

COSI 136A Project Part 1 – Report: Vietnamese Speech Corpus

Team members: Nhi Le and Zhihan Li

Background context:

Why we chose Vietnamese as the language for this project:

- Our team consists of a Vietnamese and a Chinese. Hence, we wanted to choose a language which one of us can speak, making it possible to check the accuracy of our corpus and also to make it easier to work with.
- Vietnamese has one advantage: spoken and written modern Vietnamese uses Latin-based alphabet, hence it is, in our opinion, more accessible to non-Vietnamese. Moreover, the way that Vietnamese is written is very similar to how it is actually pronounced (e.g., “cá” (pronunciation: [link](#)), which means “fish”, is literally pronounced “ca”, just with a special tonal mark. For more information on the six tonal marks of Vietnamese, please see this [link](#)). This makes it easier for the other team member who doesn’t speak Vietnamese to still at least understand the output TextGrid and how it maps to the actual audio.
- Since Vietnamese has a lot of Chinese-based words (stemming from nearly a thousand years of being dominated by ancient Chinese dynasties, not fun we know!), most of which are pronounced very similarly to how they are pronounced in Mandarin and Cantonese, we thought it would be a suitable language for our group to work with.
- Lastly, NLP capabilities for Vietnamese language are still rather in their infancy, despite 85 million people speaking the language, with Siri currently not supporting Vietnamese, and Google Assistant performing not incredibly well for complex topics. Despite this (or because of this), AI research labs and companies in Vietnam are constantly researching new technologies and publishing papers on this topic. Hence, we think it would be an interesting language to work on since there are still many areas unexplored and questions left unanswered.

Corpus Information:

1. Corpus metadata:

Language: Vietnamese

Number of Speakers: 2

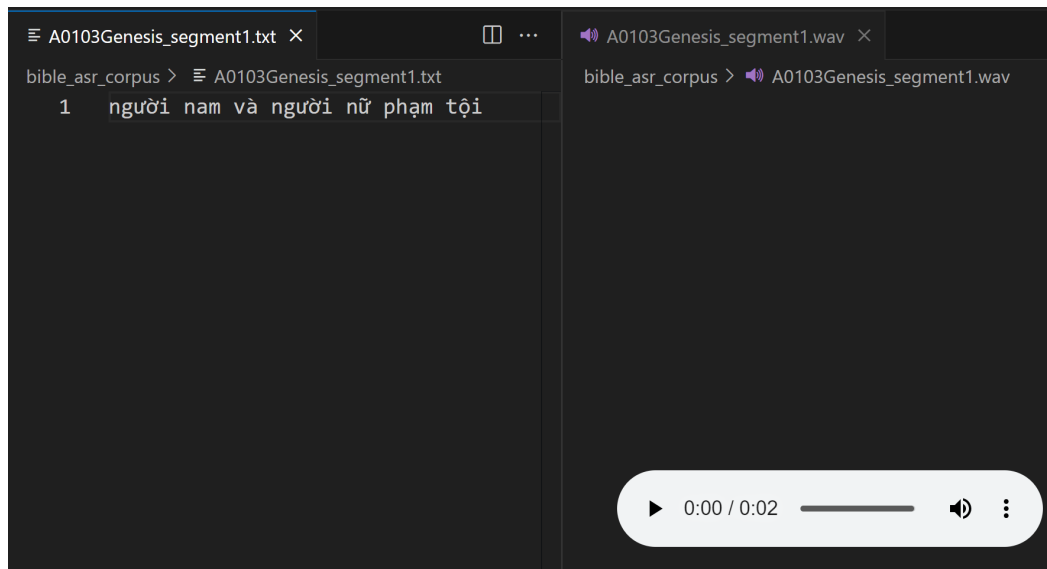
Speaker Gender: Male

Noise: sourced from audiobook and the Bible, almost noise-free.

Data: Vietnamese Bible and an Audiobook

Final form: SampleRate = 16000 Hz, Channels = 1, wav

Length of wav files: more than 2 seconds and less than 10 seconds each



Time Duration:

- Total: 7:28:54
 - Audiobook: 1:34:43
 - Bible: 5:54:11

2. Strengths & weaknesses:

Strengths:

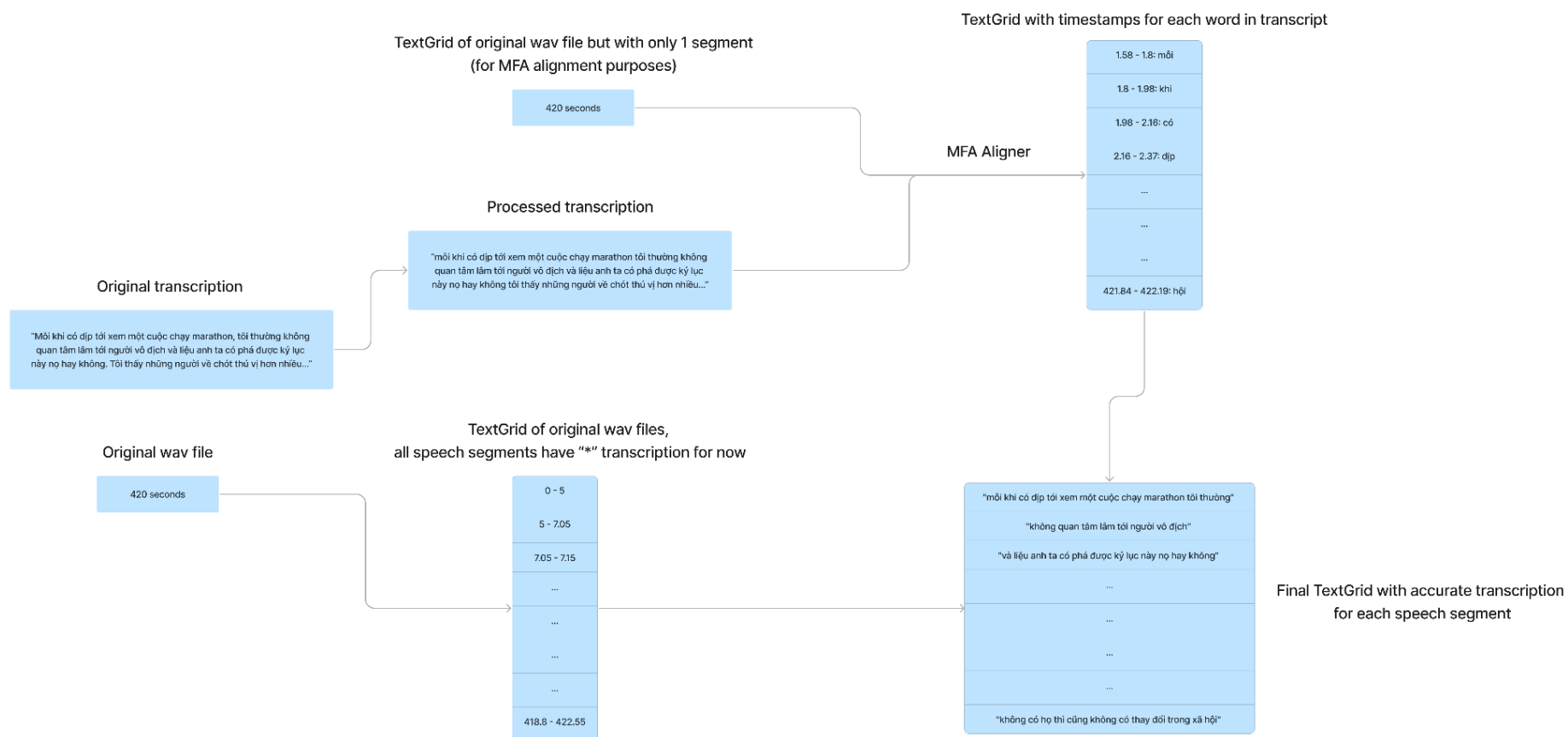
- Noise-free because they are recorded in a perfect environment (i.e. not in everyday situations, no background music or background noise)
- No speaker overlap: both the Bible and the audiobook were read by only one person
- Standard Vietnamese dialect: both the Bible and audiobook were recorded in standard Vietnamese dialect, which pronounces the words very clearly and accurately. It's also representative of how you would commonly hear Vietnamese being spoken.

Weaknesses:

- No female speakers: it was difficult to source high-quality audio that has an accurate accompanying transcript. So unfortunately, we have not been able to source any audio recorded by female speakers that would pass the criteria we have for our speech corpus.
- Noise-free: this is both a strength and weakness of this corpus. Being sound-free might help with training the ASR system, but when it comes to actual deployment and inference, we suspect that the ASR system might not perform very well when there is background noise.
- Lack of variety in dialects. Again, this is both an advantage and disadvantage. Even though we want the ASR system to first be trained on standard Vietnamese dialect, in the future, if we want to expand the scope of this model, we would need to introduce more dialects into the corpus, which is a challenging task as Vietnam has a lot of different dialects that differ greatly in terms of how they say the same words.
- Duration: the duration of our corpus is currently only 7 hours. We aim to increase its size for part 2 of the assignment by sourcing more data from various different contexts and topics of speech.

How the corpus was created:

Here is a flowchart of the process to create this corpus, detailed explanation to follow:



Step 1: Gather the data:

Step 1.1: Bible

- Audio: [link](#)

The mp3 files could be downloaded directly from the above link. I used code (convert_mp3_to_wav.py) to convert audio files into .wav format with sample rate = 16000, and 1 channel.

- **Transcription**

Link: ([link](#))

Using code(extract_bible_text.py) to extract Bible transcripts from the above link. This website has a delay mechanism and the code cannot capture all text content. Some chapter text is missing, so I use code(get_all_matched_data.py) to remove the .wav files that do not have the corresponding transcript.

Step 1.2: Audiobook:

This audiobook is a popular non-fiction book in Vietnam where the author discusses various societal issues and topics. We chose this audiobook because it uses a more modern and commonly used language than the bible or other literature books, which we think would be beneficial when it comes to building the ASR system for conversational language. It should be noted, however, that the audiobook contains some English words such as “marathon”, or names of foreign authors.

- **The audio:** downloaded from Youtube ([link](#)) and the .wav files are extracted using an online tool that converts Youtube videos to .wav files ([link](#)). Further modifications to the .wav files (e.g., cutting out the music intro, or cutting out speech that doesn't appear in the transcription) are made using the Audacity app on Windows. From the Youtube playlist, we were able to extract the audio for 14/17 book chapters (since the other chapters have low audio quality).
- **The transcription:** the transcriptions for the 14 book chapters were obtained from the PDF version of the book that we found online ([link](#)). We extracted this text manually. This text was challenging to work with since it has numbers, dates, percentages, etc. that we need to convert to written format in step 3.1. However, this proved to be a useful step because we can potentially use the code in step 3.1 to apply to any Vietnamese text with special characters like that.

Ideally, we want to obtain more audiobooks of different speakers (e.g., female speakers, speakers from different regions of Vietnam, with different dialects) to create a more diverse speech corpus, since audiobooks have one major advantage being their transcription is accurate (unlike other forms of media like podcast or news where the transcription is either auto-generated with low accuracy, or it is non-existent). However, obtaining more audiobooks proved to be a harder task than we anticipated, since most of them either don't have audio files available for download, or they have background music.

Step 2: Install Montreal Forced Aligner

- **Install Miniconda**
- **Open Miniconda prompt and create a new virtual environment:**
 - `conda create -n aligner -c conda-forge montreal-forced-aligner`
- **Activate the virtual environment:**
 - `conda activate aligner`
- **Using MFA to align a speech corpus with existing (i.e. pretrained) pronunciation dictionary and acoustic model:**
https://montreal-forced-aligner.readthedocs.io/en/latest/first_steps/index.html#first-steps-align-pretrained
- **Install MFA's pretrained models for Vietnamese:**
 - mfa model download acoustic vietnamese_mfa ([link](#))
 - mfa model download dictionary vietnamese_mfa ([link](#))

Step 3: Prepare data for MFA validation and alignment steps:

Step 3.1: preprocess the original transcription:

We need to make sure:

- The transcript doesn't have any special characters !@#\$%&().,
- The numbers in the transcript are in their full written form (e.g., number 25 should be expanded into “hai mươi lăm” in the transcript). This applies to dates, percentages, numbers, etc.
- Lowercase all words
- Remove all line breaks

An example of what the transcript looks like before and after preprocessing:

Mỗi khi có dịp tới xem một cuộc chạy marathon, tôi thường không quan tâm lắm tới người vô địch và liệu anh ta có phá được kỷ lục này nọ hay không. Tôi thấy những người về chót thú vị hơn nhiều. Lần nào cũng vậy, khi những người thắng cuộc đã lên bục nhận giải, chụp ảnh, trả lời truyền hình, rồi đã về nhà tắm rửa xong, thì nhóm người này vẫn hì hục, nhằn nại ở những cây số cuối cùng. Tôi đứng ở ven đường để ngắm lòng quyết tâm đầy đau đớn của họ. Thường khi họ rẽ vào khúc ngoặt cuối cùng dẫn tới đích thì các băng rôn đã được tháo xuống từ lâu, cũng không còn ai đứng ở vạch đích để bấm thời gian cho họ, và người xem cũng đã ra về gần hết. Bám sát gót những người đang lê lét này là các nhân viên vệ sinh khua chổi quét đường.

Tôi không để ý tới những người về đầu vì họ là dân chuyên nghiệp, họ sinh ra để dẫn đầu, họ có tố chất để làm điều siêu phàm. Những người về cuối thì hiểu rằng họ không có vai trò gì trong cái cuộc thi thố này. Họ chẳng đem lại vinh quang cho ai, mà thất bại của họ cũng không làm ai máy may quan tâm. Động cơ để họ cản rằng lê bước tiếp không phải là những gủi gắm của một tập thể, chẳng phải là danh dự của một quốc gia, hay danh tiếng của bản thân mà họ cần phải bảo vệ. Họ đơn thuần bước binh và có thể hơi điên rồ. Họ tiếp tục chỉ vì bỏ cuộc không phải là lựa chọn của họ.

Cái bước binh và điên rồ của những con người bình thường này có cái gì đó thật lôi cuốn tôi. Nó làm tôi liên tưởng tới câu chuyện mà tôi mới được biết về em bé sáu tuổi da đen Ruby Bridges - cũng là một cuộc chạy marathon, nhưng ở dạng khác. Vào cuối những năm 1950, bang New Orleans ở Mỹ đã xóa bỏ sự phân biệt màu da ở các thư viện, trên xe buýt và ở các công viên công cộng, duy ở các trường học thì vẫn không. Năm 1960, một tòa án liên bang ra quyết định bắt chính quyền bang này phải cho phép học sinh da đen tới các trường vốn dành cho da trắng. Ruby đăng ký học lớp một ở một trường gần nhà. Em sẽ là học sinh da đen đầu tiên và duy nhất của trường vào năm đó. Ngày nhập trường, bốn cảnh sát tòa án liên bang hộ tống Ruby và mẹ em tới trường trong một chiếc xe limousine lớn. Đội họ ở cổng trường là một đám đông da trắng giận dữ, gào thét, chửi rủa. Xuống xe, hai cảnh sát đi trước, hai đi sau để bảo vệ, họ đi dọc những bức tường đầy vết cà chua và những dòng chữ thóa mạ. Một người đàn bà da trắng gào lên: "Tao sẽ đầu độc mày, tao sẽ tìm được cách." Nhớ lại hành trình đi qua đám đông hung dữ đó, một cảnh sát liên bang nói về Ruby: "Em không khóc. Em không thút thít. Em chỉ rào bước đi, như một người lính bé nhỏ. Tất cả chúng tôi đều rất tự hào về em."



Mỗi khi có dịp tới xem một cuộc chạy marathon tôi thường không quan tâm lắm tới người vô địch và liệu anh ta có phá được kỷ lục này nọ hay không tôi thấy những người về chót thú vị hơn nhiều lần nào cũng vậy khi những người thắng cuộc đã lên bục nhận giải chụp ảnh trả lời truyền hình rồi đã về nhà tắm rửa xong thì nhóm người này vẫn hì hục nhằn nại ở những cây số cuối cùng tôi đứng ở ven đường để ngắm lòng quyết tâm đầy đau đớn của họ thường khi họ rẽ vào khúc ngoặt cuối cùng dẫn tới đích thì các băng rôn đã được tháo xuống từ lâu cũng không còn ai đứng ở vạch đích để bấm thời gian cho họ và người xem cũng đã ra về gần hết bám sát gót những người đang lê lét này là các nhân viên vệ sinh khua chổi quét đường tôi không để ý tới những người về đầu vì họ là dân chuyên nghiệp họ sinh ra để dẫn đầu họ có tố chất để làm điều siêu phàm những người về cuối thì hiểu rằng họ không có vai trò gì trong cái cuộc thi thố này họ chẳng đem lại vinh quang cho ai mà thất bại của họ cũng không làm ai máy may quan tâm động cơ để họ cản rằng lê bước tiếp không phải là những gủi gắm của một tập thể chẳng phải là danh dự của một quốc gia hay danh tiếng của bản thân mà họ cần phải bảo vệ họ đơn thuần bước binh và có thể hơi điên rồ họ tiếp tục chỉ vì bỏ cuộc không phải là lựa chọn của họ cái bước binh và điên rồ của những con người bình thường này có cái gì đó thật lôi cuốn tôi nó làm tôi liên tưởng tới câu chuyện mà tôi mới được biết về em bé sáu tuổi da đen ruby bridges cũng là một cuộc chạy marathon nhưng ở dạng khác vào cuối những năm một nghìn chín trăm năm mươi bang new orleans ở mỹ đã xóa bỏ sự phân biệt màu da ở các thư viện trên xe buýt và ở các công viên công cộng duy ở các trường học thì vẫn không năm một nghìn chín trăm sáu mươi một tòa án liên bang ra quyết định bắt chính quyền bang này phải cho phép học sinh da đen tới các trường vốn dành cho da trắng ruby đăng ký học lớp một ở một trường gần nhà em sẽ là học sinh da đen đầu tiên và duy nhất của trường vào năm đó ngày nhập trường bốn cảnh sát tòa án liên bang hộ tống ruby và mẹ em tới trường trong một chiếc xe limousine lớn đội họ ở cổng trường là một đám đông da trắng giận dữ gào thét chửi rủa xuống xe hai cảnh sát đi trước hai đi sau để bảo vệ họ đi dọc những bức tường đầy vết cà chua và những dòng chữ thóa mạ một người đàn bà da trắng gào lên tao sẽ đầu độc mày tao sẽ tìm được cách nhớ lại hành trình đi qua đám đông hung dữ đó một cảnh sát liên bang nói về ruby em không khóc em không thút thít em chỉ rào bước đi như một người lính bé nhỏ tất cả chúng tôi đều rất tự hào về em cả ngày hôm đó hai mẹ con không dám bước chân ra khỏi phòng hiệu trưởng qua vách kính họ chứng kiến cảnh các phụ huynh da trắng xông vào trường và giận dữ kéo con mình ra

Step 3.2: Create a TextGrid that partitions the original wav file into multiple speech segments:

Using the code from lecture 2, we create a TextGrid that partitions each wav file into multiple speech segments based on a specified energy threshold (this threshold was chosen based on trial and error until we were able to get an output with speech segments that are neither too long nor too short). Final output of this step should look like this:

```
File type = "ooTextFile"
Object class = "TextGrid"

0
424.56725
<exists>
1
"IntervalTier"
"words"
0
424.56725
239
0
5
"*"
5
7.05
"*"
7.05
7.15
""
7.15
10.25
"*"
10.25
10.700000000000001
""
```










Step 3.3: Create a TextGrid with a single speech segment that contains the preprocessed transcript

- This will be the input for MFA aligner, which will help us detect the timestamps for each word in the transcript. The output TextGrid files at this step should look like this, where there is one single speech segment:

```
File type = "ooTextFile"
Object class = "TextGrid"

0
397.977125
<exists>
1
"IntervalTier"
"speech"
0
397.977125
1
0
397.977125
"vào buổi sáng chủ nhật đẹp trời tuần trước trên đoạn quốc lộ 1a thuộc tỉnh Quảng Nam một tài xế xe tải chở xoài rụng về tránh một xe máy cùng chiều cua tay lái gấp và làm xe lật nghiêng hơn chục tấn xoài đổ tràn ra đường người dân xung quanh xúm lại chia nhau bảo vệ hiện trường và giúp tài xế thu gom xoài nằm vung vãi các báo đăng một tin ngắn về sự việc hôm sau câu chuyện rơi vào quên lãng điều này làm tôi suy nghĩ tôi tin chắc rằng nếu như người dân lao vào hái xoài thì ngay lập tức dư luận sẽ dậy sóng trên các mặt báo lại đây những cảnh báo về đạo đức xã hội suy đồi và người ta lại hồ thẹn lẫn cho nhau trước các bạn quốc tế vậy mà sao hôm đó các ống kính báo chí không chụp cận cảnh những người dân đang tươi tắn nhể nhại mò hôi khuôn xoài hộ tài xế như họ đã từng zoom vào các khuôn mặt tươi tắn và nhể nhại mò hôi hôi bìa cách đây mấy tháng sao không có ai phỏng vấn anh tài xế thờ phào nhẹ nhõm vì không phải đến hàng và mấy hôm sau sao không có người dân nào căng băng rôn ven đường tôi tự hào là người Quảng Nam câu chuyện nhỏ này cho ta thấy là chúng ta một cách vô thức thờ ơ với các tin tốt nhưng lại quan tâm đặc biệt tới các tin xấu hôm chủ nhật kia thậm chí ta còn gần như nghe được tiếng thở dài tiếc rẻ của các nhà bình luận trên mạng vì vụ xoài đó này đã không cho họ một cơ hội để khẳng định lại một lần nữa rằng chúng ta thật là những con người tồi tệ sống trong một môi trường thật tồi tệ tin tức về những tệ nạn hay bất cập trong xã hội cung cấp cho chúng ta những cái cớ để than phiền và kêu ca phàn nàn bực dọc cấu kỉnh chê bai đang trở thành những trạng thái thường trực trong dư luận các trạng thái này được gói ghém một cách tài tình trong từ bức xúc không từ tiếng việt nào lại có một sự nghiệp thăng tiến ngoạn mục như vậy từ chỗ vô danh cách đây bảy tám năm bảy giờ nếu gõ bức xúc vào google ta sẽ được hai mươi chín triệu kết quả gần gần mười lần ngọc trình một con số ấn tượng cho một từ có làn da xấu xí như vậy vì sao chúng ta lại ưu tiên tin xấu đâm đuổi với chúng thay vì chú ý tới những điều tốt lành vì sao chúng ta muốn kêu ca phàn nàn thay vì vui tươi chuyện tay nhau những tin vui những câu chuyện đẹp hội chúng bức xúc mỗi nghe thoát tường vô lý nhưng nó có những lý do tâm lý đáng sau trước hết khi lên tiếng phê bình hay than phiền về một điều gì đó chúng ta chứng tỏ cho người khác và cho bản thân là chúng ta không thờ ơ vô cảm mà vẫn còn quan tâm lo lắng hơn nữa khi chê trách người khác chúng ta cảm thấy ưu việt về mặt đạo đức và tự hài lòng vì thấy mình tốt đẹp hơn càng có nhiều vụ cướp tiệm vàng bác sĩ vút xác bệnh nhân bảo mẫu đánh trẻ hồi của bé hoa chúng ta càng có nhiều cơ hội để tự nhủ là chúng ta không phải họ chúng ta chỉ không may bị chung sống cùng họ nhưng thực chất chúng ta ưu tú hơn họ nhiều một điểm quan trọng nữa là khi bức xúc chúng ta phát ra tín hiệu là chúng ta vô can và vô tội các chính trị gia đã nhận ra điều đó các bạn có thấy gần đây các quan chức cũng bức xúc rất nhiều đại biểu quốc hội bức xúc trước tình trạng tham nhũng báo điện tử chính phủ hai mươi chín/mười/hai nghìn không trăm mười ba bộ trưởng bộ y tế bức xúc về đồng nghiệp tuổi trẻ bốn/mười một/hai nghìn không trăm mười ba bộ trưởng bộ nông nghiệp và phát triển nông thôn bức xúc về thực phẩm độc báo đầu tư hai/một/hai nghìn không trăm mười bốn khi bày tỏ sự bức xúc một cách khéo léo chúng ta tuyên bố là mình không thể thuộc về bên thủ phạm được mà mình đứng về phía bị thiệt thòi mình cũng là nạn nhân dần dần chúng ta đâm ra nghiện những cái lắc đầu những cái mình tốt đẹp đây sự quan tâm cộng với sự vô can không liên đới không chịu trách nhiệm là một cảm giác êm ái nó cũng giúp xoa dịu những bút rút lương tâm thi thoảng nói lên khi chúng ta lơ mơ cảm thấy mình không đủ dũng cảm để làm hết những gì có thể làm trước những sai trái trong xã hội những lúc đó cách trấn an bản thân hiệu nghiệm là tỏ ra bức xúc một cách gay gắt nhưng chúng ta không vô can cuộc sống của mỗi cá nhân chúng ta đang đặt trên nền của bao nhiêu bất công và phi lý những viên gạch xây nên ngôi nhà của ta được đóng bởi những đứa trẻ có tuổi thơ vất vả cái ti vi ta dùng được làm bởi những người công nhân di cư có một cuộc sống buồn tẻ và khốn khổ con cái họ bị khó dễ khi tới trường vì không có hộ khẩu cho nên lần tới khi ngồi trong quán ăn ở một resort bên bờ biển bạn cũng đứng bức xúc với mấy người phục vụ què mùa đang lóng ngóng rót rượu vang vào cốc uống nước cam nữa hãy chụp mất xuống khiêm nhường khi họ đứng trước mặt bạn bởi có thể gia đình họ đã bị đuổi ra khỏi nơi này và ngôi làng mà họ đã sống ở đó nhiều đời đã bị xóa sổ để biến thành nơi bạn đang tới nghỉ có thể chúng ta không phải là những kẻ trực tiếp tạo ra bất công nhưng cuộc sống của chúng ta đang phụ thuộc vào những kẻ đó chúng ta ngồi cùng bàn tiệc với họ ý thức về điều đó là việc tối thiểu mà ta có thể làm ngoài việc chúng ta chuyền tay nhau những câu chuyện đẹp những tin tốt để chúng nhận được sự công nhận và giá trị xứng đáng"
```

The folder of data that should be input into MFA aligner should look like this: where for each .wav file, we have a corresponding TextGrid file:

| | | |
|--|---------------------|---------------|
|  chapter1 | 11/10/2023 12:34 PM | TEXTGRID File |
|  chapter1 | 11/9/2023 12:31 PM | WAV File |
|  chapter5 | 11/10/2023 12:34 PM | TEXTGRID File |
|  chapter5 | 11/2/2023 3:29 PM | WAV File |
|  chapter6 | 11/10/2023 12:34 PM | TEXTGRID File |
|  chapter6 | 11/9/2023 12:35 PM | WAV File |
|  chapter7 | 11/10/2023 12:34 PM | TEXTGRID File |
|  chapter7 | 11/9/2023 12:44 PM | WAV File |

Step 3.4: Using MFA's aligner to detect the timestamps for each word in the transcript:

- Open Miniconda prompt, Run as Administrator (this will help with permission to write new files into folder):
 - `conda activate aligner` (if not already activated the virtual environment)
 - `mfa validate input/path/to/speech/corpus/folder/specified/above vietnamse_mfa vietnamse_mfa`

The last two arguments “vietnamse_mfa vietnamse_mfa” are the pretrained MFA models for the pronunciation dictionary and the acoustic model. This step essentially validates our corpus (wav and TextGrid files) to see if they are valid inputs for the MFA aligner to work.

This step also tells us other information such as number of speakers detected, number of utterances (speech segments), out-of-vocabulary words (e.g., non-Vietnamese words, words that are legitimate Vietnamese words but are not included in the MFA's pretrained dictionary):

```
(base) C:\Windows\System32>conda activate aligner

(aligner) C:\Windows\System32>mfa validate C:\BRANDEIS\cosi-136a\project\audiobook1\audiobook1_speech_corpus vietnamese_mfa vietnamese_mfa
Please be aware that you are running an alpha version of MFA. If you would like to install a more stable version, please visit https://montreal-forced-aligner.readthedocs.io/en/latest/installation.html#installing-older-versions-of-mfa
INFO      Setting up corpus information...
INFO      Loading corpus from source files...
0% ----- 0/100 [ 0:00:01 < -:--:-- , ? it/s ] INFO      Found 1 speaker across 14 files, average number of utterances per speaker: 14.0
INFO      Initializing multiprocessing jobs...
WARNING   Number of jobs was specified as 3, but due to only having 1 speakers, MFA will only use 1 jobs. Use the
--single_speaker flag if you would like to split utterances across jobs regardless of their speaker.
INFO      Normalizing text...
79% ----- 11/14 [ 0:00:02 < -:--:-- , ? it/s ] INFO      Generating MFCCs...
100% ----- 14/14 [ 0:00:22 < 0:00:00 , 1 it/s ] INFO      Calculating CMVN...
INFO      Generating final features...
64% ----- 9/14 [ 0:00:02 < 0:00:01 , 45 it/s ] INFO      Creating corpus split...
0% ----- 0/14 [ 0:00:02 < -:--:-- , ? it/s ] INFO      Corpus
INFO      14 sound files
INFO      14 text files
INFO      1 speakers
INFO      14 utterances
INFO      8052.819 seconds total duration
INFO      Sound file read errors
INFO      There were no issues reading sound files.
INFO      Feature generation
INFO      There were no utterances missing features.
INFO      Files without transcriptions
INFO      There were no sound files missing transcriptions.
INFO      Transcriptions without sound files
INFO      There were no transcription files missing sound files.
INFO      Dictionary
INFO      Out of vocabulary words
WARNING   62 OOV word types
WARNING   2424total OOV tokens
WARNING   For a full list of the word types, please see:
C:\Users\halan\Documents\MFA\audiobook1_speech_corpus\oovs_found.txt. For a by-utterance breakdown of missing
words, see: C:\Users\halan\Documents\MFA\audiobook1_speech_corpus\utterance_oovs.txt
INFO      Training
INFO      Initializing training for monophone...
INFO      Compiling training graphs...
INFO      Generating initial alignments...
100% ----- 14/14 [ 0:00:03 < 0:00:00 , 4 it/s ] INFO      Initialization complete!
INFO      monophone - Iteration 1 of 40
INFO      Generating alignments...
7% ----- 1/14 [ 0:11:59 < -:--:-- , ? it/s ] INFO      Accumulating statistics...
14% ----- 2/14 [ 0:00:02 < -:--:-- , ? it/s ] INFO      monophone - Iteration 2 of 40
INFO      Generating alignments...
7% ----- 1/14 [ 0:04:35 < -:--:-- , ? it/s ] INFO      Accumulating statistics...
14% ----- 2/14 [ 0:00:02 < 0:00:02 , 9 it/s ] INFO      monophone - Iteration 3 of 40
INFO      Generating alignments...
7% ----- 1/14 [ 0:06:26 < -:--:-- , ? it/s ] INFO      Accumulating statistics...
14% ----- 2/14 [ 0:00:02 < -:--:-- , ? it/s ] INFO      monophone - Iteration 4 of 40
INFO      Generating alignments...
21% ----- 3/14 [ 0:05:29 < 0:00:14 , 1 it/s ] INFO      Accumulating statistics...
```

Step 4: Using MFA to create aligned TextGrids:

Once we have passed the validation in step 3.4, we can finally start using MFA aligner:

- Open Miniconda prompt as administrator (if not already).
 - `conda activate aligner` (or any name of your virtual environment)
 - `mfa align --clean --overwrite path/to/speech/corpus/created/in/step/3.3 vietnamese_mfa vietnamese_mfa path/to/output/directory`

```
(aligner) C:\Windows\System32>mfa align --clean --overwrite C:\BRANDEIS\cosi-136a\project\audiobook1\audiobook1_speech_corpus vietnamese_mfa vietnamese_mfa C:\BRANDEIS\cosi-136a\project\audiobook1\audiobook1_mfa_aligned_textgrid
Please be aware that you are running an alpha version of MFA. If you would like to install a more stable version, please visit https://montreal-forced-aligner.readthedocs.io/en/latest/installation.html#installing-older-versions-of-mfa
INFO     Setting up corpus information...
INFO     Loading corpus from source files...
0%      Found 1 speaker across 14 files, average number of utterances per speaker: 14.0
INFO     Initializing multiprocessing jobs...
WARNING  Number of jobs was specified as 3, but due to only having 1 speakers, MFA will only use 1 jobs. Use the --single_speaker flag if you would like to split utterances across jobs regardless of their speaker.
INFO     Normalizing text...
86%      Generating MFCCs...
INFO     Generating MFCCs...
100%     Calculating CMVN...
INFO     Calculating CMVN...
93%      Generating final features...
INFO     Generating final features...
0%      Creating corpus split...
INFO     Creating corpus split...
0%      Compiling training graphs...
INFO     Compiling training graphs...
INFO     Performing first-pass alignment...
INFO     Generating alignments...
100%     Calculating fMLLR for speaker adaptation...
INFO     Calculating fMLLR for speaker adaptation...
100%     Performing second-pass alignment...
INFO     Performing second-pass alignment...
INFO     Generating alignments...
100%     Collecting phone and word alignments from alignment lattices...
INFO     Collecting phone and word alignments from alignment lattices...
93%      Alignment analysis not available without using postgresql
WARNING  Alignment analysis not available without using postgresql
INFO     Exporting alignment TextGrids to C:\BRANDEIS\cosi-136a\project\audiobook1\audiobook1_mfa_aligned_textgrid...
100%     Exporting alignment TextGrids to C:\BRANDEIS\cosi-136a\project\audiobook1\audiobook1_mfa_aligned_textgrid...
INFO     Finished exporting TextGrids to C:\BRANDEIS\cosi-136a\project\audiobook1\audiobook1_mfa_aligned_textgrid!
INFO     Done! Everything took 261.909 seconds
```

The output from this step will be a TextGrid that looks like this, where there are timestamps for where in the audio file each word from the transcription was spoken:

```
File type = "ooTextFile"
Object class = "TextGrid"

xmin = 0
xmax = 424.56725
tiers? <exists>
size = 2
item []:
  item [1]:
    class = "IntervalTier"
    name = "words"
    xmin = 0
    xmax = 424.56725
    intervals: size = 1671
    intervals [1]:
      xmin = 0
      xmax = 1.58
      text = ""
    intervals [2]:
      xmin = 1.58
      xmax = 1.8
      text = "mỗi"
    intervals [3]:
      xmin = 1.8
      xmax = 1.98
      text = "khi"
    intervals [4]:
      xmin = 1.98
      xmax = 2.16
      text = "cô"
    intervals [5]:
      xmin = 2.16
      xmax = 2.37
      text = "đip"
    intervals [6]:
      xmin = 2.37
      xmax = 2.45
      text = ""
    intervals [7]:
      xmin = 2.45
      xmax = 2.6
      text = "tối"
    intervals [8]:
      xmin = 2.6
      xmax = 2.94
      text = "xem"
```

Step 5: Combine outputs from step 3.2 and step 4 into the final TextGrid

Finally, we can simply combine the TextGrid with empty speech segments in step 3.2, and the TextGrid with speech segments for each individual word in the transcript, into a final TextGrid. This step simply involves collecting all the words that belong in a speech segment based on their timestamps. The final output looks like this:

```
File type = "ooTextFile"
Object class = "TextGrid"

0
424.56725
<exists>
1
"IntervalTier"
"speech_segments"
0
424.56725
239
0
5
"mỗi khi có dịp tới xem một cuộc chạy marathon tôi thường"
5
7.05
"không quan tâm lắm tới người vô địch"
7.05
7.15
""
7.15
10.25
"và liệu anh ta có phá được kỷ lục này nọ hay không"
10.25
10.700000000000001
""
10.700000000000001
13.100000000000001
"tôi thấy những người về chót thú vị hơn nhiều"
13.100000000000001
13.350000000000001
""
13.350000000000001
18.35
"lần nào cũng vậy khi những người thắng cuộc đã lên bục nhận giải chụp ảnh trả lời truyền hình"
18.35
20.650000000000002
"rồi đã về nhà tắm rửa xong"
20.650000000000002
20.700000000000003
""
20.700000000000003
24.750000000000004
"thì nhóm người này vẫn hì hục nhằn nặn ở những cây số cuối cùng"
```

Step 6: Split original wav file into smaller .wav files, and create the respective transcript .txt files:

At this step, we simply use the TextGrid created in step 5 to look up the timestamps of each speech segment. Using a Python script, we were able to split the original wav file into smaller wav files, and export the transcription of each speech segment into a .txt file with the same name.

Future work

To prepare for training an ASR model in part 2 of the project, we aim to collect more data for the speech corpus. Specifically, we aim to collect audio with female speakers and speakers of different ages (where there is a difference in pitch). Moreover, we also aim to increase the number of topics covered by our corpus. Currently, the topics covered are religion (the Bible) and societal issues (the audiobook). We also want to incorporate more conversational language into our corpus, such as from podcasts.

Source code

All the step-by-step code used to create our corpus can be found in the Box repository. We have a folder of Python scripts used for the Bible, and another folder for the audiobook.