Web Crawler for Multilingual ASR Engine

Speaker : Mo Shi, Zekun Zhang, Ziqi Wang, Chaoji Zuo, Minh Duc Le

01

Introduction

01 About Us

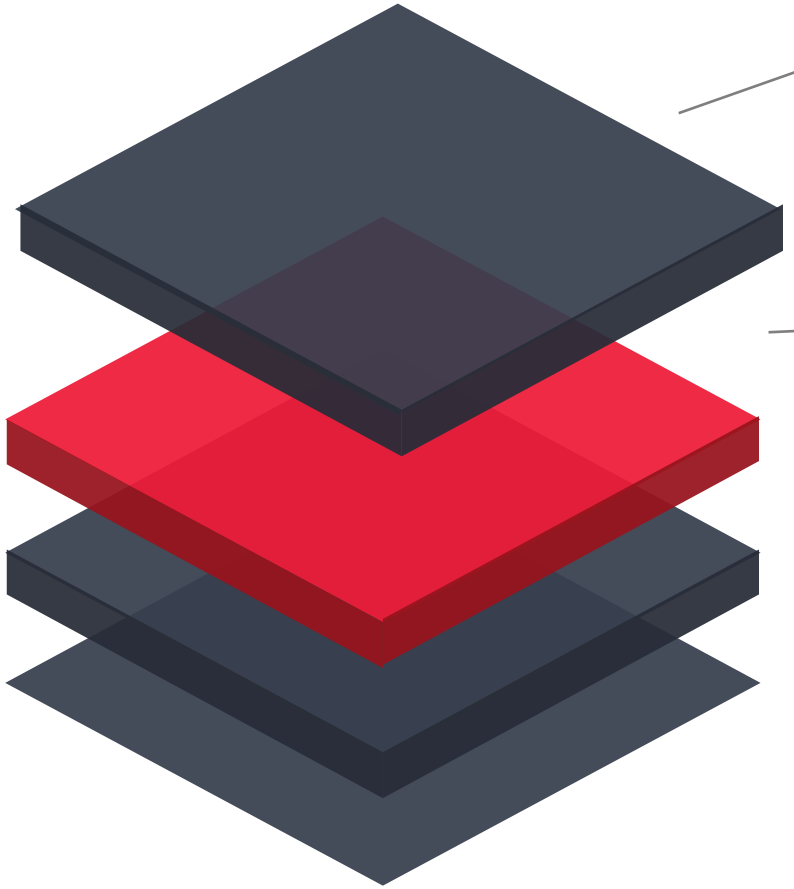
- We are **Rutgers Students** from Department of Electrical and Computer Engineering
- Most of us are from **China** and the others is from **Vietnam** (Each of us knows at least **2 languages**, which help a bit to work on this project)
- We share the same interest in **Machine Learning** and **Natural Language Processing**.
- In the Capstone Expo at Rutgers University, we are very lucky to be awarded **2 prizes**.

Prizes

Top 10 Capstone Projects Award

Best in Research Award

02 Our Capstone Project



Motivation

The Automatic Speech Recognition **(ASR)** engine does **not work** well on **low-resource** languages

Objectives

- Provide **open-source** minority language web crawler
- Decrease the existing **error** rate for the low-resource language ASR

Goals

- **Create** a web crawler to collect text and audio data
- **Filter** the text and audio data, the text needs to be proportionally as long as the corresponding audio
- **Process** the data to produce the aligned text-audio file, or **TextGrid file**

01 Outline

Resource Finding

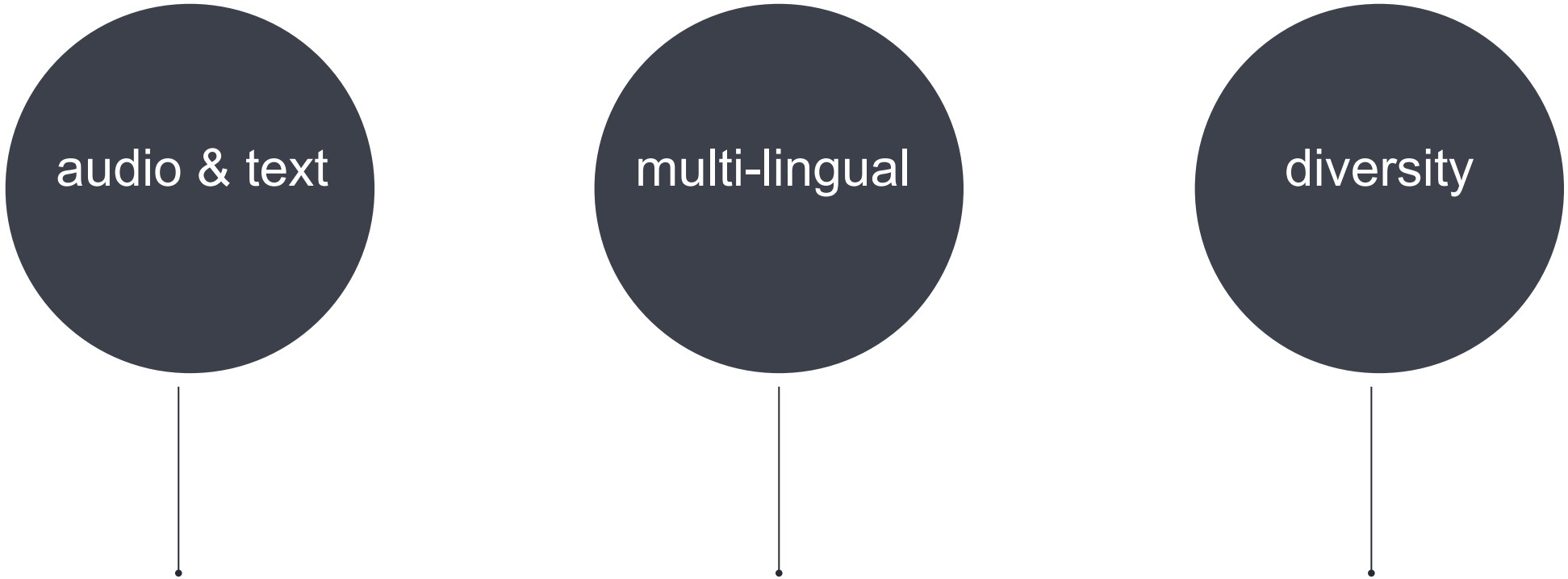
Web Crawling

Data Processing

02

Resource Finding

02 Requirements & Limitations



02 WordProject

Bible website with multiple language

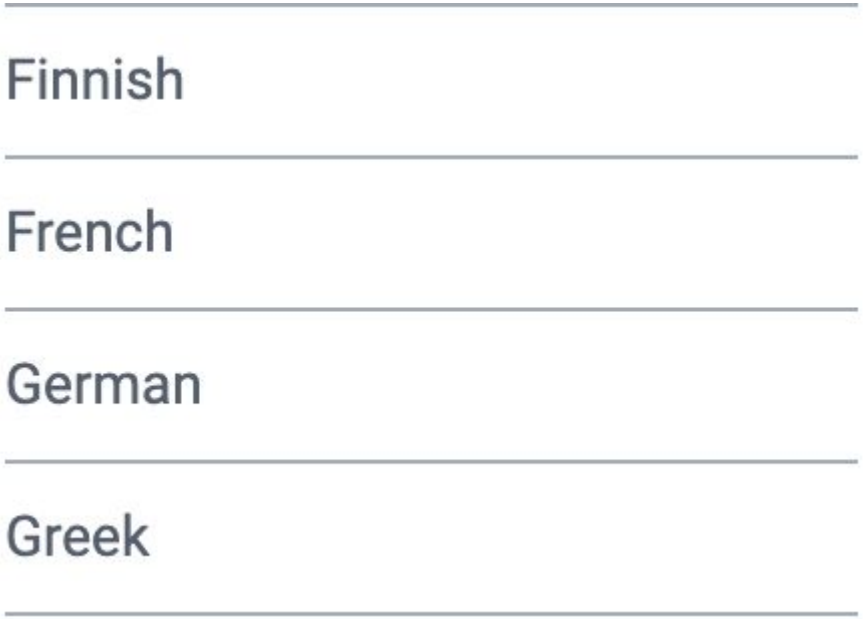
https://www.wordproject.org/bibles/verses/

- support 37 languages
- perfect matching rate
- words are old fashion
- frequency is always the same

02 SBS Radio

multiple language audio news website

https://www.sbs.com.au/yourlanguage/



- support 60 languages
- multiple speakers
- close to real life
- low matching rate

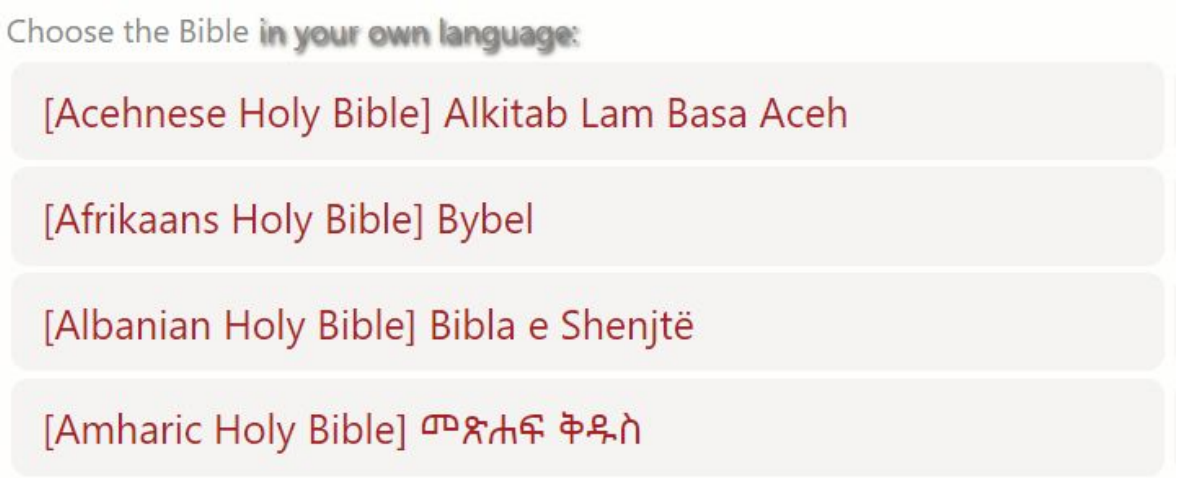
03

Web Crawling

03 Language Selection

Language List Page

- Obtain a **list** of languages to achieve language selection in our program



- **Manually** check if each language is supported - some language contains invalid resources

03 Web Analysis

BeautifulSoup

Transform HTML documents into a “Tag Tree”

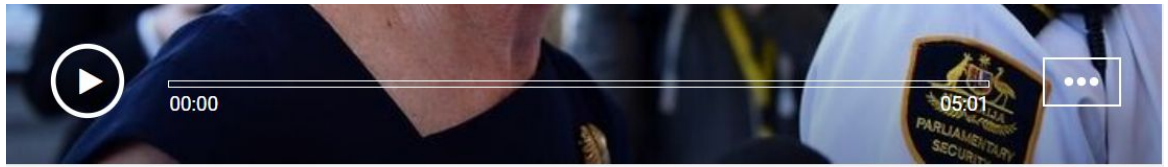
Sending Internet requests to the target website

Request

Requests
http for humans

03 Text Crawling

- The summary of the news which can be used as the text
  - Specific tag <p> that stores the text
    - Corresponding location in the HTML file



The Liberal party needs to look to its own history and Pauline Hanson. (AAP) (AAP)

格拉顿研究所（Grattan Institute）最近公布的研究结果显示，澳洲国民信任政府以至全体政界人士的程度，达至50年来的最低点；当中，大部份被访者均认为从政者只顾一己私利及特定利益集团的益处，而漠视公众利益。有人认为，政界应与其他行业一样，制定一套职业道德守则。这个建议可如何实行？余睿章在今集「时事百宝箱」为大家讲解。

```
<p> == $0
"格拉顿研究所（Grattan Institute）最近公布的研究结果显示，澳洲国民信任政府以至全体政界人士的程度，达至50年来的最低点；当中，大部份被访者均认为从政者只顾一己私利及特定利益集团的益处，而漠视公众利益。有人认为，政界应与其他行业一样，制定一套职业道德守则。这个建议可如何实行？余睿章在今集「时事百宝箱」为大家讲解。"
</p>
```

03 Text Crawling

Text Normalization

Kapitola 15

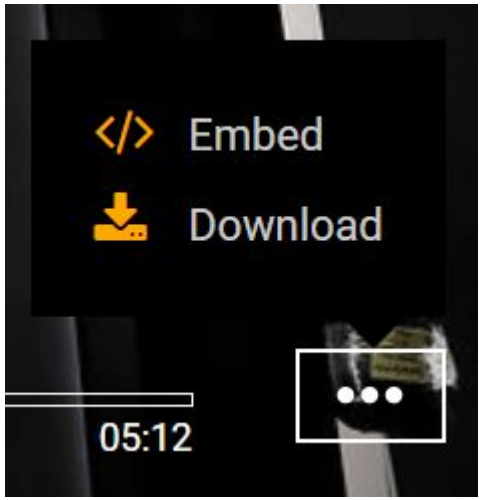
1 Když pak ty věci pominuly, sta
pavéza tvá, a odplata tvá velmi
2 Jemužto řekl Abram: Panovníč
svého, bude Damašský Eliezer?
3 Řekl ještě Abram: Aj, mně jsi n
4 A aj, slovo Hospodinovo k něr
tvým bude.
5 I vyvedl jej ven a řekl: Vzhlédn
tvé.
6 I uvěřil Hospodinu, a počteno

Original text with punctuations and numbers

Text after normalization: punctuations and numbers are eliminated

KAPITOLA
KDYŽ PAK TY VĚCI POMINULY STALO
JEMUŽTO ŘEKL ABRAM PANOVNÍČE H
ŘEKL JEŠTĚ ABRAM AJ MNĚ JSI NEDA
A AJ SLOVO HOSPODINOVO K NĚMU
I VYVEDL JEJ VEN A ŘEKL VZHLÉDNIŽ
I UVĚŘIL HOSPODINU A POČTENO MU
NEBO BYL ŘEKL JEMU JÁ JSEM HOSP
I ŘEKL PANOVNÍČE HOSPODINE PO Č
I ODPOVĚDĚL JEMU VEZMI MNĚ JALO
KTERÝŽTO VZAV TY VŠECKY VĚCI ZRC
PTÁCI PAK SEDALI NA TA MRTVÁ TĚLA
I STALO SE KDYŽ SLUNCE ZAPADALO
ŘEKL TEDY BŮH ABRAMOVI TO ZAJIS
VŠAK NÁROD JEMUŽ SLOUŽITI BUDC
TY PAK PŮJDEŠ K OTCŮM SVÝM V PO
A ČTVRTÉ POKOLENÍ SEM SE NAVRÁT

03 Audio Crawling




```
▼<div class="sm2-button-bd">
    <a href="http://
    wpaorg.wordproject.com/
    bibles/app/audio/32/1/15.mp3"
```

wget

Fast download of web resources

Audio Transformation

ffmpeg: mono/stereo, 44100 Hz, .wav format

```
Audio: pcm_s16le ([1][0][0][0] / 0x0001), 44100 Hz, 1 channels, s16, 705 kb/s
```

03 Filtering

Text too short and duration too long

More Advanced Filtering Strategy

Compute the rate of # of words / min

04

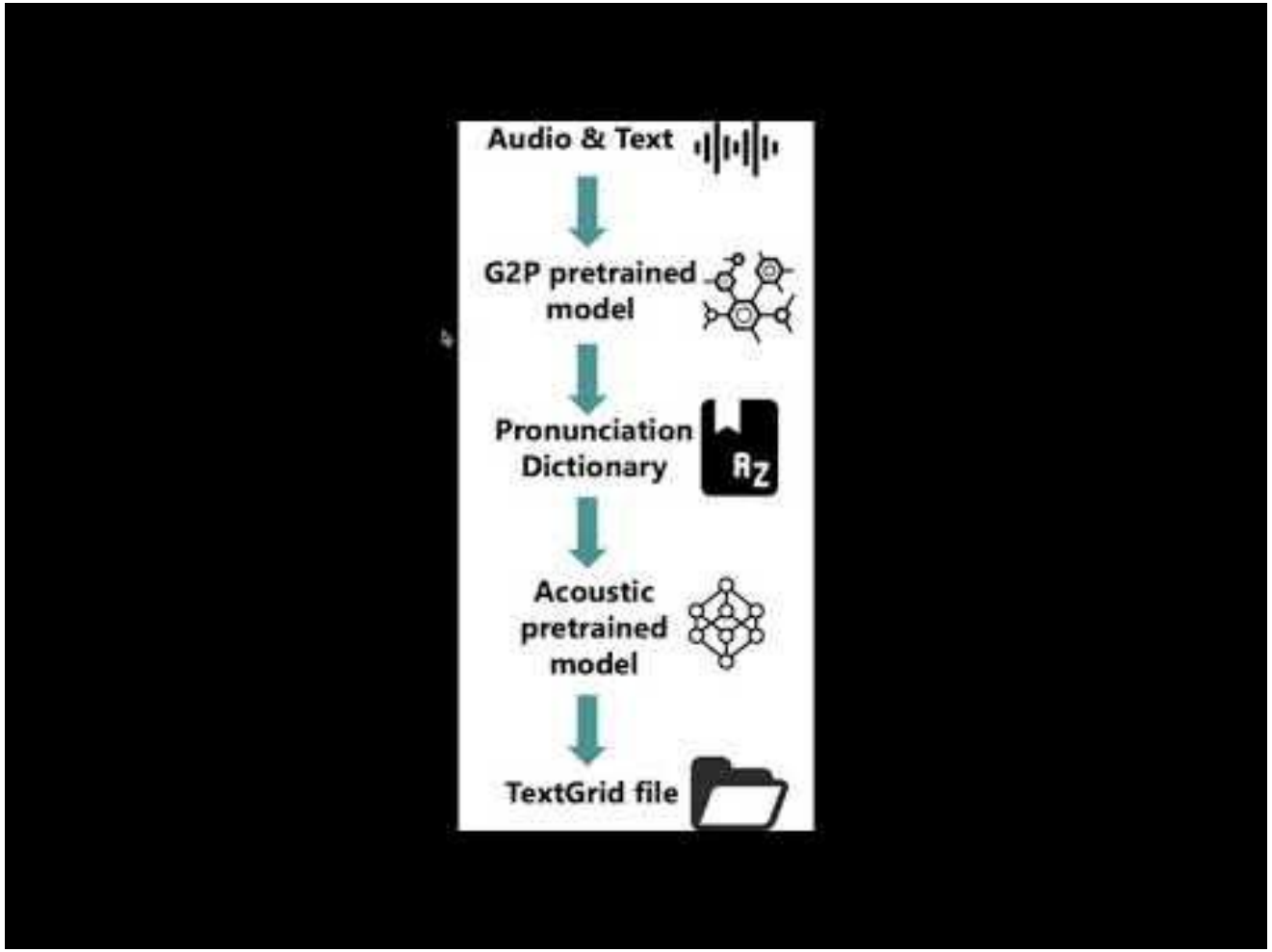
Data Processing

04 Work flow

Montreal Forced Aligner

https://montreal-forced-aligner.readthedocs.io/en/latest/introduction.html#montreal-forced-aligner

- Forced align tools kit based on Kaldi
- Developed by McGill university.



05

Results & Analysis

05 Results

Resources	Match Rate (beam 200)	Match Rate (beam 400)	Match Rate (beam 800)
Ukrainian (SBS)	20.7%	26.9%	25.7%
Ukrainian (Bible)	24.7%	66.3%	77.8%
Vietnamese (SBS)	1.0%	3.6%	5.3%
Vietnamese (Bible)	15.5%	15.7%	16.1%
German (SBS)	16.9%	39.6%	44.5%
German (Bible)	39.9%	55.6%	75.2%

Match Rate:
Aligned time / Total time

Success Rate:
Aligned files / Total files

General Conclusion:
With beam value increased, match rate / success rate also increased

Problem:
Miss-Aligned: the words aligned does not actually match with the contents in the audio

05 Analysis

Beam

different beam value for each resource to test how it influence the result

Pronunciation dictionary

self-trained dictionary & pre-trained dictionary

Acoustic model

the quality of acoustic model

Resource Quality

the inner features of the resource

05 Resources

Length of the audio

Long audio in SBS always have a short corresponding text

Quality of the resource

1. The **diversity** of each audio
2. The text quality(**Long enough** and **almost perfect match the audio**)
3. The total number of resources on the websites
   => There is a trade-off between the match rate and quality of the resource. The higher the match rate, the less quality we expect from the resource

05 G2P Model & Dictionary

```
乃是照着他的形像造男造女      ai3 sh ii4 zh ao4 zh uo2 t a1 d i4 x i2 ng x ia4 ng z ao4 n a2 n z ao4 n v3
也要管理海里的鱼    ie3 iao4 g ua3 n l i3 h ai3 l i3 d i4 v2
事就这样成了        sh ii4 j iu4 zh ei4 ia4 ng ch e2 ng l iao3
于是地发生了青草    v2 sh ii4 d i4 f a1 sh e1 ng l iao3 j i1 ng c ao3
于是神造了两个大光  v2 sh ii4 sh e2 n z ao4 l iao3 l ia3 ng g e4 d a4 g ua1 ng
于是神造出野兽      v2 sh ii4 sh e2 n z ao4 ch u1 ie3 sh ou4
作记号    z uo4 j i4 h ao4
使他们管理海里的鱼  sh ii3 t a1 m e5 n g ua3 n l i3 h ai3 l i3 d i4 v2
使旱地露出来        sh ii3 h a4 n d i4 l u4 ch u5 l ai5
充满海中的水        ch o1 ng m a3 n h ai3 zh o1 ng d i4 sh uei3
全赐给你们作食物    q va2 n g ei3 n i3 m e5 n z uo4 sh ii2 u4
分别明暗  f e1 n b ie2 m i2 ng a4 n
又对他们说iou4 d uei4 t a1 m e5 n sh uo1
又造众星  iou4 z ao4 zh o4 ng x i1 ng
又造出各样飞鸟      iou4 z ao4 ch u1 g e3 ia4 ng f ei1 d iao3
可以分昼夜k e3 i3 f e1 n ie4
各从其类  g e4 c o2 ng q i2 l ei4
和全地    h e2 q va2 n d i4
和地上各样行动的活物h e2 d i4 sh a4 ng g e3 ia4 ng x i2 ng d o4 ng d i4 h uo2 u4
和结种子的菜蔬      h e2 j ie2 zh o3 ng z ii3 d i4 c ai4 sh u1
地上一切昆虫        d i4 sh a4 ng i1 q ie4 k ue1 n ch o2 ng
地上的牲畜d i4 sh a4 ng d i4 i1 sh e1 ng
地是空虚混沌        d i4 sh ii4 k o1 ng x v1 h ue2 n
```

```
can2    c a2 n
can3    c a3 n
can4    c a4 n
cang1   c a1 ng
cang2   c a2 ng
cao1    c ao1
cao2    c ao2
cao3    c ao3
ce4     c e4
ceng2   c e2 ng
cha1    ch a1
cha2    ch a2
cha4    ch a4
chai1   ch ai1
chai2   ch ai2
chan1   ch a1 n
chan2   ch a2 n
chan3   ch a3 n
chan4   ch a4 n
chang1  ch a1 ng
chang2  ch a2 ng
chang3  ch a3 ng
chang4  ch a4 ng
chao1   ch ao1
chao2   ch ao2
chao3   ch ao3
che1    ch e1
che3    ch e3
che4    ch e4
chen2   ch e2 n
```

05

Analysis

Acoustic model

Pretrained model

Corpus: GlobalPhones

Pros:

- Quality
- Amount

Cons:

- Old resources
- Speed
- Success rate

Self-trained model

Pros:

- Flexible
- Fast

Cons:

- Accuracy
- Time in training

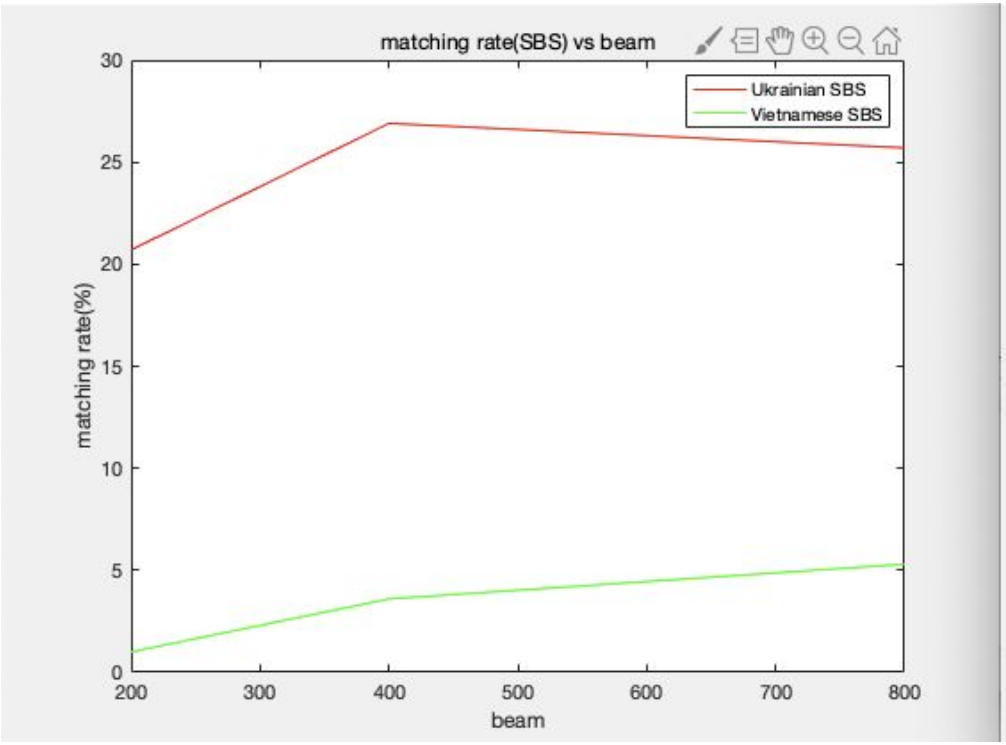
05 Analysis

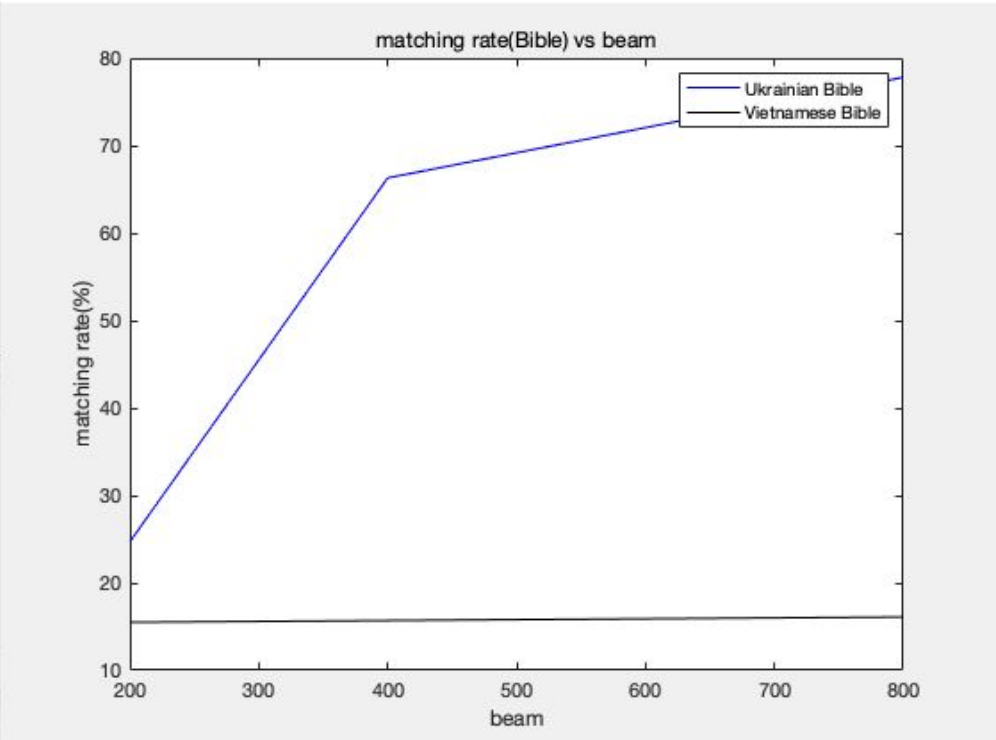
Beam

- Beam is a decoder to decode the output sequence. More intuitively, it’s somehow like the search range in the generated g2p dictionary.
- the increasement of beam leads to the increasement of align time.

05 Analysis

Beam





06

Demo

07

Future Work

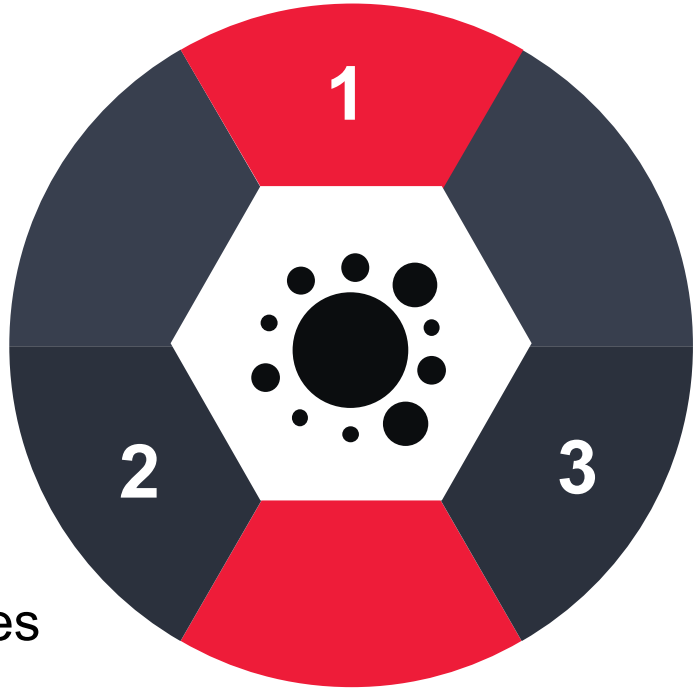
05 Future Work

Stage Model Training

Use output TextGrid files as the input for second-stage forced alignment

Slicing Strategy

Automatically slice original audios/texts to smaller pieces to improve performance



Better Resources

Hard to obtain resources for all languages on just one or two websites

Find resources under different channels in YouTube

Use a variety of resources generated in different scenes
(news, audio book, talk show)

Keep in touch

Project Github Link:

https://github.com/ciuji/RU_capstone

Department:

Rutgers Electrical and Computer Engineering

Rutgers Capstone Projects:

https://www.ece.rutgers.edu/ece-capstone

Project Github QR Code

THANK YOU !