Modelling Lips-State Detection Using CNN for Non-Verbal Communications

Abtahi Ishmam, Mahmudul Hasan, Md. Saif Hassan Onim[0000−0002−7228−2823], Koushik Roy, Md. Akiful Hoque Akif, and Hussain Nyeem[0000−0003−4839−5059]

Military Institute of Science and Technology (MIST)
Mirpur Cantonment, Dhaka–1216, Bangladesh
abtahiishmam3@gmail.com mahmud108974@gmail.com saif@eece.mist.ac.bd
rkoushikroy2@gmail.com mohammadaxif5717@gmail.com h.nyeem@eece.mist.ac.bd

Abstract. Vision-based deep learning models can be promising for speech-and-hearing-impaired and secret communications. While such non-verbal communications are primarily investigated with hand-gestures and facial expressions, no research endeavour is tracked so far for the lips state (*i.e.*, open/close)-based interpretation/translation system. In support of this development, this paper reports two new Convolutional Neural Network (CNN) models for lips state detection. Building upon two prominent lips landmark detectors, DLIB and MediaPipe, we simplify lips-state model with a set of six key landmarks, and use their distances for the lips state classification. Thereby, both the models are developed to count the opening and closing of lips and thus, they can classify a symbol with the total count. Varying frame-rates, lips-movements and face-angles are investigated to determine the effectiveness of the models. Our early experimental results demonstrate that the model with DLIB is relatively slower in terms of an average of 6 frames per second (FPS) and higher average detection accuracy of 95.25%. In contrast, the model with MediaPipe offers faster landmark detection capability with an average FPS of 20 and detection accuracy of 94.4%. Both models thus could effectively interpret the lips state for non-verbal semantics into a natural language.

Keywords: Lips-state detection· DLIB· MediaPipe· CNN· non verbal communications· human-robot interaction.

1 Introduction

Lips-state (*i.e.*, *open* or *close*) detection can be promising for vision-based non-verbal communication system, which has traditionally been investigated with the head movement and gestures, and facial expression. Lips-state detection and interpretation is a key step in many security, surveillance and law-enforcement applications. For example, lips reading can be helpful in emergency hostage situation. Such communications also enable a speech and hearing impaired person (with a disorder like *stuttering*, *apraxia*, *dysarthria*, or *muteness*) to communicate using lips-state with a minimal effort. In support of developing such a system

for non-verbal communication for translating simple lip-state combinations to a complete instruction as illustrated in Fig. 1, we aim to start with the development of a lip-state detection model in this paper. The envisaged model with higher possible detection accuracy thus could be promising to a cost-effective and as simple solution as a user-friendly mobile application.

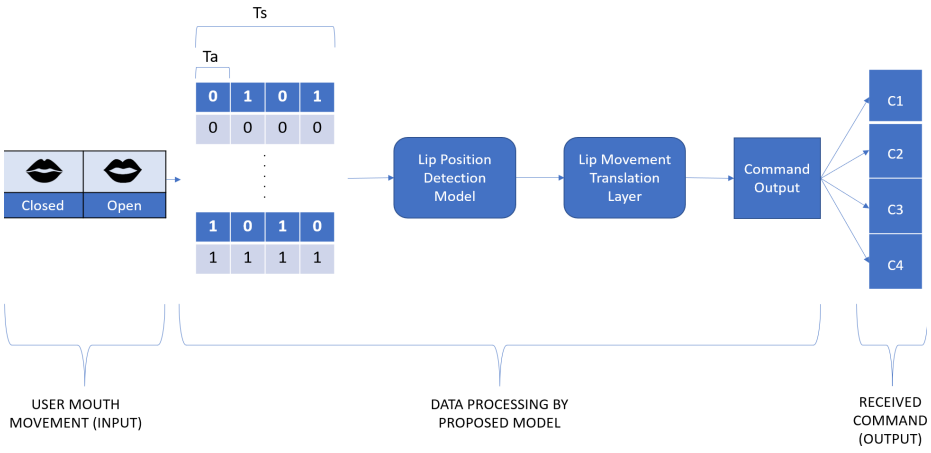


Fig. 1: A general framework of the proposed lips-state detection based interpretation/translation model for non-verbal communication system.

Despite having an obvious potential for nonverbal communication, no or a little research endeavours can be tracked in the literature on lips-state detection. For example, lip-reading was studied for verbal communications using Convolutional Neural Network (CNN) models [6]. Besides, lips states and movements have been partially considered for the facial detection and emotion recognition, both with and without the landmarks of lips. The case of without lips landmarks has higher dependency on the image contrast and spatial resolution, and thus, performance of this approach significantly varies [2]. In contrast, the lips landmarks based detection is widely investigated facial detection and emotion recognition. For example, Sharma *et al.* [9] considered face alignment and feature extraction in their face recognition model using Convolutional Neural Network (CNN) and DLIB face alignment.

To better tackle the challenges in lips detection, including varying skin colour, appearance and different lighting conditions, Juan *et al.* [11] considered the relative distances among faces, eyes and mouth to locate the mouth region. Their model can segment the lips more efficiently than some other prominent models. Amornpan *et al.* [1] applied transfer learning for feature extraction for a face recognition application using a pre-trained deep learning model and validated with the public face datasets like Extended Yale Face Database B (Cropped) and the Extended Cohn-Kanade Dataset (CK+).

Similarly, Singh *et al.* [10] proposed to use Viola Jones algorithm to localise the face and mouth in the image with an iterative and adaptive construction of merging threshold to make the model image quality invariant. Krause *et al.* [5] introduced a highly portable, and automatic solution for extracting oral posture from digital video using an existing face-tracking utility and OpenFace2. Later, Xu *et al.* [12] improved the facial landmark localisation across large poses using a split-and-aggregate strategy.

The above vision-based models focus on the face tracking and have a partial consideration of lips detection. In addition to the contextual information extracted from the landmarks of eyes, nose and face-shape, lips were considered to complement the features required for the face detection and recognition and emotion classification. However, as mentioned in the beginning of this section, no lips-state detection is considered so far for developing the envisaged translation model for non-verbal communication system.

In this paper, two new lips-state detection models have been proposed for non-verbal communications. Both models use landmark detectors for the localisation of lips. Two highly accurate and robust facial landmark detectors: DLIB [4] and MediaPipe [7] are considered, which can gather facial landmark information from both real-time and static images. Building upon these detectors, we simplify modelling the lips-state with a set of six key landmarks and their distances to detect relative variations in lips.

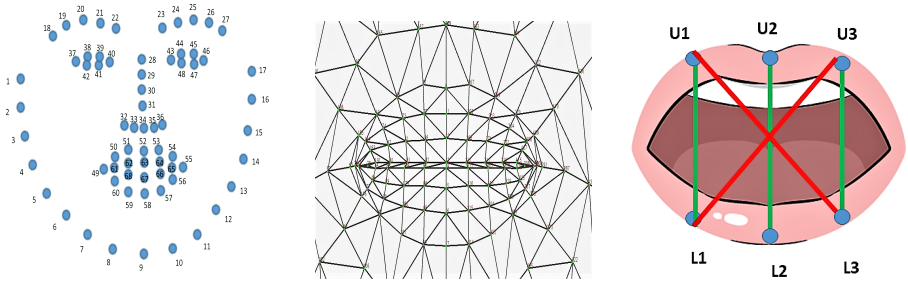
Thereby, we develop two models to accurately track the lips-states. Particularly, these models start with the detection of human faces from the captured frames, and isolate the face region followed by the extraction of the landmarks of lips. This identified landmarks and the distance among the landmark points are then used for the successive decision making, approximation and classification of the lips-state. Performance of these proposed models are finally analysed to learn their merits for the non-verbal communication system (see Sec. 3).

2 New Lips-State Detection Models

Our research aims at developing an alternative interpretation or translation system for the speech and hearing impaired communication based on lips-state. To this end, we have modelled lips-states streamlining the landmarks, and thus, developed lips-states classifier model using two prominent landmark detectors, and thus, we construct two models as illustrated in Fig. 3. The first model that we call model-I uses the popular landmark detector *DLIB* to feed data into our neural network based prediction model to predict the lips-states. Similarly, we use the MediaPipe face-mesh landmark detector to develop the second model called model-II with Support Vector Regression (SVR) prediction block. Having similar network architecture, the lips-state classification accuracy and speed may primarily depend on their underlying detectors. Prior to analysis of these performances in Sec. 3, we now present below the necessary technical details of the proposed models.

2.1 Landmark Detection

DLIB. For the first model, pre-trained face feature point detector that comes with the DLIB library is used to obtain 68 Cartesian coordinate points corresponding to a specific area of the face. This 68-points shown in Fig. 2a comes from the DLIB model which is trained on the iBUG 300-W dataset [8]. From those landmark point 6 landmark points were selected for lip distance calculation.



(a) (b) (c)

Fig. 2: Facial landmark points: (a) DLIB, (b) MediaPipe, and (c) simplified lips landmarks and distances of our model.

MediaPipe. MediaPipe face mesh is a face geometry detector based on the blaze-face model. It can estimate 468 3D face landmarks as shown in Fig. 2b, from both recorded and real-time video streams even on mobile devices. The detector employs machine learning to infer the 3D surface geometry, requiring only a single camera input without the need for a dedicated depth sensor, a feature suitable for our purpose, accurate tracking of lips. Utilising lightweight model architectures together with GPU acceleration throughout the pipeline, the detector also can deliver real-time performance even on weaker processing machines such as *raspberry pi* or *jetson* devices. Out of the 468 landmarks, 6 were used for lips distance calculation.

2.2 Distance Calculation

After detection of landmarks, five distances were calculated Fig. 2c from the (x_i, y_i) coordinates of the points. Six landmark points were selected for distance calculation shown in Table. 1. These distances are computed using Eq. (1a) - (1e).

$$LD = \alpha \times \sqrt{(U_1[x_1] - L_1[x_2])^2 + (U_1[y_1] - L_1[y_2])^2} \tag{1a}$$

$$MD = \beta \times \sqrt{(U_2[x_1] - L_2[x_2])^2 + (U_2[y_1] - L_2[y_2])^2} \tag{1b}$$

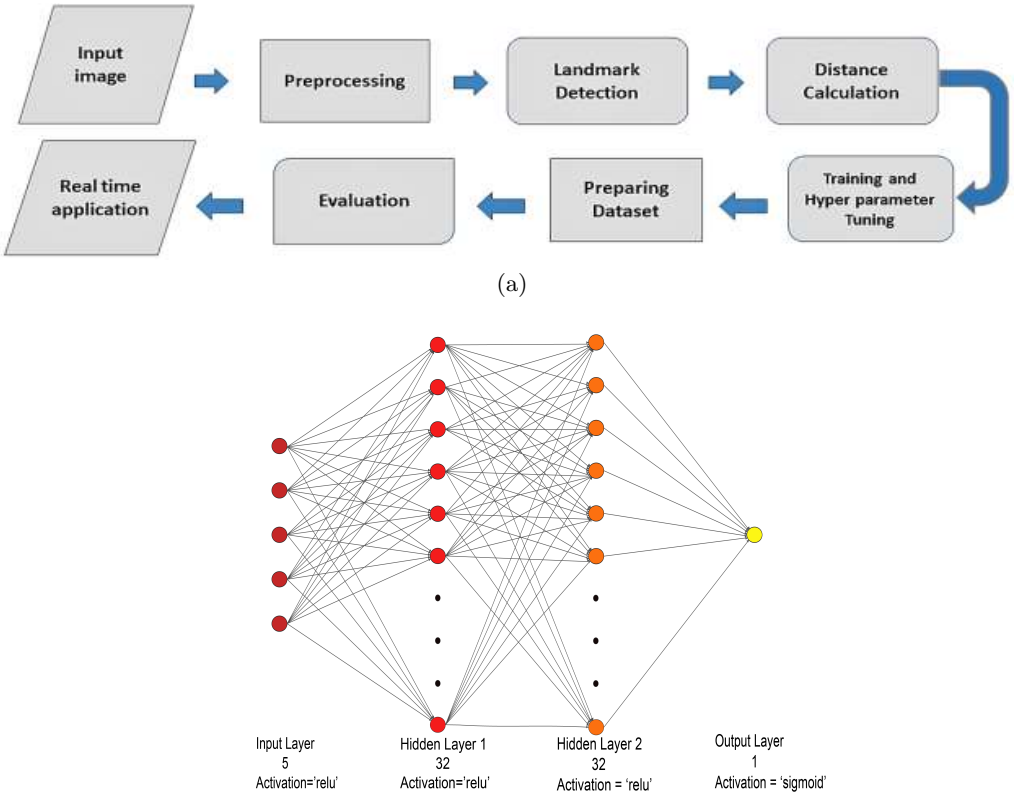


Fig. 3: Proposed model: (a) key processes and (b) neural network architecture.

Table 1: Landmark used for distance calculation

Lips landmark point	**Our model-I**	**Our model-II**
U1	51	37
U2	52	267
U3	53	0
L1	59	84
L2	58	314
L3	57	17

$$RD = \gamma \times \sqrt{(U_3[x_1] - L3[x_2])^2 + (U_3[y_1] - L_3[y_2])^2} \tag{1c}$$

$$D_1 = \delta \times \sqrt{(U_1[x_1] - L_3[x_2])^2 + (U_1[y_1] - L_3[y_2])^2} \tag{1d}$$

$$D_2 = \epsilon \times \sqrt{(U_3[x_1] - L_1[x_2])^2 + (U_3[y_1] - L_1[y_2])^2} \tag{1e}$$

Here, α, β, γ, δ, ϵ are the length control coefficient, which can be tuned for improved observation; U_1 = upper lip left point; U_2 = upper lip middle point; U_3 = upper lip right point; L_1 = lower lip left point; L_2 = lower lip middle point; L_3 = lower lip right point; LD = lip left distance; MD = lip middle distance; RD = lip right distance; $D1$ = lip diagonal 1 distance; $D2$ = lip diagonal 2 distance. These selected landmark points demonstrate the maximum variation upon lips-state . Although the middle points U2 and L2 can be sufficient for our purpose, the other points were considered to tackle the variety of lips shapes during their-states.

2.3 Dataset Collection

We collected video recording of 15 individuals for training. For each person, video was captured for both closed and open mouth positions. The lips training dataset consists of total 23,412 frames. All of these frames were passed through the preprocessing stage to extract landmarks and calculate the 5 distances mentioned in distance calculation. Thus a dataset consisting of 5 lips landmark distances and mouth position, either open or close for all frames was created and used to train the data on the train dataset. The validation data consists of 2325 frames of 3 people excluded from the training set.

2.4 Lips-State Detection

Proposed model-I using DLIB data. The proposed model-I utilises a typical and very lightweight dense neural net with one dense layer of 32 units and Rectified Linear Unit (RELU) activation. Early stopping was employed for better optimisation of training time and to avoid over-fitting. For compilation of the model, the optimiser was set to Adam and loss was binary cross-entropy. After training and tuning, it was used in real time application.

Even though the model is very lightweight, the underlying DLIB detector in the data pre-processing stage was adding enough complexity to turn the frames per second (FPS) down. We got about 5 to 9 FPS in real time application, that is usable at best but not robust enough to work in every situation.

Proposed model-II using MediaPipe data. Proposed model-II uses the landmark distance values obtained from MediaPipe and plugs it into a Support Vector Classifier (SVC) model. The model was used with the default parameters as it was able to provide sufficiently accurate results.

2.5 Training Performance Evaluation

The mouth position detection model was trained for 350 epochs with early stopping employed to stop the model training if the accuracy improvement is negligible. It can be seen from the loss vs epoch curve that the model is very quick to get to a lower loss value (see Fig. 4). After a while the improvement as well as the learning rate gets saturated.

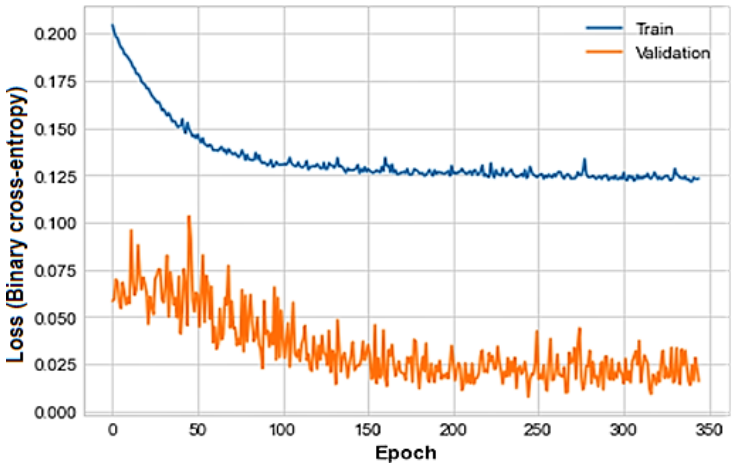


Fig. 4: Average loss vs epoch

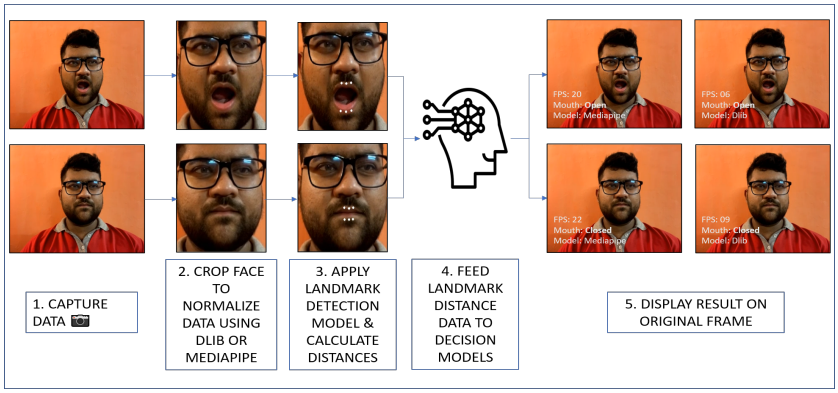


Fig. 5: Performance of the proposed model in a realtime application scenario.

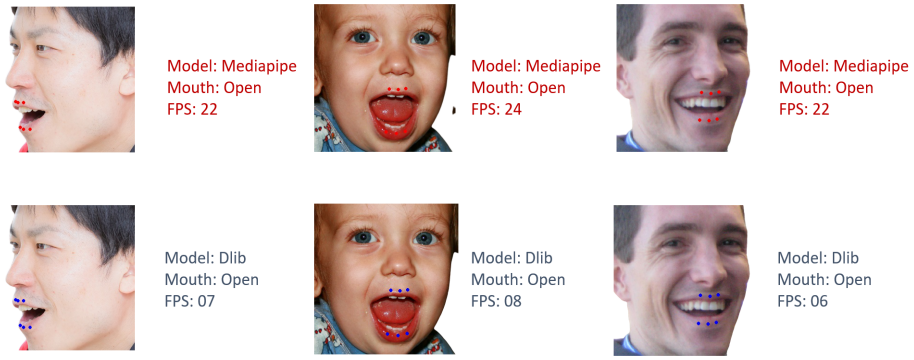
2.6 Time Slots and Commands

Each of the detected actions are placed into time slot for converting into commands as shown in Fig. 1. If the time slots are field with N number of actions. The command will be known as N actions per time slot window. The total number of actions per window will increase the maximum number of commands. This is limited by the maximum rate of frame captured and the maximum number of actions taken by the user. Their overall relation can be expressed as follows:

$$T_a = T_s/N_a \tag{2}$$

T_s = time for N_a; T_a = time for action; N_a = total number of actions per time slot. Since it is binary action, $N_c = 2^{Na}$, where N_c = total number of commands. For example if $T_a = 0.6$ sec, $N_a = 4$, then T_s will be $0.6 \times 4 = 2.4$ sec.

(a) Closed lips



(b) Open lips

Fig. 6: Examples of lips-state detection using Proposed model-I (*bottom row*) and model-II (*top-row*) for opening and closing lips.

3 Results

The model was tested on both recorded video footage with video resolution of 1920×1080 at 30 fps as shown in Fig. 5 and still images from Flickr-Faces-HQ Dataset (FFHQ) [3]. A model trained with a particular dataset may not give similar results when tested with data from real world. The model is therefore evaluated for the cases unknown to it.

A step by step working procedure of our models for real-time application scenario is illustrated in Fig 5. A few cases of visual detection of lips-state ares also illustrated for the considered datasets in Fig. 6. The images in the figure captured varying conditions of face, angle, appearance and lighting that illustrates the performance of the model for detecting the opening and closing of the lips.

Similarly, the performance of the models while applied on faces turned in different angles have been given in Table. 2, which shows that the effect of face rotation can be quite significant and similar for both models. The detection accuracy is 100% from -40 degree to +40 degree angle. But the results are affected when face is rotated beyond that. The accuracy starts declining at +-60 degrees still staying at a quite appreciable 90%. But if the face is rotated any further, both the models fail to detect the facial landmarks.

Table 2: Detection accuracy of proposed model at varying face angles

Angle (degree)	**Our model-I (%)**	**Our model-II (%)**
+0	100	100
+20	100	100
+40	100	100
+60	95	95
+80	not detected	not detected
-0	100	100
-20	100	100
-40	100	100
-60	95	95
-80	not detected	not detected

Table 3: Detection accuracy at varying lips-state detection rate

Movements per sec	**Our model-I (%)**	**Our model-II (%)**
1	100	100
2	100	100
3	95	100
4	85	95
5	75	90
6	60	80

The models' accuracy at different speeds of changing the lips-states is also evaluated and given in the Table. 3. We observed that the accuracy decreases for both models with the increase in the lips-state per second. For instance, as in Table 3, both of the models can detect the lips-states, while the lips open/close at a rate of 2 per second. For a different case, when the rate increases to six per

Table 4: Overall performance of lips-state detection

Parameters	**Our model-I**	**Our model-II**
Best Validation Accuracy	95%	94%
Best Training Loss	0.122%	0.130%
Best Training Accuracy	95.47%	94.808%
Average Accuracy	95.25%	94.40%
Average FPS	6	20

Table 5: Sample of training data

Left	**Middle**	**Right**	**Diagonal-1**	**Diagonal-2**	**Output**
30.06659	29.123	28.01785	31.38471	33.10589	1
30.06659	30.01666	29.01724	32.31099	33.54102	1
34.05877	34.0147	33.01515	36.05551	37.16181	1
17.02939	17.02939	17.02939	21.40093	21.40093	0
18.02776	17.03876	17.02939	21.63331	21.40093	0
18.02776	17.01345	16.03122	21.63331	21.40093	0

second, nearly 40% of the state-changes are not detected. However, compared to model-I, the model-II is found more accurate for the states being changed at a higher rate.

The overall training and test accuracy of the models is given in Table. 4, where we observed that model-I is slightly more accurate, while model-II is much faster, almost 3 times than model-I. So, proposed model-II can be a better choice for fast moving scenarios. In contrast, model-I is more suited for situations where accuracy is the main priority. In other words, Table 4 indicates that with only 6 frames per second, model-I has an average accuracy of 95.25%, and with 20 frames per second, model-II has an average accuracy of 94.40%.

Additionally, some values of the dataset are given in Table. 5 that suggest that the higher distance values correspond to mouth open position whereas lower distance values normally indicate closed mouth position. For example, in Table 5, we see that the maximum distance between the landmarks are for *diagonals* and the values are comparatively higher in case of open lips than that of the closed lips. The shortest distances, in contrast, are between the landmarks pair of left-most and rightmost sides, and they are relatively lower for the case of closed lips.

Finally, the lips-state detection for varying conditions are illustrated in Fig. 6. The trends in accurately detecting the lips-states also holds for the other test images. Capability of accurately detecting the lips opening and closing at different angles and lighting conditions means that the proposed model have the potential for a lips-state based electronic translator of a nonverbal communication system.

4 Conclusion

In support of developing an alternative interpretation/translation system for non-verbal communication system, we have introduced two new models for lips-state detection. Building upon two popular facial landmark detectors, DLIB and MediaPipe, the proposed models have been investigated for the classification of opening and closing of lips for the standard datasets. Being faster, computationally efficient (in terms of FPS) and reasonably accurate, the proposed model-II with MediaPipe can be a promising candidate for the envisaged translation system. Our research continues to further develop the detection accuracy of the proposed models, particularly for extreme facial rotations and very poor lighting conditions.

References

1. Amornpan, P., Praisan, P.: Face recognition using transferred deep learning for feature extraction. In: 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (2019)
2. Bouvier, C., Benoit, A., Caplier, A., Coulon, P.Y.: Open or Closed Mouth State Detection: Static Supervised Classification Based on Log-polar Signature. In: ACIVS 2008 - International Conference on Advanced Concepts for Intelligent Vision Systems. vol. Volume 5259/2008, pp. 1093–1102. Springer Berlin / Heidelberg, Juan-Les-Pins, France (Oct 2008), https://hal.archives-ouvertes.fr/hal-00372148
3. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks (2019)
4. King, D.E.: Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research **10**, 1755–1758 (2009)
5. Krause, P.A., Kay, C.A., Kawamoto, A.H.: Automatic motion tracking of lips using digital video and openface 2.0. Laboratory Phonology: Journal of the Association for Laboratory Phonology **11**(1) (2020)
6. Lu, Y., Li, H.: Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory. Applied Sciences **9**(8), 1599 (Apr 2019). https://doi.org/10.3390/app9081599, http://dx.doi.org/10.3390/app9081599
7. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
8. Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: Database and results. Image and vision computing **47**, 3–18 (2016)
9. Sharma, S., Shanmugasundaram, K., Ramasamy, S.K.: Farec—cnn based efficient face recognition technique using dlib. In: 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT). pp. 192–195. IEEE (2016)

10. Singh, B., Sahoo, S., Kumar, V., Issac, A., Dutta, M.K.: Improved lip contour extraction using k-means clustering and ellipse fitting. In: 2016 2nd International Conference on Communication Control and Intelligent Systems (CCIS). pp. 99–103. IEEE (2016)
11. WenJuan, Y., YaLing, L., MingHui, D.: A real-time lip localization and tacking for lip reading. In: 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE). vol. 6, pp. V6–363. IEEE (2010)
12. Xu, Z., Li, B., Geng, M., Yuan, Y., Yu, G.: Anchorface: An anchor-based facial landmark detector across large poses. arXiv preprint arXiv:2007.03221 (2020)