WATCH WHAT YOU SAY, YOUR COMPUTER MIGHT BE LISTENING: A REVIEW OF AUTOMATED SPEECH RECOGNITION

Stephen De Gennaro
IBM Thomas J. Watson Research Center
Yorktown Heights, New York

Spoken language is the most convenient and natural means by which people interact with each other and is, therefore, a promising candidate for human-machine interactions. Speech also offers an additional channel for hands-busy applications, complementing the use of motor output channels for control.

Current speech recognition systems vary considerably across a number of important characteristics, including vocabulary size, speaking mode, training requirements for new speakers, robustness to acoustic environments, and accuracy. Algorithmically, these systems range from rule-based techniques through more probabilistic or self-learning approaches such as Hidden Markov Modeling and Neural Networks.

This tutorial begins with a brief summary of the relevant features of current speech recognition systems and the strengths and weaknesses of the various algorithmic approaches.

Issues critical to the successful application of speech recognition in human-machine interactions will be discussed in more detail. These issues include: "raw" recognition vs. speech understanding, grammar-based, constrained dialog vs.. natural-language, free discourse and the imposition of appropriate constraints to make the task technologically feasible, with acceptable accuracy. The potential delay in recognition, due to algorithmic features or to interactive verification of recognized words for error correction, must also be considered. The need for verification is determined by both the expected error rate and the cost of each error. The cost function could be dynamic if errors will have significantly different impact given the state of the system.

As in any human-machine interface, not all of the constraints are due to the "-machine" half of the interaction; the "human-" partner also imposes a share of burdens. For example, people will not always say exactly what is expected of them, even when carefully prompted with a small set of possible choices. Stress, attention, and general physical state can have strong effects on system performance.

For applications that are not currently speech-driven, it might also be necessary to reformulate the basic human-machine interface. For example, moving a cursor over a document by voice in the most straightforward way (up-up-up-left-left- ...) is not as efficient (and probably not as acceptable) as moving the cursor by keys or mouse. Recasting the basic problem as "go to the word HELLO is the second paragraph" provides a better match between the task and speech control. Navigating a robot by voice instead of by joystick might require similar reformulation. Of course, while "go to the word ..." is more natural, it imposes new requirements to understand the command and map it into the appropriate action.

Given these characteristics, it is not yet reasonable to expect that an off-the-shelf speech recognizer can be used as a transparent replacement for a keyboard or joystick as input to any arbitrary application. The highest performance levels will be achieved when the application and recognition processes are more tightly coupled.

Using existing speech recognition technology, it is certainly possible to build successful virtual environments. This is particularly true for systems that mimic environments in which speech is already the primary communication mode, and where the control "protocols" have been well defined. For example, Air Traffic Control training consoles have been demonstrated using continuous speech, but the task is constrained to deal only with objects (i.e., specific flights) on the trainee's console to provide reasonable accuracy and response delay. A general problem with coupling, however, is that it leaves the system vulnerable to unanticipated scenarios, with concomitant risk that the system will not respond gracefully to "exceptional" interactions.

This tutorial will conclude with some pointers to conferences and publications that reflect the current state of commercially available speech recognizers, and the state of research in this field.