Lip-Synchronization for Dubbed Instructional Videos

Abhishek Jha*[1] Vikram Voleti*[1] Vinay P. Namboodiri[2] C. V. Jawahar[1]
[1] Center for Visual Information Technology, KCIS, IIIT Hyderabad, India
[2] Department of Computer Science and Engineering, IIT Kanpur, India
{abhishek.jha@research, jawahar@}.iiit.ac.in, vikram.voleti@gmail.com, vinaypn@iitk.ac.in

Abstract

Online instructional video lectures such as MOOCs are often limited by linguistic constraint of different demographies. Students from backgrounds that are non-native to the accent or language of the instructor often find it difficult to comprehend the full lecture, which leads to lower retention rates of the courses. Simple audio dubbing in the accent or language of the student makes the video appear unnatural.

In this paper, we propose two lip synchronization methods — one for audio dubbed in the non-native accent of the student, and another with audio in the foreign language of the student. We describe an automated pipeline to synchronize the lip movements of the instructor with the audio in both cases. With the help of a user-based study, we verify that our method is is preferred over unsynchronized videos.

1. Introduction

Online instructional videos, especially Massive Open Online Courses (MOOCs), are prime examples of how education can help skill development beyond the boundaries of conventional classrooms. Yet the retention rates in these courses can be as low as 10%. One of the major reasons for this is a cultural gap between the linguistics of the student and the instructor. Students from different parts of the world often find it difficult to understand the accent and language of the instructors, owing to their non-familiarity with it. This results in slow learning curves as well as dropouts from such online courses. Subtitles in different languages do not lend enough help since they divert the attention of the student. A quick-fix solution to this would be to dub instructional videos in the accent or language of the student. However, dubbing without lip synchronization makes the video appear unnatural.

In this paper, we propose 'Visual Dubbing' for synchronizing lip motion in instructional videos according to the language it is dubbed in. Our main ideas and contributions are two-fold: 1) we propose an English-to-non-Native-English approach to dubbing online educational tutorial videos originally in English to a non-native English accent, such as Indian accent or French accent; 2) we propose an English-to-Foreign-Language approach to dubbing videos such that the lip movements warp to match the audio in the new language. Lastly we show how the generated lip-motion or 'Visual Dubbing' makes the instructional video more engaging based on a used-based-study.

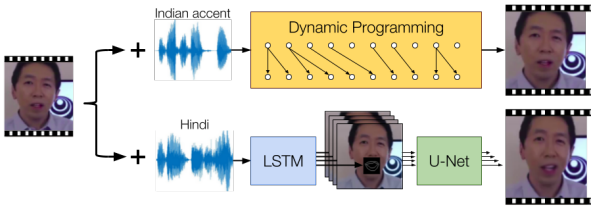

Figure 1: (top) Dynamic Programming to non-native English accent, (bottom) Visual Dubbing to other language

Some of the recent work in this area focuses on synthesizing photo-realistic lip motions and facial expressions. Face2Face [4] morphs the facial landmarks of a person based on those of another actor. But it requires a human in the loop which can be quite expensive and erroneous.

Most similar to our work are [3, 2] which use speech audio represented as MFCC features [3] and text [2] to train an LSTM to produce a sequence of lip landmark points. The lip landmarks are then used to generate mouth texture. Finally this mouth texture is merged with the face in the original frame. Our work is different from [3, 2] in that our method synchronizes lip motion across two different languages, in contrast to just English-to-English. Hence, our challenges include learning higher-level viseme-phonemic relations across two different languages.

2. Method

Instructional videos provide a controlled framework for this problem, since the speakers usually speak scripted dialogues in good lighting facing the camera. The challenge is to model the lip movements given the dubbed audio, and generate new lip movements for the same speaker.

*these authors contributed equally to this work.

2.1. English to non-Native English

Given the original instructional video in native English (like Andrew Ng's machine learning tutorials), and audio of the same dialogues in a non-native English accent (like French or Indian), we use Dynamic Programming [1] to create a dynamic map between the MFCC features of the original audio (speaker's English) and the target audio (listener's non-native English). This is illustrated in Figure 1, where every frame in the target is non-linearly mapped to the appropriate frame in the original video. Using this mapping, we render a new video to match the dubbed audio. This not only makes it easier for a student to comprehend the lecture in their non-native accent, the mental fatigue of looking at an unsynchronized video is avoided.

2.2. English to Foreign Language

Major challenges in lip-syncing audio of a foreign language (Hindi) on video of original language (English) are the differences in their grammatical structure, and set of phonemes. To solve these, we first learn a mapping between Hindi audio and lip landmarks. From the predicted lip landmarks, we synthesize mouth regions over the original English video to match the Hindi audio. This entire pipeline can be seen in Figure 1 (bottom).

To learn audio-to-lip-landmarks mapping, we train a time-delayed LSTM on MFCC features of Hindi audio. We curated a dataset of 2.5 hours of Hindi speech consisting of a speaker reciting Hindi news articles and stories. We input MFCC features to the LSTM, and predict the ground truth lip landmarks. To learn lip-landmarks-to-mouth mapping, we train a U-Net separately on 16,000 frames of Andrew Ng from his deeplearning.ai tutorials, similar to ObamaNet [2]. We replace the mouth region in each frame with a polygon connecting the lip landmarks. Conditioning on this image, we train the U-Net to generate the original frame.

3. Results

3.1. English to non-Native English

To show the implementation of our proposed method, we collect 15 videos clips of upto 1 minute each from publicly available Machine Learning tutorials of Andrew Ng. Since the lip-motion synchronization's effect on mental fatigue could be dependent on the listener, we conduct a user-based study to evaluate our system. We select 5 subjects familiar with the content in the instructional videos, and 5 others who are new to the content. To each of these subjects, we show 5 randomly selected video pairs of un-synced dubbed video and our dynamically synced videos (Figure 2, top) in non-native English accent. We then ask them to rank each video between 1 (Hard to understand) to 5 (Easy to understand). As shown in Table 1 row 1, the mean score was higher for our dynamically synced version.

Figure 2: Results of (top) Dynamic Programming to non-native English accent, (bottom) Visual Dubbing to other language.

3.2. English to Hindi

We perform the same user-based experiment with Hindi lip-synchronized videos: we show the same video with Hindi audio naively overlaid (un-synced), and Hindi visually-dubbed (lip synchronized) (Figure 2, bottom). As shown in Table 1 row 2, the mean score is higher for the visually-dubbed version.

	US(N)	S(N)	US(F)	S(F)
Dynamic	3.0	**4.6**	1.9	**3.2**
VDub	2.5	**3.5**	1.5	**3.5**

Table 1: Mean scores for Dynamic Programming (Dynamic) on Indian-English and Visual Dubbing (VDub) on Hindi: for un-synced speech overlay ('US'), and lip-synced version ('S'), by naive (N) and familiar (F) listeners.

4. Discussions

In this work, we assume the availability of dubbed audio, which can be automated using Machine translation (MT) systems and text-to-speech (TTS) synthesizers. However imperfections in MT translations and lack of personality in the TTS-synthesized speech could make them unsuitable for instructional videos. Furthermore, handling multiple speakers, extreme head poses, and robust key point tracking present future scope of improvement. Lastly, we believe this work can help expand the reach of instructional videos across diverse linguistic groups.

References

[1] T. H. Cormen. *Introduction to algorithms*. 2009. 2

[2] R. Kumar, J. Sotelo, K. Kumar, A. de Brébisson, and Y. Bengio. Obamanet: Photo-realistic lip-sync from text. 1, 2

[3] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM TOG*, 36(4), 2017. 1

[4] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016. 1