ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse58942020.pdf
https://doi.org/10.30534/ijatcse/2020/58942020

A Pipeline to Data Preprocessing for Lipreading and Audio-Visual Speech Recognition

Hea Choon Ngo[1], Ummi Raba'ah Hashim[2], Raja Rina Raja Ikram[3], Lizawati Salahuddin [4], Mok Lee Teoh[5]
[1, 2, 3, 4, 5]Center for Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM), 76100 Durian Tunggal, Melaka, Malaysia,
[1]heachoon@utem.edu.my, [2]ummi@utem.edu.my, [3]raja.rina@utem.edu.my, [4]lizawati@utem.edu.my, [5]teohmoklee@gmail.com

ABSTRACT

Studies show that only about 30 to 45 percent of English language can be understood by lipreading alone. Even the most talented lip readers are unable to collect a complete message based on lipreading only, although they are often very good at interpreting facial features, body language, and context to find out. As you can imagine, this technique affects the brain in different ways and becomes exhausting over a period of time. If a person who is deaf, uses language and is able to read lips, hearing people may not understand the challenges they are facing just to have a simple one-on-one conversation. The hearing person may be annoyed that they are often asked to repeat themselves or to speak more slowly and clearly. They could lose patience and break off the conversation. In our modern world, where technology connects us in a way never thought possible, there are a variety of ways to communicate with another person. Deaf people come from all walks of life and with different backgrounds. In this study, a lipreading model is being developed that is able to record, analyze, translate the movement of lips and display them into subtitles. A model is trained with GRID Corpus, MIRACL-VC1 and pre-trained dataset and with the LipNet model to build a system which deaf people can decode text from the movement of a speaker's mouth. This system will help the deaf people understand what others are actually saying and communicate more effectively. As a conclusion, this system helps deaf people to communicate effectively with others.

Key words: Conversation, facial features, lipreading model

1. INTRODUCTION

Language is tool for humans to communicate with each other and to express their feelings. However, not everyone can express their opinion, some of them cannot communicate with others. Deaf people refer to someone who cannot hear as well as someone with normal hearing. Deaf people have a minority population in every country in the world [1]. They usually have very little or very severe hearing loss, so they have to use sign language to communication with others [2]. When deaf people meet up together, they communicate directly by drawing based on their knowledge of signed language and find common ground in the movement between signed language [3].

Lipreading is the language that translates the speaker's lip movement into speech. Lipreading is usually provided with face expressions and body language to enable a more precise understanding. Lipreading plays an important role in daily communication for deaf people. Deaf people can understand what the speaker said by lipreading and translate it into speech.

In this study, three datasets are used which are GRID dataset, MIRACL-VC1 and pre-trained dataset. During preprocessing, the video is converted into a frame with OpenCV. Then, by using OpenCV, the frame is converted into a grayscale image for the next stage. After changing the frame to grayscale image, shape_predictor_68_face_landmarks.dat is used to identify the face in the frame and crop lip portion of the entire image. Dlib is a modern C++ toolkit that includes machine learning algorithm and tools to solve real problems. This toolkit is used to recognize the face and mark facial contours such as the eye, nose and lips. For the last step of preprocessing, the cropped frame is written in an image.

This cropped image is used as training data for the LipNet model. LipNet is the first end-to-end lipreading model at the sentence level. LipNet is consistently trained to predict sentences. This model uses spatio-temporal convolutional neural network, recurring neural networks and the connectionist temporal loss of classification [2]. After training this model, it can accept entering a new test video without language. For the evaluation phase, 4 datasets are used for testing and for further use in the training. Performance is measured in terms of accuracy using the LipNet model we have developed. Finally, the output of this system is the predicted subtitle of the speakers.

2. RELATED WORK

Most of the existing automated lipreading does not apply deep learning. Deep learning requires a preprocessing process, such as extracting the image features from the frame and extracting the frame from the video features. According to Iain Matthews et al., (2002), there are some problems in developing better visual and audio-visual speech recognition [4]. One of the problems is getting speech recognition from an audio signal. This problem is solved due to the speech recognition systems that are often used by personal computers. However, there are some factors that need to be improved, such as speaker type and microphone, accent and ambient noise. The main problem with generating visual features is the amount of data in a video. Thousands of pixels may need to be extracted for each video, and other variable such as different speakers, head position and environment also affect the result. In this work, the "bottom-up" approach is used to estimate the facial feature from the image, and "top-down" holds the important information and assumptions in a model. They combined two "top-down" methods, the Active Shape Model (ASM) lip contour tracker and the Active Appearance Model, with a "bottom up" model to improve lipreading performance because the ASM model is not enough features for lip reading [5]. In this work, they also record their own aligned audio-visual database called AVletter. This database contains three repetitions by 10 speakers.

Gerasimos Potamianos et al., (2003) states that two main aspects lie in automatic speech recognition (ASR), namely visual front-end design and audio-visual speech integration [6]. Automatic speech recognition is part of the human computer interfaces for identifying and processing the human voice. The visual front-end design creates visual speech features upon receipt of video input, which are generally customized based on three categories: appearance-based features, shape-based features, and a combination of both. Appearance-based features use region of interest ROI to ensure that the spoken utterance are stored. The ROI is a square with image pixels in the mouth area of the speaker. The image may have included the lower face or even the entire face. Shape-based feature extraction displays information about facial contours such as the eye, nose and lip. Finally, both categories were combined to create an appearance model for speech reading [7]. They concluded that extracted visual features provide more accurate speech information.

According to Ziheng Zhou et al., (2014) there are three issues related to features extraction [8]. The three topics are speaker dependency, body position and temporal information. The mouth position and size of each person looks different when extracted from the images. Problems can arise if we want to extract important information such as lipreading from the image of the cropped mouth [9]. To solve this problem, techniques such as normalization of the vocal tract are used to solve the variability in different speakers. When taking pictures or video recordings of the speaker's face, it should not only take a frontal view from every angle. The camera view affects the appearance of the mouth, so it should be taken at more than one angle. There are several techniques that can be used to vary poses. For example, extracting pose-dependent features from non-frontal view images and using them for speech recognition or converting pose-dependent feature to pose- independent features before each classification process. Within the visual speech signal, not only the visual appearance of the speaker's mouth is important, but also the temporal information. In order to extract the temporal information, they took into account the local pixel-level spatio-temporal structures of the structure and the model in the extracted visual features.

2.1 Classification with Deep Learning

In recent years, several research and work has been done to apply deep learning to lipreading. These approaches carry out a word or phoneme classification. According to Hiroshi Ninomiya et al., (2015), that is an increase in the number of studies on the use of deep learning in automatic speech recognition [10]. Based on their research, they show that deep bottleneck features(DBNFs) that apply a deep neural network concept can reduce the error rate of an automatic speech recognition system. Bottleneck features are created by multi-layer perceptron, internal layers, also known as the "bottle neck layer", and a small hidden unit. In this work, a DNN is initially trained unsupervised and later fine-tuned [11] to predict the probabilities of classes. After testing, the result shows that the word error rate was improved from 60.3% to 35.9%. using DBAFs. As a conclusion, a system with DBNFs performs better than a pure audio ASR system.

Joon Son Chung et al., (2016) proposed a spatial and spatio-temporal convolutional neural network for word classification [12]. They developed a system for predicting the words spoken by a speaker using video without audio as input. The main task for this network is to predict when a word will be spoken and the input format is a sequence of mouth regions. They developed four architectures that are 3D convolutional with early fusion (EF-3), 3D convolution with multiple tower(MT-3), early fusion (EF) and multiple tower (MT). EF-3 and MT-3 is a 3D convolution with a small partial and temporal kernel size. These two architecture go well with a spatio-temporal feature such as the lip shape whereas 2D convolutional wastes parameters or redundancy when trying to respond to spatio-temporal features. If testing with 333-word, it gains 65.4 accuracy and the use of CNNs in pre-trained words helps to recognize sentences rather than single words [13]. Convolutional Neural Network (CNN) are commonly been applied to detect diabetics and cancer [14] [15].

According to Michael Wand et al., (2016), the neural network used in the lipreading system has better word accuracy than a conventional extraction of processing pipeline features [16]. Feed-forward and recurrent neural network layers, which are referred to as Long Short-Term Memory (LSTM), together with the LSTM sequence classifier results in a word accuracy of approximately 80% when predicting the word spoken by the speaker.

Yannis M. Assael et al., (2016) defined that LipNet as the first end-to-end model to perform sentence-level prediction in visual speech recognition [17]. LipNet use a variable from video frames as input and trained end-to end to the text sequence. In the LipNet architecture, a sequence of T-frames is used as input video and processed by 3 layer of STCNN and spatial pooling. The lip features are extracted by 2 Bi-GRU, the output of GRU is processed by a linear layer and a Softmax. This model is trained using the connectionist temporal classification.

Lipreading is a technique for understanding speech by interpreting the movement of speaker's lips. Lipreading is usually used by deaf people to express their needs and feelings. Lipreading can be processed and implemented using image processing and neural network technique.

3. METHODOLOGY AND ANALYSIS

This system was started with video acquisition. In this study, we used the GRID audio-visual sentence corpus and MIRACL-VC1 as a dataset. We have to use the function of Dlib and cv2 to complete this task. Dlib is a modern C++ toolkit that includes machine learning algorithm and tools for creating complex software to solve real problems. OpenCV or cv2 is an open source software library for computer vision and machine learning. In this study, we used cv2 as a tool for editing video and frame images.

At the beginning of this project, we need to set the path of the dataset so the system can load the file. In GRID Dataset, there are 35 folders have to be loaded and only the file with number 21 is skipped because the file has been damaged. After the dataset is loaded, we use cv2.VideoCapture to convert the video into a frame. In this case, we convert the video to a 75 frame image. For MIRACL-VC1 and pre-trained dataset, we have to load the folders and make sure that every folder contains images of every phrase. Then, save it in a list to crop out the lip portion. There are 68 specific points called landmarks that are on every face including nose, eyes and lips as well as the edge of each point. The Dlib library is an open source library that recognizes facial landmark point of a face. When recognizing facial markings in the Dlib library, the images used for the training are labeled manually and indicate the coordinate point of the regions that surround each facial structure. After that, specify the distance between pairs of the input pixels.

The training data is used to train a classifier through an ensemble of regression trees to determine the position of the face marker directly from the pixel intensities. With this face marking detector, the face marking can be recognized in real time with high accuracy [18]. In this case, we take the image part from 49 to 55 point. These points are the lip section in a face. After we recognize the lip section, the specific section is cropped into an image. We repeat these steps until the entire image is cropped. The entire image is then combined with the function cv2.VideoWriter_fourcc(*XVID) to form a video. The image is converted to a video in XVID format. Finally, this module is developed and integrated into LipNet. Figure 1 shows the flowchart of main process.

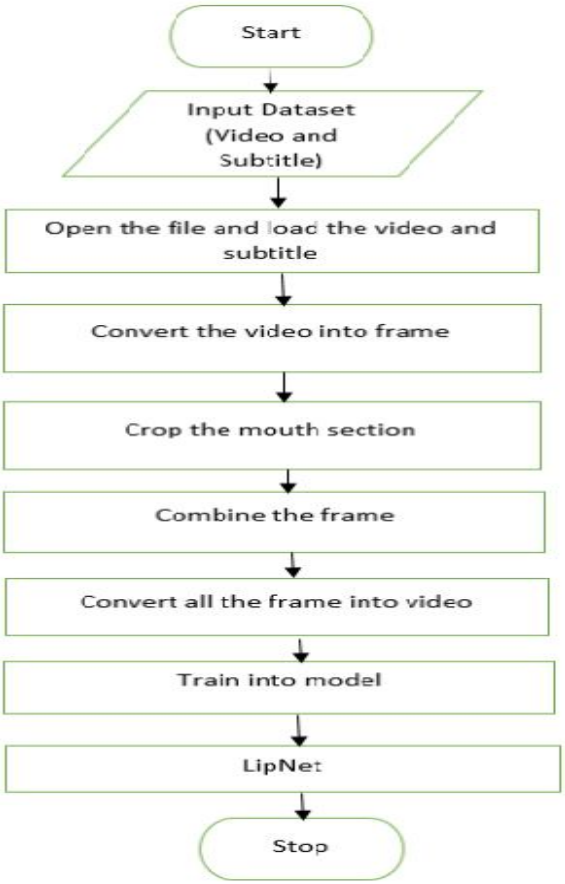

Figure 1: Flowchart of Main Process

3.1 Data Requirement

Grid corpus is a large audio-visual sentence corpus with multiple speaker, which was collected to support collaborative studies of computer behavior in speech perception. In the grid corpus, each sentence is formed by a six-word sequence as shown in Table 1. For example, sentences like "place blue now at F 9" is on the form of <command: 4> <color:4> <preposition:4> <letter:25> <digit:10> <adverb:4>, number indicates the number of choices for each part. Grid has good options and is large enough to be used as a training requirement for an audio speech recognition system. Each speaker produced all combination of the three keywords, a total of 1000 sentences per speaker. The remaining components such as command, preposition and adverb were fillers. There are 16 female and

18 male speakers involved in generating this database. Among all of them, 2 grew up in Scotland, 1 from Jamaica and others from England. All of them are staff and students in the Departments of Computer Science and Human Communication Science at the University of Sheffield. English is their main language. The ages were between 18 to 49 years.

Table 1: GRID Corpus Dataset

command	color*	preposition	letter*	digit*	adverb
bin	blue	at	A–Z	1–9, zero	again
lay	green	by	excluding W		now
place	red	in			please
set	white	with			soon

MIRACL-VC1 is a lipreading dataset that contains color images [19]. This dataset can be used in areas such as speech recognition, face recognition and biometrics. There are fifteen speakers that are five men and ten women in this dataset. Each speaker speaks ten sentences and the image are snapped and recorded. This dataset contains a total number of 3000 instances. The sentences spoken by the speaker shown in Table 2.

Table 2: MIRACL-VC1 Dataset

ID	Phrases
1	*Stop navigation.*
2	*Excuse me.*
3	*I am sorry.*
4	*Thank you.*
5	*Good bye.*
6	*I love this game.*
7	*Nice to meet you.*
8	*You are welcome.*
9	*How are you?*
10	*Have a good time.*

In this study, we used a total of three datasets to train the LipNet model. The database used as mentioned are trained GRID Corpus dataset, trained MIRACL-VC1 dataset and pre-trained dataset. This dataset consists of 11 speakers, 8 speakers are used for the training and 3 speakers are for testing. Each of the speakers speaks nine sentences and the video is recorded. Figure 2 shows the phrases spoken in this dataset. This dataset consists of 99 video files and each video file's duration is 3 seconds.

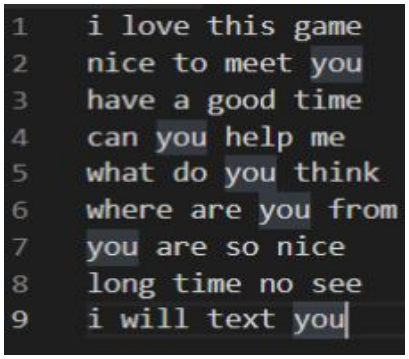

Figure 2: Pre-trained Dataset

To use this dataset for the training model, a pre-processing step is required to apply it. Figure 3 shows the steps for pre-processing the pre-trained dataset. First, the speaker records the video and saves it in a folder. The video is then converted to audio for further use. In this step, VLC Media Player converts the video into an audio file. The video is also converted to a frame and saved in a folder. The cropped frame is subjected to an image processing process to crop out the mouth portion of each frame. By using the audio and frame file, we record the alignment for every second and the image that occurs. Figure 4 shows an example of alignment file.

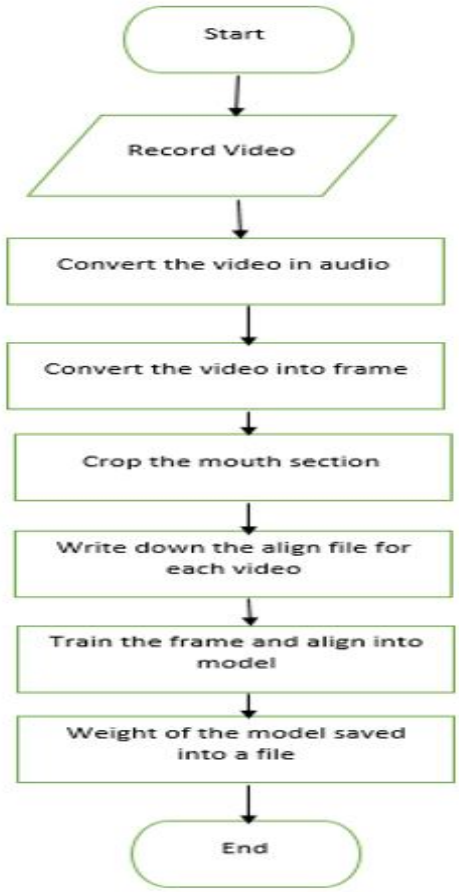

Figure 3: Steps for Pre-processing the Pre-trained Dataset

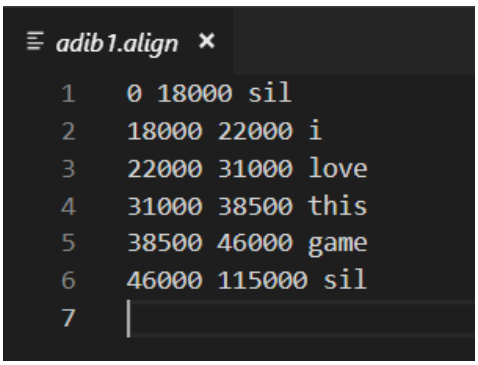

Figure 4: An Example of Alignment File

3.2 Image Processing Method

Image processing is a method of performing an operation on an image to improve image quality and extract information from the image. Nowadays, image processing has become one of the common technologies that are growing rapidly in the area such as healthcare, agricultural, computer science and engineering [20] [21]. There are major steps in image processing where images are imported using image acquisition tools, and images are analyzed and manipulated. The output where the result can be changed is an image or a report based on an image analysis. There are two types of methods in image processing, namely analogue and digital image processing. Analogue image processing, which is used for printing such as photocopies, while digital image processing helps manipulate digital images using computers [22]. The focus is on improving the image information and processing the image data for storage.

Digital Image is a representation of a two-dimensional image as a finite set of digital values called pixels. Pixels represent greyscale, colors and opacities. Image processing techniques can be used to solve problems such as recognition, detection and verification. The better the preprocessing, the better the image quality and result. Figure 5 shows the general steps in digital image processing. The image acquisition is the step for acquiring the frame from the GRID dataset and the MIRACL-VC1 dataset. For image enhancement, we use the spatial domain method and adjust the pixel of the image in width and height. Then, we change the image to grayscale to avoid unnecessary noise when restoring the image. Facial contour marking points are used in the image in the morphological processing. Local segmentation is used for segmentation to crop the image from 49 to 55 marker points, which are the lip section. The cropped image is saved for further processing.

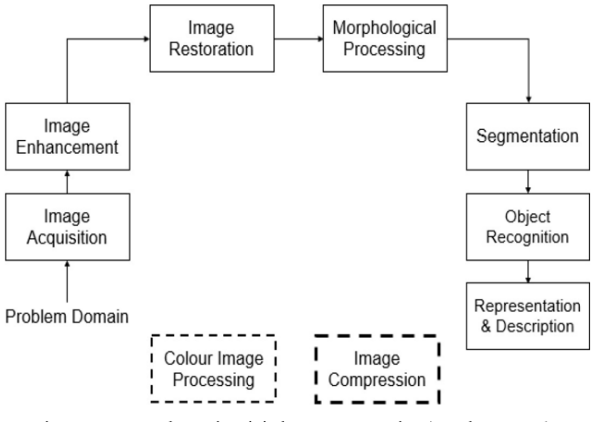

Figure 5: General Steps in Digital Image Processing

In this study, we use the Dlib software to execute the shape_predictor_68_face_landmarks model to perform facial contour recognition. In shape_predictor_68_face_landmarks, it applied the histogram of oriented gradients (HOG) and Support Vector Machines (SVM) algorithm. The histogram of oriented gradient is an image processing algorithm that performs a feature extraction function. It will have the information of 68 points mark-up in the facial contour, if there is a new input frame it will have the similarity on the face and mark the 68 points. Support Vector Machines (SVM) is a machine learning algorithm. HOG data is used to classify the landmark. In this study, facial markings from 49 to 55 (the mouth section) are cropped for the further process. Figure 6 shows the 68-point mark-up used for annotations [23].

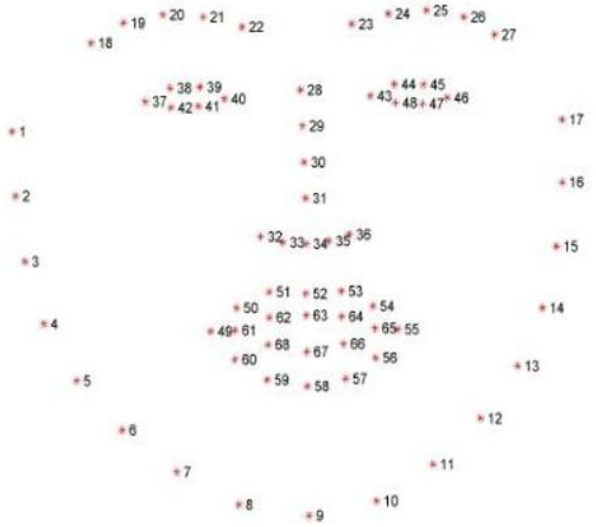

Figure 6: The 68-points Mark-up Used for Annotations [23]

3.3 Classification with Deep Learning

SVM is used for face recognition. To recognize the lip portion obtained, a vector of the facial HOG feature is extracted. This vector is then used in the SVM model to determine a matching score for lip portion input vector. The SVM returns the label with the maximum point of scores, which represents the confidence to the closet match within the training data. The neural network is used in LipNet, a model for performing lipreading. This network consists of spatio-temporal convolutional layers (STCNN), bi-directional gated recurrent units (Bi-GRU), connectionist temporal classification (CTC). The first STCNN records and processes 75 frames at the same time. Each STCNN is followed by a max-pooling layer 2x2 mask. After the third STCNN layer, the feature vector is used to enter Bi- GRU. Bi-directional RNN differs from standard RNN because it can access future information that cannot be reached by the current status. The GRU output is followed by a fully connected layer and a SoftMax classification. This end-to-end model is then trained with CTC. Figure 7 shows the architecture of LipNet [17].

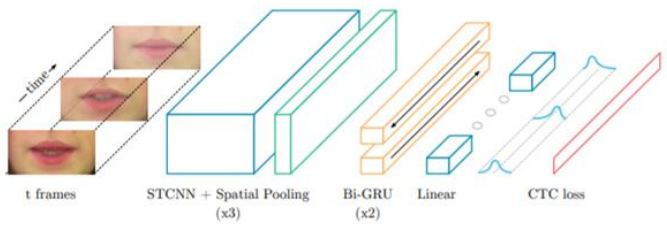

Figure 7: The Architecture of LipNet [17]

4. RESULT AND DISCUSSION

This section discusses about the accuracy and the correctness of the model. Input is required to obtain the performance of the results. Figure 8 shows the example of inputs.

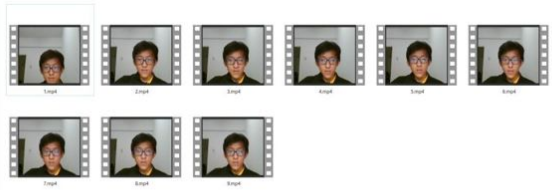

Figure 8: Example of Inputs

Performance or accuracy is calculated using the Word Error Rate (WER). WER is a performance measured in speech recognition or machine translation.

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \quad (1)$$

where

S = number of substitutions
D = number of deletions
I = number of insertions
C = number of correct words
N = number of words in the reference ($N = S + D + C$)

We invite three UTeM (Universiti Teknikal Malaysia Melaka) students to participate in this test. Each of the testers speaks three word from each dataset, namely the GRID corpus, the MIRACL-VC1 dataset and the pre-trained dataset. The following table, Figure 9-11 shows the test result of the LipNet model in each dataset. The value of WER is in percentage (%).

Sentence \ Student	Student A	Student B	Student C	Mean
1	75%	100%	100%	91.66%
2	50%	50%	75%	58.33%
3	100%	100%	100%	100%
Mean	75%	83.33%	91.66%	83.33%

	A	B	C
1	Tester1	Predicted Word	WER
2	i love this game	how a i you	0.75
3	nice to meet you	nice I you	0.5
4	have a good time	nice I you	1
5			
6	Tester2		
7	i love this game	stop navgtyoe	1
8	nice to meet you	nice I a you	0.5
9	have a good time	you aigtiou	1
10			
11	Tester3		
12	i love this game	how are you	1
13	nice to meet you	thank I you	0.75
14	have a good time	you are you	1

Figure 9: Test Result on MIRACL-VC1 dataset

Sentence \ Student	Student A	Student B	Student C	Mean
1	100%	100%	83.33%	94.44%
2	83.33%	100%	66.67%	83.33%
3	83.33%	66.67%	83.33%	77.78%
4	100%	83.33%	33.33%	72.22%
Mean	91.67%	87.5%	66.67%	81.94%

	A	B	C
1	Tester1	Predicted Word	WER
2	lay green by a 1 row	bin red with c one soon	0.8333
3	place red on b 2 please	place red i b set soon	0.5
4	set blue at c 4 soon	place red i b set soon	0.8333
5	bin white with d 6 again	place red i b set soon	1
6			
7	Tester2		
8	lay green by a 1 row	place red i s four soon	1
9	place red on b 2 please	place red i c four soon	0.6667
10	set blue at c 4 soon	place r with b set soon	0.8333
11	bin white with d 6 again	bin blue with b four soon	0.6667
12			
13	Tester3		
14	lay green by a 1 row	bin green by n five now	0.6667
15	place red on b 2 please	place red at p six please	0.5
16	set blue at c 4 soon	set blue in e nine soon	0.5
17	bin white with d 6 again	bin white with g six again	0.1667
18			

Figure 10: Test Result on GRID Corpus dataset

Student / Sentence	Student A	Student B	Student C	Mean
1	0%	100%	100%	66.66%
2	50%	100%	100%	83.33%
3	100%	100%	0	66.66%
Mean	50%	100%	66.66%	72.30%

	A	B	C
1	Tester1	Predicted Word	WER
2	i love this game	i love this game	0
3	nice to meet you	i love this game	1
4	have a good time	i love this game	1
5			
6	Tester2		
7	i love this game	have this game	0.5
8	nice to meet you	i love this game	1
9	have a good time	i love this game	1
10			
11	Tester3		
12	i love this game	have a good time	1
13	nice to meet you	have a do time	1
14	have a good time	have a good time	0

Figure 11: Test Result on pre-trained dataset

The result shows that a pre-trained dataset has the lowest percentage of word error rate compared to others dataset. The performance of the pre-trained dataset is the best compared to others. The performance of the GRID dataset is low and can be caused by different pronunciation of Malaysian English, as the dataset was created in England. The accuracy is lowest for the MIRACL-V1 dataset. This is due to the dataset has fewer frames and the file is small. Without sufficient frame, the performance of the LipNet model will be affected.

5. CONCLUSION

We have proposed to apply deep learning in this model that maps sequences of image frames of a speaker's mouth to whole sentences. The end-to-end model eliminates the need to segment videos into words before a sentence is predicted. This model does not require handcrafted spatio-temporal visual features nor a separately model.

ACKNOWLEDGMENT

This research work is supported and funded by the Center for Research and Innovation Management (CRIM), Universiti Teknikal Malaysia Melaka (UTeM).

REFERENCES

1. Munday, J., **"Introducing translation studies: theories and applications (4th Edition),"** Routledge, 2016. https://doi.org/10.4324/9781315691862
2. World Health Organization, **"Deafness and hearing loss,"** Accessed: 20 July, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.
3. Kusters, A, Spotti, M. Swanwick, R. & Tapio, E., **"Beyond languages, beyond modalities: Transforming the study of semiotic repertoires,"** *International Journal of Multilingualism*, 14 (3), 219-232, 2017.
4. Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., and Harvey, R., **"Extraction of visual features for lipreading,"** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2), 198–213, 2002. https://doi.org/10.1109/34.982900
5. Graves, A., Fern´andez, S. Gomez, F., Schmidhuber, J., **"Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,"** In *Proceedings of the 23rd International Conference on Machine Learning*, 369–376, 2006.
6. Potamianos, G., Neti, C., Gravier, G., Garg, A. and Senior, A. W., **"Recent advances in the automatic recognition of audiovisual speech,"** *Proceedings of the IEEE*, 91(9), 1306–1326, 2003.
7. C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, Mashari, and J. Zhou, **"Audio-visual speech recognition,"** Center Lang. Speech Process., Johns Hopkins Univ., Baltimore, MD, 2000.
8. Zhou, Z., Zhao, G., Hong, X., and Pietikäinen, M. **"A review of recent advances in visual speech decoding,"** *Image and Vision Computing*, 32(9), 590–605, 2014. https://doi.org/10.1016/j.imavis.2014.06.004
9. Cox, S. J., Harvey, R. W. Lan, Y., Newman, J.L., and Theobald B. J., **"The challenge of multispeaker lip-reading,"** In *AVSP*, 179–184, 2008.
10. Ninomiya, H., Kitaoka, N., Tamura, S., Iribe, Y., and Takeda, K., **"Integration of deep bottleneck features for audio-visual speech recognition,"** In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
11. Hinton, G. E., Osindero, S. and Teh, Y. W., **"A fast learning algorithm for deep belief nets,"** *Neural Computation*, 18(7), 1527– 1554, 2006.
12. Chung, J. S., and Zisserman, A., **"Lip reading in the wild,"** In *Asian Conference on Computer Vision*, 87-103, Springer, Cham, 2016.
13. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T., **"Long-term recurrent convolutional networks for visual recognition and description,"** In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2625–2634, 2015. https://doi.org/10.1109/CVPR.2015.7298878
14. P. Santhi, N. Deeban, N. Jeyapunitha, B. Muthukumaran and R. Ravikumar, **"Prediction of**

Diabetes using Neural Networks," *International Journal of Advanced Trends in Computer Science and Engineering*, 9(2), 985-990, 2020. https://doi.org/10.30534/ijatcse/2020/13922020

15. KollaBhanu Prakash, Lakshmi Kalyani. N, Pradeep Kumar Vadla and Naga Pawan YVR, **"Analysis of Mammography for Identifying Cancer Cells using Convolution Neural Networks,"** *International Journal of Advanced Trends in Computer Science and Engineering*, 9(2), 1184-1188, 2020. https://doi.org/10.30534/ijatcse/2020/44922020
16. Wand, M., Koutnik, J., and Schmidhuber, J., **"Lipreading with long short-term memory,"** In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 6115–6119, 2016.
17. Assael, Y. M., Shilingford, B., Whiteson, S., and De Freitas, N., "**LipNet: End-to-end sentence-level lipreading,"** 2016.
18. Adrian Rosebrock, **"Facial Landmarks with DLIB, OpenCV and Python,"** Accessed: 8 May, 2020. [Online]. Available: https://www.pyimagesearch.com/2017/04/03/facial-landmarks-dlib-opencv-python/
19. Rekik, A., Ben-Hamadou, A. & Mahdi, W., **"A New visual speech recognition approach for RGB-D cameras,"** In *International Conference Image Analysis and Recognition*, 21-28, Springer, Cham, 2014.
20. Ngo, H.C., Ummi Raba'ah, Sek, Y.W., Yogan J. Kumar, Ke, W.S., **"Weeds Detection** in **Agricultural Fields using Convolutional Neural Network**," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, ISSN: 2278-3075, 8(11), 2019.
21. Rahiddin, R. N. N., Hashim, U. R., Ismail, N. H., Salahuddin, L., Choon, N. H., & Zabri, S. N. **"Classification of wood defect images using local binary pattern variants,"** *International Journal of Advances in Intelligent Informatics*, *6*(1), 36-45, 2020. https://doi.org/10.26555/ijain.v6i1.392
22. Joshi, M. A., **"Digital image processing: An algorithmic approach,"** PHI Learning Pvt. Ltd., 2018.
23. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M., **"300 faces in-the-wild challenge: The first facial landmark localization challenge,"** In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 397-403, 2013. https://doi.org/10.1109/ICCVW.2013.59