HLT: Speech & Language Technology: Topics for Research and Bachelor & Master Theses

(email: zeineldeen@cs.rwth-aachen.de, ney@cs.rwth-aachen.de **)**

framework:

- **helpful: courses on speech & language technology and on machine learning (beyond tool using and programming) or equivalent knowledge**
- **understand the worldwide state-of-the-art (by experiments and reading papers)**
- **learn to use, extend and/or build software tools (both public and RWTH-HLT)**
- **work in a team (HLT RWTH, AppTek Aachen; UP Valencia for HTR; ...)**
- **work on (difficult) real-life data (public and company)**
- **challenge: generate insights or competitive results that are relevant for the worldwide scientific community**
- **typical procedure: start as a student worker (’Hiwi’ job) to prepare (8-12 weeks) your thesis registration**

HLT: Speech & Language Technology: Topics for Research and Bachelor/Master Theses

(tentative) directions for research and BA/MS theses
(from theoretical to practical aspects):

- **classification error bounds:**
  - **framework: Bayes decision theory:**
    mismatch conditions, string context, training criterion, language model, ...
  - **objective: understand the (practical) error rate from first principles**
  - **area: basics of machine learning and information theory/statistics**
  - **approach: computer simulations and (ideally) analytic calculations**
- **unsupervised training (in the strict sense):**
  assumption: *perfect* LM and huge amounts of speech data without transcriptions
  question: how to build an ASR system?
  challenges:
  - **what is a suitable training criterion? relation to classification error?**
  - **computational complexity? combinatorial search?**

  some relations to pre-training, semi-supervised training
  and decipherment (homophonic substitutions)

- **LLM: large language models**
  possible directions:
  – mathematical improvements and refinements
  – supervised fine-tuning
  – reward modelling and reinforcement learning
  – RAG: retrieval augmented generation
  – specific variant: LLM for dialog
  data and (real-life) tasks: public and AppTek
- **(regular) language models for ASR:**
  measure WER as a function of language model perplexity:
  – acoustic models: hybrid HMM, CTC, transducer, attention
  – language models: count-based, LSTM RNN, transformer
  verify: no search errors
- **verifying training speech data:**
  – challenge: improve existing ASR by filtering speech training data
  – approach: use combination of WER and recognition scores on transcribed data
  some relation to semi-supervised training

- **extending seq-to-seq architectures (attention, FST/HMM):**
  challenge: improve existing high-performance systems
  – attention: strict monotonicity
  – segmentation and streaming (as opposed to batch model)
  – FST/HMM: combination with attention-like structures
- **handwriting text recognition (HTR, 'OCR')**
  transfer concepts from ASR to HTR
  – starting point: RWTH's FST-HMM engine for ASR
  – switch from acoustic to image features
  – define/select suitable HWR tasks (English, Arabic, Chinese, ...)
  ideally: cooperation with UP Valencia