# Speech Recognition Is Here At Last!

**by Joe Weber**

**The technology we've all been anxiously awaiting has finally arrived.** *There are now several major hospitals across the country that have implemented enterprise-wide speech recognition. At each of these hospitals, all the physicians are dictating into PCs so that powerful software can instantly display their spoken words. The physicians then review this recognized text, correct the very few mistakes, and electronically sign the report immediately. No transcription delay. No transcription cost. The hospital executives are thrilled, and the word is just beginning to spread. Within a year or two, this will surely be the way clinical documentation is handled everywhere.*

Just kidding. I figured this was an effective way to get your attention, albeit personally hazardous, given the readership of this magazine. That whole first paragraph is a lie, except for the first sentence. I do hope you're chuckling rather than gritting your teeth or clutching your chest because we will now turn our attention to a speech recognition approach that actually is appealing to the medical transcription industry.

## The Reality

The simple reality is that most physicians refuse to correct the mistakes made by a speech-recognition engine, even when there's only a 2-3% error rate. Therefore, the primary users of what we shall call "front-end" recognition are only those physicians who are both progressive thinkers and would otherwise be paying for transcription out of their own pockets. And there are just not that many progressive physicians.

The vast majority of physicians want to keep doing exactly what they've always been doing, and nothing more. They also won't embrace structured/codified input, which is arguably the most powerful weapon available to us for advancing the science of medicine as well as increasing both the quality and cost effectiveness of healthcare. But that's a discussion for another day.

## The Physician-Friendly Approach

There is a way to implement speech recognition that is actually quite palatable for physicians. It's palatable because they don't even have to know that it's going on, and they don't have to change their dictation behavior at all. Physicians love innovation but they hate change. So if you want to stay friends with physicians, don't force them to change anything. This stealthy approach, happily, is also friendly to transcription folks. It's not the Holy Grail, but it is beginning to make a significant impact on the industry.

We shall refer to this approach as "back-end" recognition. Physicians dictate as they always have, typically over the telephone into a digital dictation system. The resulting digital-voice file is routed through a speech-recognition engine to produce a draft of the text report. This draft is transmitted to a medical editor, who listens to the playback while reading the draft and corrects any mistakes observed. Dictation can also occur via portable digital recorder or PDA, uploaded to a PC, or physicians can dictate directly into a PC. In these cases, the voice file is digitally transmitted to the server for recognition.

## What's the Payoff?

The benefit of this back-end approach is simple: increased productivity, resulting in reduced cost. But does it really have this impact? There are a number of vendors offering back-end recognition solutions. They claim productivity improvements ranging from about 30% to 100%. The latter figure represents a doubling of output, which is big! With the lower figure, however, it doesn't seem worth going through the process change, not to mention that it's not worth the cost. The most advanced speedtyping software, i.e., Stedman's Smartype and Instant Text, averages about a 40% productivity boost for less than $200, one-time cost per transcriptionist. It's rather foolish to pay thousands of dollars per transcriptionist each and every year to achieve less than that. [Disclosure: Stedman's Smartype is my company's product.]

So, for the sake of our argument, let's assume that we're only interested in this technology if it approaches an average of 100% productivity increase. Let's run the numbers from the perspective of a medical transcription company. If the company now pays 9 cents per line (cpl) to its transcriptionists for straight transcription, and it can promise them a doubling of productivity, then its transcriptionists should be willing to accept 5 cpl. If the technology costs less than 3 cpl, and actually does double productivity, it makes sense to implement it.

Unfortunately, there have been implementations where the cost is 2-4 cents per line and productivity is increased by just 30-40%. There's no value proposition in that scenario. If you're considering implementing a back-end recognition solution, it's critically important to come up with a reasonably accurate estimate of productivity increase (through a controlled test), determine what that's worth in cost savings, and then match that to the price being charged by the technology vendor. If the latter is not at least 1 cent per line, preferably more, lower than the former, walk away. Run away.

## The Total Solution

Most of the vendors of this application technology provide a total solution. This means that you place your entire dictation/transcription operation on their platform. If you're looking to purchase a new dictation system, transcription software,

and document-distribution system, then this is a reasonable course to evaluate. The cost for this total solution, however, is likely to be rather high, potentially requiring some up-front dollars plus maybe 3-4 cents per line.

There are two major basic-technology providers of speech recognition: ScanSoft (Dragon) and Philips (SpeechMagic). The accuracy of their engines is rather close. Each one does certain things better than the other, but, overall, there is not a substantial difference. I believe that one is a little better, but in this article I'm not going to tell you which one that is.

The application providers use technology from one of the technology providers, ScanSoft or Philips, or they utilize a proprietary technology. Remember that while accuracy is an important contributing factor, the only variables that really count are productivity increase and cost per line. Make sure you know the magnitude of both and how they compare before you sign any agreement that requires a substantial investment.

### The Other Approach

If you have a workflow which makes you happy, and prefer not to go through the process-change and expense to change it, then you should consider speech-enabling your existing platform. This means that you acquire just the speech engine and the toolkit to integrate it. It will require some work but you won't have to go through a major platform shift, and you should save many thousands of dollars. The only observable difference from what happens today is that your transcriptionists will receive a draft text report, along with the synchronized voice, and will now be editors rather than typists.

In order to maximize productivity, you will need to incorporate some complementary software, such as automatic formatting. If the editor needs to format the report as well as correct misrecognitions, there is not likely to be a speed advantage over straight transcription. If all they have to do is make the corrections, this will have a profound positive impact on overall productivity.

It is also important to optimize the acoustic and language models of the speech-recognition engine. The acoustic model represents how each dictator pronounces the sounds (phonemes) of the English language. The language model determines what words the dictator uses and how s/he puts them together in context.

In front-end recognition, the acoustic model is initially customized by having the dictator read displayed text for 5-20 minutes. But in this back-end approach, the dictator is kept in the dark. So the acoustic model is formed by matching the words in prior voice dictations to the words in the associated transcribed text. The language model is put together by analyzing a relatively large number of prior reports for each worktype for each dictator. There is no need for any effort on behalf of the dictating physicians. Lucky for us! Because we know exactly how cooperative physicians can be when asked to make any effort to improve clinical documentation. Nonetheless, it doesn't hurt to ask the physicians to be a little more careful with their dictations. You'll probably have to incent them with money or doughnuts.

The accuracy, naturally, will vary by dictator. Those who enunciate most clearly will achieve the highest accuracy. The systems do surprisingly well with accents as long as the dictator doesn't mumble or manifest substantial dysfluency. Some physicians are such bad dictators that it will be years before the technology advances enough for it to make sense to even attempt to edit their drafts generated by these recognition engines, but the majority of dictators should qualify immediately.

To get started on this process, it seems advisable to acquire some outside expertise to help assure an elegant integration into your workflow and to optimize your abilities to construct the best acoustic and language models, which are critical for maximizing accuracy and thus productivity. Make sure that the consultants will transfer their skills and knowledge to your staff, once the process is running smoothly and effectively. [Disclosure: My company provides autoformatting and other software, as well as implementation expertise for speech-enabling existing workflows.]

If you choose to go this route, you are likely to find the pricing extremely attractive. The software, when amortized over 3 years, can come out to well less than 1 cent per line. If you can double productivity for that price, you don't need complex math to recognize the value proposition.

### Is It Really Here?

Since 1982, when I sat in the living room of the founders of Dragon Systems excitedly observing their initial alpha software run on an 8086 IBM PC (anyone remember those?), I've been watching this technology very, very closely. Starting in the mid-1980s, lots of folks fell prey to the rolling 3- to 5-year window: In 3-5 years, the first paragraph of this article will be reality. Well, that window rolled for a couple of decades, and that reality is still not here. But what is here, as described in this article, is something rather powerful. And it is something that should resonate with the souls of everyone in the medical transcription industry.

If transcription productivity can be doubled for 1 cent per line, the entire face of the transcription industry should be transformed overnight. In any industry related to healthcare, nothing ever happens as fast as we think it will, but the handwriting is on the wall. Or, in more apropos verbiage, the words are appearing on the screen. If you can save several cents per line for most of your dictators, that's an opportunity you should grab sooner rather than later. As stated in that first true sentence of this article, "The technology we've all been anxiously awaiting has finally arrived."

Joe Weber is CEO of Lexicore, provider of software and consulting services to optimize speech recognition applications for medical transcription companies and healthcare organizations. E-mail: **joeweber@alum.mit.edu**.