CAPSTONE DEFENSE

ENGLISH TO VIETNAMESE SUBTITLE GENERATION SYSTEM

Students

Le Hoang Phuc

Ngo Anh Kiet

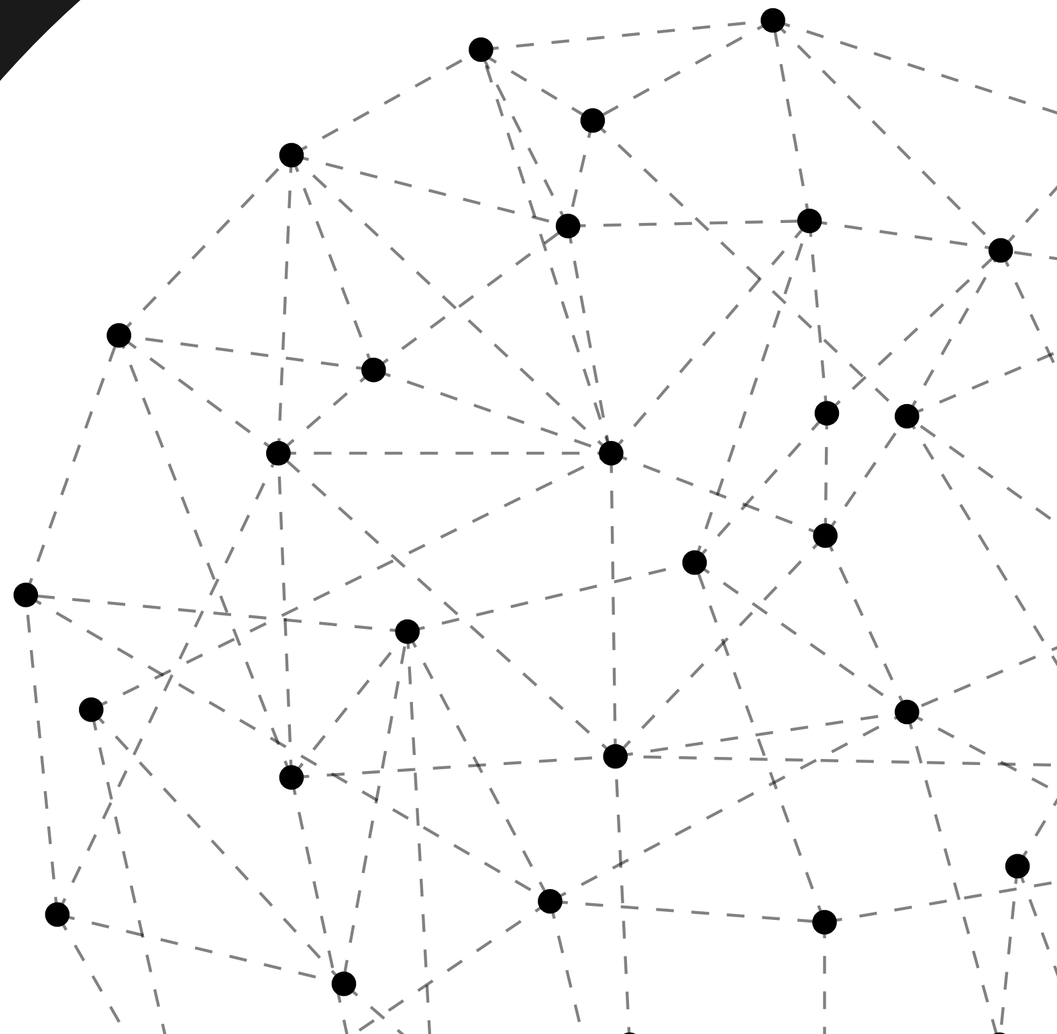
Kieu Minh Duy

Instructor

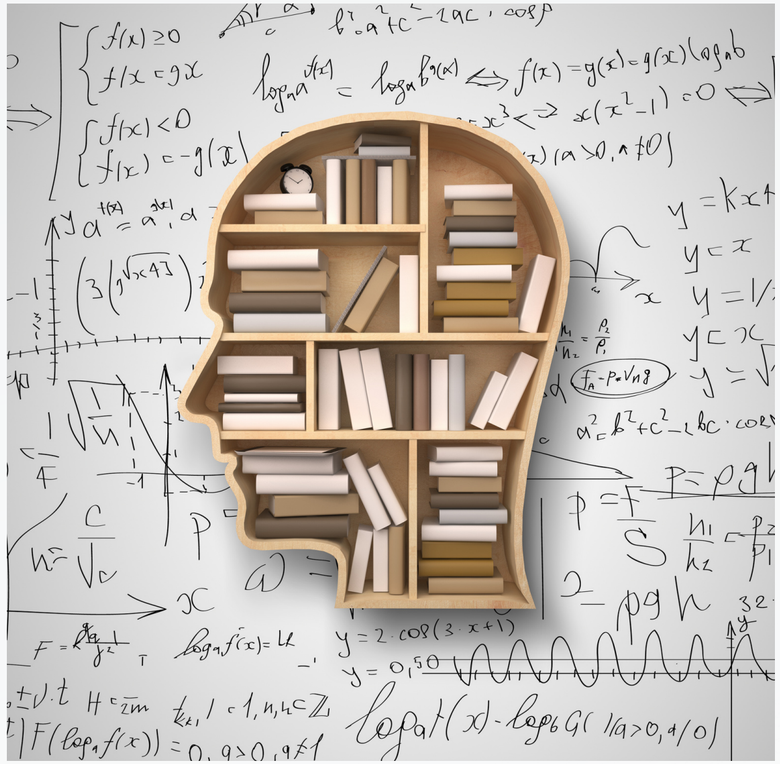
Assoc. Prof. Phan Duy Hung

CONTENTS

ABSTRACT & INTRODUCTION



ABSTRACT



ABSTRACT

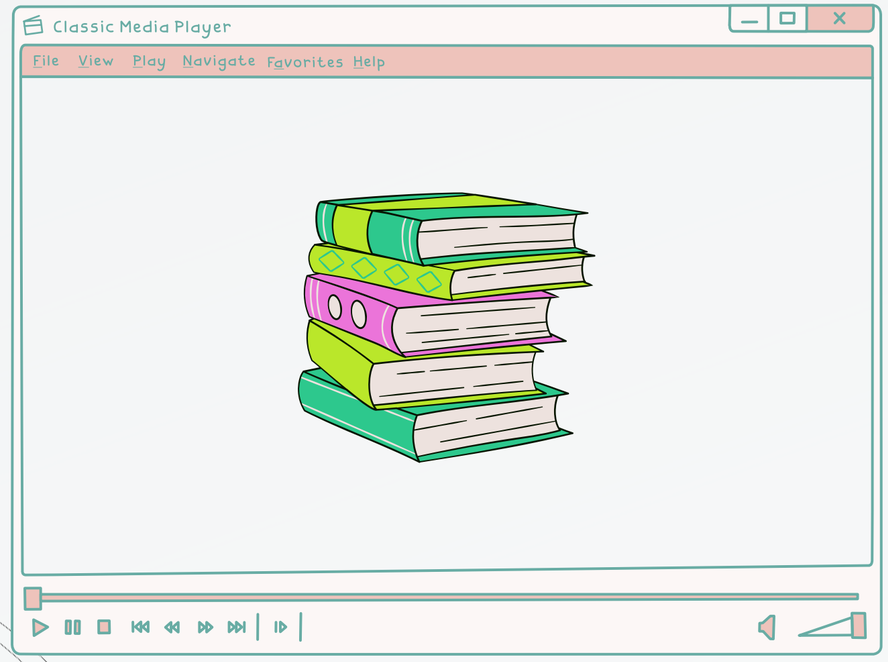
YouTube

At least 800 million videos

Two main sub-task:

- Recognition
- Translation

INTRODUCTION



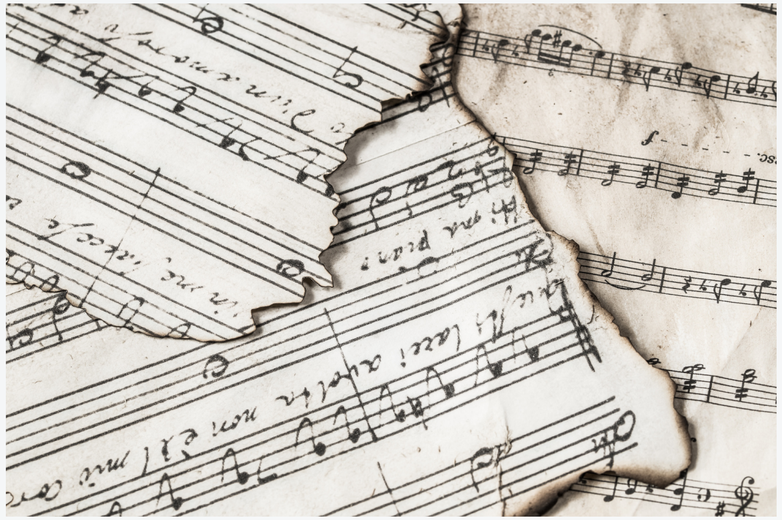
- Terminology
- Some accents hard to listen

INTRODUCTION

Obstacles' Vietnamese users:

- Researching requires a significant amount of English knowledge.
- Services available are hard to access

INTRODUCTION

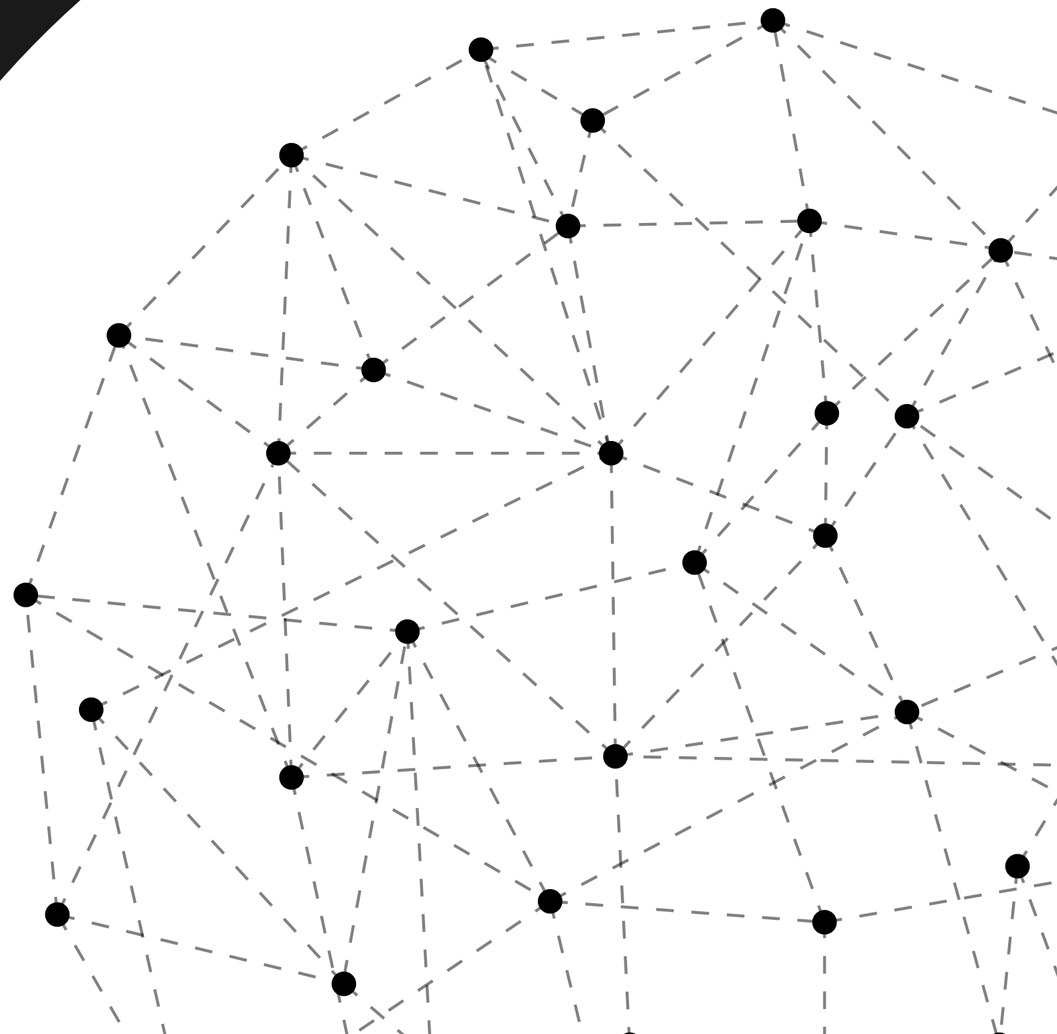


Support for learning

INTRODUCTION

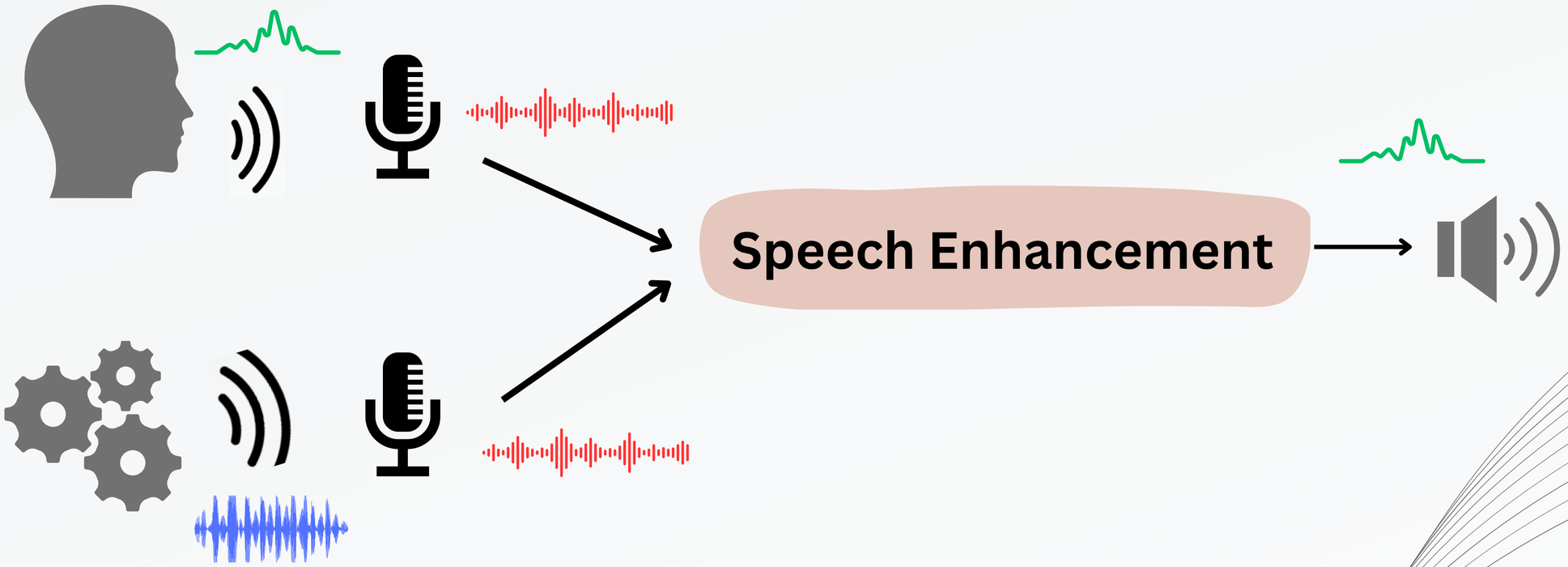
This thesis has focused on combining the latest optimized algorithms to create a robust application for generating Vietnamese subtitles from a random video

RELATED WORKS



Speech Enhancement

According to J. Benesty, S. Makino, and J. Chen, speech enhancement means "improving the intelligibility and quality of a degraded speech signal"



Speech Enhancement

Karman filtering, spectral subtraction, and Liljencrants–Fant

Wu, C., Li, B., & Zheng, J. (2011). A Speech Enhancement Method Based on Kalman Filtering

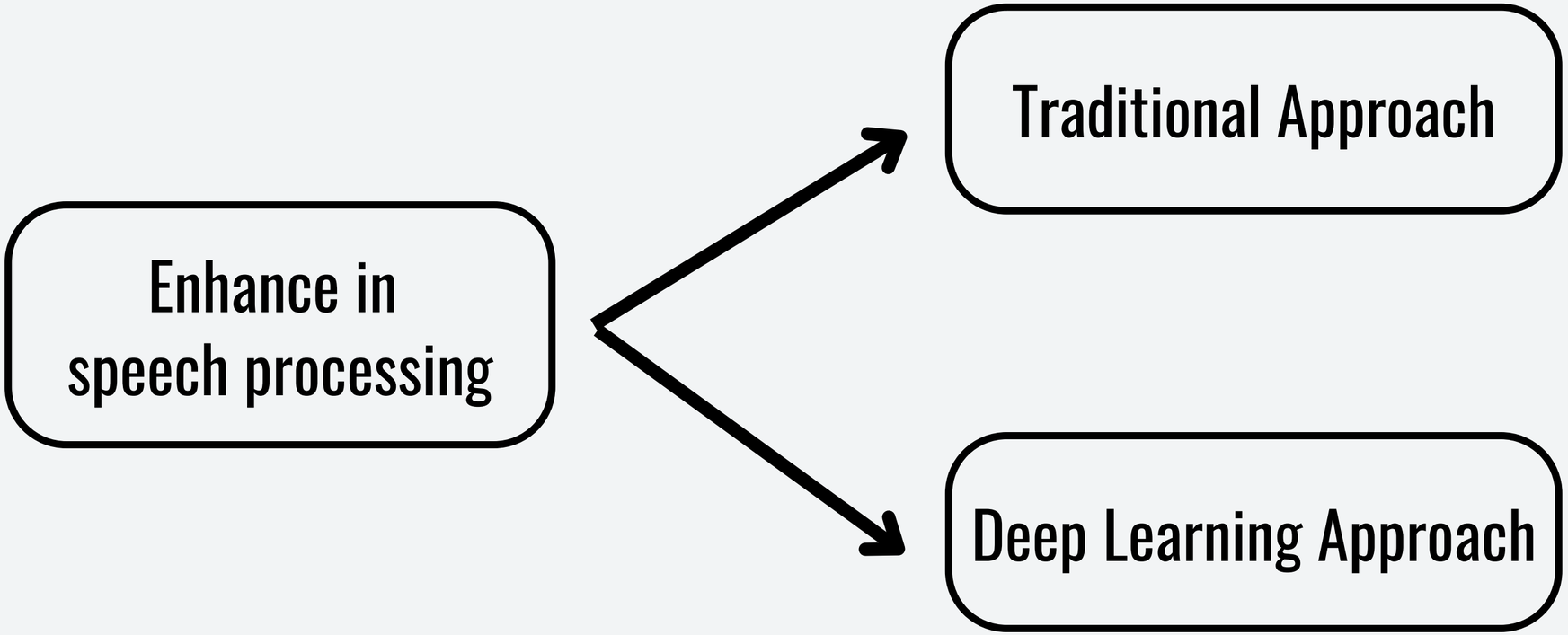
Spectral subtraction algorithm and use SNR to evaluate

Kaladharan N. (2014) Speech enhancement by spectral subtraction method

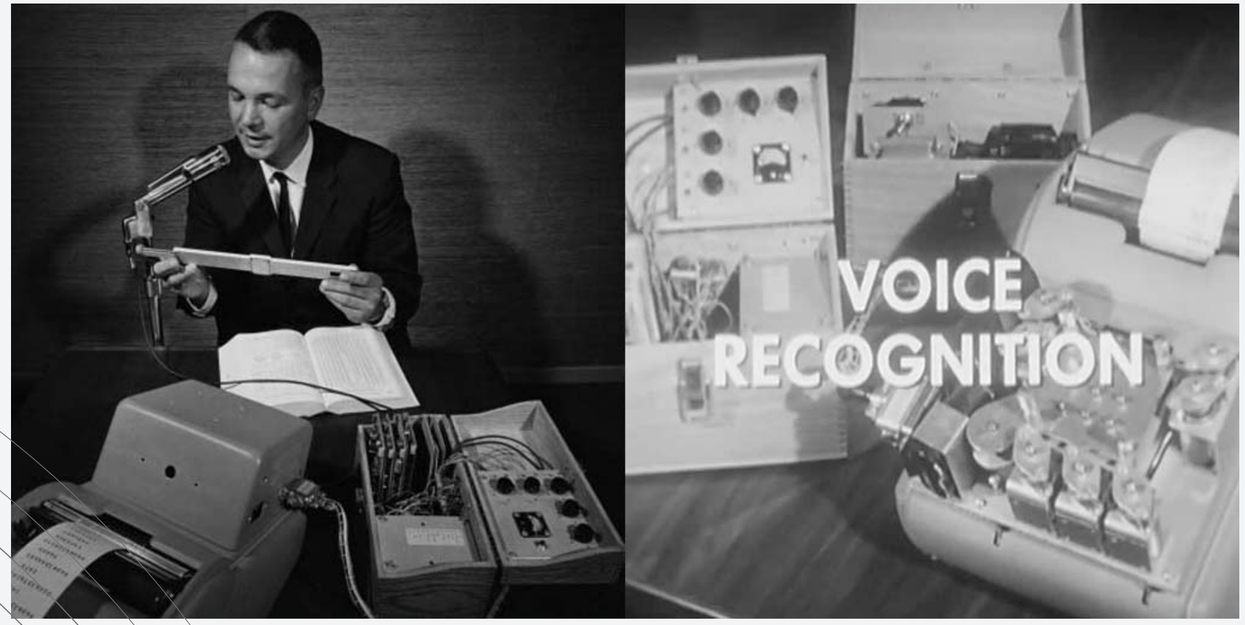
Magnitude and phase spectrum compensation method

Li, Z., Wu, W., Zhang, Q., Ren, H., & Bai, S. (2016). Speech enhancement using magnitude and phase spectrum compensation

Speech Enhancement



Speech Recognition



- Audrey system
- IBM Shoebox

Speech Recognition

OpenAI Whisper

Facebook Wav2vec

Facebook Wav2vec 2.0

WER for evaluation - LibriSpeech dataset

Speech Translation

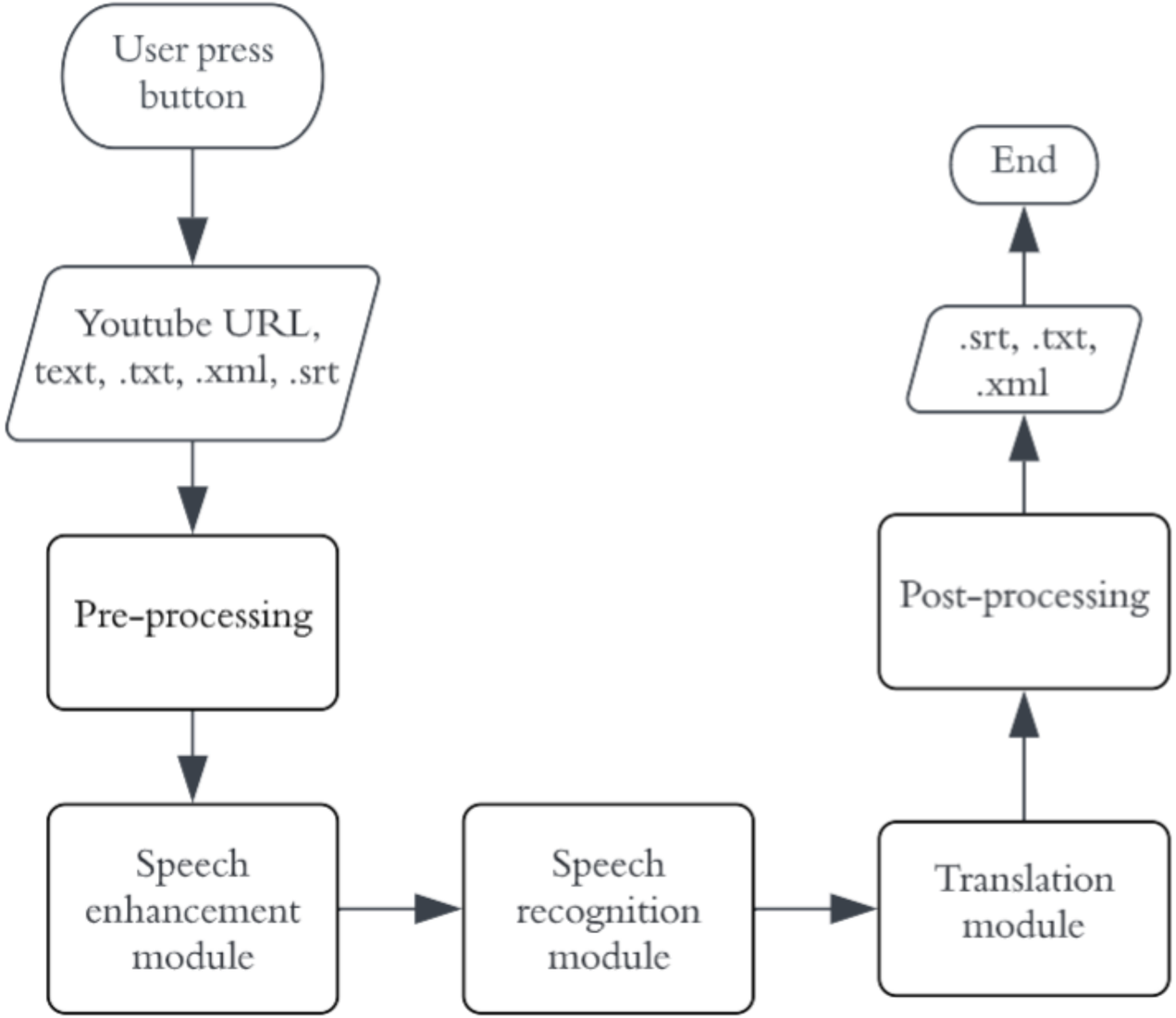
Context

Speech Translation

Transformers + VietAI + T5 framework

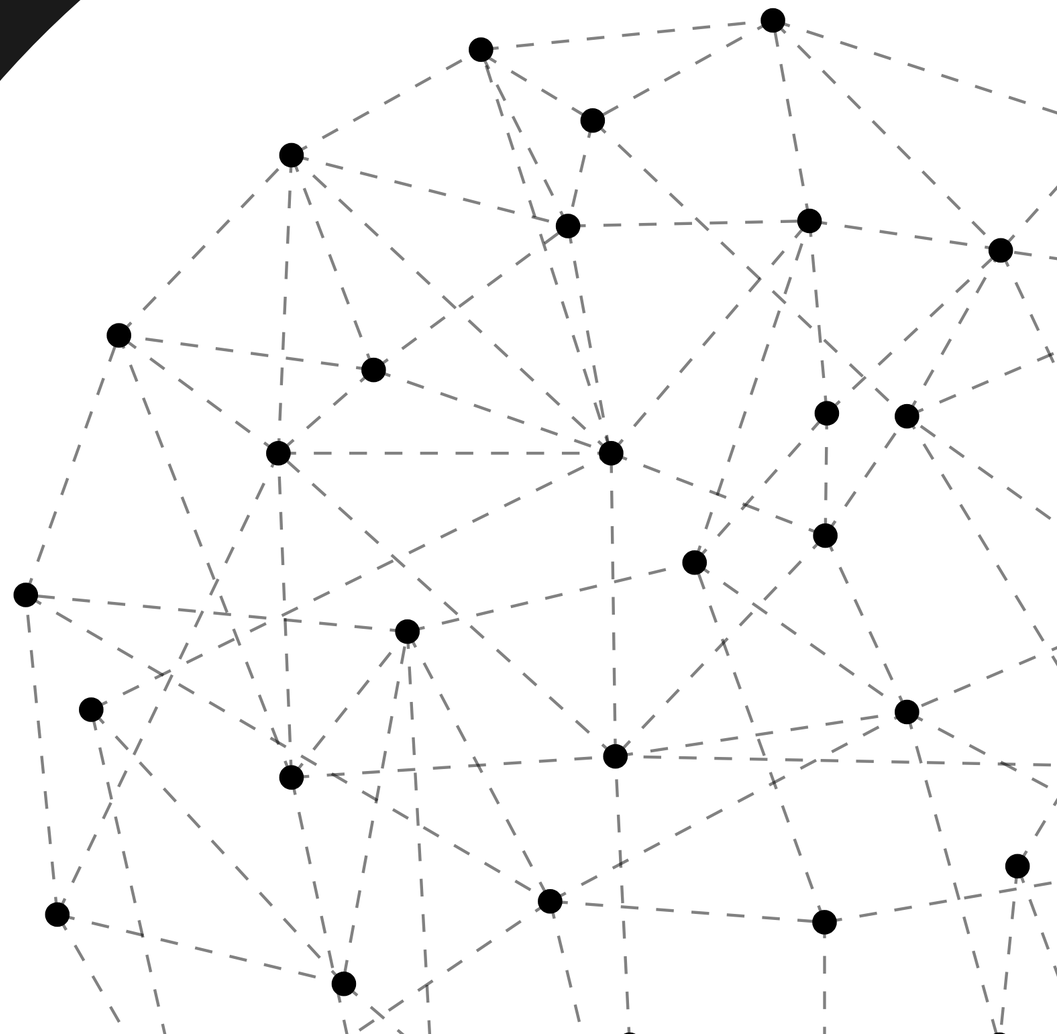
↓

EnViT5



Pipeline of Automatic Subtitle Generation Application

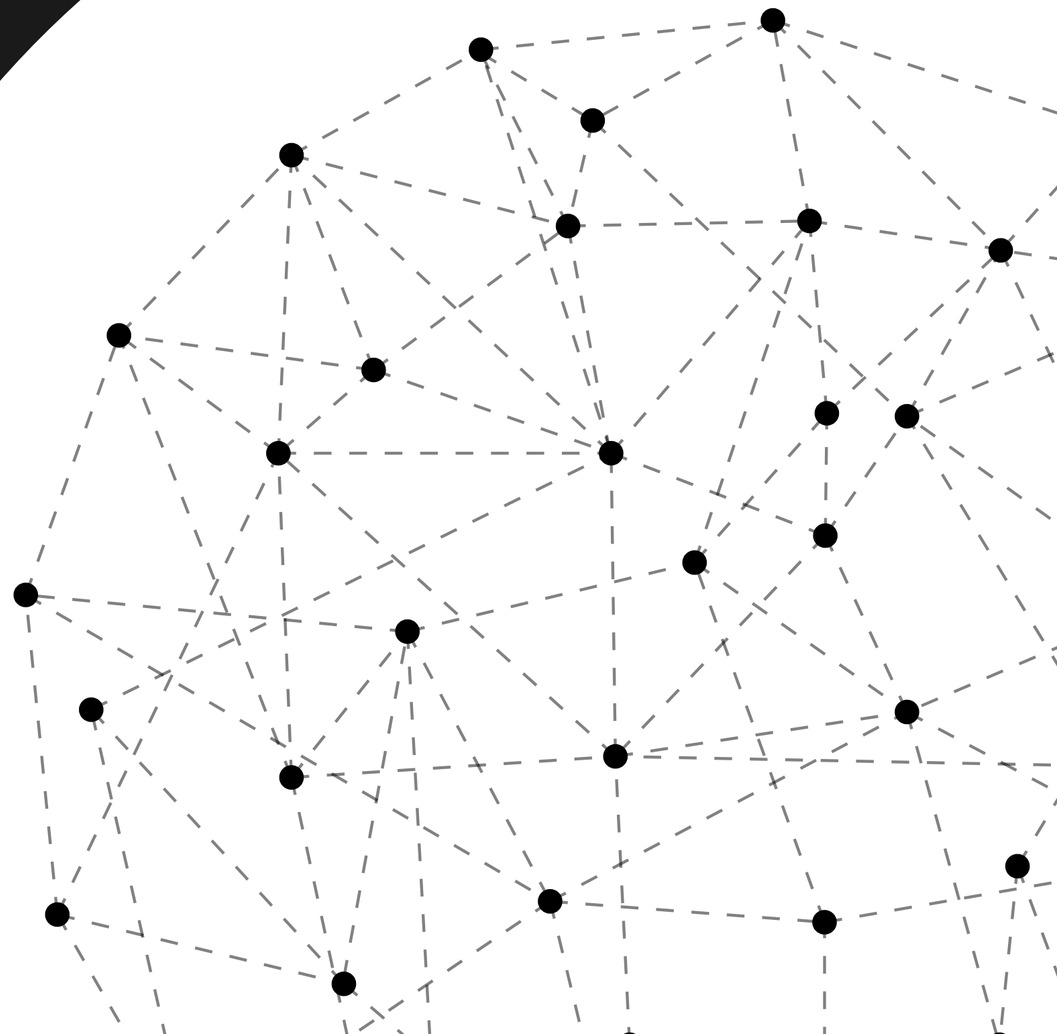
OBJECTIVES & CONTRIBUTIONS



Objectives & Contributions

- This thesis results in an effective pipeline for creating Vietnamese subtitles based on speech processing, and the latest models.
- Create an N2Vi subtitle generation application that is easy to access for Vietnamese users.

DATA PREPARATION



DATA PREPARATION

Youtube

More than 40 hours video

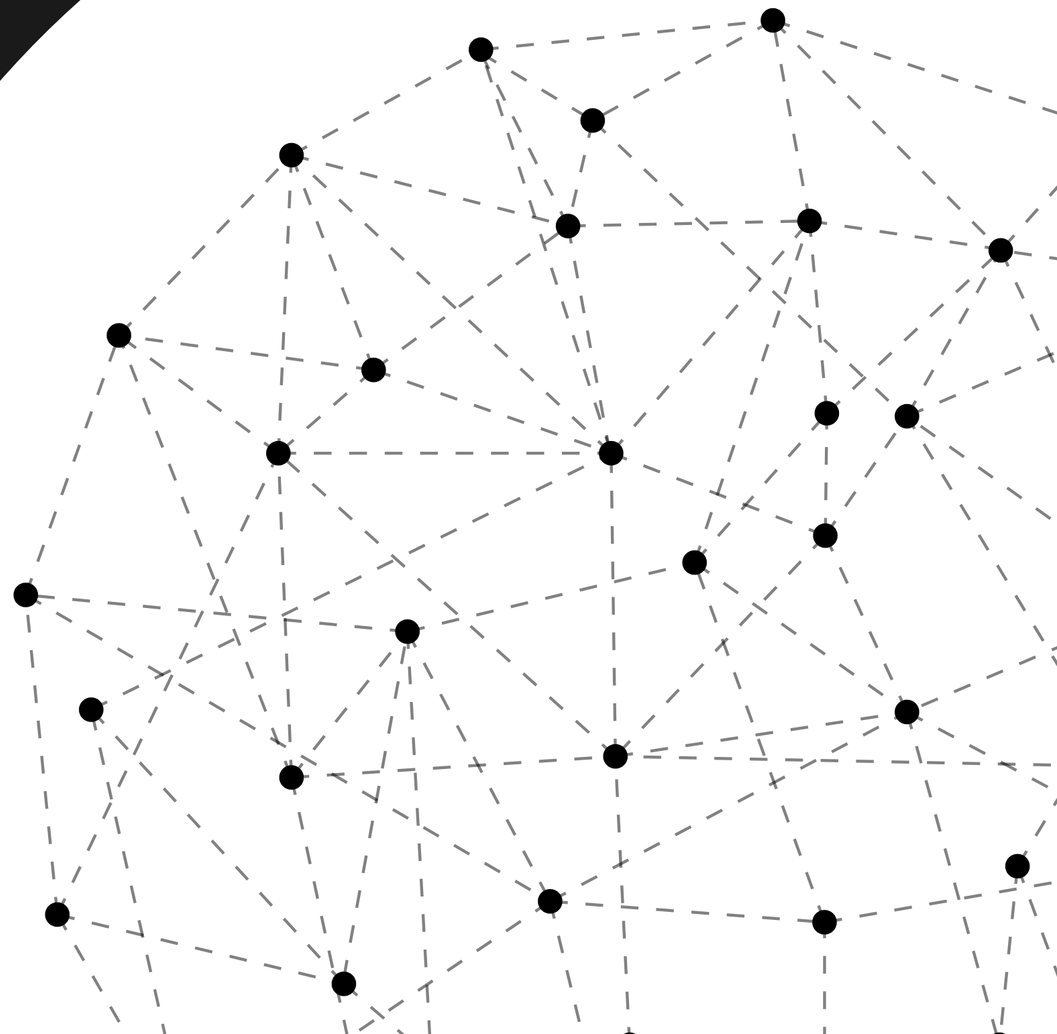
- Extract and select the desired video data from Youtube
- A diverse set of videos with various speakers, many accents, background environments and topics

- One audio file (.mp4 format)
- Two subtitles files corresponding to English and Vietnamese (.xml format)

Describe

Get from
Ted-ed Channel

MODULES & FLOWS

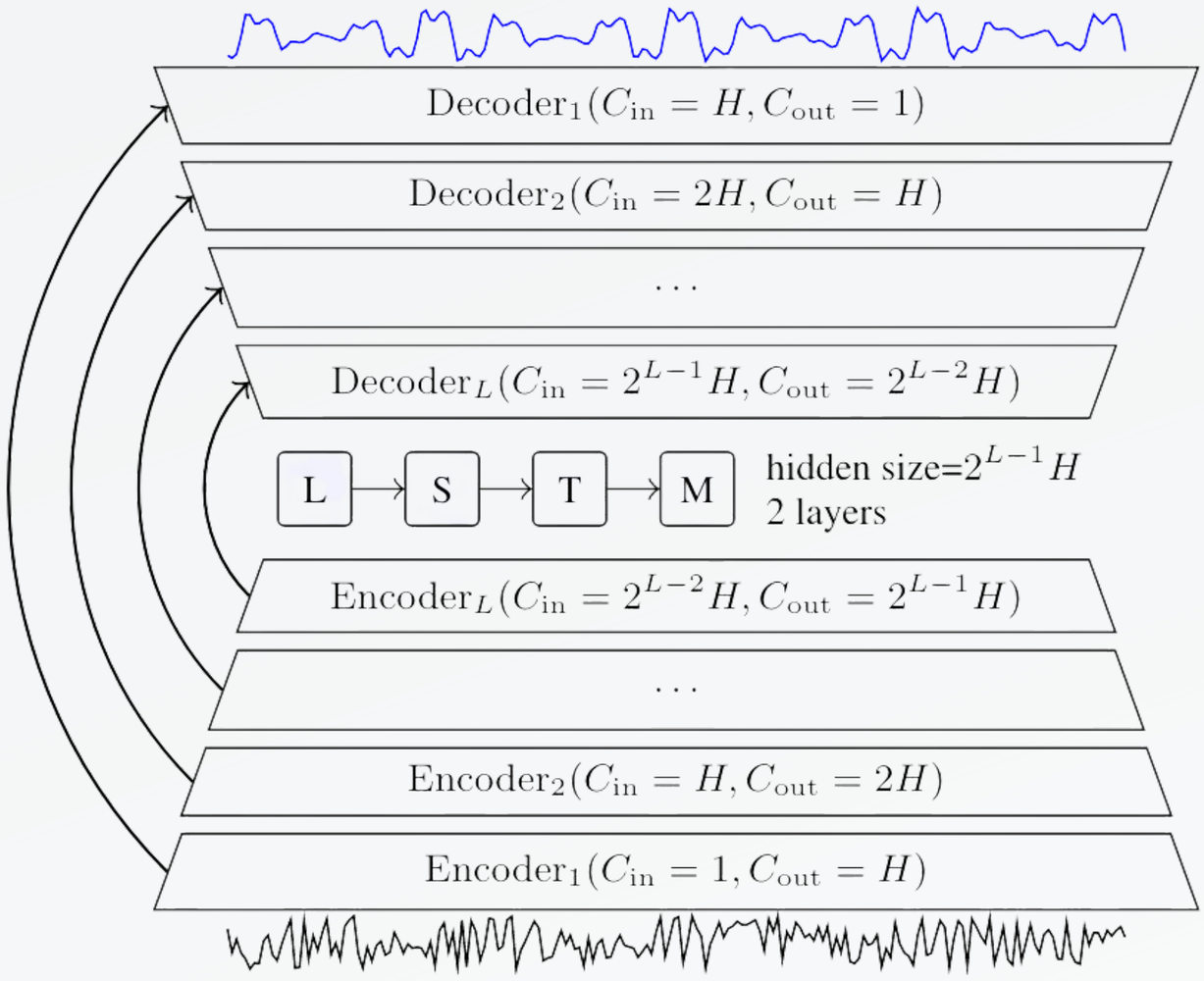


OVERVIEW

English to Vietnamese subtitle generation application

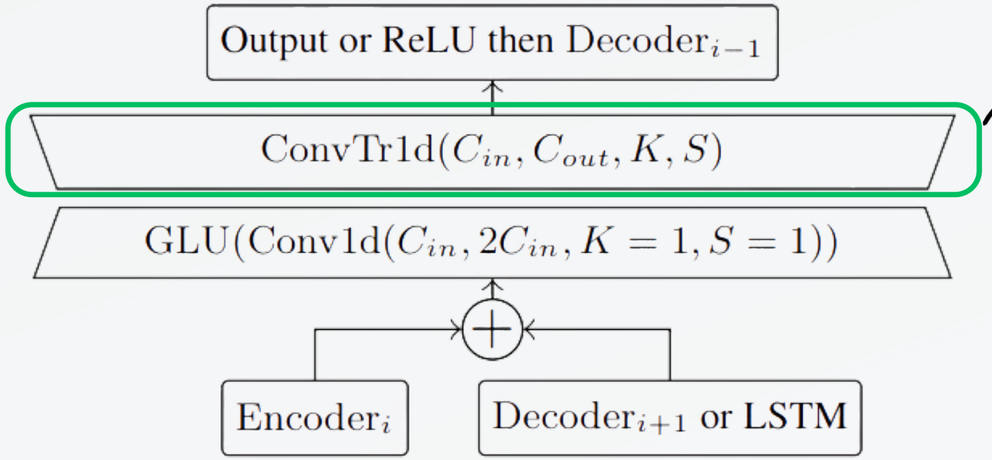
- Enhancement
- Recognition
- Translation

ENHANCEMENT



DEMUCS architecture

ENHANCEMENT

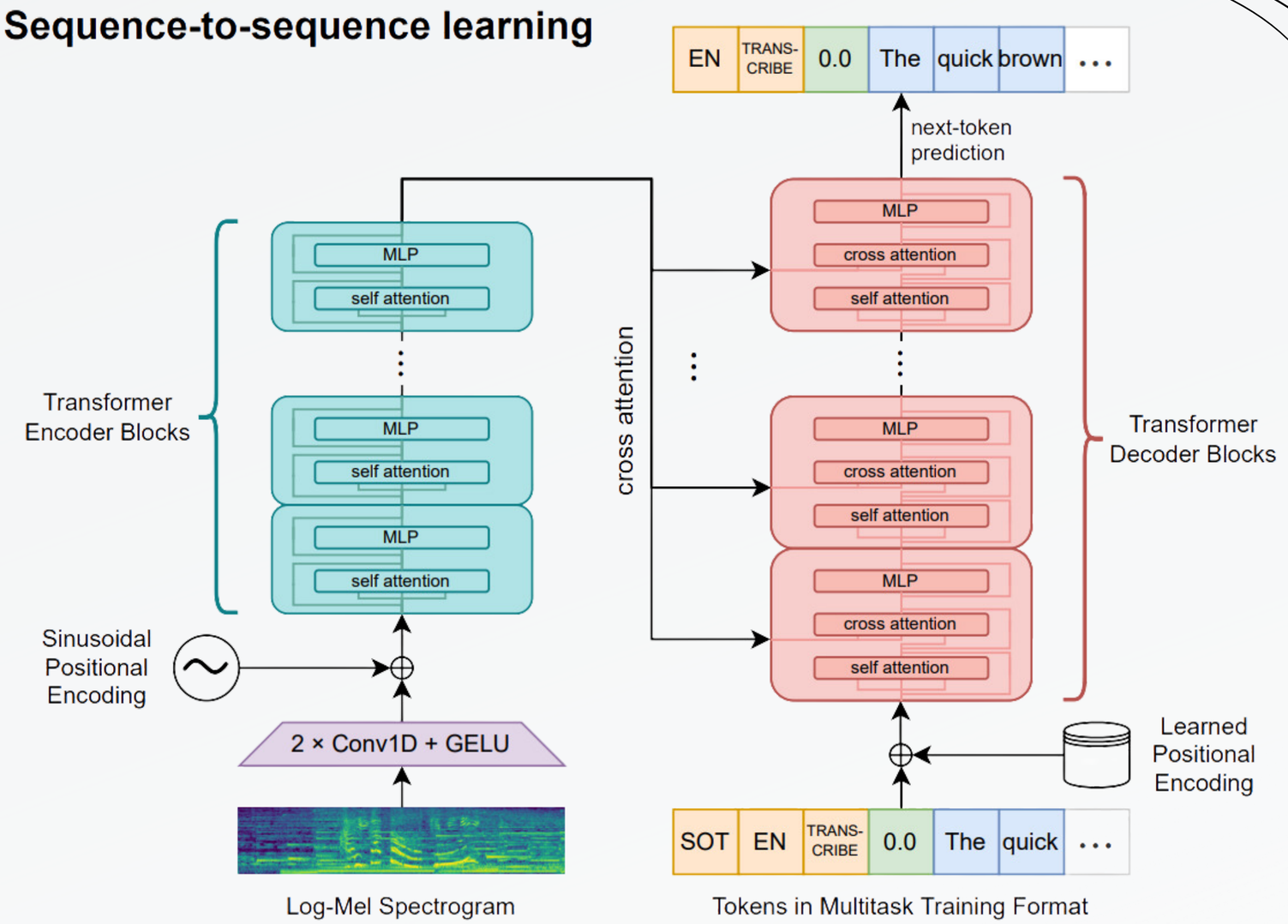


DEMUCS decoder architecture

require 4 x operations and memory less than Wave-U-Net

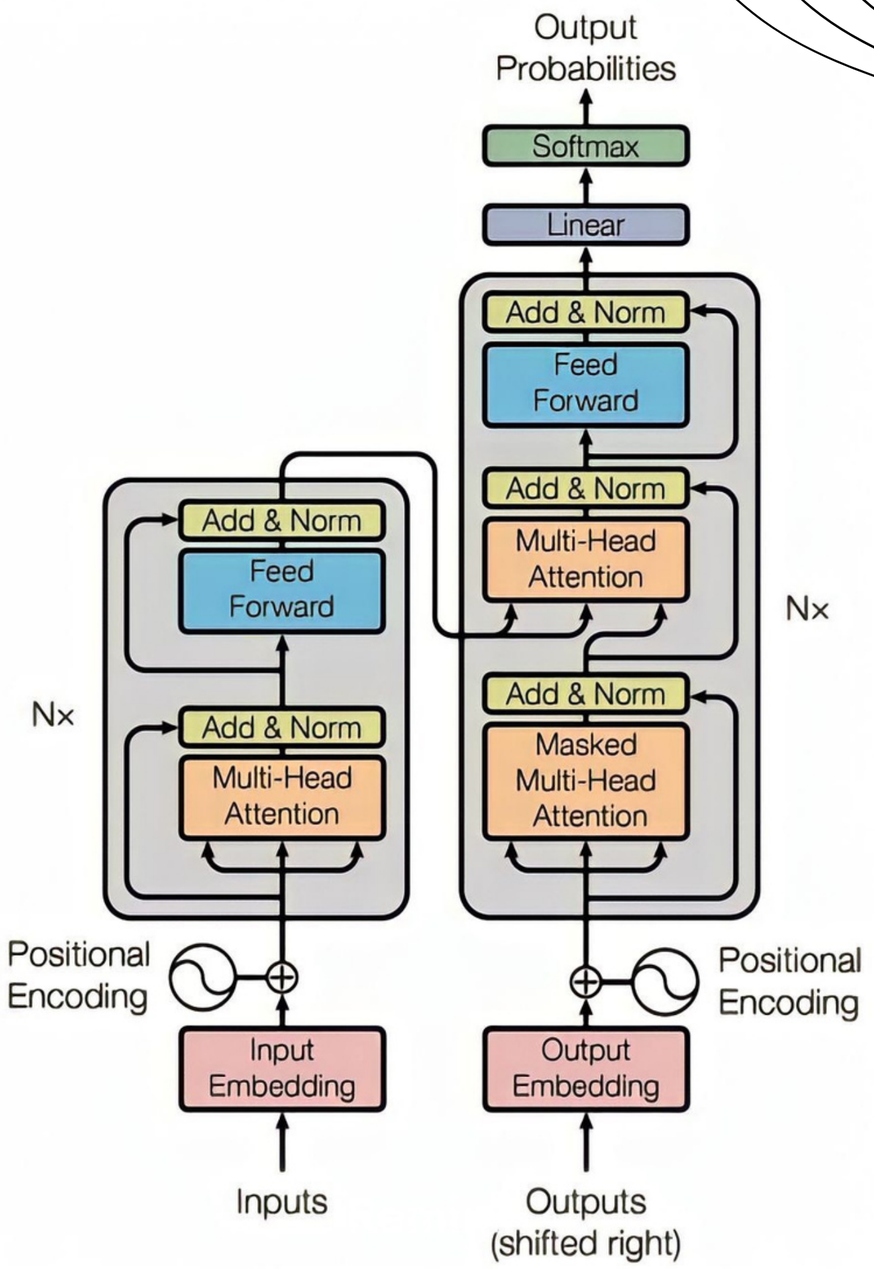
RECOGNITION

Whisper architecture



TRANSLATION

EnViT5 inherits transformer's architecture with N = 12



TRANSLATION

Iterative outputs

Long input text

Long required output length

TRANSLATION

This thesis results in an effective pipeline for creating Vietnamese subtitles based on speech processing, and the latest models to achieve this goal. Besides, we also collected a dataset and used it to evaluate some recent state-of-the-art models and then come up with the most suitable ones for our system. We also dig down into those models to carry out processing methods to improve the outcomes.

398 > 256

TRANSLATION

This thesis results in an effective pipeline for creating Vietnamese subtitles based on speech processing, and the latest models to achieve this goal.

150 < 256

Besides, we also collected a dataset and used it to evaluate some recent state-of-the-art models and then come up with the most suitable ones for our system. We also dig down into those models to carry out processing methods to improve the outcomes.

249 < 256

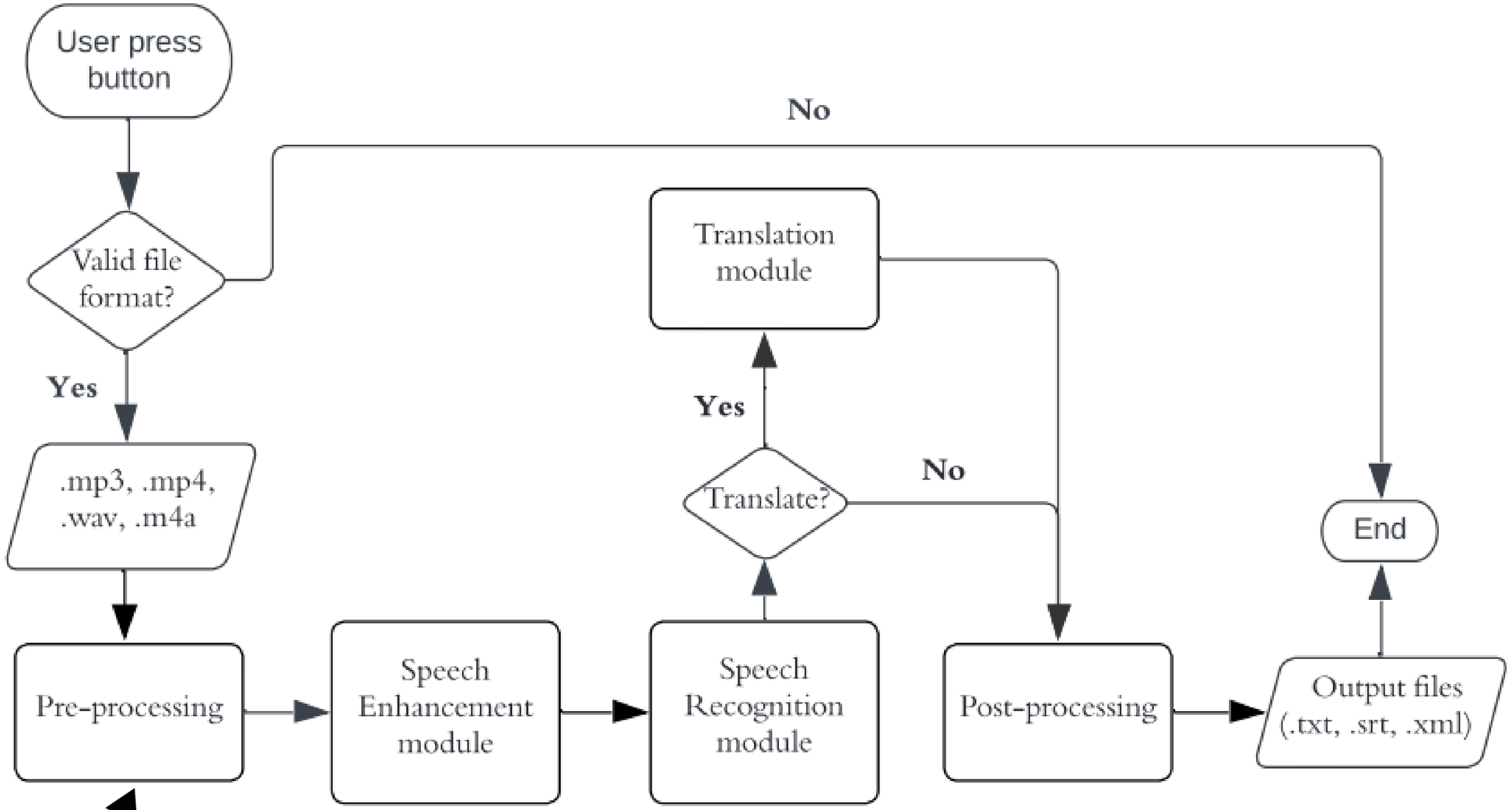
Split long input text into chunks with threshold!

OVERVIEW

English to Vietnamese subtitle generation application

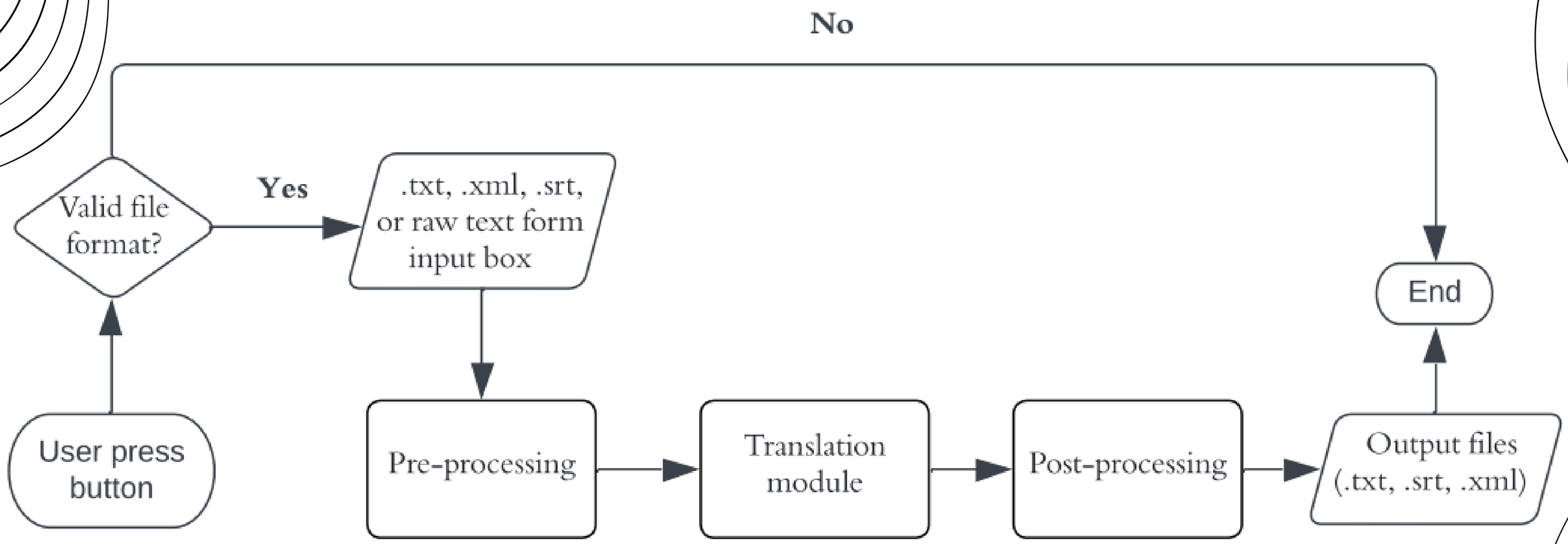
- Recognition feature
- Translation feature
- Tubescribe feature

RECOGNITION FLOW



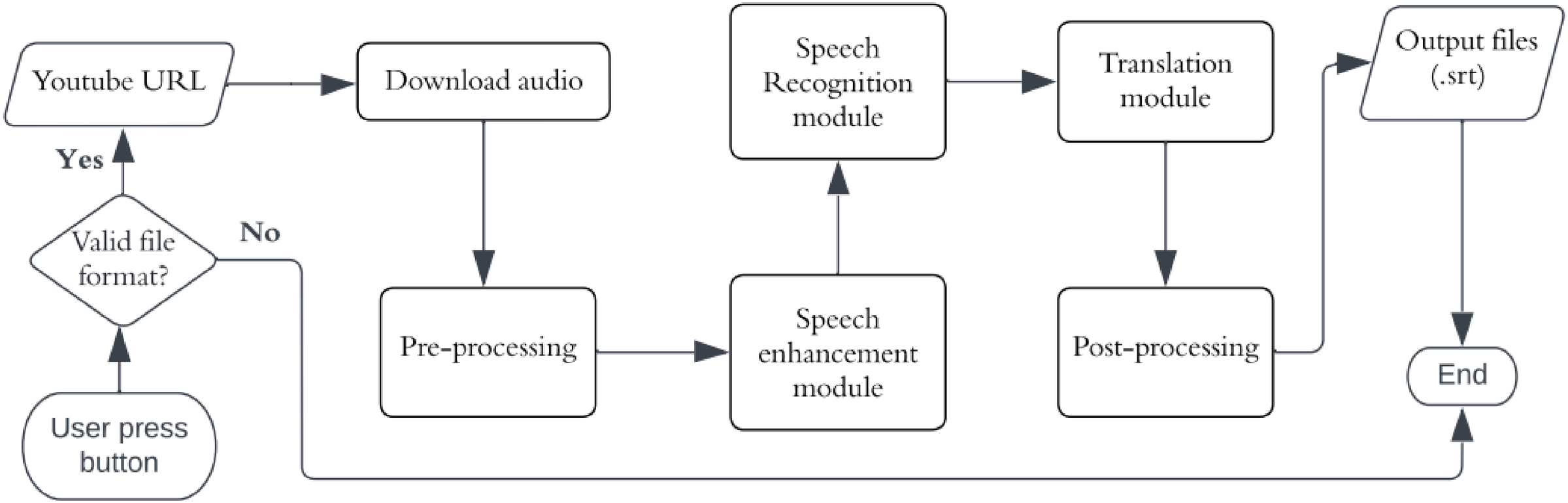
If input is video, audio is extracted here!

TRANSLATION FLOW

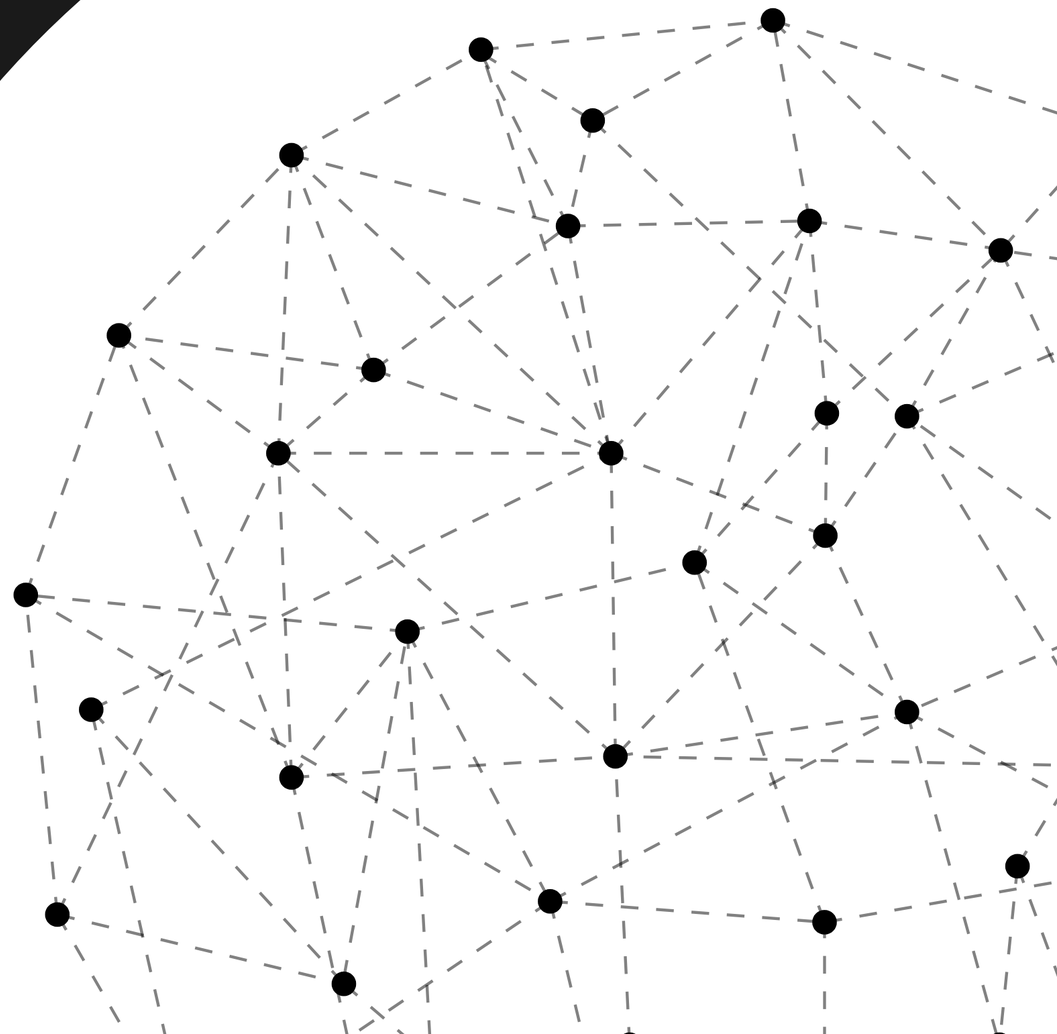


Chunk splitting here!

TUBESCRIBE FLOW



EXPERIMENT RESULT



DATASET OVERVIEW

- For the speech recognition task, we have 500 audio files approximately 40 hours of audio, and their corresponding captions

- For the machine translation evaluation, after filtering valid data, we used 453 bilingual subtitle files. On average, each file contains about 3000 tokens, which means we have approximately 1.300.000 tokens for the evaluation machine translation model task.

EVALUATION METRIC

1. WER (Word Error Rate) score

$$WER = \frac{S + D + I}{N}$$

- A substitution (S) occurs when a word gets replaced.
- An insertion (I) occurs when a word is added that is not in the ground truth.
- A deletion (D) happens when a word is left out of the transcript completely.
- N is the total words of the actual/reference sentence.

EVALUATION METRIC

2. BLEU (Bilingual Evaluation Understudy) score

$$\textit{Brevity Penalty} \times \textit{Geometric Average Precision } (N)$$

$$\begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c <= r \end{cases}$$

Target Sentence: The guard arrived late because it was raining

Predicted Sentence: The guard arrived late because of the rain

COMPARISON

1. Speech recognition

- facebook/wav2vec2-large-960h-lv60-self output:

```
excuse me were you in the military by chance fes thank you for your service o maybe in rorward
to im very old oh that's awsul not ninety five years old i'll tell you what if you can tell me
a story from worldwardnto i'll pay for your cart to day a he ha ha a you're af fello no i want
to first time i rot out to sea i got so sea fick i thought i couldna die halfo that everything
was o gay and i was aboard jifor clatawhile and that i was on the island me guam for a maughty
```

- openai/whisper-medium output:

```
 Excuse me. Were you in the military by chance? Yes. Thank you for your
service. Oh, maybe in World War two. I'm very old. Oh, that's awesome. I'm
95 years old. I'll tell you what if you can tell me a story from World War
two. I'll pay for your cart today. No, I want to first time I ever went out
to sea I got so seasick. I thought I was gonna die after that everything
was okay. And I was a board chump for quite a while and then I was on the
```

COMPARISON

1. Speech recognition

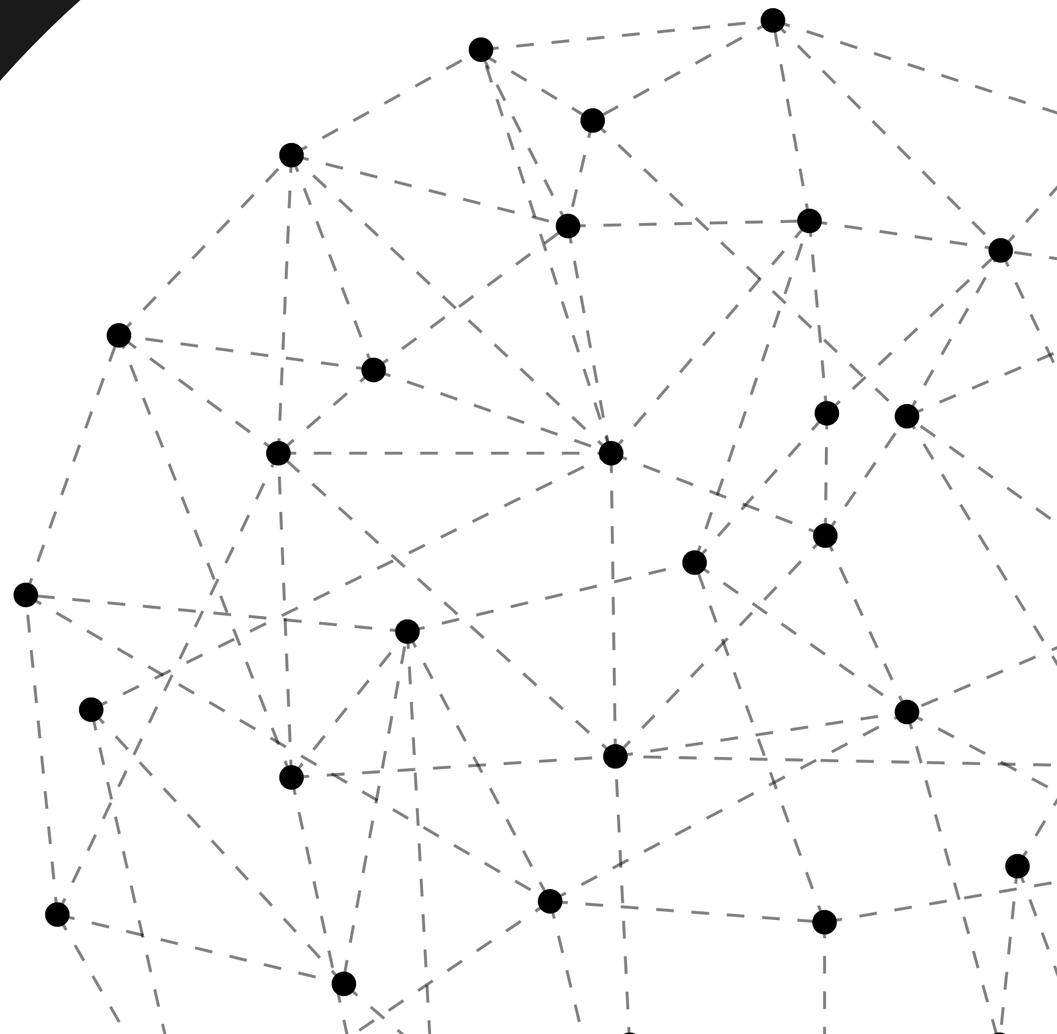
Model name	*Params*	*WER*
openai/whisper-medium	769M	1.0065
openai/whisper-tiny.en	39M	1.0312
jonatasgrosman/wav2vec2-large-xlsr-53-english	315M	1.0327
facebook/wav2vec2-base-960h	94M	1.0766
facebook/wav2vec2-large-960h-lv60-self	315M	1.0788

COMPARISON

2. Machine translation

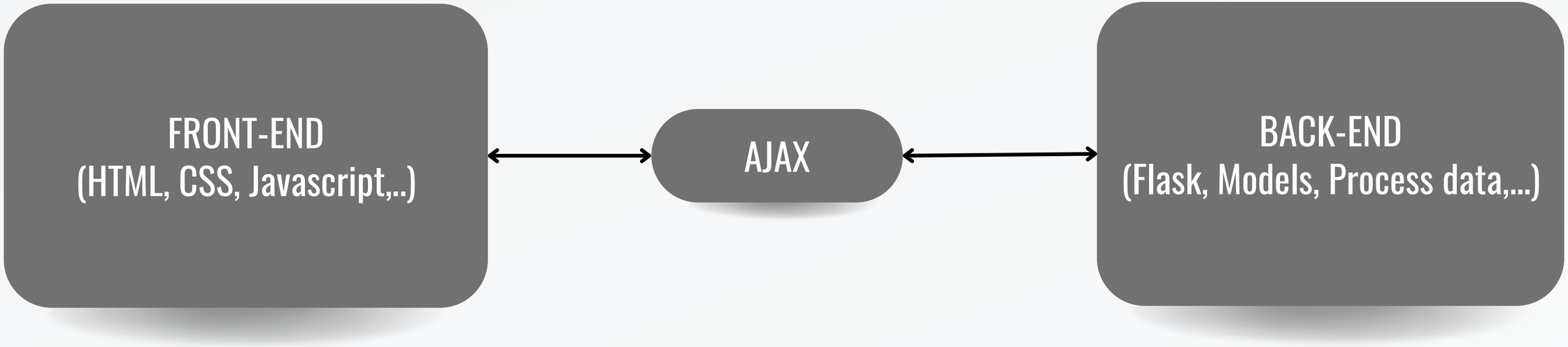
Model name	*BLeU*
Google translate	0.2453
Amazon translate	0.2969
EnViT5-base	0.3192
EnViT5-base + Our preprocessing method	0.3255

USER INTERFACE



USER INTERFACE

1. Overview



USER INTERFACE

1. Overview

TRANSLATE

Similar to google translate with one upgrade that users can translate their own subtitle files

RECOGNIZE

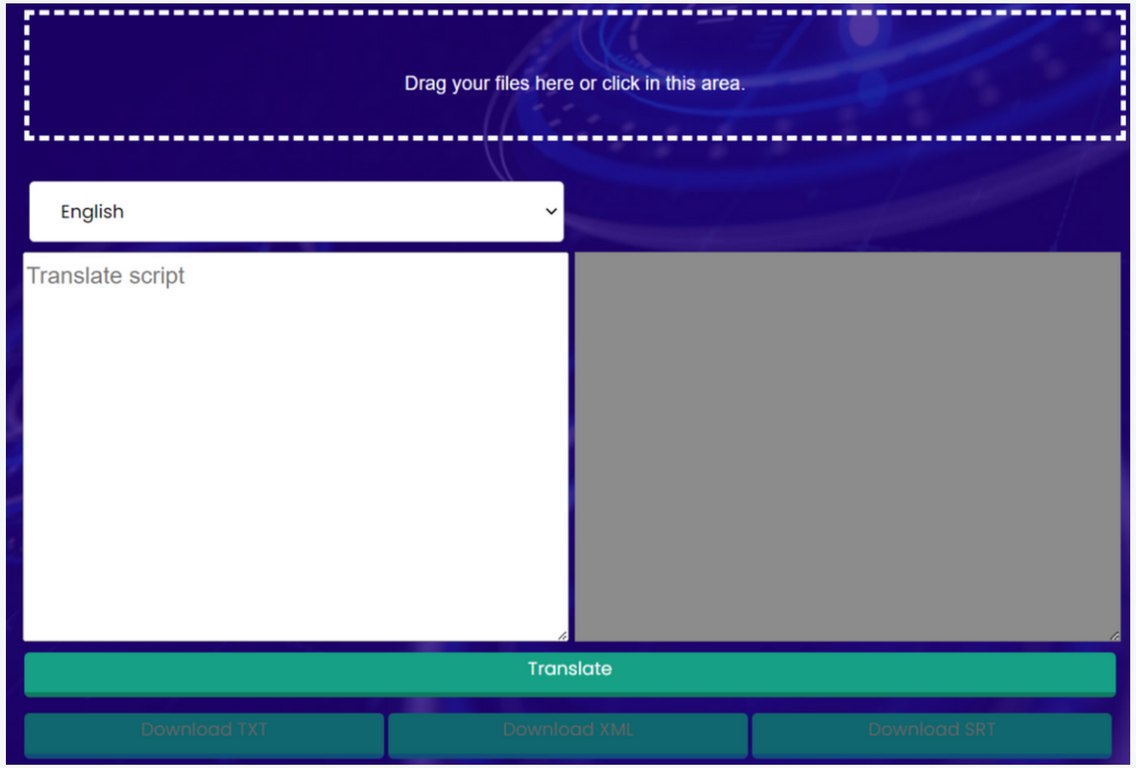
Recognize user's own audio. The result can be in English or Vietnamese

TUBESCRIBE

Create Vietnamese subtitle files for Youtube video by parsing the video's URL

USER INTERFACE

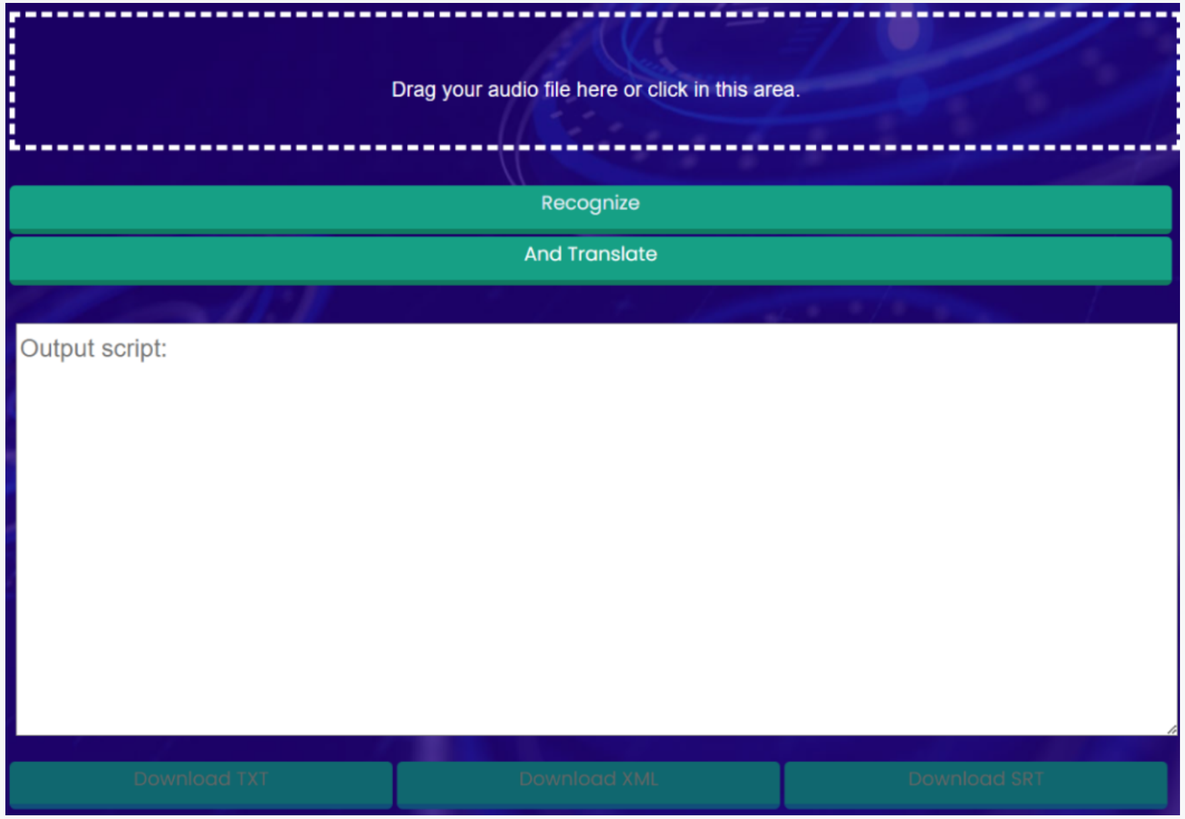
2. Translation



- Users can drag or upload text, XML or SRT files. The file contents will be displayed in the white box.
- The output will be displayed on the right grey box for users checking the content before choosing to download.
- The output file format is based on the input file format the user uploaded:
  - TXT: .txt
  - XML: .txt, .xml, .srt
  - SRT: .txt, .srt

USER INTERFACE

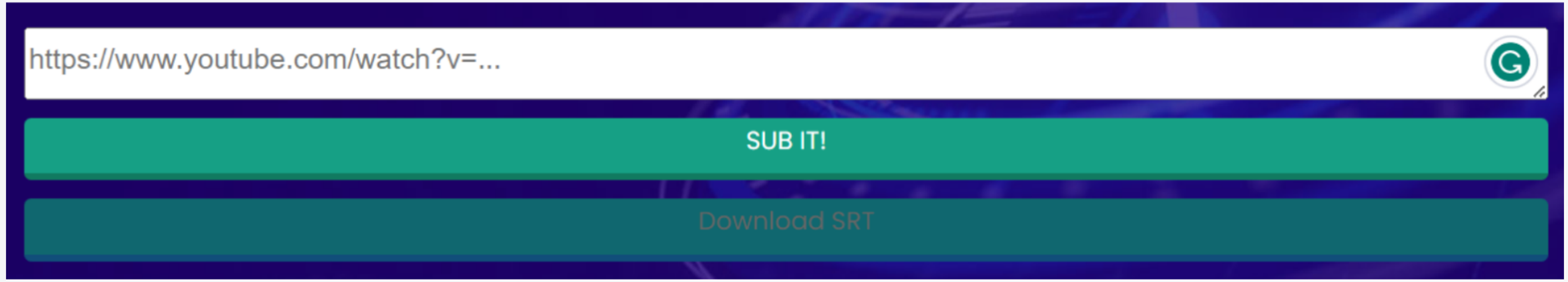
3. Recognition



- Users can drag or upload audio files (.mp3. .mp4, .m4a, .wav)
- The output will display on the below box for users checking the content script before choosing to download
- The output file formats are text, XML, and SRT

USER INTERFACE

4. Tubescribe



- Users enter the URL of their desired Youtube video
- After generating the subtitle done user can download the SRT file and then uses some third-party extension to attach SRT file to the Youtube video

CONCLUSION AND FUTURE WORK

FUTURE WORK

The project remains some functions that can be upgraded in the future:

- The recognition phase is now limited, it only supports English input. In the future, we can integrate the Vietnamese as well as other languages as input for speech recognition model.
- The application can support more languages than only English - Vietnamese in translation module
- The application can be optimized to become lighter and stronger to deploy in real-time.

CONCLUSION

In this study, we have learned about the relevant tasks of a Vietnamese subtitle generation application. We have tested, compared, and improved in inferring those pre-trained models. At the same time, an end-to-end “N2Vi” captioning application is designed, connected, and implemented. Up to the present, our application can mostly meet users' basic usage needs and the expectation for great improvements in the future.

DEMO

QUESTION & ANSWER

THANK FOR LISTENING