INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Technology Pipeline for Large Scale Cross-Lingual Dubbing of Lecture Videos into Multiple Indian Languages

*Anusha P[1], Arun K[2], Ashish S[2], Bhagyashree M[1], Ishika G[1], Jom K[1], Jordan F[2], K V Vikram[1], Mano R Kumar M[1], Metilda S M[2], M Wajahat[2], Mohana N[1], Mudit Batra[2], Navina K[1], Nihal G[1], Nithya R[2], Pruthwik M[3], Sudhanshu S[1], Vasista L[2], Vandan M[3], Vineeth K[1], Vrundha S[2], Dipti Mishra[3], Hema A Murthy[1], Pushpak Bhattacharya[4], Srinivasan Umesh[2], Rajeev Sangal[3]*

[1] SMT Lab, Department of Computer Science and Engineering, IIT Madras, India
[2] Speech Lab, Department of Electrical Engineering, IIT Madras, India
[3] IIIT Hyderabad, India [4] IIT Bombay, India

hema@cse.iitm.ac.in, jom@cse.iitm.ac.in

## Abstract

Cross-lingual dubbing of lecture videos requires the transcription of the original audio, correction and removal of disfluencies, domain term discovery, text-to-text translation into the target language, text-to-speech synthesis followed by isochronous lipsyncing to the original video. This task becomes challenging when the source and target languages belong to different language families, resulting in differences in generated audio duration. This is further compounded by the original speaker's rhythm, especially for extempore speech. This paper describes the challenges in regenerating English lecture videos in Indian languages semi-automatically. A prototype is developed for dubbing lectures into Indian languages. A demo for dubbing with supervision is available online[1].

**Index Terms**: Lecture transcreation, lipsyncing, video to video translation.

## 1. Introduction

A cross-lingual dubbing system typically involves four modules – (1) An automatic speech recogniser (ASR) to transcribe the English lectures. (2) A machine translation (MT) system translates the transcribed text to a target language. (3) A text-to-speech (TTS) system synthesises the translated text. (4) A lip-syncing module to synchronise the TTS audio with the original video. The modules involved in dubbing lecture videos are shown in Figure 1. Each module is prone to errors. Each of the modules leads to various challenges. We now briefly list the challenges and later discuss novel ways in which they are addressed.

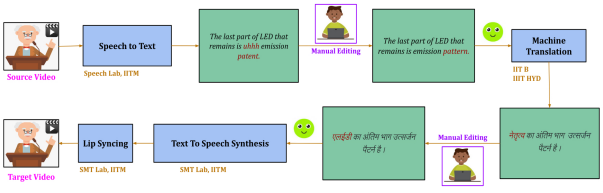


Figure 1: *Video to video pipeline for dubbing of technical lectures*

Thanks to MeiTY Govt of India for the project: 11(1)/2022-HCC(TDIL)

[1]Video to video translation and dubbing online demo page: https://asr.iitm.ac.in/demo/

## 2. Challenges in transcreation

### 2.1. Automatic Speech Recognition (ASR) of lecture videos

State-of-the-art ASR systems work well for general Indian English but fail for technical lectures. Technical lectures are more or less conversational and are replete with esoteric terms. Moreover, the sentences are often incomplete or ill-formed. This is compounded by the fact that Indians are bilingual and tend to switch between English and their native tongue. The ASR outputs are often time-stamped using voice-activity detection (VAD) to produce SRTs (sub-rip-text), which are used for subtitling the videos. The SRTs are seldom grammatically correct, which makes the task of translation difficult.

### 2.2. What to translate? What to transliterate?

While most speech recognition systems are able to transcribe verbatim the content spoken by the speaker, the text is quite unreadable owing to the presence of disfluencies. ASR outputs also often lack proper punctuation. ASR outputs are post-processed to remove disfluencies and summarised to make meaningful sentences for machine translation. Many of the technical words in these lectures need to be transliterated rather than translated, owing to the nonavailability of these terms in the vernacular, or the terms are rather esoteric. This requires the identification of domain terms in the source. Occasionally, even non-domain terms must be kept as is.

### 2.3. Isochronous lip syncing

Word order differences between English and Indian languages, and the length of translated text lead to a mismatch in the duration of the original audio and the synthesised audio. The audio and video need to be synced back so that the audio-visual experience of the lecture is preserved even in the target language.

## 3. Automatic speech recognition

In this work, NPTEL (nptel.ac.in) lecture videos for which manual transcriptions are available are used to train the speech recognition systems from scratch. A state-of-the-art conformer model is trained with 12 encoder and 6 decoder layers using 5000 byte-pair encodings (BPEs) as targets. Since Indian English mannerisms are similar, the English text obtained is more or less verbatim despite significant differences in accents. The word error rate (WER) of the transcription is quite good $< 10\%$. While the transcription is accurate, the text needs to be post-processed for disfluency removal and chunked appro-

priately so that the text translation is accurate.

## 4. Domain term detection

Technical lectures have a large number of mathematical symbols, equations and abbreviations. Handling such symbols without any loss of information is a challenging task in the MT and TTS stages of the V2V pipeline, even if the ASR is able to predict the text/symbols correctly. The domain terms need to be discovered for each technical domain separately for each topic. The term "Python" can have different meanings in Computer Science and life sciences. Identification of such terms based on relevant domain concepts for a given document is a highly challenging task. These terms usually lead to different vocabulary selections in machine translation. As a special case to the domain term, we also include mathematical symbols, variables and equations as technical expressions for the identification. The automatic system identifies domain terms and their domains by using both unsupervised (i.e. TextRank[1] and TFIDF[2]) and supervised (Token classification using ALBERT[3]) methods built on inhouse domain dictionaries and TermTraction corpora[2]. An F1-score of 60% was obtained for domain term identification for a finite set of domains. The dictionary is continuously evaluated and augmented.

## 5. Machine translation for the technical domain

Terminology integrated specialised machine translation system for translating English lecture text to Indian languages is used. These Machine Translation systems dynamically translate or do not translate specific domain terms/ expressions based on Domain Term Identifier indication. It uses symbolic placeholders around domain terms as an indicator to the machine translation system to treat them specially and perform one of the actions of translate, transliterate or keep as it is (do not translate). For English to Indian Language Machine Translation, we first translate from English to Hindi and Telugu by ensuring that the semantic information of the source text is preserved. Hindi and Telugu were chosen since they belong to two of the prominent language families in India, namely Aryan and Dravidian, and both have considerable representation in terms of language resources compared to other Indian Languages. Most Aryan languages share similar linguistic features, for example, word order, vocabulary, aspiration, schwa deletion, etc., even though they do not share a script. Similarly, Dravidian languages are agglutinative and also share properties. Therefore to leverage this language typology, we use machine translation systems from Hindi to other considered Indo-Aryan languages and Telugu to other considered Dravidian languages for better automatic translation.

## 6. Text to speech synthesis in the target language and Isochronous lip syncing

The translated text is synthesised using Fastspeech2 model and HiFiGAN vocoder. The video and TTS audio durations are significantly mismatched. Since these are technical lectures, clarity in the audio is quintessential. Further, the lecturer does a lot of board work or uses view graphs. Therefore, we interpolate the video (also timestamped at the original SRT) to match the audio duration of the target language rather than the audio to the video duration. Since the matching is performed at the syllable-level, and syllable-rate and viseme [4] rates are more or less equivalent, the quality of the video generated is more or less natural. Since the word order is similar intra-language family with perhaps some variability in audio duration, the same algorithm is used for other language videos.

## 7. Online Video Translation and Dubbing platform

An online tool for multi-lingual translation and dubbing is developed, incorporating the semi-supervised approach to video-to-video translation with manual intervention. The screen grab of the tool is shown in Figure 2.

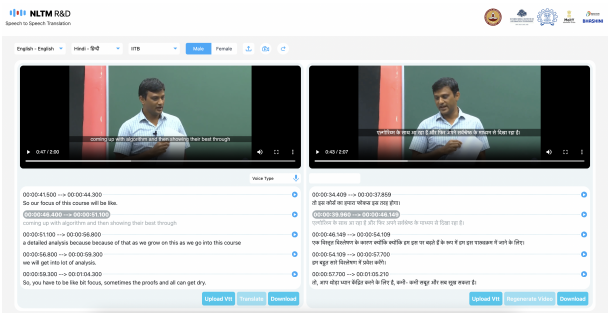


Figure 2: *Online video dubbing demo page (`https://asr.iitm.ac.in/demo/`) screenshot.*

## 8. Conclusions

Informal evaluations suggest that the dubbed videos do give an audio-visual experience in the native tongue. The prosody and timbre of the original speaker in terms of intonation, emphasis, and mannerisms are still missing. Conversational speech voice conversion is still nascent [5]. Prosodic modification based on domain terms may be explored to improve the quality of the videos generated. About 70 videos have been generated in different Indian languages [3].

## 9. References


[1] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.

[2] T. Joachims, "A probabilistic analysis of the rocchio algorithm with tfidf for text categorization." Carnegie-mellon univ pittsburgh pa dept of computer science, Tech. Rep., 1996.

[3] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=H1eA7AEtvS

[4] P. H, K. C, T. G, and G. J, "Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility," *Elife*, 2016.

[5] B. Mukherjeee, A. Prakash, and H. A. Murthy, "Analysis of conversational speech with application to voice adaptation," in *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*. IEEE, 2021, pp. 765–772. [Online]. Available: https://doi.org/10.1109/ASRU51503.2021.9688146


[2] `https://ssmt.iiit.ac.in/TermTraction`

[3] Links to videos generated:
`https://drive.google.com/drive/folders/1hByrzH_lSceDaOQ71pWLqj6WevxaOxTz`