Summary of accomplishments

Dan Povey

Initial objectives (SGMM)

- I had done experiments with this style of model and had got very large improvements
    - 25% relative with ML training on 50h data
    - ~5-10% relative with ML training on 1000
    - Discriminative training -> smaller gains.
- Wanted to
    - Popularize the technique
    - Show even larger improvements with very small amounts of data (if trend held)
    - Use out-of-language data to train shared parameters

Initial goals (Lexicon)

- Nagendra had had experience with the difficultly and expense of obtaining a pronunciation dictionary (Lexicon) for new language
- Wanted to develop tools and techniques to make this easier or cheaper.
- Projects were merged under the common theme of "limited data".
- There was an open-source component: the plan to release the software developed in the project under an open source license.

SGMM accomplishments

- Demonstrated that the SGMM approach works: got better results than the baseline system.
- Developed new techniques in the SGMM framework:
  - Constrained MLLR estimation with parameter subspace
- Have worked on the basic SGMM approach:
  - E.g. looked at issues relating to poorly conditioned matrices that arise in estimation
  - Filled out the derivations

Multilingual applications

- Have demonstrated that it is possible to improve results for a group of languages by sharing non-state-specific parameters
- Better results than systems trained on the individual languages.
- Have demonstrated that with very small amounts of data (1h), we can dramatically improve results by using other languages to train shared parameters.

Open-source framework

- We have developed software that we can release.
- Software does all the required modeling, and does it quite efficiently.
- Need to do some cleanup, documentation and packaging before it would be useful to others.

Parts of the framework (1/2)

- HTK scripts to build baseline system (Martin, Samuel, Petr, Lukas…)
- Scripts to build language models based on SRILM tools (Samuel, Nagendra, Martin…)
- Scripts to improve lexicon coverage (Nagendra, Samuel, Pinar…)
- OpenFST based scripts (Lukas, Samuel) and C++ programs (Ariya) to generate WFST's for transcriptions and language models.
- C++ based code framework to read in HTK-like models and WFSTs and decode and train (Lukas, Ondrej, Petr,…)
- Matlab (Lukas) and C++ (Ariya, Tony) and HTK-based scripts (Mohit) to obtain UBM from baseline HMM set.

Parts of the framework (2/2)

- Matrix/vector template code wrapping various combinations of BLAS/ATLAS/CLAPACK (Lukas, Ondrej, Dan)
- Unit-testing code (Dan, Lukas, Ondrej,)
- C++ class for subspace GMM training and likelihood evaluation (Lukas, Dan, Arnab)
- Command line tools to use the above (Petr, Lukas, Arnab, Dan)
- Subspace GMM training and testing scripts (Petr, Lukas, Arnab, Dan)
- An entirely different HTK-based framework to implement this model (Rick, Shou-Chun)

Future of SGMM approach

- I believe the SGMM approach has a bright future
- It is a special case of a GMM which means basically all standard techniques apply.
- But extra techniques are possible with SGMM.
- It is more compact and gives better results
- But - it is more mathematically difficult which may limit the rate of uptake.
- In speaker identification, people were forced to use Factor Analysis style systems because they work better, even though they have the same issue.

Lexicon learning work

(summary of Nagendra’s summary)

- Have demonstrated that it is possible to do reasonably well starting from a 1000 word pronunciation dictionary and using g2p to get the remaining words.
- Should be useful in languages where there are relatively few resources
- We plan to include the scripts and tools developed as part of the software setup we release

Further reading and resources

- For a (long) tutorial introduction to SGMMs: http://dpovey.googlepages.com/ubmtutorial.pdf
- For pre-release versions of our software, contact dpovey@microsoft.com (because I always respond promptly to email; but the software is based in Brno).
- We will most likely make two or three journal and/or conference papers out of this work.
- Many of us plan to continue research using the systems we developed.