2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC) | 979-8-3503-9938-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISACC56298.2023.10084209

Review of Various Machine Learning and Deep Learning Techniques for Audio Visual Automatic Speech Recognition

Arpita Choudhury
Dept. of CSE
NIT Silchar
Silchar, India
arpitachoudhury_rs@cse.nits.ac.in

Pinki Roy
Dept. of CSE
NIT Silchar
Silchar, India
pinki@cse.nits.ac.in

Sivaji Bandyopadhyay
Dept. of CSE
NIT Silchar
Silchar, India
sivaji.cse.ju@gmail.com

***Abstract*—The visual cues obtained from the face and mouth region of a speaker provide valuable information for speech perception. The idea of audio visual speech recognition is to combine visual information with acoustic speech signals to enhance the intelligibility of speech in the presence of ambient noises. In audio visual speech recognition lip image sequences of speakers are used along with acoustic signals to convert speech into text. Researchers are exploring ways to upgrade the performance of audio visual speech recognition and solve certain real life problems like designing voice dialling systems, highly secured biometric systems for authentication etc. A review of the latest research findings on audio visual automatic speech recognition using traditional machine learning, neural networks and other deep learning techniques is presented in this work. This paper describes future research opportunities through a comparative analysis of the various techniques used in the literature for the different stages of audiovisual speech recognition, including the region of interest detection, audio and visual speech feature extraction and fusion of the modalities.**

***Index Terms*—audio visual speech recognition, feature extraction, deep learning**

I. Introduction

The current trend in research is the development of state-of-the-art systems to improve human-machine interaction by replacing the traditional use of mouse and keyboards. Speech, the primary means of communication among humans, can be exploited for hands free computation in the era of artificial intelligence. Speech recognition by machines has grabbed the attention of researchers lately, where the utterance of a speaker is translated into the corresponding text. The visual speech perception of human beings is helpful when the reliability of speech recognition using acoustic waves becomes uncertain due to ambient noise. McGurk's experiment [1] explains how the intelligibility of speech relies on visual cues. However, visual articulations vary among speakers and are less informative compared to audio signals which make tracking the lip movement for lipreading a challenging task. Also, the variation in illumination, position of the head, etc. affects visual speech recognition. The pitfalls of both modalities are solved by integrating visual information with audio, which induces a more reliable speech recognition system. One of the earliest research in this domain by Petajan [2] showed that Audio visual speech recognition (AVSR) by applying dynamic time warping and using features extracted from the speakers' mouth region performs better than stand alone audio based or vision based speech recognition. Hidden Markov Model (HMM) [3] was applied by Goldschen [4] in 1993 for AVSR with a significant gain in recognition accuracy. Speech recognition tasks are either speaker independent or speaker dependent. Speaker independent recognition is more challenging since the users of the speaker independent AVSR system do not belong to the training set. Also, based on the style of recognition AVSR is classified as phoneme or viseme based recognition, isolated word or digit recognition, connected digit recognition, continuous speech recognition, and spontaneous speech recognition. The general flow of AVSR is shown in Fig 1. At first, the audio and visual information are separated from one another. Image frames are obtained from the video of speakers and lip regions are extracted from the image frames for lip tracking using a ROI (region of interest) detection algorithm. After ROI detection visual feature extraction methods are applied to the lip images prior to classification. Acoustic features are extracted from audio signals after preprocessing. There are two approaches for the integration of audio and visual modalities, feature fusion and decision fusion. In feature fusion, the acoustic and visual features are concatenated before being supplied to the classification model. In decision fusion, the acoustic and visual features are separately supplied to classification models and the final prediction is made based on some decision criteria.

There are numerous applications of speech recognition in the era of smart devices [5] [6]. AVSR benefits differently abled persons through hands free computation. Building a highly secured authentication system based on AVSR will ensure more reliable online transactions. AVSR can be employed to design smart home appliances and improve the performance of voice dialling and medical documentation as well. However, designing a real time robust AVASR system is computationally complex, but as per the survey carried out by us, it is observed that the number of publications over the years addressing the

challenges of AVSR are on a rising trend. This indicates the fascination of researchers to develop an AVASR system for intelligent human-machine interaction.

The motivation of this review is to study the progress of research to design an AVASR system for intelligent human-machine interaction. Accordingly,the complex problem of AVSR is described in a simplified way with the complete architecture of AVSR in this study. Research findings with significant performance and novel ideas corresponding to the challenges associated with every step of AVSR are reviewed and presented in a concise manner. The various audio and visual feature extraction, classification and modality integration techniques for AVSR are briefly discussed in our work under section II to motivate researchers who are interested to work in this domain. The performance of the different machine learning and deep learning models for audiovisual speech recognition is available in this review. The analysis done in this work lead to research opportunities mentioned in the section III. This study will assist researchers in gaining an overall understanding of AVSR and identifying areas of research interest related to AVSR.

II. Literature Review

A. Acoustic speech feature extraction

Preprocessing of the audio signal is an essential step before extracting audio features. For audio preprocessing, the primary work is to convert speech signals from analog to digital. Preprocessing includes voice activity detection(VAD) [7] [8], noise removal [9], pre-emphasis, framing, windowing and normalization [10]. VAD is necessary to determine the presence of human speech in an acoustic signal. In framing the continuous acoustic signal is segmented into a number of blocks, termed frames, each of 20 to 40 ms approximately. Windowing is the process of smoothing the discontinuities generated at the edges of a frame during framing. A relative analysis of various windowing functions is presented by K. M. Prabhu [11]. The Hamming window is widely used for speech recognition and is computed as

$$w(n) = 0.54 - 0.46cos(2\pi n/M - 1) \quad (1)$$

Here, M is the total length of the filter and n ranges from 0 to M-1. The last step of preprocessing is normalization to balance the signal spectrum and transform the speech data to normal form depending upon a threshold [12]. The different normalization functions are discussed in detail by Labied, M. [10].

After preprocessing features are extracted from the audio signal for classification. MFCC [13], PLP [14], and LPCC [15] are to name some of the prominent audio features. In 1980 Davis and Mermelstein developed Mel Frequency Cepstral Coefficients (MFCC) [16] which are widely preferred for speech recognition by researchers. The use of time derivatives along with static parameters makes recognition more efficient. In the frequency domain, MFCC uses the mel scale that relates

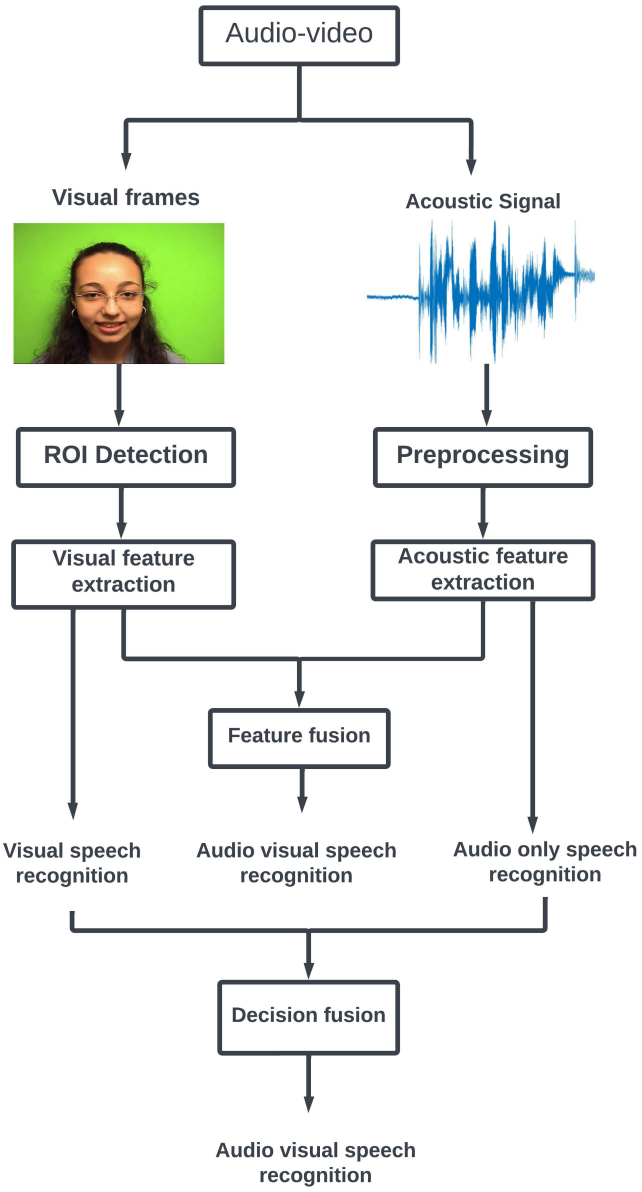

Fig. 1. Workflow of Audio Visual Speech Recognition

to the scale of our auditory system. The relation between mel scale and frequency of speech is represented as:

$$m = 2595 \log_{10}(1 + \frac{f}{700}) \quad (2)$$

Linear Predictive Codes (LPC) compute the signal's power spectrum and are used for formant analysis [17]. Perceptual Linear Prediction introduced by Hermansky is more adapted to human hearing compared to LPC, it eliminates unnecessary information and increases the recognition accuracy of speech. The various aspects of speech recognition from acoustic waveforms are discussed by N. Dave [18]. A review of various acoustic features and their properties along with the machine learning models is performed by Gaikwad, S. K et al. [19].

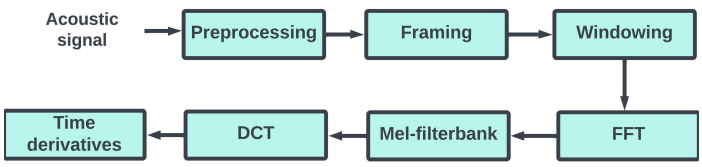

Fig. 2. Block diagram of MFCC

B. Visual Speech Recognition

1) Region of Interest detection: The first step of lipreading is identifying the region of interest (ROI) i.e. detection of the face and mouth area of the speaker. This is followed by lip movement tracking from consecutive image frames to recognize the word being spoken. Illumination independent robust lip tracking approach was adopted by Stiefelhagen, R. et al. [20] for lipreading that involved locating and tracking of eyes and nostrils along with lips to ensure recovery in case of tracking failure. A robust framework for object detection using an overcomplete wavelet representation and SVM classifier was employed for face and pedestrian detection in cluttered scenes by Constantine P. Papageorgiou et al. [21]. The advantage of motion cues was applied to boost the detection accuracy. A rapid object detection technique with high detection accuracy is presented by Paul Viola and Michael Jones which is popularly used in AVSR for ROI detection [22]. For selecting critical visual features from a large set of features AdaBoost is employed here. Skin color based face detection is another feasible choice. A histogram skin color model that uses histogram back projection for skin color segmentation is used by Liu, Q., and Peng, G. Z. for robust face detection [23]. Here, morphological and blob analysis optimizes the result of segmentation for images in Hue Saturation color space and an accuracy of 88.9% is obtained. An improvement over Violla Jones face detection method is observed [24] by implementing a hybrid of haar feature and skin color based model. Skin color post filtering after color compensation nullified the influence of varying illumination and gave a precision of 95.75 with the Bayesian classifier. Depth data captured using Microsoft Kinect Device improved the performance of AVSR over traditional approaches [25]. Degradation of visual data due to brightness and contrast conditions and Gaussian and block noise are considered here for the English digit data recognition task on the BAVCD dataset. Based on the intuition that the human gaze in a face-to-face conversation is not restricted to the speaker's lip region, extraoral information is obtained from the whole face, upper face, and cheek region of the speaker, which improved the performance of visual speech recognition [26]. They established the claim with their ROI selection method on SOTA models 3D ResNet18 and LipNet and achieved an accuracy of 85.02% for visual speech recognition.

2) Visual speech feature extraction and modelling: Visual speech features should be extracted from lip movement images for lipreading. Visual features are categorized as appearance and shape based features. Appearance based features are obtained by applying image transformation whereas shape based features consider the lip geometry of the speaker. A combination of these two types of features is seen in the literature to enhance the performance of AVSR.

S. Dupont and J. Luettin developed an appearance based lip model for visual speech recognition [27] where lip tracking and visual feature extraction are done with the help of the point distribution model [28]. With PLP and J-RASTA-PLP audio features they implemented three HMM based models on the M2VTS database. They observed that the combination of the streams at the state level significantly reduced the word error rate.

A segmentation technique is introduced by T. F. Cootes et al. [29] to extract lip contours as visual features using the Chain code marking algorithm. The database used in this work contains a total of 150 samples where ten numerals are uttered in Brazilian Portuguese by more than twenty speakers. This novel lipreading approach modelled with the nearest neighbour algorithm and Euclidean distance classifier achieved an overall success rate of 35%.

An approach to deal with visual speech recognition using side profiles of the speaker is seen in the work by P. Lucey. and G. Potamianos [30]. Static and dynamic visual features are obtained using DCT (discrete cosine transform) and LDA (linear discriminant analysis) from a connected digit database. Profile based recognition turned out to be less efficient than frontal view based recognition, but combining the profile and frontal view as a multiview AVASR system enhanced the performance in the presence of noise.

In order to minimize the dimensionality of DCT coefficients PCA (principal component analysis) according to energy is used by X. Hong et al. [31] and it is found that the lipreading accuracy is raised when the ultimate dimension is under a certain range. They compared the effect of entire DCT coefficients with block based DCT coefficients and concluded that block DCT works slightly better than entire DCT. Their speaker dependent lipreading system implemented on HIT Bi-CAVDB dataset acquired 77% accuracy.

Two types of visual features, active appearance model(AAM) and Sieve features are extracted from the English letter database for speaker independent lipreading [32]. HMM with Gaussian Mixture Model is employed for classification purposes. AAM performed better than Sieve with a maximum accuracy rate of 0.21 and a standard error of 0.05.

Optical flow is the measure of spatiotemporal dissimilarity between consecutive frames which is calculated by AA Shaikh et al. as a visual feature [33]. SVM (support vector machine) classifiers learn each viseme using the radial basis function as kernel function with an accuracy of 95.9%. Discrete English phonemes corresponding to the visemes defined in the Facial Animation Parameters (FAP) of MPEG4 standard are collected from 4 males and 3 females in constrained environments for this work.

Prashant Borde et al. carried out audio visual speech recognition for isolated words using the vVISWA dataset [34] using Zernike moment as shape based visual feature and MFCC as an acoustic feature. Euclidean distance classifier used for speech recognition to achieve a recognition rate of 100% and 63.88% for ASR(Acoustic speech recognition) and VSR(Visual speech recognition) respectively.

A novel visual speech recognition (VSR) technique using a hybrid of appearance based and shape based visual speech features with their time derivatives is proposed by S. Tamura et al [35]. Appearance based visual features PCA, DCT, LDA and GIF (GA based informative feature) are calculated in this work. The modeling is done using GMM-HMM. The accuracy of their deep bottleneck feature extraction approach for VSR and AVSR is found to be 73.66% and 90% respectively. Japanese dataset CENSREC-1-AV is introduced in this paper for the experiment. The importance of voice activity detection (VAD) to upgrade the performance of AVSR is also discussed.

Reconstructing intelligible acoustic speech from visual information is a challenging task. B. Milner and T. Le Cornu extracted AAM and 2D-DCT visual features to address this problem [36]. From these visual features, LPC coefficients and filterbank configuration are estimated to generate an acoustic signal by computing the time-frequency surface. GRID [37] dataset is used for this experiment which obtained an intelligibility of 49.02% only from visual information.

J. Wang et al. used Microsoft Kinect multi sensory device to collect 3D lip information and extract the proposed V3DL joint visual feature [38]. This work was conducted for isolated Chinese word recognition and resulted in 79.21% visual speech recognition accuracy with HMM-GMM architecture.

Appearance based visual features DCT and LBP-TOP (Local Binary Patterns on Three orthogonal Planes) for speaker independent connected digit and isolated phrase recognition are used by C. Sui et al. [39]. Here, a cascaded hybrid appearance visual feature (CHAVF) is proposed. To reduce the dimensionality of LBP-TOP features, different Mutual Information Feature Selectors(MIFS), namely MMI(Maximum Mutual Information), mRMR(minimal Redundancy Maximal Relevance) and CMI(Conditional Mutual Information) are experimented and mRMR is found to be the most suitable dimensionality technique. HMM classifier is used in this work for recognition to accomplish an accuracy of 69.18%. This work used the AusTalk dataset for AVSR and OuluVS for phrase classification.

Isolated digit recognition is performed by P. Borde et al. [40] using appearance based visual features DCT and LBP(Local Binary Pattern) with the aid of random forest classifier. LBP is a texture operator calculated by assigning binary labels to the neighbourhood pixels (generally 3x3 neighborhood) of an image by thresholding each pixel with the value of the centre pixel as shown in Fig 3. Their AVSR system was implemented over CUAVE and vVISWa datasets. Their AVSR approach using the vVISWa dataset after feature normalization obtained 100% recognition accuracy.

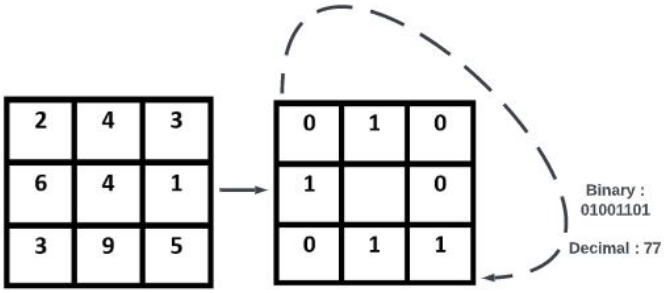

Fig. 3. Local Binary Pattern Computation

S. Debnath and P. Roy extracted pseudo-Zernike moment as shape based visual LBP-TOP and DCT as appearance based feature [41]. Based on the majority voting technique, hybrid classification methods are opted here using Artificial Neural Network, multiclass support vector machine, and Naive Bayes classifier. Singular value decomposition is considered for dimensionality reduction. The proposed approach performed better when a hybrid ANN+SVM classifier is used compared to SSH approach [42]. The vVISWA dataset is used in this experiment.

To overcome the ambiguity among conflicting classes a novel scheme is introduced in a recent work by Gonzalo D. Sad et al [43]. Cascaded classifiers are implemented with four different models HMM RF, SVM, and AdaBoost with three different audiovisual datasets. For the AV-CMU database, five different parabolic lip contours are extracted as visual features. Mouth height, mouth width, and the area between lips are extracted from AV-UNR video database. The visual features estimated for the AVLetters database are local spatiotemporal descriptors [44]. For feature normalization, the wavelet feature extraction technique was adopted. The proposed early integration method worked better with HMM and RF classifiers.

A comparative analysis of different visual feature extraction techniques that performed well for visual speech recognition can be drawn based on TABLE III. Optical flow is a motion based visual feature extraction technique that has produced outstanding speech recognition results. However, the use of optical flow as visual feature for AVSR is limited in the literature. On the other hand, DCT is computationally efficient and popularly used feature extraction technique which has produced impressive results for AVSR. From TABLE I, it is observable that appearance based visual features are preferred over shape based features, since shape based features depend upon the manual marking of facial landmarks and lip contour. Dimensionality reduction is a very important for the efficient processing of features. PCA and LDA are very efficient and frequently used as dimensionality reduction methods in the literature. Several traditional machine learning models are employed for visual speech recognition out of which the Hidden Markov Model(HMM), Support Vector Machine(SVM), and Random Forest(RF) performed well for lipreading and can be

TABLE I
TYPE OF VISUAL SPEECH FEATURE EXTRACTION TECHNIQUE USED IN LITERATURE

Paper	Appearance based visual feature	Shape based visual feature	Model based visual feature
S. Dupont J. Luettin [27]	x	x	✓
L.G. Da Silveira et al.	x	✓	x
X. Hong et al.	✓	x	x
P. Lucey G. Potamianos [30]	✓	x	x
AA Shaikh et al. [32]	x	x	✓
P. Borde et al. [34]	x	✓	x
C. Sui et al. [39]	✓	x	x
P. Borde et al. [40]	✓	x	✓
S. Debnath P. Roy [41]	✓	✓	x

experimented with to develop noise robust AVASR system. Different types of noises are superimposed with the clean speech in literature as shown in TABLE II to evaluate the performance of AVSR system in the presence of noise.

TABLE II
TYPE OF SPEECH AND NOISE DATA USED IN LITERATURE FOR AVASR

Paper	Noisy Speech/ Clean Speech	Type of Noise
S. Dupont, J Luettin [27]	Noisy Speech	Gaussian white noise
P. Lucey, G. Potamianos [30]	Noisy Speech	Babble Noise
AA Shaikh et al. [32]	Clean Speech	NA
P. Borde et al. [34]	Noisy Speech	-
S. Tamura et al. [35]	Noisy Speech	Interior car noise and musical waveform
C. Sui et al. [39]	Noisy speech	Additive white noise
P. Borde et al. [40]	Clean speech	NA

C. Integration of audio and visual modality

The diminishing performance of ASR in the presence of noise prompted researchers to concentrate on the bimodal perception of speech by humans and apply it to design intelligent human-machine interaction systems. There are different strategies adopted by researchers for combining audio and visual modalities. These strategies include early integration or feature fusion, modal fusion, and late integration or decision fusion [45] [46]. The early integration or feature fusion techniques rely upon concatenating the acoustic and visual features extracted from the audio waveforms and the speaker's image sequences of various utterances. However, there are certain limitations of feature fusion. In the case of continuous speech recognition, the extracted audio and visual features become extremely large, therefore the concatenation of audio and visual features can hamper the processing speed. Synchronization of acoustic and visual information is also non-trivial. Corruption of either visual or acoustic information can lead to miss classification if feature fusion is applied. Modal fusion is a mid level integration approach that overcomes the issues with feature fusion. MS-HMM is the widely used modal integration technique for AVSR. Decision fusion allows independent classification of speech from the two different modalities and is simpler to implement compared to previous strategies. However, the correlation between acoustic and visual configuration is overlooked by this approach.

To integrate audio and visual configuration two Dynamic Bayesian networks, based on a statistical model namely Coupled HMM and Fractional HMM are introduced by A. V. Nefian et al. [47]. To perform a speaker independent isolated word recognition task they extracted 2D DCT visual features followed by multiclass LDA and MFCC from the CMU database. CHMM is observed to perform better than MSHMM with the highest recognition rate of 98.1 %.

Decision fusion is computationally complex for continuous speech recognition tasks, and feature fusion is indifferent to the asynchronous nature of acoustic and visual speech. Georg F. Meyer et al. addressed this problem by applying the N best decision fusion method for audio visual speech recognition of digit data [48].

D. Deep learning based Audiovisual speech recognition

In recent times, deep learning architectures are gaining much popularity in the field of artificial intelligence when compared with traditional machine learning algorithms for their ability to process massive amounts of complex data and produce more accurate predictions. Likewise, audio visual speech recognition using deep learning techniques is found beneficial.

A. Sagheer et al. [49] proposed a shifting and rotation invariant lipreading approach on a Japanese sentence database. Here, visual speech features are extracted using the unsupervised neural network based hypercolumn model(HCM) and HMM. Their approach performed better than self-organizing-map(SOM) [50] and DCT based systems [51].

The idea of cross modality learning and shared representation learning from audio and visual data is introduced by J. Ngiam et al. [52]. The issues raised here are resolved using deep autoencoders and the performance is compared with an RBM based model and other existing models. They observed that their video only deep autoencoder implemented on CUAVE and AVLetters dataset performs better than the bimodal deep autoencoder.

AVSR for connected digit sequences is performed using Deep Belief Network (DBN) by J. Huang and B. Kingsbury [53]. Two categories of DBN are implemented, 1 frame and 3 frame DBN and a held-out set of data is used to adjust the learning rate. For the fusion of audio and visual modality, a fusion of mid level features learned by modality specific DBN is considered here. Their work lags behind the baseline model for clean data due to small amounts of training and held-out data but outperforms in noisy conditions.

TABLE III
VISUAL SPEECH RECOGNITION PERFORMANCE FOR DIFFERENT FEATURE EXTRACTION METHODS

Paper	Visual Feature	Model	Dataset	Result
Xiaopeng Hong et al. [31]	DCT+PCA	SCHMM	Bi-CAVDB	Accuracy: 77.1%
Ayaz A. Shaikh et al [33]	Optical flow	SVM	English Phoneme	Accuracy: 95.9%
Prashant Borde et al. [34]	Zernike moments + PCA	Euclidean Distance	vVISWa	Recognition rate : 63.88%
S. Tamura et al. [35]	DBVF-PDLGC	HMM	CENSREC-1-AV	Accuracy: 77.80%
Chao Sui et al. [39]	CHAVF	HMM	AusTalk	Accuracy : 69.18%
Prashant Borde et al. [40]	DCT+LBP+LDA	Random Forest	vVISWa	Recognition rate:83.75%
S. Debnath, P. Roy [41]	PZM+DCT+LBP-TOP+SVD	ANN+SVM	vVISWa	Accuracy : 78.45%

Noise robust acoustic features, Log mel scale filterbank and MFCC with temporal derivatives are acquired by K. Noda et al. for AVSR [54] utilizing deep denoising autoencoder. Visual features are extracted from visual frames utilizing Convolutional Neural Network(CNN). Their visual speech recognition approach with independent CNN for each speaker outperformed when a combination of 32 Gaussian components was implemented on the Japanese audiovisual dataset [55].

Di Hu & X Li introduced a novel deep learning architecture named Temporal multimodal Restricted Boltzman Machines (RTMRBM) [56]. This architecture models multimodal sequences by transforming the sequences of connected MRBs into a probabilistic series model. MRBMs define the joint distribution over audio modality, visual modality, and shared hidden units. The accuracy of the designed audio visual speech recognition system implemented on AVLetters, AVLetters2 and AVDigits dataset is 66.04% and has outperformed MDAE [57] and CRBM [58] based approach.

Multimodal recurrent neural networks based AVSR system considering the sequential characteristics of audio and visual modalities is implemented by W. Feng et al. [59] on AVLetters dataset. Acoustic speech recognition has been performed with bidirectional LSTM-RNN (Long Short Term Memory-Recurrent Neural Network). The visual speech recognition is done with CNN and bidirectional LSTM RNN. The weighted state layer in both audio and visual parts generates semantically consistent output for fusion. The video only recognition part worked better for CNN+unidirectional LSTM based work with 57.7% accuracy compared to CNN+bidirectional LSTM.

Fei Tao, Carlos Busso [60] observed AVSR does not always perform better than ASR when the speech is clean. To maintain the performance of AVSR under all conditions this work introduced Gating Neural Network as an observation model with HMM which outperformed the conventional GMM-HMM and DNN-HMM models. They used the CRSS-4ENGLISH-14 Corpus which contains read and spontaneous speech.

Bidirectional gated Recurrent Units classification is performed [61] on features extracted from raw image pixels and acoustic signals to design an end-to-end audio visual speech recognition system. Their deep learning based approach has given a classification rate of 98% on LRW corpus and was found to be efficient in a high noise environment.

An RNN transducer based model for speaker independent AVSR using Youtube videos is proposed by T. Makino et al. [62]. The encoder part of RNN-t architecture takes a stack of mel-filterbank coefficients from the audio frames and V2P features from video frames. The encoder consists of a 5-layer stack of bidirectional LSTM and the decoder has 2 layers of unidirectional LSTMs. The word error rate achieved by this model for visual only, audio only, and audio visual inputs is 33.6%, 4.8%, and 4.5% respectively which outperforms the CTC-V2P and TM-Seq2Seq model.

Integration of audio and visual speech modality for isolated word recognition using Long short-term memory(LSTM) and feedforward convolutional neural network is experimented on by S. P. Kulkarni [63]. The classification for the audio signal is done employing a deep neural network and the same for visual speech is done using long short-term memory recurrent neural network. The accuracy of their AVSR model was found to be 92.38%.

An end-to-end audio visual speech recognition model is designed by P. Ma et al. [64] with ResNet-18 and a convolution-augmented transformer, called conformers. At the front end, features are directly extracted from raw pixels and acoustic signals using audio and visual encoders. Their approach implemented on the LRS2 and LRS3 datasets resulted in a significant reduction in word error rate.

E. Performance measures

Accuracy, recognition rate(RR), and word error rate(WER) are popular performance measures used by researchers to evaluate the performance of their audiovisual speech recognition model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

In equation 3 TP, TN, FP and, FN represent true positive, true negative, false positive, and false negative respectively.

$$RR = \frac{C_s}{T_s} \times 100\% \tag{4}$$

C_s and T_s stand for total supplied and correctly identified test sample in equation 4.

$$WER = \frac{I + D + S}{N} \tag{5}$$

I,D, and S in equation 5 refer to the number of insertions, deletions and substitutions, and N refers to the total number of words in the vocabulary.

F. Audio visual speech database

Audio visual speech recognition problems are classified in the literature as phoneme based recognition, isolated word and digit recognition, connected digit recognition, continuous speech recognition, and spontaneous speech recognition. Based on the type of recognition researchers have developed various audiovisual corpora to conduct their experiments.

Fig. 4. Sample frames from CUAVE Dataset

Fig. 5. Sample frames from vVISWa Dataset

1) **M2VTS [65]**: A total of 185 recordings of 37 subjects among which 12 females and 25 males are there who continuously pronounced French digits from zero to nine. Five recordings were taken from each speaker with an audio track recorded at 48kHz sampling frequency and 16-bit PCM coding. From these recordings, 27000 visual frames are generated and converted to a gray level image. From the five utterances of each speaker, three pronunciations are considered for the training set, one for the development set and one for the test set.
2) **CMU [66]**: Carnegie Mellon University generated this audio visual corpus from 10 subjects, 7 males and 3 females who uttered 78 isolated words, every ten times.
3) **IBMSmartroom [67]** : 38 speakers uttering connected digit strings recorded inside IBM Smartroom using two microphones, one head mounted, another omnidirectional, and three pan-tilt-zoom cameras for frontal and two side views. Two synchronous audio streams at 22kHz and three visual streams at 30Hz and 368x240 pixel frames are available in this dataset.
4) **HIT Bi-CAVDB [31]** : Harbin Institute of Technology Bimodal Chinese Audio Video database contains utterances of 10 people, 5 males, and 5 females pronouncing 96 different Chinese syllables.
5) **CUAVE [68]**: Clemson University Audio Visual Experiment dataset has two sections, in one section there are individual speakers and in the other section, there are speaker pairs. The videos are taken for 36 hours, 19 males and 17 females uttered connected and continuous digits spoken under different situations.
6) **Tulips 1.0 [69]**: 12 Undergraduate students from the Cognitive Science Program at UCSD are made to utter the first 4 digits in English. This dataset was compiled at Movellan's laboratory at the department of Cognitive Science, UCSD.
7) **AVLetters [70]**: Isolated English alphabets from A to Z are uttered by 10 talkers, five male, and five female. A total of 780 utterances were recorded without any head restraint considering three repetitions from each speaker.
8) **BAVCD [71]**: Bilingual Audio Visual Corpus with depth information contains connected digit utterances of 15 subjects in English and 6 subjects in Greek. 500 connected digit utterance is recorded for the English part and 2200 connected digit utterance is recorded for the Greek part.
9) **vVISWA [72]**: Visual Vocabulary of Independent Standard Words recorded from 48 native speakers and 10 non-native speakers. They have captured the full frontal, 45 degree, and side profile of each speaker. Their dataset encompasses three languages Marathi, Hindi, and English.
10) **CENSREC-1-AV [35]**: Audio visual Corpora and environments for noisy Speech Recognition are established in the information processing society of Japan where Japanese connected digit utterances are recorded from 20 female and 22 male speakers constructing a total of 3234 utterances.
11) **AVLetters2 [32]**: 26 letters of the English alphabet are pronounced seven times by each of the five subjects. The videos are captured using a tri-chip Thomson Viper Filmstream high definition Camera and the head movement and emotion of the speakers were constrained.
12) **AVDigits [56]**: This dataset is recorded from 53 participants 41 male and 12 female from three different views frontal, 45 degree, and side profile. The dataset has two parts, the digit part consists of zero to nine uttered in English in random order five times by each speaker. In the other part selective short phrases are repeated 5 times by each of the 39 speakers, 32 male, and 7 female in three different modes, neutral, whisper, and silent speech.
13) **AusTalk [73]**: This dataset contains words, digit strings, sentences, stories and spontaneous speech recorded by 861 adult speakers whose ages range from 18 to 83 years, belonging from 15 different locations in Australia. This dataset was recorded between June 2011 to June 2016 and the speakers have spoken in Australian English.
14) **OuluVS [44]**: 20 people, 17 males and 3 females of different nationalities have spoken ten everyday greetings, one to five times to construct this audiovisual dataset of 817 image sequences. A SONY DSR-200AP 3 CCD

camera with a frame rate of 25 fps was used to collect the data.

15) **XM2VTSDB [74]**: This is the extended M2VTS database containing utterances of 295 subjects. The subjects were asked to read three given phonetically balanced sentences each twice. Sony VX1000E digital camcorder and DHR 1000UX digital VCR are used to record the data.
16) **AVTIMIT [75]**: Frontal view of 223 speakers, 117 male, and 106 female uttering TIMIT SX sentences are captured using SONY DCR-VX200 video camcorder.
17) **LRS2 [76]**: Thousands of spoken sentences from BBC television are collected for this database where each sentence is approximately 100 characters in length. Speaker labels do not exist for this database and there are large variations in the head pose.
18) **LRS3 [77]**: This audiovisual dataset contains 400 hrs of TED and TEDx talks.

III. Research Opportunities

The efficiency of AVSR system depends upon the performance of each step associated with it. The audio and visual feature extraction techniques that can optimize the performance of AVSR is still inconclusive. It is observed that for the same set of features and classification model, the performance of AVSR varies for different datasets. Concatenation of different features for modeling improves the performance of AVSR, however, the increased size of the feature vector effects the processing time. Therefore, feature dimensionality reduction is an important aspect to design a real time AVSR system. Deep learning is a relatively new approach in the area of AVSR and is recommendable because of its ability to handle the massive amount of raw data efficiently. Instead of using handcrafted features with traditional machine learning models, deep learning and neural network techniques extract robust features for AVSR. Integration of modalities is the most challenging task in AVSR. Although, the concatenation of audio and visual feature seems to simplify the task of classification, synchronization of audio and visual frames is a little explored subject that researchers should look into in the future. For continuous speech recognition decision fusion or model fusion is a better strategy than feature fusion. VSR is unaffected by noise, therefore it should be given higher weight in decision fusion in the presence of noise. Adaptive weight selection for audio and visual modality based on SNR is a crucial and least explored area in AVSR. Very few research works considered and dealt with the possible visual noises which require attention. AVSR should not only upgrade the performance of speech recognition in the presence of ambient noise, but it should also retain the performance of ASR in clean environments. Also, we discovered that the majority of the studies in this sector only addressed the individuals' frontal profiles, which may not be viable in real life scenarios and hence may attract the attention of researchers. The ultimate challenge of AVSR is designing a real-time robust AVASR application that will be less computationally complex where robustness indicates that the system should be necessarily illumination, rotation, and scaling invariant to give high performance in the presence of noise at any form and intensity.

IV. Conclusion

Knowledge of the importance and demand of intelligent human-machine interaction in this era, motivated us to conduct this survey and throw light on the latest research accomplishments in the direction of AVSR to find future research opportunities. Accordingly, the various feature extraction techniques for both acoustic and visual speech signals and their contribution to audiovisual speech recognition are discussed. Different classification techniques suggested by researchers for AVSR are investigated and compared in our work depending upon their performance. The set of acoustic and visual features that can maximize the performance of audiovisual speech recognition in both clean and noisy environments is yet inconclusive. Real life examples of applications taking the advantage of audiovisual speech recognition are rare. Our contribution to this study is to summarise the progress of research for each stage of AVSR in a straightforward manner. Deep learning possesses the capability of deep feature extraction from raw data and it is relatively new in the domain of AVSR. Thus, deep learning appears to be the potential solution to designing efficient noise-robust audio visual speech recognition applications.

References

[1] McGurk, H. and MacDonald, J., 1976. Hearing lips and seeing voices. Nature, 264(5588), pp.746-748.

[2] Petajan, E., Bischoff, B., Bodoff, D. and Brooke, N.M., 1988, May. An improved automatic lipreading system to enhance speech recognition. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 19-25).

[3] Rabiner, L. and Juang, B., 1986. An introduction to hidden Markov models. ieee assp magazine, 3(1), pp.4-16.

[4] Goldschen, A.J., Garcia, O.N. and Petajan, E., 1994, November. Continuous optical automatic speech recognition by lipreading. In Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers (Vol. 1, pp. 572-577). Ieee.

[5] Rudrapal, D., Das, S., Debbarma, S., Kar, N. and Debbarma, N., 2012. Voice recognition and authentication as a proficient biometric tool and its application in online exam for PH people. International Journal of Computer Applications, 39(12), pp.6-12.

[6] Singh, S. and Yamini, M., 2013, July. Voice based login authentication for Linux. In 2013 International Conference on Recent Trenthe ds in Information Technology (ICRTIT) (pp. 619-624). IEEE.

[7] N. Xu, C. Wang, and J. Bao, "Voice activity detection using entropy-based method," in 2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS), Dec. 2015, pp. 1–4, doi: 10.1109/ICSPCS.2015.7391751.

[8] Bharath Y.K, Veena S, Nagalakshmi K.V, M. Darshan, and R. Nagapadma, "Development of robust VAD schemes for Voice Operated Switch application in aircrafts: Comparison of real-time VAD schemes which are based on Linear Energy-based Detector, Fuzzy Logic and Artificial Neural Networks," in 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2016, no. 1, pp. 191– 195, doi: 10.1109/ICATCCT.2016.7911990.

[9] S. Lee and H. Kwon, "A Preprocessing Strategy for Denoising of Speech Data Based on Speech Segment Detection," pp. 1–24, 2020, doi: 10.3390/app10207385.

[10] Labied, M., Belangour, A., Banane, M., & Erraissi, A. (2022, March). An overview of Automatic Speech Recognition Preprocessing Techniques. In 2022 International Conference on Decision Aid Sciences and Applications (DASA) (pp. 804-809). IEEE.

[11] K. M. Prabhu, Window Functions and Their Applications in Signal Processing. CRC Press, 2014.

[12] R. Singh, U. Bhattacharjee, and A. K. Singh, "Performance Evaluation of Normalization Techniques in Adverse Conditions," Procedia Comput. Sci., vol. 171, no. 2019, pp. 1581–1590, 2020, doi: 10.1016/j.procs.2020.04.169.

[13] Ravikumar, K.M., Rajagopal, R. and Nagaraj, H.C., 2009, June. An approach for objective assessment of stuttered speech using MFCC. In The international congress for global science and technology (p. 19).

[14] Hönig, F., Stemmer, G., Hacker, C. and Brugnara, F., 2005. Revising perceptual linear prediction (PLP). In Ninth European Conference on Speech Communication and Technology.

[15] Ananthi, S., and P. Dhanalakshmi. "SVM and HMM modeling techniques for speech recognition using LPCC and MFCC features." In Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014, pp. 519-526. Springer, Cham, 2015.

[16] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on acoustics, speech, and signal processing, 28(4), 357-366.

[17] B. P. Yuhas, M. H. Goldstein Jr., T. J. Sejnowski, and R. E. Jenkins, "Neural network models of sensory integration for improved vowel recognition," Proc. IEEE, vol. 78, Issue 10, pp. 1658–1668, Oct. 1990.

[18] Dave, N. (2013). Feature extraction methods LPC, PLP and MFCC in speech recognition. International journal for advance research in engineering and technology, 1(6), 1-4.

[19] Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2010). A review on speech recognition technique. International Journal of Computer Applications, 10(3), 16-24.

[20] Stiefelhagen, R., Meier, U. and Yang, J., 1997. Real-time lip-tracking for lipreading. In Fifth European Conference on Speech Communication and Technology.

[21] Papageorgiou, C. P., Oren, M., & Poggio, T. (1998, January). A general framework for object detection. In Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271) (pp. 555-562). IEEE.

[22] Viola, P., & Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001 (Vol. 1, pp. I-I). Ieee.

[23] Liu, Q., & Peng, G. Z. (2010, March). A robust skin color based face detection algorithm. In 2010 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR 2010) (Vol. 2, pp. 525-528). IEEE.

[24] Erdem, C. E., Ulukaya, S., Karaali, A., & Erdem, A. T. (2011, May). Combining Haar feature and skin color based classifiers for face detection. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1497-1500). IEEE.

[25] Galatas, G., Potamianos, G. and Makedon, F., 2012, June. Audio-visual speech recognition using depth information from the Kinect in noisy video conditions. In Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments (pp. 1-4).

[26] Zhang, Y., Yang, S., Xiao, J., Shan, S., & Chen, X. (2020, November). Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 356-363). IEEE.

[27] Dupont, S. and Luettin, J., 2000. Audio-visual speech modeling for continuous speech recognition. IEEE transactions on multimedia, 2(3), pp.141-151.

[28] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models—their training and application," Comput. Vis. Image Understand., vol. 61, pp. 38–59, Jan. 1995.

[29] Da Silveira, L.G., Facon, J. and Borges, D.L., 2003, October.Visual speech recognition: a solution from feature extraction to words classification. In 16th Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2003) (pp. 399-405). IEEE.

[30] Lucey, P. and Potamianos, G., 2006, October. Lipreading using profile versus frontal views. In 2006 IEEE Workshop on Multimedia Signal Processing (pp. 24-28). IEEE.

[31] Hong, X., Yao, H., Wan, Y. and Chen, R., 2006, December. A PCA based visual DCT feature extraction method for lip-reading. In 2006 International Conference on Intelligent Information Hiding and Multimedia (pp. 321-326). IEEE.

[32] Cox, S.J., Harvey, R.W., Lan, Y., Newman, J.L. and Theobald, B.J., 2008, September. The challenge of multispeaker lip-reading. In AVSP (pp. 179-184).

[33] Shaikh, A.A., Kumar, D.K., Yau, W.C., Azemin, M.C. and Gubbi, J., 2010, October. Lip reading using optical flow and support vector machines. In 2010 3Rd international congress on image and signal processing (Vol. 1, pp. 327-330). IEEE.

[34] Borde, P., Varpe, A., Manza, R. and Yannawar, P., 2015. Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition. International journal of speech technology, 18(2), pp.167-175.

[35] Tamura, S., Ninomiya, H., Kitaoka, N., Osuga, S., Iribe, Y., Takeda, K. and Hayamizu, S., 2015, December. Audio-visual speech recognition using deep bottleneck features and high-performance lipreading. In 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA) (pp. 575-582). IEEE.

[36] Milner, B., & Le Cornu, T. (2015). Reconstructing intelligible audio speech from visual speech features. Interspeech 2015.

[37] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audiovisual corpus for speech perception and automatic speech recognition," Journal of the Acoustical Society of America, vol. 150, no. 5, pp. 2421–2424, Nov. 2006

[38] Wang, J., Zhang, J., Honda, K., Wei, J., & Dang, J. (2016). Audio-visual speech recognition integrating 3D lip information obtained from the Kinect. Multimedia Systems, 22(3), 315-323.

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

[40] Borde, P., Kulkarni, S., Gawali, B., & Yannawar, P. (2020). Recognition of Isolated Digit Using Random Forest for Audio-Visual Speech Recognition. Proceedings of the National Academy of Sciences, India Section A: Physical Sciences, 1-8.

[41] Debnath, S. and Roy, P., 2021. Appearance and shape-based hybrid visual feature extraction: toward audio–visual automatic speech recognition. Signal, Image and Video Processing, 15(1), pp.25-32.

[42] Liu, G.H., Yang, J.Y. and Li, Z., 2015. Content-based image retrieval using computational visual attention model. pattern recognition, 48(8), pp.2554-2566.

[43] Sad, G.D., Terissi, L.D. and Gómez, J.C., 2022. Complementary models for audio-visual speech classification. International Journal of Speech Technology, pp.1-19.

[44] Zhao, G., Barnard, M. and Pietikainen, M., 2009. Lipreading with local spatiotemporal descriptors. IEEE Transactions on Multimedia, 11(7), pp.1254-1265.

[45] Katsaggelos, A. K., Bahaadini, S. & Molina, R. Audiovisual Fusion: Challenges and New Approaches. Proc. IEEE 103 (2015), 1635–1653.

[46] Seong, T. W., & Ibrahim, M. Z. (2018). A review of audio-visual speech recognition. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10(1-4), 35-40.

[47] Nefian, A.V., Liang, L., Pi, X., Liu, X. and Murphy, K., 2002. Dynamic Bayesian networks for audio-visual speech recognition. EURASIP Journal on Advances in Signal Processing, 2002(11), pp.1-15.

[48] Meyer, G.F., Mulligan, J.B. and Wuerger, S.M., 2004. Continuous audio–visual digit recognition using N-best decision fusion. Information Fusion, 5(2), pp.91-101.

[49] Sagheer, A., Tsuruta, N., Taniguchi, R.I. and Maeda, S., 2005, March. Visual speech features representation for automatic lip-reading. In Proceedings. (ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. (Vol. 2, pp. ii-781). IEEE.

[50] N. Tsuruta, H. Iuchi, A. Sagheer, T. Tobely, "Self–Organizing Feature Maps for HMM Based Lip-reading,"The 7th Int. conf. on Knowledge-Based Intelligent Information & Engineering Sys, KES03, 2: 162-168, 2003.

[51] M. Heckmann, K. Kroschel, C. Savariaux, F. Berthommier,"DCT-Based Video Features For Audio-Visual Speech Recognition," Proc. Of Inter. Conf. on Spoken Language Processing, ICSLP: 1925-1928, 2002.

[52] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. and Ng, A.Y., 2011, January. Multimodal deep learning. In ICML.

[53] Huang, J. and Kingsbury, B., 2013, May. Audio-visual deep learning for noise robust speech recognition. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 7596-7599). IEEE.

[54] Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H.G. and Ogata, T., 2015. Audio-visual speech recognition using deep learning. Applied Intelligence, 42(4), pp.722-737.

[55] Kuwabara, H., Takeda, K., Sagisaka, Y., Katagiri, S., Morikawa, S. and Watanabe, T., 1989, May. Construction of a large-scale Japanese speech database and its management system. In International Conference on Acoustics, Speech, and Signal Processing, (pp. 560-563). IEEE.

[56] Hu, D., & Li, X. (2016). Temporal multimodal learning in audiovisual speech recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3574-3582).

[57] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In Proceedings of the 28th international conference on machine learning, pages 689– 696, 2011.

[58] M. R. Amer, B. Siddiquie, S. Khan, A. Divakaran, and H. Sawhney. Multimodal fusion using dynamic hybrid models. In Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on, pages 556–563. IEEE, 2014.

[59] Feng, W., Guan, N., Li, Y., Zhang, X. and Luo, Z., 2017, May. Audio visual speech recognition with multimodal recurrent neural networks. In 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 681-688). IEEE.

[60] Tao, F., & Busso, C. (2018). Gating neural network for large vocabulary audiovisual speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(7), 1290-1302.

[61] Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., & Pantic, M. (2018, April). End-to-end audiovisual speech recognition. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 6548-6552). IEEE.

[62] Makino, T., Liao, H., Assael, Y., Shillingford, B., Garcia, B., Braga, O. and Siohan, O., 2019, December. Recurrent neural network transducer for audio-visual speech recognition. In 2019 IEEE automatic speech recognition and understanding workshop (ASRU) (pp. 905-912). IEEE.

[63] Kulkarni, S.P., 2021. Integration of Audio video Speech Recognition using LSTM and Feed Forward Convolutional Neural Network.

[64] Ma, P., Petridis, S. and Pantic, M., 2021, June. End-to-end audio-visual speech recognition with conformers. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7613-7617). IEEE.

[65] Pigeon, S., & Vandendorpe, L. (1997, March). The M2VTS multimodal face database (release 1.00). In International Conference on Audio-and Video-Based Biometric Person Authentication (pp. 403-409). Springer, Berlin, Heidelberg.

[66] Advanced Multimedia Processing Lab, http://amp.ece.cmu.edu/projects/AudioVisualSpeechProcessing/, Carnegie Mellon University, Pittsburgh, Pa, USA

[67] Potamianos, G., & Lucey, P. (2006, September). Audio-visual ASR from multiple views inside smart rooms. In 2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (pp. 35-40). IEEE.

[68] Patterson, E. K., Gurbuz, S., Tufekci, Z., & Gowdy, J. N. (2002, May). CUAVE: A new audio-visual database for multimodal human-computer interface research. In 2002 IEEE International conference on acoustics, speech, and signal processing (Vol. 2, pp. II-2017). IEEE.

[69] Morade, S. S., & Patnaik, S. (2015). Comparison of classifiers for lip reading with CUAVE and TULIPS database. Optik, 126(24), 5753-5761.

[70] Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., & Harvey, R. (2002). Extraction of visual features for lipreading. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2), 198-213.

[71] Galatas, G., Potamianos, G., Kosmopoulos, D., McMurrough, C., & Makedon, F. (2011). Bilingual corpus for AVASR using multiple sensors and depth information. In Auditory-Visual Speech Processing 2011.

[72] Borde, P., Manza, R., Gawali, B., & Yannawar, P. (2016). vVISWa–A Multilingual Multi-Pose Audio Visual Database for Robust Human Computer Interaction. International Journal of Computer Applications, 137(4), 25-31.

[73] Wagner, M., Tran, D., Togneri, R., Rose, P., Powers, D., Onslow, M., ... & Ambikairajah, E. (2011). The big australian speech corpus (the big asc). In SST 2010, Thirteenth Australasian International Conference on Speech Science and Technology (pp. 166-170). ASSTA.

[74] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. In Second International Conference on Audio and Video-based Biometric Person Authentication, 1999.

[75] Hazen, T. J., Saenko, K., La, C. H., & Glass, J. R. (2004, October). A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. In Proceedings of the 6th international conference on Multimodal interfaces (pp. 235-242).

[76] Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2018). Deep audio-visual speech recognition. IEEE transactions on pattern analysis and machine intelligence.

[77] Afouras, T., Chung, J. S., & Zisserman, A. (2018). LRS3-TED: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496.