Migration Letters

Volume: 21, No: S3 (2024), pp. 735-747

ISSN: 1741-8984 (Print) ISSN: 1741-8992 (Online)

www.migrationletters.com

Survey And Anlysis On Automated Speech Reading Techniques On Various Languages Using Deep Learning

Divya[*1], Dr.Suresha D[2], Dr. Sanjeev Kulkarni[*3,] Navya Rai[4], Jyothi Prasad[5]

Abstract

Speech reading is a process of understanding the language with the help of lip movements with or without sound especially useful for hearing impaired, or aged people. In the presented paper, we explore variation of lip texture features for recognition of visemes. Aiding the cause, Deep learning, and Machine learning sign language recognition approach is used for Regional Indian Language word recognition and achieved high accuracy at an estimate of computation. A study of comparisons on automated Speech reading approaches is presented in this paper with the main focus being with regard to deep learning and related methodologies with promising result for both feature extraction and classification schema for lip-reading sentences. It might investigate the efficiency of different neural network topologies such as (CNNs), (RNNs), in lip reading exercises, or attentional strategies. RNN is a recurrent kind of artificial neural network that utilizes time series data or sequential data. A neural network type called a convolutional network architecture, or CNN or ConvNet, is particularly adept at processing input with a grid-like architecture, like an image. Work focuses on benefits of Attention-Transformers and Temporal Convolutional Networks for classification compared to Recurrent Neural Networks. Paper includes analysis of well-organized lip reading techniques for varied linguistic systems. Comparison of various Algorithms, discussion of the trade-offs between many algorithms, adopted methodologies performance and limitations have been formulated. The study focuses more on Machine Learning & Deep Learning technologies regardless of particular application areas for speech reading domain.

Keywords—*Lip Reading, Feature extraction, Spatio-Temporal Residual CNN, Feature classification, Hidden Markov Model (HMM), MTCNN, D3D (DenseNet 3D)*

Introduction

Human action recognition is an important area of research which aims to precisely explain human behavior or act only from visual information is an impressive skill but an uphill battle for the novice. With the advancing technologies like AI, Deep Learning, Lip Reading plays a supreme role in grasping human speech using various techniques and approaches. The task of lip reading focuses more elementary language structures which include characters or phonemes[1] to boost machine-powered lip reading.

The main challenges to speech reading sentences are:

[1] Dept. of Information Science & Engineering, AJIET, Mangalore, Karnataka, India, ORCID: 0009-0002-4305-1521
[2] Dept. of Information Science & Engineering, AJIET, Mangalore, Karnataka, India, ORCID: 0000-0003-2578-0552
[3]Dept. of Computer Science & Engineering, Srinivas University Institute of Engineering & technology, y, Mukka, Mangalore, Karnataka, India,
ORCID: 0000-0002-3957-1711
[4]Dept. of Information Science & Engineering, AJIET, Mangalore, Karnataka, India,
[5]Department of Computer Science & Engineering, MITE, Moodbidri.Karnataka, India
ORCID iD: 0009-0003-6187-6356

- Lip reading systems can predict the words or characters that users are taught or skilled to predict. Speech reading need “listening” to observe the speaker's face or lip movement to determine speech patterns, movements, gestures and expressions for the prediction of words.

- Lip reading systems may have to work on noisy environments which involve further pre-processing, speech recognition systems fails or performs poorly, because of the extra noise signals

- Speech reading models must be trained to cover a wide range of accents, similar letter shapes vocabulary which requires a significant number of arguments in optimized models and a trained significant volume of data to be used.

Notice the growth of this field, many cumulative array of articles include new algorithms for interpreting speaker lips and classifiers to detect, decipher individual words, and recognize lip movements. In this survey, the review and evaluation of speech reading research works on a variety/multiple of languages, highlighting the progression of using deep learning models to assist deaf and hard-of-hearing. The advances and overcomes in Deep Lip Reading has been highlighted in this paper focusing on the accuracy of Audio Speech Recognition systems, problems and challenges of current lipreading methods, detailed information and the methods applied. Paper gives survey of using machine learning by applying deep learning and neural networks to devise an automated lip-reading system.

In this survey, we review the primary methodology in the development of lipreading used in various languages. The paper focuses on language like Kannada, English, Hindi, Urdu, etc. Primary there are two methods- traditional manual speech reading and deep learning speech reading. Automated deep learning lip-reading involving stages:

- Reading image file representing speech to be decoded.
- Data processing and obtaining high-quality data to train your model, where the region of interest (ROI) lip is extracted.
- Feature Extraction (Frontend) - This involves extracting visual, temporal and relevant features from redundant features.

- Classification is done at Backend this involves classify lip movements into phonemes, lower dimensional feature vector is then stitched phonemes back into words.

The main contributions and organization insights of this survey are:

i) Comparison of various automated lip-reading approaches to locate and extract the subject of interest (ROI) from the video image in different languages.

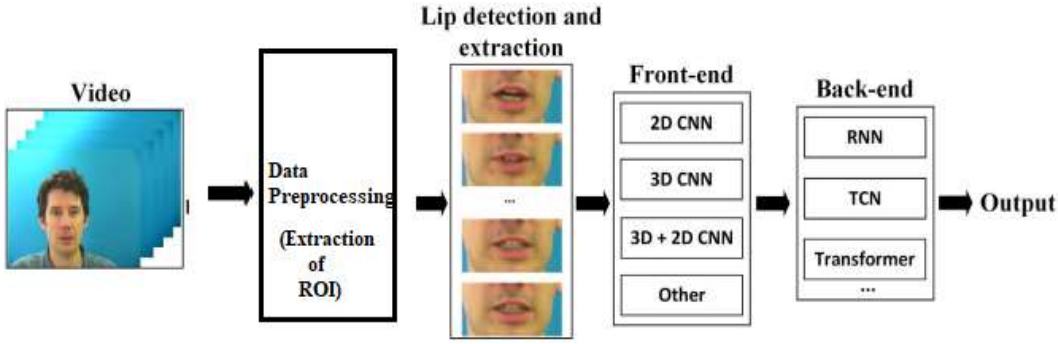
ii) A comparison of different feature extraction schemas used by various works and the algorithms. Some of the main models include color information-based, Recurrent Neural Network (RNN), Boltzmann Machines (BM), Temporal Convolutional Network (TCN), and Transformers, Feedforward Neural Network (FNN), Deep Belief Network (DBN), Auto-encoder, and 3D Convolutional Neural Network (CNN)., model-based and mixed feature extraction.

iii) Critical review on the feature extraction transformation to lower the dimensionality of features. Common methods -Linear Discriminant function analysis (LDA), Artificial Neural Network (ANN), Discrete Cosine Transform (DCT), and Discrete Wavelet Transform (DWT), PCA and to process the extracted features and encode them into equal length feature vectors.

iv) A comparison of different most up-to-date classification networks and classification schemas used for speech-reading in different languages. Common methods include Template Matching, Recurrent Neural Networks (RNNs). DTW, Naive Bayes classifiers,

Bi-LSTMs, GRUs and LSTMs with attention, HMM, Support Vector Machine (SVM), and Time-Delay Neural Network (TDNN).

Figure 1 gives framework on Deep learning lipreading process: Firstly, the lip region (ROI) is located and extracted from the video, learn deeper features from the experimental data then the temporal and spatial features are pulled out by the front-end, and finally, input these pulled out characteristics as input to the back-end and are concatenated at each time



step for classification.

Figure 1: **General framework for automated lip-reading**

The challenges associated with lip reading in tonal languages, such as Chinese, and the lack of Cantonese lip reading datasets necessitate further research. In this paper [32], we address these issues by introducing CLRW, a comprehensive Cantonese lip reading dataset.

I. PREPROCESSING

Preprocessing methods are essential for lip reading because of the raise in the visual input standard , reduce noise, and extract important information for subsequent investigation. Here are a few typical lip-reading preprocessing methods:

1. Video stabilization: Lip reading frequently entails watching videos of people speaking. Smoother and more stable lip movement trajectories are produced using video stabilisation techniques to decrease undesirable camera jitter and motion.
2. Frame selection: Processing a series of video frames may be necessary while lip-reading. Key frames that capture crucial lip movements or linguistic clues can be found using frame selection algorithms. This lessens the complexity of processing and concentrates analysis on useful frames.
3. Face detection and tracking: Face detection and tracking techniques accustomed to locate and follow the face throughout the movie before the lip area is pulled out. These methods provide accurate localization of the lip region of interest, allowing for accurate study of lip movements.
4. After the face has been identified and tracked, the region around the lips has to be isolated for additional study. The precise region around the lips is extracted using methods like face alignment or form modelling while taking into consideration variations in stance, scale, and lighting conditions.
5. Normalisation of illumination: The way that lip motions seem to the eye depends a lot on the lighting. By changing the colour channels of the lip region or balancing the brightness and contrast between frames, illumination normalisation approaches try to lessen the effects of lighting changes.
6. Noise reduction: Accurate lipreading can be hampered by noise, including camera noise and outside disruptions. Wavelet denoising or filters (such as the Gaussian, median, or bilateral filters) are used as noise reduction techniques to reduce undesired noise while maintaining crucial lip characteristics.
7. Motion estimate: Accurate motion estimation is necessary for analyzing lip motions. The motion of lip contours between two next frames may estimated using optical flow methods, which reveals details about the temporal dynamics of lip motion.

8. After being isolated and preprocessed, the lip region is next subjected to a variety of feature extraction techniques in order to gather pertinent visual information. These could be geometric features like motion vectors or lip shapes, features based on how something looks, such texture descriptors, or deep learning-based information gleaned from previously trained models.

II. FEATURE EXTRACTION

Feature extraction is essential for extracting pertinent data from lip photos or video frames in deep learning-based lip reading. Here are a few widely used deep learning strategies for lipreading feature extraction:

A. (ConvNet/CNN)

Simulated neuron layers are used by CNNs built on ImageNet to perform a number of complicated vision-related tasks, including corner extraction and lipreading. CNNs extracts detailed visual information from lip images or video frames for lip reading. While the following pooling layers aid in reducing spatial dimensions and extracting dominating features, the convolutional layers in CNNs learn to capture local spatial patterns and structures.

B. Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs): RNNs, such Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU), are frequently employed for capturing temporal dependencies in sequential data, which qualifies them for lip activity examination. To mimic the temporal dynamics in lip motions, RNNs process lip sequences or temporal representations taken from lip films.

C. 3D Convolutional Neural Networks (3D CNNs)

Lip reading involves examining both the spatial patterns and the time evolution of quick and involuntary movements of lips using 3D Convolutional Neural Networks (3D CNNs). In collecting both spatial and temporal variables at once, 3D CNNs extend the notion of CNNs to three dimensions. The volumetric structure of lip films can be utilised by 3D CNNs to directly learn spatiotemporal representations from the data.

D. Transformer-based Models

Transformer models, originally developed for challenges involving natural language processing, encouraging findings have been obtained with relation methods for visual speech comprehension. These models make use of self-attentional techniques to identify cross-lip frame global relationships. Transformer-based models can successfully extract features related to the lips for tasks like lip reading by paying attention to informative frames and learning contextual relationships.

E. Pretrained Models and supervised Transfer Learning

For the purpose of geometric feature extraction of lip textures during lipreading, trained mannequins, like those using sizable image or video datasets (like ImageNet or Kinetics), can be employed. The pretrained models can capture generic visual features through the use of transfer learning, which can then be adjusted or used as fixed feature extractors for particular lip reading tasks. This method is especially helpful when there aren't enough labelled samples or when the lip reading dataset is tiny.

F. Autoencoders and Variational Autoencoders (VAEs)

Unsupervised learning algorithms like autoencoders and VAEs are capable of learning condensed representations of video frames or lip pictures. These models capture crucial information and eliminate extraneous noise by encoding and recreating the input data. Later lip reading challenges can make use of the learnt representations as features.

The difficulty of the lip reading problem, the accessibility of labelled data, the availability of computational resources, and the performance requirements all play a role in the decision of which deep learning approach to use for feature extraction in lip reading. Researchers

frequently test out various architectures and modify them to fit the special features of datasets including lip reading.

III. METHODOLOGY

Automated lip-reading approaches and a variety of algorithms have been developed using Neural Network Methodologies , cutting-edge techniques of deep learning for better & optimized lip detection.

In paper[3] demonstrates Myanmar consonant identification implementing CIELa*b* color shifts, Moore Neighborhood Mapping Algorithm, and linear SVM classification technique for one syllable sounds to extract accurate lip motions localizing each lower and upper lip region. First, Lip localization is carried out to segment lip region, tracking contours between the top and lower lip lines. CIEL*a*b* color-shifting technique is used for segment lip region, Lip tracking results for utterance of Ga Gyi for tracing the boundaries of the lip (a two-syllable consonant) on only a chosen frame, use Moore neighborhood. More features can be considered for recognising additional consonants in Myanmar's syllables.

This study will provide a novel, visually-based approach to teaching and learning the Myanmar language for deaf people. Figure 2 show color enhanced, transformed and histogram equalized image.

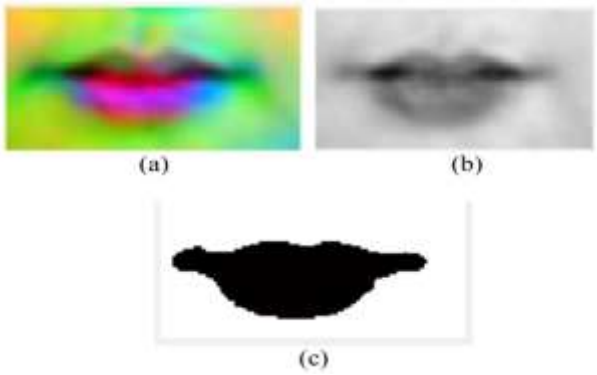


Figure 2: **a) Color enhanced image b) Transformed image c) histogram equalization**

Data preprocessing [5] MTCNN or Multi-Task CNN is utilized to identify facial landmarks when analyzing photos using a neural network.

In paper [3], the authors address the issue of automatic lip-reading by introducing a novel form of optical sensor, event cameras. A biological influence can be seen in the optical sensors used in event cameras. Event cameras, in contrast to ordinary cameras, record per-pixel brightness changes asynchronously at the microsecond level. Event cameras have substantial technological and practical advantages over ordinary cameras for the ALR challenge that necessitates the observation of fine-grained spatiotemporal features: Since only variations in the scene's luminance are captured, event cameras' output does not contain much redundant visual information. Additionally, they are able to do so because of their excellent temporal resolution to capture finer-grained movements. Finally, they are low-power and can function in difficult lighting conditions, which are crucial in practical applications.

Paper [7] the Multi-Temporal Lip-Audio Memory (MTLAM) technique is used to offer the best use of audio signals to supplement the lack of information provided by lip movements. Here, contrastive loss and reconstructive loss are used. The goal of the reconstruction loss is to close the gaps between the recovered multi-temporal audio components. The contrastive loss is then used to store various audio attributes in various value memory slots.

Specifically, the question of automatic speech reading, provides event cameras [8] as a novel sort of optical sensor. Event cameras, in contrast to ordinary cameras, record per-pixel brightness changes asynchronously at the microsecond level. Event cameras can record the high temporal resolution needed to capture finer-grained movements for the automatic lip-reading task, which calls for the perception of fine-grained spatiotemporal features. Only brightness changes of the scene are recorded, removing redundant visual information, and they can operate in difficult lighting conditions, which are crucial in real-world applications. Asynchronous event data is individually output at each pixel by an event camera. In this paper, we choose for synchronous frame-like representations from asynchronous event data. We suggest a multi-branch network with message flow modules connecting several branches, each of which takes as input event frames with various temporal resolutions. The model's many branches concentrate on learning Spatio-temporal properties at various granularities. Finer temporal information is contained in the features from the higher-rate branch, which might direct learning from the lower-rate branch to concentrate on particular features. The bidirectional Gate Recurrent Unit that the sequence model employs accepts the output of the multibranch network as input.

In [2] D3D algorithm is used on Tibetan lip reading dataset TLRW-50, Connectionist Temporal Classification, a method for training neural networks with more layers than two. Method is tested on 20 speakers to record these 50 kinds of

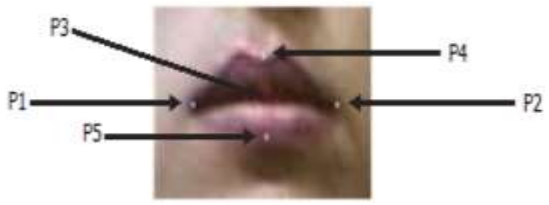
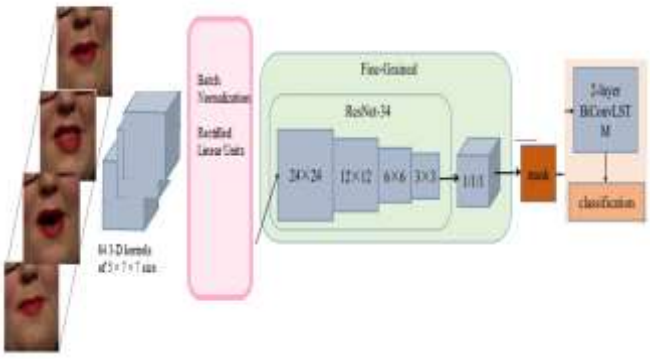


Figure 3: lip contour points to compute lip geometric features.



Fg. 4 The architecture of D3D with attention-based long short-term memory (LSTM)

Tibetan Words based on psychological, spectral, and experimental analysis. In order for the input and output sequences to be aligned during decoding and for the network to train autonomously, the Connectionist Temporal Classification (CTC) loss was employed.

To localize the face and lips[4] from the given image, Haar cascade classifiers. In Haar Feature-Based Cascade Classifier calculation is done by finding out the difference of the average of the positive and negative images based on cascade function.

Lip Reading based on neural network [4] helpful for hearing impaired by extracting only lip contour points is proposed. Geometric lip features used to extract lip size, extremities

and inner lip features using Pseudo Hue and Luminosity for each frame of the video. Figure 3: lip contour points to compute lip geometric features. The neural network is used for Hindi word identification and is expandable to more languages, improving performance.

Andrei [5] suggests a lip-reading system that is not language-dependent D3D and Visual attention autoencoder. The experiments are carried on LRM (Lip Reading Multilingual) dataset for English, Romanian, and Mandarin language showing the feasibility and effectiveness of the systems. Here two popular lip-reading methods are used - the D3D (DenseNet 3D) and visual attention autoencoder network. D3D (DenseNet 3D) model make use of 3D CNN layers as the front-end network. Next method Visual attention autoencoder is used top of the ResNet-34 network. Figure 4 shows D3D (DenseNet 3D) with a 3-layer Sequential data processing of bidirectional RNNs.

In the paper [6] a powerful Lip2Speech method is developed for wild environment using multimodal supervision. Connectionist Temporal Classification (CTC) loss maintain the visual-audio synchronization. Synthesizing the acoustic features can focus more on modelling speech content. Moreover, we employ a deeper speech synthesiser made of 1D CNN layers in order to properly integrate the speaker features in the speech output.

Proposed method using LRS2 wild dataset consist of 142,000 utterances including pre-train and train sets. LRS3 dataset consist of 1,308 utterances are used for testing., and LRW word-level English BBC news program datasets constrained to 500 words. The total training data length is about 157 hours.

Event-based Action Recognition Methods: Vision sensors for cameras with dynamics of a scene have been used in numerous other applications to remove duplicate information, however, the two that matter the most are gesture recognition [13] and gait recognition [14]. They fall within the category of action recognition tasks, just like lip-reading [9, 10, 11, 12].

In this essay [16], we concentrate on the particular difficulties associated with lip reading and suggest specialized solutions. This strategy focuses on enhancing how photos or videos are represented visually. For various classification and detection tasks, [17, 18] use attention-weighted averages of visual characteristics as building blocks, while OCNet [19] models the context between pixels for semantic segmentation using self-attention. Early efforts in the field of visual speech detection relied on manually created visual characteristics and statistical modelling techniques like HMMs, GMMs, and PCA [20,21–23,24–25,26–28]. In more recent publications, algorithms based on optical flow [29] or a CNN and LSTM combination [30, 31] were presented.

On sub-clips [16], a spatio-temporal residual CNN is applied to extract visual spatial feature maps. To extract basic visual information, the input video frames are run through a spatio-temporal CNN. The next step is the independent processing of each input frame by a shared Visual Transformer Pooling (VTP) block. It is then combined with spatial positional encodings (SPE). A visual attention mask is then extracted using a learnable query vector. The self-attended feature map is weighted averaged using the attention mask. One token at a time, a text token sequence is predicted using an encoder-decoder Transformer model from the source video embedding sequence. A beam search is used to eventually infer an output sentence from these distributions.

We [32] suggest a brand-new two-branch network called TBGL that has both a local and a global branch. While the local branch divides the feature into three sections to capture modest local lip motion, the global branch concentrates on modelling the entire lip. We achieve comparable performance on several lip reading datasets, including LRW, CAS-VSR-W1K, and CLRW, by jointly training these branches. We first submitted the input image sequence into the front-end in order to extract visual spatiotemporal features. Then, we employ a back-end with two branches; the global branch of TBGL simulates the entire motion of the lip, while the local branch splits the feature into three sections to concentrate on the subtle local motion of the lip. Finally, we use a bidirectional knowledge distillation

loss to add more supervision while the two branches of the TBGL are being trained together.

IV. PERFORMANCES IN LIP READING

In this section, the evaluation results are summarized in Table I by comparing the performance of some lipreading algorithms over the different data subsets for languages like Romanian, Mandarin, English, Kannada, Hindi etc .

Methodology [1] works on visualizing the lip movement and recognizes the Kannada words and analyzing the movement of lips for different words of a Kannada Dictionary.

Deep weighted feature descriptions to encode and decode the type of Kannada words like Vottakshara and Deerga spoken by a Kannada speaking person. The ROI is extracted by using the features and Texture shape of Lips with active contours. For calculations using complex math models, deep nets are crucial for estimation of shape features by assigning labels to the shapes of lips in each and every video frame. The words of Kannada language or "Kagunitha" are combined together to form words of a sentence. Similarly, each "Kannada Kagunitha" matches specific lip movement and label is assigned to a particular sequence of shapes of lip movement. These lip assigned labels are compared with similarity scores to finalize the type of labels to be generated as a text output. The challenging issues of Recognizing Kannada words from tongue twisting, Kannada Kagunitha which can be solved using e texture and shape. Dataset Lip Reading in the Wild (LRW) is used to to learn the features of shapes and texture, using deep weighted feature descriptors method. The proposed method has yielded a classification accuracy of 84.82% using decision tree classification.

The accuracy achieved in [4] by the proposed approach is 97%. The proposed algorithm is applied for recognition of ten words of Hindi language and can be easily extended to include more words of other languages. The presented approach will be helpful for hearing impaired or dumb people to communicate with humans or machines. The proposed algorithm is fast as well as robust to various occlusions

In [5] dataset LRW for English, LRW-1000 for Mandarin, and LRRo for Romanian is used. LRRo has 47 words which include 27 easy and 20 hard words contains a total of 5,386 utterances. Dataset LRW consist of 49,199 utterances for English language. LRW-1000 dataset contains a total of 20,606 utterances for the Mandarin language. Multilingual dataset using 3 different language English, Mandarin, and Romanian is formed together of 75,191 utterances.

The paper [6] multiple speakers with unconstrained sentences are verified for effectiveness using LRS2, LRS3, and LRW datasets. Proposed model achieves comparable performances with the other popular methods using word-level dataset, LRW. On our DVS-Lip dataset, all the experiments in this study were carried out. They will see indiscriminative spatio-temporal features as a result of discarding the fine-grained temporal and spatial information in this situation [8]. The evaluation on the DVS-Lip test set shows that our suggested MSTP is noticeably better than the most advanced event-based and video-based approaches. Current CNN-based techniques [9,11,12,14,15] only accept event frames with a fixed frame rate as input, making them unsuitable for tasks like lip-reading that call for the perception of fine-grained spatiotemporal details. In contrast, by feeding multi-grained event frames into various branches, our MSTP can learn both comprehensive spatial features and fine temporal features.

TABLE I: Comparison of different lip reading techniques in different databases.

SI. No.	**Technique Used**	**Dataset**	**Accuracy(%)**

1	Proposed Deep Weighted Feature Representation (Kannada words)[1]	Lip Reading in the Wild(LRW)	84.82
2	CIELa*b* and Moore Neighborhood tracing algorithm (TWO SYLLABLE CONSONANTS- eg: ಂ (Ga Nge) [3]	Database of Myanmar consonants	91.66
3	Vector Quantization neural network (identification of 10 Hindi words) [4]	Hindi Words	97%
4	D3D (DenseNet 3D) [5]	LRW for English	83.2
		LRRo for Romanian	41.8
		LRW-1000 for Mandarin	37.3
		LRM	74.5
	Visual attention autoencoder network	LRW for English	82.0
		LRW-1000 for Mandarin	39.3
		LRRo for Romanian	62.9
		LRM	74.6
5	Automatic Speech Recognition (ASR) model using CTC and CNN [6]	LRW dataset	13.86
6	Multi-grained SpatioTemporal Features Perceived Network (MSTP) [8]	Event-based lipreading dataset (DVS-Lip)	72.10
7	Spatio-Temporal CNN + Visual Transformer Pooling (VTP)+ Transformer encoder-decoder + Beam search decoding and rescoring. [16]	TED and TEDx talks dataset	88.2
8	TBGL MS-TCNs (Back-end)	LRW and CAS-VSR-W1K datasets.	88.4 49.1

	3D CNN+ ResNet-18(Front-end)[32]		
9	MS-TCNs ensemble MS-TCNs (Back-end) 3D CNN+ ResNet-18(Front-end) self-distillation[33]	LRW and CAS-VSR-W1K datasets.	88.5 46.6

[16] LRS2 includes video excerpts from several British television programs, including Country file and Top Gear; the transcribed content totals over 224 hours. Over 5,000 English-language TED and TEDx presentations totaling 475 hours have been compiled into LRS3. Each movie is available at a resolution of 224 x 224 and a frame rate of 25. The input videos are first downsized to a resolution of 160 square pixels, and then a central 96 square pixel crop is taken. To determine if someone is speaking in that frame or not, VTP applies the previously introduced Linear Transformer, an additional fully connected (FC) layer, and a sigmoid activation on top of the frame-level encoder outputs.

Because Chinese is a tonal language, it is challenging for Chinese speakers to discriminate between lexical meanings, unlike English. This research presents CLRW dataset, an 800-word class Cantonese lip reading dataset with 400,000 samples. As in [16], even here the 3D CNN layer will conduct a preliminary 455 spatial-temporal alignment for our front-end network. The following action is to extract 458 discriminative features using a ResNet-18 module.

In order to create CLRW [32], a wide variety of samples had to be gathered, covering things like gender, age, postures, lighting, and speaking speed. The distribution of real-world scenarios is reflected by CLRW, which increases its practical applicability by not restricting these variables. We benchmark CLRW and provide a thorough analysis of the outcomes. The evaluation highlights the CLRW dataset's difficulty and how it might help with upcoming Cantonese speech reading tests.

The survey demonstrates the significant contributions of deep learning techniques to lip reading research. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms have shown promising results in improving lip reading accuracy across different languages. The availability of language-specific datasets is crucial for accurate lip reading performance. The survey highlights the importance of creating comprehensive datasets that encompass various demographics, accents, speech speeds, and environmental conditions. Such datasets enable the development of robust lip reading models that can handle real-world scenarios.

V. CONCLUSION

The survey on lip reading in various languages, using different datasets, provides valuable insights into the challenges and advancements in this field. The survey reveals that lip reading varies significantly across languages due to factors such as phonetic diversity, tonal variations, and distinct lip movements. Tonal languages like Chinese pose unique difficulties, while languages with complex phonetic systems require specialized models and training techniques.

In conclusion, the lip reading survey across various languages and datasets underscores the language-specific challenges, the importance of diverse datasets, advancements in deep learning techniques, the potential of transfer learning, and the practical applications of lip

reading technology. Further research and development in this field will continue enhancing the precision and applicability of lip reading systems across different languages and real-world scenarios.

ACKNOWLEDGEMENTS

I **Author1 Mrs. Divya** would like to thank to Research Guide Dr. Suresha D for his constant support and guidance helped to complete this research article..

AUTHOR CONTRIBUTION

Author 1: Divya is responsible for analyzing the collected data, summarizing findings, and drawing conclusions from the survey results. She also designed the survey methodology, including data collection methods, criteria for selecting papers, and the overall research approach. Have made a substantial contribution in conducting statistical analyses or other formal investigations

Author 3: Dr. Sanjeev Kulkarni who worked out almost all of the technical details, and performed the numerical calculations for the suggested experiment. Contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

Author4 Navya Rai who extensively reviewed existing literature on speech reading techniques to identify gaps, trends, and relevant studies. She took lead role in creating figures, charts, or other visual representations of data. She carried out the experiment and were in charge of overall direction and planning.

Author5 Jyothi Prasad Contributed in planning, supervised the work. He processed the experimental data, performed the analysis, drafted the manuscript and designed the figures.

REFERENCES

[1] Nandini M S, Trisiladevi C. Nagavi , Nagappa U.Bhajantri, “Deep Weighted Feature Descriptors for Lip Reading of Kannada Language” 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), 13 May 2019

[2] Zhenye Gan; Xinke Yu; Hao Zeng; Tianqin Zhao “Tibetan lip reading based on D3D” 2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), 03 February 2022

[3] Thein Thein; Kalyar Myo San, “Lip movements recognition towards an automatic lip reading system for Myanmar consonants”, 2018 12th International Conference on Research Challenges in Information Science (RCIS), 09 July 2018

[4] Neeru Rathee “A novel approach for lip reading based on neural network”, 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), 18 July 2016

[5] Andrei-Cosmin Jitaru; Liviu-Daniel Ştefan; Bogdan Ionescu, “Toward Language-independent Lip Reading: A Transfer Learning Approach”, 2021 International Symposium on Signals, Circuits and Systems (ISSCS), July 2021

[6] Minsu Kim, Joanna Hong, Se Jin Park, and Yong Man Ro, “Lip-To-Speech Synthesis In The Wild With Multi-Task Learning,” IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP), Vol. 24, Pp. 4342–4355, 2022.

[7] Jeong Hun Yeo, Minsu Kim, Yong Man Ro, “MULTI-TEMPORAL LIP-AUDIO MEMORY FOR VISUAL SPEECH RECOGNITION”, ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 05 May 2023

[8] Ganchao Tan; Yang Wang; Han Han; Yang Cao; Feng Wu; Zheng-Jun Zha, “Multi-grained Spatio-Temporal Features Perceived Network for Event-based Lip-Reading”, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 27 September 2022

[9] Zhengwei et al. Wang. Action-net: Multipath excitation for action recognition. In CVPR, 2021. 3, 6, 7

[10] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. Advances in Neural Information Processing Systems, 34, 2021. 3

[11] Joao Carreira et al. Quo vadis, action recognition? a new model and the kinetics dataset. In CVPR, 2017. 3, 6, 7

[12] Zhaoyang Liu et al. Tam: Temporal adaptive module for video recognition. In ICCV, 2021. 3, 6, 7

[13] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In CVPR, pages 7243–7252, 2017. 2, 3

[14] Yanxiang Wang, Bowen Du, Yiran Shen, Kai Wu, Guangrong Zhao, Jianguo Sun, and Hongkai Wen. Ev-gait: Event based robust gait recognition using dynamic vision sensors. In CVPR, pages 6358–6367, 2019. 2, 3, 6, 7

[15] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014. 5

[16] K R Prajwal; Triantafyllos Afouras; Andrew Zisserman,"Sub-word Level Lip Reading With Visual Attention", 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 27 September 2022

[17] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. In Proc. ICLR, 2018. 3

[18] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In Proc. CVPR, pages 7794–7803, 2018. 3

[19] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. In arXiv preprint arXiv:1809.00916, 2018. 3

[20] Andrew Aubrey, Bertrand Rivet, Yulia Hicks, Laurent Girin, Jonathon Chambers, and Christian Jutten. Two novel visual voice activity detectors based on appearance models and retinal filtering. In 2007 15th European Signal Processing Conference, pages 2409–2413, 2007. 3

[21] Peng Liu and Zuoying Wang. Voice activity detection using visual information. In 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 1, pages I–609, 2004. 3

[22] Qingju Liu, Andrew J. Aubrey, and Wenwu Wang. Interference reduction in reverberant speech separation with visual voice activity detection. IEEE Transactions on Multimedia, 16(6):1610–1623, 2014. 3

[23] Qingju Liu, Wenwu Wang, and Philip Jackson. A visual voice activity detection method with adaboosting. In Sensor Signal Processing for Defence (SSPD 2011), pages 1–5, 2011. 3

[24] Foteini Patrona, Alexandros Iosifidis, Anastasios Tefas, Nikolaos Nikolaidis, and Ioannis Pitas. Visual voice activity detection in the wild. IEEE Transactions on Multimedia , 18(6):967–977, 2016. 3

[25] Gerasimos Potamianos, C. Neti, Juergen Luettin, and Iain Matthews. Audio-visual automatic speech recognition: An overview. Issues in audio-visual speech processing, 2004. 3

[26] Spyridon Siatras, Nikos Nikolaidis, Michail Krinidis, and Ioannis Pitas. Visual lip activity detection and speaker detection using mouth region intensities. IEEE Transactions on Circuits and Systems for Video Technology, 19(1):133–137, 2009. 3

[27] David Sodoyer, Bertrand Rivet, Laurent Girin, Christophe Savariaux, jean-luc Schwartz, and Christian Jutten. A study of lip movements during spontaneous dialog and its application to voice activity detection. The Journal of the Acoustical Society of America, 125:1184–96, 03 2009. 3

[28] D. Sodoyer, B. Rivet, L. Girin, J.-L. Schwartz, and C. Jutten. An analysis of visual speech information applied to voice activity detection. In 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, volume 1, pages I–I, 2006. 3

[29] Andrew J. Aubrey, Yulia A. Hicks, and Jonathon A. Chambers. Visual voice activity detection with optical flow. Iet Image Processing, 4:463–472, 2010. 3

[30] Sylvain Guy, Stephane Lathuili ´ ere, Pablo Mesejo, and Radu ` Horaud. Learning visual voice activity detection with an automatically annotated dataset. In 2020 25th International Conference on Pattern Recognition (ICPR), pages 4851– 4856. IEEE, 2021. 3

[31] Rahul Sharma, Krishna Somandepalli, and Shrikanth Narayanan. Toward visual voice activity detection for unconstrained videos. In 2019 IEEE International Conference on Image Processing (ICIP), pages 2991–2995, 2019. 3

[32] Lip Reading in Cantonese,"Yewei Xiao, Lianwei Teng , Aosu Zhu, Xuanming Liu, And Picheng Tian " ,10.1109/ACCESS.2022.3204677, 05 September 2022
[33] J. S. Chung and A. Zisserman, ''Out of time: Automated lip sync in the 726 wild,'' in Proc. Asian Conf. Comput. Vis. Springer, 2016, pp. 251–263.