A Systematic Study and Empirical Analysis of Lip Reading Models using Traditional and Deep Learning Algorithms

R Sangeetha[1*] and D. Malathi[1]

[1]*Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur-603203*

Abstract

Despite the fact that there are many applications for analyzing and recreating the audio through existing lip movement recognition, the researchers have shown the interest in developing the automatic lip-reading systems to achieve the increased performance. Modelling of the framework has been playing a major role in advance yield of sequential framework. In recent years there have been lot of interest in Deep Neural Networks (DNN) and break through results in various domains including Image Classification, Speech Recognition and Natural Language Processing. To represents complex functions DNNs are used and also they play a vital role in Automatic Lip Reading (ALR) systems. This paper mainly focuses on the traditional pixel, shape and mixed feature extractions and their improved technologies for lip reading recognitions. It highlights the most important techniques and progression from end-to-end deep learning architectures that were evolved during the past decade. The investigation points out the voice-visual databases that are used for analyzing and train the system with the most common words and the count of speakers and the size, length of the language and time duration. On the flip side, ALR systems developed were compared with their old-style systems. The statistical analysis is performed to recognize the characters or numerals and words or sentences in English and compared their performances.

Keywords: Audio visual Automatic Speech Recognition, Automatic Lip Reading, Hidden Markov Model,Active Shape Model.

1. Introduction

Humans used to communicate via speech. Every group of people have their own kind of languages. An interpreter is used for communication between different kinds of languages spoken by people. In such cases, the interpreter needs well-versed knowledge in both languages. To overcome this issue lip reading system was evolved. Initially, the traditional system was based on video and voice clippings. The procedure involves feature extraction and segmentation of video for clear recognition of pixel and shape. Noise filtration of audio clip is also more important for analysing the lip reading. Later, deep learning techniques are implemented for enhancing and improved performance for the task. Despite the fact the visual channel is the only source of ear disabled persons and meanwhile the audio channel for eye disabled persons.This interest has achive to the development of coding and decoding of speech and encryption of video.

The knowledge of lip understanding was projected by Sumby in 1954. In the year 1984, the lip reading organization was erected by Petajan, University of Illinois. It turns into a universalperformancetill the late 1980s.The foremost lip-reading method was just to recognize the alphabetic and numeric character recognition. The only source of communication is conversation. The research in lip reading paved a way for Audio-Visual Automatic Speech Recognition (AV-ASR) systems using deep learning. The similar sound is produced for the two different words called as homophones (e.g., blue and blew). Here arises the major challenge for the system that aims to recognize the lip reading other than characters. Speechreading system works based on the horizontal and vertical axis of the lip nodal structure. The face is illuminated with their shadow and the lip is localized. Gray level value is mapped for both the lip image and teeth shown lip image.

The audio-video databases for 1920's different kinds of recognition are used for alphabet recognition, AVLetters database and biggest multi-speaker databases with 295 participants are used for digit recognition. The improvement leads with sentence recognition. The earliest database used is IBM via Voice T M. In 2020's earliest VIDTIMIT that can analysis for 10sentences. AV-TIMIT was the database. The system that initiated from the 1980's to still on progress encompasses the basic terms of image processing and computer vision such as detection of lip and extraction of features and recognition. It is a sequential progress shown in Figure 1.

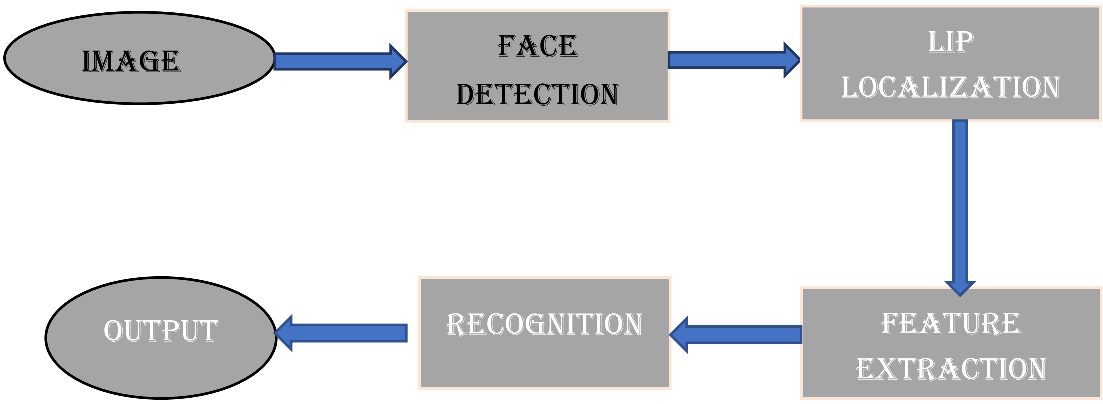

Figure 1: Framework of a Lip Reading Analysis System.

In lip reading analysis system, primarily from the audio visual, the speaker is identified via OpenCV. Once the face detection is done then localization is followed up in order to identify the location of the lips. The next process is to extract the basic information from the image and the lip movement. And from the extracted feature the visual data is found out and classified.

Feature extraction before the recognition falls under two categories: pixel-based and model-based methods. Both the methods were induced. PBE type uses the entire mouth directly. This method is inclined to identify the way spelled out and shadow in correspondence to the light. The other one is MBE that takes the lip delineation and the arguments that describe the contour. And also, it includes the origin of the word and its destruction, zoom and rotation of mouth. The next segment is recognition. The basic recognition evolved during the tradition methods are template matching, Time Warping Dynamically (DTW), Artificial Neural Networks (ANN) and Hidden Markov Model (HMM). Template matching matches the data collected from the previous results whereas the time wrapping means the context is stretched till the length is comfortable with the reference length. Artificial neural network is a non-linear method and the HMM is a statistical approach for recognition.

This successful mechanism can be directly applied to the recurrent network that yields a successful connection between the contents as to include with most of the analysing technologies. Rather than this some of the complex terms such as long-term short memory is applied with the multilayer networks. Then the structure gets elevated. Hybrid level data synchronization is used for the sequential frames so that the key frame weights are more expressed. Effective result is obtained by combining the same network with approach. LSTM neural networks are recurrent neural networks that are designed with a neural that stores the network information for certain duration. It is planned to evade biggest term for a long time which is practically a complex task. It has the skill to append or to remove the data to the cubicle that are controlled by structure gates.

2. Related Works

Eric Petajan et.al.,[1] have developed a lip reading system that overcomes the issues of the previously developed system which uses distance measures and warping time dynamically and quantization. Theyattempted to recognize the speech from multiple persons under ideal circumstances. Initially the image is pre-processed like segmentation, localization and normalization. Furthermore, the word is visually captured and the mouth images vectors were quantized. Frame by frame, the picture vector values are observed. But it fails to recognize the similarpronouncing characters (e.g., B-P). The speech recognized and the result from the visualized section was combined to obtain the result. The author stated that the future work can be addressing the lighting condition, camera angle and recognition through telephone or handset and mike.

In 2018, Kai Xu[26]developed the automatic lip reading with Cascaded attention-CTC that works well with movement parameters like face, lip and facial features. A peer-to-peer neural network-based system for lip recognition that follows encoding strategy. The network that encodes the video frame by frame using fixed 3DCNN. The 3DCNN stores the information and the cascaded CTC decoder is used to generate the text. They have used Character Error Rate (CER), Word Error Rate (WER), and Bilingual Evaluation Understudy (BLEU) to measure the performance ofLCANet.The hidden layers in the neural network eradicates the fault and that improves the performance and it conjunction faster. The state-of-the-art method in deep learning is beforehand implemented technique. The research is established with GRID database. The result analysis shows improvement in 12.3% in BLEU and 1.3% in character error rate and 3.0% in word error rate.

Lip Localization Technique towards an Automatic Lip-Reading approach for Myanmar Consonants Recognition are introduced by Thein et.al.[25]. The system is mainly based on changes of colour and the nearby tracking algorithm. Using SVM classifier the lip movement is characterized for the hearing improvement. In this approach the input image is correlated and stretched for segmentation and next ROI extraction was carried out and the algorithm is applied and followed up by shape information extraction and the information was classified. The classified image obtained was converted as a text. CIELa accuracy rate is found to be more than YCbCr.

Bor-ShingLinet.al., [22] have developed Novel Lip-reading Recognition Algorithm which recognizes the difference between the English language vowels while speaking. The parameter for the system to be framed for analysing is shortlisted so that the accuracy of the result will be higher. Parameters for detecting the breadth, altitude, edge points, range and the dimensions while talking are considered. The flow for analysis is that the image is captured, contrast is enhanced for pre-processing. Face is detected and ROI value is captured and a lip contour was also detected, simultaneouslyundesired objects were removed. Information is extracted and normalized and the vowel is identified. This shortlisted parameter will spot the ROI of the lip and Spotting of the value ROI automatically is attained without prior training. Same system is tested with different circumstantial conditions and the calculated accuracy is 80%.

Lip-reading via Deep Neural Network Using Appearance-based Visual Features anticipated by FatemehVakhshitehet.al. [23]. is stressed out the feature mining and credit of structures. Deep belief network is casted off for extraction. The author used the most complicated dataset CUAVE. This step was taken to overcome the challenges using deep learning. Some of the challenges facedby a lip reading system are combining the whole context sentences through a process, similar words effect and lacking of training statistics for each class. Selection of feature for recognition is also a challenge and the main challenge falls to the speaker. The most important advantage of the system is that the model used for recognition is predictable Hidden Markov Model which means conventional HMM and thus the accuracy is increased by 45%. From the visual stream, the feature is extracted and the belief network is applied whereas it is encoded and the aimed result is decoded to obtain the visual phoneme recognition result.

S.L. Wang et al. [2] developed a real time automatic lip reading system for recognizing isolated English digits from 0 to 9 using parameter set of 14 points in Active Shape Model (ASM).The various methods are used in automatic lip reading system to the colour transformation by using RGB colour space method and then lip region segmentation developed FCMS (Fuzzy C Means with Shape Function) and lip model method uses 14 ASM.A real-time automatic lip reading system has been successfully implemented on a 1.9-GHz PC. The lip reading system provides an average processing rate of 40 frame/sec.

EvangelosSkodras and Nikolaos Fakotakis[9] designed an approach for lip reading that works under the lesser criteria,soit is as an unconstrained method for lip detection in colour images. Generally, the lip movement visual data that are mined for the information can be used in many applications. It provides many features of sound provoking in recognition system. K-means clustering algorithm is used by means of colour clustering in order to find the area of the lips. The information obtained is processed by means of its formation and structure with their specific features and snugged. The mouth corners are tuned for the detection of corners. Two databases CID and GTAV are used. The detection rate is obtained as 97.5%. Furthermore, the research work should be carried to find the inner points of the lip and prominence of teeth and validating its performance in audio-visual appreciations.

WaqqasuRehman Butt and Lombardi L [8] have focused the importance of lip-reading recognition for machines. Artificial programs that take the comment from human voice through pictorial representation requires lip recognition. In order that the author prioritize the Active Appearance Model (AAM) & Hidden Markov Model (HMM), AAM is used for analysing the facial features and their points whereas HMM is used for lip movement and its feature detection. Knowledge-based and Feature invariant, Template matching and appearance based are four kinds of face detection mechanisms whereas lip detection are based on model and image methods. AAM is an extension of active shape model (ASM). This paper shortly compares both the HMM and AAM. The result proves that AAM model are more effective than the other. Reference point in accurate to the location were observed in AAM. For future work, more features are need to be extracted in other sense parts.

Table 1: Summary of works done on Lip Reading Systems

Author/Year/Ref. No.	Methodology	Dataset	Result	Limitation/Future work
R. Seymour, et. al., /2008/[3]	Densely connected convolutional networks (DenseNets)	XM2VTS database	Performance of the model achieved an accuracy of 98%	Proposed model can be implemented in mobile platform

Jamal Ahmad Dargham, et. al., /2008/[4]	pixel intensity normalization method, Neural Network	In-house database, WWW database	Pixel-intensity normalization has low error rate compared to maximum normalization method. Increase in scale factor decrease the error	-
G. Papandreou, et. al., /2009/[5]	Vector Quantization neural network	Ten words of Hindi language	Performed well and fast to different occlusions	Experiment is performed using only 10 words. It can be experimented on other well-known datasets
Jongju Shin, et. al., /2011/[6]	HMM, ANN, and K-NN, neural network classifier	30 isolated Korean word	Achieved an accuracy of 92.67% for dependent on person and 40.06% for independent on person word correct rate	-
N.Puviarasan, et. al., /2011/[7]	k-means clustering	GRID database	Experiments show that proposed method performs well under various environmental conditions	Robustness can be increased for automatic speech recognition
R. Navarathna, et. al., /2011/[8]	Convolutional Neural Network, Long Short-term Memory (LSTM)	OuluVS2	Accuracy of 91.38% is obtained. Compared to current methods proposed system performs significantly	-
Kamil S, et. al., /2013/[12]	Deep Neural Networks (DNN), LDA, SAT and fMLLR	New dataset created containing 12 speakers	Proposed method is viable for speaker-independent lip-reading	Different DNN architectures can be used
Sunil S. Morade, et. al., /2014/[14]	Ergodic hidden Markov model (HMM)	Videos have been recorded from 0 to 9, Cuave database In-house database	Obtained results shows that proposed HMM model with 3 states provides good results with less complexity	-
Jong-Seok Lee/2014/[15]	Visual-speech pass filtering (VSPF)	DIGIT dataset, CITY dataset, e AVletters	VSPF method works effectively in noisy conditions and detected the lip movements	Real world datasets can be utilized for performing the experiments
Ahmed Rekik, et. al., /2014/[16]	3D face pose tracking	BIWI Kinect Head Pose database, MIRACL-VC1, OuluVS, and CUAVE	Proposed system achieved competitive accuracy on various datasets compared to state of art methods	In continuous video flow speech portions can be spotted
M.Z. Ibrahim, et. al., /2015/[17]	Enhanced dynamic time warping technique, convex hull	CUAVE database	Accuracy of 71% is obtained with visual information from lip height	Scale invariant features are not used. Speaker Independent experiments are not performed

Abhishek Jha, et. al., /2016/[19]	WLAS architecture, VGG-M convolution module, attention based sequence-to-sequence LSTM	LRW dataset	An improvement of 95% accuracy is achieved over current baseline methods	-
Dominic Howell, et. al., /2016/[20]	Weighted finite-state transducer	Dataset consisting of videos of 3000 sentences spoken by a single speaker is recorded, RM-3000,ISO-211 Dataset	System is effective in a speaker-independent environment.	-
Ashley D. Gritzman, et. al, /2016/[21]	Adaptive threshold optimization (ATO) algorithm	AR Face Database	Experiments performed with and without ATO. Obtained results show that using ATO produces significant results	As a future work negative examples can also be utilized for ATO matching
Joon Son Chung, et. al., /2018/[24]	Two-stream convolutional neural network	Dataset is developed with millions of words spoken by different people	Proposed model shows modest improvement over state of art models	Architecture can be varied and different lip-reading profiles can be used
Thein Thein, et. al., /2018/[25]	CIELa*b* color transformation, Moore Neighborhood Tracing Algorithm and linear SVM classifier	An own AV database is created for Myanmar consonants	Localization of lip movement has been performed successfully	Support Vector machine can be utilized for better classification of the features
Abderrahim Mesbaha, et. al., /2019/[28]	Hidden Markov Model (HMM), Deep Belief Network (DBN)	CUAVE	Obtained results show that proposed method outperforms HMM baseline recognizer	-
Yuanyao Lu, et. al., /2020/[29]	Convolution Neural Network and Bi-directional Long Short-term Memory	Own database is created containing lip movements of six speakers	Proposed network effectively predicts the words from area of mouth in image sequence	Further study can be performed for complex cases.
AnandHanda, et. al., 2020/[30]	Convolutional Neural Networks and Long Short Term Memory network	LRW (Oxford-BBC), MIRACL-VC1, OuluVS, GRID, and CUAVE	Obtained results shows significance performance in all evaluated datasets	Proposed model can be utilized for medical image segmentation.
Ashley D. Gritzman, et. al., /2021/[31]	Support vector regression (SVR) · Histogram threshold · Shape-based adaptive thresholding (SAT)	AR Face Dataset	Proposed method obtained significant improvement in improving the accuracy of color based lip segmentation	-

Lip Reading Based on Background Subtraction and Image Projection, the research is proposed in 2015 by FatchulArifin et.al. [1] This research focuses mainly on two topics namely Background subtraction and image projection. The first concept is about detecting the movable target in a video. The camera is in motionless position and the moving entity to be detected is matched up with the orientation frame. (E.g. Surveillance camera). The implementation process is of four levels in ANN and they are Gray scaling, image size alteration i.e., resizing in each frame and background subtraction, dimensional projection of image whereas in SVM there are three stages and they are background subtraction with horizontal projection and background subtraction with vertical projection and finally without subtraction with both projections. Artificial neural network and support vector machine are used as classifiers. 5-fold cross authentication is used which means the entire dataset is divided into 5 parts and subdivides to test data and training data. Experimental results prove that the accuracy for ANN is 67%.

Anacquaint with fuzzy logic network that overcomes the disadvantages of neural network classifiers are introduced by in [10]. In general, the steps involved in lip reading are feature extraction from face area and distinguishing the speech. The motive of the fuzzy logic is that it contains memory layer after each layer that contains all the information about the layer and Hierarchical organization is evolved. The ultimate goal is to create the simple networks that forms a dynamic data main class network. It uses the memory unit to hold the signal about the input class. In this, a tactic for speech acknowledgment has been introduced. It is noticed that the performance of the developed model is found to be90% and it needs to be improved further.

A Robust Geometrical-Based Lip-Reading using Hidden Markov Model is developed by [11]. The pre-processing of such as mean face and mouth detection is performed. Then the features such as skin and contour detection and the convex hull detection are extracted. CUAVEdatabase is used and the system produces word recognition up to 68% which provides better performance than predictable appearance-based Distinct Cosine Transform practice. The comparative analysis of Accuracy obtained in various papers is shown in Figure 2.

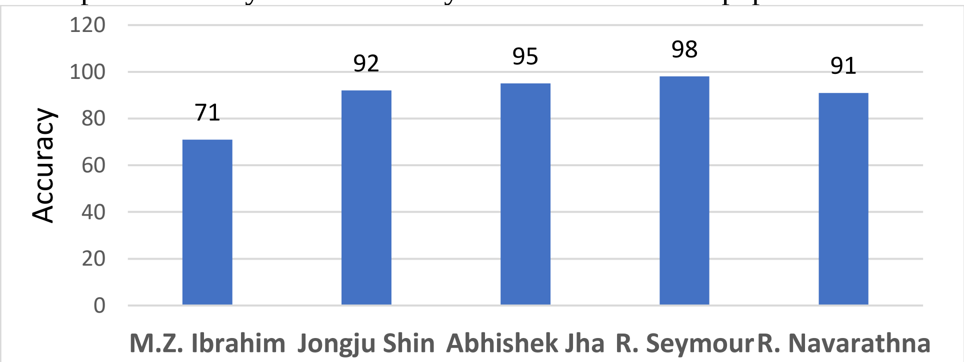

Figure 2: Comparative analysis of various models.

The Digits Datasets used in ALR Systems in various papers is shown in Table 3.

Table 3: Digits Datasets used in ALR Systems

S.No	Dataset Name	Year	Cites	Language	Speakers	Task	Classes
1	XM2VTS	1999	1466	English	295	Digits	10
2	BANCA	2003	507	Multiple	208	Digits	10
3	IBMIH	2004	37	English	79	Digits	10
4	AVOZES	2004	55	English	20	Digits	10
5	CUAVE	2004	248	English	36	Digits	10
6	VALID	2005	33	English	106	Digits	10
7	IBMSR	2008	15	English	38	Digits	10
8	CENSREC-1-AV	2010	20	Japanese	42	Digits	10
9	QuLips	2010	11	English	2	Digits	10
10	AusTalk	2014	6	English	1000	Digits	10

The Alphabets Datasets used in ALR Systems in various papers is shown in Table 4.

Table 4: Alphabets Datasets used in ALR Systems

S.No	Dataset Name	Year	Cites	Language	Speakers	Task	Classes
1	AVLetters	1998	455	English	10	Alphabet	26
2	AV@CAR	2004	26	Spanish	20	Alphabet	26
3	AVICAR	2004	150	English	86	Alphabet	26
4	AVLetters2	2008	44	English	5	Alphabet	26

The WordsDatasets used in ALR Systems in various papers is shown in Table5.

Table 5: Words Datasets used in ALR Systems

S.No	Dataset Name	Year	Cites	Language	Speakers	Task	Classes
1	MIRACL-VC	2014	10	English	15	Words	10
2	AusTalk	2014	6	English	1000	Words	966
3	MODALITY	2015	2	English	35	Words	182
4	LRW	2016	30	English	1000	Words	500

The : Phrases/Sentences Datasets used in ALR Systems in various papers is shown in Table6.

Table 6: Phrases/Sentence Datasets used in ALR Systems

S.No	Dataset Name	Year	Cites	Language	Speakers	Task	Classes
1	IBMViaVoice	2000	295	English	290	Sentences	10,500
2	VIDTIMIT	2002	45	English	43	Sentences	346
3	AV-TIMIT	2004	112	English	233	Sentences	510
4	GRID	2006	520	English	34	Phrases	51
5	IV2	2008	13	French	300	Sentences	15
6	UWB-07-ICAV	2008	9	Czech	50	Sentences	7550
7	OuluVS	2009	164	English	20	Phrases	10
8	WAPUSK20	2010	12	English	20	Phrases	52
9	LILiR	2010	49	English	12	Sentences	200
10	BL	2011	7	French	17	Sentences	238

3. Conclusion:

In this paper, acontemporary analysis on automatic lip-reading techniques has been performed. Active appearance models performed well and achieved more accuracy. Hidden Markov models observed the features of the lip in a sequence since the features are trained and tested. Other deep learning models developed also produce promising results. An analysis on visual feature extraction is also performed. Though many techniques are available still automatic lip reading need to achieve more accuracy. As a future work an ensemble model will be developed to outperform the various models discussed in this paper.

Conflict of interest: The authors declare that they have no known competing financialinterests or personal relationships that could have appearedto influence the work reported in this paper.

References

[1]. Petajan, E., Bischoff, B., Bodoff, D., & Brooke, N. M. (1988).“An improved automatic lip reading system to enhance speech recognition”. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,Washington D.C - CHI ’88. https://doi.org/10.1145/57167.57170

[2]. S. L. Wang, W. H. Lau, S. H. Leung and H. Yan, "A real-time automatic lipreading system," 2004 IEEE International Symposium on Circuits and Systems (IEEE Cat. No.04CH37512), 2004,Vancouver, BC, Canada pp. II-101, https://doi.org/10.1109/ISCAS.2004.1329218

[3]. Seymour, R., Stewart, D., & Ji, M. (2008). Comparison of Image Transform-Based Features for Visual Speech Recognition in Clean and Corrupted Videos. EURASIP Journal on Image and Video Processing, (2008) 810362. https://doi.org/10.1155/2008/810362

[4]. Dargham, J. A., Chekima, A., &Omatu, S.. Lip detection by the use of neural networks. Artificial Life and Robotics, 12/1–2 (2008) 301–306. https://doi.org/10.1007/s10015-007-0494-0

[5]. G. Papandreou, A. Katsamanis, V. Pitsikalis, P. Maragos, Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition,IEEEACMTrans. AudioSpeechLang. Process.17/3(2009)423–435.

[6]. Shin, J., Lee, J., & Kim, D. Real-time lip reading system for isolated Korean word recognition. Pattern Recognition, 44/3 (2011) 559–571. https://doi.org/10.1016/j.patcog.2010.09.011

[7]. N. Puviarasan S. Palanivel, Lip reading of hearing impaired persons using HMM, Expert Systems with Applications 38 (2011) 4477-4481.

[8]. R. Navarathna, T. Kleinschmidt, D.B. Dean, S. Sridharan, P.J. Lucey, “Can audiovisual speech recognition outperform acoustically enhanced speech recognition in automotive environment?” Proceedings of Interspeech, (2011) 2241–2244.

[9]. Skodras, E., &Fakotakis, N. (2011). An unconstrained method for lip detection in color images. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),Prague, Czech Republic. https://doi.org/10.1109/icassp.2011.5946578

[10]. Badura, S., Klimo, M., &Skvarek, O. (2012). “Lip reading using fuzzy logic network with memory” 2012 6th International Conference on Application of Information and Communication Technologies (AICT), Georgia, Tbilisi. https://doi.org/10.1109/icaict.2012.6398471

[11]. Ibrahim, M. Z., &Mulvaney, D. J. (2013). “Robust geometrical-based lip-reading using Hidden Markov models” Eurocon 2013, Zagreb, Croatia. https://doi.org/10.1109/eurocon.2013.6625256

[12]. Kamil S. TALHA, Khairunizam WAN, S.K.Za’ba, Zuradzman Mohamad Razlan and Shahriman A.B, “Speech Analysis Based On Image Information from Lip Movement” IOP Conference Series: Materials Science and Engineering 53 (2013) 1-9.

[13]. Ur Rehman Butt, W., & Lombardi, L. (2013). A survey of automatic lip-reading approaches. Eighth International Conference on Digital Information Management (ICDIM 2013),Islamabad, Pakistan. https://doi.org/10.1109/icdim.2013.6694023

[14]. Morade, S. S., & Patnaik, S.. A novel lip reading algorithm by using localized ACM and HMM: Tested for digit recognition. Optik, 125/18 (2014) 5181–5186. https://doi.org/10.1016/j.ijleo.2014.05.011

[15]. Lee, J. S.. Visual-speech-pass filtering for robust automatic lip-reading. Pattern Analysis and Applications, 17/3 (2014) 611–621. https://doi.org/10.1007/s10044-013-0350-x

[16]. Campilho, A., &Kamel, M. (2014). Image analysis and recognition: 11th International Conference, ICIAR 2014 Vilamoura, Portugal, October 22–24, 2014 proceedings, part II. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8815, 21–28. https://doi.org/10.1007/978-3-319-11755-3

[17]. Ibrahim, M. Z., &Mulvaney, D. J.. Geometrical-based lip-reading using template probabilistic multi-dimension dynamic time warping. Journal of Visual Communication and Image Representation, 30 (2015) 219–233. https://doi.org/10.1016/j.jvcir.2015.04.013

[18]. Arifin, F., Nasuha, A., &Hermawan, H. D. (2015). “Lip reading based on background subtraction and image projection” 2015 International Conference on Information Technology Systems and Innovation (ICITSI). IEEE Indonesia. https://doi.org/10.1109/icitsi.2015.7437727

[19]. Gritzman, A. D., Aharonson, V., Rubin, D. M., &Pantanowitz, A. (2016). Automatic computation of histogram threshold for lip segmentation using feedback of shape information. Signal, Image and Video Processing, 10/5 (2016) 869–876. https://doi.org/10.1007/s11760-015-0834-9

[20]. Howell, D., Cox, S., & Theobald, B.. Visual units and confusion modelling for automatic lip-reading. Image and Vision Computing, 51 (2016) 1–12. https://doi.org/10.1016/j.imavis.2016.03.003

[21]. Gritzman, A. D., Aharonson, V., Rubin, D. M., &Pantanowitz, A.. Automatic computation of histogram threshold for lip segmentation using feedback of shape information. Signal, Image and Video Processing, 10/5 (2016) 869–876. https://doi.org/10.1007/s11760-015-0834-9

[22]. Lin, B.-S., Yao, Y.-H., Liu, C.-F., Lien, C.-F., & Lin, B.-S.. Development of Novel Lip-Reading Recognition Algorithm. IEEE Access, 5 (2017) 794–801. https://doi.org/10.1109/access.2017.2649838

[23]. Vakhshiteh, F., &Almasganj, F. (2017). Lip-Reading via Deep Neural Network Using Appearance-Based Visual Features. 2017 24th National and 2nd International Iranian Conference on Biomedical Engineering (ICBME),Iran, Islamic Republic of Tehran. https://doi.org/10.1109/icbme.2017.8430230

[24]. Chung, J. S., & Zisserman, A. Learning to lip read words by watching videos. Computer Vision and Image Understanding, 173 (2018) 76–85. https://doi.org/10.1016/j.cviu.2018.02.001

[25]. Thein, T., & San, K. M.. Lip movements recognition towards an automatic lip reading system for Myanmar consonants. Proceedings - International Conference on Research Challenges in Information Science, 2018-May(1), 1–6. https://doi.org/10.1109/RCIS.2018.8406660

[26]. Xu, K., Li, D., Cassimatis, N., & Wang, X. (2018). LCANet: End-to-End Lipreading with Cascaded Attention-CTC. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xian, China. https://doi.org/10.1109/fg.2018.00088

[27]. Thein, T., & San, K. M. (2018). Lip localization technique towards an automatic lip-reading approach for Myanmar consonants recognition. 2018 International Conference on Information and Computer Technologies (ICICT),DeKalb, Illinois, United States of America. https://doi.org/10.1109/infoct.2018.8356854

[28]. AbderrahimMesbaha, AissamBerrahoub, HichamHammouchia b, Hassan Berbiab, Hassan Qjidaaa, Mohamed Daoudic, Lip reading with Hahn Convolutional Neural Networks, Image and Vision Computing, 88(2019) 76-83.

[29]. Lu, Y., & Yan, J. "Automatic Lip Reading Using Convolution Neural Network and Bidirectional Long Short-Term Memory". International Journal of Pattern Recognition and Artificial Intelligence, 34/1 (2020) 2054003. https://doi.org/10.1142/S0218001420540038

[30]. Handa, A., Agarwal, R., &Kohli, N. (2020). A multimodel keyword spotting system based on lip movement and speech features. Multimedia Tools and Applications, 79/27–28 (2020) 20461–20481. https://doi.org/10.1007/s11042-020-08837-2

[31]. Gritzman, A. D., Postema, M., Rubin, D. M., &Aharonson, V. Threshold-based outer lip segmentation using support vector regression. Signal, Image and Video Processing, 15/6 (2021) 1197–1202. https://doi.org/10.1007/s11760-020-01849-3