Overview of Lip Reading Methods: Issues, Current Developments, and Future Prospects

Lida K Kuriakose
Assistant Professor, Dept of CSE
AJCE,Kanjirapally
Kottayam, Indiia
lidakkuriakose@amaljyothi.ac.in

***Abstract*— Lip reading, also known as speech reading, is a technique for understanding spoken words by interpreting the speaker's facial expressions and lip movements. It is a helpful tool for those who are deaf or hard of hearing and can also be utilised in noisy settings where it is challenging to understand auditory communication. The development of automatic lip-reading algorithms employing computer vision and machine learning methods has advanced significantly in recent years. We will discuss about some of the most popular lip-reading techniques in this review.**

Keywords—lip reading; front-end; back-end; extraction

I. INTRODUCTION

Communication between individuals frequently involves the use of vocal and visual messages. Visual speech recognition can identify the substance of speech from the movement of the lips of the speaker. In recent years, advances in computer vision and machine learning have enabled the development of automated lip reading systems. These systems use video data of the speaker's face to recognize the spoken words. Automated lip reading systems have the potential to be more accurate and efficient than human lip readers, and can be used in various applications such as speech recognition, speech therapy, and security. Lip reading has potential applications in various fields such as speech recognition, speech therapy, and security. In speech recognition, lip reading can be used as a complementary source of information to improve the accuracy of speech recognition systems in noisy environments or when the audio signal is corrupted. In speech therapy, lip reading can be used to improve the communication skills of individuals with speech disorders and a help people with hearing impairments. In security, lip reading can used in detecting speech n silent cctv footages and can be used in various biometric password systems also. To understand the history of lip reading, we must go back to 1954, the year sumby [1] submitted his first piece that was related to lip reading. A new lip contour reading technology was later produced by petajan [2] and became well-liked in the 1980s. Many studies in the area of lip-reading have already been conducted. An artificial neural network (ann) and pixel-based approach was suggested in a recognition model [3] in 1989. Because audio signals are sensitive to environmental noise. A phrase identification rate of 25% was attained by goldschen [4] and colleagues in 1993 using hidden markov models (hmms) in their lipreading systems. Chiou [5] demonstrated a colour motion-video lip-reading system that included the snake model, hmm, and principal component analysis (pca) to attain an accuracy of roughly 94% for 10 words. Using several layers for visual entities, a context-based deep neural networks (dnn) system was developed to improve continuous lip-reading performance. This system achieved word accuracy of roughly 84.7%, a significant 33% improvement over baseline hmm. The focus of lip reading has shifted over the past several years as deep learning technology has achieved spectacular results in numerous domains. Researchers exploited the deep network's[6] potent representation learning capability to automatically learn good features in accordance with the task objectives rather than attempting to create some feature extraction algorithms to extract features. These traits frequently exhibit high generalization properties and work well in a number of contexts.

However, lip reading is a challenging task due to factors such as lighting, occlusion, and variability in lip movements. In recent years, researchers have developed various lip reading methods that aim to improve the accuracy of speech recognition from lip movements. In this paper, we review the existing techniques used for lip reading and the challenges faced in this field. We also highlight the recent trends and future directions in this area.

II. LIP READING AND RECOGNITION

A. Background

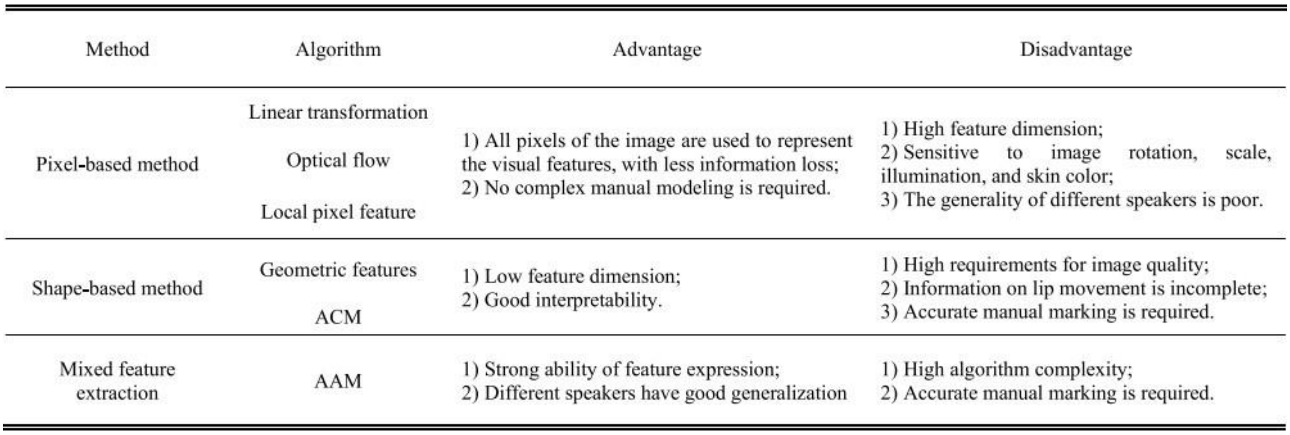
Representation learning for VSR had been studied for a very long time prior to the advent of deep learning. Traditional feature extraction techniques can be divided into three groups from the viewpoint of feature engineering: pixel-based, shape-based, and mixed extraction [7-31]. Traditional representation learning techniques, while straightforward and comprehensible, typically do not perform well, especially in

DOI:10.5281/zenodo.8016645

uncontrolled circumstances. TABLE 1 gives an outline of various feature extraction techniques. The main goal of this study is to summarize and analyses representation learning techniques supported by deep learning technologies.

Many issues arise as the database gets bigger and more complicated, including an increase in the number of speakers, a wide range of postures, and shifting lighting and backdrop environments. The standard manual features are not available everywhere. The Deep Neural Network (DNN) technique, which has strong robustness in the context of massive data, may learn deeper features from the experimental data, according to the researchers.

TABLE I. Traditional feature extraction techniques



Method	Algorithm	Advantage	Disadvantage
Pixel-based method	Linear transformation Optical flow Local pixel feature	1) All pixels of the image are used to represent the visual features, with less information loss; 2) No complex manual modeling is required.	1) High feature dimension; 2) Sensitive to image rotation, scale, illumination, and skin color; 3) The generality of different speakers is poor.
Shape-based method	Geometric features ACM	1) Low feature dimension; 2) Good interpretability.	1) High requirements for image quality; 2) Information on lip movement is incomplete; 3) Accurate manual marking is required.
Mixed feature extraction	AAM	1) Strong ability of feature expression; 2) Different speakers have good generalization	1) High algorithm complexity; 2) Accurate manual marking is required.

Figure 1 illustrates the deep learning-based lip reading flow framework, which differs from the conventional lip reading method. End-to-end methodologies are used in the deep neural network-based lip reading method, which automatically learns the properties of lip movement data from the video to achieve categorization. The first step involves finding and extracting the lip region from the source video. The processed data is then processed by the deep neural network.

The module can be split into the following two components based on its functional characteristics:

1) Front-end: It primarily uses various types of deep neural networks to extract aspects of lip movement from lip photos. The front-end network's feature extraction heavily influences the classification effect. Boltzmann Machines (BM) [36], Auto-encoder[35], Deep Belief Network (DBN) , Feed forward Neural Network (FNN) [33-34], and Convolutional Neural Network[37] are all components of the front-end model.

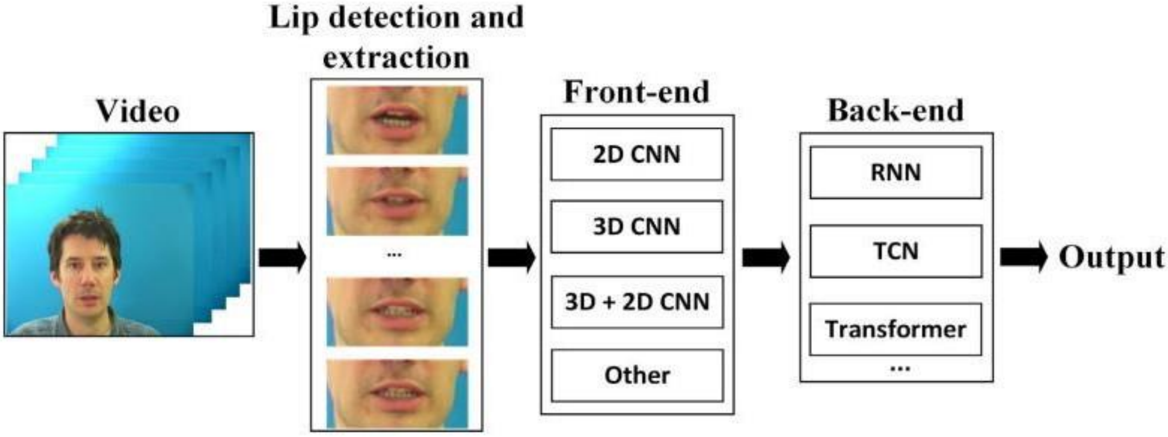


Fig 1: Deep learning-based lip reading flow framework.

Back-end: It basically models the features taken from the front-end network over time. Recurrent Neural Networks (RNN), Temporal Convolutional Networks (TCN), and Transformers are the principal models.

B. Lip Detection and Extraction

Deep learning-based lip reading also requires lip ROI extraction, just like conventional lip reading techniques do. Yet many pre-trained models can now be used because facial detection technology has matured. As seen in Figure 2, the Dlib library [32] is used as an example to identify 68 facial landmarks. Only the area comprising these 20 lip landmarks is used as the input of the front-end network.

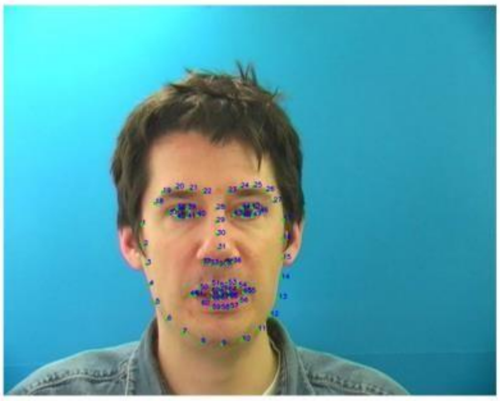


Fig 2: 68 landmarks of the face, among them, 49-68 are lip points.

C. Front-End

The front-end network is used to learn representations of the lip picture and extract features from them. Although many additional deep neural network topologies, including Boltzmann Machine [36], Auto encoder [35], Feed forward Neural Network [33-34] are initially applied to the front-end network. To extract features, Convolutional Neural Networks (CNN) [37] have long been the most popular and efficient network architecture. CNN is a particular type of feed forward neural network that uses convolutional computation and deep structure. It is one of the deep learning methods that best exemplifies the field.

The neurons in a convolution neural network only need to perceive their immediate surroundings before synthesizing the local data at a higher level to obtain global data. In order to extract and combine the collected information, the convolution layer overlay over the structure is helpful for picture recognition. Later, 3D CNN [38] was suggested to handle the information of the time dimension in video. For the purpose of separating visual features from the face landmark points, researchers suggested using Graph Convolution Networks (GCNs) [39]. CNN's 2D, 3D,GCN and 3D + 2D networks make up the bulk of the front-end network. An end-to-end visual transformer-based pooling mechanism created by Afouras et al. [40] learns to track and aggregate the lip movement representations. By reducing the requirement for intricate preprocessing, the suggested visual backbone network can increase the visual representation's robustness. The ablation investigation unequivocally demonstrates that the performance of Visual Speech Recognition is greatly improved by the visual transformer-based pooling technique.

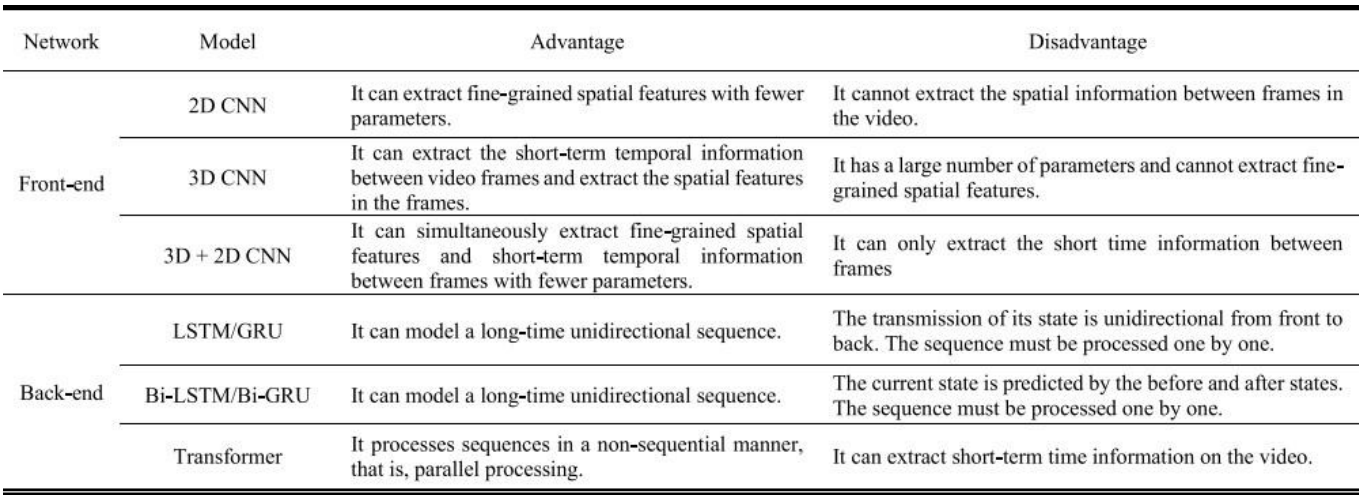
D. Back-End

Since lip movement is a process of movement, the features that were taken from the front-end network in time are primarily modelled in the back-end network. Back end learns the long-term reliance and predict. This section's primary structural component is the recurrent neural network (RNN).The RNN [41] is primarily used to address the time issue. The prior state or the first n time steps mostly determine the current information. The common RNN, however, struggles to learn the long-term reliance problem, which frequently results in the issue that the gradient fades and leaves the RNN with only a short-term memory. Based on RNN, the Long Short-Term Memory (LSTM) [42] is improved. The input gate, forgetting gate, and output gate are the three "gates" that are used to control the state and output at various stages in the LSTM. By combining long-term memory and short-term memory, the "gate" structure helps to solve the gradient disappearance issue. The LSTM's structure is strengthened by the Gate Control Unit (GRU) [43]. The GRU reduces the LSTM's three gates to just two: the update gate and reset gate.

Transformers [44] provide substantial advantages over RNN-based systems in terms of long-term reliance and parallel computing. Transformers typically have some shortcomings, though. First, in small-scale datasets, transformers are more prone to over fitting than RNNs and TCNs [45]. Second, there are some specific tasks (such word-level tasks with short-term context) where transformers are restricted. Transformers are therefore better suited for sentence-level tasks as opposed to word-level tasks.

RNNs and Transformers require a lot of memory and processing power when used with deep sequence models. Another class of deep sequence model called a Temporal Convolutional Network (TCN), with several modifications have been used to make it more suitable for Visual Speech Recognition.

TABLE:2 Comparison of front-end and back-end networks



Network	Model	Advantage	Disadvantage
Front-end	2D CNN	It can extract fine-grained spatial features with fewer parameters.	It cannot extract the spatial information between frames in the video.
	3D CNN	It can extract the short-term temporal information between video frames and extract the spatial features in the frames.	It has a large number of parameters and cannot extract fine-grained spatial features.
	3D + 2D CNN	It can simultaneously extract fine-grained spatial features and short-term temporal information between frames with fewer parameters.	It can only extract the short time information between frames
Back-end	LSTM/GRU	It can model a long-time unidirectional sequence.	The transmission of its state is unidirectional from front to back. The sequence must be processed one by one.
	Bi-LSTM/Bi-GRU	It can model a long-time unidirectional sequence.	The current state is predicted by the before and after states. The sequence must be processed one by one.
	Transformer	It processes sequences in a non-sequential manner, that is, parallel processing.	It can extract short-term time information on the video.

III. PERFORMANCE EVALUATION

The most often used corpus for alphabet recognition is the AVLetters database. Zhao et al. [46] achieved a 62.80% word accuracy rate using LBP-TOP for feature extraction and a Support Vector Machine (SVM) for classification (WAR). The highest WAR, 69.60%, was recorded by Pei et al. [47] using a lip-reading system based on RFMA. Petridis and Pantic [48] employed an LSTM for the backend and a frontend that integrated Deep Belief Network features and DCT features to achieve a classification accuracy of 58.10%.Recurrent Temporal Multimodal Restricted Boltzmann Machines, a system Hu et al. [49] designed based on multimodal RBMs, achieved a WAR of 64.63%.

The most used database for recognizing digits is CUAVE. Using an AAM for feature extraction and an HMM for classification, Papandreou et al. [50] achieved a word recognition rate of 83.00% while performing digit recognition. Using an RBM-Auto encoder, Ngiam et al. [51] achieved a word recognition rate of 68.70%. To reach a WAR of 63.40%, Rahmani and Almasganj [53] retrieved deep bottleneck features and subsequently used a GMM-HMM for the language model. Petridis et al[52] .'s use of the dual flow approach led to a WAR of 78.60%.

One of the first and most popular phrase prediction databases is GRID. For their backend, Wand et al. [54] experimented with Eigenlips, HOG, and feed forward neural networks, three different feature extraction algorithms. Whereas the lip-reading system with the feed forward network in the frontend utilized an LSTM for the backend, the lip-reading systems using Eigenlips and HOG for their respective frontends employed an SVM. Performance findings show that the feed forward network and LSTM combo was the most effective model. Spatiotemporal convolutional networks and Bidirectional RNNs were used by Assael et al. [55], Xu et al. [57], and Margam et al. [58] to achieve word accuracies of 95.20, 97.10, and 98.70%, respectively. The most used multi-view database is OuluVS2. A frontend that incorporated DCT and PCA features, an HMM, and other techniques were employed by Lee et al. [59] to achieve a 63.00% word accuracy rate for phrase prediction. Also, they developed a lip-reading system that employed an LSTM for classification and a CNN for feature extraction, yielding a word accuracy rate of 83.80%. Wu et al[60] .'s method of combining STLP and SDF characteristics with an SVM for classification led to an accuracy of 87.55%. Based on the three-stream technique, Petridis et al. [61] attained a 96.90% word recognition rate.

One of the most difficult datasets for word categorization, according to Chung and Zisserman [62], is LRW, utilised for validation and training. They employed a spatiotemporal CNN to get a word accuracy rate (WR) of 61.10%, whereas Torfi et al. [63] used a linked 3D CNN to achieve a WAR of 98.50% for their lip-reading system. The systems proposed by Martinez et al. [64] and Ma et el. [65], [66] that all used a 3D CNN and ResNet for the frontend with a TCN for the backend and they consequently attained WARs of 85.30%, 88.36%, and 88.50%, produced the best results for the validation on the LRW set. TCNs have advantages over RNNs, as was covered in Section V, and they are expected to take the place of RNNs for many tasks involving sequence processing. A system designed by Fenghour et al. [67] decoded videos in two stages, with the first stage predicting visemes using a 3D-CNN plus ResNet with a Linear Decoder Transformer, and the second stage predicting words using a converter that calculated perplexity scores using the pre-trained GPT transformer. A 64.0% WAR was obtained by Fenghour et al..

IV. DIFFICULTIES AND CHALLENGES OF LIP READING

Lip reading is difficult primarily because the input is a video, often known as an image sequence, and the majority of the visual content is unchanging. The alteration in lip movement is the primary distinction. Action recognition, however, can be categorized using just one image and is a part of video categorization. While lip reading frequently requires that the elements associated with the speech content be extracted from a single image and the timing relationship between the entire sequence of images be examined in order to infer the content. These are the primary challenges with lip reading:

A. External Influencing Elements

The variety of external factors, such as lighting, skin tone, and facial hair can affect lip reading. Because different speakers have varying skin tones, wrinkles, beards or no, variations in background and external light, lip reading will be hampered. It greatly affects feature extraction.

B. Variations in Speech

Different people speak differently, and their lip movements and facial expressions may vary. Some people may speak rapidly, mumble or have accents, making it difficult to understand what they are saying. It difficult to develop a single, universal lip reading system that works for everyone.

C. Context Dependence

Lip reading is heavily dependent on context, as the same lip movements can be used to produce different sounds depending on the surrounding words and sentence structure. This makes it challenging to develop lip reading systems that are accurate and reliable across different contexts.

D. Limited Training Data

Training lip reading systems requires large amounts of data, but high-quality datasets that accurately capture lip movements can be difficult to obtain. A large-scale database with multiple speakers and diverse postural backgrounds plays a major role in the development of lip reading technology.

V. CONCLUSION

This survey reviews automated lip-reading systems running from 2007 to 2021. One can see a progressions of visual speech recognition systems moving from the use of traditional algorithms for letter and digit classification to the use of deep neural networks for predicting words and sentences. Greater sophistication corpuses such BBC-LRS2, LRS3-TED, LSVSR, and LRW-1000 not only contain greater vocabulary sets with thousands of words spoken by thousands of people, but also talk in a variety of positions, lighting, and resolutions. Systems for lip reading include feature extraction and classification components. Despite the fact that Auto encoders do have the advantage of being able to map visual feature data from higher dimensional space into lower dimensional space without the need for any labelled classification, 2D+3D CNNs are the most frequently used network for frontends due to their capacity to learn spatial and temporal features.

The vast majority of classification networks are RNNs in the form of LSTMs and GRUs. Transformers and TCNs, however, have begun to displace RNNs in recent years because of their improved parallel processing, capacity to recognize long-term dependencies, and speed at which they can be trained. The usage of phonemes and visemes could theoretically allow lip-reading algorithms to be lexicon-free, predicting words spoken by a person that did not appear in the training phase.

There are still other issues preventing automatic lip-reading from progressing. They include the necessity to predict unseen words—spoken words that did not exist during training and are not included in the lexicon—as well as visual ambiguities, in which the semantic and syntactic characteristics of words can be learnt for words that sound the same when spoken. There are still difficulties from a visual standpoint, such as speaker dependency, especially when trying to generalize to speakers who did not appear in the training data; the need to generalize to videos with different spatial resolutions; and the need to generalize to videos with different frame rates while consisting of different quantities of temporal data.

References

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

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

[3] 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

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

[5] 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

[6] K. Thangthai, R. Harvey, S. Cox et al., ―Improving Lipreading performance for robust audiovisual speech recognition using DNNs‖, in Faavsp-the Joint Conf. Facial Analysis, 2015.

[7] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, ''Audio-visual automatic speech recognition: An overview,'' in Issues in Visual and AudioVisual Speech Processing. Cambridge, MA, USA: MIT Press,2004.

[8] C. Lee, E. Lee, S. Jung, and S. Lee, ''Design and implementation of a real-time lipreading system using PCA and HMM,'' J. Korea Multimedia Soc., vol. 7, no. 11, pp. 1597–1609, 2004.

[9] J. Yao and Z. Kaifeng, ''Evaluation model of the artist based on fuzzy membership to improve the principal component analysis of robust kernel,'' in Proc. Int. Conf. Big Data Secur. Cloud, Apr. 2016, pp. 322–326.

[10] G. Sterpu and N. Harte, ''Towards lipreading sentences using active appearance models,'' in Proc. Int. Conf. Auditory-Vis. Speech Process, 2017, pp. 70–75.

[11] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, ‘‘A comparison of model and transform-based visual features for audio-visual LVCSR,’’ in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Aug. 2001, pp. 825–828.

[12] S. S. Morade and S. Patnaik, ‘‘Lip reading using DWT and LSDA,’’ in Proc. IEEE Int. Advance Comput. Conf. (IACC), Feb. 2014, pp. 1013–1018.

[13] J. He, H. Zhang, and J. Z. Liu, ‘‘LDA based feature extraction method in DCT domain in lipreading,’’ Comput. Eng. Appl., vol. 45, no. 32, pp. 150–155, 2009.

[14] Y. Liang, W. Yao, and M. Du, ‘‘Feature extraction based on LSDA for lipreading,’’ in Proc. Int. Conf. Multimedia Technol., Oct. 2010, pp. 1–4.

[15] I. Almajai, S. Cox, R. Harvey, and Y. Lan, ‘‘Improved speaker independent lip reading using speaker adaptive training and deep neural networks,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2016, pp. 2722–2726.

[16] G. Potamianos, J. Luettin, and C. Neti, ‘‘Hierarchical discriminant features for audio-visual LVCSR,’’ in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., vol. 1, May 2001, pp. 165–168.

[17] A. A. Shaikh, D. K. Kumar, W. C. Yau, M. Z. C. Azemin, and J. Gubbi, ‘‘Lip reading using optical flow and support vector machines,’’ in Proc. 3rd Int. Congr. Image Signal Process., vol. 1, Oct. 2010, pp. 327–330.

[18] L. Cappelletta and N. Harte, ‘‘Viseme definitions comparison for visualonly speech recognition,’’ in Proc. 19th Eur. Signal Process. Conf., 2011, pp. 2109–2113.

[19] Z. Zhou, G. Zhao, and M. Pietikainen, ‘‘Towards a practical lipreading system,’’ in Proc. CVPR, Jun. 2011, pp. 137–144.

[20] G. Zhao and M. Pietikainen, ‘‘Dynamic texture recognition using local binary patterns with an application to facial expressions,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 6, pp. 915–928, Jun. 2007.

[21] A. Rekik, A. Ben-Hamadou, and W. Mahdi, ‘‘A new visual speech recognition approach for RGB-D cameras,’’ in Image Analysis and Recognition. Cham, Switzerland: Springer, 2014, pp. 21–28.

[22] X. Ma, L. Yan, and Q. Zhong, ‘‘Lip feature extraction based on improved jumping-snake model,’’ in Proc. 35th Chin. Control Conf. (CCC), Jul. 2016, pp. 6928–6933.

[23] J. Luettin and N. A. Thacker, ‘‘Speechreading using probabilistic models,’’ Comput. Vis. Image Understand., vol. 65, no. 2, pp. 163–178, Feb. 1997.

[24] T. F. Cootes, G. J. Edwards, and C. J. Taylor, ‘‘Active appearance models,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 6, pp. 681–685, Jun. 2001.

[25] T. Watanabe, K. Katsurada, and Y. Kanazawa, ‘‘Lip reading from multi view facial images using 3D-AAM,’’ in Proc. Asian Conf. Comput. Vis., 2016, pp. 303–316.

[26] H. L. Bear, S. J. Cox, and R. W. Harvey, ‘‘Speaker-independent machine lip-reading with speaker-dependent viseme classifiers,’’ 2017, arXiv:1710.01122. [Online]. Available: http://arxiv.org/abs/1710.01122

[27] H. L. Bear, R. W. Harvey, and Y. Lan, ‘‘Finding phonemes: Improving machine lip-reading,’’ 2017, arXiv:1710.01142. [Online]. Available: http://arxiv.org/abs/1710.01142

[28] A. Biswas, P. K. Sahu, and M. Chandra, ‘‘Multiple camera in car audio– visual speech recognition using phonetic and visemic information,’’ Comput. Electr. Eng., vol. 47, pp. 35–50, Oct. 2015.

[29] H. L. Bear and R. Harvey, ‘‘Decoding visemes: Improving machine lip-reading,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2016, pp. 2009–2013.

[30] D. Howell, S. Cox, and B. Theobald, ‘‘Visual units and confusion modelling for automatic lip-reading,’’ Image Vis. Comput., vol. 51, pp. 1–12, Jul. 2016

[31] H. L. Bear and R. Harvey, ‘‘Phoneme-to-viseme mappings: The good, the bad, and the ugly,’’ Speech Commun., vol. 95, pp. 40–67, Dec. 2017

[32] D. E. King, ‘‘Dlib-ml: A machine learning toolkit,’’ J. Mach. Learn. Res., vol. 10, pp. 1755–1758, Jan. 2009

[33] M. Wand and J. Schmidhuber, ‘‘Improving speaker-independent lipreading with domain-adversarial training,’’ 2017, arXiv:1708.01565. [Online]. Available: http://arxiv.org/abs/1708.01565

[34] M. Wand, J. Schmidhuber, and N. T. Vu, ‘‘Investigations on end-to-end audiovisual fusion,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp. 3041–3045.

[35] S. Petridis, Z. Li, and M. Pantic, ‘‘End-to-end visual speech recognition with LSTMS,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 2592–2596.

[36] C. Sui, M. Bennamoun, and R. Togneri, ‘‘Listening with your eyes: Towards a practical visual speech recognition system using deep Boltzmann machines,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 154–162

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘‘ImageNet classification with deep convolutional neural networks,’’ in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2012, pp. 1097–1105.

[38] S. Ji, W. Xu, M. Yang, and K. Yu, ‘‘3D convolutional neural networks for human action recognition,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221–231, Jan. 2013

[39] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in CVPR, 2014, pp. 1867–1874

[40] T. Afouras, A. Zisserman et al., “Sub-word level lip reading with visual attention,” arXiv:2110.07603, 2021.

[41] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, ‘‘Learning phrase representations using RNN encoder-decoder for statistical machine translation,’’ in Proc. EMNLP, 2014, pp. 1–14

[42] J. S. Chung and A. Zisserman, ‘‘Out of time: Automated lip sync in the wild,’’ in Proc. Asian Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 251–263.

[43] D. Lee, J. Lee, and K.-E. Kim, ‘‘Multi-view automatic lip-reading using neural network,’’ in Proc. Asian Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 290–302

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ‘‘Attention is all you need,’’ in Proc. NIPS, 2017, pp. 5998–6008

[45] B. Martinez, P. Ma, S. Petridis, and M. Pantic, ‘‘Lipreading using temporal convolutional networks,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2020, pp. 6319–6323.

[46] G. Zhao, M. Barnard, and M. Pietikainen, ‘‘Lipreading with local spatiotemporal descriptors,’’ IEEE Trans. Multimedia, vol. 11, no. 7, pp. 1254–1265, Nov. 2009.

[47] Y. Pei, T.-K. Kim, and H. Zha, ‘‘Unsupervised random forest manifold alignment for lipreading,’’ in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 129–136.

[48] S. Petridis and M. Pantic, ‘‘Deep complementary bottleneck features for visual speech recognition,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2016, pp. 2304–2308

[49] D. Hu, X. Li, and X. Lu, ‘‘Temporal multimodal learning in audiovisual speech recognition,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 3574–3582

[50] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, ‘‘Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition,’’ IEEE Trans. Audio, Speech, Language Process., vol. 17, no. 3, pp. 423–435, Mar. 2009

[51] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, ‘‘Multimodal deep learning,’’ in Proc. 28th Int. Conf. Mach. Learn., (ICML), 2011, pp. 1–8.

[52] S. Petridis, Z. Li, and M. Pantic, ‘‘End-to-end visual speech recognition with LSTMS,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 2592–2596

[53] M. H. Rahmani and F. Almasganj, ‘‘Lip-reading via a DNN-HMMhybrid system using combination of the image-based and model-based features,’’ in Proc. 3rd Int. Conf. Pattern Recognit. Image Anal. (IPRIA), Apr. 2017, pp. 195–199.

[54] M. Wand, J. Koutník, and J. Schmidhuber, ‘‘Lipreading with long shortterm memory,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2016, pp. 6115–6119.

[55] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, ‘‘LipNet: End-to-end sentence level lipreading,’’ in Proc. ICLR Conf., 2016, pp. 1–13.

[56] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson, ‘‘3D convolutional neural networks for cross audio-visual matching recognition,’’ IEEE Access, vol. 5, pp. 22081–22091, 2017.

[57] K. Xu, D. Li, N. Cassimatis, and X. Wang, ‘‘LCANet: End-to-end lipreading with cascaded attention-CTC,’’ in Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), May 2018, pp. 548–555.

[58] D. Kumar Margam, R. Aralikatti, T. Sharma, A. Thanda, P. A K, S. Roy, and S. M Venkatesan, ‘‘LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models,’’ 2019, arXiv:1906.12170. [Online]. Available: http://arxiv.org/abs/1906.12170

[59] D. Lee, J. Lee, and K.-E. Kim, ‘‘Multi-view automatic lip-reading using neural network,’’ in Proc. Asian Conf. Comput. Vis. Cham, Switzerland:Springer, 2016, pp. 290–302.

[60] P. Wu, H. Liu, X. Li, T. Fan, and X. Zhang, ‘‘A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion,’’ IEEETrans. Multimedia, vol. 18, no. 3, pp. 326–338, Mar. 2016

[61] T. Mohammed, R. Campbell, M. Macsweeney, F. Barry, and M. Coleman, ‘‘Speechreading and its association with reading among deaf, hearing and dyslexic individuals,’’ Clin. Linguistics Phonetics, vol. 20, nos. 7–8, pp. 621–630, Jan. 2006

[62] J. S. Chung and A. Zisserman, ‘‘Lip reading in the wild,’’ in Proc. AsianConf. Comput. Vis., 2015, pp. 87–103.

[63] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson, ‘‘3D convolutional neural networks for cross audio-visual matching recognition,’’ IEEE Access, vol. 5, pp. 22081–22091, 2017.

[64] B. Martinez, P. Ma, S. Petridis, and M. Pantic, ‘‘Lipreading using temporal convolutional networks,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2020, pp. 6319–6323.

[65] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic, ‘‘Lip-reading with densely connected temporal convolutional networks,’’ in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2021, pp. 2857–2866

[66] P. Ma, B. Martinez, S. Petridis, and M. Pantic, ‘‘Towards practical lipreading with distilled and efficient models,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Jun. 2021, pp. 7608–7612.

[67] S. Fenghour, D. Chen, K. Guo, and P. Xiao, ‘‘Lip reading sentences using deep learning with only visual cues,’’ IEEE Access, vol. 8, pp. 215516–215530, 2020.

[68] Deep Learning-Based Automated Lip-Reading: A Survey SOUHEIL FENGHOUR 1 , DAQING CHEN 1 , KUN GUO 2 , BO LI 3 , AND PERRY XIAO 1 1School of Engineering, London South Bank University, London SE1 0AA, U.K. 2Xi’an VANXUM Electronics Technology Company Ltd., Xi’an 710065, China 3School of Electronics and Information, Northwestern Polytechnical University, Xi’an 710129,
China

[69] A Survey of Research on Lipreading Technology MINGFENG HAO 1 , MUTALLIP MAMUT2 , NURBIYA YADIKAR1 , ALIMJAN AYSA3,4 , AND KURBAN UBUL 1,4, (Member, IEEE)