

Main conclusion: *Noise and filtering related robustness issues may be separated from context (language) issues, and may (should) be evaluated independently.*

The raw score before the context processing has important value in ASR evaluation. First it gives us a measure of the front-end, and second it allows us to evaluate the context processing. Here is the reasoning:

1. Context effects, as measured by the score, as a function of the task entropy, are strong [Miller et al.]. Context affects the score dramatically, for both HSR and ASR.
2. The high-context score (word, sentence) is a one parameter function of the low-context score (mean-phone). This parameter accounts for the entropy reduction due to context, and depends on the corpus [Boothroyd].
3. At both ends of the entropy scale (sentences having high context versus nonsense syllables), the relative performance, as measured by the ratio of the HSR to ASR probability correct score, is typically >10 . *For the case of nonsense syllables, this number is probably a very conservative lower bound.* Another way of quantify the relative performance is to find the SNR difference in dB such that the the HSR and ASR score is equal. This difference in SNR can be as much as 30 to 40 dB.
4. Since entropy is a critically important variable, it is necessary to control for it, by evaluating ASR and HSR under similar source conditions.
5. HSR is our gold standard.
6. Context processing is dependent on HSR robustness to noise and filtering in a manner that is highly unfavorable to ASR:
 - (a) The smaller the error at the input to the context processor, the better the processor will perform. Context models are much better at fixing a few errors than fixing many errors.
 - (b) Human context processing is superior to that of machine.
 - (c) HSR error is lower than ASR error at the input to the context processor.

⇒ Modeling context will not solve the ASR robustness problem until we improve the raw error rate. This error may be estimated from nonsense syllable error rates.

Discussion: Context cannot compensate for the poor performance of ASR acoustic front end. ASR's lack of robustness is directly traceable to its poor performance on low-context material. Using language models to obscure this error does everyone a disservice, as it tends to cover up and confuse the real problem, while making no fundamental contribution to the solution. *We must treat the disease, not the symptoms.*

HSR's low relative error to nonsense sounds means that for a given condition, context will be more effectively utilized in HSR in reducing the error than ASR. The lower the error, the greater the language advantage. Since human language process is superior to machine, this lower HSR raw error magnifies the effect of context processing.

Ideally, when controlling for context, we would like to remove all source entropy dependence. Nonsense syllables are one way of doing this since it was shown early that humans perform well with little to no context, even under conditions of filtering and noise. Both Fletcher and Miller have quantified HSR performance, in a controlled manner, under such conditions. It is difficult to find published ASR results for the high entropy source with filtering and noise, presumably because the performance is so bad.

In the one case that I measured with Mazin, for a wideband, no noise condition, machine error was about 50% when the corresponding human performance was about 10%. In this example I later found that *all* the HSR error was due to poorly formed subset of utterances. A large subset of the utterances have no (i.e., *zero*) HSR errors under the wideband quiet condition. While I have not yet evaluated the machine error on the well pronounced subset of sounds, I suspect that removing the slightly malformed utterances has modest impact on the ASR error.

Final statement: It is essential to start working with nonsense speech (CVC, CV, etc.), under conditions of noise and filtering, which is exactly where ASR dramatically fails. But then this is exactly the point. Doing so will increase our awareness to the magnitude and nature of the problem.

The present practice of training under noise, and then publishing the scores, is delusion. This practice should be discouraged if not stopped. Rather than wasting time and money in this exercise, the resources should be spent redesigning the acoustics to improve the robustness, so that the ASR score does not depend on filtering, and performs at realistic SNRs.

–Jont Allen, May 6, 2002