PROJECT OVERVIEW

Project

- **Aim:** *Build a small ASR (character level) for your own language with only 1 hour of labeled data!*
- **Preparation (tutorials)**
  - Create your own 2h dataset individually (DONE)
  - Learn how to pretrain a CPC system (DONE)
  - Fine tune it using CTC (DONE)
  - Compute a CER (DONE)
- **Project**
  - in small teams (2-3) -- can also be done individually
    (the data has to be collected individually, but the code can be done in small group)
  - Prepare a train and test set split for each of your own data (e.g. 1h/1h); further, prepare a 20% validation set from the training set (early stopping, etc)
  - Take a pretrained CPC and fine tune it on your train set with CTC, compute the CER
    - If you are in a group, you can start with one of your dataset, and then reapply to the other(s) in order to compare the results

- Try one or several additional optional steps to improve the results:
  - Change the architecture
  - Data augmentation for the fine tuning (you can use noise datasets like MUSAN)
  - Multingual fine tuning by using some datasets of the rest of the class
  - Multilingual pretraining (using CPC) with these datasets
  - Study the effect of language proximity
  - Add a LM and compute WER
- Write a 2-3p report
  - Explain what you did, the problems encountered and steps to address them
  - Figures and tables to illustrate your results (you can use an appendix if it does not fit into 3 pages
  - Include some errors analysis
  - Summarize what you've learned from this project
  - State what directions you did not have time to address and which ones seem most promising

- Provide and document your code
  - Ipython notebook / Github
  - Readmes and comments
- Provide your dataset
  - See how on https://github.com/besacier/AMMIcourse/
- Deadline: Friday July 3!

Good luck!