Enhanced Video Articulation (EVA)—A Lip-Reading Tool

M. Sri Geetha, M. Sujay Sudharshan, S. Sunderesh, P. Aruna, and M. Sanjana Sri

Abstract Lip-reading depends much on the context and the knowledge of language as it does on visual clues and is notoriously challenging. Human lip-reading is a tedious task that requires not only the knowledge about the language but also the visual evidences to approximately or appropriately predict the spoken words. Even the speech recognition in the noisy environment can be made possible through the lip-reading technique. The accuracy achieved through manual lip reading is 40% which can be improvised using AI deep learning techniques. To improve the accuracy of predicting phrases from the low-resolution videos. Enhanced video articulation (EVA) is an articulatory lip-reading technique determining the labial movements of a person. The face of the person is recognised and is segmented for the labia. The process counts on neural networks, AI algorithms comprising many simple computing components coupled together that learn and process information in a way similar to the human brain. As a result, we infer that the deep learning architectures perform similarly to the traditional methods but will improve the word recognition rates.

Keywords Video enhancement · Labial segmentation · Lip reading · Audio-visual speech recognition · SegNet

1 Introduction

Enhanced video articulation (EVA)- a lip reading tool is an AI based tool that is used to predict/generate the words spoken by the speaker from the input low-clarity video. This model can be used to develop a variety of applications, such as a voice replacement for people with laryngectomies. This would also help the people with

M. Sri Geetha (✉) · M. Sujay Sudharshan · S. Sunderesh · P. Aruna · M. Sanjana Sri
Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, Coimbatore—22, India
e-mail: srigeetha.m@srec.ac.in

P. Aruna
e-mail: arunapalaniappan@srec.ac.in

X.-Z. Gao et al. (eds.), *Advances in Computational Intelligence and Communication Technology*, Lecture Notes in Networks and Systems 399,
https://doi.org/10.1007/978-981-16-9756-2_14

hearing disorder to easily recognise the spoken words instead of manual methods of lip reading. EVA does this by converting the movement of the labia to text in real time with enhanced accuracy. This model can also be used to provide quality of experience [1] to the users or the audience of live stream watchers and the online video streaming applications, and the ambiguity and difficulty of the task might be used to replace/overdub genuine speech. Enhancement of the input video does not alter the original size of the file to a greater extent which helps us to overcome the problem of space complexity.

2 Related works on lip-reading

2.1 Video Resolution Enhancement

The study suggested super resolution (SR) methodology that retained more features, resulted in a better super resolved video sequence. Studies on a number of well-known video sequences revealed that the suggested method outperformed both traditional and state-of-the-art approaches.

3 Lip-*R*eading

There are many research works on lip reading using pre-deep learning techniques. Convolutional neural network (CNNs) have been employed in a number of articles to anticipate phonemes or visemes [2–4] from still pictures, rather than recognising entire words or phrases. A viseme is the visual equivalent of a phoneme, "which is the smallest recognisable unit of sound that makes up a spoken word". Petridis et al. train an LSTM [5] classifier using a discrete cosine transform (DCT) and profoundly deep bottleneck features to perceive the entire word.

Wand et al. [6] identified brief words using an LSTM with HOG input features. The persistent use of superficial characteristics is most likely due to a lack of training data in lip reading. However, because the word boundaries must be known ahead of time, the environment is still remote from the real world, as it is with any word-level classification assignment. To determine labelling, it employs a CNN and LSTM-based network and connectionist temporal classification (CTC).

4 Audio–Video Speech Recognition

"Lip reading and audio–visual speech recognition (AVSR) are strongly correlated issues. Using a huge non-public audio–visual dataset, Mroueh et al. conduct phoneme categorization using feedforward deep neural networks (DNNs). Hidden Markovian models [7] works well with manually created or pre-trained visual features and encodes input images using database records; few researchers used discrete cosine transforms (DCT) and pre-trained convolutional neural networks (CNN) to characterise phonemes; and every one of the three joins these elements with HMMs to group spoken digits or segregated words. As with lip reading, there has been little work put into developing AVSR systems that generalise to real-world situations.

Earlier work can be arranged into two. The main kind utilises connectionist fleeting classification(CFC), in which the model predicts outline-by-outline marks prior to deciding the best arrangement between the casing-by-outline forecasts and yield succession. The output labels are not conditioned on each other, which is a flaw. Sequence-to-sequence models, on the other hand, read the entire input sequence before attempting to predict the output phrase."

The above-mentioned papers predicted the spoken word/phrases from a low-resolution video which was a major drawback; this resulted in low test data accuracy of the models.

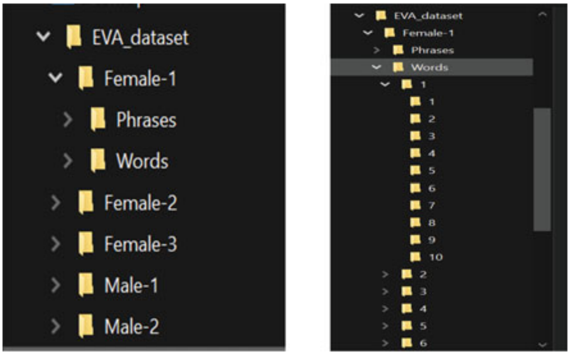
5 Dataset Description

For training and testing of EVA, we have created our own dataset. Unlike MIRACL-VC1 dataset [7, 8], which consists of each ten words and phrases spoken by five men and ten women, our dataset consists of both words and phrases uttered by three women and two men. The speakers were positioned against the camera and were instructed to utter the words and phrases as in the table given below. Each dataset instance is composed of a synchronised sequence of colour photographs. Each words and phrases are further decomposed into frames of each 30 and are arranged in the folders for further processing. The EVA dataset contains a total number of 1250 instances (Fig. 1).

Fig. 1 Description of EVA dataset

ID	Words	ID	Phrases
1	As	1	But I do really know
2	Simple	2	But I didn't like you
3	That	3	But I don't know
4	Know	4	It's that simple
5	No	5	It's simple
6	You	6	As simple as that
7	Really	7	As it is
8	But	8	It's time
9	Well	9	All is well
10	Time	10	All are welcome

Fig. 2 Structure of EVA dataset and detailed structure of the dataset



EVA_dataset
Female-1
Phrases
Words
Female-2
Female-3
Male-1
Male-2

EVA_dataset
Female-1
Phrases
Words
1
1
2
3
4
5
6
7
8
9
10
2
3
4
5
6

The structure of EVA dataset and its detailed structure are given in the Fig. 2. The words and phrases uttered by each of three women and five men are put into their respective hierarchical structure.

6 Proposed model

Enhanced video articulation (EVA) takes low-resolution video with poor or no audio quality as the input. Video super resolution is still an issue which is not solved and is being researched in full capacity. EVA is dissected into three modules, namely video enhancement, labial segmentation, and articulation of the spoken words (Fig. 3).

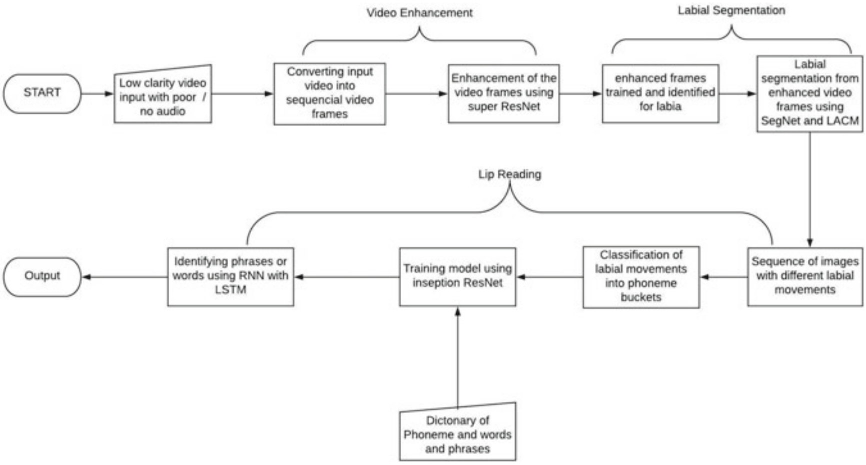

Video Enhancement
Labial Segmentation
START
Low clarity video input with poor / no audio
Converting input video into sequencial video frames
Enhancement of the video frames using super ResNet
enhanced frames trained and identified for labia
Labial segmentation from enhanced video frames using SegNet and LACM
Lip Reading
Output
Identifying phrases or words using RNN with LSTM
Training model using inseption ResNet
Classification of labial movements into phoneme buckets
Sequence of images with different labial movements
Dictonary of Phoneme and words and phrases

Fig. 3 Architecture of EVA (enhanced video articulation)

7 Video Enhancement

There has been great progress in the field of image resolution in which the image is enhanced from low resolution to high resolution using convolutional neural networks (CNN). The first step is to process the low-resolution videos and convert them into sequential frames. To overcome these problems, we came up with a solution which helps us to convert a low-resolution video to a high-resolution video using deep learning neural networks (DLNN). Hence, the frames of the video are put as the input for our model which uses residual network (ResNet) CNN which enhances the resolution of each frame and stores it into another file. This file is later used as an input file for our final part which converts these HR frames to a video. Thus, the proposed model focuses on enhancing the video at lesser computational cost and consuming lesser time as compared to previous trivial models [8]. The transformed high-resolution frames are then fed into the labial segmentation module where the face is detected using Haar classifier and labial part is cropped using mouth points.

Our model is based on super-resolution convolutional neural networks (SRCNN), but after tonnes of research on various research papers [9], we figured that residual network CNN was a better model as it gives a better image quality with the minimum time complexity. This model uses six residual blocks and 2X upscaling which is a refined version of its parent model SRResNet uses 15 residual blocks and 4X upscaling. After training the model, we tested the model which changed the resolution and the quality to a great extent. Figure 4 depicts the work of super-resolution CNN.

This model can be used in many aspects of work like it can be used to convert low-resolution CCTV footage into high-resolution video which might turn out to be a very important evidence to put criminals behind the bars. It can be used in the field of medicine to convert the videos taken to capture the interiors of the human body with tiny cameras (Figs. 5 and 6).

$$Sum = X(C) + X(A) - X(B) - X(D) \tag{1}$$

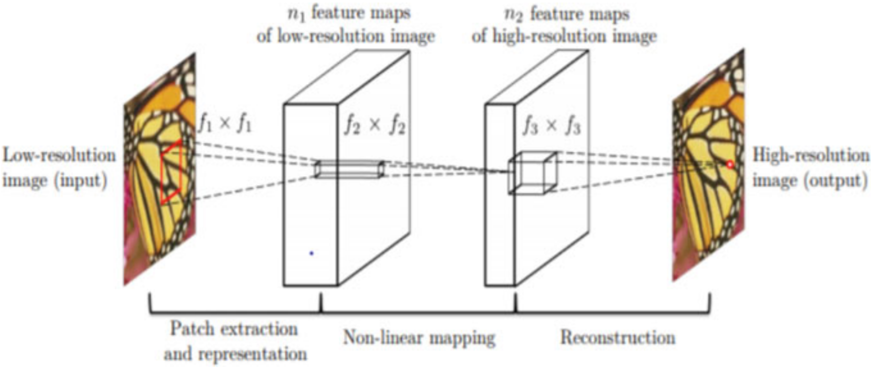


Fig. 4 Super-resolution CNN

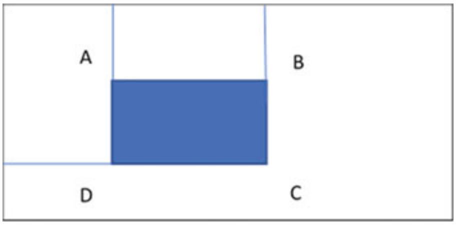


Fig. 5 Finding the sum of the shaded rectangular area

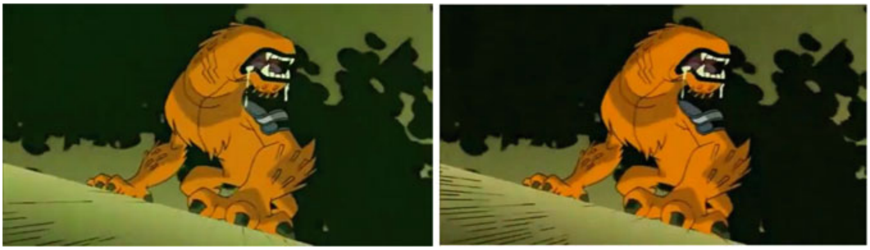


Fig. 6 Input image and enhanced image

where *A*, *B*, *C*, *D* are the points belonging to the integral image *X*.

8 Labial Segmentation

"The precision of the segmentation findings has a direct impact on the identification rate in visual lip-reading systems, hence labial segmentation is crucial. Labial segmentation strategies include the MAP-MRP framework, a clustering method, and an active contour model (ACM). Because of its benefits over traditional image segmentation approaches, the LACM model is adopted. ACM can acquires object boundary precision down to the sub-pixel level, energy minimization and accurate outcomes.

Active contour model (ACM) is a vital part image segmentation and computer vision. ACM is classified as edge-based model and region-based model. An image gradient is frequently used in edge-based models to drive the active contour to migrate towards the object's intended bounds. The suggested models were not employed for labial segmentation since they had an incomplete convergence problem due to weak object boundaries and picture noise. The active contour model's scope is limited to a particular region, which limits the effects of unrelated elements. The localised area-based active contour model (LACM) is proven to be capable of creating superior segmentation results where ACM fails.

The labial segmentation is done with the help of a local active contour model, which generates the initial contour automatically. The evolving curve in LACM divides the local neighbourhoods into two areas: the local region that lays inside and the local region that lays outside. The limited energy for advancing and separating would then be able to be done. However, with LACM, incorrect parameters such as high radius or long-evolving curves cause incorrect prediction. According to studies, the labial is generally of an elliptical region. As a result, the labia can be approximated by various elliptical outlines based on their particular structure. The min-bound oval shape as the underlying developing bend is of critical significance to separate the labial shapes. To track down the base bouncing oval of labial district, the identification of labial corner specks is required. In particular, $X(x, y)$ address a pixel esteem at arrange (x, y), m and n are the most extreme upsides of lines and sections. The left corner, right corner, upper corner, lower corner is signified as L a, L b, V a, and V, individually (Fig. 7).

The steps involved in the extraction or segregation of labial contour are,

- Locate the labial area;
- Acquire the min-bounded oval shape;
- Proceed and develop iteratively;
- Segregate the labial contour area.

SegNet core trainable segmentation engine comprises of networks that actually encodes and decodes and a layer that does the classification pixel-wise. The decoder network maps the low-resolution encoder feature to full input resolution feature. SegNet's distinct quality is that it provides us with the low-resolution feature map that serves as input and it has been decoded too. To conduct non-linear up-sampling, the decoder leverages pooling indices obtained in the matching encoder's max-pooling step."

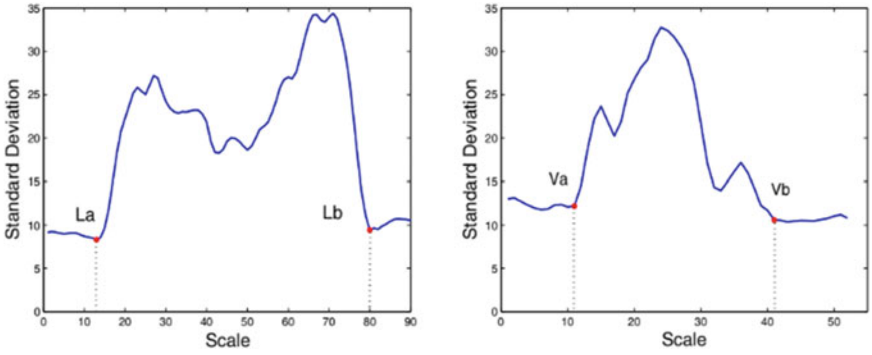


Fig. 7 Standard deviation of column and rows

9 Articulation

The feature-extracted enhanced frames are applied to RNN with LSTM architecture to predict the words in the respective frames. RNN and LSTM, as well as their variations, require time-based sequential processing. Unlike neural turning machines, its design saves all prior representations in memory. This can be inefficient: consider saving the representation of each frame in a movie; most of the time, the representation vector does not vary from frame to frame, thus we are holding an abundance of absolutely similar pictures.

RNN and LSTM necessitate memory-bandwidth-bound computation that will restrict the relevance of neural networks solutions. Long short-term memory (LSTM) with long storage unit acts as an agent that used reinforcement to educate its parameters and to memorise the letters for predicting the term.

The articulation part is structured into three:

- Labial localisation system to locate the labia in the digital input.
- Feature extraction system, which evaluates the appropriate labial features.
- Classification system, which maps feature vectors to terms.

Labial localization: We took an unconventional approach to labial localisation. The primary components were computed using a series of digitised labial pictures. The fixed portions of the image were compressed using these primary components. The labia will be in the region with the least amount of information loss since the major components were fine-tuned for labia.

Feature extraction: A normalised rectangle comprising the labial segment is fed into the feature extractor. For the rectangular input frame, principal component analysis is employed to build a feature vector. A single-layer feedforward neural network is used to estimate the main components.

Translation to words: Words (visemes) must be categorised from the feature extractor's feature vectors. A time-delay neural network was utilised to categorise the feature vectors in order to incorporate temporal features into the classification process.

10 Result

Recurrent neural network with long short-term memory (RNN with LSTM) is used, and the model is fine-tuned with $L2$ regularisers with value 0.1 and 0.5. Optimizer-SGD with learning rate 0.005 is set. The model shows result in 80th epoch with training accuracy as 93% in the end for $L2 = 0.1$. The validation accuracy gets better result (51%) when epoch is 21 and $L2 = 0.1$.

In Fig. 8, Female-I uttered the phrase “But I do really know”. Our model predicted the phrase as “But I don’t really know” (Fig. 9).

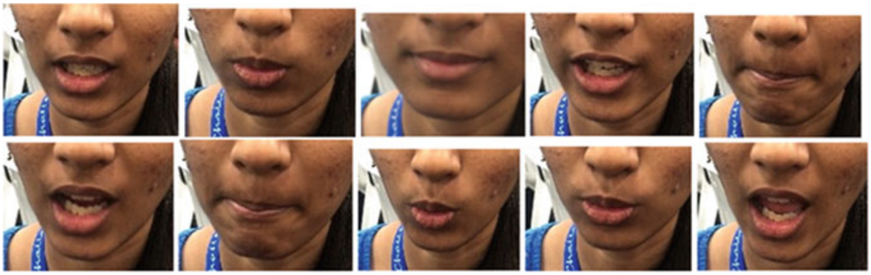
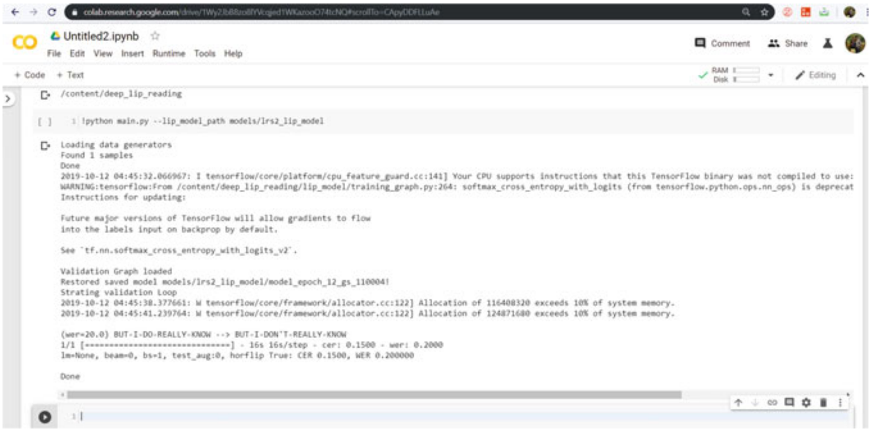


Fig. 8 Clip that contains the phrase “But I don’t know” by Female-1



colab.research.google.com/drive/1Wy2Jb88zo8fYVcqjed1WKazooO74tcNQ#scrollTo=CApyDDFLLuAe
CO
Untitled2.ipynb
File Edit View Insert Runtime Tools Help
Comment
Share
+ Code
+ Text
RAM
Disk
Editing

```
/content/deep_lip_reading
```

```
!python main.py --lip_model_path models/lrs2_lip_model
```

```
Loading data generators
Found 1 samples
Done
2019-10-12 04:45:32.066967: I tensorflow/core/platform/cpu_feature_guard.cc:141] Your CPU supports instructions that this TensorFlow binary was not compiled to use:
WARNING:tensorflow:From /content/deep_lip_reading/lip_model/training_graph.py:264: softmax_cross_entropy_with_logits (from tensorflow.python.ops.nn_ops) is deprecat
Instructions for updating:

Future major versions of TensorFlow will allow gradients to flow
into the labels input on backprop by default.

See `tf.nn.softmax_cross_entropy_with_logits_v2`.

Validation Graph loaded
Restored saved model models/lrs2_lip_model/model_epoch_12_gs_110004!
Strating validation Loop
2019-10-12 04:45:38.377661: W tensorflow/core/framework/allocator.cc:122] Allocation of 116408320 exceeds 10% of system memory.
2019-10-12 04:45:41.239764: W tensorflow/core/framework/allocator.cc:122] Allocation of 124871680 exceeds 10% of system memory.

(wer=20.0) BUT-I-DO-REALLY-KNOW --> BUT-I-DON'T-REALLY-KNOW
1/1 [==============================] - 16s 16s/step - cer: 0.1500 - wer: 0.2000
lm=None, beam=0, bs=1, test_aug:0, horflip True: CER 0.1500, WER 0.200000

Done
```

Fig. 9 Predicted output

11 Application

This model can be used to create different application that can be proposed such as an alternative voice for laryngectomised patients. This would also help the people with hearing disorder to easily recognise the spoken words instead of manual methods of lip reading. The output of EVA can be taken for spying purposes, that is, to generate the words spoken in an audio-less video from CCTV cameras with low/poor quality. This can be integrated with the live videos to generate subtitles for that instance accurately. This acts as a valuable communications tool for deaf and hard of hearing people.

12 Conclusion

In this work, we proposed a novel feature approach utilising an enhanced version of SRResNet CNN to convert low-resolution to high-resolution videos. Our experimental results demonstrate a supervised learning approach that recognises ten self-curated phrases. EVA (enhanced video articulation) is a model which performs the labial articulation of enhanced video. SegNet and LACM are used to extract the labia from enhanced frames. These enhanced frames are applied to RNN with LSTM architecture to predict the words in the respective frames.

This work leads to a fundamental understanding of existing models and overcoming their drawbacks. Future work can be developed that is independent of speakers' phoneme by creating larger dataset, increasing the number of phrases uttered by the speakers.

References

1. Desai D, Agrawal P, Parikh P. Soni MP (2020) Visual speech recognition. Int J Eng Res Technol (IJERT)
2. NadeemHashmi S, Gupta H, Mittal D, Kumar K, Nanda A, Gupta S (2018) A lip reading model using CNN with batch normalization. In: 2018 Eleventh international conference on contemporary computing (IC3)
3. Barkowsky M, Sedano I, Brunnström K, Leszczuk M, Staelens N (2014) Hybrid video quality prediction: reviewing video quality measurement for widening application scope. Multimedia Tools Appl 74(2):323–343
4. Bear HL, Harvey R (2016) Decoding visemes: improving machine lip-reading. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2009–2013
5. Chung JS, Zisserman A (2016) Lip reading in the wild. In: Asian conference on computer vision. Springer, pp 87–103
6. International Phonetic Association (1999) Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet. Cambridge University Press
7. Wand M, Koutn J et al. (2016) Lipreading with long short-term memory. In ICASSP'16, pp 6115–6119
8. Ephrat A, Peleg S (2017) Vid2speech: speech reconstruction from silent video. arXiv preprint arXiv:1701.00495
9. Mei Y, Fan Y, Zhou Y, Huang L, Huang TS, Shi H (2020) Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5690–5699