started off being "Speaker Dependent". That meant that before someone could expect reasonable results from SR, they first had to spend an hour or more reading a specially prepared script… that was called "training". A user would have to carefully read that script, sometimes more than once, in order for the SR system to build the individual's "voice model". That script was specially crafted to make the user say words that contained the few hundred unique sounds (utterances) that allowed the SR system to best understand that person's unique pronunciation.

But it was quickly realized that no SR user, especially a busy professional, wanted to spend the time necessary to train the system to his or her voice. It was a very tedious and time consuming prerequisite that had to be done before the SR system could be used. So to overcome this problem, SR systems started to tout themselves as "Speaker Independent". They didn't accomplish this through any new technical breakthrough, but by moving the "training" to a background task. How, you may ask did they accomplish this? Well, simply by sacrificing accuracy for a few weeks while using a "feedback loop" (which consisted of a Transcriptionist correcting the SR output) to build the individuals "voice model" over time. Instead of an hour or so of "training" so that the SR system could learn a user's pronunciation, the SR system learned over weeks of best-guessing what the user said, and then seeing what the user actually said by noting the corrections made by the Transcriptionist.

Now we come to a very important thing you should know, about what accuracy rates you can expect from SR systems and the reason why SR alone does not work without the support of Transcription.

As we’ve all seen, in the futuristic world of Star Trek, the Enterprise Computer would never misinterpret anything anyone said to it, unlike the famous Microsoft Bill Gate’s example of an errant SR system interpreting someone saying “Recognize Speech” for “Wreck a Nice Beach”. But in fact, that’s just the kind of mistake that even the most sophisticated SR products continue to struggle with today.

SR accuracy rates vary from product to product, but typically start around 80% and increase over usage to around 90% percent. Some SR manufacturers may say that their system is even more accurate, but they all will agree, even if they won’t admit it, that their SR systems are not 100% accurate and will never… its worth repeating… will NEVER be!

And that’s the reason why SR alone does not work without Transcription… because SR isn’t 100% accurate, which therefore requires a Transcriptionist to correct its output!

The SR system’s feedback loop is called by some manufacturers the “Correction Editor”. This sounds like it is a software utility, but in fact is a skilled typist... the Transcriptionist. A Transcriptionist is the person that fixes the output of the SR system. After the SR system does the best job it can of typing what it thinks it heard… because we know there may be one or more wrong words in each document, the Transcriptionist must listen to the entire audio dictation while visually proofing the SR-typed document, in order to type-over and correct the wrong words. It’s this correction that the Transcriptionist makes that is the feedback loop that helps the SR system to “learn” how each individual pronounces words. While that feedback loop helps improve the SR accuracy from around 80% to the low to mid 90%... that is pretty much the best overall accuracy that can be expected.

Even if SR systems were 99% accurate, that would mean that for every 100 spoken words (which is barely a paragraph or two), 1 word will be wrong! If, for example, an SR system misses the prefix “non” in non-malignant… without the Transcriptionist being in the loop to correct that mistake, that simple mistake would make for a very bad day for the patient, the doctor, and the hospital.

Over the decades I’ve designed dictation products and SR systems, and have sat in many “Transcription Solution” meetings where a hospital’s Transcription Department feared that SR technology was going to put them out of business… that is was going to replace them with computers that have no need for food, rest, or a paycheck. But in each case, I’ve never seen that happen. That is because the hospital still needs their transcriptionists to proof and fix the output of the SR system.

Every SR Transcriptionist I’ve met has said “… while I sit at the keyboard, with my fingers-at-the-ready, ready to correct the SR system’s mistakes, I could have just as quickly and easily typed the entire document myself!”.

And that brings into focus a very important consideration when debating whether to implement an SR system…and that is to consider if it will be cost effective to implement an SR system for the transcription of dictation.

From what I’ve described thus far, I think you’ll agree that in most cases the answer will be “no”. In most dictation applications, it is not cost effective to implement an SR system… because you still need your transcription staff to correct the SR system’s mistakes.

Ahhh, but you say, they only have to type a few words instead of the whole document… that is true, but they still have to listen to the entire document… and since they can type as fast as they listen, they could have simply typed the whole thing themselves. In fact by implementing an SR system, you’ve just added the (usually high) cost of the SR system to the transcription costs you already have. And with the same number of transcriptionists listening to all of your dictations, you have no net-gain in efficiency brought about by the use of the (imperfect) SR system.

Some SR sales people will say, “...but you can get rid of your skilled MT’s (medical transcriptionists) and replace them with lower paid, lower skilled typists. But that doesn’t work in reality. An MT is an MT because they are keenly familiar with the medical terminology being used. A lower skilled typist will spend many more times the time of a skilled MT in proofing work, as well as constantly interrupting other typists to ask them to listen to help them to determine what a doctor has said. Instead of efficiently proofing and correcting documents, a lower skilled typist’s head will be buried in the PDR (Physician’s Desk Reference) constantly trying to look-up which word they thought the doctor said. If you’re willing to, and can afford to let them struggle through a few years of hard earned experience, they might actually end up becoming reasonably proficient MT’s after a very long period of apprenticeship.

And just about every SR sales person will say, as they have for the two decades I’ve been involved with this discussion of “Transcription versus Speech Recognition” is, “…just one more generation of faster computer and SR will become the Holy Grail.” We’ll, as I’ve described earlier, SR is a very complicated process of analyzing the audio file, which results at best in an imperfect best-guess of what was said.

To put the complexity of the SR task into perspective, you just have to look at the size of a typical SR system’s lexicon (the vocabulary list of words for a particular application). In a very effective yet limited application of SR, such as in an Airline Reservation System, the lexicon contains only about 100 words. A few words like Flight Number, Arrival, Departure, the names of Airlines and Cities, and the numbers 0 through 9 are sufficient for a person to ask an SR based Airline Reservation System if their flight will depart on time. **But in an application like Radiology, or Pathology, or in a legal or business application, the typical lexicon can contain 20,000 to 40,000 words!** And forget the ubiquitous “talking typewriter” application that we all wish we had… where you can speak on any subject not restricted to a particular application and have the machine type a perfect document for you… **because for a “Conversational English” application, the size of the lexicon grows to 1,000,000 words and beyond!**

With that simple comparison of the size of the lexicon required for particular applications, it becomes evident why it’s such a daunting task, and why SR’s typed results are less than perfect.

As a design engineer and one who has spent nearly two decades in the dictation and transcription industry, I am a fan of SR technology. But I am also a practical person when it comes to product marketing. **As such I have found that today’s SR technology, and what we can expect it to become in**

the future, falls short of being a cost effective technology to implement in a typical dictation environment.

And here's an interesting reality to note… that the opposite task, that of "Speech Synthesis" which has been successfully used in talking computer applications, talking toys, talking appliances and other talking products, is actually pretty simple to design by comparison to an SR system. The talking toys of decades ago (like Texas Instruments' Speak'N' Spell) were very robotic sounding… but the talking applications of today are relatively natural sounding. For an example of this, take a moment to listen to the synthesized voice you'll hear at www.DigiTelStore.com.

The reason Speech Synthesis is easy to accomplish, relative to SR, is because it takes only a few hundred "utterances" played in their proper sequence to form any word that can be spoken in the English language. (To oversimplify how Speech Synthesis works… these "utterances" are syllable-like sounds that the mouth, tongue, and larynx make as air passes through them. To learn more about this technology, search the Internet for the key words "phonemes" and "allophones".)

So now that you have a brief understanding of how these technologies work, it becomes very easy to understand why Speech Synthesis is a practical application of technology, while unfortunately for medical, legal and general business dictation applications, Speech Recognition is not practical.

With this understanding of how complex the problem is in trying to create the perfect SR system, it reinforces my amazement at how intricate and unique the human brain is, especially when it comes to interpreting speech. From a few months old, a child begins to understand the spoken word… yet after decades of the most brilliant researchers working on SR techniques, and companies investing millions upon millions of dollars into its research, it seems that we have reached a pinnacle of the expectations for the maximum accuracy we can expect of an SR system. **Because that accuracy expectation is less than 100%, and because an SR implementation cannot completely take the human element (the Transcriptionist) out of the loop… this shows that today's SR technology is generally not a more cost effective solution (to turning dictation into documents) versus simply using traditional Transcription techniques alone.**

Thank you for taking the time to read this article. I hope you have enjoyed it. I would love to hear your comments and of your experiences working with Transcription and SR technologies.

Mini Bio: Andy Braverman is the President and Owner of Apptec Corporation. Andy has been involved in the design and marketing of dictation and transcription systems for nearly two decades… one decade of which was devoted to designing *next generation* dictation and transcription products for Philips Speech Processing of Vienna, Austria. In the past decade, Andy has devoted his talents to bringing to market feature rich and cost effective dictation and transcription products for medical, legal, and general business applications. His company, Apptec Corporation, based on Long Island, is also involved in developing custom products to suit their client's specific needs, from software development to circuit design. If you have a question for Andy, or a problem that needs solving, he invites you to contact him at 1-631-828-1245 or at Andy@DigiTelStore.com. Or to see his latest adventures in the field of Speech Processing please visit www.DigiTelStore.com.