Audio-Visual Speech Recognition and Synthesis

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering
by Research

by

Abhishek Jha
20162310
abhishek.jha@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
April 2019

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**Audio-Visual Speech Recognition and Synthesis**" by **Abhishek Jha**, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Adviser: Prof. C. V. Jawahar
Date	Adviser: Dr. Vinay P. Namboodiri

To ‘Maa’ and ‘Papa’

Acknowledgments

It fills me with immense pleasure and joy to write this section acknowledging the people without whom, this thesis would not have been possible. First and foremost I would like to thank my advisors Dr. Vinay P. Namboodiri and Prof. C. V. Jawahar, for their continuous guidance and support through out my MS studies. They have been a constant source of inspiration and encouragement, throughout this journey.

I thank the anonymous reviewers who took their time to give quality feedback on my research papers. I also thank the thesis committee for their careful evaluation of this thesis and providing me with valuable comments.

I cannot be grateful enough to CVIT lab for providing me with a nurturing environment to grow as a competent researcher. I sincerely thank Yashaswi Sir, Anand Sir, Praveen, Aditya, Aniket and Pritish for the stimulating discussions, reviewing papers and sharing their research experiences with me. I owe special thanks to Tejaswi for her continuous encouragement throughout my years of MS studies and through the process of researching and writing this thesis. I would also like to thank my friends Sahil, Vamsi, Harish for all the fun we had in last two years.

Finally, I must express my very profound gratitude to my parents and sister for their patience, encouragement and unfailing support.

Abstract

Understanding speech in the absence of audio, from the visual perception of lip-motion can aid a variety of computer vision applications. System comprehending 'silent speech presents a promising potential for low bandwidth video-calling, speech transmission in auditory noisy environment to aid for hearing impaired. While presenting numerous opportunities, it is highly difficult to model lips in silent speech video by observing lip-motion of speaker. Albeit developments in automatic-speech recognition (ASR) has yielded better audio-speech recognition systems in last two decades, in the presence of noise their performance drastically deteriorates. This calls for a computer vision solution to the speech understanding problem. In this thesis, we present two solutions for modelling lips in silent speech videos.

In the first part of the thesis, we propose a word-spotting solution for searching spoken keywords in silent lip-videos. In this work on visual speech recognition our contributions are twofold: we develop a pipeline for recognition-free retrieval, and show its performance against recognition-based retrieval on a large-scale dataset and another set of out-of-vocabulary words. 2) We introduce a query expansion technique using pseudo-relevant feedback and propose a novel re-ranking method based on maximizing the correlation between spatio-temporal landmarks of the query and the top retrieval candidates. The proposed pipeline improves baseline performance by over 35% for word-spotting task on one of the largest lipreading corpus. We demonstrate the robustness of our method through a series of experiments, by investigating domain invariance, out-of-vocabulary prediction and careful analysis of results on dataset. We also present the qualitative results showing success and failure cases. We finally show the application of our method by spotting words in an archaic speech video.

In the second part of our work, we propose a lip-synchronization solution for 'visually redubbing speech videos in a target language. Current methods of adapting a native speech video in foreign language either through placement of subtitle in the video, which distracts the viewer or through audio redubbing the video in the target language. This causes unsynchronized lip-motion of the speaker with respect to the redubbed audio, resulting in video appearing unnatural. In this work, we propose two lip synchronization methods: 1) cross-accent lip-synchronization for change in accent of the same language audio dubbing, and 2) cross-language lip-synchronization for speech videos dubbed in a different language. Since viseme remains the same in cross-accent dubbing, we propose a dynamic programing algorithm to align the visual speech from the original video with the accented speech in the target audio. In cross-language dubbing overall linguistics changes, hence we propose a lip-synthesis model conditioned upon on the redubbed audio. Finally, a user-based study is conducted, which validates our claim of better viewing experience in comparison to baseline methods. We present the application of

both these methods by ‘visually redubbing Andrew Ngs machine learning tutorial video clips in Indian accented English and Hindi language respectively.

In the final part of this thesis, we propose an improved method of 2D lip-landmark localization method. We investigated the current landmark localization techniques in facial domain and human-pose estimation to discover the shortcoming in adapting these methods for the task of lip-landmark localization. Present state-of-the-art methods in the domain considers lip-landmarks as a subset of facial landmarks and hence doesn't explicitly optimizes for it. In this work we propose a new lip-centric loss formulation on the existing stacked-hourglass architecture which improves the baseline performance. Finally we use 300W and 300VW faces dataset to show the performance of our methods and compare them with the baselines.

Overall, in this thesis we examined the current methods of lip modelling, investigated them for their shortcomings and proposed solutions to overcome those challenges. We perform detailed studies, and ablation studies to study our proposed methods and reported both success and failure cases for the same. We compare our solutions with the current baseline on challenging datasets, reporting quantitative results and demonstrating qualitative performances. Our proposed solutions improves the baseline performances in their individual domains.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Internet, in the last two decades, has resulted in unprecedented changes in the way humans share information. Advancement in semiconductor technology, with more affordable visual sensors and faster signal processors, have been at the forefront of utilizing this internet revolution. This has resulted in an unprecedented data boom, where the end user itself is both content creator and consumer. A large portion of this visual data comprises of speech videos. Speech videos come in various forms like instructional videos, music videos, news, talk shows, video calls and many more forms which contain an entity speaking in front of the camera. With the availability of such humongous amount of speech data comes the challenge of developing systems which can efficiently, index, organize, understand and retrieve it. While there are systems available for efficient management of text documents and still images, videos system for such caliber are still in nascent stage.

Modern search engines often index videos through the present mata-data. While annotating videos with such meta-data is an expensive practice, it also fails to provide adequate information of the video. Speech video on the other hand intrinsically contains spoken speech in the form of audio, describing the content of the video. Audio speech hence can be a linguistically rich semantic method for describing the video, and may be independently used to understand speech videos.

In the case of noise in the channel, this audio signal might get corrupted compromising quality of the video for speech understanding task. This demands a solution that can leverage visual modality as well.

In this thesis we present a set of solutions which models the visual modality of speech video for language understanding task. Our proposed solutions tries to solve challenges pertaining to understanding the linguistics of a speech video in the absence of audio modality, i.e. silent speech video. Our first work gives solution for spotting words in silent speech videos. In our second work, we show the lip-motion reconstruction conditioned upon the audio, thereby morphing the lips of the speaker based on an audio clip in a different language. Finally we show ways to improve the fine-grained lip-landmark localization. In all the three tasks our main objective is to model the visual speech in the absence of audio, which can host a range of application from lipreading to re-dubbing of visual of a movie in a different language.

1.1 Motivation

Audio as a form of communication is not exclusive to human, in fact many birds, animals and insects use some form of audio communication to interact with each other. These communications are not as complex as humans. Humans communicate through a much more complex system of rules, called language. Language being one of the most important invention in human kind, allowed humans to not only communicate with each other regarding physical objects but also facilitated the conception of abstract concepts. Expressing these diverse set of concepts required us to move from simple gestures and sounds to a richer system of speech, verbal or spoken speech.

This new system consists of complex rules governing the placement of each phonemes, one of the units of sound that distinguishes one from another, to form a meaningful sentences. This sound production occurs when air flows through vocal tract situated in the neck region, as shown in Figure 1.1 . The air interacts with various muscles in vocal tract and mouth to produce different frequencies leading to voice generation such as talking, singing, laughing, crying, screaming, etc.

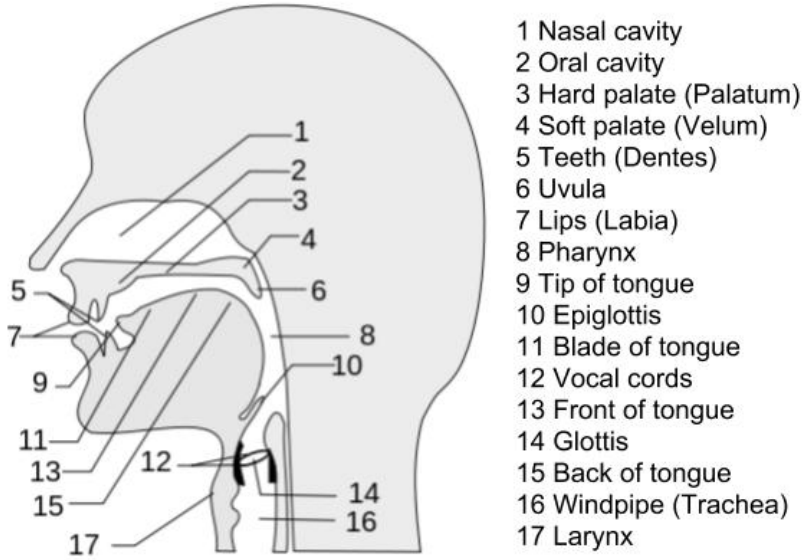

Figure 1.1 Anatomy of human vocal tract [67]: The diagram shows the sagittal section of human vocal tract labelled with different constituting parts on the right.[2]

In English language, there are 44 phoneme, with 24 consonant phonemes and 20 vowel phonemes. Each of these phonemes is uttered by certain orientations of different organs in vocal tract. Moreover number of phonemes may also change with different dialect of the same language. In Table 1.1 we show, 44 phonemes corresponding to English language, along with their International Phonetic Alphabet (IPA) denoting the phoneme, and few example words containing the particular phoneme.

[2]Image source: `https://commons.wikimedia.org/wiki/File:VocalTract_withNumbers.svg` This image is licensed under the Creative Commons Attribution 3.0 Unported license.

Consonants		
Phoneme	**IPA Symbol**	**Examples**
1	b	bulb, bulge
2	d	dam, milled
3	f	phone, often
4	g	guest, egg
5	h	hop, who
6	dʒ	sledge, exaggerate
7	k	kit, cat
8	l	lip, shell
9	m	manner, combination
10	n	know, pneumonic
11	p	pin, peacock
12	r	carrot, rhyme
13	s	citric, psycho
14	t	matter, train
15	v	vine, five
16	w	when, choir
17	z	zen, scissors
18	ʒ	measure, division
19	tʃ	culture, righteous
20	ʃ	shirt, conscience
21	θ	thought
22	ð	feather
23	ŋ	ring, tongue
24	j	uniform, onion

Vowels		
Phoneme	**IPA Symbol**	**Examples**
25	æ	cat, plaid
26	eɪ	bay, foyer
27	e	lend, many
28	i:	key, people
29	ɪ	england, women
30	aɪ	sky, island
31	ɒ	honest, maul
32	oʊ	open, coat
33	ʊ	could, wolf
34	ʌ	monkey, blood
35	u:	group, who
36	ɔɪ	join, oil
37	aʊ	how, shout
38	ə	about, honour
39	eə^{r}	chair, prayer
40	ɑ:	arm
41	ɜ:^{r}	bird, journey
42	ɔ:	war, paw
43	ɪə^{r}	ear, tier
44	ʊə^{r}	sure, tourist
-	-	-
-	-	-
-	-	-
-	-	-

Table 1.1 List of phonemes: (left) presents the list of consonant phonemes and their corresponding IPA symbols along-with two examples each. (right) presents the list of vowel phonemes and their corresponding IPA symbols along-with two examples each.

While phonemes constitute audio speech, the orientation of facial muscles while uttering a particular phoneme is called viseme. As many of the organs and muscles responsible for speech generation in human is inside the vocal tract all the way to inner mouth, different phonemes can have same viseme. For example: p and b have the same viseme in 'park' and 'bark'. This many to one relation between phonemes and visemes is one of the reason visually deciphering spoken speech is non-trivial. Figure 1.2 shows lip shape corresponding to different visemes.

The ability to perceive viseme into language is called lipreading. Lipreading can be a very important skill for hearing impaired people. Since lip-motions are fast during speaking, along with the uncertainty in phoneme viseme relation, lip-reading is non-trivial and often takes years of practice to achieve expertise.

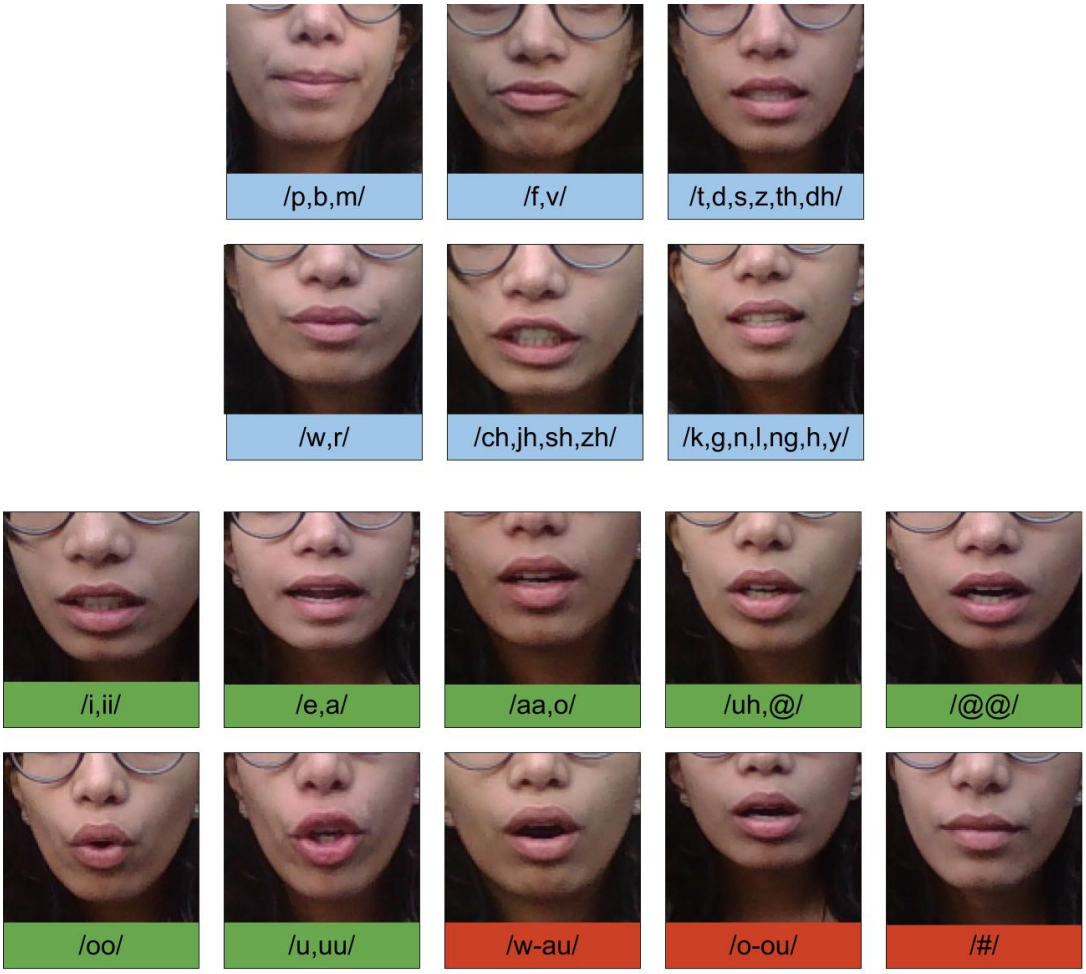

Figure 1.2 Visemes corresponding to different phonemes [26]: The figure shows lip shapes while speaking consonant visemes (with blue labels), monophthong visemes (with green labels) and diphthong visemes (with red labels).

Since words are structured entities of phonemes, probability of a particular viseme being present is dependent on its neighbouring visemes. This makes the problem of lipreading a sequence learning problem of spatio-temporal audio-visual signal.

In this thesis, we have investigated different methods of modelling audio-visual speech data for efficient lipreading task. We have also explored the application of sequence learning and generative models for synthesizing lip-motion, conditioned upon input speech audio.

1.2 Scope of the Thesis

1.2.1 Problem statements and challenges

We will be focusing our work on comprehending visual speech from audio-visual data, by modelling lip-motions. Particularly, we will discuss about three challenging problems in this domain:

Lipreading: The aim of this work is to predict the spoken word, given video of a person speaking, in the absence of audio. Video of person speaking, called speech video, contains person's face with lip-motion visible.

In natural environment, a speaker can move his head, changing the head-pose resulting in partially occluded lip region. Quality of the video play another important role in the lipreader's accuracy. Lipreading models, being recognition based models, are constraint by a fixed size vocabulary. Moreover, for similar sounding words, lipreading performance deteriorates [86]. Recurrent neural network (RNN) has emerged as a useful tool for modelling such sequential data, but training these models requires, large scale annotated datasets. With all these challenges, we find lipreading to be non-trivial and data intensive problem.

Lip-synchronization: The Internet in last 25 years has revolutionized the way content is being made and distributed. Every minute hundreds of hours of video data is uploaded on Youtube, across different parts of the world. Within these videos, Movies, TV shows and instructional videos are majorly watched across different linguistic and cultural demographic regions. Hence, it becomes necessary for the videos created in one language to be appended with audio in another target language for better penetration of video content across demography, through audio dubbing. As the redubbed video and audio contains visemes and phonemes in two different languages, it may lead to discomfort for viewers. Subtitles on the other hand, leads to distraction from the main content of the video. A possible solution for this problem can be lip-synchronization.

Lip-synchronization for person in the video with respect to the redubbed audio, requires generation of lip-motion for that particular person. Present systems allows the lip-synchronization across two different face poses, but such a system would require parallel lip-videos in two different languages for the same speech content. This requires a human in the loop, and therefore can be expensive. Hence, we aim at cross-language and cross-accent synchronization of lip-motion of the speaker in the speech video based on the redubbed audio.

Lip-landmark localization: Many applications of visual speech recognition require modelling of lip-motions. Hence, an accurate structural representation of lips can be vital for improving their performances. Lips being elastic in nature present deformable contours and change in skin tone near the edges. Moreover presence of facial hair and occlusion due to microphone or hand during conversation make the lip-landmark localization problem more challenging. Hence, we aim to develop solutions for fine-grained lip-landmark localization task.

1.2.2 Contributions

For the first problem, instead of classifying the given video clip as one of the words in the fixed set of vocabulary, we propose a novel pipeline for spotting word in silent speech videos. Given a silent speech video, we intend to localize the instances of a word spoken by another person, in the target video. We propose a recognition-free retrieval pipeline over the features learned through a recognition-based model, to perform this localization. This allows us to spot the query exemplar in the target video,

without explicitly recognizing the words, allowing us to overcome the fixed-vocabulary constraints and change in accent. Our contribution for work are as follows:

1. We develop a pipeline for recognition-free retrieval, and show its performance against recognition-based retrieval on a large-scale dataset and another set of out-of-vocabulary words.

2. We introduce a query expansion technique using pseudo-relevant feedback and propose a novel re-ranking method based on maximizing the correlation between spatio-temporal landmarks of the query and the top retrieval candidates.

Our proposed method improves the word spotting performance on a standard large scale lip video dataset — Lipreading in the wild (LRW) [17]. We also show domain invariance on another standard dataset GRID corpus [21] and show the application of our method by spotting words in the famous Charlie Chaplin: The great dictator speech video.

Second problem requires translating audio data from the redubbed audio into photo-realistic video of the speaker's lip complementing the phonemes in the target audio. We aim to translate English language video clip appended with Hindi language audio, into lip-synchronized (synced) Hindi language video. To address this problem, we created datasets of speech video clips containing instructional videos in English language and their Hindi translated counterpart with mouth of dubbing artist visible.

We first extract the lip-landmarks from English and Hindi language videos. We then train an RNN to map audio data into lip-landmark space. We also train a U-Net on English language dataset to reconstruct photo-realistic lip region of interest (ROI) from English speaker's lip-landmarks. During inference we use the predicted lip-landmarks as a prior for generating lips. Our main ideas and contributions are two-fold:

1. We propose a model for lip-syncing target video with audio dubbing in a different accent of the same language, such as Indian English accent or French English accent.

2. We propose another model to lip-sync the speech video based on the different language audio dubbing, for instance English video with Hindi audio dubbing.

3. We also propose a scalable pipeline for dataset creation, used to train our models.

We show the application of our method on Andrew Ng's Machine Learning tutorial videos [1], by lip-syncing it with a Hindi audio translation of the English dialogue-transcript. To examine the efficacy of our method we conducted human-based study.

Our third problem requires fine-grained prediction of key-points for lips in an image. Current methods of key-point localization for human pose-estimation uses stacked convolutional network to predict heatmaps corresponding to each key-point. We exploit this network architecture for fine-grained landmark detection task. Our contribution are as follows:

1. We propose a stack-hourglass model with a modified loss formulation to predict lip-landmarks.

2. We propose a novel metric, suitable for evaluating fine-grained lip-landmark localization task.

We compare the results of the proposed method against the baseline network using the proposed new metric. Our proposed network improves the localization performance for lip-landmarks on 300W dataset [84]. We also show the application of the method by locating key-point on the lips in 300VW dataset [85].

1.3 Background

Lipreading

Recent onset of Deep learning, and its application in lipreading domain has yielded some impressive performances over the classical methods of using HMMs and other sequence modeling methods. The basic step in those methods has been training an end-to-end deep network with a classifier on the top for classifying words. Petridis et al [78] trains an LSTM classifier on discrete cosine transform (DCT) and deep bottleneck features (DBF), and thereby predicting full words. Wand et al [97] uses an LSTM with HOG features to predict short phrases. Chung et al [17] uses 3D convolution with different feature fusion techniques to shows word level lipreading on a much larger dataset than previously used, Lipreading in Wild (LRW) dataset [17]. Recent work by Chung et al [19] uses a temporal CNN followed by a attention based LSTM to show character level prediction on LRW dataset and 'Lipreading sentences in the wild' (LRS) [19]. Another work by Stafylakis and Tzimiropoulos [86] shows state-of-the-art performance on LRW dataset using a residual network over a 3D convolutional layer followed by two parallel LSTM modules. The performances of these networks [86] are constraint by the vocabulary size of the dataset.

Lip-synchronization

Lip-synchronization is widely used in animation movies, where the dubbing artist's audio is required to be aligned to the lip-motion of the animated character. These methods often require a human in the loop to manually assign lip shapes to the target character. Some of the most initial work in the domain of lip-synchronization has been for animations [56, 98, 101]. Before deep learning HMMs [24] has been widely used for sequence learning problem and therefore has been used in the early works in the domain of speech modelling [100, 101]. Face2Face [92] originally developed for morphing facial expression between two subject uses lip-landmarks for mapping the expressions. These methods can be exploited to synchronize lips as well, but requires a human in the loop, for creating input feed of target lip-motions. This limits the applicability of Face2Face [92] for large scale lip-synchronization of speech videos. Our proposed generation method is only conditioned upon target language audio during inference.

The dawn of deep learning gave way to better sequence learning methods. RNNs particularly LSTMs [46] have become de facto method for sequence learning problems. Similarly CNNs can be used to learn representation of static two dimensional images and three dimensional objects. Hence, CNNs followed by RNNs has been widely used for modelling spatio-temporal signals like videos. Many recent works

have exploited these kind of networks for facial expression generation [80, 57]. Some of the similar work to our employs end-to-end conditional encoder-decoder architecture for lip-synthesis from audio [20], and two-step RNN and deep generator networks for lip-synthesis conditioned upon either audio [91] or text[64]. Our method is different from them in the sense that we try to generate lip from audio dubbed by another person in a different language.

Active speaker detection

One of the crucial step for modelling visual speech data is creation of speech video dataset with lip ROI visible. Here active speaker detection methods are employed to find who is the speaker among a number of possible candidates in a frame for the associated speech audio. Hu *et al* [50] uses a cascade of features based on: mean-square-distance (MSD) of all the lip-fiducial across subsequent frames, centered face of candidate speakers, length of speech face tracklets and length of subtitles, and audio-visual synchronization of lip-motion and speech audio for active localization. Hong *et al* [49] Haar based lip detector and uses MSD across frames to detect active speakers among other candidates. Moaci *et al* [72] uses peaks in audio waveform and the motion is the mouth region to compute synchronization score by taking the scaler product between audio and visual features to localize active speaker. Izadinia *et al* [52] proposed a cross-modal approach where they projected first derivative MFCC features of audio and motion features of individual frames into a joint embedding, using CCA [8], to compute the correlation between the two modalities and thereby detecting the sound source.

In [81], a multi-modal variant of LSTM [34] was proposed to capture both correlation between speech audio and lip video, and the temporal relationship within each modality. This cross-modal temporal correlation between the two modality has been used to localize the active speaker.

In one of the recent work by Chung and Zisserman [18], used a siamese [47] like network having two parallel convolutional network: one with lip videos of all the candidates as input modality and another with MFCC of the speech audio as input modality. Both of these modality had been projected into a joint embedding by minimizing the contrastive loss [41] between them. The nearest candidate to the speech audio in the joint space is assigned as active speaker. We use a similar method to that of [18], ours being different in the sense that domain is movie videos. A movie video may contain video segments with thematic music in the background and simultaneous speech . Moreover, any video segment in the movie can contain zero to multiple speaking candidates with speech audio in the background. Hence our requirement is a robust active speaker localization system which can address these challenges.

1.3.1 Datasets

Deep neural networks requires large amount of annotated data to learn complex functions. Lipreading and lip-synthesis require the network to learn high-level semantic association between the linguistics and the visual ques of the audio-visual data. Hence, a scalable solutions in this space intrinsically demands large-scale lip datasets. In this section, we will discuss about the major datasets available in the space of lipreading.

Grid Corpus:

It [21] consists of 1000 phrases, spoken by each of the 33 speakers, resulting in a total of 33k utterances. The speakers set contains both male and female speakers. Each video sample is of length about 3 second at 25 frames per second. Each annotation sentence/phrase contains 6 words in fixed syntactic order: command(4) + color(4) + preposition(4) + letter(25) + digit(10)+ adverb(4). Few samples also contains annotation for in between silence. Annotations are being provided at word level. Some of the examples of samples annotations from grid corpus: 'put red at G 9 now', 'set white with p2 soon', 'Place red in a zero now', 'Place red at J 2, please'.

Lipreading in Wild:

This dataset [17] contains 500 word vocabulary, with training set containing 1000 samples each for every word whereas validation and test set containing 50 samples each for every word. This results in a total of around 550K utterances. Each video clip is 116 second long at 25 frames per second, with actual word spoken around the center of the video clip. The dataset has been curated from BBC news videos, and hence contains both male and female speakers of different ethnicities. The annotations are given at word level. Few example of words in vocabulary: 'AFTERNOON', 'BORDER', 'FORWARD', 'MISSING', 'SHORT', etc.

Lipreading Sentences in Wild:

Also abbreviated as LRS dataset [19] contains video clips annotated for sentences. There are around 118k utterances with more than 780k word utterances spread across pre-training, training, validation and test sets. Dataset comprises of a total vocabulary size of 17,428. This dataset is again curated from BBC news videos.

1.4 Organization

Following three chapters (chapters 2, 3 and 4) provide a self contained description of our contributions on word spotting in silent speech videos, lip-synchronization, and lip-landmark localization respectively. Brief outline of text in this thesis is as follows:

- In chapter 2, we explain our novel word spotting pipeline for spotting word spoken by one person in another target silent speech video.
- In chapter 3, we explain our cross-accent and cross-language lip-synchronization method for synchronizing lips of a speaker in a speech video based on a target audio clip.
- In chapter 4, we detail the recent works in the domain of landmark detection and explain our proposed method for lip-landmark localization.
- Finally, we end the thesis with concluding remarks and future work.

Chapter 2

Word Spotting in Silent Lip Videos

Our goal is to spot words in silent speech videos without explicitly recognizing the spoken words, where the lip motion of the speaker is clearly visible and audio is absent. Existing work in this domain has mainly focused on recognizing a fixed set of words in word-segmented lip videos, which limits the applicability of the learned model due to limited vocabulary and high dependency on the model's recognition performance.

Our contribution is two-fold: 1) we develop a pipeline for recognition-free retrieval, and show its performance against recognition-based retrieval on a large-scale dataset and another set of out-of-vocabulary words. 2) We introduce a query expansion technique using pseudo-relevant feedback and propose a novel re-ranking method based on maximizing the correlation between spatio-temporal landmarks of the query and the top retrieval candidates. Our word spotting method achieves 35% higher mean average precision over recognition-based method on large-scale LRW dataset. Finally, we demonstrate the application of the method by word spotting in a popular speech video ("*The great dictator*" by Charlie Chaplin) where we show that the word retrieval can be used to understand what was spoken perhaps in the silent movies.

2.1 Introduction

Parsing information from videos has been explored in various ways in computer vision. Recent advances in deep learning have facilitated many such tasks. One such parsing requirement is of reading lips from videos. This has applications in surveillance or aiding improvements in speech recognition in noisy outdoor settings. Solving this problem has been attempted using methods based on recurrent neural networks (RNN) [46] and spatio-temporal deep convolutional networks [54]. However, for practical applications, recognizing lip motion into words is still in its nascent stages, with state of the art models [86] being limited to a constrained vocabulary. In this work, we adopt a recognition-free 'word-spotting' approach that does not suffer from the vocabulary limitations. Unlike text documents, where the performance in character recognition [106], word recognition [23] and spotting research [87] has seen a great boost in the post deep learning era, this approach has been rarely pursued for lipreading task.

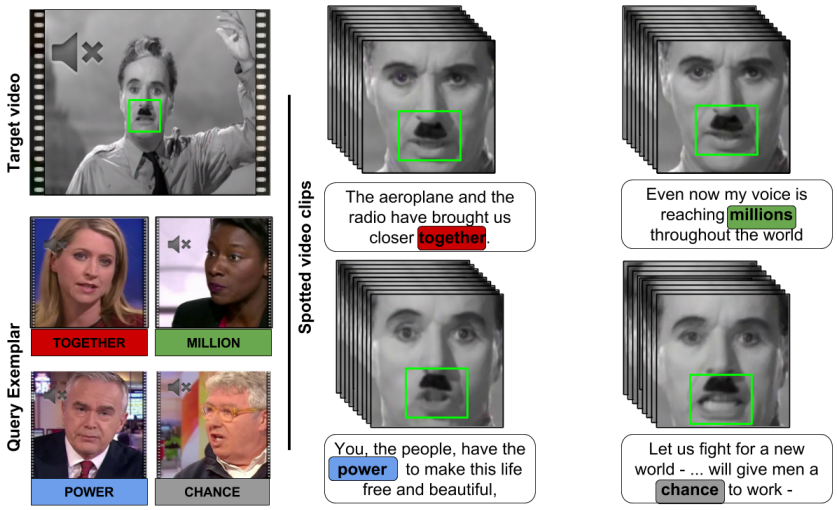

Figure 2.1 Example of word spotting in black and white Charlie Chaplin silent video: (left) target is the silent video and queries are the exemplars spoken by different people;(right) retrieved video clip segments where the words 'together', 'million', 'power' and 'chance' are present.

Training a lipreader requires careful word level annotation, which is expensive even for a small vocabulary set. Although progress in speech recognition [104] has resulted in better audio-to-text prediction and can be used for annotation, such methods are often prone to changes in accent and presence of noise in the audio channel. Lipreader's performance is also susceptible to similar sounding words [86]. In recognition-based retrieval, we use a lipreader to predict the word spoken in a video clip. Evidently, if the word is wrongly predicted due to variations in visual appearance, it would never appear in the top results. In contrast, for recognition-free retrieval, the 'word spotting' i.e. matching of words is based on the feature representation of the target word without explicitly predicting the word itself. It intrinsically compares the features of the target word with the query word. Hence, even if the target word is misclassified it appears in the top results.

We are motivated by the fact that for handwritten documents word spotting has shown better performance for retrieving target words in different handwriting styles than word recognition [70]. Likewise, we show that recognition-free retrieval can also be useful for spotting words when target words come from a different source than the data used for training a lipreader, like archaic black and white documents films in Figure 2.1. We further investigate the applicability of recognition-free pipeline for out-of-vocabulary word spotting, for a different domain of data with respect to what has been used for training the lipreading model. Figure 2.1 shows few sample results of our pipeline for spotting different query words in the black and white video clip in four spoken sentences.

We further show that the word spotting performance can be improved by using a novel re-ranking method for top-k retrieval candidates. We also adapt the standard pseudo-relevance feedback query expansion method for lipreading task. Our pipeline takes silent speech videos as input and retrieves a queried word that is provided again as a video clip from the target input dataset. The target video is first densely segmented into 'word proposal clips', where these clips may or may not contain any word. Any

'word proposal clip' is considered a spotted word if the similarity measure between the query and the target 'word proposal clip' is greater than a particular threshold.

We show improvement in word spotting on a standard large scale lip video dataset Lipreading in the wild (LRW) [17], and another standard dataset GRID corpus [21] for showing domain invariance. We also assess our pipeline's performance in a popular speech video by Charlie Chaplin: "*The great dictator*".

2.2 Related Work

Research in visual speech recognition has been pursued for at least two decades [7, 13, 44] with earlier approaches focusing mainly on handcrafted features and HMM-based speech recognizers [9, 45, 65]. Some of these approaches have been thoroughly reviewed in [42, 108]. Wand *et al.* [97] showed word level lipreading using an LSTM [46] stacked over two-layered neural network on GRID corpus dataset [21]. Recently, Chung and Zisserman [17] have used multiple lipreading models that fuses the temporal sequence at different layers of underlying VGG-M model [15] to classify the input video clip into 500 words. Assael *et al.* [5] uses a Connectionist Temporal Classification (CTC) [40] to show one of the best results on GRID corpus [21].

Lipreading involves modeling temporal sequences of lip video clips into phonemes or characters, hence better sequence learning models using deep networks proved to be pivotal in lipreading research. Chung *et al.* [19] have proposed Watch Listen Attend and Spell (WLAS) architecture that leverages attention model [6] for doing character level prediction of input lip videos. They provide the best results on Lipreading in the Wild (LRW) dataset and GRID corpus [21]. They however use a much larger Lipreading Sentences (LRS) dataset that is not widely available [19] for pretraining, hence making it a data intensive model that is not accessible. In a recent work, Stafylakis and Tzimiropoulos [86] trained a model entirely on LRW dataset to give state-of-the-art result for word level prediction. This model consisted of three parts: a spatio-temporal convolutional front-end, followed by a Resnet-34 [43], and a bidirectional LSTM [39] at the end. Since this model has been trained to classify lip videos into one of 500 word classes, it does not address out-of-vocabulary words. Our pipeline employs recognition architectures based on [17] and [86] as feature extractors to show how recognition-free leverages these features spaces for improved retrieval performance.

Initial work in word spotting appeared in speech recognition community, majority relying on HMMs [36, 83]. Kernel machines and large margin classifiers introduced by Keshet *et al.* [58] in discriminative supervised setting resulted in an improvement over the previous methods. Post deep learning, RNNs with CTC objective functions gave a major improvement over the HMMs [29] for modeling temporal audio speech signals. Unlike audio speech, visual speech is spatio-temporal signal. Hence, our choice of feature extractors contain VGG-M [15] and Resnet-34 [43] modules for modeling facial features, and uses LSTM and temporal convolution for modeling temporal information.

Word spotting is a well defined problem in document analysis and retrieval [35]: hand writing recognition [30, 33, 70, 87], word image retrieval [62], scene-text [99] etc. Although a large corpus of work exists for word spotting for documents, images and audio speech, the visual speech domain has been largely ignored. The work that is closest to our approach is by Wu *et al.* [103]. In their approach, the authors use geometric and appearance based features to build their word spotting pipeline and they rely on the knowledge of optimal handcrafted feature. In our work, though we also adopt a recognition-free retrieval approach, we do so using recognition-based features and show that the recognition-free approach improves on the recognition-based approach. We further also improve the base recognition-free pipeline by using query expansion and re-ranking extensions. We benchmark our work on standard datasets.

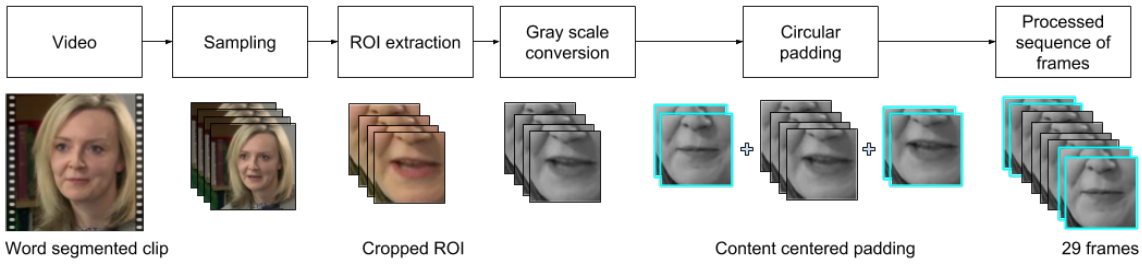

Figure 2.2 Preprocessing: The pipeline which takes a variable length word clip and converts it into a fixed length sequence of frames.

2.3 Proposed Method

In this section, we will discuss the individual components of our proposed word spotting pipeline and move along to develop a holistic overview of the method.

2.3.1 Recognition-free Retrieval

Recognition-based retrieval relies on recognizing words in lip videos by completely depending on the lipreading model. During testing a video clip containing a word is classified as one of the word in the vocabulary it is trained on. Moreover, modeling a lipreader with open vocabulary is an active area of research.

Retrieving a word from a set of candidate silent videos without directly recognizing each candidate words being spoken is recognition-free retrieval or word spotting. This opens up an opportunity to use a sub-performing lipreader with incorrect word recognition. In a recognition-free setup, the user formulates a query and a rank list is computed based on its distance from all the clips in the target corpus (retrieval set), such that most similar candidate is given the highest rank. Since word spotting

systems rely heavily on computing similarity, the quality of the feature representation is more important than the classification of input clips.

Word spotting based on the modality of query are of two types: query by string (QbS) where the input query is string and the retrieval is video, and query by exemplar (QbE), where query is video and retrieval is also video. In this work, our query will be through exemplar.

2.3.2 Preprocessing

We use the recognition models as described in [19] and [86] as feature extractors. These models takes inputs as a fixed length input of spatial dimension 225×225 and 112×112 respectively with a sequence length of 29 frames. The feature extractors are trained on LRW [17] dataset which consists of fixed length video clips of 29 frames and 1.16 sec duration, with actual word at the center. Hence it is required to preprocess the input videos (other than that of LRW) before feeding them to the feature extractors. As shown in Figure 2.2, the preprocessing step proceeds by sampling the input video at 25 frames per second. We extract the facial-landmarks using dlib library [60] and localize the lip using them. The sampled region of interest (ROI) frames are then converted to grayscale. Since words can be of different length we circular pad grayscaled sequence of frames on both the side such that the actual content is at the center of the sequence. Circular padding of length 2 for a sequence: $\{1, 2, 3, 4, 5\}$ on both sides gives $\{4, 5, 1, 2, 3, 4, 5, 1, 2\}$.

2.3.3 Video Features

Our first feature extractor only uses the visual stream of the WLAS architecture and hence called Watch, Attend and Spell (WAS) model [19]. Chung *et al.* [19] train WLAS model on LRS dataset [19] and fine tune it on LRW dataset [17]. As LRS dataset [19] is not yet publicly available, we trained our WAS model entirely on LRW dataset. WAS contains two modules: a VGG-M convolution module and an attention-based sequence to sequence LSTM module, followed by 28 neurons with softmax non-linearity. Our output sequence for a lip video clip is maximum 20 character long, 28 dimensional(D) (A to Z, *eos*, *padding*) ground truth (GT) word label. Using early stopping we achieve a word accuracy of 53%.

We also employ another network 'N3' as described by Stafylakis and Tzimiropoulos [86] for feature extraction. This network is composed of three modules: A layer of 3D convolutions followed by three dense layers (fully connected layers), and finally a temporal convolution layer. The final layer has 500 neurons with softmax non-linearity. The classification accuracy of this model is 69.7%. We will address this model as CMT in this work.

In both the feature extractors, the choice of features are the softmax scores or the probabilities of a lip videos belonging to different words in the vocabulary, instead of sparsely belonging to only one word. We also experimented with the output of the last dense layer as feature representation for the input video, and found softmax scores to be empirically better.

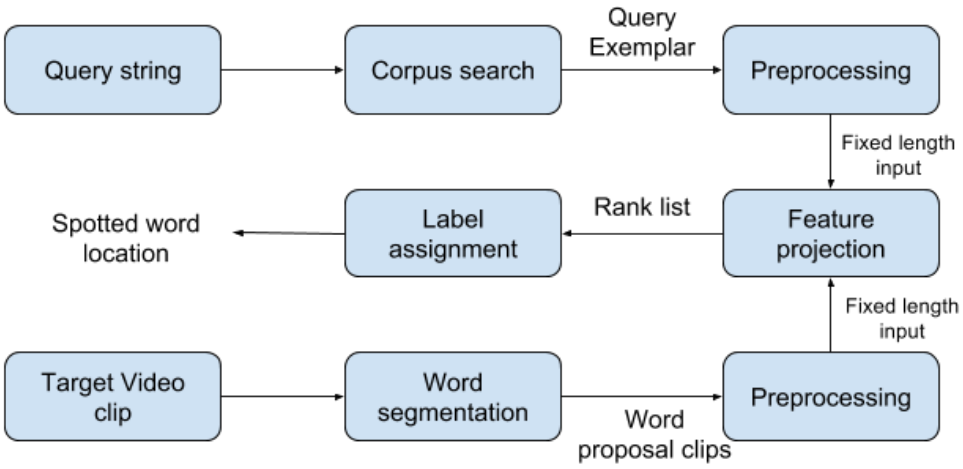

Figure 2.3 Overall pipeline: First a string is searched in an annotated corpus to formulate an exemplar which is then preprocessed, and projected into feature space. Target video is then segmented into word clips, either using given time-stamp or dense segmentation, preprocessed and projected in the same feature space. A ranking is computed based on the cosine similarity between query exemplar and the word proposal clips. Label is transferred based on majority voting, as discussed later in Subsection 3.4

2.3.4 Overall Pipeline

In this section, we propose a pipeline for spotting words in silent lip videos. In order to demonstrate generic nature of our pipeline, we first train our two different feature extractors on LRW dataset. We project the query set, consisting of preprocessed annotated video clips, and retrieval set video clips which do not have any labels into the feature space. The label of the query is assigned to a particular candidate clip in the retrieval set, only if the mean similarity score of that candidate with all the same label queries is greater than a threshold, otherwise it is assumed the candidate word proposal clip does not contain a full word. In Figure 2.3 we show our overall pipeline.

More precisely, if q_i^c is the feature representation of i^{th} query belonging to label c and r_j is feature representation of the j^{th} word proposal clip, the similarity score between the two is given by n_{ij}^c in Equation 3.1.

$$n_{ij}^c = \frac{(q_i^c)^T . r_j}{\|q_i^c\| \, . \, \|r_j\|} \tag{2.1}$$

The average similarity between all the queries q^c belonging to label c and the candidate r_j is given by s_j^c in the below Equation 3.2.

$$s_j^c = \frac{\sum_{|q^c|} n_{ij}^c}{|q^c|} \tag{2.2}$$

Finally, the label assignment for the candidate r_j is c if the mean similarity score between all the queries belonging to label c, i.e. s_j^c, is greater than ρ. Otherwise, we consider the word proposal clip is either noise or does not contain the whole word, as represented by ϕ.

$$label_{r_j} = \begin{cases} c & \text{if } s_j^c > \rho \\ \phi & \text{otherwise} \end{cases} \tag{2.3}$$

Hence, these word proposal clips are spotted as word c using the queries q_i^c in the target video. We can further use enhancements over this pipeline to improve the retrieval performance, which we will discuss in the next section.

2.4 Enhancements

In this section, we discuss a query expansion technique to search videos with semantic relevance to the given query, followed by re-ranking method to improve ordering of top-k results.

2.4.1 Query Expansion and Re-ranking

Query expansion, in image retrieval [4], has been widely used to improve retrieval performance by increasing the recall and obtain additional documents which might get missed with the original query. Similar to documents, we first feed a *seed* query to our retrieval system which gives us a ranked list of all the candidates from the retrieval set. From this set, top-k candidates are selected to construct a new query based on the weighted sum of the query and top-k candidates feature vectors as the pseudo-relevance feedback to improve the retrieval results.

Re-ranking is used to improve the ranking of top retrieval results for a given query. Some of the prominent re-ranking method [28, 95] relies on geometrical consistency between query and its top retrieval candidates. Fergus *et al.* [28] uses RANSAC [31] to re-rank top results from Google Image search engine. Unlike images, lip videos are temporal in nature with each word consisting of a specific set of phonemes. To adapt such a method for lip videos, we extract spatio-temporal features. Out of total 68 facial landmarks [60], we first compute the distance between all the 20 landmarks associated with lip and the lip-central landmark (landmark no. 63), as shown by 'red' color landmark in Figure 2.4(a). Both landmark no. 63 and 67, being in the center, are clearly visible for different head poses and hence can be chosen for computing distances. However, on an average, the motion of the upper lip is lesser than the lower lip for most of the word utterances, makes landmark 63 more stable and a better choice.

This geometric feature extraction results in a 20D spatial feature for each frame, or $20 \times 29D$ spatio-temporal feature for the video clip. We then re-rank our candidate using their temporal lip landmark correlation with the query lip video, as shown in Figure 2.4(b). Using recognition-free retrieval top-k candidates are selected for a given query. Spatio-temporal features for both top-k candidates and query are extracted. The correlation of landmark of the lip region of these top-k candidates with the query is computed, the re-ranking is done in the order of decreasing correlation.

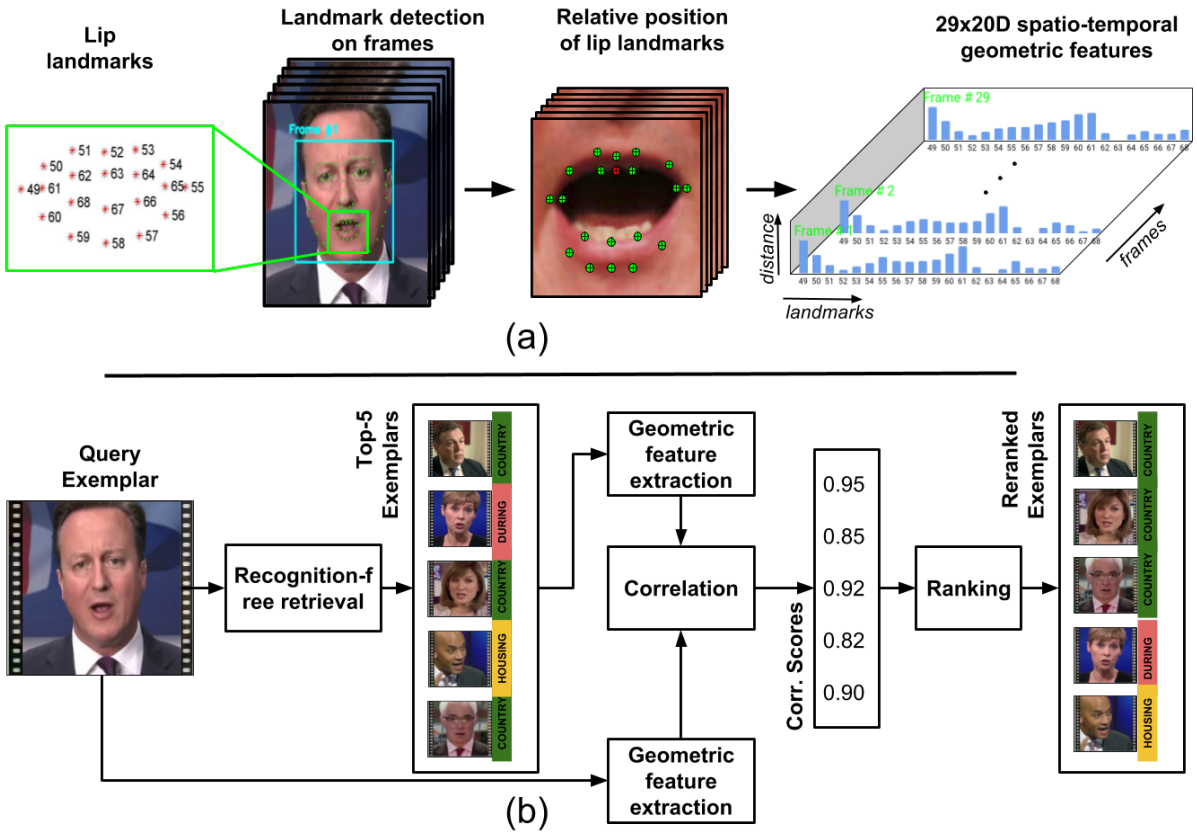

Figure 2.4 Re-ranking using geometric cues of lip video: (a) shows method of extracting spatio-temporal feature using lip landmarks of each frame of the video clip; (b) shows re-ranking of top-5 retrieved candidates based on the correlation between spatio-temporal features of top-5 candidates and that of the query.

2.5 Experiments

2.5.1 Datasets

Lipreading in Wild (LRW) [17] has 500 words classes with 1000 clips for training, 50 for testing, and 50 for validation for each of the words, which has been curated from BBC news videos. Each word clip is of length 1.16 second duration containing 29 frames. We use the LRW to train both feature extractors. The proposed retrieval pipeline only uses the test set for querying and validation set for retrieval, since training set has been used to train feature extractors.

GRID corpus [21] contains 1000 phrases, spoken by each of 33 speakers. Each phrase has a fixed syntax containing 6 words: $command(4) + color(4) + preposition(4) + letter(25) + digit(10) + adverb(4)$; an example of which is 'put red at G 9 now'. We use speakers 10-19, similar to [97], in our experiment. For showing domain invariance, we randomly sample 1000 phrases from these speakers to create our query set. Similarly, we sample another 1000 phrases from the same speakers to create

Figure 2.5 Random frames from LRW dataset (top row), GRID corpus (middle row) and Charlie Chaplin "*The great dictator*" speech video (bottom row).

our retrieval set. All the speech videos are word segmented and preprocessed before feeding to feature extractors.

For qualitative results we show lipreading on **Charlie Chaplin's** famous "The great dictator" speech video. We only use the video, without audio cues for our experiment.The video is segmented into sentence level video clips using the timestamps provided by Youtube subtitles, which also gives the ground truth annotations. The retrieval corpus is made by densely segmenting these sentence videos into word proposal clips. Randomly selected frames from these three datasets are shown in Figure 2.5.

2.5.2 Implementation

For WAS, we use the pretrained VGG-M model from Chung and Zisserman [18], and only train attention sequence-to-sequence LSTM module, while freezing the weights of VGG-M module. We use the LRW training set for training our model, with validation set used for parameter tuning. The network has been trained with batch size 64, cross-entropy loss and SGD optimizer. Initial learning rate was set to 0.1 with a decay of 0.01% every two iterations. No data augmentation was used.

For training CMT, we follow the similar procedure as mentioned in Stafylakis and Tzimiropoulos [86] to train our model end-to-end. Again the batch size of 64 was taken with cross-entropy loss and SGD optimizer was used. Initial learning rate was set to $3e^{-3}$ with exponential decay in learning rate when the validation loss does not decrease for 2 epochs. We also perform data augmentation with

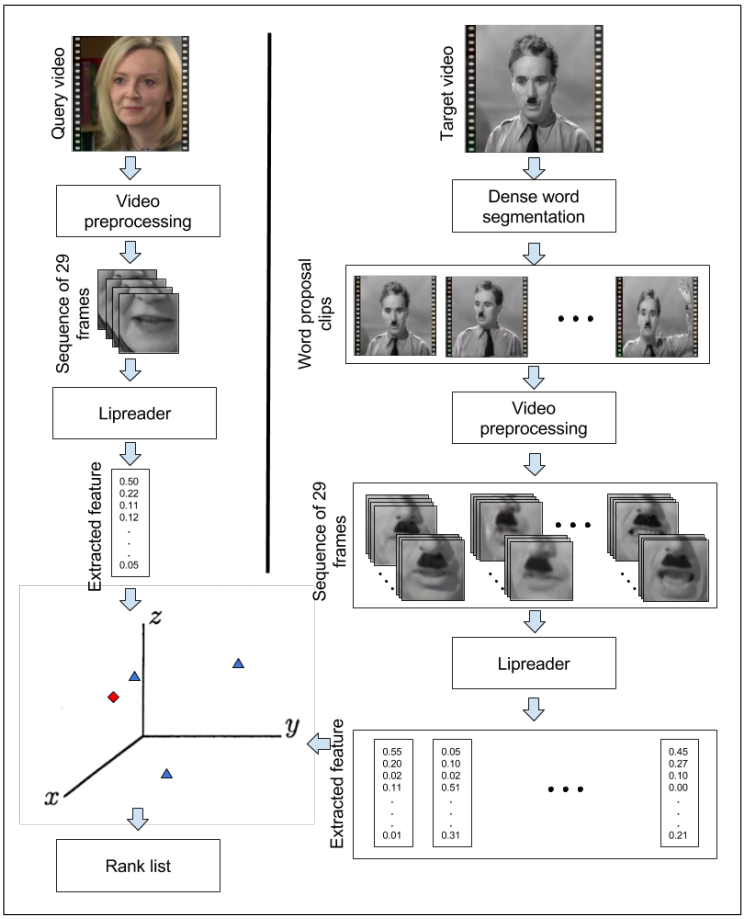

Figure 2.6 Word spotting in Charlie Chaplin video: (left) a query exemplar with known annotation is preprocessed into fixed length input and fed to the feature extractor. (right) the Charlie Chaplin video is first densely segmented into word proposal clips and fed to the feature extractor. All the word proposal clips and query exemplar is projected into feature space and ranking is computed based on cosine similarity.

random cropping of 4 pixels around the lip region of interest (ROI), and horizontally flipping all frames of randomly chosen input clips. For both the networks, WAS and CMT, early stopping was employed if validation accuracy failed to improve over 3 consecutive epochs. We implement both the networks in Keras deep learning library [16].

Word spotting on LRW dataset has been shown considering LRW test set as query set and LRW validation set as retrieval set. Here, we want to assign label to the query video clips, considering we know the GT label for retrieval set. Both the query and retrieval set are first preprocessed, as discussed in Section 3.2. Since all the video clips are 29 frames long, circular padding is not required during preprocessing. After feature extraction, the query is searched in the retrieval set, the candidate with highest cosine similarity is ranked highest. To transfer word label from retrieval set the query, we take the majority vote of top-5 candidates in the retrieval set.

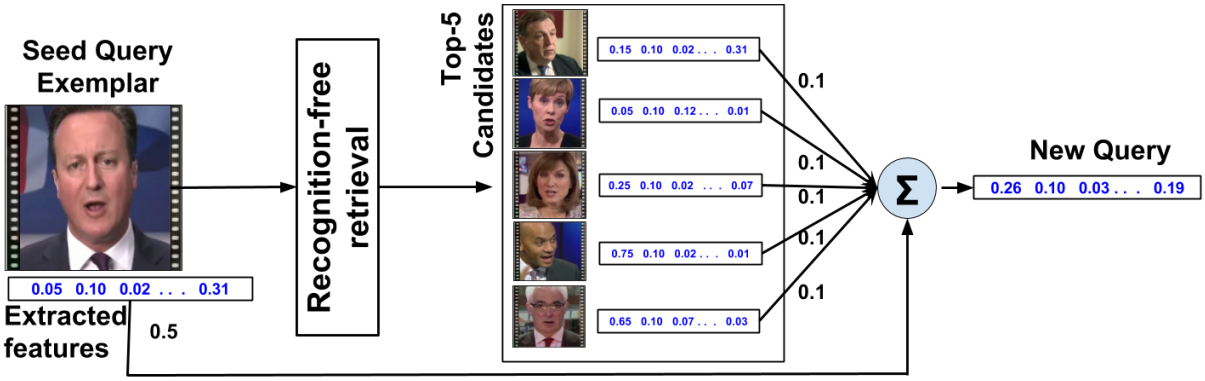

Figure 2.7 Formulation of new query: The weighted sum of the feature representation of *seed* query and its top-5 retrieved candidates becomes the new query.

During query expansion, we first search a *seed* query in the retrieval set to get top-5 candidates. The 'New query' is the weighted sum of the top-5 candidates with weights 0.1 each and *seed* query with weight 0.5, as shown in Figure 2.7. This query is then used to retrieve a new set of candidates which becomes our final retrieval for the *seed* query.

For each query video coming from LRW test set, we retrieve top-10 candidates from LRW validation set using recognition-free retrieval. For Re-ranking, we then extract spatio-temporal feature for both query video and its top-10 retrieval candidates using DLib [60] and OpenCV [10] libraries. Correlation between spatio-temporal features of query and candidates were computed and were used to re-rank the top-10 candidates. This method proves to be effective in refining the search results for our retrieval pipeline.

For showing word spotting in Charlie Chaplin video, as shown in Figure 2.6, the sentence videos are densely segmented into fixed length (29 frames) word proposal clips by taking stride of 3 frames. We spot the words in retrieval corpus consisting of these clips. Since the segmentation is dense there will be very few word proposal clips which will entirely cover actual words spoken in the video. As discussed in Section 3.4, we calculate the average similarity score between all the query exemplars coming from LRW validation set belonging to a particular word label and a word proposal clip from Charlie Chaplin video. If the average similarity is more than a threshold (ρ), we assign the word label to the word proposal clip. We empirically selected the value of $\rho = 0.3$ for this experiment.

2.5.3 Baselines

We compare our pipeline with recognition-based retrieval. WAS [19], in the original paper, was first pretrained on LRS dataset, and later fine-tuned on LRW dataset, gives a word accuracy of 76.2%. Our WAS model trained solely on LRW dataset gives the word accuracy of 53%. The recognition-based baseline of our WAS is given in Table 1, column 1. Another lipreader CMT, gives the word accuracy of 69.7%. The recognition-based baseline is given in Table 1 column 3.

For GRID corpus we do not fine-tune our LRW trained base feature extractors on GRID corpus. The recognition-based baseline for the domain-invariance out-of-vocabulary retrieval is shown in Table 3, column 1 and 3.

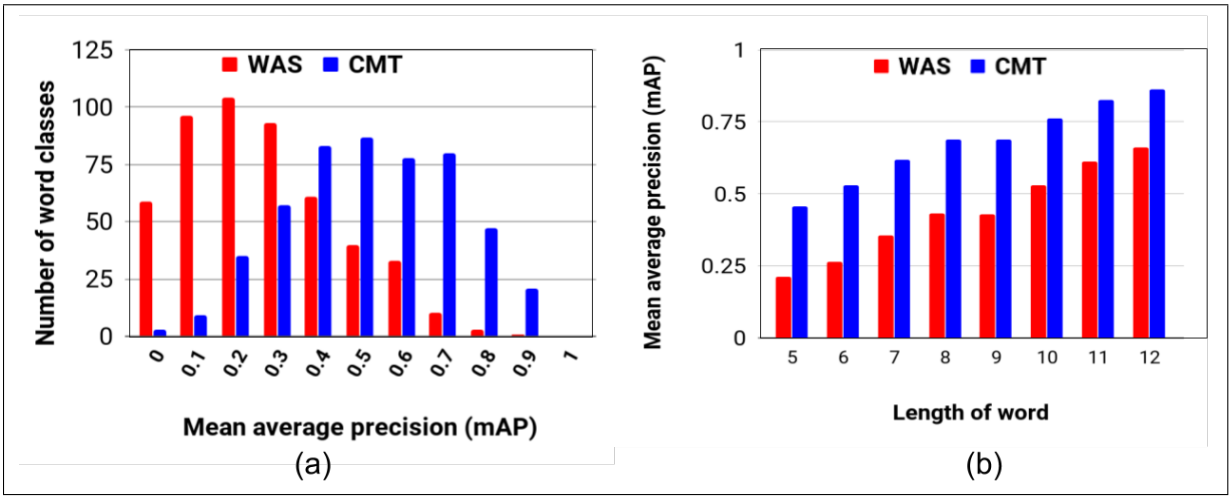

Figure 2.8 (a) number of words below a certain mAP for WAS and CMT based pipeline: y-axis is the number of words, and x-axis is the mAP; (b) variation of mean average precision (mAP) with the length of the word for CMT and WAS based pipeline: y-axis is average mAP and x-axis is word length in LRW vocabulary

2.5.4 Evaluation metric

For search based applications, the most important performance factor is: how many good results are in the top search results. Hence, Precision@K, which measures the precision at fixed lower levels of retrieval results, makes sense as an important performance metric. It considers the number of desirable results out of the top-k retrieval results without taking into account the overall rank ordering of the search results.

Recall@K is another important evaluation metric that we show, which is the number of desired results retrieved among top-k search results, with respect to the total number of available positive results.

While Precision@K and Recall@K give specific insights into the performance of the retrieval system, both measure performance for a fixed number of retrievals (K) and are insensitive to the overall rank ordering of the search results. We therefore also report the Mean Average Precision (mAP) for our retrieval system. mAP provides a measure of the quality of retrieval across different recall levels. mAP has been shown to have especially good discrimination and stability, and is one of the most standard evaluation measures for word spotting.

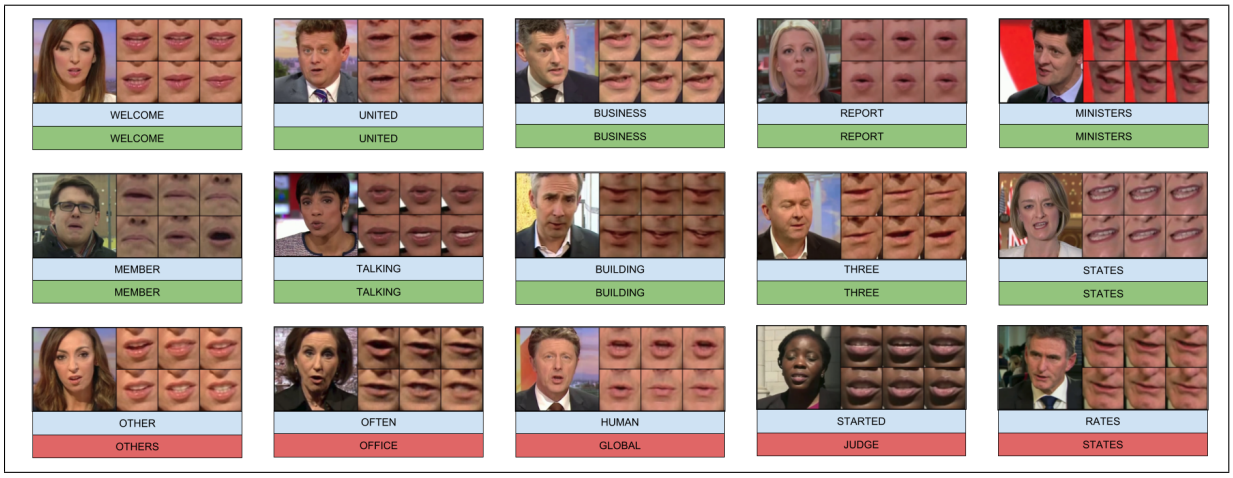

Figure 2.9 Qualitative results on LRW dataset: Each image depicts the central frame of the query video clip (left) and a sequence of lip ROIs of 6 consecutive frames around central frame, shown in raster order (right); (middle) blue boxes are the ground truths; (bottom) green boxes are correct predictions while the red ones are incorrect predictions. Label is propagated to a query based on the majority label present in the top-5 retrieval candidates.

2.6 Results

2.6.1 Comparison with Baseline Methods

Recognition-free retrieval or word spotting on LRW dataset when the base lipreader is WAS gives an absolute improvement of 35.9% over the recognition-based baseline of mAP 0.23; Table 1, column 2. Similarly, for recognition-free retrieval using CMT lipreader there is an improvement of 49.67% over the recognition-based baseline of mAP 0.38; Table 1, column 4. For recognition-free retrieval using WAS (in red) and CMT (in blue) feature extractor, Figure 2.8(a) shows the number of words below a certain mAP value. The variation of average mAP with the length of the words in the LRW vocabulary is shown in Figure 2.8(b). It can be seen that the average mAP value increases with the increase in word length. The qualitative results for word-spotting on LRW dataset using CMT features can be seen in Figure 2.9.

Query expansion on LRW dataset using two lipreaders: WAS and CMT give a mAP of 0.3146 and 0.5722 respectively; Table 2, column 2 and 5. Although the mAP results are comparative to the recognition-free method, we see an overall increase in recall@10. Also, re-ranking using spatio-temporal cues improves the retrieval performance for WAS and CMT, giving a mAP of 0.3179 and 0.5709 respectively; Table 2, column 3 and 6.

Charlie Chaplin "*The great dictator*" speech video, contains 39 words from LRW vocabulary. It has a total of 54 spoken sentences, out of which 33 sentences actually contains LRW vocabulary words. Hence, the query set contains 50 exemplars, from LRW validation set, belonging to each of these 39 common vocabulary words. Using our CMT based recognition-free pipeline we were able to correctly

	WAS		CMT	
	RB (BL)	RF (ours)	RB (BL)	RF (ours)
mAP	0.2317	**0.3149**	0.3807	**0.5698**
P@10	0.2928	0.4566	0.3253	0.6519
R@10	0.0586	0.0913	0.0651	0.1304
% imp.in mAP	–	**35.90**	–	**49.67**

Table 2.1 Retrieval performance for LRW dataset: Left two columns show recognition-based (RB) baseline and recognition-free (RF) performances for WAS features; right two columns show the similar results for CMT features. Across columns (first row) mAP is mean average precision, (second row) P@10 is precision at 10, (third row) R@10 is recall at 10, and (last row) % imp.in mAP is percentage mAP improvement of recognition-free retrieval over baseline.

Figure 2.10 Qualitative results on Charlie Chaplin "*The great dictator video*": Each image is one of the frames in the sentence clips extracted from the speech video. The top text box in blue color contains the subtitles with **bold** text showing the common LRW vocabulary word present in the subtitle. The bottom text box shows the correctly spotted word.

spot instances of 13 instances of the common vocabulary words in 11 sentences. Whereas on using recognition-based pipeline, only 6 instances of common vocabulary words in 6 sentences are correctly predicted. The qualitative results can be seen in Figure 2.10, where we spot the sentences which contain the query words.

2.6.2 Domain Invariance

Domain Invariance provides us the robustness of the pipeline for target data distribution different from the one it is trained on. GRID corpus contains 51 words with only 1 common word available in LRW dataset vocabulary. Hence this experiment also shows out-of-vocabulary retrieval performance of the proposed pipeline.

	WAS			CMT		
	RF	QExp	ReR	RF	QExp	ReR
mAP	0.3149	0.3146	**0.3179**	0.5698	0.5722	**0.5709**
P@10	0.4566	0.4591	0.4566	0.6519	0.6572	0.6519
R@10	0.0913	**0.0918**	0.0913	0.1304	**0.1314**	0.1304

Table 2.2 Different recognition-free performance for LRW dataset: Left three columns are recognition-free (RF), query expansion (QExp) and re-ranking (ReR) performances for WAS features; right three columns show similar results for CMT features. Across columns (first row) mAP is mean average precision, (second row) P@10 is precision at 10, and (last row) R@10 is recall at 10.

On GRID corpus, the recognition-based baseline is 0.033 (mAP) for WAS features and 0.06 (mAP) for CMT features, while the recognition-free performance is 0.068 (mAP)for WAS and 0.177 (mAP) for CMT; Table 3, column 2. This signifies the utility of recognition-free retrieval for out-of-vocabulary words when the underlying lipreader is constrained by vocabulary size.

	WAS		CMT	
	RB (BL)	RF(ours)	RB (BL)	RF(ours)
mAP	0.033	**0.068**	0.060	**0.177**
P@10	0.034	0.219	0.224	0.322
R@10	0.002	0.016	0.019	0.020
% imp.in mAP	–	**106**	–	**195**

Table 2.3 Domain invariance results on Grid corpus dataset (for both WAS and CMT): Left column has recognition-based (RB) baseline performance and right has our recognition-free (RF) performance where (first row) mAP is mean average precision, (second row) P@10 is precision at 10, (third row) R@10 is recall at 10, and (last row) % imp.in mAP is the percentage mAP improvement of our proposed method over baseline.

2.6.3 Discussions

Many conclusions can be drawn from the result presented in Subsection 6.1. Recognition-free retrieval performed better than recognition-based counterpart for spotting words in LRW dataset. From Figure 2.8(b), we see that quality of retrieval improves when the length of word increases, as longer the word is more the number of phonemes it contains, and less is the chance of it being similar to other words. Errors in similar sounding words are more likely, as can also be seen in Figure 2.9.

Performance of recognition-based retrieval on GRID corpus is inferior to that on LRW dataset, the reason being neither of the two feature extractors in our experiments were fine-tuned on GRID corpus. Still, the recognition-free retrieval showed an improvement over recognition-based. Quality of lip video is also important, as some words in Charlie Chaplin videos were not spotted, due to lower contrast and quality of the lip ROI, as shown in Figure 2.10.

2.7 Summary

We proposed a recognition-free retrieval pipeline and showed its precedence over recognition-based retrieval for the task of word-spotting. The base features from WAS and CMT lipreading models have been used to spot words in LRW dataset with an improvement of about 36% and 50% over the recognition-based counterpart. Pseudo-relevance feedback and re-ranking techniques, using spatio-temporal geometrical cues available in the lip videos, has been incorporated in the pipeline to further improve the retrieval results. We also showed domain invariance of our pipeline through out-of-vocabulary word spotting on GRID corpus dataset with an improvement of 106% and 195% over the baseline using WAS and CMT features respectively. Lastly, we presented the practical applicability of our proposed pipeline by spotting words in 11 out of 33 sentences in the "Charlie Chaplin, *The great dictator*" speech video.

Chapter 3

Cross-Language Speech Dependent Lip-Synchronization

Speech videos such as movie dialogues, public speech, online courses are a great source of infotainment. These videos are often limited by the linguistic constraint of audiences being from different demographies. Vernacular backgrounds that are non-native to the accent or the language of the content producer often makes it difficult for a listener to comprehend the full essence of the content. Such videos are often supplemented with foreign language subtitles which hamper viewing experience. Otherwise, simple audio dubbing in a different language makes the video appear unnatural due to unsynchronized lip motion.

In this work, we try to address two problems that commonly arise in lip-synchronization: 1) Lip-synchronization for change in accent of the same language audio dubbing, and 2) Cross-language lip-synchronization for speech videos dubbed in a different language. We describe an automated pipeline to synchronize the lip movements of the speaker conditioned upon the audio in both cases. Our quantitative evaluation shows high SSIM index between generated cross-language lip-synchronized videos and the original videos. With the help of a user-based study, we verify that our method is preferred over unsynchronized videos.

3.1 Introduction

Speech videos are an effective way of story-telling. Many socio-political changes are caused by public speeches, movies depict the cultural aspect of societies. Similarly, 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. Moreover, the Internet has made it possible for a global inter-cultural exchange of ideas where any information is just a click away. Yet we find limited penetration for these speech videos when they cross international boundaries. For instance, the retention rates in MOOC courses can be as low as 10% [48]. One of the major reasons for this is a cultural gap between the linguistics of the audience and the content producer. 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 them. 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

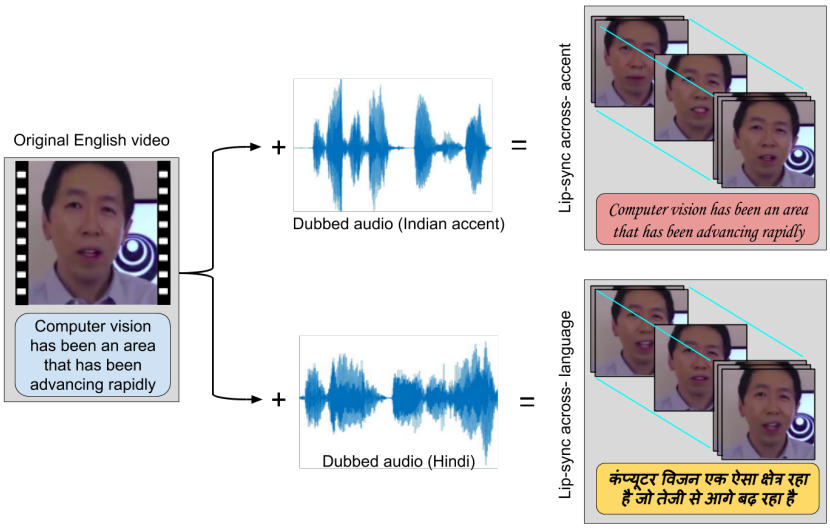

Figure 3.1 Lip synchronization on Andrew Ng Machine learning tutorial video based on dubbed audio: (top-right) shows Dynamic Programming to synchronize lip-motion of the original English video (left) into Indian English accent, (bottom-right) cross-language lip-sync to synchronize into different language (Hindi).

the audience. One way to alleviate this problem would be to dub speech videos in the accent or language of the audience. The current process of dubbing involves *translation* to the target language to resemble the lip motion of the source language as much as possible, *recording* of the dubbed content in pace with the original performance, and *editing* of the dubbed soundtrack and lip motion to be temporally close. This process is performed by production companies, and is both time-consuming and expensive. Even after such effort, the result of dubbing is not visually pleasing to the viewer because of clear visual discrepancy between the lip motion and the audio track.

This discrepancy between the speech information of the dubbed track and the facial motion of the video track is due to differences in correspondence of phoneme sequences and lip motions [88]. It causes a strong discomfort for viewers, and is also a huge distraction for those who are hearing-impaired, as they rely significantly on lip reading [76, 66]. This difference is one of the reasons why people dislike watching dubbed content [59], because it alters the sound perceived by the observer [71].

Such a problem can be solved by synchronizing the lip motion of the speaker in the target video to be coherent with the dubbed audio track. A similar problem exists in the field of computer animation, where lip-motion of the animated characters are constrained upon the textual script of the character. This usually required a human in the loop to manually lay the visemes, hence such a system cannot be scaled for photo-realistic lip-synchronization (lip-sync). These findings are the motivation for our work to solve the problem of lip-synchronization.

With recent developments in deep neural architectures like Generative Adversarial Networks (GANs) [37, 51], we are now able to generate photo-realistic images conditioned upon an input prior. Similarly, Recurrent Neural Networks (RNNs) [46] have given way to better sequence learning methods which

has become the de facto in time-series data modeling such as speech videos. Both of these methods are quintessentially data-driven approaches, and require large amount of training data.

In this work, we propose address the issue of synchronizing lip motion in speech videos according to the language it is dubbed in. Our main ideas and contributions are two-fold: 1) we propose a model for lip-syncing a target video with audio dubbing in a different accent of the same language, such as Indian English accent or French English accent, as shown in Figure 3.1 (top); 2) we propose another model to lip-sync the speech video based on audio dubbing in a different language, for instance English video with Hindi audio dubbing, as shown in Figure 3.1 (bottom).

The input to both our models is speech video where the lip-motion of a speaker is clearly visible, and the dubbed audio. The output is a video generated with synchronized lip motion. We also propose a scalable pipeline for dataset creation, which will later be used to train our models. Unlike audio-dubbing which requires professional dubbing artists to give their voices, our method does not depend on human visual input for lip-synchronization of a target video. Figure 3.1 shows the lip-synchronization for a video clip of an Andrew Ng MOOC video [1].

We evaluate our generative model based on the structural similarity (SSIM) index of the lip-synchronized videos with respect to the original English videos. Lastly, with the help of a user-based study we show that lip-synchronization makes the speech video more engaging while preserving photorealism.

3.2 Related Work

The essential component of speech perception in videos and animations is visual cues, such as visemes [89] and phonemes. In the past few years, a number of work have appeared in the field of visual speech recognition, where the target is to recognize speech by observing the speaker's lip-motion. This requires modeling the viseme-phoneme relationship to solve problems like lipreading [97, 19, 86], word-spotting [53], and speech reconstruction [25]. However, our problem is to synthesize lip from dubbed audio, which is essentially an inverse problem to visual speech recognition. Hence, we will restrict our focus to the works in the domain of lip-synchronization and generation.

The earliest work related to ours could be animation of facial movements in avatars modeled either from audio [56] or text [98, 101]. These work mostly used HMM [24] for sequential lip trajectory generation [100, 101]. One of the first systems for animating a virtual avatar's face directly from speech was proposed by [11], which models the joint distribution of acoustic and visual speech.

The work by Bregler et.al.[12] learns a mapping between the visemes and phonemes for one specific actor and language, and synthesizes new lip movements through image warping. However, the results of this method fail to dub between different languages and different individuals. Some of the recent work focus on synthesizing photo-realistic lip-motions and facial expressions. Face2Face [92] morphs the facial landmarks of a person in a target video based on those of another actor. However, these kind of models require a human in the loop, which can be quite expensive to scale, and prone to errors.

The advent of Recurrent Neural Networks, especially LSTMs [46], gave way to better sequence learning, which made generation of features from speech more efficient. Pham *et al* [80] used CNN followed by an LSTM to generate face parameters from input audio waveform. Karras *et al* [57] proposed a network consisting of a spatial convolution layer followed by a temporal convolution network on top of fully connected layers to convert speech audio into facial expressions. Chung *et al* [20] proposed an encoder-decoder convolutional network to jointly embed face and audio. The encoder network consisted of an audio stream and a video stream which merges in a bottleneck representation. This representation is used by the decoder to synthesize lips-motion video frames.

Most similar to our work are [91, 64] which use speech audio represented as Mel Frequency Cepstral Coefficients (MFCC) features [91] and text [64] 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 [91, 64] 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 languages.

3.3 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 lips, and generate new lip movements for the same speaker, given the dubbed audio. We consider dubbing to be of two types: a) in the same language as the original speech video but with a different accent, to alleviate the nuances related to unfamiliar accent; and b) where the video is re-dubbed into an entirely different language, to cater to non-native audiences. Both the dubbing types pose different challenges in lip synchronization. For English to non-native English, words remain the same while the pitch, tone and timing of the spoken word change across accents. On the other hand, for cross-language dubbing both words and phonemes change. In this section we will discuss two different methods to address these challenges.

3.3.1 Cross-accent lip-sync

While dubbing for different accent, since the same words are spoken by the instructor, all the required viseme sequences are already present in the original video. However, the time instances of the words being spoken might change. Hence, this problem of cross-accent lip-sync can be reduced to a non-linear alignment between original audio and the dubbed audio. This can be done by creating a mapping between different segments of the two audio clips.

In our setup, we have Andrew Ng's Machine Learning audio-visual clip [1], along with a dubbed audio clip of the same dialogues in Indian-accented English. Spoken words can be broken down into sequences of phonemes. Therefore, we densely segment these audio clips into 25 millisecond (ms) clips. We then use Dynamic Programming [22] to create a dynamic map between the segments of the

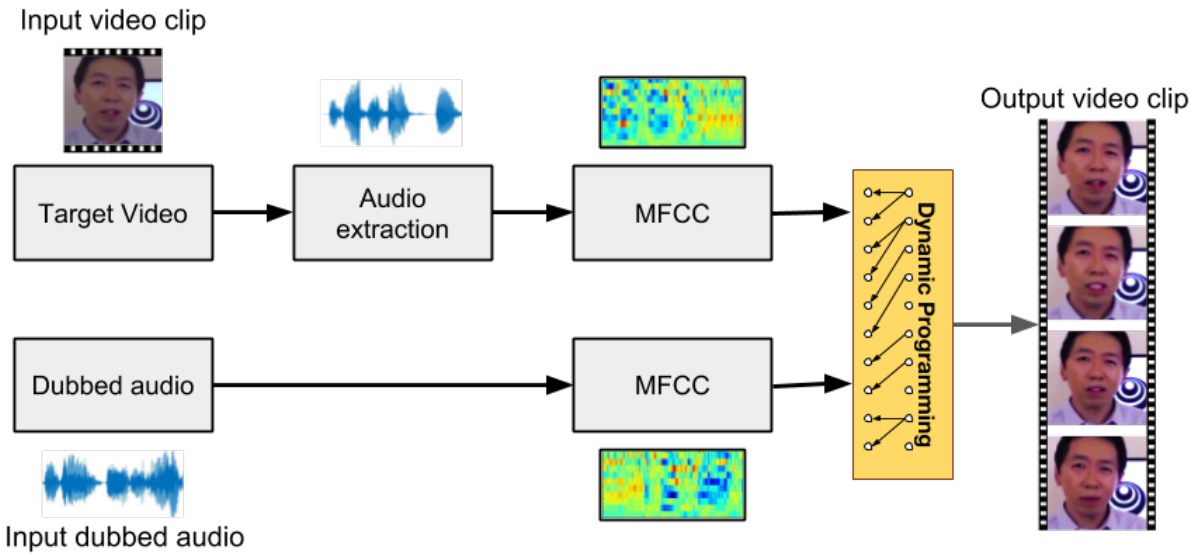

Figure 3.2 Pipeline for Dynamic Programming: Inputs are the target video and the dubbed audio in Indian accent, which are dynamically synced based on the similarity between the MFCC features of the voice of the native English speaker and the dubbed audio.

two audio clips in the feature space. MFCC has been widely used by the speech community, as they provide an optimal encoding of those audio bands which are most relevant to spoken speech. Therefore our mapping is based on finding the nearest neighbor of a segment from one audio clip in another. This is then converted into a mapping between the corresponding video frames. This procedure is illustrated in Figure 3.2, where we show the overall pipeline for non-linear alignment between original English audio and non-native English accent dubbed audio using dynamic programming.

Each segment of the original video clip is assigned a new time-stamp based on the mapping. This creates a non-uniform separation between adjacent segments of the original video clip. We render a new lip-synced video by interpolating frames in the original clip to fill the voids, and down-sampling to remove excess frames. Here we propose to use a Dynamic Programming algorithm (Algorithm 1), where x and y are the two inputs (in this case, the MFCCs of original and dubbed audio), L_x and L_y are the lengths of the respective audio segments. C represents the cumulative cost of the dynamic mapping with time, and M is used to find the optimal path for the mapping between input and output frames. $xTOy$ is the variable that records the mapping from x to y. Similarly, to find $yTOx$, the following can be appended at line 18: $yTOx[q-1] \leftarrow p-1$.

3.3.2 Cross-language lip-sync

Major challenges in lip-syncing audio of a foreign language (e.g. Hindi) on video of original language (e.g. English) are the differences in their grammatical structure and set of phonemes. One way is to directly generate lip-images conditioned upon the foreign language audio and target video. But such

Algorithm 1 Dynamic Programming

$init \leftarrow 5$
$C[i][j] \leftarrow 0, i = 0toL_x, j = 0toL_y$
for $i = 1toL_x$ **do** $C[i][0] \leftarrow i * init$
for $j = 1toL_y$ **do** $C[0][j] \leftarrow j * init$
$M[i][j] \leftarrow 0, i = 0to(L_x - 1), j = 0to(L_y - 1)$
for $i = 1toL_x$ **do**
    for $j = 1toL_y$ **do**
        $min1 \leftarrow C[i-1][j-1] +$ **cost**$(input_x[i-1], input_y[j-1])$
        $min2 \leftarrow C[i-1][j] + init$
        $min3 \leftarrow C[i][j-1] + init$
        $C[i][j] \leftarrow$ **min**$(min1, min2, min3)$
        if $C[i][j] = min1$ **then** $M[i-1][j-1] \leftarrow 1$
        if $C[i][j] = min2$ **then** $M[i-1][j-1] \leftarrow 2$
        if $C[i][j] = min3$ **then** $M[i-1][j-1] \leftarrow 3$
$p \leftarrow L_x, q \leftarrow L_y$
while $p \neq 0$ and $q \neq 0$ **do**
    if $M[p-1][q-1] = 1$ **then**
        $xTOy[p-1] \leftarrow q-1, p \leftarrow p-1, q \leftarrow q-1$
    else if $M[p-1][q-1] = 2$ **then** $p \leftarrow p-1$
    else if $M[p-1][q-1] = 3$ **then** $q \leftarrow q-1$

end-to-end systems require large amount of training data to learn the complex audio-visual relation between the two modalities [20]. At the same time, recent developments in generative networks [37, 51] have yielded impressive results. Considering these factors, we first learn an embedding between Hindi audio and lip-landmarks. This allows us to predict lip-landmarks from a relatively smaller speech corpus. From these predicted lip-landmarks, we generate mouth images over the original English video to match the Hindi audio. This entire two-step pipeline can be seen in Figure 3.3 (which also includes an intermediate processing step, discussed in Section 4.3).

3.3.2.1 Audio to Lip Landmarks

The first step is to encode audio into lip-landmarks. For each phoneme there exists a viseme, and the lip-motion responsible for the transition between two different visemes depends on the its location in the viseme-sequence that constitutes the spoken word. This makes audio to lip-landmarks a sequence modeling problem. Hence, similar to [91, 64], we use an LSTM [46] to encode audio. This LSTM takes audio MFCC features at each time step as input, and predicts lip-landmarks at time step 't' after each input, and is therefore called Time-Delayed LSTM (TD-LSTM) [38]. During training, the TD-LSTM is trained on Hindi speech audio-visual. The input is MFCC for every 25ms audio segment at 10ms time step, and the output is the lip-landmark of of the Hindi speaker at the 200ms delayed frame, as shown in Figure 3.3 (left). During inference, given a new audio sample, we use the predicted lip-landmarks as the prior to generate the mouth (lip-region) in the second step.

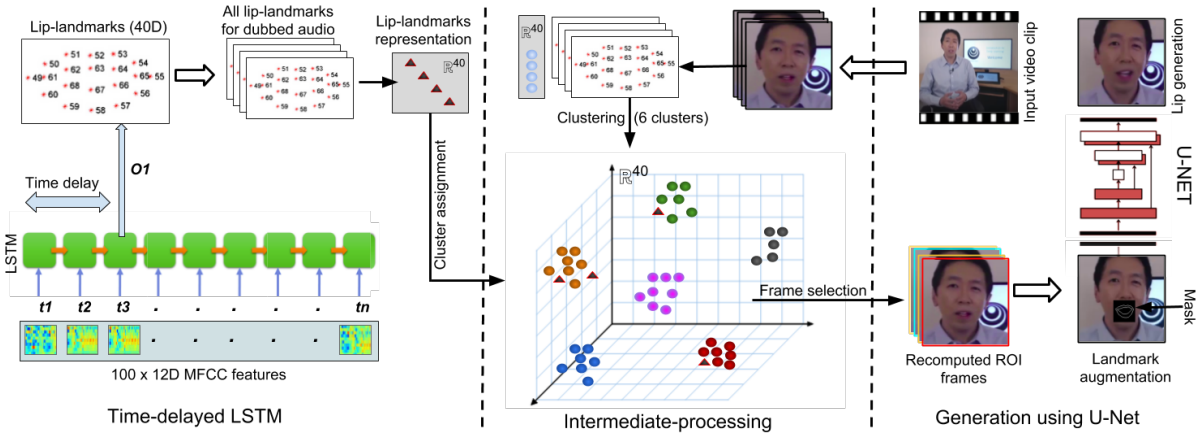

Figure 3.3 Cross-language lip-sync: (left) Pipeline for training LSTM with Hindi speech and lip landmarks, (center) shows reassignment of frames for each predicted lip-landmark using intermediate-processing step, (right) pipeline for inferring using U-Net on frames from English video.

3.3.2.2 Lip Landmarks to Generated Faces

Once lip landmarks are predicted from the audio in foreign language, in the second step the lips of the speakers in the original video must be modified to match these landmarks. To solve this problem, we use a U-Net similar to pix2pix generator [64] to generate mouth of the speaker based on an encoded prior. During training, the input to the network is the face image of the speaker, with the mouth masked by a black box of constant size, and the original landmarks in the face drawn as a white polygon, see Figure 3.3 (right). The output of the network is the original face image. This allows the network to learn to generate actual face of the speaker with the lip-region conditioned upon the lip-polygon on the masked face.

As L1 loss is commonly used while generating images, we use this to train the U-Net. In addition, since our main focus is on correctly generating mouth region of the speaker, we add another loss term to penalize wrongly predicted pixels in that region. Considering the mean of the black mask as the center of the mouth region, we add a Gaussian weight kernel G_{ij} to the L1 loss such that the weight of this loss decreases radially from the center of the mouth to the face extremities. Formally, for a ground truth $\hat{y}$ and the predicted face frame y, where any pixel location is represented by (i, j), our loss is defined as:

$$L = \sum_{i,j} (L1_{ij} * (1 + G_{ij})) \tag{3.1}$$

where;

$$L1_{ij} = \|\hat{y_{ij}} - y_{ij}\| \tag{3.2}$$

$$G_{ij} = c * \exp \frac{(i - u_i) * (j - u_j)}{v_{ij}} \tag{3.3}$$

Figure 3.4 Generation results for the U-Net: (top) input to U-Net, (middle) generated images, (bottom) ground truth

In equation 3, c is a normalization constant, u_i and u_j represent the mean pixel location of the black mask (mouth region), and v_{ij} represents the covariance.

During inference, the mouth in every frame of the video is replaced by a constant black square, and a white polygon of the lip landmarks predicted by the LSTM network from the previous step. Thus, the U-Net will then generate faces according to the Hindi dubbing audio. Unlike [91, 64], we train our network on multiple sources, allowing the network to generalize over multiple speakers. In Figure 3.4, we show some of the generation results of U-Net.

3.3.3 Dataset

In this section, we discuss our dataset curation pipeline used in our cross-language lip-synchronization method. We require two different datasets, one for each language: (i) Hindi speech dataset, for training time-delayed LSTM, and (ii) English videos dataset, to train U-Net for lip-generation.

3.3.3.1 Hindi speech dataset

As we wish to learn an encoding from Hindi audio to lip landmarks, we require a dataset consisting of Hindi audio to train a time-delayed LSTM. Since parallel audio speech corpus is difficult to find, and because dubbing is mostly a post-production phenomenon, we record 5 hours of audio-visual data of a native Hindi speaker narrating articles from Hindi newspapers and stories. Using voice activity detection [2], the video clips are segmented to give continuous segments of speech clips. For 5 hours of speech data, we get 5000 video clips of average length 2 seconds. Each such video clip is then sampled at 25 frames per second (fps). To obtain landmarks, we use a HOG-based face detector to find the speaker's face in the clip, and predict 68 face landmarks using Dlib [60]. We then choose the

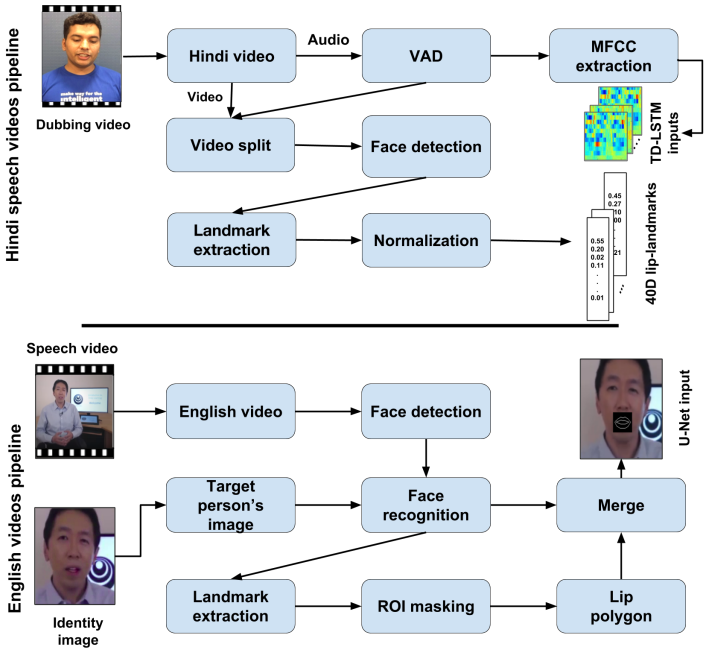

Figure 3.5 Dataset curation: (top) shows pipeline to create dataset from Hindi speech audio to give MFCC features and lip landmarks for training Time delayed LSTM (TD-LSTM); (bottom) shows pipeline to create dataset from Andrew Ng (and other) videos in English to give masked frames, for training U-Net.

landmarks corresponding to mouth region (landmarks 49 to 68) a.k.a. lip-landmarks, and normalize them. For each video clip segmented by voice activity detection, these normalized lip-landmarks are saved. Similarly, for each video clip we extract audio and sample it at 100Hz. We extract MFCCs for each sampled segment of the extracted audio clip. The training set consisted of 90% of total dataset, validation was done on rest of 10% of the dataset. The dataset curation pipeline for Hindi Speech can be seen in Figure 3.5 (top).

3.3.3.2 English speech dataset

As our aim is to generate lip-synced Andrew Ng's machine learning videos with Hindi dubbing, we use 20 Andrew Ng videos to create a dataset of English speech videos. The input to our U-Net is frames from instructional video clips from the English speech dataset, where the pixels in the mouth region of the instructor are masked with the wire-frame structure of the lip-landmarks. For each frame where the face of the speaker has been detected, the face region is noted as the bounding box of the 68 landmarks, similar to that in Hindi speech dataset. The square region around the face with 1.5 times the face width is extracted. This results in images with full visibility of the instructor's face. The mouth region of each face is considered as the bounding box around the mean of the mouth landmarks (49 to 68), and of width 0.25 times the width of the face region. It is then replaced with a black mask and a white polygon connecting the lip landmarks, and resized to the input shape of the U-Net. The output of the U-Net is

Figure 3.6 Random frames from Telugu movie dialogue clips (top-left), GRID corpus (top-right), English speech dataset(bottom-left), Hindi speech dataset (bottom-right).

the original face image. The training set consisted of 10 video clips while the validation was done on remaining 10 video clips. This pipeline is shown in Figure 3.5 (bottom).

It was important to not let the network overfit on the input images. We therefore used multiple sources of images as the training dataset for the U-Net — frames extracted from 1) Telugu movies, 2) videos of Andrew Ng's deep learning.ai lectures, 3) GRID corpus [21], 4) Hindi Speech dataset. Table 1 details the number of frames from each source. Figure 3.6 shows some randomly sampled frames from these datasets.

Source	Train	Validation
Telugu movies	37130 frames	4159 frames
English speech	24359 frames	16035 frames
GRID	13350 frames	1500 frames
Hindi speech	37790 frames	3714 frames

Table 3.1 Number of images in Train and Validation sets for training U-Net.

For each source of images, we detected faces and predicted landmarks using [14].

3.3.4 Representation

Audio: The input to dynamic programming algorithm and time-delayed LSTM is raw audio from the Hindi speech video clips represented as Mel-frequency cepstral coefficients (MFCC). The audio was sampled at 100Hz with each sample being of length 25ms. We extract 13 coefficient MFCC feature for each sampled segment, but only use 12 coefficient as the feature representation, discounting the first

feature. For TD-LSTM, length of each input training sample is 100 (or total 1 second), with shape 100 x 12D.

Lip-landmarks: The output of the TD-LSTM are lip-landmarks of the speaker of Hindi speech video clips i.e. 40 dimensional (D) normalized lip-landmarks for each frame (20 lip-landmarks x 2D). The output is computed at a delay of 200ms, and is represented as 1 x 40D vector.

Images: The input to our U-Net is a masked face image of size 256 x 256 x 3D similarly output is a 256 x 256 x 3D image consisting original face of the speaker.

Figure 3.7 Qualitative results for cross-accent and cross-language lip-sync: (a) In both the examples, frames are sampled at 3 fps from the original instructional video in English (top) and cross-accent lip-synced video (bottom). (b) Each of the 6 images depicts original English video (left) along with its enlarged ROI, (right) shows our generated Hindi lip-synced video.

3.4 Implementation

3.4.1 TD-LSTM

Our proposed TD-LSTM model consists of a single layer LSTM with 60 neurons in the hidden layer, followed by a 40D dense layer. We also up-sample each video clip at 100Hz to compute lip-landmarks. In each forward pass, the network takes 100 time-steps MFCC features (100 x 12D) and predicts the 20th up-sampled lip-landmark frame (1x40D). This is done densely for each Hindi audio-visual clip. This results in an offset in the prediction of lip-landmarks of 200ms at the beginning, and 800ms at the end of the video. We compensate for this by replicating the first and the last frame's predicted landmarks respectively for the appropriate number of frames. We also implemented TD-LSTM with 500ms and 800ms time-delays. But we found very little perceptual difference between the results, and therefore chose 200ms delay. Using Bidirectional TD-LSTM also did not perceptually affect the results.

We implemented the network in Keras deep learning framework [16], with a batch size of 64, mean square loss, and Adam [61] as the choice of optimizer. We trained our network for 20 epochs, with a total time of around 4 hours on Nvidia GTX 1080 Ti, when the loss started plateauing. Pre-training with 10% videos randomly sampled from GRID corpus resulted in faster saturation of loss. As the TD-LSTM learns to encode dubbed audio into lip-landmarks, we observed that it captures the artifacts corresponding to the dubbing artist, such as thickness of lips and span of mouth. Hence we found pre-training with GRID corpus also helps in generalization, in case of different dubbing artists.

3.4.2 U-Net

We use U-Net architecture similar to [64] and trained the it on 4 NVIDIA TitanX GPUs, using a batch size of 16, counting 4 batches per iteration, until ≈ 5000 iterations. This took ≈ 2 seconds per iteration, and occupied ≈ 3.3GB of memory including model weights and images kept in the buffer. As U-Net has been trained on Andrew Ng's lip-landmarks to predict original frame, it also learns an undesirable mapping between jaw location and the shape of lip in the output. This badly affects the generation of lip in the instances where the predicted landmarks correspond to a closed lip while the target frame has mouth open, or vice-versa. To overcome this, we introduce an *intermediate-processing* step between TD-LSTM and U-Net.

3.4.3 Intermediate processing

We normalize the lip-landmarks from all the frames in the target instructional video in English, and group them into 6 clusters. The frames in the target video nearest to the cluster centroids are chosen to represent the clusters, with their landmarks as their new centroids. All the lip-landmarks predicted by TD-LSTM are then assigned to a cluster based in their distances from the new centroids. This allows the predicted lip-landmarks to be assigned appropriate face frames. Only these 6 frames are then fed to the U-Net (after masking the mouth region). This results in a set of generated frames consisting of lip-synced mouth regions but in only 6 distinct facial poses corresponding to the new centroids. This results in a jittery face video with proper lip-synchronization.

3.4.4 Homography computation

The frames generated from the U-Net are slightly blurred, therefore we use a pre-trained CNN deblur network as in [20], trained on facial images for sharpening. We then compute the pairwise homography between each generated video frame and that of the original instructional video clip using all the 3D face-landmarks predicted using [14], except those corresponding to eyes and lip region. This gives a transformation matrix between the frame pairs. We then crop a rectangular mouth region in the generated video frames, with its center at the mean of lip-landmarks, and its length twice of that between the mean of lip-landmarks and the landmark corresponding to the tip of the nose. This region of interest

(ROI) is augmented over the original video using the the computed transformation matrix. All these ROI augmented frames along with Hindi dubbing are then used to create the final Hindi lip-synced video.

3.4.5 Evaluation metric

Audio-visual data generation is done for human consumption, hence its evaluation is subjective. Hence, we conducted a user-based study to evaluate the outputs of our cross-accent and cross-language methods. This allowed us to get an average subjective evaluation.

Structural similarity (SSIM) index [102] is a widely used metric to evaluate the quality of generated images and videos. We compare the SSIM index to evaluate the generation of the U-Net.

3.5 Results

3.5.1 User-based study

To check the quality of our proposed models, we asked 20 different people to rate the lip-unsynced video and the generated lip-synced video pairs, for each of 10 Andrew Ng's ML video clips, for each of the two cases of cross-accent and cross-language. The 10 video clip-pairs were of upto 1 minute duration each.

3.5.1.1 Cross-accent lip-sync

For each of the pairs of un-synced dubbed video where the Indian English-accented dubbed audio is naively overlaid on the original English video, and our dynamically lip-synced videos, we randomly selected 5 video pairs for each subject and asked them to rank the videos between 1 (Hard to understand) to 5 (Easy to understand) based on their comfort. As shown in Table 2, users preferred our dynamically programmed lip-synchronized videos.

	US(N)	S(N)	US(F)	S(F)
Dynamic	3.0	**4.6**	1.9	**3.2**

Table 3.2 Mean scores for Dynamic Programming (Dynamic) on Indian-English: for un-synced speech overlay ('US'), and lip-synced version 'S', by naive (N) and familiar (F) users.

3.5.1.2 Cross-language lip-sync

We perform a similar user-based experiment with cross-language lip-synchronized videos — we showed 10 videos, with Hindi audio naively overlaid (un-synced), and Hindi lip-synced video to 20 users. Since comfort is subjective and ill-defined, we asked users to rate the percentage of lip-synchronization perceived by them for each pair. As shown in Table 3, the means of the comfort score and percentage lip-synchronization were higher for our cross-language lip-synced videos. The

average comfort rating across users for each video pair can be seen in Figure 3.8(a), where as average percentage lip-synchronization can be seen in Figure 3.8(b). We also show qualitative results of cross-accent and cross-language lip-synchronization in Figure 3.7(a) and (b) respectively.

	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

Table 3.3 Mean scores and standard deviation for Cross-language lip-sync on Hindi: (C) comfort level for (US) un-synced speech overlay, and (S) lip-synced version; (LS%) Lip-Sync percentage for (US) un-synced and (S) lip-synced versions.

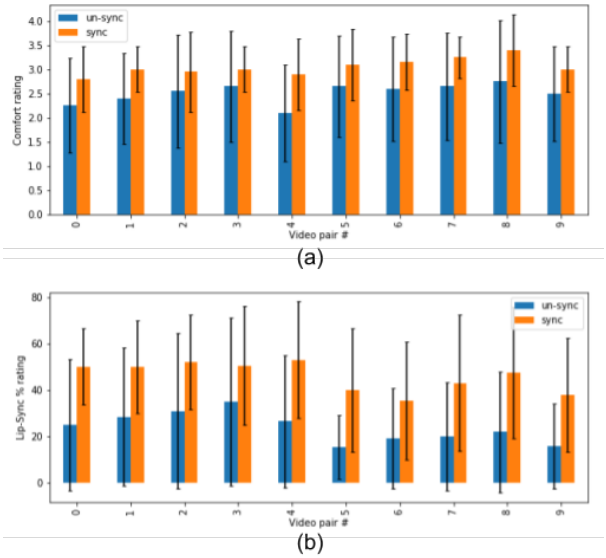

Figure 3.8 User feedback for cross-language lip-sync corresponding to 10 video pairs - (a) shows average comfort rating and its standard deviation for lip-unsynced (blue) and lip-synced videos (orange), (b) shows average percentage of perceived lip-synchronization and its standard deviation for lip-unsynced (blue) and lip-synced videos (orange).

3.5.2 Quality of generation

We compare SSIM index to evaluate the frame quality in original English video and its Hindi lip-synced version for each of the 10 video pairs. The average SSIM index across all the generated frames w.r.t the frames in the original videos was 0.98, with an overall standard deviation of 0.01. To evaluate the generation output of our cross-language model, we also computed SSIM scores for the 4 datasets we used to train U-Net. The average SSIM score for each of these dataset can be seen in Figure 3.9, with mean average SSIM score for all the dataset to be 0.58 with the overall standard deviation of 0.05.

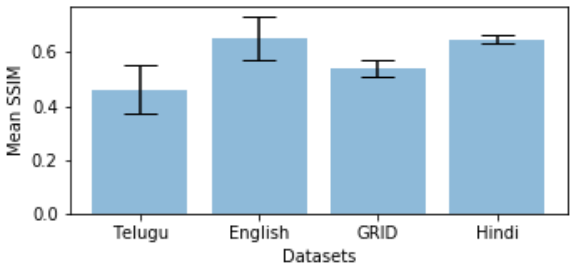

Figure 3.9 Mean and standard deviation of SSIM scores for various datasets used to train U-Net

3.6 Discussion

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.

3.7 Summary

We propose two different lip-synchronization methods for educational videos for same language with different accents, and two different languages to improve instructor-student engagement during online video lectures. We detail our pipelines for dataset creation for audio-to-lip-landmarks as well as lip-landmarks-to-mouth-generation. Our user-based-study shows that lip-synchronization can improve effectiveness of content delivery through dubbed speech videos.

Chapter 4

Towards accurate lip-landmark localization

Many applications in the lip-space like lipreading and lip-motion synthesis require localization of region of interest i.e. to model the lips. A better lip-landmark localization method, may therefore improve the performance of these systems. Current methods of lip-localization assumes lip-landmarks to be a subset of facial-landmarks which might lead to sub-optimal optimization for localizing lip-landmarks.

In this work we propose a solution for better 2D lip-landmark localization. We investigate the current approach of facial-landmark localization and adapt it for lip setting. Just by changing the loss formulation to account for lip region we find a boost in performance. Our method performs better in comparison to state-of-the-art stacked hourglass network giving resulting in better performance. For performing this comparison, we formulate a novel metric similar to that used for 2D facial landmark localization task. Finally we show the performance of our method on 300W faces and 300VW faces dataset and compare it against the baseline approach.

4.1 Introduction

Faces have been at the center of many Computer vision applications — emotion recognition, lipreading, surveillance, enactment and many more. Many of these applications uses some form of facial key-point localization to register face.

One one hand we find a large number of work addressing key-point localization for Human pose estimation and facial landmark prediction, on the other hand lip-landmark is relatively untouched domain. Lip being elastic in nature can present a number of pose variation. These different lip-poses during speech are called viseme. Moreover mouth region's location keeps on changing because of change in head pose.

While most of the current literature don't explicitly localize points on the lips. They, however, optimize for overall facial landmarks with lip making a subset of these.

Advent of deep neural network gave way to better function learning methods. These networks like convolutional neural network (CNN) [63] and recurrent neural network (RNN) [46] are able to learn

complex structural relationship from data. As deep learning is a data driven approach, it requires humongous amount of data to build generalized systems.

Curating a large scale annotated dataset of lip-landmarks requires careful annotation of lip contours. It is often difficult to correctly annotate lips due to different facial features, skin tones and presence of facial hair and other possible occlusions. This makes the dataset curation an expensive and time consuming process.

All these challenges motivated us to work towards solving for lip-landmark localization problem. In this work, our contributions are as follows:

1. We propose a new loss-function on an existing facial-landmark localization architecture, i.e. stacked hourglass network [14], to explicitly optimize for lips.

2. We also introduce a new evaluation metric for lip-landmarks similar to normalized mean square method used in stacked hourglass network.

We show the performance of our network on 300W [84] and 300VW [85] datasets and compare the results with the existing baseline method used for facial landmark detection [14]. We perform a comparative study between the performances of lip and facial landmark models and report our results with qualitative analysis.

Finally we also discuss the dataset curation challenges and provide a possible thread for future work.

4.2 Related work

Lip-landmark detection problem is similar to key-point detection problem in facial domain and human pose estimation. In this section we will present the related work done in these domain in addition to the available work in the field of lip or mouth detection and lip-landmark localization.

Human pose estimation as a tree based problem has been attempted to solve in the earlier computer vision literature [73] with solution like pictorial structure (PS) by Fischler et al [32], and deformable part model (DPM) by Felzenszwalb and Huttenlocher [27]. Later work by Yang and Ramanan [105] posed the problem of landmark localization as a flexible mixture of template. Another similar work by Johnson and Everingham [55] propose cascaded body part detection using a mixture of pictorial structures.

With the inception of deep convolutional network came better landmark detection models. Specially CNNs which paved the way for better classification and detection problems has been used for regressing landmarks. Toshev and Szegedy [93] used a CNN based regressor to recursively predict one landmark at a time conditioned on input image and previously predicted landmarks. Pfister et al [79] uses convnets for estimating human pose by exploiting the temporality across the frames in gesture videos. Bulat and Tzimiropoulos [14] proposes a two-step convolutional heatmap regressor based solution to address this problem. One of the recent work in the domain is stacked hourglass network proposed by Newell et al

[74]. Each hourglass unit in this network refines the heatmap prediction coming from the previous unit and the final unit's output is used during inference.

Since previous 2 decades, various solutions has been proposed addressing the ch alleges in the domain of mount detection. Some of the earliest work in this domain relied on RGB pixel values for computing low level features to detect mouth [77] in the face images. Viola and Jones [96] proposed boosted cascaded classifier to detect faces in the scene. Later a similar method was used by Lienhart et al [68] detect mouth region.

Ahonen et al [3] proposed the use of local binary patterns (LBP) for face detection, which was later used by Rodriguez [82] to show face verification. Kazemi and Josephine uses an ensemble of boosted regression trees to iteratively align an initial average face estimate on a target face.

Since the beginning of deep learning era CNNs has been predominantly used for facial landmark localization. Sun et al [90] uses cascade of convolutional networks to performance coarse-to-fine prediction of key-points. Zhang et al [107] proposed a task oriented deep model for joint prediction of facial landmarks and attributes. Mnemonic descent proposed by Trigeorgis [94] uses RNNs to iteratively refine the initial facial landmark prior.

Most similar to our work is the use of stacked hourglass network for facial landmark localization by Bulat and Tzimiropoulos [14]. They adapt the stacked hourglass network proposed by Newell et al [74] for predicting lip-landmarks. They predict 68 facial landmarks including 20 landmarks associate with lips or mount region. Our method is different from that in the terms of loss function which explicitly optimized for lip-landmarks instead of face landmarks.

4.3 Method

In this section we discuss in detail our proposed architecture, the loss function used, dataset details, a novel metric for performance evaluation and the implementation details to develop a key-point localization solution for lips.

4.3.1 Stacked hourglass network

Stack hourglass network used by Bulat and Tzimiropoulos [14], i.e. 2D face alignment network or 2D-FAN consist of four hourglass networks where each hourglass network improves the performance of previous hourglass network. The input of this network is an RGB image while the output is the 68 channels two-dimensional (2D) Gaussian distribution with mean representing the actual location of the key-point. Each of these 68 channels is regressed for one landmark each out of total 68 facial landmarks. We employ the a similar architecture for lip-localization task. Since the number of landmarks corresponding to lip are 20, we replace the last convolution layer with another convolution network with 20 filers instead of 68. Hence the loss is computer for 20 channels instead of 68 for each sample. We use the same hourglass architecture as that by Newell et al [74] and use a stack of 4 hourglass in our final model. We will refer to this network as LAN. Figure 4.1 shows the network architecture of LAN.

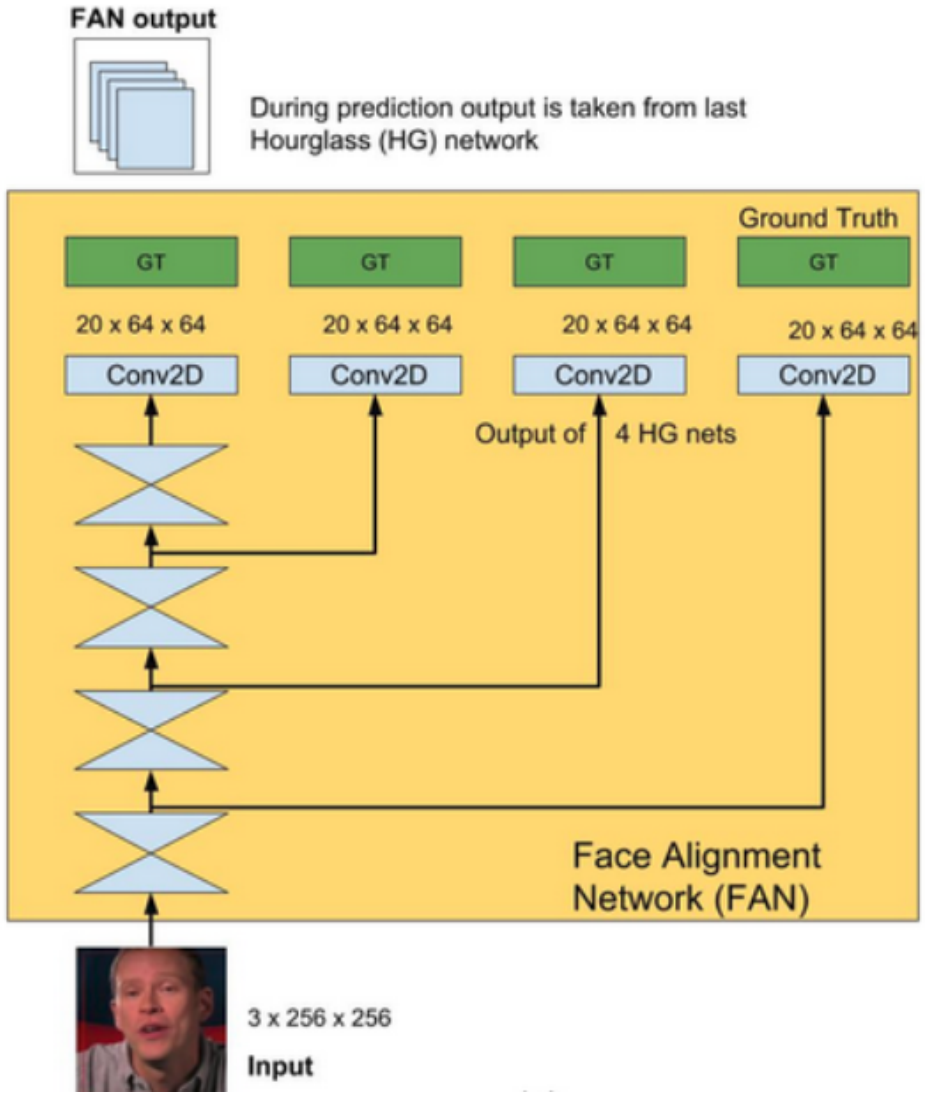

Figure 4.1 Proposed network for lip alignment network (LAN): The basic architecture is similar to FAN except the last layer output has 20 channels, one for each lip-landmarks.

4.3.2 Datasets

As we are only interested in lip-landmarks, the facial-landmarks dataset with lip annotations can be used to train out network. We use 300W-LP dataset [109], for training our final model. We also use this dataset for validation. The dataset has a total of 61,225 with face pose variation between -90°. to +90°. The samples are annotated with 68 facial landmark points.

For testing purpose, we use 300W [84] and 300VW [85] datasets. 300W has 4000 samples images with face pose variation between -45°. to +45°. It contains images in both indoor and outdoor conditions. For 300VW dataset there are 218,595 sample frames with with face pose variation between -45°.

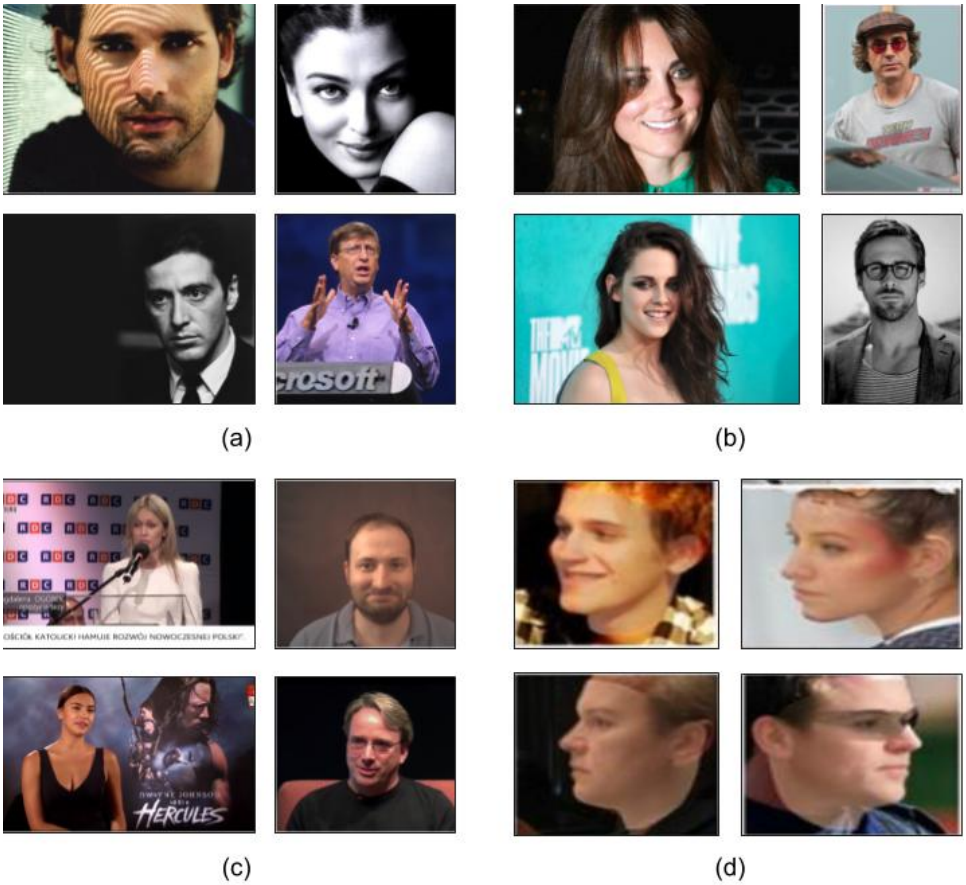

Figure 4.2 Randomly selected images from used datasets: (a) 300W Indoor , (b) 300W Outdoor [84], (c) 300VW [85] and (d) 300W-LP [109]

to +45°. In Both of these dataset, each image has 68 key-points as facial landmark annotations. Figure 4.2 shows few samples from each of these datasets.

4.3.3 Metric

The metric used for evaluating 2D face alignment network (2D-FAN) is normalized mean error((NME). It is computed as the average euclidean distance between predicted landmarks and the ground truth, normalized by area of face bounding box. The face area may not be the appropriate denominator for fine-grained lip evaluation, as it is much greater than the area of lip region of interest (ROI). We define a new evaluation metric 'lip-normalized' mean error (L-NME). Mathematically, L-NME can be defined as:

$$\text{L-NME} = \frac{1}{N} \sum_{i} \frac{\|x_i - y_i\|_2}{l_{bb}} \tag{4.1}$$

where x_i denotes the i^{th} ground truth lip-landmark, where as y_i denotes the i^{th} predicted lip-landmark. N is the total number of landmarks on the lips i.e. 20; and l_{bb} is the lip bounding-box area.

We compare NME with L-NME for lip-landmark localization using 2D-FAN predictions. We found that NME because of larger denominator causes the overall error to minimize while L-NME having smaller denominator causes the overall error to become large, with the same numerator, which is desirable for evaluating fine-grained landmark-localization. In figure 4.3 shows the comparison between NME and L-NME metric for 300W dataset. We can see that, for NME the error converges to maximum lot early, making it unsuitable metric for lip-landmarks localization.

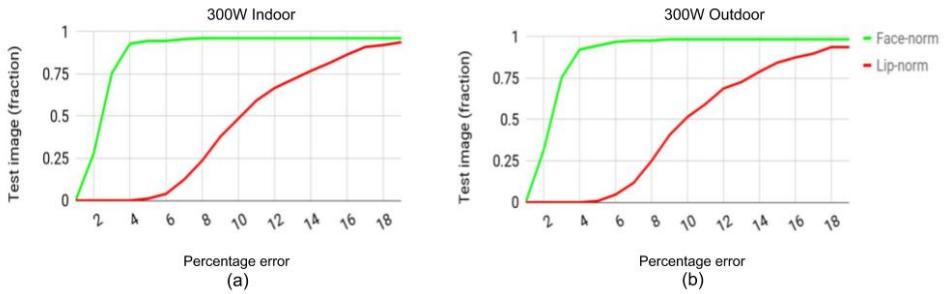

Figure 4.3 Comparison between NME and L-NME: On x-axis we have percentage test population, on y-axis we have percentage error. (a) shows percentage of total test population falling under certain percentage of error, where green curve denotes NME error and red curve denotes L-NME error for 300W Indoor dataset. Similarly, (b) represent the same for 300W Outdoor dataset.

4.3.4 Implementation details

Preprocessing

The input to our model is RGB images of resolution 256×256. For each input image we first search for faces in them using dlib [60] library. We then crop each face and resize them to 256×256 resolution. For each such new face image we normalize the image between 0 and 1, and subtract 0.5. This normalized $3 \times 256 \times 256$ becomes the input to our network. It should be noted that dlib library also provide facial landmark, but we don't use them at all, neither for face detection nor any initialization.

Training

We use 4 hourglass network as our final architecture, the last layer is a 2D convolutional layer with 20 filters each for one landmark. We implement the network in Keras library [16]. The loss used is mean square error (mse) and our choice of optimizer is Adam. We train the network for 6 epochs after which the loss starts saturating.

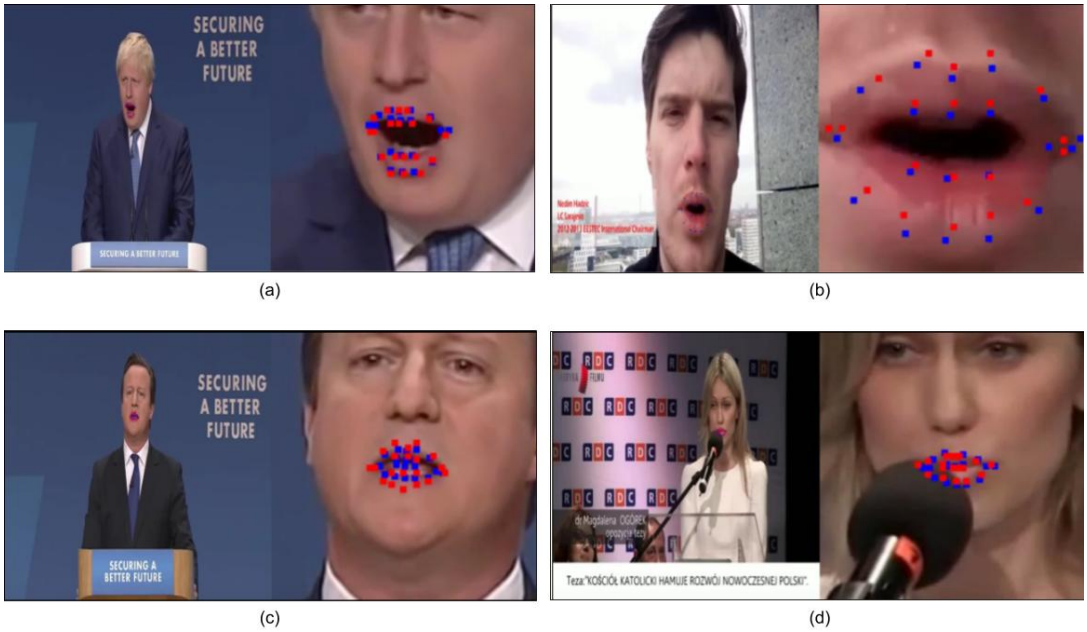

Figure 4.4 Qualitative results for LAN on 300VW dataset: (a-d) shows LAN lip-localization results and the ROI of lip region on randomly selected frames from 300VW dataset. Blue points denotes the ground truth annotations while red dots are predicted landmarks. It can be seen that proposed method is able to predict lip-landmarks for occluded lip region in image (d)

4.4 Results

4.4.1 Comparison between LAN and FAN

One of the baseline for comparison is lip-landmarks predicted through FAN [14]. We predict the lip-landmarks using both and LAN and FAN on 300W dataset. Figure 4.5 shows the quantitative comparison between our LAN and baseline FAN.

It can be seen from Figure 4.5 that our network performs better than baseline for lip-landmark localization task. Qualitative results on randomly selected frames from 300VW dataset can be seen in figure 4.4.

4.4.2 Discussion

From figure 4.2 and 4.4, it can be seen that faces comprises of very small region of the scene. While lip in itself is a small part of face, it can be very difficult for the annotators to annotate the key-points on the lips. At such a small scale effect of lighting, skin stretching, facial hair can significantly deteriorate the annotation quality. Hence, a more robust solution is require for lip-representation.

While lip-landmark annotation can be very expensive, annotating speech is relatively cheaper. Moreover, availability of large scale lipreading datasets like LRW [17] and LRS [19], makes it easier to exploit word and sentence level annotation of speech videos in multi-task setting to improve lip-landmark local-

ization. Hence, use of language as a week supervision for lip-landmark localization may be a potential future line of research.

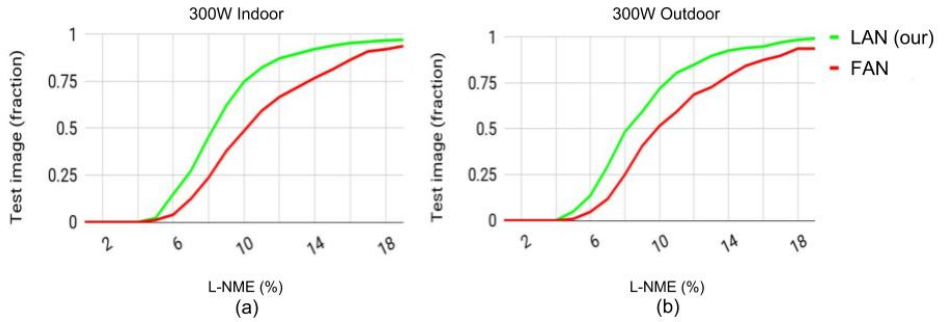

Figure 4.5 Comparison between performance of Lip alignment network (LAN) and face alignment network (FAN) on 300W dataset using L-NME metric: On x-axis we have percentage test population, on y-axis we have percentage L-NME. (a) shows percentage of total test population falling under certain percentage of L-NME, where green curve denotes lip-landmarks predicted using our proposed LAN, and red denoted lip-landmarks predicted using FAN on 300W Indoor dataset. Similarly, (b) represent the same for 300W Outdoor dataset.

4.5 Summary

We proposed an improvement on an existing key-point localization network to adapt it for lip-landmark localization task. We proposed a novel lip-normalized mean error metric suitable for fine-grained lip-landmark localization task. Our proposed lip alignment network (LAN) improves the lip-landmarks prediction of baseline face alignment network (FAN) on 300W dataset. Finally we show the qualitative results on 300VW dataset. We believe our network can be used to improve the performance of the lip key-point centric applications.

Chapter 5

Summary, Conclusions, and Future Directions

Here we summarize this thesis, derive conclusions and discuss about possible new research threads in the domain of audio-visual speech synthesis.

5.1 Summary

We started with motivating the reader about general problems in the domain of visual speech recognition. We illustrated the mechanism of speech generation in human, discussed about phonemes, viseme and their relationships. We also discussed how this many to one phoneme-viseme relation makes the problem of visual speech recognition non-trivial. Audio-visual speech recognition and synthesis while having a huge potential in various speech application host a variety of problems and subproblems. We narrowed down our focus to two main problems in this domain: lipreading and visual speech synthesis. After formally defining our problem statements and corresponding challenges, we briefly described our contributions and proposed solutions for the same. We also introduced reader to the related work in this space based on which we proposed our solutions. We also briefly presented major datasets available in this domain.

Subsequent chapters in this thesis, presented our detailed approach towards addressing the defined two problems, with chapter 2 presenting a word spotting pipeline for spotting words in silent speech videos and chapter 3 presenting our proposed solution for visually dubbing speech videos across different accents and languages. In chapter 2, we started with discussing general problems associated with discriminatively classifying lip-motion into words. As most of the current work presents lip-reading solutions which are vocabulary bound for small set of words, those solutions may not be scalable lacking to generalize for new words. We proposed a method of spotting spoken words in silent lip-videos, to address these issues. Then we discussed the related work in this domain and compared them with our proposed method. The following section in chapter 2 explained the intricacies and details of the pipeline. The performance of the pipeline is evaluated against the state-of-the-art architecture for word spotting task on three matrices recall at k, precision at k and mean average precision. We tested our method for domain-invariance on GRID corpus [21] dataset. We also proposed two extension over our baseline method, i.e. query expansion using pseudo relevance feedback and reranking. The chapter

was concluded with qualitative results and related graphs on LRW [17] dataset and an archaic speech video. Chapter 3, started with a note on importance of speech videos in education sector, and how Instructional videos are changing the paradigm of conventional classroom education. The challenges of adapting these videos for different linguistic groups were discussed. The conventional procedure of audio-dubbing, in multimedia post processing was also explained. Since audio-redubbed videos does synchronize lips of the speaker in the dubbing language, it looks unnatural and may hamper viewing experience. Here we formally proposed two lip-synchronization methods: cross-accent lip-synchronization for synchronizing lips of a speaker when the dubbed audio contained a different accent in the same language, and cross-language lip-synchronization when the dubbed audio is in a different language altogether. We then, discussed about the previous works in this domain, drawing inspiration and comparing our work with them. The next section in this chapter detailed on implementation details of the two-methods: dynamic programming for cross-accent synchronization and TD-LSTM based UNet architecture for cross-language synchronization. We then proposed a scalable dataset curation pipeline to train our models. We showed the performance of our network on Andrew Ngs ML tutorial videos, showing qualitative performance with both positive and negative results, and finally conducted human-based studies to evaluate the performance of methods again the unsynchronized counterparts. We concluded with a discussion on the shortcomings in our proposed methods and a possible future direction.

In chapter 4, we started with introducing readers with the novelty of the problem of lip-landmark localization and its challenges. This is followed by a detailed literature survey of the work in the domain of key-point detection — human pose estimation, facial-landmark localization and lip detection. In the subsequent section, we elaborated the details of the proposed networks along with the implementation details. We then proposed a new metric for the evaluation of lip-landmark localization. We also reported the performance of our proposed networks and compared them with the baseline. Finally, the chapter concludes with a discussion on challenges pertaining to lip-landmark dataset curation and a possible direction for future research in this domain.

5.2 Conclusion

In this thesis, we used a recognition-based model to show recognition-free retrieval thereby spotting spoken words in silent videos, without explicitly recognizing the words. Our method improved the state-of-the-art lip word spotting baseline by more than 35% on one of the largest lipreading datasets [17]. Application of our method on Charlie Chaplin ‘the great dictator, encourages its applicability on silent archaic videos. Our recognition-free word spotting method is also robust against change in domain when compared with its recognition based counterparts. Since words are not explicitly recognized rather queried in the feature space, this query by exemplar technique allows change in domain, in terms of accents, video quality and presence of noise. This might also be useful for spotting out-of-vocabulary words, we showed this through spotting words in GRID corpus [21].

Through our subsequent work, we investigated the use of dubbed audio for lip-synchronization. This work is essentially an inverse problem of our previous work, where we try to comprehend the lip-motion. Through our proposed method we lip-synchronized Andrew Ngs ML tutorial videos, with Hindi audios, which were preferred over unsynchronized counterparts in the human-based evaluation. Our motivation is to extend this work to make English Movies and TV shows more immersive to the native audience, in non-English speaking countries.

Finally we addressed the challenging problem of accurate lip-landmark localization. Our proposed solution results in stable lip-landmarks, improving the current baseline on 300W and 300VW datasets. Our landmark localization methods might result in improving the performance of lip-modelling techniques in the applications like lipreading, emotion classification, digital avatars etc.

We hope that this thesis would help the readers in appreciating various challenges and possibilities in the field of audio-visual speech modelling, and in driving future line of research in this direction.

5.3 Future directions

Through the course of works presented in this thesis, we discovered few challenges present in the domain which can be further discussed and worked on. Below, we present few possible direction for future research based on this thesis:

Audio speech generation for silent lip videos:

Generate audio speech from silent lip videos, with realistic voice impersonation of the speaker. Speech in terms of audio output is more practically useful and easily accessible for conversational application. Moreover training lipreading model against an audio output is less expensive than first annotating frames of lip video and training, as most of the modern camera has inbuilt microphones and generally video clips contains corresponding sound clip.

Vid2Speech [25] does the same using GRID corpus dataset[21] by using 2D convolutional network(CNN) which takes each frame of fixed length the silent lip video as one seperate channel. The last layer of this CNN is of size 18 which corresponds to the size of the sound representation vectors from Linear Predictive Coding (LPC). The network is trained with backpropagation using mean squared error (MSE) loss. After predicting 18D audio vector they use audio synthesizer. Though the output is audio but it doesn't produce same phonetic characteristics of the speaker. Recent work done in the mobile app 'Lyrebird' [69], the developers has been able to generate the audio voice of any speaker. Hence these recent two developments in the field has a greater potential for future use of lipreading to generate realistic voice impersonation of any individual.

Visual speech question answering:

Voice assisted technologies have seen an increased demand in past few years, allowing people to perform intelligible conversation with voice-based assistants. Such technology thrive mostly in home

and office environment where the presence of external noise in the background is minimum. To adapt these technology for industrial setting would require more robust solution. Vision, particularly live lip-videos can be a possible modality. In this thesis we have shown word spotting for lip-videos, a similar system for spotting sentences would require long term sequence modelling of lip-representations, reasoning and understanding of language semantics.

Visual speech summarization:

In chapter 3, we investigated conditional lip-synchronization, a possible lead to this project can be visual speech summarization. Given a theme, finding only relevant sections from a big speech and stitching them without aberrant changes in the new video. Some attempts has already be done for video summarization like Otani et al. [75] but not in Lipreading domain. While working on the problem of cross-language lip synchronization, we realized that UNet memorizes the viseme from the overall jaw position, and required us to design an intermediate clustering stage. Similar measures are also taken by Suwajanakorn et al [91]. Formulation of simpler generation methods might yield more robust and easily scalable lip-synthesis solutions.

We would like to conclude this thesis with a possible application direction hoping that our work will inspire future readers in working towards better audio-visual speech solutions and applying them for societal good.

Aid to hearing impaired:

Lipreading can also host novel applications for hearing impaired people, by assisting them in communication. India has about 1.1 million hearing impaired with literacy rate about 2% among them (Census of India states it to be 1.2 million with about 25% overall population under primary level of education). Moreover, Indian Sign Language (ISL) is in English which is not a major language among rural population. A mobile application for hearing impaired people, which can facilitate two way communication between hearing impaired and normal person in local languages can directly cater to the needs of this section of society.

Related Publications

Journal

1. **Abhishek Jha**, Vinay Namboodiri and C.V. Jawahar, Spotting Words in Real World Videos : A Retrieval based approach, Journal of Machine Vision and Applications (MVA), Springer, February, 2019.

Conference publications

1. **Abhishek Jha**, Vinay Namboodiri and C.V. Jawahar, Word Spotting in Silent Lip Videos, IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, CA, USA, 2018.

2. **Abhishek Jha***, Vikram Voleti*, Vinay Namboodiri and C.V. Jawahar, Lip-Synchronization for Dubbed Instructional Videos, Fine-grained Instructional Video undERstanding (FIVER), CVPR Workshop, Salt Lake City, Utah, USA, 2018. (*equal contribution)

3. **Abhishek Jha**, Vikram Voleti, Vinay Namboodiri and C.V. Jawahar, Cross-Language Speech Dependent Lip-Synchronization, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019.

Other publications during MS which are not part of this thesis

1. Yashaswi Verma, **Abhishek Jha** and C. V. Jawahar, Cross-specificity: modelling data semantics for cross-modal matching and retrieval, International Journal of Multimedia Information Retrieval, Springer, June 2018.

Bibliography

[1] Machine Learning Course, Andrew Ng. https://www.deeplearning.ai.

[2] WEBRTC, VAD. https://webrtc.org/.

[3] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE TPAMI*, (12):2037–2041, 2006.

[4] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*. IEEE, 2012.

[5] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas. Lipnet: Sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016.

[6] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[7] S. Basu, N. Oliver, and A. Pentland. 3d modeling and tracking of human lip motions. In *ICCV*. IEEE, 1998.

[8] M. Borga. Canonical correlation: a tutorial. 2001.

[9] H. A. Bourlard and N. Morgan. *Connectionist speech recognition: a hybrid approach*, volume 247. Springer Science & Business Media, 2012.

[10] G. Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11), 2000.

[11] M. Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 21–28. ACM Press/Addison-Wesley Publishing Co., 1999.

[12] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360. ACM Press/Addison-Wesley Publishing Co., 1997.

[13] S. S. Brooke N.M. Pca image coding schemes and visual speech intelligibility. In *Proceedings of the Institute of Acoustics*, volume 16, 1994.

[14] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *ICCV*, pages 1021–1030. IEEE, 2017.

[15] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.

[16] F. Chollet et al. Keras, 2015.

[17] J. S. Chung and A. Zisserman. Lip reading in the wild. In *ACCV*, pages 87–103. Springer, 2016.

[18] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *ACCV*, pages 251–263. Springer, 2016.

[19] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In *CVPR*, pages 3444–3453. IEEE, 2016.

[20] J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017.

[21] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5), 2006.

[22] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. 2009.

[23] P. Doetsch, M. Kozielski, and H. Ney. Fast and robust training of recurrent neural networks for offline handwriting recognition. In *ICFHR*, 2014.

[24] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[25] A. Ephrat and S. Peleg. Vid2speech: speech reconstruction from silent video. In *ICASSP*, pages 5095–5099. IEEE, 2017.

[26] T. Ezzat and T. Poggio. Miketalk: A talking facial display based on morphing visemes. In *Computer Animation*, pages 96–102. IEEE, 1998.

[27] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61 (1):55–79, 2005.

[28] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *ECCV*, pages 242–256. Springer, 2004.

[29] S. Fernández, A. Graves, and J. Schmidhuber. An application of recurrent neural networks to discriminative keyword spotting. *ICANN*, 2007.

[30] A. Fischer, A. Keller, V. Frinken, and H. Bunke. HMM-based word spotting in handwritten documents using subword models. In *ICMR*, 2010.

[31] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 1981.

[32] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on computers*, 100(1):67–92, 1973.

[33] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke. A novel word spotting method based on recurrent neural networks. *IEEE TPAMI*, 34(2), 2012.

[34] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. 1999.

[35] A. P. Giotis, G. Sfikas, B. Gatos, and C. Nikou. A survey of document image word spotting techniques. *Pattern Recognition*, 68, 2017.

[36] H. Gish and K. Ng. A segmental speech model with applications to word spotting. In *ICASSP*, volume 2. IEEE, 1993.

[37] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[38] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *NEURAL NETWORKS*, pages 5–6, 2005.

[39] A. Graves, S. Fernández, and J. Schmidhuber. Bidirectional LSTM networks for improved phoneme classification and recognition. *ICANN*, 2005.

[40] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006.

[41] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, pages 1735–1742. IEEE, 2006.

[42] A. B. Hassanat. Visual words for automatic lip-reading. *arXiv preprint arXiv:1409.6689*, 2014.

[43] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE, 2016.

[44] M. E. Hennecke. Audio-visual speech recognition: Preprocessing, learning and sensory integration. *PhD thesis, Stanford Univ.*, 1997.

[45] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6), 2012.

[46] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[47] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.

[48] K. S. Hone and G. R. El Said. Exploring the factors affecting mooc retention: A survey study. *Computers & Education*, 98:157–168, 2016.

[49] R. Hong, M. Wang, M. Xu, S. Yan, and T.-S. Chua. Dynamic captioning: video accessibility enhancement for hearing impairment. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 421–430. ACM, 2010.

[50] Y. Hu, J. Kautz, Y. Yu, and W. Wang. Speaker-following video subtitles. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(2):32, 2015.

[51] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arxiv*, 2016.

[52] H. Izadinia, I. Saleemi, and M. Shah. Multimodal analysis for identification and segmentation of moving-sounding objects. *IEEE Transactions on Multimedia*, 15(2):378–390, 2013.

[53] A. Jha, V. P. Namboodiri, and C. Jawahar. Word spotting in silent lip videos. In *WACV*, pages 150–159. IEEE, 2018.

[54] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE TPAMI*, 35(1), 2013.

[55] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, pages 1465–1472. IEEE, 2011.

[56] P. Kakumanu, R. Gutierrez-Osuna, A. Esposito, R. Bryll, A. Goshtasby, and O. Garcia. Speech driven facial animation. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–5. ACM, 2001.

[57] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):94, 2017.

[58] J. Keshet, D. Grangier, and S. Bengio. Discriminative keyword spotting. *Speech Communication*, 51(4), 2009.

[59] R. Kilborn. Speak my language': current attitudes to television subtitling and dubbing. *Media, culture & society*, 15(4):641–660, 1993.

[60] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10 (Jul):1755–1758, 2009.

[61] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[62] P. Krishnan and C. V. Jawahar. Bringing semantics in word image retrieval. In *ICDAR*, 2013.

[63] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

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

[65] J.-S. Lee and C. H. Park. Robust audio-visual speech recognition based on late integration. *IEEE TMM*, 10(5), 2008.

[66] K. P. Lesner S.A. Visual vowel and diphthong perception across speakers. *Journal of the Academy of Rehabiitative Audiology 14*, pages 252–258, 1981.

[67] P. Lieberman, S. Fecteau, H. Théoret, R. R. Garcia, F. Aboitiz, A. MacLarnon, R. Melrose, T. Riede, I. Tattersall, and P. Lieberman. The evolution of human speech: Its anatomical and neural bases. *Current Anthropology*, 48(1):39–66, 2007.

[68] R. Lienhart, L. Liang, and A. Kuranov. A detector tree of boosted classifiers for real-time object detection and tracking. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 2, pages II–277. IEEE, 2003.

[69] Lydebird. www.lydebird.ai.

[70] R. Manmatha, C. Han, and E. M. Riseman. Word spotting: A new approach to indexing handwriting. In *CVPR*, pages 631–637. IEEE, 1996.

[71] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746, 1976.

[72] G. Monaci. Towards real-time audiovisual speaker localization. In *Signal Processing Conference, 2011 19th European*, pages 1055–1059. IEEE, 2011.

[73] R. Nevatia and T. O. Binford. Description and recognition of curved objects. *Artificial Intelligence*, 8(1):77–98, 1977.

[74] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016.

[75] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, and N. Yokoya. Video summarization using deep semantic features. In *ACCV*, pages 361–377. Springer, 2016.

[76] E. Owens and B. Blazek. Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech, Language, and Hearing Research*, 28(3):381–393, 1985.

[77] M. Pantic, M. Tomc, and L. J. Rothkrantz. A hybrid approach to mouth features detection. In *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, volume 2, pages 1188–1193. IEEE, 2001.

[78] S. Petridis and M. Pantic. Deep complementary bottleneck features for visual speech recognition. In *ICASSP*. IEEE, 2016.

[79] T. Pfister, K. Simonyan, J. Charles, and A. Zisserman. Deep convolutional neural networks for efficient pose estimation in gesture videos. In *ACCV*, pages 538–552. Springer, 2014.

[80] H. X. Pham, Y. Wang, and V. Pavlovic. End-to-end learning for 3d facial animation from raw waveforms of speech. *arXiv preprint arXiv:1710.00920*, 2017.

[81] J. S. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan. Look, listen and learn-a multimodal lstm for speaker identification. 2016.

[82] Y. Rodriguez. Face detection and verification using local binary patterns. Technical report, Ecole Polytechnique Fédérale de Lausanne, 2006.

[83] J. R. Rohlicek, W. Russell, S. Roukos, and H. Gish. Continuous hidden markov modeling for speaker-independent word spotting. In *ICASSP*, pages 627–630. IEEE, 1989.

[84] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *CVPR Workshops*, pages 896–903. IEEE, 2013.

[85] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *ICCV Workshops*, pages 50–58. IEEE, 2015.

[86] T. Stafylakis and G. Tzimiropoulos. Combining residual networks with LSTMs for lipreading. *arXiv preprint arXiv:1703.04105*, 2017.

[87] S. Sudholt and G. A. Fink. Phocnet: A deep convolutional neural network for word spotting in handwritten documents. In *ICFHR*, 2016.

[88] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, 1954.

[89] Q. Summerfield. Lipreading and audio-visual speech perception. *Phil. Trans. R. Soc. Lond. B*, 335(1273):71–78, 1992.

[90] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, pages 3476–3483. IEEE, 2013.

[91] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):95, 2017.

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

[93] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, pages 1653–1660. IEEE, 2014.

[94] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, pages 4177–4187. IEEE, 2016.

[95] S. S. Tsai, D. Chen, G. Takacs, V. Chandrasekhar, R. Vedantham, R. Grzeszczuk, and B. Girod. Fast geometric re-ranking for image-based retrieval. In *ICIP*. IEEE, 2010.

[96] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, volume 1, pages 511–518. IEEE, 2001.

[97] M. Wand, J. Koutník, and J. Schmidhuber. Lipreading with long short-term memory. In *ICASSP*, pages 6115–6119. IEEE, 2016.

[98] A. Wang, M. Emmi, and P. Faloutsos. Assembling an expressive facial animation system. In *Proceedings of the 2007 ACM SIGGRAPH symposium on Video games*, pages 21–26. ACM, 2007.

[99] K. Wang and S. Belongie. Word spotting in the wild. In *ECCV*. Springer, 2010.

[100] L. Wang, X. Qian, W. Han, and F. K. Soong. Synthesizing photo-real talking head via trajectory-guided sample selection. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[101] L. Wang, W. Han, F. K. Soong, and Q. Huo. Text driven 3d photo-realistic talking head. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[102] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, Nov 2003. doi: 10.1109/ACSSC.2003.1292216.

[103] P. Wu, H. Liu, X. Li, T. Fan, and X. Zhang. A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion. *IEEE TMM*, 18(3), 2016.

[104] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*, 2016.

[105] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392. IEEE, 2011.

[106] X.-Y. Zhang, F. Yin, Y.-M. Zhang, C.-L. Liu, and Y. Bengio. Drawing and recognizing chinese characters with recurrent neural network. *IEEE TPAMI*, 2017.

[107] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, pages 94–108. Springer, 2014.

[108] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen. A review of recent advances in visual speech decoding. *Image and vision computing*, 32(9), 2014.

[109] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *CVPR*, pages 146–155. IEEE, 2016.