Swayam: Transcription, Translation and Subtitling Project

1. Objective

Objective is to transcribe, translate and subtitle MHRD swayam Course video lectures (1300 hrs) in 8 Indian Languages which are detailed as follows.

1.1. **Transcription** of MHRD swayam Course Videos (English)

1.2. **Translation** and **subtitling** into following Indian languages

- 1.2.1. Bangla
- 1.2.2. Gujarati
- 1.2.3. Hindi
- 1.2.4. Kannada
- 1.2.5. Malayalam
- 1.2.6. Marathi
- 1.2.7. Tamil
- 1.2.8. Telugu

2. Method

The work is carried out by the three R&D teams jointly by engaging startups from the beginning of the project. The work is being done in a consortium mode in collaboration with the following institutes (chief investigator and individual Project Investigators (PI) are also mentioned in the table).

Institute	**Role**	**Languages(for Translation)**
IIIT-Hyderabad	Lead Institute	Hindi, Marathi, Kannada, Telugu
C-DAC , Noida	Member	Gujarati, Bangla, Hindi
AUKBC, Chennai	**Member**	**Tamil, Malayalam**

3. Aim

This project aims at maintaining high quality for all three tasks by making the process efficient and consistent at each individual step by conveying exact meaning as it is spoken in the english with consistent translation (with domain terms) in eight languages.

1. Ensuring the good quality throughout the process
2. Providing three translation options for domain terms
3. Reducing human efforts by ⅓ at the end of the project thus making future translation faster.

4. Project Pipeline

4.1. Transcription

Task is to transcribe spoken content as it is being spoken out. The process of transcription turned out as follows.

- Use ASR (automatic speech recognition) system for speech to text

- Manual validation and gap filling for the wrong content removal and to add missed out content
- Manual validation to fill the Technical gap

4.2. Translation

For this task, our main objective is to provide translation into indian languages without losing the conveyed meaning and to provide domain terminology into three options (Target language translated term, Transliterated term , English term) for each language as translation.

It is very difficult for the automatic MT systems to translate technical content and convey the same meaning as it is spoken in English. Therefore the role of human translator and validator becomes more critical here. The turned out process is as follows for the translation task.

- **Automatic Translation using MT systems**
- **Domain Term Marking (with three options)**
- **Post - Edit for the MT output**
- **Translation Validation**
    - **To check faithfulness with the spoken content**
    - **To check the fluency of the translated text (so the text does not sound jarring)**

4.3. Translation to SRT

In this, we perform the subtitling of the translated text for each video in each Language.

4.4. Validation at each stage

The overall task can be viewed as the below figure. All the above tasks involve pre-processing, post-processing **and verification at each stage for each language**. Annotation and pre-processing tasks are being carried out by external vendors/companies, and a thorough detailed verification of each tasks is done by the in-house team/s.

5. Targets Achieved

- Total Courses Translated and SRT Generated: 82
    - Total PG courses: 27
    - Total UG courses: 55
- **Total Hours of Lectures : 1300 Hrs**

From Domains :

- Biology, Computer
- Science, Human
- rights, Justice,
- Chemistry,
- Mathematics,
- Economics etc

Total number of manpower trained in this project for Translation post editing: 115