SPRAAK: an open source “SPeech Recognition and Automatic Annotation Kit”
Kris Demuynck, Jan Roelens, Dirk Van Compernolle, Patrick Wambacq
Katholieke Universiteit Leuven, Belgium
kris.demuynck@esat.kuleuven.be

SPRAAK is a toolkit intended for
• Speech recognition research
• Development of speech recognition applications (niche markets)

Licensing
• Free source code license for academic usage (open source concept)
• Commercial license

Development History
• Derived from HMM75 that has a >15 year development history at KULeuven
• Modernization & Conversion was supported by the STEVIN programme
• Development Partners: Radboud U. Nijmegen (NL), Twente U. (NL), TNO (NL)
• SPRAAK and/or HMM75 widely used by Belgian & Dutch universities
Software Environment

- Software modules [for efficiency]
- Python scripts [as glue, fast prototyping, ...]
- Development platform: Linux + gcc + Python
- Doxygen for automatic extraction of programmer’s manual from the code
- Run-time platforms:
  - Linux, Windows
  - Other Unix platforms (including Mac OS X) feasible with much effort

Programming philosophy

- Large components “preprocessing, acoustic models, language models, decoder” act as objects with well defined, very flexible, interfaces
- Plug and play concept
- High reusability of underlying modules
- Multi-threading safe (everywhere)
- Multi-processing ready for all time critical functions (e.g. Viterbi)

Interfaces

- File Formats: Proprietary & support for HTK (AM) and SRI (LM)
- Low level API:
  - access to low-level routines giving full control
  - can be accessed directly in high level scripting via Python interface
- High level API:
  - control over run-time functionalities
  - client-server model + equivalent C functions

Functionality, highlights

- Implementation of modern probabilistic speech recognition framework
- Feature extraction:
  - Many building blocks built in
  - A scripting language to construct new feature extraction schemes starting from elementary building blocks
- Acoustic modeling:
  - Efficient tied Gaussian system
  - MIDA feature optimization (mutual information based maximum phone discrimination)
- Feature decorrelation
- Pronunciation modeling:
  - Cross-word context-dependent quin-phones
  - Assimilation rules
  - FST encoded lexicon with prefix and suffix sharing
  - Weighted pronunciation variants
- Language Modeling:
  - N-grams with N>3 (virtually no constraints)
  - BNF grammar compiler for FSG’s
- Text normalization toolkit
- Training:
  - Distributed training on multiprocessors and multiple hosts
  - Automatically resuming of experiments after computer/software crashes
- Decoder:
  - lexicon and CD tied state information is combined into a compact optimized FST
  - LM (N-gram, FST, ...) is applied on-line by multi-tag token passing decoder

Reference Implementations (SPRAAK or HMM75)

- AURORA-4 clean speech reference (WSJ0): 4.9% WER
- TIDigits: 0.17% WER using word models (11-17 states)
- N-Best
  - Flemish Broadcast News (dev-07): 13%
  - Commercial Greek Dictation System with 540k words

Status & Future Plans

- Completed
  - code conversion + programmer’s manual
  - Windows version
  - regression tests w.r.t. HMM75 on classic benchmarks
  - train/test on new data/benchmark (N-Best)
- In progress / ‘test’-phase
  - High level API for client server implementations
  - Parts of the user’s manual
- Future Developments
  - ‘Flavor’ (graph based) decoder
  - High-level flow-chart scripting language to configure multi-stage recognizers, parallel recognizers (rover alike) with the available processing blocks
  - Missing data based feature extraction