The ASR System for the EML LISTEN Demonstrator

Wei Zhou

European Media Laboratory GmbH

www.eml.org

Introduction

- Me

  - 2015: Master of Science in Communication and Multimedia Engineering
    University of Erlangen-Nuremberg, Erlangen, Germany
  - since then: software engineer at EML European Median Laboratory
    - real-time speech recognition engine
    - acoustic and language modeling for Mandarin and Cantonese
    - PhD: RNNLM for ASR

- LISTEN project

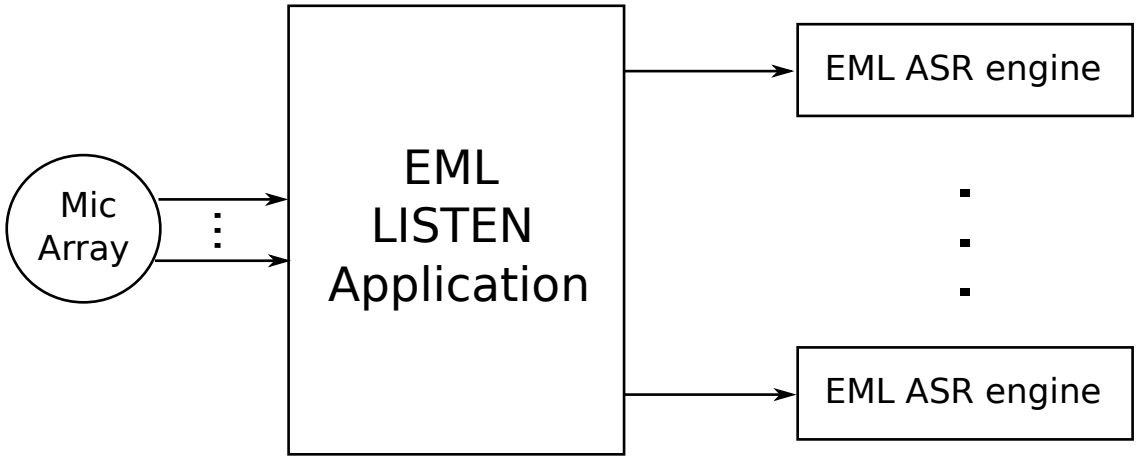
  - hands-free voice-enabled interface to web applications for smart home environments
  - 2015.06 – 2019.06
  - Cedat85, EML, FORTH, RWTH

Outline

- System Overview
- Voice Activity Detection
- Acoustic Model
- Multiple Search Spaces
- Outlook

System Overview

- EML LISTEN demonstrator



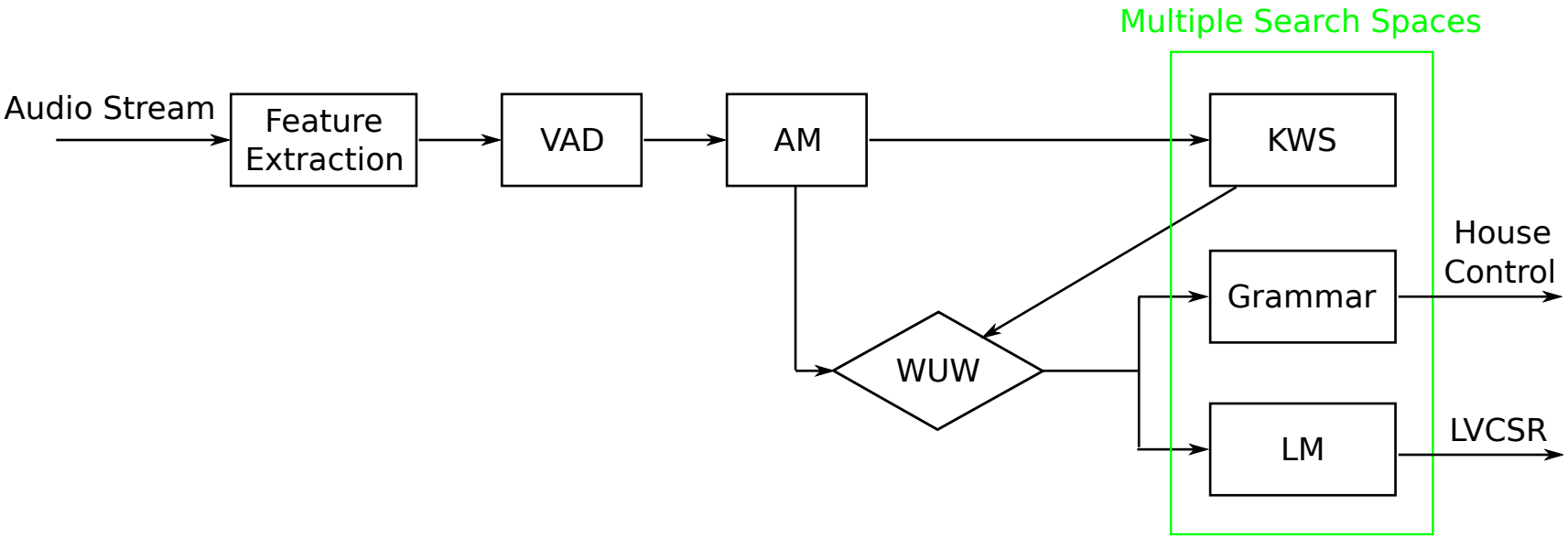
- Mic array

  - real-time multi-source DOA estimation and beamforming

  [D. Pavlidi et al., 2013, 'Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures']

System Overview

- Single engine:



- Dynamic network decoder

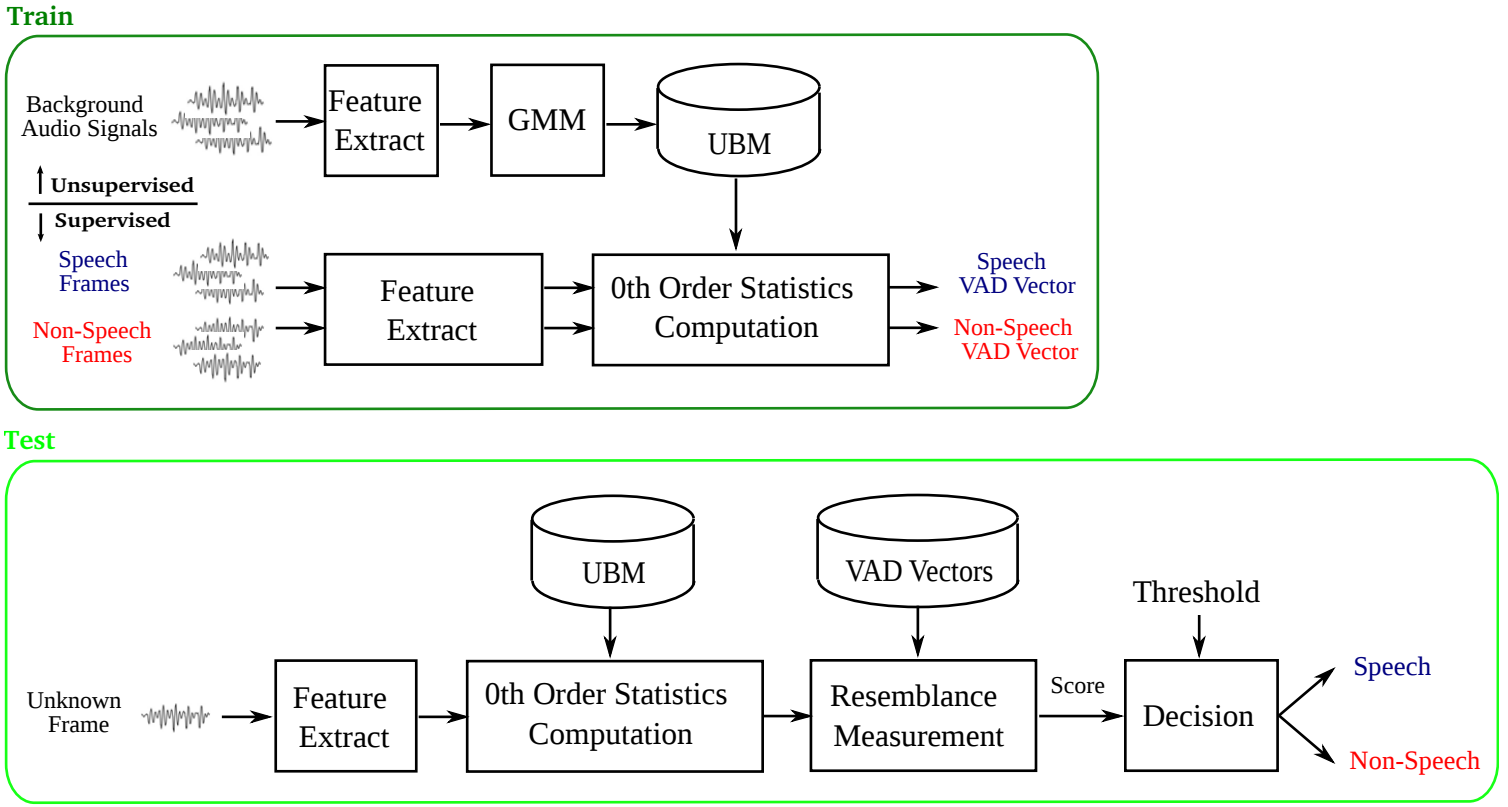
  - history conditioned tree search

  [D. Rybach et al., 2011, 'RASR – The RWTH Aachen University Open Source Speech Recognition Toolkit']

- Always listening mode

Voice Activity Detection

- Online VAD



- decision every segment of 0.1 – 0.3s
- very low computation load

[O. Gahabi et al., 2018, 'A Robust Voice Activity Detection For Real-time Automatic Speech Recognition']

- Accuracy: for 20 frames segment about 15% EER

Acoustic Model

- BLSTM
  - 5 layers (each of 1024 units)
  - 4500 outputs (tied triphone states)

- Bilingual
  - German: about 600h
  - English: about 600h

- RETURNN
  - optimizer: adam
  - dropout: 0.1
  - regularizer: L2

[P. Doetsch et al., 2016, 'RETURNN: The RWTH Extensible Training framework for Universal Recurrent Neural Networks']

Multiple Search Spaces

- Parallel multiple back-end search
  - asynchronous running
  - share same front-end and acoustic scoring

- Grammar and Language Model (LM) based search
  - idle before wake-up word is detected

- Keyword Spotting (KWS)
  - simple WFSA with garbage modeling and only one keyword
  - run time dynamic lexicon update: arbitrary wake-up word
  - fast (0.2 RTF) and low computation load
  - accuracy
    - 'Scotty': 94% precision and 94% recall
    - 'Elisa': 98% precision and 89% recall

Multiple Search Spaces

- Grammar (house control)
  - simple FST with garbage modeling
  - very small vocabulary: German 100 , English 100
  - fast: about 0.5 RTF and <1s latency
  - noise robust: restricted paths
  - accuracy: about 90% action complete rate, 98% correct rejection rate

- Language Model
  - 4-gram class-based bilingual LM
  - large vocabulary: German 680k , English 440k
  - speed: about 1.5 RTF
  - accuracy: about 30% WER (data mismatch)
  - simple action parser: about 79% action complete rate

*[H2020-MSCA-2014-RISE 644283 **LISTEN D4.2**: 'Report on recognition evaluation, technologies, tools']*

Outlook

- Better LM
  - in domain data
  - RNNLM
- Online acoustic adaptation
  - environment
  - speaker
- Privacy
  - speaker verification: only registered user

Thank you for your attention!