Data Collector for Multi-lingual ASR Engine

Capstone Group S19-38

Team Members:

Mo Shi, Chaoji Zuo, Zekun Zhang, Ziqi Wang, Duc Le

Advisor:

Dr. Shahab Jalavand from Interaction LLC.

ASR & Problem Description

(Automatic Speech Recognition)

Extract the text from audio. Speech, product launch, video subtitle, etc.

Core **models: acostic** model, **language** model

Relative **algorithms: Hidden Markov** Model(HMM), **neural networks**

1. Can our system works on different languages?

2. What kind of language should we choose?
   a. too many users?
   b. too limited?(so few users, doesn't even have its own charactors)

3. where to find the resources?

   Target: make sure found enough resources to train ASR system

Finding Resources & Constructing Dataset

- Data we need : **Audio** with corresponding **Abstract** or **Transcript**
- Goal Languages (from different languages categories) :
  - Italiano, Hindi, Cantonese
- Type of Websites :
  - News websites, Languages learning websites, Video websites, e.g. YouTube, TED Talks
- Methods of constructing dataset
  - **Web Scraping: extract** data from websites and **save** to database **automatically**.
  - Focus is on websites that have **both the video and text.**
  - **Length** of video and text need to be long enough.
  - **Forced alignment:** match the words in the text to the corresponding part of audio.
  - Each member is responsible for one website.

Scope of Work

- Finding resources - 1 Week (2/11-2/17)
- Building small programs & Primary estimation - 2 Weeks (2/18-3/2)
- Integration & Final estimation - 1 Week (3/3-3/10)
- Advanced target: Generalization & ASR related learning (3/11-)

General Plan & Allocation

- Each one find his own resource website & design a specific program to obtain the data set
- Try to combine all programs together into a gathered version. i.e. building a template that works for all small programs
- Estimation & Generalization

THANK YOU!