Abstract

This paper reviews past work comparing modern speech recognition systems and humans to determine how far recent dramatic advances in technology have progressed towards the goal of human-like performance. Comparisons use six modern speech corpora with vocabularies ranging from 10 to more than 65,000 words and content ranging from read isolated words to spontaneous conversations. Error rates of machines are often more than an order of magnitude greater than those of humans for quiet, wideband, read speech. Machine performance degrades further below that of humans in noise, with channel variability, and for spontaneous speech. Humans can also recognize quiet, clearly spoken nonsense syllables and nonsense sentences with little high-level grammatical information. These comparisons suggest that the human-machine performance gap can be reduced by basic research on improving low-level acoustic-phonetic modeling, on improving robustness with noise and channel variability, and on more accurately modeling spontaneous speech.