Lip-Reading Techniques: A Review

Sooraj V, Hardhik M, Nishanth S Murthy, Sandesh C, Shashidhar R

Abstract: Lip reading is a skill of determining a person's words by watching lip movements without having heard the sound, or in other words it is a method of determining speech by looking at the movements of the lips. Audio visual speech recognition (AVSR) is an approach that uses image-processing abilities in lip-reading to assist speech recognition systems. It is combination of both audio part and visual part, which implies integration of both lip-reading and speech recognition processes working separately. In this paper, we go through different methods of lip reading and discuss the steps involved in lip reading which includes face detection, lip localization followed by feature extraction and recognition. Audio-visual speech recognition is helpful in an area having audio noise. We look out for performance of hybrid models used for AVSR and trace out the limitations of different approaches which may be helpful for further research in this field. We compare and analyze with various databases of AVSR and their functions, and also discuss the challenges faced, and extend our perceptivity into direction of future research for different types of lip-reading.

Index Terms: Lip-reading, image-processing, face detection, lip localization, feature extraction, audio-visual speech recognition (AVSR), Hybrid Models.

1 INTRODUCTION

In recent trends, pattern-recognition has proved to be an important topic of discussion which emphasizes on the use of computers to mimic people's ideas regarding different items to convey some valuable information. When matched with other recognition systems such as fingerprint, gesture or facial recognition, audio visual speech recognition is more beneficial and robust which makes it important building block of Human Computer interface [22,23]. The other important areas of research in lip-reading are pattern recognition [24,25], image processing and computer-vision [26]. Nowadays, lip reading is becoming very important technique implemented in recognition systems where several lip-reading techniques may be used to improve performance of recognition models. Lip reading finds great applications in the field of information security [27,28], speech recognition[29,30,31] and driver assistance systems[32]. Looking at history of lip reading, we will have to go back to 1954 when Sumby[33] proposed his first work associated with lip reading. Later Petajan[34] introduced a different lip contour reading system which was popular in 1980s. After that there has been a number of researches in the field of lip- reading. Since audio signal is susceptible to noise in the environment, a pixel based method combined with artificial neural network (ANN) was proposed in a recognition model[35] developed in 1989. In 1993, Goldschen and others used Hidden Markov Models (HMMs) in their lip reading systems to achieve sentence recognition rate of 25%[36]. Chiou[37] gave a lip-reading system which used colour motion-video combining snake model, HMM and principal component analysis (PCA) to achieve accuracy of about 94% for 10 words.

- *Sooraj V is pursuing Bachelor of Engineering in Department of Electronics and Communication, JSS Science and Technology University, Mysuru-570006, India. His research interests include speech signal processing and image processing.*
- *Hardhik M is pursuing Bachelor of Engineering in Department of Electronics and Communication, JSS Science and Technology University, Mysuru-570006, India. His research interests include digital signal processing and programming using java*
- *Nishanth S Murthy is pursuing Bachelor of Engineering in Department of Electronics and Communication, JSS Science and Technology University, Mysuru-570006, India. His research interests include CMOS VLSI circuits and digital signal processing.*

For improving performance of continuous lip-reading, a context-based deep neural networks (DNN) system[38] was realized with many layers for visual entities to achieve word accuracy of about 84.7% with a massive 33% increase when compared to baseline HMM. A number of companies and institutions are investing on researches in the field of lip reading. Haar feature and Adaboost cascade classifiers [39] were used to detect the facial gestures and lip movements of the speaker in an open source system invented by Intel. This system has got the ability to enhance word recognition accuracy and processing speed. A kind of computer for lip reading was designed to differentiate between various languages such as German, Arabic, Italian, Polish etc with great accuracy. Google and Oxford universities have discovered tremendous lip-reading software based on artificial intelligence which may be known to find out the lip movements of the speaker on BBC-TV shows. It turned out to be great with 46.8% accuracy when compared to trained lip specialist which was merely 12.8% in a similar test. The organization of the paper is as follows: Section 2-Lip-reading system, Section 3-Database and Section 4-Conclusion.

2 LIP-READING SYSTEM

The existing lip-reading system emphasizes on face detection, lip localization, followed by feature extraction and recognition blocks as shown in Figure 1. After identifying speakers face, lip region has to be found and then information has to be extracted by analyzing movements of the lips.

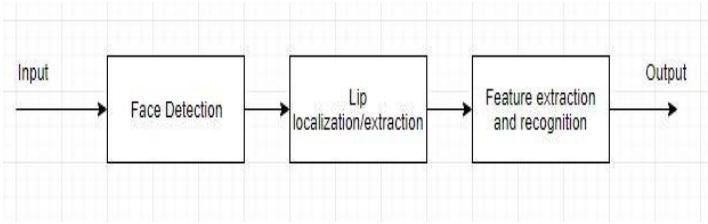

Figure 1: *Block diagram of existing lip-reading system*

Here, first step is to detect the face of the speaker and identify the region of the lips. Next step is to minimize the image data and extract the feature related to movements of the lips. The last step would be to identify the visual data from the extracted lip movement and classify it using a high efficiency classifier.

2.1 Face Detection

Face detection may be visualized as a computer vision issue which involves locating faces in images. Face detection is the primary step in face-biometrics, and its efficiency has a great impact on the performance of additional operations . It is a challenging task for humans to solve, so we have feature-based techniques such as the cascade classifier to solve it. At present deep learning methods have achieved good results on standard face detection datasets. Face detection is generally considered to be the introductory step towards a number of face-oriented technologies like face identification or recognition. Face detection has many beneficial applications. Face detection may be termed as a particular event of object-class recognition. In object-class recognition or detection, the main job is to determine the positions and sizes of all entities in an image that are part of a given class. Face detection algorithms concentrate on the discovery of human faces. Face detection is similar to image recognition in which the image under observation is compared bit by bit. Any face related feature changes in the database will not give a valid comparison. Face recognition has an important role in any sort of face related image-processing applications. Nowadays, a lot of works related to face detection or recognition have been proposed to make it more progressed and efficient.

2.2 Lip Extraction

Lip area extraction is the most significant part of the process to get good recognition rate. Many innovations have been made for extracting facial image from the face. The active appearance model (AAM) is a kind of model in which the shape as well as grey-level appearance (the one which shows only shades of grey colour with no other colours) may be determined. It is hard to directly identify or recognize lip regions because various other parts like moustache, eyes, nose, eyebrows, and body are observed in target image. Face detection and lip localization may be seen in Figure 2.

Figure 2: *Lip region extraction observed for CUAVE database[40]*

Takeshi Saitoh and Ryosuke Konishi[12] proposed active-appearance model (AAM) to extract the area of lips. This model gives the idea of how lip extraction is done in order to interpret the characters. As said earlier, it is hard to directly identify or recognize lip regions because various other parts like moustache, eyes, nose, eyebrows and body are observed in target image. Therefore, we first extract face region from the target image, and then the region of interest is set to identify the lip-region after the face-recognition process. Use of AAM is considered to extract the face and lip regions. In this work, Hidden Markov model along with dynamic programming (DP) matching methods are implemented, where both are shown to be giving high recognition accuracy. The upgraded analysis of lip extraction in real time was proposed by Takeshi Saitoh [14], and in this analysis the videos of changing lip movement were captured using a camera and database was developed. This system is shown to operate in two modes namely the registration and recognition modes. Registration mode is used when the person records a speech sample before recognizing it, and recognition mode is preferred when the person communicates. This paper suggests two automatic processes namely automatic spoken section extraction while the other one being a camera control to decrease the number of steps. In Automatic-spoken section extraction, the system concentrates on phrase recognition, studies the lip shape (closed) that can be seen before and after a letter is spoken. Shape can also be seen when a person speaks some consonant letter. To distinguish the shapes, the threshold time is set. In camera control method they make use of camera in order to extract or capture the image. In Initial-mode we see that region extraction is not applied, instead the rectangular area of 80×80 pixels lying between a 320×240-pixel image is considered, and the extracted rectangular region is utilized. A Lip-reading study of English alphabets as expressed by Filipino Speakers which made use of Image analysis was propounded[15], here English letters pronounced consisted of letters from A-Z. The data gathered were processed using MATLAB in order to convert video into a sequence or collection of images. Image sequences were developed using pre-recorded video for image analysis. Processing was done in MATLAB, where a folder was automatically created to convert video into images. Twelve frames were considered for image processing. Finally, images were represented in .jpg format. Lip detection and extraction was performed using Viola-Jones method, and KLT (Kanade Lucas Tomasi) algorithm was used for the purpose of point-plotting. A procedure for powerful lip-detection and feature extraction which used appearance-based models was explained[4]. This approach was a combination of visual and acoustic information used in the design of an audio-visual speech recognition system, which aimed to enhance recognition rates. The system was segregated into three parts i.e., an acoustic module, a visual module and a sensor-fusion module, and it was tested for different noise sources and acoustic levels. Results indicated that system decreased the error rate when there was noise, and even in the case when powerful noise related acoustic features were considered.

2.3 Feature Extraction methods

Snakes or active-contour models are usually used for shape analysis and object detection by making use of deformable templates [21]. The extracted target contour is transformed into energy minimization to make it optically fit. The comparison between pixel based methods and model based methods for feature extraction[41] and important feature extraction methods are shown in Table 1. Hybrid models are the models which are a combination of two or more methods in order to interpret and analyze the data. These models give more accurate results with high accuracy rate. Hidden Markov model (HMM) is an important statistical method for continuous-sequence categorization like speech recognition, dynamic hand-gesture identification and face related data (facial expressions) recognition.

Typical Methods	Typical Algorithms	Description
Pixel based Methods	a. Direct pixel	This method used the scanning lines centered on the lips as eigen vector, but it is sensitive to changes in light and is weak in high complexity computation.
	b. Image transformation	This method used all pixels transform results as feature vectors, while it takes out high-frequency components which it represents detailed information.
	c. Optical flow Method	This method extracts lip motion parameters and analyses the motion law, but it requires an accurate positioning in the pre-processing.
Model based Methods	a. Deformable template	This method moves close to the object by adjusting the model parameters, but it is easy to fall into local minimum and is sensitive to the initial position.
	b. Snake	This method defines a closed curve to achieve energy minimization, the position of the initial model needs to be determined manually.

Table 1: *Comparison of different algorithms used for feature extraction [41]*

Fatemeh Vakshiteh, Farshad Almasganj, Ahmad Nickabadi[16] proposed lip-reading via deep neural network using hybrid visual features. In this paper they make use of DBN-HMM hybrid models for feature extraction. This paper focuses on lip-reading model having effectively developed processing blocks to recognize highly distinct visual features. In this model, use of structured Deep Belief Network (DBN) oriented recognizer is emphasized. Speaker-independent (SI) and Multi-speaker (MS) works were carried out over CUAVE database, to get phoneme recognition rates (PRRs) of 73.40% and 77.65% respectively. Considering word recognition rates as the point of interest, it was shown that the best values obtained for SI and MS works were 76.91% and 80.25%. Accuracies that may be observed in the results prove that the suggested technique overcomes all disadvantages faced by the conventional Hidden Markov Model (HMM). The phoneme recognition rates(PRR)and word recognition rates(WRR) obtained in this work, were similar and the related accuracies were found to be high. An appearance based feature extraction process[1] which included Deep Belief Network(DBN) supported recognizer was introduced. It was known to perform better than HMM baseline recognizer. Visual based features were extracted in the automatic speech recognition system to give a baseline accuracy of 29.8%. Another interesting point of visual features is that using them as inputs resulted in best DBN architecture achieving an accuracy of 45.63%.The system of continuous AVSR built using hybrid ANN-HMM model[2] was proposed by Martin Heckmann and others. Audio extraction was done using RASTA-PLP and video extraction by a chroma-keying process, where the lips were coloured blue so that it may be located easily and extracted in real time. Continuous word recognition using this hybrid model gave good recognition results compared to pure HMM systems. AAM is a hybrid method which combines both pixel and model-based methods. The advantage of this model is that from any position or angle it is able to recognize the words, making use of the extracted lip data. It describes the gray level change of object with a collection of model variables to detect the lips[18]. Active-shape models(ASM) are numerical models of the structures of the objects. The set of labelled landmark points (reference) are taken as a parameter to define the shape of the object. X and Y coordinates are used to locate each landmarked point. Principal component analysis is used for building statistical shape model by taking a trained set of reference objects in images. The shape of an object deviated from the mean shape is detected from the eigen values and vectors of a covariance matrix. Discrete Cosine Transform (DCT) is broadly used in image and signal processing because of data-compression property[20]. It uses cosine-based function to transform the input into low frequency component of an image. The pixel colour, intensity, corners, edges are the features recognized from the image related detection procedure. This is also called as colour-based methods because of colour difference between face and lips. RGB model consists of red, green and blue components which are transformed and filtered using high pass filter (HPF) and converted into binary image to recognize the lip[21].The hue value difference between lip pixel and face pixel is used as a criteria to recognize the lip in HSV (Hue saturation value) model. In YCbCr model, the differences in blue and red chroma component is used as a fact to locate the lip. As lips are usually known to be the reddest (or pink) part of our face having Cr value in the range 140 to 165 and Cb value in the range 140 to 195[19], it may be used as a colour component in the process of identification. The lip reading system using HMM where Direct Cosine Transform(DCT) and Discrete Wavelet Transform (DWT) related features were extracted [9] from the mouth region and compared with DCT plus DWT related features which inferred that HMM with DWT related features gave good results showing 97% performance when compared to HMM with DCT which gave only 91%. The main objective of this paper was to improve communication between a normal and a hearing-impaired person. DCT plus DWT based features were taken from the mouth part.

2.4 Recognition models

If we consider recent researches in the field, there are different methods of lip reading like Dynamic Time Warping (DTW), template matching, Hidden Markov model (HMM) and Artificial Neural Networks (ANN). The description regarding different recognition methods[41] is as shown in Table 2.

Classification	Description
Template Matching	This method is based on static image and ignores the change in lip dynamic characteristics and words and sentences recognition rate are very poor.
Dynamic Time Warping (DTW)	This method can solve the inconsistency of pronunciation length and speech speed,but require accurate starting point primitively.
Artificial Neural Networks (ANN)	This method has the ability to imitate human cognitive system,but lacks solid mathematical theory and needs the help of experience.
and Hidden Markov model (HMM)	This method is suitable for sequence classification due to its capabilities in modelling and analyzing temporal process.

Table 2: *Recognition models[41]*

HMM is the appropriate model for correctly depicting the

information related to movement of the lips. The training and testing phases of HMM system includes extraction of features using DWT or DCT from mouth part, which may then be given as inputs to figure out the variables of the system, after which the word may be identified during testing phase. We can look out for a two-channel technique built for HMM in Figure 3.

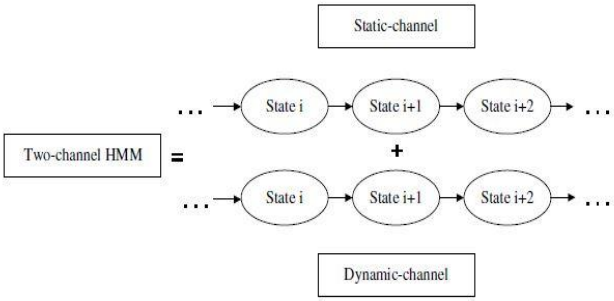

Figure 3: *Block diagram of two channel HMM*

2.5 Audio visual speech recognition system

Audio visual speech recognition (AVSR) is an approach using image processing techniques in lip reading to assist speech recognition systems. Figure 4 illustrates the fundamentals of AVSR system which emphasizes on combining audio and video processing of speech signal.

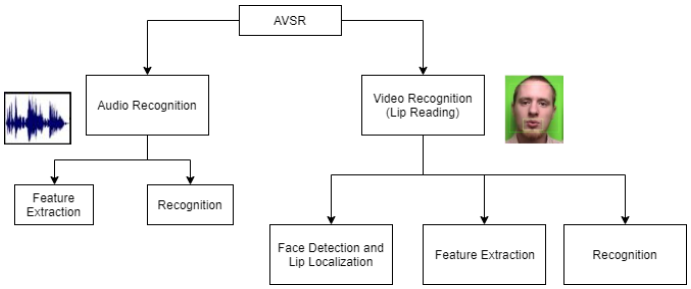

Figure 4: *General illustration of an Audio-Visual Speech Recognition (AVSR) system*

The Audio-Visual Speech Recognition related to multi stream DBN Models with articulatory features was initiated [13], and this method presents articulatory features being supported for a multi stream dynamic Bayesian network for audio-visual speech recognition. When compared to state synchronous and asynchronous DBN models, this system is found to be more efficient with better recognition rates. The performance of the system may be good even though there is presence of noises introduced by audio. State synchronous and asynchronous DBN models had recognition rates of 87.02% and 88.32%, but DBN models designed using articulatory features enhanced recognition rates to 89.38%.There is another approach which emphasizes on discriminative training of HMM stream exponents for AVSR[17]. In this paper we can find the use of a generalized probabilistic descent (GPD)algorithm in order to find out hidden Markov model (HMM) stream exponents necessary for AVSR. By combining two single streams of HMM and by adding exponents related to each stream it is possible to design a dual stream HMM. The issue with respect to multi stream HMM is training the exponent with respect to AVSR system. Good performance and gains were obtained in bimodal ASR when compared to a single stream HMM. Ara V. Nejian[3] and others proposed a unique approach for audio-visual speech recognition, which used a coupled hidden markov model (CHMM). This model is different from HMM in the way that video and audio sequences are considered distinctly and it is not required to merge both the observations, which is a difficult task. CHMM based systems perform better than multi-stream HMM based systems by achieving better recognition rates. In addition to that, probabilities for visual and audio streams can be determined separately, so this model can be considered appropriate for machines supporting parallel processing. A standard CHMM [3] used in Audio Visual Continuous Speech Recognition (AVCSR) may be seen in figure 5. Here classical coupled HMM (CHMM)consists of a number of HMMs. There are three types of nodes: mixture nodes, observation nodes and backbone nodes. Continuous observable nodes are shown by observation nodes while the hidden discrete nodes are shown by the mixture and backbone nodes. Comparing this strategy with ASR and VSR systems, it could be seen that in both cases accuracy can be increased by 10% and 20% respectively.

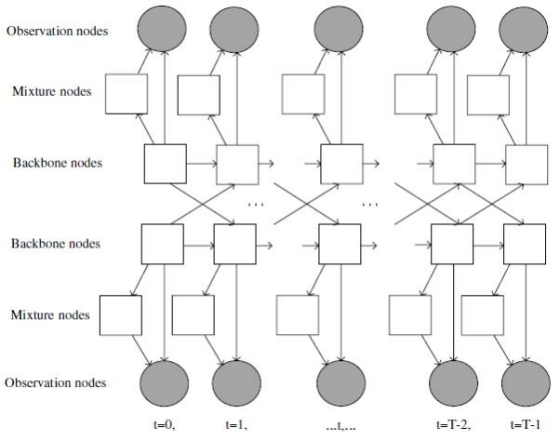

Figure 5: *Audio visual coupled HMM*

The method for Audio-Visual Speech Recognition integrating facial depth information captured by the Kinect was given by Georgios Galatas[6] and others. It used BAVCD(Bilingual Audio Visual Corpus with Depth information) database which consisted of spoken words in both English as well as Greek. Feature extraction was done by using an appearance-based approach, by applying DCT on the mouth image extracted from each video and depth frame. Two stage LDA applied to visual and depth features gave rise to increase in recognition accuracy. The lip reading system that concentrated on well-known multi-stream HMM was proposed[5]. Lip-reading was done using modular BDPCA related feature extraction procedure. After this extraction, Visemic LDA was employed to form end visual features. The lip-reading system was efficient enough at low SNR achieving an accuracy of 68.13%. Both lip-reading and audio-only speech recognition processes had great word identification rates in clear environment, but experimental results showed that lip-reading system performed better than audio-only system in a noisy environment. Ahmad B. A. Hassanat[7] described VSR (visual speech recognition) as a speaker-dependant problem. This inference was drawn by comparing the word error rates (WER) of both speaker-dependant as well as speaker-independant experiments. It was found to be 76.38% and 33% respectively. Speaker dependant experiments gave better results than speaker independent experiments. An approach to

simultaneous speaker-speech recognition, where it was concluded that speaker-dependant audio recognizers performed better than the speaker-independent recognizers [8]. A unique method for speech-recognition that captures audio-visual sensor array was given by Hari Krishna Maganti[10] and others. Audio-Visual multiple speaker tracer was preferred to differentiate speakers with good accuracy. It was taken as input to a super directive beamformer to enhance speech signal. Lip motion features may be extracted for Speech-Reading and Speaker-Identification[11] using a two-stage discriminative feature extraction method.

3 DATABASE COMPARISON

Database is an important foundation for lip-reading systems because it may have a direct impact on the recognition rates of the systems. It is very hard to construct a standard and practical database. The comparison between different databases used for lip reading systems may be seen in Table 3.

Language	Databases	Year	Resolution	FPS	Content	Additional Features
English	TULIPS	1995	100*75	30	Numerals 1-4	No
	XM2VTS	1999	720*576	25	3 sentences	Head Rotations glass, hats
	AVLetters	2002	376*288	25	Letters A-Z	Moustaches
	CUAVE	2002	720*480	30	7000 utterances	Simultaneous speech
	VIDTIMIT	2002	512*384	25	10 sentences	Different countries
	Grid	2006	720*576	25	1000 sentences	No
	OuluVs	2009	720*576	25	10 phrases	Head Rotations
Chinese	CAVSR	2001	352*240	30	78 single syllable	No
	HIT	2005	256*256	25	96 phrases and 200 sentences	No
French	M2VTS	1997	286*350	25	Numerals 0-9	Head Rotations

Table 3: *Comparison between existing databases[41]*

3.1 CUAVE

CUAVE(Clemson University Audio-Visual Experiments) database was captured by E.K. Patterson, Department of Electrical & Computer Engineering, Clemson University, United States. This database was shot with a resolution of 720 x480 in an isolated sound booth with 1MP-CCD (charged coupled device) camera (having 29.97 fps NTSC Standard). It had two parts: one of individuals and the other one of speaker pairs. Till date, most work carried out using this database is known to have low resolution and pre-segmented video considered only for lip part.

3.2 TULIPS

It is a tiny audio-visual database which consists of 12 subjects indicating the initial 4 digits of English alphabets. Subjects are developed by students pursing undergraduate courses from Cognitive Science Program at UCSD. This database is known to be gathered at R. Movellan's laboratory, Department of Cognitive Science, UCSD. Sunil S. Morade, Suprava Patnaik[40] considered CUAVE and TULIPS databases for experimentation and comparison of results obtained using different methods and concluded that SVM performed better than the rest for CUAVE database.

4 CONCLUSION

This paper explains different methods of lip reading and discusses the steps involved in lip reading which includes face-detection, lip-localization followed by feature extraction and recognition. Performance of hybrid models used for audio-visual speech recognition (AVSR) has been assessed for different approaches so that it may help for further research in the field.

5 REFERENCES

[1] Fatemeh Vakhshiteh, Farshad Almasganj, “Lip-reading via Deep Neural Networks using appearance based visual features”, 2017 24th national and 2nd International Iranian Conference on Biomedical Engineering (ICBME), Amirkabir University of Technology, Tehran, Iran, 30 November - 1 December 2017

[2] Martin Heckmann, Frederic Berthommier, Kristian Kroschel, “A Hybrid ANN/HMM Audio-Visual Speech Recognition System”, AVSP 2001 International Conference on Auditory-Visual Speech Processing.

[3] Ara V. Nejian, Luhong Liang, Xiaobo Pi, Liu Xiaoxiang, Crusoe Mao and Kevin Murphy, “A Coupled HMM for Audio-Visual Speech Recognition”, 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing.

[4] Stephane Dupont and Juergen Luettin, “Audio-Visual Speech Modeling for Continuous Speech Recognition”, IEEE Trans. On Multimedia, Vol. 2, No. 3, September 2000

[5] Guanyong Wu, Jie Zhu, “Modular BDPCA based Visual Feature Representation for Lip-Reading”, 2008 15th IEEE International Conference on Image Processing

[6] Georgios Galatas, Gerasimos Potamianos, Fillia Makedon, “Audio-Visual Speech Recognition incorporating facial depth information captured by the Kinect”, 20th European Signal Processing Conference (EUSIPCO 2012), Bucharest, Romania, August 27 - 31, 2012

[7] Ahmad B. A. Hassanat, Visual Speech Recognition, Speech and Language Technologies - Edited by Prof. Ivo Ipsic, ISBN 978-953-307-322-4, Publisher InTech, Published in print edition June, 2011

[8] E. K. Patterson, J. N. Gowdy, “An Audio-Visual approach to simultaneous speaker-speech recognition”, 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).

[9] N. Puviarasan, S. Palanivel, “Lip reading of hearing impaired persons using HMM”, Expert Systems with Applications 38 (2011) 4477–4481, 2010 Elsevier Ltd. - Science Direct

[10] Hari Krishna Maganti, Iain McCowan, “Speech Enhancement and Recognition in Meetings With an Audio–Visual Sensor Array”, IEEE Trans. Audio, Speech, And Language Processing, Vol. 15, No. 8, November 2007

[11] H. Ertan Çetingül, Yücel Yemez, Engin Erzin and A. Murat Tekalp, “Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading”, IEEE Trans. Image Processing, Vol. 15, No. 10, October 2006

[12] Takeshi Saitoh, Ryosuke Konishi “A Study of Influence of Word Lip Reading by Change of Frame Rate”. ISCA

Archive on Audio-Visual Speech Processing :speech.org/ archiveHakone, Kanagawa, Japan"

[13] Dong-mei JIANG, Peng WU WANG, Hichem SAHLI, Werner VERHELST "Audio Visual Speech Recognition Basedon MultiStream DBN Models with Articulatory Features", ISBN 978-1-4244-6245-2/10/$26.00 ©2010 IEEE

[14] Takeshi Saitoh "Development of Communication Support System Using Lip Reading" IEEJ Transactions On Electrical And Electronic Engineering, IEEJ Trans 2013; 8: 574–579 Published online in Wiley Online Library (wileyonlinelibrary.com). DOI:10.1002/tee.21898

[15] Cruz, Hans Miguel Puente, Jofet Kane T Santos, Christian, Vea Larry A., Rajendaran Vairavan "Lip Reading Analysis of English Letters as Pronounced by Filipino Speakers Using Image Analysis" 1st International Conference on Green and Sustainable Computing (ICoGeS) 2017 IOP PublishingIOP Conf. Series: Journal of Physics: Conf. Series 12345678901019 (2018) 012041 doi :10.1088/1742-6596/1019/1/01204

[16] Fatemeh Vakhshiteh, Farshad Almasganj, Ahmad Nickabadi "Lip-Reading via deep neural networks using hybrid visual features", Image Anal Stereol 2018;36:159-171 doi: 10.5566/ias.1859 Research Paper

[17] Gerasimos Potamianos and Hans Peter Graf "Discriminative Training Of Hmm Stream Exponents For Audio-Visual Speech Recognition" AT&T Labs-Research, 180 Park Ave, Florham Park, NJ 07932-0971, U.S.A. Labs-Research, 100 Schulz Drive, Red Bank, NJ 07701-7033,U.S.A.email:{makis ,hpg}@research.att.com

[18] L. R. Aran, F. Wong and L. P. Yi, "A review on methods and classifiers in lip reading," 2017 IEEE 2nd International Conference on Automatic Control and Intelligent Systems (I2CACIS), Kota Kinabalu, 2017, pp. 196-201.

[19] Rein-Lien Hsu, M. Abdel-Mottaleb and A. K. Jain, "Face Detection In Color Images", in IEEE Trans. On Pattern Analysis And Machine Intelligence, vol. 24, no. 5, pp. 696-706, May 2002.

[20] Salma Pathan et al, "Recognition of spoken English phrases using visual features extraction and classification" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (4) ,2015, 3716-3719

[21] Indian Journal of Science and Technology, Vol 9(32), DOI: 10.17485/ijst/2016/v9i32/98737, August 2016

[22] P. Polycarpou, A. Andreeva, A. Ioannou et al., Don't Read My Lips: Assessing Listening and Speaking Skills Through Play with a Humanoid Robot, in Int. Conf. Human-computer Interaction, (2016), pp. 255–260.

[23] J. Shin, H. I. Kim and R. H. Park, New interface for musical instruments using lip reading, Image Process. Iet, 9(9) (2015) 770–776.

[24] S. Tamura, H. Ninomiya, N. Kitaoka et al., "Audio-visual speech recognition using deep bottleneck features and high-performance lipreading", Asia-pacific Signal & Information Processing Association Summit & Conf., 2015, pp. 575–582.

[25] T. Watanabe, K. Katsurada and Y. Kanazawa, "Lip reading from multi view facial images using 3D-AAM", Asian Conf. Comput. Vis. (2016) 303–316.

[26] M. Baart and A. G. Samuel, Turning a blind eye to the lexicon: ERPs show no cross-talk between lip-read and lexical context during speech sound processing, J. Memory Language 85 (2015) 42–59.

[27] F. S. Lesani, F. F. Ghazvini and R. Dianat, Mobile phone security using automatic lip reading, in 9th Int. Conf. e-Commerce in Developing Countries: With Focus on e-Business (ECDC), 2015, pp. 1–5.

[28] S. Mathulaprangsan, C. Y. Wang, A. Z. Kusum et al., "A survey of visual lip reading and lip-password verification", Int. Conf. Orange Technologies 2015, pp. 22–25.

[29] D. Bahdanau, J. Chorowski, D. Serdyuk et al., "End-to-end attention-based large vocabulary speech recognition", Comput. Sci. (2016) 4945–4949.

[30] J. T. Huang, J. Li and Y. Gong, "An analysis of convolutional neural networks for speech recognition", in IEEE Int. Conf. Acoustics, 2015, pp. 4989–4993.

[31] Y. Miao, M. Gowayyed and F. Metze, EESEN: "End-to-end speech recognition using deep RNN models and WFST-based decoding", in Automatic Speech Recognition and Understanding, 2016, pp. 167–174.

[32] C. Hyunmin, C. M. Kang, B. Kim et al., "Autonomous Braking System via Deep Reinforcement Learning", 2017.

[33] W. H. Sumby and I. Pollack, Erratum: Visual contribution to speech intelligibility in noise, J. Acoust. Soc. Am. 26(2) (1954) 212–215.

[34] E. D. Petajan, "Automatic lipreading to enhance speech recognition", Proc. IEEE Communication Society Global Telecommunications Conf. (Atlanta, Georgia, 1984), pp. 26–29.

[35] B. P. Yuhas, M. H. Goldstein and T. J. Sejnowski, Integration of acoustic and visual speech signals using neural networks, IEEE Commun. Mag., (1989) 65–71.

[36] A. J. Goldschen, O. N. Garcia and E. D. Petajan, Continuous Automatic Speech Recognition by Lipreading (George Washington University, 1993), pp. 321–343.

[37] G. I. Chiou and J. N. Hwang, "Lip-reading by Using Snakes, Principal Component Analysis, and Hidden Markov Models to Recognize Color Motion Video", IEEE Trans. Image Processing. 6(8) (1997) 1192–1195.

[38] K. Thangthai, R. Harvey, S. Cox et al., "Improving Lip-reading performance for robust audiovisual speech recognition using DNNs", in Faavsp-the Joint Conf. Facial Analysis, 2015.

[39] Available at: https://arxiv.org/pdf/1611.05358v1.pdf.

[40] Sunil S. Morade Suprava Patnaik (2015), "Comparison of classifiers for lip reading with CUAVE and TULIPS database". Optik - International Journal for Light and Electron Optics, 126(24), 5753–5761.

[41] Yuanyao Lu, Jie Yan and Ke Gu,Review on "Automatic Lip Reading Techniques" , International Journal of Pattern Recognition and Artificial Intelligence Vol. 32, No. 7 (2018) 1856007