Show what you know: musings on the reporting of negative results in speech recognition research

Hynek Hermansky[1] and Nelson Morgan[2]

[1] *Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), Martigny, Switzerland*
[2] *International Computer Science Institute (ICSI), Berkeley, USA*

WHAT IS A "NEGATIVE RESULT"?

In a sense, well-designed experiments never have a completely negative result, since there is always the opportunity to learn something. In fact, unexpected results by definition provide the most information. Conventionally, negative results refer to those that do not support the hypothesis that an experiment has been designed to test; that is, results that are unable to disprove the null hypothesis (e.g., that the distinction between results from novel and baseline approaches can be explained by chance variability). Such a result can certainly be due to many causes, including bugs, and does not by itself confirm any hypothesis. However, learning about negative as well as positive results can be instrumental in providing the context for the development of new hypotheses to be tested. Hearing only about the successes is equivalent to throwing away half of the information. Personally, we have often been more intrigued with reports of significant unexpected failures than with the usual reports of method A being 5% better than baseline method B. Such reports often provide little surprise at all. We hope that the new journal will provide a forum for experimenters who have unexpected results from well-designed experiments.

WHITHER SPEECH RECOGNITION

More than thirty years ago, John Pierce wrote a short but very strongly worded letter to editor of the Journal of the Acoustical Society of America (JASA) (Pierce 1969). In the letter he questioned research in automatic speech recognition (ASR), which he viewed as more of an art than a science. At that time, the letter did not make him many friends in the field, given its critical perspective. The fury of ASR researchers may have been justified. Given Pierce's standing in the scientific community and his high-level position in the Bell Labs management hierarchy, the letter had an extremely negative effect. The letter was stark in its criticism, and its tone might have been more appropriate for a private communication rather than for a public forum such as JASA. While the letter and the attitude of its influential author apparently had some significant effects, research and development in ASR eventually recovered. So today we may have the luxury of taking some distance and perhaps even appreciate that many issues that were raised in the letter were valid (Jelinek 1996).

Particularly critical was Pierce's accusation that ASR researchers do not behave like scientists but rather like *mad inventors or untrustworthy engineers (sic)*. To avoid that, his advice was ... *If there was no clear experimental evidence, ... (one* It seems that Pierce's advice is still sometimes forgotten.

SCIENTIFIC METHOD

What is involved in designing a good experiment? Briefly, the scientific method requires:

1. Observe the phenomenon.
2. Form the hypothesis.
3. Make the prediction.
4. Test the hypothesis (run experiment to see whether you prediction is valid).

FIGURE 1. Just as it is easier to go down the hill, it seems safer to keep reducing the error rates.

5. If prediction not met, modify the hypothesis and go to 3.
6. Repeat the entire process.

The most important part of the method is in making the hypothesis and in designing experiments for testing it. Part of the current scientific practice is to publish results of experiments. In our field, how does our choice of publication topics relate to the scientific method? Going through a typical collection of ASR papers, e.g. in the IEEE Proceedings of ICASSP, one quickly finds that when the scientific method is at least partially followed, the hypothesis almost always turns out to be "the performance of the system improves". Even though this is not the only hypothesis that can be made, it is a perfectly valid hypothesis. However, when people make the positive hypothesis and the experimental results do not support it (it is "not a win"), why do we so seldom see this result published? It would be easy to say that the publication review process does not let such a result through, but it is our experience as reviewers and editors that we have rarely seen submissions describing negative results (without an additional redeeming positive result to ensure paper acceptance). So is the lack of reporting due to reviewer perspectives or to self-censorship on the part of the authors? We suspect that the latter is at least a major cause.

TOWARDS INCREASING ERROR RATES

Negative results can sometimes be an outcome of a conscious strategy to accept short-term losses in order to improve the long-term outcome. In the mid-90's a number of us argued that incremental system improvements might not be sufficient to significantly increase the capabilities of ASR technology, as we described in (Bourlard, Hermansky and Morgan, 1996). The paper was somewhat lengthy and focused to a large extent on approaches that we were proposing at the time, but we believe that its fundamental message is worth repeating here.

Suppose that there exists an analytic formula for error=$F(P)$, where P represents the vector of some system parameters that influence the final error rate. As long as the system function $F(.)$ has a single local minimum, the technique gradual modification of P while gradually decreasing the system error, should work. However, the $F(.)$ is almost certainly more complex than that. Then it is clear, that the strategy of the systematic error decrease is sub-optimal (Fig. 1) and that some temporal error increase may be necessary in order to reach the global optimum. Restricting ourselves from reporting results that move up on this error curve may discourage innovation.

WHAT COMES UP MUST GO DOWN

Admittedly, it is unlikely that any particular experiment will reveal the global optimum in ASR performance given the inherent difficulty of the problem. However, even if we would be satisfied to reach a local optimum, it would be extremely useful to know when we are getting out of such a "valley". In that case, why is it that we do not at least strive to demonstrate the limits of our claimed improvements? Why is it that one much more often sees Fig. 2a rather than the Fig. 2b?

A perfectly valid hypothesis of the scientific method would also be "the performance of the system gets worse". So why is it that we so seldom make this prediction? It is true that random errors are more likely to hurt performance than

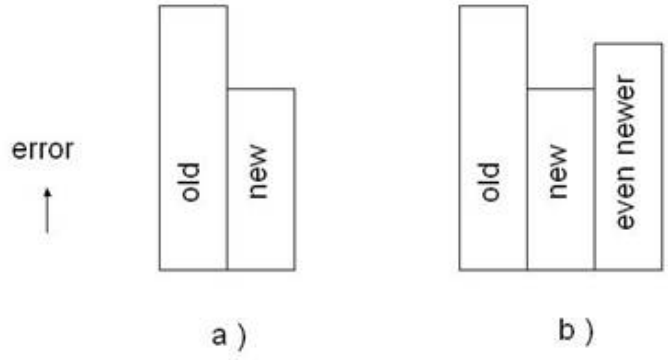

FIGURE 2. As with almost everything, there must be a point where any technique starts breaking. The point of the local optimum is of interest and may not be always reached (since it is not often reported).

to enhance it, thus leading to skepticism about the conclusions that might be drawn from a poor result. Nonetheless, repeated efforts can still lead to the conclusion that one has at least difficulty extending the range of improvement past the observed error minimum. Such results can be extremely helpful in guiding the direction of future research. Surely this information is commonly passed to colleagues within the same lab; how much better it would be for colleagues elsewhere to also know about technological limitations.

Of course, it is difficult to argue with the engineering goal of improving the performance of ASR systems. However, to investigate the range for which the claimed improvement applies is also more important. Why don.t we try to break our ASR system more often?

Our early schooling was in the days of the preeminence of analog devices, and for most applications the key consideration was to make sure that one operated in the linear part of device characteristics. One of us (Hynek) was fortunate in that one of his teachers insisted on pushing the limits of the experiment until the system broke, ("that is when things get interesting" (Pinos 1970)) and we believe that this advice is still valuable until today. To improve even a little on the best performance is good, but to fix things that don.t work at all is even better; this may not be possible without understanding where they don.t work.

AESTHETICS AND SCIENCE

In a question period at the most recent ASRU 2003 Workshop, Jordan Cohen asked where we could find "beauty in our field". What did he mean? We believe he was referring to the way the current ASR systems are put together as a collection of often mutually interfering modules, each of which "was a win" at the time of its introduction (meaning it improved the error rate), but without any globally optimizing principle. It may be inevitable that the best such systems will be complex, but it is hard to avoid the feeling that there should be some unifying principles (other than minimum error rate or maximum mutual information, etc., which are difficult to optimize over the entire heterogeneous system) that would lend greater coherence to both the systems and the experiments to improve them. We recall an earlier observation that "it seems we are attempting to do long division using Roman numerals". It is possible that a more complete disclosure of limitations of the technology might encourage the development of more parsimonious and effective models.

ARE WE WORKING ON THE RIGHT STUFF

From the early days of ASR, the primary task was to recognize one item out of the closed set of items. Some efforts have been applied to word spotting, in which rejection of other items was critical. What if all of the efforts from the field's start in the early 1950's had been focused on identifying elements of some vocabulary without restricting the input, both from other words and from nonspeech sounds? What if rejection had been a key issue even as we had made

the bridge from small orthogonal vocabularies upwards? We have sometimes mused about how the field might differ if this had been the emphasis.

Many authors have noted the large remaining gap between machine and human performance, especially when it comes to generalization to new unseen conditions (robustness). This certainly discourages their use in certain applications. Still, should we attempt to "achieve human level of performance" in ASR? We're not sure. Why should we follow the steps of Dr. Frankenstein (Shelley 1818) and attempt to build an artificial human being? Should not we rather work on useful tools that could potentially far exceed what human can do in certain situations? Nonetheless, there may still be enormous clues for solving our machine recognition puzzle that await us from improved understanding of the human mechanisms.

CONCLUSION

We have discussed a range of ideas in this piece, not all of which were fully coherent with one another. If there is a key point, though it is this: we are unlikely to blindly stumble across paradigm-changing improvements to speech recognition and other spoken language tasks by the "mad inventor" approaches derided by Pierce. It is necessary that we propose experiments to improve understanding of something, whether it is the function of human mechanisms for recognition or understanding of spoken language, or the availability of desired information in the speech signal or chosen representations for that signal. It is then necessary that we inform one another of the results of these well-formed experiments, whether they are positive or negative. The decision on publication of these results should be based on the importance of the question being asked and the skill applied towards answering it, rather than on whether the increment in a particular measure (such as word error rate) is positive or negative. It is our hope that this journal will encourage greater dissemination of "the other half" of the results beyond the lab-internal discussions that already occur.

REFERENCES

BOURLARD, H., HERMANSKY, H. and MORGAN, N. (1996) Towards increasing error rates, Speech Communication, V. 18, pp. 205-231.

JELINEK, F. (1996) Five speculations (and a divertimento) on the themes of H. Bourlard, H. Hermansky, and N. Morgan, Speech Communication, V. 18, pp. 242-246.

PIERCE, J. R. (1969) Whither Speech Recognition, J. Acoust.Soc. Am, V. 46, pp. 1049-1050.

PINOS, Z. (1970) Personal communications.

SHELLEY MARY WOLLSTONECRAFT (1818) Frankenstein, or, The Modern Prometheus.