From a Sketch to a Real LISTEN Demonstrator

Jozef Ivanecký

www.eml.org
Who I am …

• Studies:
  - PhD in Speech Recognition, TU Košice, Slovakia
  - MsC in Electrical Engineering at Department of Cybernetics and Artificial Intelligence, TU Košice, Slovakia

• IBM
  - Embedded Via Voice development, Watson Research Group, Czech Republic
  - Acoustic modeling for embedded German and ML ASR systems, IBM R&D, Germany

• EML
  - Language modeling tools
  - Embedded ASR
Outline

- Plans
- Technology
- Integration
- Platforms
- Demo
• Plans for the demonstrator:
  - Always listening mode
  - Multiple audio sources (strong background noise)
  - High accuracy
  - Bilingual (if possible)
  - Real time recognition
  - No active cooling
Technology

- Key technological enhancements:
  - Stream based segmentation
  - Stream based speaker diarization
  - Improved voice activity detection (VAD)
  - Partial traceback support
  - Dynamic lexicon adaptation
  - Grapheme-to-Phoneme (G2P) service
  - Always-Listening Mode
  - Multiple Search Spaces
Voice Activity Detection

• Efficient filtering non-speech segment before processing the audio signal by the decoder.
  Load for Intel NUK:
  - Speech segment, KwS: 20-30%
  - Non speech segment: 3-5%
• Improved WER
• ALM trigger
Always Listening Mode

• A prerequisite for any hands-free voice enabled interface
• New extension of EML decoder
  - Utilizes:
    - VAD
    - Partial traceback
• Just Key word spotting
Multiple Search Spaces

- The demonstrator uses 3 search spaces (SS):
  - SS #1: Key word spotting – with dynamic lexicon update
  - SS #2: Large vocabulary language model
  - SS #3: Small vocabulary grammar (FST)

- After initialization:
  - SS #1 – ON
  - SS #2 & SS #3 - OFF
Multiple Search Spaces

• Why not just 1 SS?
  - Resources saving: Large vocabulary LM vs. simple KwS system
    - KwS load for Intel NUK: 20-30%
    - LV LM load for Intel NUK: 100%
  - Privacy issue

• Why not just 2 SSs?
  - Small grammar - better accuracy for noisy environment:
    - In a quiet environment SS #2 and SS #3 produce usually identical result
    - In a noisy environment are SS #3 results better
    - SS #3 is faster
Hi Scotty, all lights in Jozef’s room on.

KWS

LM

FSA
Demonstrator v1

- The first version of the demonstrator was using two recognition engines:
  - Engine #1: SS #1 - KwS + small PCM buffer
    - ALM mode
  - Engine #2: SS #2 + SS #3
    - P2T mode – controlled by Engine #1
    - VAD included
  - Different AM for Engine #1 and Engine #2
Demonstrator Architecture v2+

Audio client with multiple streams

Audio inputs manager

Audio data collector

EML decoder – N
EML decoder – ...
EML decoder – 2
EML decoder – 1

Search space #1
Search space #2
Search space #3

Result processor N
Result processor ...
Result processor 2
Result processor 1

Demo Application

House Control
Web Applications
Platforms

• Several options
  - EML Transcription platform
  - Intel based NUK 3 – current demo choice
  - Cortex A-9 based Odroid U-3 – for a limited version
Audio Clients

- Simple directional microphone
- Microphone array
  - Sources tracking
  - Sources separation
• Microphone array with multiple audio sources
• Bilingual system: German-English
  - Key word: [Hi/Hallo] Scotty
  - LM vocabulary size: 680314+438061
  - Grammar vocabulary size: 109+96
• Platform: Intel NUK 3
  - 32G RAM
  - 4 cores i7-7567U CPU @ 3.50GHz
Thank you for your attention!