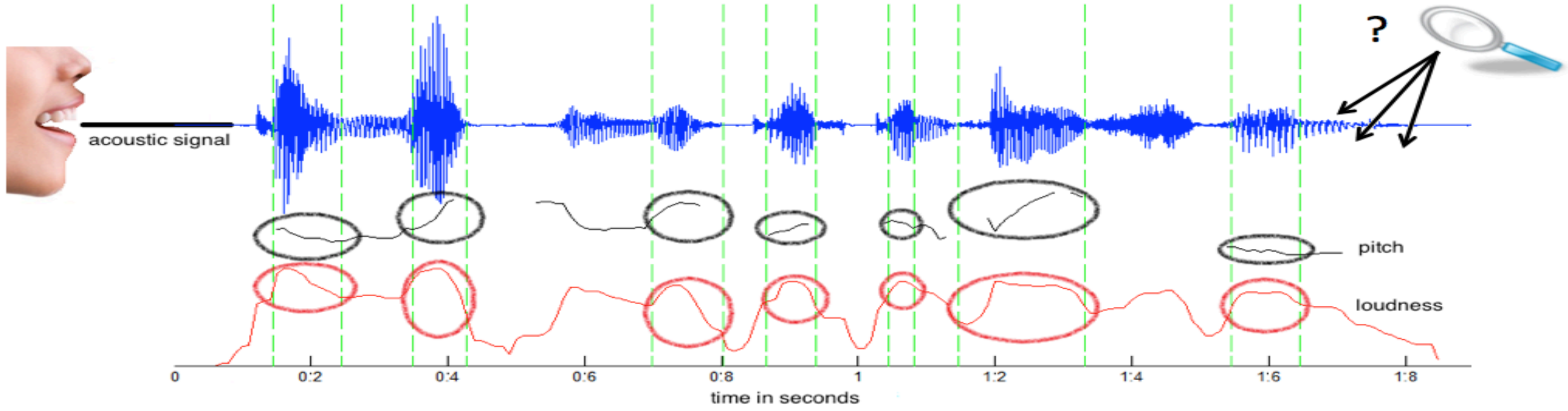
DEEPSER-toolkit

Jaebok Kim
Human Media Interaction
University of Twente

Goals

- Off-the-shelf training models and building speech emotion recognition (SER) applications
- Customizing own models and reproducing experiments
- 100% python codes for soft programmers



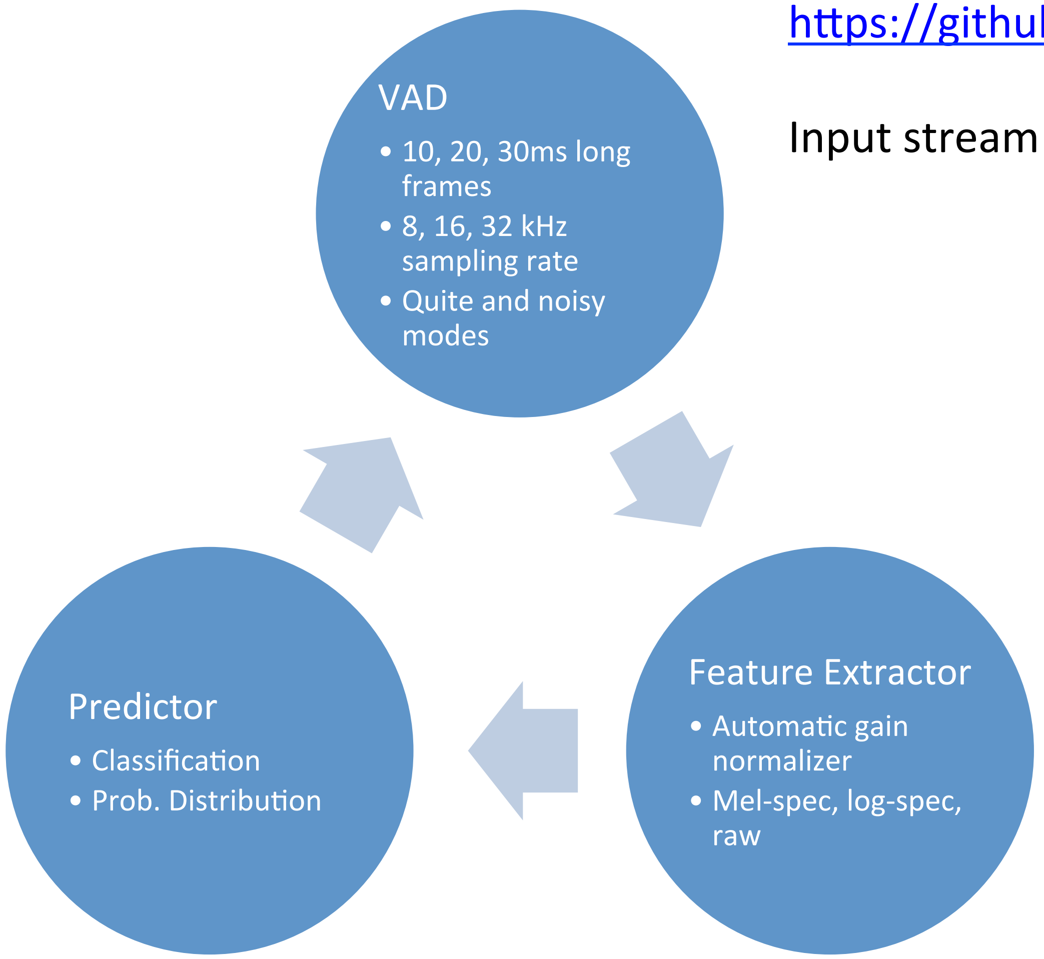
Requirements

- OS
  - MAC OSX >= 10.10.5 (Yosemite)
    - Required packages will be installed by “brew”
  - Ubuntu >= 16.04
    - Required packages will be installed by “apt-get”
  - Windows ? Not tested but may work…
- Python
  - Tested on 2.7x and 3.5x
  - Required packages will be installed by “pip”
- Details of installation can be found in README.md files of each git-hub repository.

Components

- Recognizer (https://github.com/batikim09/LIVE_SER/)
  - Recognizing emotion by using voice activity detection and a trained keras/tensorflow model
- Trainer
  - Extracting features, building and optimizing a keras/tensorflow model

Recognizer



https://github.com/batikim09/LIVE_SER/

Input stream can be both file and live microphone.

An example script (OSX)

- Find your available microphones
  - python ./src/offline_ser.py
- Run recognizer
  - python ./src/offline_ser.py -d_id 1 -p_mode 1 -f_mode 1 -log ./output/live.wav.csv -md ./model/AIBO.si.ENG.cw.raw.2d.res.lstm.gpool.dnn.1.h5 -c_len 1600 -m_t_step 16000 -tasks 'arousal:3,valence:3' –vd 1000 –s_ratio 0.7

-d_id 1: you find your device index is 1

-p_mode 1: classification mode

-f_mode 1: raw wave form, depending on your trained model

-log …: store all recognized results into a file

-md …: your trained model

-c_len 1600: time-steps in each contextual window

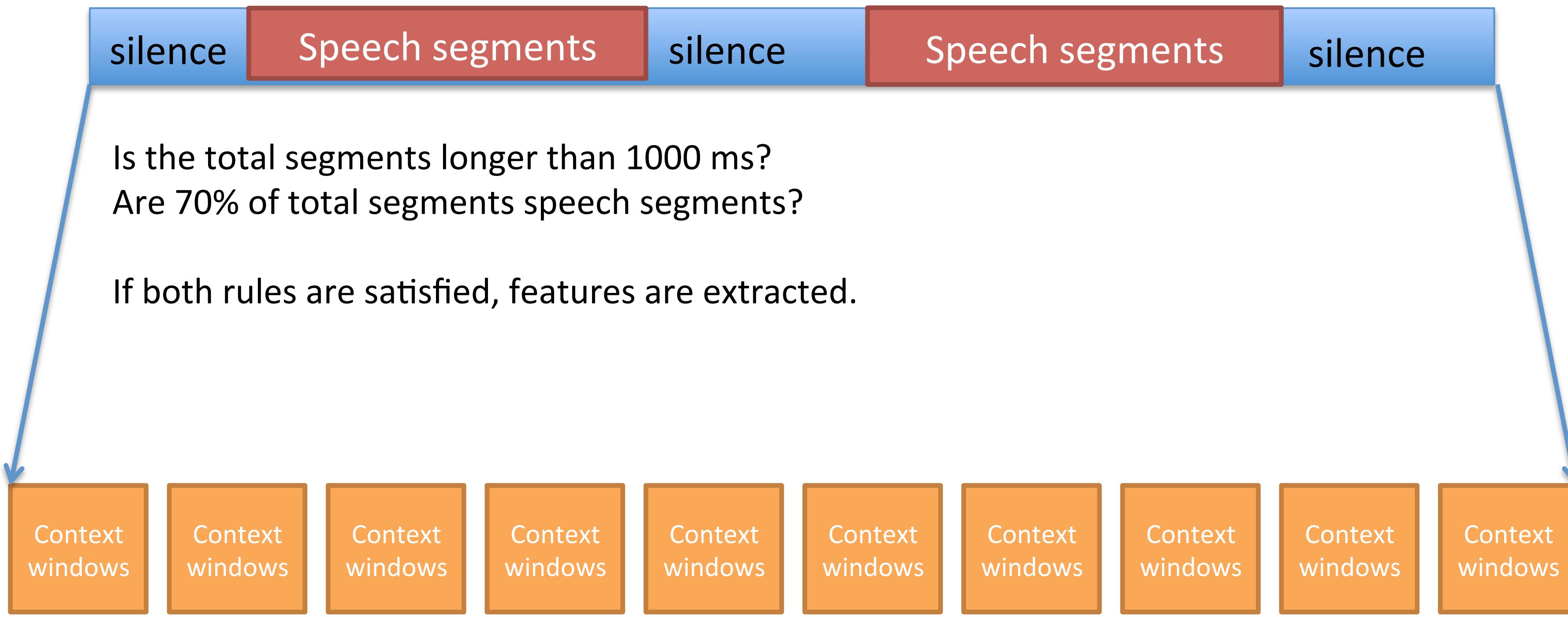
-m_t_step 16000: maximum time-steps for each utterance

-tasks …: your classification tasks and their number of classes

-vd 1000: minimum duration of speech to recognize

-s_ratio 0.7: minimum proportion of speech segments in total segments

VAD and feature extraction



Is the total segments longer than 1000 ms?
Are 70% of total segments speech segments?

If both rules are satisfied, features are extracted.

The time steps in a contextual window and maximum time steps per utterance depend on features & model.

For example, log spectrogram windows have 10 time steps of each 25ms long frame but overlaps, 10 windows per utterance make maximum time steps 100 (~= 1sec).
Raw form windows have 1600 time steps (1600 samples ~= 100ms), 10 windows per utterance make maximum time steps 16000 (=1sec).

Pre-trained models

- Two English models provide arousal and valence (3 class each) predictions
  - ./model/AIBO.si.ENG.cw.raw.2d.res.lstm.gpool.dnn.1.h5
    - Raw, RESNET-LSTM-DNN
    - Based on “Deep Temporal Models using Identity Skip-Connections for Speech Emotion Recognition, ACMMM17”
  - ./model/si.ENG.cw.mspec_mm.3d.rc3d.1.h5
    - Mel-spec + RESNET-3DCNN
    - Based on “Learning spectro-temporal features with 3D CNNs for speech emotion recognition, ACII17”
- Their performances are more less same.

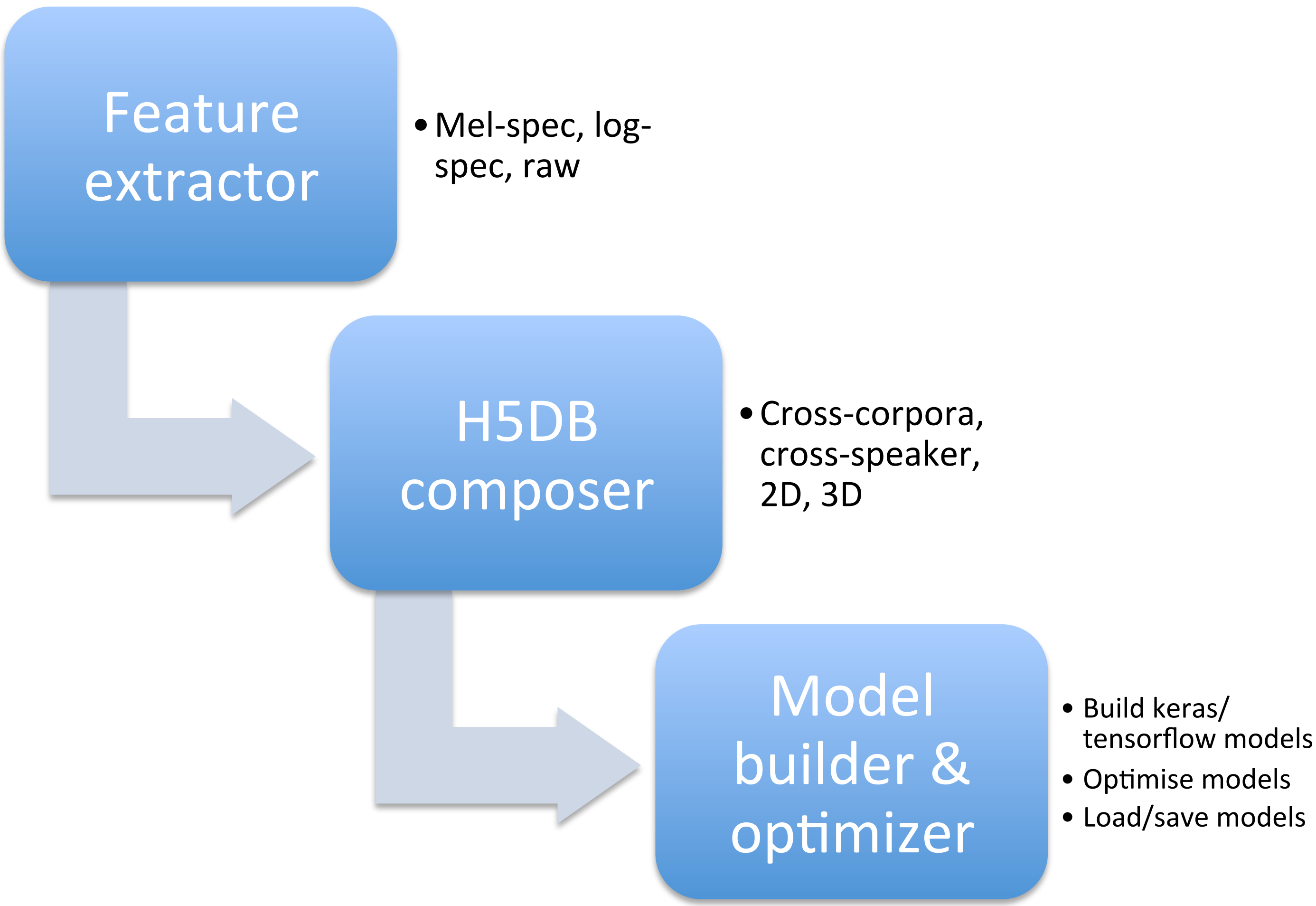
Automatic Gain Normalization

- Gains are really crucial to performance.
- After starting the script, practice several utterances with various gains.
- It collects gains and finds minimum and maximum gains for min-max normalization.

Next steps

- Nice demos, visualization? Students' projects?
- Provide more models recognizing various contexts: gender…

Trainer (TBD)


Feature extractor
• Mel-spec, log-spec, raw
H5DB composer
• Cross-corpora, cross-speaker, 2D, 3D
Model builder & optimizer
• Build keras/ tensorflow models
• Optimise models
• Load/save models