Cross-Language Speech Dependent Lip-Synchronization

Abhishek Jha 1  Vikram Voleti2  Vinay Namboodiri 3  C. V. Jawahar 1

1CVIT, KCIS, IIIT Hyderabad, India  2Mila, Université de Montréal, Canada  3Department 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.

Approach

1. Audio -> Lip landmarks
2. Video frame + Lip landmarks -> New video frame

❖ 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

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.

Dataset curation - automated pipeline

Hindi videos pipeline

Video split
Voice Activity Detection
MFCC extraction
Norm

Hindi video
Dubbing video

English videos pipeline

Identity image
Frame extraction
Face detection
ROI masking

English video
Speech video

Results

User-based study

<table>
<thead>
<tr>
<th></th>
<th>C - US</th>
<th>C - S</th>
<th>LS% - US</th>
<th>LS% - S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.51</td>
<td>3.1</td>
<td>23.86</td>
<td>45.95</td>
</tr>
<tr>
<td>Std. dev.</td>
<td>1.07</td>
<td>0.6</td>
<td>25.9</td>
<td>24.1</td>
</tr>
</tbody>
</table>

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

Datasets

❖ Hindi Speech corpus: 2.5 hours of audio visual speech
❖ Andrew Ng ML videos: 16000 image frames extracted from deeplearning.ai MOOC videos.
❖ Telugu movies: scenes with protagonist’s face, extracted from Telugu movies

Email: abhishek_jha@research.iiit.ac.in
vikram.voleti@gmail.com
Project page: preon.iiit.ac.in/~abhishek_jha/lipsync_inst_vid