An Overview of the ABCP1 Recognizer

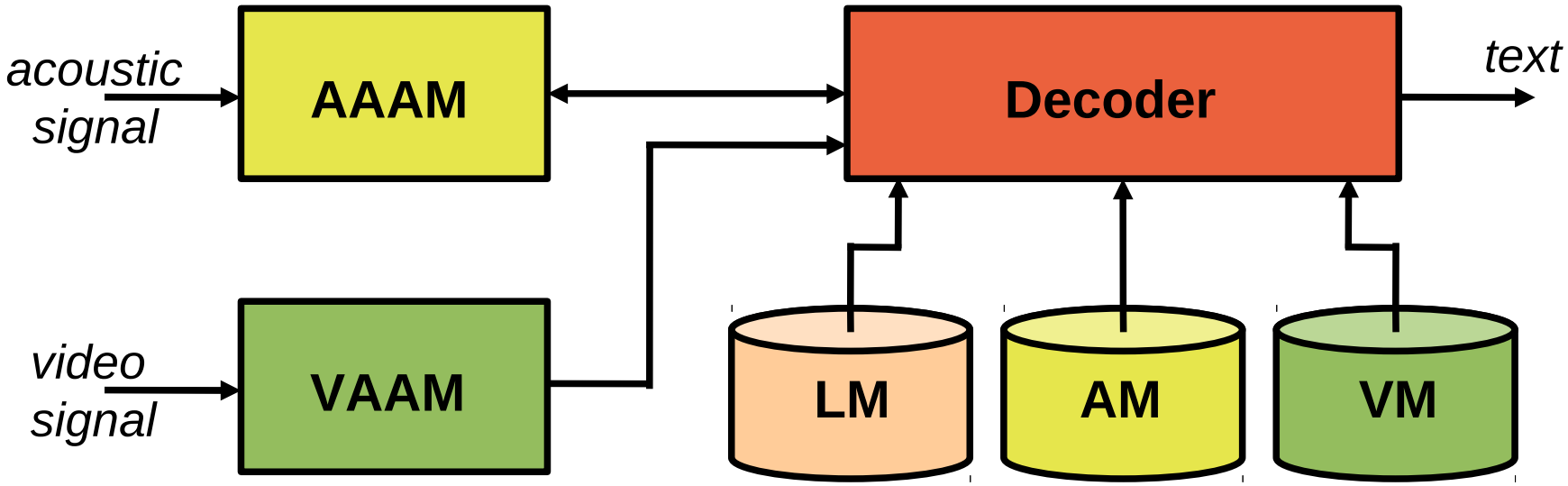
Summary

- Speech Application
- System Structure
- System Modules
  - Speech Acoustic Acquisition & Analysis (SAAAM)
  - Speech Visual Acquisition & Analysis (SVAAM)
  - Acoustic Model (AM)
  - Visual Model (VM)
  - Language Model (LM)
  - Decoder (Dec)

Speech Application

- Continuous Speech – (European) Portuguese language (paused reading-like utterances)
- Speaker dependent (eventually w/ non-severe disarthria)
- Medium/Large vocabulary: up to 20K words
- LM with (real) PP approx. 100
- Use of both Acoustic and Visual speech cues
- Real Time operation (std PC (2012))
- Performance level w/ WER bellow 5%

Main Structure



The ABCP1 Modules diagram

Acoustic Acquisition & Analysis (SAAAM)

- Operation modes
  - Offline
  - Online (~½ sec buffer w/ Dec control signal)
- RTAudio platform (std params)
- Analysis
  - MFCC + E + 1st derivatives
  - (Low-level) + LLE algorithm
- Future Work (exp. other analysis methods)

Visual Acquisition & Analysis (SVAAM)

- Acquisition (V4L/OpenCV w/ std params)
- Face detection (w/ Boosted Cascade Classifier (BCC))
- Two-pass ROI tracking method
  - Pass-1 quick (BCC method) approx. ROI:
    - auxiliary face marks tracking
    - geometric normalization (face rotations)
  - Pass-2 accurate ROI (Template Matching method over approx. ROI)
- Analysis
  - DCT2 coef.s + basic geometric features
  - Exp. and compare w/ higher level available methods/tools
- Pos-processing techniques
- Future Work (aux. face marks robust meth.s; other analysis meth.s)

Acoustic Model

- #38 phone-like (EP) basic units
- Two SC-HMMs sets
  - “Main” AM w/ approx. 1K tri-phones
  - “Broad” AM w/ approx. 40 mono-phones (w/o context)
- Techniques to speedup computations
  - Initial log-LL approx.s using w/o context units (“broad” AM)
  - Multilevel Prunning
- Effects (+) of Speaker Dependency
- HTK - the main development tool
- Future Work (non-implemented known progressive refinement strategies)

Visual Model

- Visyllabic units
  - Vocabulary based on mapping phones/visemes into syllables/visyllables
  - Comparing w/ visemes: (pros) context robustness; (cons) hundreads (in EP NOT prohibitive; and Dec method “helps”)
- Visyllables modelled w/o context through SC-HMMs
  - Good discriminability x robustness trade-off
- HTK - the main development tool
- Future Work (Improve non-fully automatic existing syllabification process; New techniques using visyllables at decoding pass-2)

Language Model

- Pass-dependent LMs
- LM1 (Pass-1):
  - Lexicons (acoustic): up to 20K entries, based on (#38) phonelike units; tree structured and hybrid tree/linear lexicons available
  - Syntactic grammars: unigram w/ word-pair restrictions
- LM2 (Pass-2):
  - Lexicons: Linear (acoustic) phone-based; plus Linear Visyllabic
  - Grammars (several versions: training params, etc.): bigrams, w/ PP near 100 (lower, over shrinked search spaces); exps. w/ artificial low-PP trigrams
- Hybrid grammars: standard word-sequences frequencies combined w/ morphosyntactic plus gender and number inflections information
- CMU-Cambridge Language Model toolkit – the main used tool
- Future Work (improve Hybrid grammars; LM adaptation new approach)

Decoder Module

- Two-passes approach – why?
  - Streams (Acoustic/ Visual) differences in Nature; Discriminability; Temporal resolution; Misalignement variability; (open-)systems such as Julius w/ very good performance
- Pass-1
  - AM only (acoustic-cues) + LM1
  - search lexicon-tree w/ Viterbi alg.
- Pass-2
  - “Ignition” criteria (acoustic signal behaviour and time restrictions)
  - AM + VM + LM2
  - search word-trellis w/ stack decoder alg.
- Future Work (computation acceleration; address “near” single one-best)

The End