ARTIFICIAL INTELLIGENCE: A SURVEY ON LIP-READING TECHNIQUES

Ms. Apurva H. Kulkarni
Dept. of Computer Engineering
J.T. Mahajan College of Engineering
Dist-Jalgaon, India
apurvakulkarni152@gmail.com

Dr. Dnyaneshwar Kirange
Dept. of Computer Engineering
J.T. Mahajan College of Engineering
Dist-Jalgaon, India
dkirange@rediffmail.com

Abstract— Lip reading is a visual way of "listening" to someone. This is done by looking at the speakers face to follow their speech patterns in order to recognize what is being said. Lip-reading technology mainly includes face detection, lip localization, feature extraction, training the classifier through corpus and finally recognition of the word/sentence through lip movement. An intelligent system will be trained by giving user's lip-movement frames sequences as input and will identify lip movement and the said word using either visual information or both audio and visual information. Deep learning is an emerging branch of artificial intelligence which mimics the human brain. It has different layers in the model which is used to process minute details like neurons in brain. This paper mainly focuses on the survey of different lip reading techniques and different language datasets in the era of deep learning. Various Automatic lip reading techniques are discussed and summarized.

Keywords— Learning systems, Neural networks, Artificial intelligence, Speech Recognition, Databases

I. INTRODUCTION

In our world there are so many different languages so this task of lip reading is not generic. Lip-Reading requires a great deal of concentration when done by a human. Biometric security is present in many modern devices, the most common of which is fingerprint authentication. These authentication systems aren't as fool proof as they seem. It is very difficult to mimic or forge the Lip movement hence it is more secure. There are plenty of techniques evolved for lip reading like in the fields of image processing, AI, machine learning and recently deep learning and still the work is going on. Writing computer code that can read lips is cumbersome so it is better to form model using AI, where computers learn from data and predict results. A system is feed with thousands of hour of videos and transcript and learns the patterns and makes predictions. Audio and visual lip reading datasets are available. Deep learning is currently used in most common face recognition, handwriting recognition, NLP processing and speech recognition software. Popular Deep Learning libraries such as Keras, PyTorch, and Tensorflow are used widely in industry today. Lip reading is a bimodal consisting of audio and visual components. Recently using the visual clues in combination with the sound clues are used for improved speech recognition. In the noisy environment when voice is not audible in that case visual clues can help and can improve the accuracy rate. For building this system different face detection, feature extraction, deep learning models and dataset need to be reviewed. In this paper the literature survey regarding different automatic lip recognition system is done and data is collected. Different models, datasets, summary is given in the next section.

II. LITERATURE SURVEY

A lip reading system mainly consists of three parts: lip detection and localization, lip feature extraction and lip reading recognition [38].

A. A Lip detection and localization methods:

The lip localization and detection techniques are Gray/color information-based methods where color information is used to locate lips, Geometric information-based methods where rough region of mouth is calculated corresponding to the proportion of the face[38].

B. Feature Extraction Method

Traditionally, there are two categories of the feature extraction method in a lip reading system: pixel-based methods and model-based methods [38].

The pixel based methods are direct pixel method, Image Transformation method and optical flow method.
In direct pixel method lip key points are marked and then identified. Image transformation method goal is to exploitation of statistical characteristics of the image (i.e. high correlation, redundancy). Some Image transformation techniques are Fourier Transform (FFT, DFT, and WFT) Discrete Cosine Transform (DCT) Walsh-Hadamand Transform (WHT) Wavelet Transform (CWT, DWT, FWT). In optical flow method apparent motion of lips are identified. Lip Motion parameters will be extracted and analyzed. In Model Based method Parameters based on lip counter will be obtained and send to the classifier. Various Models are Deformable template model and snake model.

Other Feature extraction methods include novel analysis to determine which areas (patches) of the mouth ROI are the most informative for visual speech [36]. It is determined that a particular area of the ROI can be more useful in terms of lip-reading.

C. Recognition models

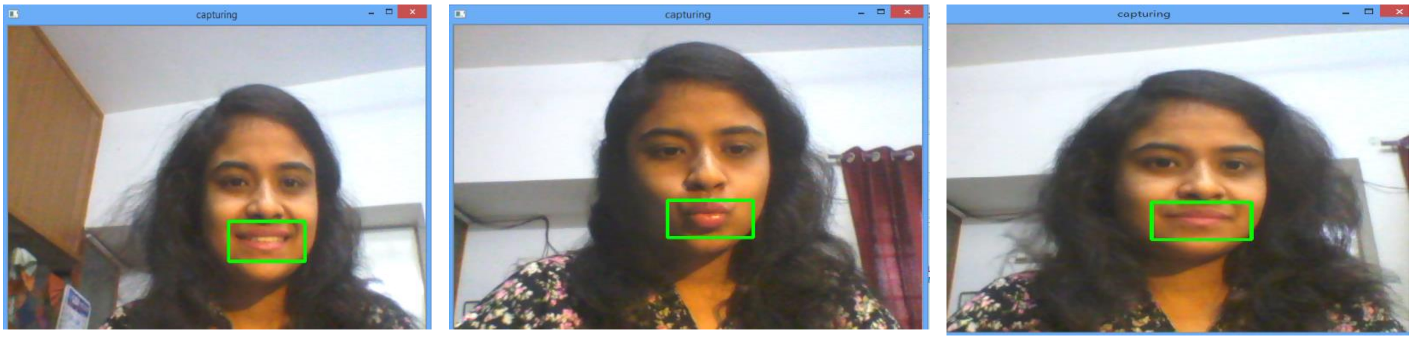
The recognition methods of lip reading are template matching, Hidden Markov model (HMM) Dynamic Time Warping (DTW), Artificial Neural Networks (ANN), DL architectures.

The Challenges in lip reading:

1. Phonemes: - characters that produces same sound.. e.g. buy, by
2. Other challenges associated to lip-reading include Light conditions, conditions, head pose variations, poor temporal resolution.
3. Classifying them by task (e.g. letters, digits, words and sentences) and by viewing angle.
4. To decode speech in multi-view lip-reading realistic scenarios
5. The available databases dire in several aspects, such as number of speakers, language, number of utterances and spatial and temporal resolutions.
6. DNNs needs big amounts of training data.
7. Creation of generalize dataset.

The motivation of this paper is to help deaf person in lip reading and it can also help interaction in a noisy environment. It can help patients with vocal cord trouble or throat injuries. It can help analyze video footage of CCTV camera and could provide key insights into what someone is saying and what is actually happening. There and many applications which can be developed after successful implementation of this Technology.

Figure 1: Region of interest (Lip area detection for different lip movement)



For Lip Reading first we need to identify the region of interest (ROI). Detection of the lip region is very important step. As in the above diagram the lip detection for various lip movement should be identified with a good accuracy. After successful identification of the Lips i.e. region of Interest lip movement is identified in the next step.

Below Table shows different Research paper based on automatic lip reading, different datasets, Face detection and localization techniques are discussed.

Table 1: Overview of Lip-reading research paper techniques.

Author	Method Used	Findings
Muhammad Rizki Aulia Rahman Maulana, Mohamad Ivan Fanany[1]	CNN + GRU AVID DATASET	Indonesian lip reading model is the first sentence- level which can handle variable-length input, as to make it applicable in real world settings.
Parth Khetarpal, Riaz Moradian , Shayan Sadar , Sunny Doultani, Salma Pathan [2]	CNN GRID dataset	Model provides accurate recognition results even when only limited training data is available.
Aparna Brahme, U. Bhadade [3]	IPC (International phoenetic chart)	phonemes visems mapping for Marathi language
Joon Son Chung, Google DeepMind [4]	CNN and GRID dataset	Created own WLAS network LRS dataset. It operates on video input, audio input or both.
Sanaullah Manzoor, Muhammad Faisal [5]	(STCNNs), (RNNs) (CTC) , LSTMs and categorical cross entropy loss	Design an audio-visual lip-reading system for Urdu language. contributed urdu corpus.

Amit Garg, Jonathan Noyola, Sameep Bagadia[6]	CNN LSTM	Proposed several new methods for performing visual speech recognition on sequences of color images with variable length.
Kuniaki Noda ,Yuki Yamaguchi Kazuhiro Nakadai Hiroshi G. Okuno Tetsuya Ogata[7]	HMM, CNN, MFCC	AVSR system based on deep Learning architectures for audio and visual feature extraction and an MSHMM for multimodal feature integration and isolated word recognition.
Xinjun Ma, Hongjun Zhang and Yuanyuan Li[8]	LBPH	Improved the accuracy of LBP algorithm
Adriana Fernandez-Lopez, Federico Sukno[9]	Deep learning models Different datasets	Deep learning models (CNN, DBN, LSTM) and datasets survey of different languages and tasks.
Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk [10]	SVM, MFCC	Speech Recognition using MFCC extraction for performing word recognition.
Ivan Fung, Brian Mak [11]	CNN+bidirectional LSTM + Maxout activation units.	End-to-end low-resource lip-reading system that does not require any separate feature extraction stage nor pre-training phase with external data resources.
Y. Mroueh, E. Marcheret, V. Goel [12]	Scattering coef, LDA, Feedforward , IBM AV-ASR dataset	Sentence recognition and improved phoneme accuracy
Y. M. Assael, B. Shillingford, S. Whiteson, N. de Freitas[13]	3D-CNN, Bi-GRU, GRID corpus	Speech recognition performed for phrases and detection accuracy is improved
J. S. Chung, A. Zisserman [14]	CNN+ LSTM attention, OuluVS2, MVLRS	Lip reading performed for phrases and sentences. Convolutional neural network improves accuracy.
M. H. Rahmani, F. Almasganj[15]	ASM, HMM, Cuave, DBNF,DNN	Three models developed with combination of techniques for digits database and Phonemes accuracy is improved
G. Sterpu, N. Harte[16]	DCT,HMM,TCD-TIMIT	Viseme accuracy is improved using Speaker dependent dataset used
K. Thangthai, R. Harvey[17]	PCA,LDA,DNN-HMM, TCD-TIMIT	Lip reading performed on sentence level
E. K. Patterson, S. Gurbuz, Z. Tufekci, J. N. Gowdy[18]	Feed forward, lstm, GRID	Defines the relationship between data fusion in the presence of audio noise and demonstrates that optimal data fusion can only be performed if both the noise level and type are considered.
K. Xu, D. Li, N. Cassimatis, X. Wang[19]	3d+CNN, Bi-GRU+attention, GRID corpus	Lip reading based on phrase level and has achieved high accuracy in recognition.
Stavros Petridis,Themos Stafylakis, Pingchuan Ma , Feipeng Cai, Georgios Tzimiropoulos, Maja Pantic [20]	3d+CNN, Bi-GRU, ResNet, LRW	A slight improvement in the classification rate over an end-to-end audio-only and MFCC-based model is reported in clean audio conditions and low levels of noise. In presence of high levels of noise, the end-to-end audio visual model significantly outperforms both audio-only models
C. Sui, R. Togneri, M. Bennamoun[21]	CHAVF, SVM, ouluVS	Evaluates the different characteristics of planar and stereo visual features, and we first show that using the stereo feature along with the planar feature can significantly boost the accuracy on a large-scale audio-visual data corpus.
D. Howell, S. Cox, B. Theobald[22]	AAM, CD-HMM, RM3000	Show a small but statistically significant improvement in recognition accuracy.
M. Gurban, J.-P. Thiran[23]	DCT,LDA,HMM,CUAVE	Perform better than linear discriminant analysis, the most usual transform for dimensionality reduction in the field, across a wide range of dimensionality values and combined with audio at different quality levels.
Y. Takashima, R. Aihara, T. Takiguchi, Y. Ariki, N. Mitani, K. Omori, K. Nakazono[24]	CBN,HMM,ATR	Word-recognition experiments in noisy environments, where the CBN-based feature extraction method outperformed the conventional methods
M. Zimmermann, M. M. Ghazi, H. K. Ekenel, J.-P. Thiran[25]	PCA,LSTM,HMM,OuluVS2	Proposed method has outperformed the baseline techniques applied to the OuluVS2 audio visual database for phrase recognition with the frontal view cross-validation and testing sentence correctness reaching 79% and 73%
T. Afouras, J. S. Chung, A. Zisserman[26]	3D-CNN,ResNet, Bi-LSTM, depth wise CNN, attention encoder , LRS.	Model improves the state-of-the-art word error rate on the challenging BBC-Oxford Lip Reading Sentences 2 (LRS2) benchmark dataset by over 20 percent.

Literature survey of Few English datasets are listed below:

Table 2: Overview of few English Databases.

Name	Year	Language	Speaker	Task	utterances	Duration
AVLetters[27]	1998	English	10	Alphabet	780	13 min
XM2VTS[28]	1999	English	295	Digits	885	59 min
IBMVIAVOICE[29]	2000	English	290	Sentences	10,500	50 h
CUAVE[30]	2004	English	36	Digits	7,000	14 min
GRID[31]	2006	English	34	Phrases	34,000	28 h
Ouluvs[32]	2009	English	150	Sentences	(N/A)	20 h
MOBIO[33]	2012	English	150	Sentences	(N/A)	20 h
AusTalk[34]	2014	English	1000	Digits Words Phrases	24,000 966000 59000	3000 h
LRS[4]	2017	English	29	Sentences	118,166	33 h
AVDigits[35]	2018	English	10	Digits Phrases	795 5850	(N/A)

As in the above table 1 findings of the research paper are summarized. It is observed that CNN, HMM, LSTM, DNN these techniques are used widely are used dominantly these days and also gives good result. HMM has hidden states and it is represented as simple dynamic Bayesian network. It is most take input of variable length. One of the applications of these cascades is to detect objects from images [2].Best features are selected from the error rate analysis. The features with widely used speech recognition technique. HMM models are also easy to train. Pattern recognition techniques, such as the Hidden Markov Model technique, are the most popular of the speech recognition techniques. HMM is a form of finite state machine having state, transitions, observations. HMM can minimum error rate best classify face and non-face regions [2]. Ziad et al. [39] implemented Lipdrive system and after testing nine different classifiers GradientBoosting, Support Vector Machine (SVM) and logistic regression got the best results. The focus of this paper was on the application area of autonomous vehicles. It is a novel system for visual speech recognition.Comparative analysis of nine different classifiers tested on LipDrive is presented. The experiment was done to provide researches with the set of guidelines for classification and preprocessing methods. Muhammad faisal et al.[5] they performed the first experiment on LipNet model and second experiment on set of 10 words first on deep neural network and second on LSTM based network. Results proved that LSTM performs better than DNN. Both the networks were also trained on Urdu digits. Their contribution is small Urdu language corpus consisting of 10 words and 10 phrases each spoken by 10 users each 10 number of times.

Ivan et al. proposed system CNN+BLSTM with the incorporation of maxout activation unit. The accuracy is 87.6% using ouluvs2 10 phrase [11].

It is observed that AVLletters2 dataset is for alphabet recognition. Among that XM2VTS is biggest multi-speaker database available 295 participants for digits. The most popular one was CUAVE though it has fewer participants.

IBMVIAVOICE – 290 speakers was also oldest database and used widely. The mostly used database is ouluvs. The mobile database is MOBIO. To perform lip-reading perfectly frontal shots along with angles slightly departing from frontal-view are always better.

Among that CUAVE is also a multiview dataset.

III. CONCLUSION

We have discussed various deep learning, machine learning techniques and approaches for lip reading. As well as we discussed various types of available datasets. Deep learning can classify, cluster, and predict anything id we have data like images, videos, sound, text etc. It is observed that lip reading systems are currently dominated by CNN features in combination with LSTM. It has provided significant improvement in terms of performance. Different types of datasets are available like character, word, sentence, digits and phrase. The datasets are also available in various languages English, French, German, Japanese etc. Datasets for Indian languages can also be prepared. In this survey we can observe that datasets are only available in few languages we can create a datasets for a regional languages and can thus contribute to the society. In India 70% people live in a rural area so for them regional database should be created and thus taking technology to the remote areas. This gave us the brief idea about the Deep learning approaches and which approach can yield good results.

REFERENCES

[1] Muhammad Rizki Aulia Rahman Maulana, Mohamad Ivan Fanany "Sentence-level Indonesian Lip Reading with Spatiotemporal CNN and Gated RNN" IEEE, 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS) .

[2] Parth Khetarpal, Riaz Moradian , Shayan Sadar , Sunny Doultani, Salma Pathan , "LipVision: A Deep Learning Approach" , International Journal of Computer Applications (0975 – 8887) Volume 179 – No.8, December 2017.

[3] Aparna Brahme, Umesh Bhadade,"Phoneme visem mapping for Marathi language using linguistic approach", International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC) , 2016.

[4] Joon Son Chung, Andrew Senior, Oriol Vinyals, Andrew Zisserman, "Lip Reading Sentences in the Wild" , IEEE conference on computer vision and pattern recognition,2017

[5] Sanaullah Manzoor, Muhammad Faisal, "Deep Learning for Lip Reading using Audio-Visual Information for Urdu Language" , Published in ArXiV 2018

[6] Amit Garg, Jonathan Noyola, Sameep Bagadia, "Lip reading using CNN and LSTM", 2016

[7] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G. Okuno, Tetsuya Ogata, "Lipreading using Convolutional Neural Network", INTERSPEECH 2014, Singapore.

[8] Xinjun Ma, Hongjun Zhang and Yuanyuan Li "Feature Extraction Method for Lip-reading under Variant Lighting Conditions" Harbin Institute of Technology , 2016

[9] Adriana Fernandez-Lopez, Federico Sukno, "Survey on Automatic Lip-Reading in the Era of Deep Learning", Image and Vision Computing, 2018.

[10] Chadawan Ittichaichareon,Siwat Suksri and Thaweesak Yingthawornsuk "Speech Recognition using MFCC" , ICIEA , July 2012

[11] Ivan fung, Brain mak, "End to End low resource lip reading with maxout CNN and LSTM", ICASSP , IEEE, 2018.

[12] Y. Mroueh, E. Marcheret, V. Goel, "Deep multimodal learning for audio-visual speech recognition" , in: Proc. International Conference on Acoustics, Speech and Signal Processing, 2015, pp. 2130–2134.

[13] Y. M. Assael, B. Shillingford, S. Whiteson, N. de Freitas, "Lipnet: Sentence-level lipreading", in: Proc. GPU Technology Conference, 2017.

[14] J. S. Chung, A. Zisserman, "Lip reading in profile", in: Proc. British Machine Vision Conference, 2017.

[15] M. H. Rahmani, F. Almasganj, "Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features", in: Proc. International Conference on Pattern Recognition and Image Analysis, 2017, pp. 195–199.

[16] G. Sterpu, N. Harte, "Towards lipreading sentences using active appearance models", in: Proc. International Conference on Auditory-Visual Speech Processing, 2017.

[17] K. Thangthai, R. Harvey, "Improving computer lip-reading via DNN sequence discriminative training techniques", Proceedings of Interspeech (2017) 3657–3661.

[18] E. K. Patterson, S. Gurbuz, Z. Tufekci, J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research", in: Proc. International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, 2002, pp.2017–2020.

[19] K. Xu, D. Li, N. Cassimatis, X. Wang, "Lcanet: End-toend lipreading with cascaded attention-ctc", in: Proc. International Conference on Automatic Face and Gesture Recognition, 2018, pp. 548–555.

[20] Stavros Petridis,Themos Stafylakis, Pingchuan Ma , Feipeng Cai, Georgios Tzimiropoulos, Maja Pantic "End-to-end audiovisual speech recognition", in: Proc. International Conference on Acoustics, Speech and Signal Processing (in press), 2018.

[21] C. Sui, R. Togneri, M. Bennamoun, "A cascade gray-stereo visual feature extraction method for visual and audio-visual speech recognition", Speech Communication 90 (2017) 26–38.

[22] D. Howell, S. Cox, B. Theobald, "Visual units and confusion modelling for automatic lip-reading", Image and Vision Computing 51 (2016) 1– 12.

[23] M. Gurban, J.-P. Thiran, "Information theoretic feature extraction for audio-visual speech recognition", Signal Processing 57 (12) (2009) 4765–4776.

[24] Y. Takashima, R. Aihara, T. Takiguchi, Y. Ariki, N. Mitani, K. Omori, K. Nakazono, "Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss.", in: Proceedings of Interspeech, 2016, pp.277–281.

[25] M. Zimmermann, M. M. Ghazi, H. K. Ekenel, J.-P. Thiran,"Visual speech recognition using PCA networks and LSTMs in a tandem GMM- HMM system", in: Proc. Asian Conference on Computer Vision, 2016, pp. 264–276.

[26] T. Afouras, J. S. Chung, A. Zisserman, "Deep lip reading: a comparison of models and an online application", in: Proceedings of Interspeech (in press), 2018.

[27] S. J. Cox, R. Harvey, Y. Lan, J. L. Newman, B.-J. Theobald, The challenge of multispeaker lip-reading., in: Proc. International Conference on Auditory-Visual Speech Processing, 2008, pp. 179–184.

[28] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre, "XM2VTSDB: The extended M2VTS database", in: Proc. International Conference on Audio and Video-based Biometric Person Authentication, Vol. 964, 1999, pp. 965–966.

[29] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A.Mashari, "Audio visual speech recognition", Tech.IDIAP, 2000.

[30] E. K. Patterson, S. Gurbuz, Z. Tufekci, J. N. Gowdy, CUAVE: A new audio-visual database for multimodal human-computer interface research, in: Proc. International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, 2002, pp. 2017–2020.

[31] M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition, Journal of the Acoustical Society of America 120 (5) (2006) 2421–2424.

[32] G. Zhao, M. Barnard, M. Pietikainen,"Lipreading with local spatiotemporal descriptors", IEEE Transactions on Multimedia 11 (7) (2009) 1254–1265

[33] Chris McCool , Sebastien Marcel , Abdenour Hadid, Matti Pietikaine , Pavel Matejka, Jan Cernock, Norman Poh, Josef Kittler, Anthony Larcher, Christophe Levy, Driss Matrouf, Jean-Francois Bonastre, Phil Tresadernk , and Timothy Cootesk , "Bi-modal person recognition on a mobile phone: using mobile phone data", in: Proc. International Workshop on Multimedia and Expo, 2012, pp. 635–640.

[34] D. Estival, S. Cassidy, F. Cox, D. Burnham, "AusTalk: an audiovisual corpus of Australian English", in: Proc. International Conference on Language Resources and Evaluation, 2014.

[35] S. Petridis, J. Shen, D. Cetin, M. Pantic, "Visual-only recognition of normal, whispered and silent speech", in: Proc. International Conference on Acoustics, Speech and Signal Processing (in press), 2018.

[36] P. J. Lucey, G. Potamianos, S. Sridharan, "Patch-based analysis of visual speech from multiple views", in: Proc. International Conference on Auditory-Visual Speech Processing, 2008.

[37] N. Harte, E. Gillen, TCD-TIMIT: An audio-visual corpus of continuous speech, IEEE Transactions on Multimedia 17 (5) (2015) 603–615.

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

[39] Ziad Thabet,Amr Nabih, Karim Azmi, Youssef samy,Ghada Khoriba Mai Elshehaly,"Lipreading using a comparative Machine Learning Approach IEEE, 2018