Review and Comments on

Phoneme Recognition Based on Long Temporal Context

by

Petr Schwarz

The thesis in question was read thoroughly and completely by the committee member so charged. The following constitutes a review and comments on the document.

1. I find the thesis to be well organized, with a logical progression of analysis, a sufficient reference set, and an excellent development.
2. The English usage is sub-standard and creates a great deal of difficulty in reading and understanding of the thoughts expressed. Numerous spelling errors occur along with incorrect articles, grammar, and word choices. I was able to understand it nonetheless, but would suggest a copy edit by someone more skilled English before disseminating the document to other readers of English. I understand that English is not the author's first language and do not level this as a criticism of the work.
3. The basic plan of the research has been to utilize a well-known (to the author) technique called TRAP (temporal patterns) analysis. He extends it and then tries to optimize virtually every possible parameter for minimum error rates on phonemes for the TIMIT data set.
    - First, he compares his baseline to a classical MFCC system with multiple frames, and finds a 0.3 percentage point improvement. Although the claim is that this is significant, I do not know the level of statistical significance. *The author should describe the level of significance and what difference constitutes a .05 level of significance.*
    - He then optimizes the length and finds the system works best between 200 and 400 ms. (gain 2 more percentage points)
    - He then verifies that soft vs hard decision TRAP is best.
    - He then experiments with length for the individual bands and finds 500 to 600 ms is best with no real explanation.
    - He then experiments on DCT/no DCT and Hamming/no-window and finds the presence of both the help. (gain 0.2 percentage points)
    - Then the weighting shape is analyzed and an exponential window is found to be best. (gain 1.1 percentage points)

- Then he looks to see if the PCA projection is different from the DCT. Here he does not use the system of the previous bullets, but compares them in a different context. Apparently in this case the 0.2 percentage point improvement is NOT considered significant. Given the DCT and window analysis presented above, I find this analysis to lack rigor.
- He then tries a 3-band vs 1-band TRAP and finds 3 bands better. But the absolute numbers come from where? The 31.3 PER is not anywhere else I see. What happened to the improvements listed above?

This chapter has left me a little flat. There are many disconnected experiments that sometimes follow an order and sometimes do not. The optimization of the length parameter and subsequent experiments with window shapes (which have inherent effective lengths) really cannot be done is such a serial order if results are to be more than empirical.

4. The author then adds more complex state models in the next chapters, different ANNs., and bigram language models.
5. Training issues and size are examined.
6. Cross-data-set comparisons are made, along with noise and network perturbations.

In summary, I am not sure that I am tally convinced of the conclusions. The author comments that one earlier reviewer stated "the TIMIT was beaten to dead by this work." (Perhaps he means "TIMIT was beaten to death…") And this is pretty much true. TIMIT is a rather unique data set that has very little correspondence to modern problems of interest. Nevertheless, it is a standard that is used to evaluate phoneme recognizers. Given this, the author has performed a notable accomplishment at achieving the minimum PER. The generalization is not so clear, however. My experience is that parameter choices in ASR that work well in one context, often are not the right ones for another, and optimizing to tenths of percentage points on one particular set does not usually carry over to another. I would have been more convinced if the same optimizations worked in other contexts as well, e.g., Switchboard.

I also would like to have seen a complexity analysis showing CPU and memory usage as well as latency, with estimates of how much slower than real time the system could operate on general purpose computers. Also, is floating point necessary? Can the algorithm be distributed (i.e., can different stages be separated in time, with intermediate results stored or communicated separately?).

Judgment:

A PhD thesis in engineering must be judged by the following criteria in my view:

1. Has the student found an important and interesting problem to work on? **Yes.**
2. Has the student demonstrated a working knowledge of the state-of-the –art in his field? **Yes.**
3. Has the student built and constructed a system of hardware and software that can be used to reproduce/extend the state-of-the-art? **Clearly yes**.
4. Has the student progressed the state-of-the-art sufficiently for him to be considered the world's leading expert in the chosen area? **Yes, I believe he has.**
5. Has the student communicated the experimental strategy, the results, and the significance of the results in such a way that others might draw upon his experience? **Yes (despite some writing deficiencies).**

I therefore deem the presented thesis sufficient fulfillment of this requirement for the Doctoral Degree, and I should congratulate Dr. Schwarz for his work.

Mark A. Clements

02-FEB-2009