



Recognition by humans and machines: miles to go before we sleep

Richard P. Lippmann¹

Room S4-121, Lincoln Laboratory MIT, 244 Wood Street, Lexington, MA 02173-9108, USA

Received 7 February 1996

Boulevard and his colleagues [1] note that much effort over the past few years has focussed on creating large-vocabulary speech recognition systems and reducing error rates measured using clean speech materials. This has led to experimental talker-independent systems with vocabularies of 65,000 words capable of transcribing sentences on a limited set of topics. Instead of comparing these systems to each other, or to last year's performance, this short comment compares them to human listeners, who are the ultimate users and judges of speech recognizers.

Despite dramatic technical advances, the goal of achieving human levels of speech recognition performance remains distant. Detailed comparisons [2] demonstrate that word error rates for humans are still almost an order of magnitude less than machine error rates on many speech tasks. For example, with a benign amount of additive noise (SNR 10 dB) human error rates on Wall Street Journal sentences are roughly 1%, while machine error rates are above 12%. Humans do not rely as heavily on constraining grammars to achieve low error rates as do current machine recognizers. We can perform accurate acoustic-phonetic analyses of words extracted from spontaneous conversations, accurately recognize nonsense syllables, and accurately recognize words

in nonsense sentences with little contextual information. Human word error rates for nonsense sentences presented in quiet are 2% while machine word error rates on the noise-free 1000-word Resource Management corpus with a null grammar are 17%. Humans can also rapidly change topics during a conversation and accurately recognize proper names, letters of the alphabet, and digit strings.

The superiority of humans over machines is even more dramatic for degraded speech. Machine error rates often rise dramatically for spontaneous speech, rapidly spoken speech, and speech with noise and channel variability. The word error rate for a high-performance recognizer tested on the spontaneous-speech Switchboard corpus is 66%. This is more than ten times greater than the word error rate of roughly 4% estimated for human transcribers. Boulevard et al. review many approaches which could be used to make machine recognizers more robust. Most of these require prior knowledge that a particular type of degradation is expected, and the best performing approaches require samples of degraded speech. Humans exhibit much more powerful types of adaptation. We immediately recognize speech with unnatural types of degradations including extreme waveform clipping, severe band-reject filtering, short-term periodic interruptions, and analog waveform scrambling. We also rapidly adapt to natural variability caused by new talkers, talking rate variations, reverberation, room acoustics, and variability

¹ Tel.: (617) 981-2711, Fax: (617) 981-0186, Email: rpl@SST.LL.MIT.EDU.

in channel characteristics due to head shadow and other acoustic effects.

Current recognizers also can not distinguish new words and non-speech environmental sounds from acceptable speech input. Even human children can distinguish environmental sounds from speech and are able to distinguish new from known words. It is estimated that a high-school graduate knows more than 160,000 words, counting roots, derivatives, and compounds (e.g. [3]). Achieving this vocabulary requires learning roughly 10 new words a day. Machine recognizers will not be capable of learning the meanings of new words until they can accurately distinguish new from known words. They also will not be able to take advantage of a large 160,000 vocabulary until low-level acoustic-phonetic analyses are vastly improved.

The above comparisons suggest three important research directions where error rates could profitably be increased initially with the goal of eventually improving performance. First, improved low-level acoustic-phonetic accuracy is essential to increase performance. This will require new features which independently model temporal-spectral regions covering multiple time-spans. To gauge progress in this area, it would be useful to report results with a

null-grammar (every word equally likely). Second, improved approaches to robustness and adaptation are needed. These might take advantage of more complex auditory representations to perform sound-source separation, make use of many redundant features, and be capable of determining when features are missing or masked. Finally, recognizers must be able to distinguish new from known words and they must be able to differentiate environmental sounds from speech. The neural network approaches described by Bourlard can serve as useful frameworks for solving these problems. Neural networks can integrate information from diverse feature sets, support global training to minimize word error rates and also to maximize new word detection rates, and help avoid the explosion of acoustic parameters that has occurred in large-vocabulary recognizers.

References

- [1] H. Bourlard, H. Hermansky, and N. Morgan (1996), “Towards increasing speech recognition error rates”, *Speech Communication*, in this issue.
- [2] R. Lippmann (1996), “Speech recognition by machines and humans”, submitted to *Speech Communication*.
- [3] G.A. Miller (1991), *The Science of Words* (W.H. Freeman, New York).