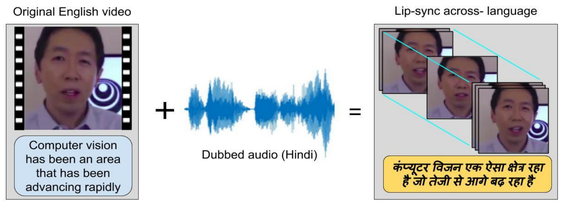
Cross-Language Speech Dependent Lip-Synchronization

Abhishek Jha [1] Vikram Voleti [2] Vinay Namboodiri [3] C. V. Jawahar [1]

[1] CVIT, KCIS, IIIT Hyderabad, India [2] Mila, Université de Montréal, Canada [3] Department of Computer Science, IIT Kanpur, India

Motivation

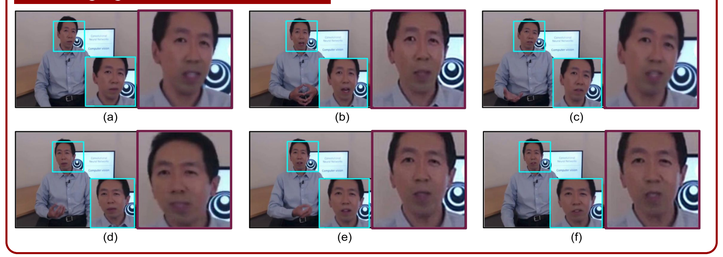
Goal: Given a video in foreign language, and dubbed speech in regional language, synchronize the lips in video to match the dubbed audio.



- Exponential growth in internet user in the last few years.
- A majority of these users are college/school-goers and young adults.
- Diverse demographic of students all over the world enroll in MOOCs

Challenges: Audio to lip fiducial generation, cross-language lip synchronization, cross identity lip synchronization.

Cross-language results



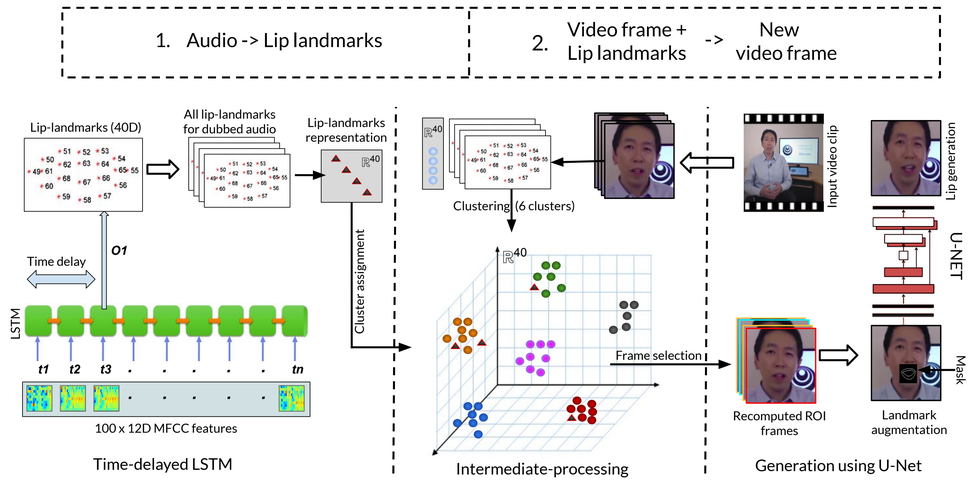
Contributions

- Developed a cross-language lip-synchronization model for dubbed speech videos (e.g. Hindi dubbing for English videos).
- Developed an automated pipeline to curate a dataset to train the proposed model.
- Conducted a user-based study, which shows learners prefer lip-sync for dubbed videos.

Email: abhishek.jha@research.iiit.ac.in
vikram.voleti@gmail.com

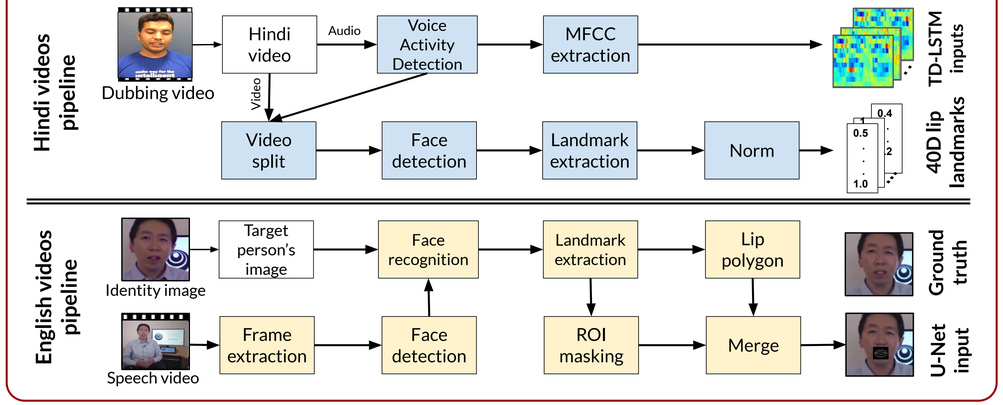
Project page: preon.iiit.ac.in/~abhishek_jha/lipsync_inst_vid

Approach



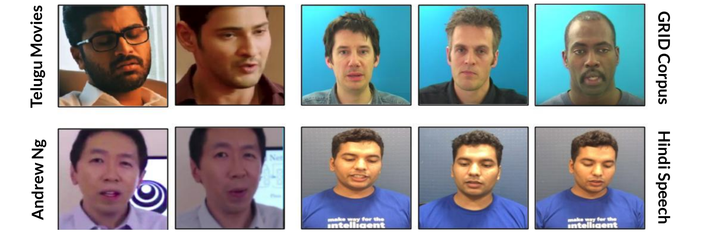
- **Audio -> Lip landmarks** : Train Time-delayed (TD) LSTM on grid corpus and Hindi speech corpus.
- **Intermediate processing:** To search for the best face to append with generated lip landmarks.
- **Face + Lip landmarks-> New face** : Train U-Net on videos from movie dataset.
- **Inference:**
  - ➢ Generate lip-landmarks from audio using TD-LSTM,
  - ➢ Generate new frames by modifying original Andrew Ng video frames using U-Net

Dataset curation - automated pipeline



Datasets

- **Hindi Speech corpus:** 2.5 hours of audio visual speech
- **Andrew Ng ML videos:** 16000 image frames extracted from deeplearning.ai MOOC videos.
- **GRID Corpus:** 33 speakers, 1000 phrase each (51 words)
- **Telugu movies**: scenes with protagonist's face, extracted from Telugu movies

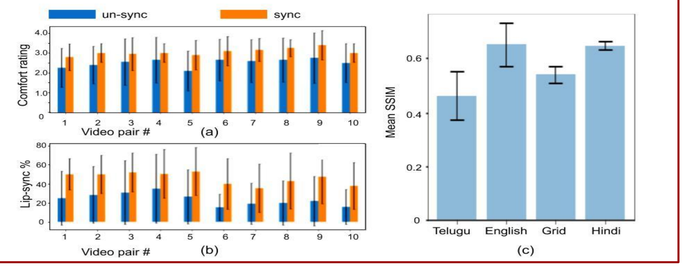


Results

User-based study

	C - US	C - S	LS% - US	LS% - S
Mean	2.51	**3.1**	23.86	**45.95**
Std. dev.	1.07	**0.6**	25.9	**24.1**

- **C:** Comfort level
- **US:** Un-synced
- **S:** Lip-synced
- **LS%:** Lip-sync percentage



U-Net results