Recent Advances in Audio-Visual Speech Recognition: Deep Learning Perspective

Diksha R. Pawar(✉) and Pravin Yannawar

Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India
dikshasalunke97@gmail.com

Abstract. Speech is the powerful engine of communication among human beings and language is meant for communicating with the world. This has motivated new researchers to study automatic speech recognition and expand a computer system so it can integrate and understand human speech. But the problem with speech recognition is the acoustic noisy environment can deeply corrupt audio speech. This polluted audio speech disturbs the whole recognition performance. So, the development of Audio-Visual Speech Recognition (AVSR) aims to solve the issues by utilizing visual pictures that are undisturbed by noise. This review paper's goal is to explain AVSR architectures, which include front-end operations, the utilized audio-visual dataset, and related studies, audio feature extraction, fusion and modeling techniques, and accuracy estimation methods.

Keywords: ASR · audio feature extraction · AVSR · audio-video fusion · HMM · accuracy estimation methods · GNN · etc.

1 Introduction

The computer is now a part of human life and has contributed significantly to creating this world digitally. Language is meant for communicating world. The largest part of human linguistic communication so far occurs as speech. Language is the most popular medium of communication and many languages are used in the world for oral and written communication. Different languages use different approaches to encoding information. Sound signals and visual lip activities are produced by the speaker's vocalization bodies, like the mouth cavity and vocal tract systems. The process of translating a human speech signal into a series of word algorithms for a human-machine interface is known as automatic speech recognition. Since 1920, researchers have been working on how computers can be made to understand the meanings of human language.

Around 1920, the first speech recognition prototype was created using a toy dog that was attached to a magnetic spring. When someone shouted the word Rex, the dog would jump, but it was energy- and frequency-sensitive, operating at around 500 Hz. Speech recognition was categorized in the study [1] as an information extraction method that was developed for the first time at Bell Labs in the 1950s [2]. Noise generally increases the ASR system's primary issue.

R. Manza et al. (Eds.): ACVAIT 2022, AISR 176, pp. 409–421, 2023.
https://doi.org/10.2991/978-94-6463-196-8_31

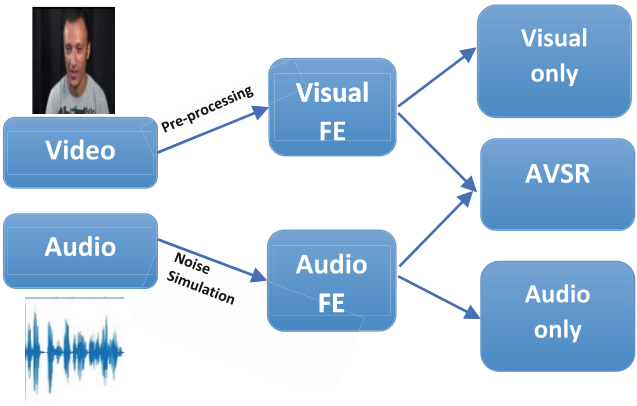

Fig. 1. Block diagram of AVSR

The primary influencing factor in studies of recognition systems is always noise [3]. The ASR, advanced methods take of visual modalities such as a combination of the speaker's lip movements and audio modality, leads to an audio-visual speech recognition (AVSR) system. In order to overcome the limitation of ASR, AVSR uses visual information from the speaker to improve speech recognition when an audio signal is corrupted by noise as shown in the Fig. 1.

To perform speech recognition tasks, and audiovisual speech recognition system (AVSR) integrates auditory and visual data. Lip syncing can be accomplished live on stage, on TV shows, on computer systems, in movies, or through other audio-visual output devices. The term can refer to any of a number of different methods and processes, in the context of live performances and audiovisual recordings. Recent years have seen an increase in the popularity of audio-visual speech recognition, attracting researchers from the fields of pattern recognition, computer vision, and signal and speech processing. However Lip-sync mistake, on the other hand, depends on the ratio timing of audio and visual components throughout production, post-production, distribution, and playback manufacturing as when a challenge or issue arises. Nowadays, Deep learning makes it feasible to convert lip movements into meaningful words. Visual information can improve speech recognition in noisy contexts also. Initially, HMM (Hidden Markov Model) was used by the researcher to represent the movement of the audio and visual patterns for face expression but now DNN (deep neural network), RNN (recurrent neural network), CNN (convolutional neural network), and the newest GAN technology make everything so easy. People were always active on online streaming websites like YouTube after 2015. That had collected hundreds of millions of daily views, Twitch had over 1.5 million broadcasters on it, and YouTube had 2.3 billion subscribers by 2020, so it appeared that the future of video technology was very bright [4].

2 Related Work

The movements of voice and video sequences were represented using simply Hidden Markov Models (HMMs) in a few of the first methods for facial expressions or for movements. Vector quantization was employed by Simons et al. and Cox et al. [5] to obtain a bond description of the audio-visual features that served as the outputs for respective HMM. Even though researchers preferred HMMs as compared to the neural network because they explicitly divide speech down into understandable states. In latest years deep learning has given proper results in neural networks being used in lots of modern approaches. Using the subject-independent method, a deep neural network (DNN) in [6] converts a graphical representation of a pattern into a sequential manner for the lower half of the facial shape.

Using deep neural networks end-to-end approach and other studies were able to solve audio-visual speech recognition "in the wild," [7, 8] which refers to unrestricted open-world speech. Sutskever et al. [9] were the first to use neural networks to solve a sequence-to-sequence challenge. After Bahdanau et al. [10] and Luong et al. [11]. Improvements were brought about with attention mechanisms by newly, Vaswani et al. [12] developed a transformer network and is based on an attention method to identify global connections among outputs and inputs. According to Johnson et al., learning multiple translating techniques raises performance overall, especially for low-resource languages [13]. Recurrent neural networks (RNNs) have relied on deep networks, which were publicized in [14, 15]. These designs produce natural output but are subject-dependent and need reconstruction and retraining procedures to adapt to new faces.

In [16] use convolutional neural networks (CNN) to convert audio data into a 3-dimensional mesh of particular speakers. The CNN approach has comment threads that are in charge of articulating dynamics and mesh point estimation in three dimensions. A CNN based on Mel-frequency Cepstral coefficients (MFCCs) developed by Chung et al. [17] to create subject-independent clips from a simple image and audio signal. This approach includes an L1 loss on the image, which blurs the image and necessitates a deblurring step too. Along with these, pixel loss discourages deviation from the training clip, which does not encourage the system to create natural emotions and results in essentially static faces with the exception of the lips.

The latest work on GANs in [18] turned machine learning techniques' attention to generative modeling. GANs used adversarial loss that has the ability to produce better images as compared to L1 and L2 losses [19] and straightforward adaptations for audio- visual datasets can be easily modified by swapping out the 2D convolutional networks for 3D convolutional networks. The generator and discriminator networks may represent periodic dependence by using three dimensions convolutional layers however, they require films of a specific length. This drawback is resolved in [20], but constrain be imposed in the latent space to create systematic and proper output. An RNN-based generator with different latent spaces for movement and information was introduced by the MoCoGAN system [21]. At last, Chen et al. in [22] give a Generative adversarial network-based encoder-decoder framework that utilizes CNNs to transform speech signals into frames and frames into spectrograms. In the year 2019- 2020 IIT Hyderabad students created Lip2Wav [60], Wav2Lip [61], LipGAN [62] model using GAN.

3 Data Corpus

Audio-Visual datasets are mostly used in industry such as the Alexa voice service for automatic speech recognition. The voice is routed over a speech-recognition machine for learning lip reading in military services and health care. There are currently many AVSR data corpora available, but some of them have defects in their word analysis, recording quality, illumination, and environmental variations. Although The Tulips1 [23], AVletters [24], M2VTS [25], CUAVE [26], (LUNA-V) [27], TIMIT [28], GRID [29] [30], vVISWa [33, 34], etc. databases are a popular databases used for voice recognition. It permits scientists to use the datasets as a reference, enabling observation and helping judgment of the results of independent tests and AVSR procedures. M2VTS [25] and GRID [29] [30] used MHMM and CHMM classification methods and gives an accuracy result near 97%. The CUAVE [26] speech database w/ith a resolution of 750 × 576 pixels was developed by Patterson et al. in 2002. Tulips1 [23] & AVletters [24] were created in 1995 and 1998 with resolutions of 100 × 75 pixels and 80 × 60 pixels correspondingly. AVletters [24] takes 10 speakers (5 Male, 5 Female) and create A to Z word datasets. Later, the newer speech database Loughborough University Audio-Visual data corpus (LUNA-V) used a geometry approach for lip-syncing audio-visual speech recognition. It has been shown through a Comparing analysis of the LUNA-V [27] and CUAVE [28] datasets that the organized and demonstrated images with high accuracy and significantly advance the task of visual-speech recognition. Vassil Panayoto et al. [31] and Anthony Rousseau et al. [32] provide large open-source speech recognition datasets.

Prashant Borde et.al has created vVISWa [33] data corpus as shown in Fig. 2 and he has explored the role of visual features from the vVISWa data corpus that are generated by Zernike events in combination with MFCC for the recognition of isolated city names [35]. An extensive collection of English, Marathi, and Hindi isolated words are read aloud in this corpus. There were 58 speakers in all that contributed to the corpus, of which 48 were native speakers as shown in Fig. 2(a) and 10 speakers were non-native shown in Fig. 2(b) that is, they are from Iraq and Yemen [33].

4 Visual Front End

Just a while ago, a lot of visual front-end designs have been mentioned in the article [42] such as appearance-based features, geometrical-based features, and a combination of appearance-geometrical-based features [36]. Most of the researchers used appearance-based features, but this feature's imperfection is that it is susceptible to environmental changes like content brightness, shine, and head attitude. However, appearance-based features build a feature vector with reduced dimensionality that contains associated speech information by taking all pixels inside the ROI (region of interest) that are instructive to speech vocalization and linearly transforming their pixel values [37–39].

Geometrical features, such as the size, length, & region of lips movement of the speakers, are used to control audio datasets [40]. More recently, Ibrahim, M. Z et al. [41] show the test-strength of the geometrical and appearance-based aspects using the head posture and brightness increment. The study found that features based on geometry

Fig (a) Different side pose of Native speakers

Fig (b) Different side pose of Non-native speakers

Fig. 2. vVISWa data corpus

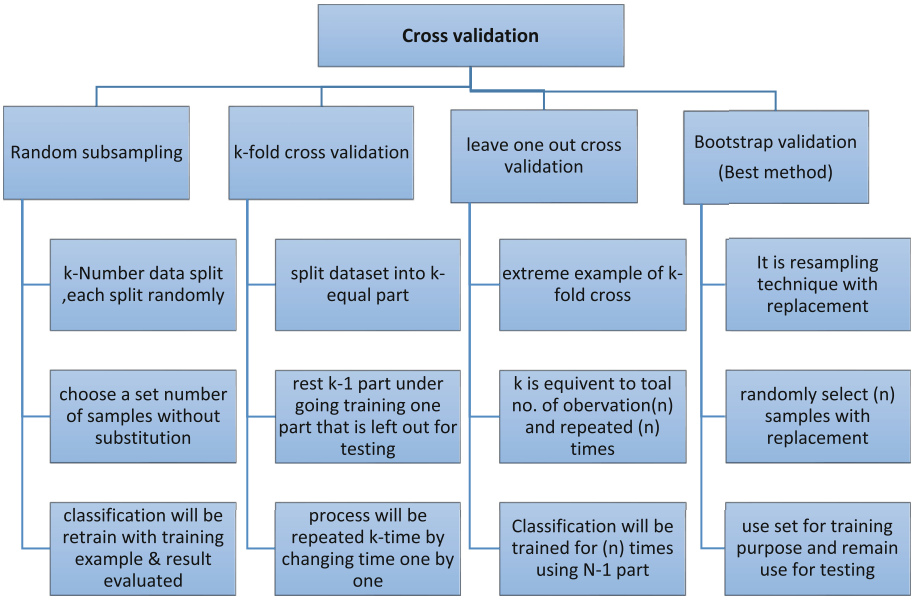

Fig. 3. Types and methods of Cross validation

are more resistant to environmental variations compared with appearance-based features and geometrical-based features may overcome the defect of appearance-based features [41].

5 Accuracy Estimation Methods

The holdout method and the cross-validation are two general methods of evaluating the classifier's accuracy. The holdout approach involves randomly dividing each sample into two training datasets and testing datasets independently. This is a train-and-test experiment, so the holdout estimation may be misleading if the training set samples take corrupted data [42]. By using cross-validation, the holdout method's flaw can be resolved. The description and difference of each method are shown in the Fig. 3

6 Audio Feature Extraction

In the literature review, there are so many features extraction techniques used like linear Predictive Coefficient (LPC) [43], Principal Component Analysis (PCA) [44, 45], Linear Discriminate Analysis (LDA) [46], Independent Component Analysis (ICA) [47] and Mel-Frequency Cepstrum Coefficients (MFCC) [43] [44]. Tripathy.s et al. used linear predictive coding (LPC) and Mel-frequency cepstral coefficient (MFCC) for Hindi speech recognition [48]. In it, datasets were classified into train databases and test databases. Speaker-dependent and speaker- independent systems were each given a portion of the tested data corpus [42]. HMM is demonstrated to perform better than LPC as a classifier for MFCC feature-extraction in the speaker-dependent environment. Therefore, this research work comes to the conclusion that while MFCC outperforms LCP in most situations, it performs worse than LPC feature extraction in speaker-independent environments [42]. In it, datasets were classified into train databases and test databases [48]. The tested data corpus was separated into two different systems: speaker-dependent and speaker-independent systems (Table 1).

7 AVSR Fusion and Modeling Techniques

In order to outperform both audio-only and visual-only recognition, AVSR aims to combine audio-visual modes data stream into a multi-modal classification. For fusion between audio and visual modalities, there are three main approaches: feature fusion, modal fusion, and decision fusion [52, 53]. The best method of integration of audio and video is model fusion, it is higher-level integration than feature fusion. Model fusion integrates both modalities and then classifies them separately. It is a middle integration method that can be demonstrated by multi-stream HMMs that utilize two or more independent streams of audio and visual performance. Decision fusion can't do interaction between two modalities during the classification process, it generally takes place after the spoken utterance is completed, becomes that results come in the delay to generate the classification result and leads to unnatural interaction sessions this is the main drawback of this approach (Table 2).

Table 1. Some popular methods of feature extraction [59]

Sr. No	Methods	Property	Procedure for implementation
1.	Linear Discriminate Analysis	Supervised linear map, rapid, Eigen-vector-based nonlinear feature extraction technique	LDA is more effective for classification than PCA [46]
2.	Independent component analysis (ICA)	Iterative non-Gaussian, non-linear feature extraction, linear map	Blind channel separation is employed to separate sources with non-Gaussian distribution [47]
3.	Principal Component Analysis (PCA)	Quick, Eigen-vector based, unsupervised linear map, nonlinear feature extraction technique	Eigenvector base method, or Karhuneu-Loeveare expansion, is a conventional technique that works well with Gaussian data [44, 45].
4.	Mel-frequency cepstral coefficients (MFCC)	By doing Fourier analysis and Spectral analysis, the power spectral is calculated.	Our characteristics are discovered via spectral analysis with a fixed resolution subjective frequency scale [43, 44]
5.	Linear Predictive coefficient	Method for extracting static features with 10 to 16 lower-order coefficients	Lower order feature extraction is done using LPC [43].
6.	Wavelet	Superior to the Fourier transform in time resolution	It improves time resolution at high frequencies compared to Fourier Transform by swapping out its fixed bandwidth for one that is proportional to frequency [49].
7.	Cepstral mean subtraction	Dependable feature extraction	The Mean Statically Parameter is used instead of the MFCC in this case [50].
8.	Integrated phoneme subspace method	PCA + LDA + ICA based on transformation	Greater Accuracy compared to the current approach [51]

Table 2. Different modeling technique

Sr. No	Approaches	Year	Ref. no	Technique
1.	Acoustic-phonetic approach	1996	[18, 22, 23]	Gaussian Mixture Modeling, SVM Classifier Classification, and Problem Phone Recognition
2	Pattern Recognition approach	1993, 1975	[24, 25]	HMM, Pattern training, the Pattern comparison
3	Template-based approach	1979	[26, 27]	Unknown speech is contrasted with a collection of recorded words.
4	Knowledge-based approach	1993	[28, 29]	Vector Quantization(VQ), the lowest distance measure using the VQ codebook
5	Statistical based approach	1998, 2004	[8, 9, 31, 40]	HMM, Statistical learning, learning-VQ, k-mean algorithm
6	Learning-based approach	2006	[14, 16, 18], 0]	Neural network, genetic algorithm, machine learning
7	Artificial Intelligent approach	1987	[26, 41]	Hybrid of acoustic-phonetic & pattern recognition
8	Stochastic Approach	1990	[43]	HMM-based chain model, temporal variability, output distribution, spectral variability

Table 3. Evaluation of Popular Works on the AVSR Speech Corpus

Sr. No.	Ref No.	Year	Dataset	Speakers	Techniques		Task	Accuracy
					Classification	*Feature extraction*		
1.	[25]	2005	M2VTS	25 males, 12 females	MHMM	LDA-PCA	Speaker Recognition	96.57%
2.	[54]	2014	XM2VTS	295 (unknown gender)	MSHMM	MFCC-DCT	Digit Recognition ara>	*89%*
3.	[55]	2014	CUAVE	19 males, 17 females	HMM	MFCC	Digit Recognition	95%
4.	[28]	2010	VidTIMIT	24 males, 19 females	DCT-MFCC	GMM	Person Recognition	EER = 5.23
5.	[56]	2010	Tulips1	7 males, 5 females	LDB	HMM-SVM	Speech Recognition	EER = 1.74
6.	[29]	2013	GRID	34 (unknown gender)	MFCC	CHMM	Speech Recognition	96.37%
7.	[57]	2014	LUNA-V	9 males, 1 female	HSV	HMM	Digit recognition	92.5% (Visual-only)
8.	[34]	2017	vVISWa	48 speakers native, 10 speaker non-native	K-Means, Random Forest & HMM	MFCC	Speech Recognition	98% (Visual-only)

8 Discussions

In this publication, we see a review of AVSR modeling methods, audio feature extraction, visual front end, and accuracy estimation methods. There are various challenges that are pertinent to the AVSR platforms are training and test datasets. Some typical issues with audio-visual data collection in video. In visual data, the existing spoken database is usually of very poor quality, but now there are high-resolution cameras are used for capturing datasets. Along with that, there are certain difficulties with integrating auditory and visual modalities for speech recognition. Syncing an audio and video and handling asynchrony modality are a big problem in real-world applications and need more work in the future. The research [58] used different validation techniques and conclude that the bootstrap validation method gives the best result compared with other validation methods, and this method is still widely used in real-world applications since it requires less computing work than other techniques like k-fold evaluation and other methods. So we can say, the out-of-sample bootstrap approach is accurate to the others in terms of accuracy and transfer error, but in actual practice, its development is laborious and computerized. In the visual front end section, we can see that researchers used appearance-based features and Geometrical-based features. Appearance-based features are imperfection in that it is sensitive to environmental variation. Compared with appearance-based features and geometrical-based features may overcome the defect of appearance-based features.

9 Conclusion

We conclude from this review paper that the future of video technology is very bright [4]. The AVSR was developed for solving the problem of ASR and the researcher firstly used HMM. After 2006 the learning approach gives the golden changes in AVSR. Recently deep learning has given proper results in neural networks being used in lots of modern approaches. It is just the starting of audio-visual speech recognition and then RNN, CNN-based models are developed by the researcher. This approach uses an L1 loss at the pixel level but the disadvantage of this is creating unnatural expressions and producing output is mostly static faces, with the mouth being the only moving part. Since adversarial loss produces finer, more detailed pictures than L1 and L2 losses, much of the current work on GANs in [18] is related to image generation. There are two networks in it: a generator that creates faces based on speech & a discriminator that determines the produced lip movement and speech are in time or it give proper results with correct syncing. As we discuss in literature review IIT Hyderabad students created Lip2Wav [60], Wav2Lip [61], LipGAN [62] model using GAN. In it, they look at the issue of lip-syncing a random speaking facial video to a specific speech piece and used various type of methods such as in-sync, out-sync and ground truth image. We see different accuracy estimation methods but bootstrap is one of the best methods in cross-validation, it tested each validation technique by looking into bias and variance [42]. According to Table 3 mostly used and accurate methods of classification and feature extraction are HMM and MFCC. Thus, we invite researchers to undertake work on this line for making robust solutions for Lip movement synchronization in a multi-pose AVSR environment.

Acknowledgment. The authors gratefully acknowledge support from the Shree Chhatrapati Shahu Maharaj Research, Training and Human Development Institute (SARTHI), An Autonomous Institute of Govt. of Maharashtra for providing financial assistance for the Major Research Project. This work was supported by Dr. Babasaheb Ambedkar Marathwada University and the vision and intelligence lab. Authors would like to thank Dr. Babasaheb Ambedkar Marathwada University for the publication support.

References

1. Ghadage, Y. H. & Shelke, S. D. Speech to Text Conversion for Multilingual Languages (2016), 236–240.
2. Morgan, N. Deep and wide: Multiple layers in automatic speech recognition. IEEE Trans. Audio, Speech Lang. Process. 20 (2012), 7– 13.
3. Tian, C., Ji, W. & Yuan, Y. Auxiliary Multimodal LSTM for Audiovisual Speech Recognition and Lipreading(2017), 1–9.
4. How many people used YouTube in 2021,backlinko.com/youtube-users.
5. A. D. Simons and S. J. Cox. Generation of mouth shapes for a synthetic talking head. Proceedings of the Institute of Acoustics, Autumn Meeting, 12(January):475482, 1990
6. S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews. A deep learning approach for generalized speech animation. ACM Transactions on Graphics (TOG), 36(93), 2017.
7. Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2017. Deep voice 3: Scaling text-tospeech with convolutional sequence learning. arXiv preprint arXiv:1710.07654 (2017).
8. Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 4779–4783.
9. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Advances in neural information processing systems. 3104– 3112
10. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
11. Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015).
12. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems. 5998–6008.
13. Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. GoogleâĂŹs multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics 5 (2017), 339–351.
14. B. Fan, L. Wang, F. Soong, and L. Xie. Photo-real talking head with deep bidirectional lstm. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4884–4888, 2015.
15. S. Suwajanakorn, S. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing Obama: Learning Lip Sync from Audio Output Obama Video. ACM Transactions on Graphics (TOG), 36(95), 2017.

16. T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics (TOG), 36(94), 2017.
17. J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? In British Machine Vision Conference (BMVC), pages 1–12, 2017.
18. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. In Advances in neural information processing systems (NIPS), pages 2672–2680, 2014.
19. Y. Li, M. R. Min, D. Shen, D. Carlson, and L. Carin. Video Generation from Text. arXiv preprint arXiv:1710.00421, 2017.
20. M. Saito, E. Matsumoto, and S. Saito. Temporal Generative Adversarial Nets with Singular Value Clipping. In IEEE International Conference on Computer Vision (ICCV), pages 2830–2839, 2017.
21. S. Tulyakov, M. Liu, X. Yang, and J. Kautz. MoCoGAN: Decomposing Motion and Content for Video Generation. arXiv preprint arXiv:1707.04993, 2017.
22. T. Chen and R. R. Rao. Audio-Visual Integration in Multimodal Communication. Proceedings of the IEEE, 86(5):837–852, 1998.
23. Luettin, J., Thacker, N. a. & Beet, S. W. Visual speech recognition using active shape models and hidden Markov models. IEEE Int. Conf. Acoust. Speech, Signal Process. 2 (1996), 817–820.
24. Matthews, I. Features for audio-visual speech recognition. Citeseer (1998).
25. Lucey, S., Chen, T., Sridharan, S. & Chandran, V. Integration strategies for audio-visual speech processing: applied to text-dependent speaker recognition. IEEE Trans. Multimed. 7 (2005), 495–506.
26. Patterson, E. K., Gurbuz, S., Tufekci, Z. & Gowdy, J. N. CUAVE: A new audio-visual database for multimodal human-computer interface research. IEEE Int. Conf. Acoust. Speech, Signal Process.2 (2002), II2017-II-2020
27. M. Z. Ibrahim, "A novel lip geometry approach for audio-visual speech recognition," Loughborough University (2014).
28. Shah, D., Han, K. J. & Narayanan, S. S. Robust Multimodal Person Recognition Using Low-Complexity Audio-Visual Feature Fusion Approaches. Int. J. Semant. Comput. 4 (2010), 155–179.
29. Ahmed Hussen Abdelaziz, Steffen Zeiler, D. K. Twin-HMM-based audio-visual speech enhancement. Digit. Signal Process (2013), 3726– 3730.
30. Receveur, S., Scheler, D. & Fingscheidt, T. A turbo-decoding weighted forward-backward algorithm for multimodal speech recognition (2014), 179–192.
31. Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 5206–5210.
32. Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2012. TED-LIUM: an Automatic Speech Recognition dedicated corpus. In LREC. 125–129.
33. P Borde, R Manza, B Gawali, P Yannawar' vVISWa ' – A Multilingual Multi-Pose Audio Visual Database for Robust Human Computer Interaction.
34. P Borde, P Yannawar. Recognition of Isolated Digit Using Random Forest for Audio Visual Speech Recognition.
35. Borde Prashant, Amarsinh Varpe, Ramesh Manza, and Pravin Yannawar. "Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition." International Journal of Speech Technology (2014): 1-9.
36. Auxiliary Multimodal LSTM for Audio-visual Speech Recognition and Lipreading (2017), 1–9. Tian, C., Ji, W. & Yuan, Y.

37. Galatas, G., Potamianos, G. & Makedon, F. Audio-visual speech recognition incorporating facial depth information captured by the Kinect (2012), 2714–2717.
38. Navarathna, R., Dean, D., Sridharan, S. & Lucey, P. Multiple cameras for audio-visual speech recognition in an automotive environment. Computer Speech and Language 27 (2013), 911–927.
39. Palecek, K. & Chaloupka, J. Audio-visual speech recognition in noisy audio environments. 2013 36th Int. Conf. Telecommun. Signal Process (2013), 484–487.
40. Ibrahim, M. Z. & Mulvaney, D. J. A lip geometry approach for feature fusion-based audio-visual speech recognition. ISCCSP 2014 - 2014 6th Int. Symp. Commun. Control Signal Process. Proc (2014), 644–647
41. Ibrahim, M. Z. & Mulvaney, D. J. Robust geometrical-based lipreading using hidden Markov models. IEEE EuroCon (2013), 2011– 2016.
42. A Review of Audio-Visual Speech Recognition Thum Wei Seong and M. Z. Ibrahim Applied Electronic and Computer Engineering Cluster Faculty of Electrical & Electronic Engineering, University Malaysia Pahang, 26600 Pekan, Pahang, Malaysia (2018).
43. Dave, N. Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition. Int. J. Adv. Res. Eng. Technol. 1 (2013), 1–5.
44. Ittichaichareon, C. Speech recognition using MFCC. Conf. Computer (2012), 135–138.
45. Hongbing Hu, Stephen. A, Z. Dimensionality reduction methods for HMM phonetic recognition (2010), 4854–4857.
46. Mohamed, A. et al. Deep belief networks using discriminative features for phone recognition. Acoust. Speech Signal Process. (ICASSP), IEEE Int. Conf (2011). 5060–5063.
47. Shrawankar, U. & Thakare, V. Feature Extraction for a Speech Recognition System in Noisy Environment: A Study. Comput. Eng. Appl. (ICCEA), Second Int. Conf. 1 (2010), 358–361.
48. Tripathy, S., Baranwal, N. & Nandi, G. C. A MFCC based Hindi speech recognition technique using HTK Toolkit. 2013 IEEE 2nd Int. Conf. Image Inf. Process. IEEE ICIIP (2013), 539–544
49. Feature Detection and Extraction Using Wavelets, Part 1: Feature Detection Video - MATLAB (mathworks.com)2020
50. The effect of reverberation on the performance of cepstral mean subtraction in speaker verification, ScienceDirect Topics(2011)
51. Sannella, M speaker recognition "Project Report" from https:\\cs.pensuu.fi/pages/tkinnu/research/index.html/viewed 23 feb 2010
52. Katsaggelos, A. K., Bahaadini, S. & Molina, R. Audiovisual Fusion: Challenges and New Approaches. Proc. IEEE 103 (2015), 1635–1653.
53. Huang, P. Sen, Zhuang, X. & Hasegawa-Johnson, M. Improving acoustic event detection using generalizable visual features and multi-modality modeling. ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. – Proc (2011). 349–352.
54. Stewart, D., Seymour, R., Pass, A. & Ming, J. Robust audio-visual speech recognition under noisy audio-video conditions. IEEE Trans. Cybern. 44 (2014), 175–184.
55. Pawar, G. S. Realization of Hidden Markov Model for English Digit Recognition. 98 (2014), 98–101.
56. Kambiz Rahbar. Independent-Speaker Isolated Word Speech Recognition Based on Mean-Shift Framing Using Hybrid HMM/SVM Classifier (2010). 156–161.
57. Ibrahim, Z. A novel lip geometry approach for audio-visual speech recognition (2014).
58. Tantithamthavorn, C., Mcintosh, S., Hassan, A. E. & Matsumoto, K. An Empirical Comparison of Model Validation Techniques for Defect Prediction Models. IEEE Trans. Softw. Eng. 5589 (2016), 1–16
59. Santosh K. Gaikwad, Dr. Bharti Gawali, Dr. Pravin Yannawar, Review of Speech Recognition Technique.(2010)

60. Prajwal K R*, Rudrabha Mukhopadhyay, Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis (CVPR, 2020) arxiv.org/abs/2005.08209.Lip2Wav: https://www.youtube.com/watch?v=HziA
61. [ACM Multimedia, 2020] Wav2Lip: Accurately Lip-syncing Videos In The Wild Wav2Lip:https://youtu.be/0fXaDCZNOJc(2020)
62. Towards Automatic Face-to-Face Translation. Authors: Prajwal K R*, Rudrabha Mukhopadhyay,LipGAN:https://www.youtube.com/watch?v=aHG6O... (2019)