Speech Recognition (Speech to Text) Software

How It Works Explained in Nearly Plain English

When speech experts talk about a "rules-based" approach, they are talking about a set of rules derived from the knowledge of experts. In the context of speech processing, these rules can be analogized to a set of English grammar rules. The most straightforward approach to creating speech recognition software would involve using a rules-based approach to compare speaker utterances to a set of English grammar rules and then searching for the closest match in a vocabulary list. In practice, no one has been able to do this with acceptable accuracy while maintaining an acceptable processing speed even though rules based approaches have produced acceptable accuracy for speech synthesis (text to speech) software.

Speaking is quite different from written text for a number of reasons. First, even individuals with very polished diction still insert the occasional "um" or awkward silence into their conversations. This is much rarer in written text unless one is dealing with a verbatim transcript. Second, regional diction is much less important in written text then it is in spoken conversation. Sports columnists from Birmingham having conversations with sports columnists from Boston may have difficulty understanding each other but their written product is likely to be very similar. Third, noise interferes with verbal speech input. All of these things make spoken speech more difficult to model and process.

When speech experts talk about a "corpus-based" approach, they are talking about combining statistical analysis with speech data. In speech processing, a corpus is a collection of audio recordings of speech and speech transcripts. Current commercially available speech recognition software programs use a corpus-based approach. They perform acoustical and statistical analysis of voiceprints to determine what a speaker has said. The typical program uses a mathematical model called a Hidden Markov Model to assign probabilities to words in its vocabulary list based on its analysis of the audio segments and chooses the word most likely to be next in the sequence.

To use a classroom analogy, the speaker/teacher isn't assigning English homework to the program; he or she is assigning math homework instead. This is why every such program will occasionally produce very strange results. The program has no knowledge of English grammar and only indirectly filters out words and phrases that make no sense in context.
Inherent Limitations

There are many people who have predicted the imminent disappearance of keyboards and mice as input devices. They may have overlooked certain issues that would impact speech recognition software even if it was much better than it actually is.

While it is possible to train speech recognition software to a very high degree of accuracy in diction and to drastically decrease the need for a keyboard and mouse, there are several significant drawbacks (at least in a work environment). First, speech input is inherently noisy. While this is considered acceptable in some work environments (e.g., telemarketing), there are many other environments where this would not be considered acceptable. Second, if one is using one's voice to dictate, one cannot be using it for other things. In many work environments, one might be on the phone at the same time that one is typing. Third, some tasks are inherently faster for able-bodied users when using a mouse or keyboard (e.g., opening Microsoft Word, changing fonts, changing margins etc.). Fourth, while it is reasonable to assume that the need for training will decrease as the software gets better, it is unlikely that one will be able to get the maximum benefit out of the software without training anytime in the foreseeable future. For many users, the time spent training will not be worth it because using a keyboard and mouse will be easier if not faster.

Improving It for Current Users

Those who have mobility issues in the upper body but retain their speaking skills could still benefit from the software. Despite the inherent limitations, certain classes of able-bodied users (e.g., radiologists) could make use of a sufficiently robust software program. Radiologists dictate more reports and make more use of transcription than perhaps any other medical specialists. Despite the fact that they would be the most obvious able-bodied customers, Nuance Communications has failed to make much headway in selling speech recognition software to them.

Radiology journals maintain an ongoing debate between those who use speech recognition software and those who do not. The speech recognition software users are a minority of radiologists. Those who do not use speech recognition software have the usual complaints regarding accuracy. It is not clear from the various articles which version of Dragon NaturallySpeaking proponents and critics are using. While Nuance markets products specifically for radiologists, it is unclear whether any radiologists have purchased them.

The fundamental barrier to improvement is processing power. The Hidden Markov Model (HMM) and other mathematical models used in speech recognition can be understood as a series of mathematical assumptions that are often but not always accurate. These assumptions are made to reduce the amount of data necessary to run
the program (thereby reducing the processing power necessary to run the program). Eliminating assumptions and incorporating data in place of those assumptions requires additional processing power.

It took nearly a quarter-century from the first application of a Hidden Markov Model to the speech recognition problem for the first commercially viable continuous speech recognition program to appear because earlier computers lacked the processing power to run such a program. There have been only incremental improvements in mathematical modeling since the mid-1970s largely because even early 21st century personal computers lack the power to run more sophisticated mathematical models while processing speech in real-time.

Research grants for the study of speech recognition and speech synthesis have waxed and waned but never entirely dried up. Unfortunately for the current users of speech recognition software, most current research is focused on the long-term rather than improving the products that are on the market today.

Though IBM funds some speech research, IBM cared so little about its ViaVoice software that it licensed it to a competitor who bought the rights to Dragon NaturallySpeaking in a bankruptcy proceeding. The two products were subsequently merged. Microsoft bundles its speech recognition software with the newer Windows operating systems. It only does that for programs that it does not value highly. Microsoft's high-value programs (e.g., Microsoft Office) are never bundled with the OS. Even the creator of Dragon NaturallySpeaking (who lost his intellectual property rights to Dragon NaturallySpeaking in bankruptcy, see below for details) is focused on long-term projects unlikely to bear fruit in his lifetime.

Nevertheless, speech recognition software may continue to see at least incremental improvements. It may benefit indirectly from ongoing development of speech recognition in other contexts. Many companies have implemented speech recognition in their voicemail menus. This task is much simpler because the vocabulary is limited. The system responds to any unknown input by asking the customer to repeat the input or offering to transfers the customer to a human customer service representative.

**Historical Dates Relevant to Current Commercial Products**

1947 – Researchers publish a book detailing their work with spectrographs that produce spectrograms. Spectrograms render speech in visual form, creating "visible speech."

---

1952 -- Researchers at Bell Labs build a machine capable of recognizing spoken digits. They claim a 97% accuracy rate.²

1956 -- Researchers at RCA Labs build a "phonetic typewriter." They claim a 98% accuracy rate.³

1960 -- An MIT researcher proposes a model for speech recognition and speech synthesis while acknowledging that building such devices is beyond 1960 capabilities. The discussion includes the recognition of problems (e.g., multiple speakers) that have not yet been solved.⁴

1966 -- Leonard E. Baum publishes the first mathematical proof for a Hidden Markov Model.⁵

1969 -- Bell Labs engineer John R. Pierce writes a letter to the editor of The Journal of the Acoustical Society of America highly critical of speech recognition research.⁶

1971 -- DARPA begins funding the Speech Understanding Research project at selected universities.⁷

1975 -- IBM computer scientist James K. Baker completes his PhD thesis requirements at Carnegie Mellon⁸ while working on the SUR project.⁹ He becomes the first person to use a Hidden Markov Model¹⁰ to address the problem of speech recognition, applying the HMM to a program known as Dragon (named after the pattern on his wedding

---

⁴ Kenneth N. Stevens, Toward a Model for Speech Recognition, 32 J. Acous. Soc'y Am. 47 (1960).
China). It is the only meaningful result from $15 million of DARPA funding for the SUR project.\textsuperscript{12}

1982 -- James and Janet Baker found Dragon Systems. Dragon Systems generates income through federal contracts and business-to-business projects.\textsuperscript{13}

1990 -- Dragon Systems releases DragonDictate, a “discrete” speech recognition program (i.e., it requires pausing between words) with a 30,000 word vocabulary.\textsuperscript{14}

1997

April -- Dragon Systems announces Dragon NaturallySpeaking, a continuous speech recognition program, offering a Personal Edition for $695 and a Professional Edition for $995. It claims a 95% accuracy rate with a 30,000 word "active" vocabulary and a 230,000 word "backup" dictionary from which words can be added to the active vocabulary during correction.\textsuperscript{15}

June -- IBM announces ViaVoice, a continuous speech recognition program for only $99.\textsuperscript{16} Dragon Systems’ immediate response is to extend a $299 special offer.\textsuperscript{17}

1998 -- Dragon Systems announces Dragon NaturallySpeaking Standard and Dragon NaturallySpeaking Preferred\textsuperscript{18} (products that would remain until 2010). Standard is

\textsuperscript{10} Baker, \textit{supra}, at 5.
\textsuperscript{11} Garfinkel, \textit{supra}, at 61.
\textsuperscript{12} Klatt, \textit{supra}, at 1354.
\textsuperscript{13} Garfinkel, \textit{supra}, at 62.
\textsuperscript{14} Id.
priced at $109\textsuperscript{19} and Preferred at $199\textsuperscript{20} (adjusted for inflation, their 2010 prices were about one third less).

**2000**

**March** -- Lernout & Hauspie announces the acquisition of Dragon Systems.\textsuperscript{21}

**August** -- The *Wall Street Journal* reports that Lernout & Hauspie misstated the extent of its Korean business.\textsuperscript{22}

**November** -- Lernout & Hauspie files for bankruptcy.\textsuperscript{23}

**2009** -- James and Janet Baker's lawsuit against Goldman Sachs (for recommending the merger with Lernout & Hauspie) survives a motion to dismiss.\textsuperscript{24}

**2010** -- Jo Lernout and Pol Hauspie are sentenced to five years in prison.\textsuperscript{25}

**2011** -- Nuance releases Dragon NaturallySpeaking 11. While it is significantly improved over the earliest versions, improvements from version 3 to version 11 are similar to the incremental improvements between Word 97 and Word 2007 rather than any fundamental change.

**Conclusion**

The inherent limitations of speech recognition software combined with fundamental barriers to its improvement limit its possibilities as a general use product. Because it is not a mass-market product, research and development will be more limited than it otherwise would be. Even so, speech recognition software is sufficiently robust to be highly useful to users whose motor skills allow voice input but discourage keyboard input. Despite its flaws, it can be useful in bridging the digital divide between able-bodied and disabled users.


\textsuperscript{22} Charles Forelle & Mark Maremont, Lernout & Hauspie Founders Guilty in Fraud, Wall St. J. Sept. 20, 2010, available at \url{http://online.wsj.com/article/SB10001424457506252545104575503500899087566.html}.


\textsuperscript{24} Id.

\textsuperscript{25} Forelle & Maremont, supra.
Selected Bibliography

**Wikipedia Articles (in suggested reading order)**


**Academic Articles and Monographs (in suggested reading order)**


- Various authors, *Part E Speech Recognition*, in Springer Handbook of Speech Processing, 521-724 (Jacob Benesty, W. Mohan Sondi, Yiteng Huang eds., 2008) [10 articles on various aspects of speech recognition].

**Radiology Articles (in chronological order)**


**Patent Applications**