Voice recognition is one of the most desirable traits that a computer can have, but also one of the most difficult features to implement. Human-computer interaction via a voice interface is an integral part of the computer of the future; just take a look at The Jetsons, 2001, or any sci-fi vision of future computing.

Voice recognition products are commercially available, and have been for several years now. Apple and IBM, most notably, manufacture products that perform voice recognition, both for dictation and computer control. However, frequently, users of such services are frustrated by the software: it requires hours of training, and even when such training is complete, voice recognition is nowhere near being flawless, and can be quite frustrating.

However, this is not a problem of inherently bad software design on the parts of IBM, Apple, and others. The problem is that voice recognition is more than just a simple parsing and analysis of sound. Most voice recognition products on the market these days have some sort of statistical context to determine the most likely words, but in some cases this just isn't enough. You need the voice recognition interface to have common sense, to be able to differentiate between homophones better, for example. Context plays an integral role in the way humans process speech, and thus, it must also play a role in the way computers process speech, if they hope to do it accurately.

Take for example, the following sentence:

“One of the elements in the Periodic Table is lead.”

From a strictly phonetic point of view, “lead,” the element, and “led,” the past tense of “to lead,” are the exact same word. However, humans are able to distinguish between these two words through an application of common sense using context. Earlier in the sentence, the speaker talks about elements and the Periodic Table, so it would follow that lead refers Pb, and not the verb.

Voice recognition software has the ability to draw on statistical context, which is useful in improving accuracy, however, it does not have the ability to draw on common sense... at least not yet!

The goal of this project is to enhance voice recognition by using contextual clues. The OpenMind Common Sense project, started here in the MIT Media Lab, is a huge repository of common sense knowledge – it contains information on how different words and concepts relate to each other. This massive repository of word relations can then be used to deduce the context in which words are used. This knowledge, coupled with the API for the IBM ViaVoice speech recognition product, can be used to enhance the quality of voice recognition.

Ultimately, the project, upon completion, would work as follows:

When ViaVoice encounters a questionable spoken word that could have multiple written options (cents/sense, to/two/too, lead/led, red/read, etc.), it would query the OpenMind repository to determine which of the homophones is the most likely option, given the context of the word (based on the previous words, sentences, paragraphs, etc.).

Currently, when a word is incorrect and the user notifies the computer, a dialog box containing a list of possible words pops up, and the user has to select the correct one from the list. However, if OpenMind can improve the accuracy of voice recognition to the point where there are only one or two really likely words, the user will be able to say “Oops” or press a key, or do something to indicate the inaccuracy to the computer – and the computer would automatically substitute the incorrect word with the next most likely word, as determined by ViaVoice and OpenMind. This is a more intuitive and natural feel, just as though you were dictating to someone, and is also much quicker than having to read through an entire list of options to select the correct word.

It is hoped that this coupling will result in significant improvement in the accuracy and ease-of-use of voice recognition, making it a viable option for dictation and every-day use as a computer interface.