LECTURE SLIDES ANNOTATION WITH CODE-SWITCHED SPEECH

Done By: Yusuf Muhammad Eissa Ahmed Eissa Ammar

Supervisors: Dr. Nada Sharaf & Dr. Caroline Sabty

TABLE OF CONTENTS

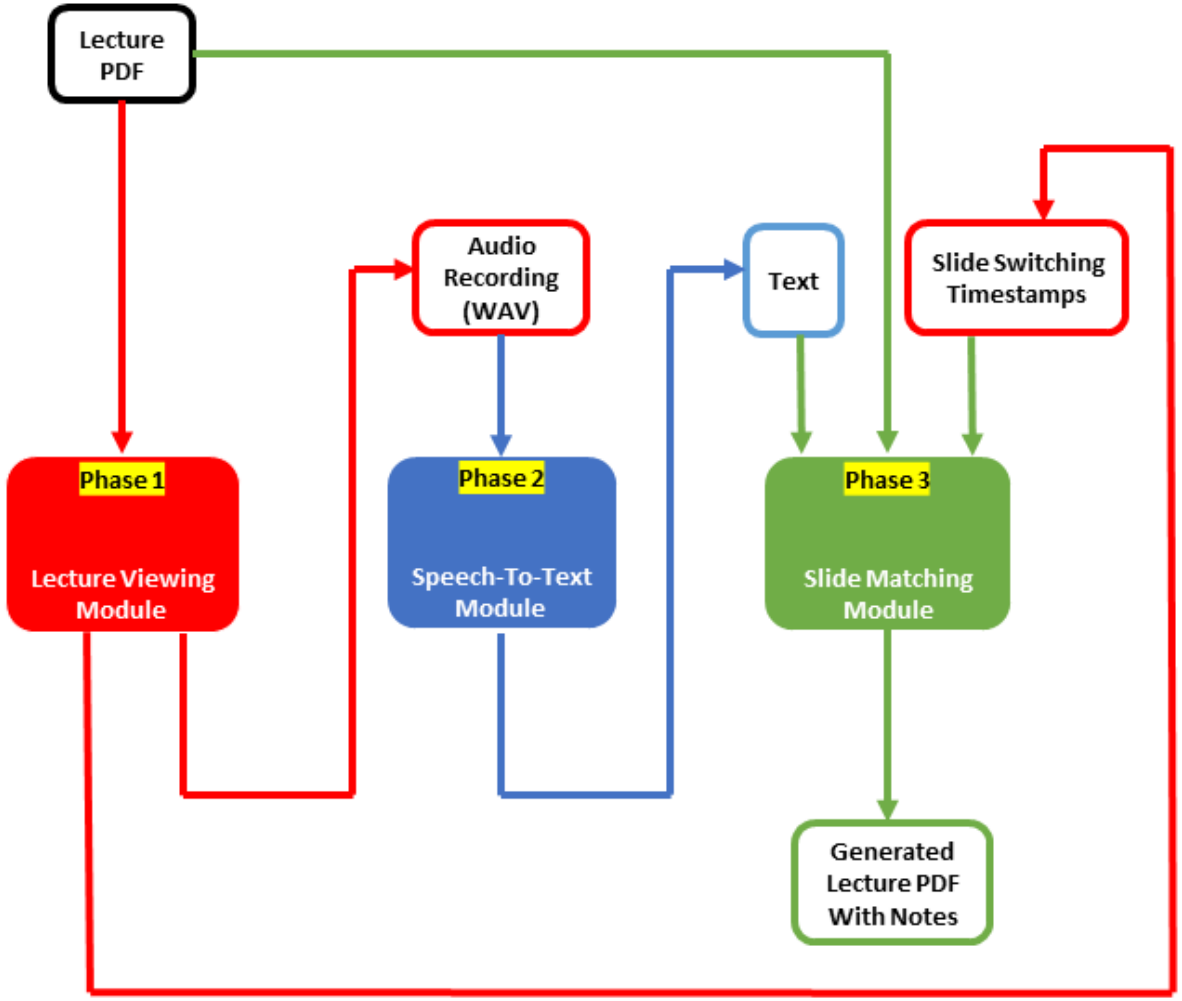
I. INTRODUCTION

INTRODUCTION

- Education & Technology
- Objective

II. METHODOLOGY

SYSTEM ARCHITECTURE



PHASE 1 - LECTURE VIEWING MODULE

❖**Description**

❖**Input:**

- Lecture PDF File

❖**Output:**

- Audio Recording (WAV)
- Slide Switching Timestamps

PHASE 1 - LECTURE VIEWING MODULE

❖**Components**

- PDF Viewer
- Sound Recorder
- Stopwatch

❖**Website Flow**

❖**Controls**

PHASE 2 - SPEECH-TO-TEXT MODULE

❖**Description**

❖**Challenges**

❖**Input:**

- Audio Recording (WAV)

❖**Output:**

- Text

PHASE 2 - SPEECH-TO-TEXT MODULE

Automatic Speech Recognition (ASR) System Selection

- **Test Set**: 20 short audio files (10 English-Only & 10 Arabic-Only)
- **Speech Recognition Open Source Libraries**

1. VOSK
2. CMUSphinx

- **Speech Recognition API(s)**

1. Google Cloud Speech To Text
2. Microsoft-Azure Speech To Text

- **Results**

PHASE 2 - SPEECH-TO-TEXT MODULE

Microsoft-Azure Speech To Text API Configuration/Inputs

- **Subscription Key and Region**
- **Audio File (WAV) Path**
- **Language Code**
- **Enabling Word-Level Confidence**
- **Enabling Word-Level Timestamps**
- **Enabling Profanity Filter**
- **Enabling Detailed Output Format**
- **Using Continuous Speech Recognition Method**

PHASE 2 - SPEECH-TO-TEXT MODULE

<u>Approach 1:</u> Continuous Speech Recognition With Continuous Language Identification

Step 1: Recognizing Audio File Using Continuous Language Identification Feature

-Candidate Languages

Step 2: Output text (includes timestamp & confidence level of each word)

PHASE 2 - SPEECH-TO-TEXT MODULE

Approach 2: Continuous Speech Recognition With Overlap Filtration

Step 1: Recognizing Each Language Separately

Step 2: Merging Outputs & Ordering by Timestamps

Step 3: Overlap Filtration

$$Overlap = FirstWordOffset + FirstWordDuration - SecondWordOffset$$

$$OverlapPercentage = Overlap/SecondWordDuration$$

Figure 3.3: Overlap Calculation

Step 4: Output text (includes timestamp & confidence level of each word)

PHASE 3 – SLIDE MATCHING MODULE

❖**Description**

❖**Input:**

- Lecture PDF File
- Text
- Slide Switching Timestamps

❖**Output:**

- Generated Lecture PDF with Notes

PHASE 3 – SLIDE MATCHING MODULE

Step 1: Time Alignment

1. Splitting text into chunks using **slide switching timestamps**
2. Matching chunks to slides correspondingly.

-After this step, every slide will have the text said when the slide was viewed, regardless of the contents of the slide

PHASE 3 – SLIDE MATCHING MODULE

Step 2: Sentence Similarity Using Deep Learning

1. Extract text from the lecture PDF file
2. Split matched notes of every slide into smaller chunks on silences that last for more than 5 seconds
3. Compare the smaller chunks of each slide to the text of all the slides, using a **sentence similarity deep learning model and cosine similarity** in an attempt to find a better match for each smaller chunk

PHASE 3 – SLIDE MATCHING MODULE

Deep Learning Model

- **Name:** Pyjay/sentence-transformers-multilingual-snli-v2-500k
- **Library:** Sentence-Transformers
- **Description:** Multilingual model that finds sentence similarity between code-switched text and English text.

Input:

- Reference Sentence
- Sentences to be compared to reference sentence

Output:

- Embeddings.

Cosine similarity is used to get a similarity score for each sentence compared to the reference sentence by using the embeddings created by the model

PHASE 3 – SLIDE MATCHING MODULE

How the deep learning model was chosen?

- **Test Set:** 100 Code-Switched sentences, 100 English-Only sentences, and 10 slides (contains English-Only text)
- **Models Tested:** 9 multilingual models and 2 English-Only models

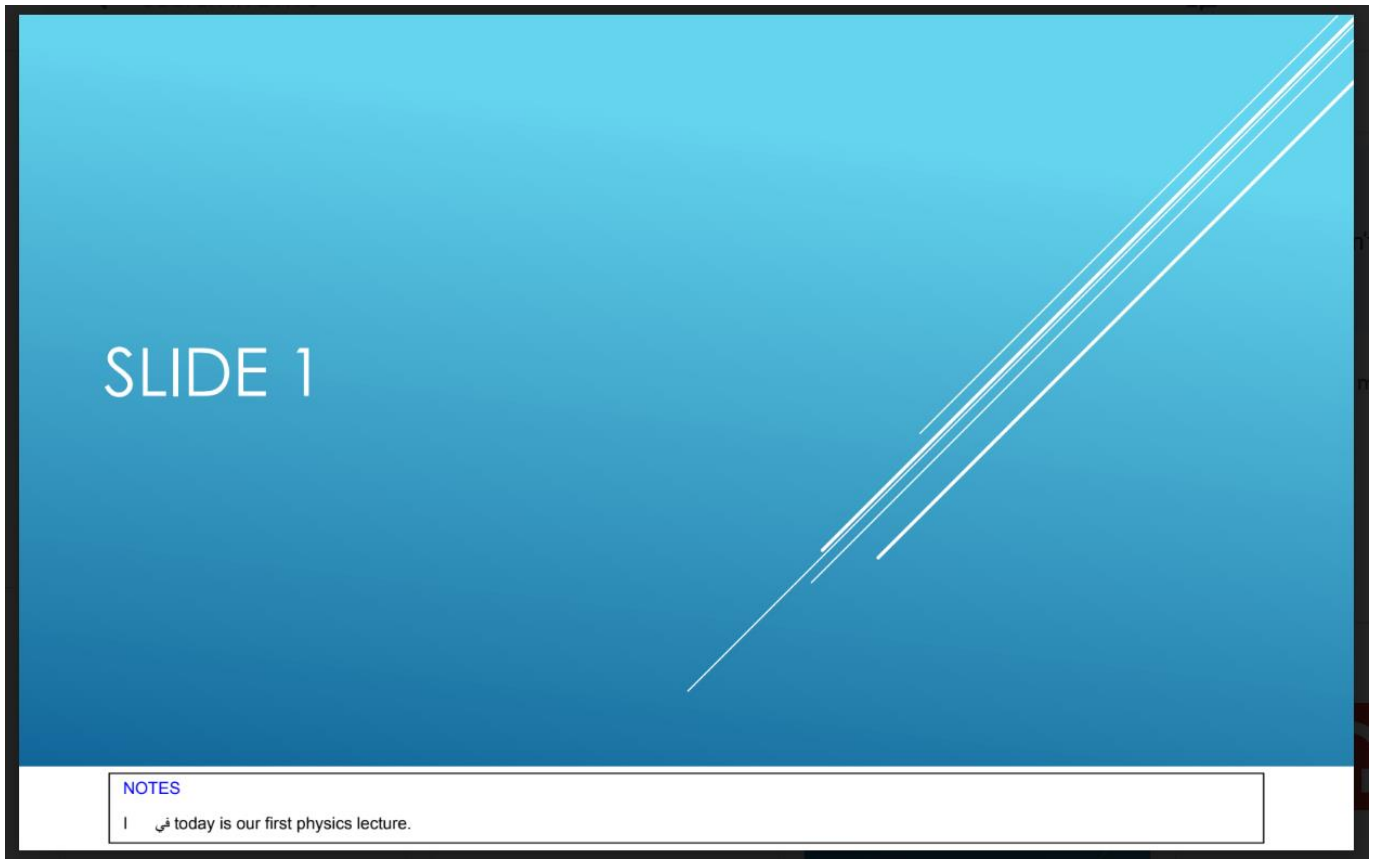
- **Test 1:** Matching Code Switched Sentences to Slides
  - Pyjay/sentence-transformers-multilingual-snli-v2-500 (multilingual model) → 87% (Highest Accuracy)

- **Test 2:** Matching English-Only Sentences to Slides
  - all-mpnet-base-v2 (English-Only model) → 85% (Highest Accuracy)
  - Pyjay/sentence-transformers-multilingual-snli-v2-500 (multilingual model) → 83% (2nd-Highest Accuracy)

- **Chosen Model & Trade-Off** (processing time)

PHASE 3 – SLIDE MATCHING MODULE

Step 3: Generating New Lecture PDF



III. RESULTS

WORD ERROR RATE (WER)

- What is WER?
- WER & System Accuracy Relation
- $$\text{WER} = \frac{\text{Added + Substituted + Deleted (words)}}{\text{Total Words (actually spoken)}}$$

SPEECH TO TEXT MODULE

Test Set:

- Short Code Switched Audio Files

  Quantity: 8 Duration: 3-6 secs

  Speaker: Me

  Description: 1 Intra-sentential code switched sentence.

- Long Code Switched Audio File

  Quantity: 1 Duration: 13 Minutes

  Speaker: Me

  Description: 100 Intra-sentential code switched sentences. Live lecture explanation during slide matching module testing.

SPEECH TO TEXT MODULE

Test Set:

- English-Only Audio File

  Quantity: 2 Duration: 1:15 & 12:30

  Speaker: Dr. Hassan Soubra (Embedded Systems Course, GUC VOD)

  Description: Speaker is French & not an native English speaker.

SPEECH TO TEXT MODULE

- **<u>Approach 1:</u> Continuous Speech Recognition With Continuous Language Identification**
- **<u>Approach 2:</u> Continuous Speech Recognition With Overlap Filtration**

Audio File	Approach 1 (WER)	Approach 2 (WER)
Short Code Switched (avg)	21%	8.8%
Long Code Switched	24.6%	13.3%
English-Only (avg)	2.18%	9.75%

Figure 4.5: WER Approaches Comparison

- Observation & Chosen Approach

SPEECH TO TEXT MODULE

DETALED RESULTS

Approach 1

Audio File	WER
CS-Short1.wav	25%
CS-Short2.wav	0%
CS-Short3.wav	44.4%
CS-Short4.wav	33.3%
CS-Short5.wav	25%
CS-Short6.wav	12.5%
CS-Short7.wav	20%
CS-Short8.wav	9%
Average	**21%**

Figure 4.1: WER of the Short Code Switched Audio Files

Audio File	WER
English-Short.wav	2.6%
English-Long.wav	1.76%
Average	**2.18%**

Figure 4.2: WER of the English-Only Audio Files

Approach 2

Audio File	WER
CS-Short1.wav	0%
CS-Short2.wav	20%
CS-Short3.wav	0%
CS-Short4.wav	16.6%
CS-Short5.wav	12.5%
CS-Short6.wav	12.5%
CS-Short7.wav	0%
CS-Short8.wav	9%
Average	**8.8%**

Figure 4.3: WER of the Short Code Switched Audio Files

Audio File	WER
English-Short.wav	5.8%
English-Long.wav	13.7%
Average	**9.75%**

Figure 4.4: WER of the English-Only Audio Files

SLIDE MATCHING MODULE

- Testing Slide Matching Module & Full System

- **Testing Procedure**
  - Creating Mock Lecture
  - Using Website

- **Results**
  - Test 1 (Speech Related To Slide Text) → 13.3% WER, 9/10 Slides
  - Test 2 (Speech **NOT** Related To Slide Text) → 13.3% WER, 9/10 Slides

IV. CONCLUSION

V. FUTURE WORK

THANKYOU FOR LISTENING

QUESTIONS?