REPORT ON DOCTORAL THESIS

Title of the thesis: TRAP-Based Probabilistic Features for Automated Speech Recognition
Ph.D. candidate: Ing. František Grézl

Detailed Comments

Overall

Very good presentation of a large number of experiments. Pretty much every interesting in-scope combination seems to have been covered. I present a few overall comments next, then a few sentences on each chapter.

The thesis could have used another editing pass, especially by a native English speaker. The typos were few enough and English usage good enough not to directly interfere with understanding, but some sentences required several readings.

Organization and layout was excellent.

Citation style was somewhat inconsistent -- sometimes the authors are named and sometimes just the citation is given (e.g. "Bilmes and Yang have shown in [8]" vs. "This issue is discussed in [35]" vs. "modulation frequencies [50]"). Not a big deal, but something to watch for in the future. Citation density seemed about right.

I'm curious if experiments were repeated multiple times with different random starting weights for the networks. I've found that you can get up to 1% variability on Number 95 (a similar corpus) just with different initial weights.

Chapter 1

The summary of ASR in the introduction is strong and concise, though I have some minor quibbles -- e.g. the fact that single pronunciation dictionaries perform about as well as multi-pronunciation dictionaries contradicts statements in 1.1.3; section 1.1.5 presents only one type of decoder.

The introduction lacks up-front presentation of what this thesis is actually about. Though the abstract contains this information, it's usually a good idea to include it in both places, especially given the complexity of an ASR system. It would help a reader less familiar with the work to focus attention on the parts of the system that are going to come up later in the thesis.

In 1.3 and 1.4, it would be nice to summarize results. Give the reader some indication of what works and what doesn't.

Chapter 2

You learn something new every day! I didn't know the history of the term "critical band".

Minor quibbles
2.1.2: Another way to capture long time context using traditional features is with a context window...

2.2.1: Do you really use a sigmoid non-linearity at the output layer and not a soft-max?

Chapter 3

People unfamiliar with the task may not know the difference between "numbers" and "digits".

I like the careful separation of training data for the different tasks. One question: does the same speaker appear in more than one part?

Again, good concise description of systems (neural net, hmm, etc).

I'm a big fan of bootstrap estimates for confidence measures (e.g. Bisani and Ney's ICASSP paper)...

Chapter 4

Nice presentation of orthogonal experiments. Given the claim of the effect of noise on separate vs. joint dimensionality reduction, a forward reference to chapter 7 would have been nice.

A preview or forward reference to section 4.6 earlier in the chapter would have been nice as well.

Why was LDA computed on separate training data?

Chapter 5

A very nice way of thinking about modifications to the critical band spectrograms. I'm quite surprised that G3 did so well. Covers the range of single modification 1d operators and a selection of 2d operators.

Chapter 6

A selection of combinations and modification operators. Though not exhaustive, covers pretty much every interesting case. Nice compelling evidence for including frequency differentiation in combinations.

One question: In section 6.2, was the "average" just the arithmetic mean? Did you try geometric (i.e. log domain)?

Chapter 7

What's the confidence interval for this test set?

Lots of (good) results here, but I doubt the difference between most of the systems is really significant.

Including some state-of-the-art competitor's results would have helped calibrate the scale of the numbers. Is 30.8 average score good? Or does the lack of speech/non-speech, Wiener filtering, etc, make the comparison unfair?

Chapter 8

Using TRAPs based features on LVASR. There are, of course, only about a million parameters one could tune for these systems, but I was surprised at the choice of equal priors for phones in the Meeting system.

Maybe I missed it in the chapter, but was the Meeting data from head-mics or tabletop?

Setting up a full ASR systems is a huge, difficult task. I applaud the effort. Though it can be hard to judge the accuracy of the analysis when the system is far from state-of-the-art, most of the analysis here seems to be consistent and likely relevant for integration with other systems.

I debate with myself if state-of-the-art numbers should be included here. It's useful for people in the field to judge the ballpark compared to similar systems, but can be misleading if you're not familiar with the huge complexity involved in ASR systems...

Chapter 9

"Febstrum". Heh. Hadn't heard of that before...

Possible Defense Questions

Isn't speech recognition already solved? I'm only sort of joking. People not in the field are often under the impression that speech recognition is much further along than it is. If there are non-ASR folks on the committee, be sure they understand that the best systems in the world still get a third of the words wrong in general settings like meetings (and are 100x slower than real-time).

If the primary benefit of TRAPs is more temporal information, why not just use conventional features with wider windows or more context?

Why use neural nets? Wouldn't <<your favorite machine learning algorithm>> be better?

All the noise experiments were on artificially applied noise. What are the issues with "real" noise? Would you expect the system to work equally well?

The neural nets are currently trained on monophone posterior targets. The gaussian mixtures are (typically) trained on triphones. What do you think would happen if the two systems were trained on the same targets?

Answers to "Official" Questions

Is the topic appropriate to the particular area of dissertation and is it up-to-date from the viewpoint of the present level of knowledge?

Yes. The topic is directly relevant and timely.

Is the work original and does it mean a contribution to the area - specify where the original contribution lies?

The work is original and contributes directly to research (and commercialization, for that matter) of speech recognition. It presents a new class of features for use with ASR, and describes a host of experiments demonstrating strengths, weaknesses, and details of implementation.

Has the core of the doctoral thesis been published at an appropriate level?

Yes. A researcher in the field could replicate the results (albeit with some effort -- ASR systems are large and complex!).

Does the list of the candidate's publications imply that he is a person with an outstanding research erudition?

Yes.

Conclusion

I believe this doctoral thesis meets the requirements leading to PhD title conferment.

Adam Janin Ph.D. - International Computer Science Institute, Berkeley, USA
http://www.icsi.berkeley.edu/index.html.
janin@icsi.berkeley.edu