Lip Reading using deep learning and neural networks

Akash Chandra
Department of CSE
VNRVJIET Hyderabad

Charitha Paruchuri
Department of CSE
VNRVJIET Hyderabad

Dr. Bheemalingappa
Department of CSE
VNRVJIET Hyderabad

A.Karthika
Department of CSE
VNRVJIET Hyderabad

P.Yogitha
Department of CSE
VNRVJIET Hyderabad

Abstract- This survey paper provides a meticulous examination of the intricacies inherent in lip reading, accentuating the formidable hurdles posed by linguistic diversity and nuanced articulatory patterns. Employing machine learning methodologies, specifically within the domains of deep learning and neural networks, we conducted training exercises on a subset of the dataset utilizing two discrete Convolutional Neural Network (CNN) architectures. The principal aim of this endeavor was to engineer an automated lip-reading framework adept at discerning spoken language exclusively through visual lip movements. It is noteworthy that even seasoned lip-reading experts encounter challenges, approximating a mere fraction of uttered words. The trained models underwent meticulous scrutiny to gauge their efficacy in predicting words accurately. Subsequently, the most performant architecture was seamlessly integrated into a real-time web application. This scholarly exploration underscores the transformative potential of neural networks in advancing the frontiers of lip-reading technology, thereby augmenting accessibility and inclusivity.

Keywords: Lip reading, Deep learning, Neural networks, CNN architectures, Word prediction, Real-time application.

I. INTRODUCTION:

Lip reading, the interpretative process of spoken language through the observation of lip movements, has emerged as a prominent focus of research, largely propelled by the availability of extensive datasets such as the Lip Reading in the Wild (LRW) corpus. These datasets not only serve as catalysts for progress in the field but also open avenues for investigating and refining lip-reading technology. A prevalent approach in this field involves a structured lip-reading pipeline, comprising a visual encoder, a temporal model, and a softmax classification layer. While the foundational architecture of this pipeline has been established, current research efforts focus on meticulous refinements of its constituent components, particularly the temporal model and associated training methodologies. The visual encoder, positioned at the core of the lip-reading pipeline, plays a crucial role in processing visual lip input. A specific visual encoder has gained widespread acceptance among researchers, prompting further investigations into optimizing the temporal model or enhancing training methodologies. Among the temporal models undergoing scrutiny are the Multi-Scale Temporal Convolutional Networks (MS-TCNs) and Bidirectional Gated Recurrent Units (BGRUs), extensively explored in the literature. However, ongoing discussions within the scientific community emphasize the necessity for a comprehensive examination to definitively determine the comparative efficacy of these models in the context of lip-reading applications.

This study aims to bridge the existing gap by conducting a systematic investigation and comparative analysis of MS-TCNs and BGRUs. A critical evaluation of their respective performances in lip-reading scenarios will be presented. In addition to temporal models, researchers have explored various data augmentation strategies to enhance the robustness and generalizability of lip-reading models. These strategies include mixup, variable length augmentation, and cutoff, all designed to

reinforce the resilience of the models by diversifying the training data. This diversification is essential for adapting to environmental variations, pronunciation nuances, and the intricacies of speech articulation.

II. LITERATURE SURVEY:

A lip-reading system utilizing neural networks was introduced by Souheil Fenghour, Daoqing Chen [1]. This system is vocabulary-free and relies solely on visual cues. It demonstrates the ability to analyze lip-reading on a wide range of phrases, even identifying words not present in the training data by training on a small set of visemes as classes. The researchers evaluated the system's performance on the LRS2 dataset and found that it achieved a word mistake rate 15% lower than many build models.

Souheil Fenghour[2] explored automated lip-reading techniques, specifically focusing on the efficacy of deep learning methods in extracting and categorizing features. This paper extensively delves into automated lip-reading systems, encompassing various aspects such as audiovisual databases, methods for feature extraction, networks for classification, and classification schemas. The survey critically assesses different classification schemas employed in lip-reading, which include ASCII characters, phonemes, and visemes. It also provides a thorough evaluation of the advantages associated with Attention-Transformers and Temporal Convolutional Networks in comparison to Recurrent Neural Networks for classification. Additionally, the paper conducts a comparative analysis of CNN with alternative neural network architectures for feature extraction, offering noteworthy contributions and insightful perspectives.

A novel approach to speech augmentation utilizing lip-reading was introduced by Ahsan Adeel, Mandar Gogate [3]. Their methodology integrates deep learning and analytical acoustic modeling, specifically employing a filtering-based strategy, which diverges from established methods predominantly relying on deep learning. The audio-visual (AV) speech enhancement framework undergoes evaluation at two levels, employing both ideal AV mapping and LSTM-driven AV mapping techniques, and is compared against conventional spectral subtraction and log-minimum mean-square error procedures. The authors assess their EVWF approach across various real-world scenarios and diverse signal-to-noise ratio (SNR) levels, utilizing ChiME3 corpora for objective testing. The evaluation encompasses perceptual assessment of speech quality and the standard mean-opinion-score approach, supported by inferential statistics for subjective testing. Comparative simulations demonstrate significant improvements in voice quality and intelligibility attributed to the fusion of lip-reading and speech enhancement methods.

Yuanyao Lu and Jie Yan[4] introduced an innovative approach to automated lip reading, asserting that conventional lip reading algorithms utilizing predefined features lack the ability to comprehensively grasp the significance of lip movement sequences. To address this limitation, the authors propose a hybrid neural network architecture that combines CNN and BiLSTM. The approach focuses on discerning local features through convolutional procedures while also capturing hidden correlations in temporal information derived from lip image sequences. Results from their lip reading recognition experiments indicate that their suggested method surpasses established techniques such as the active contour model (ACM) and hidden Markov model.

Ms. A.Sangeerani Devi, Ramya.M [5] proposed a computer-assisted instruction (CAI) for hearing-impaired students that includes two components. The first module comprises the instructional lesson, allowing learners to study the terms presented. The second module involves a multiple-choice game, requiring students to predict the word spoken by the speaker in a lip-reading video and select the correct answer option.

In their research, Apurva H. Kulkarni and Dr. Dnyaneshwar Kirange [6] conducted a thorough evaluation of various lip reading methods and language datasets within the domain of deep learning. The study covers an introduction to automated lip reading

approaches, specifically focusing on three essential components of a lip reading system: lip detection and localization, feature extraction and reading recognition. The first section dedicated to lip detection and localization techniques explores two main approaches. Firstly, methods based on gray/color information utilize color information for accurate lip positioning. Secondly, algorithms based on geometric information construct a rough mouth area based on face proportions. Feature Extraction Method: Traditional feature extraction approaches in lip reading systems are categorized into two groups: pixel-based and model-based. Recognition Models: The techniques for lip reading recognition encompass various approaches, including template matching, Hidden Markov Model (HMM), Dynamic Time Warping (DTW), Artificial Neural Networks (ANN). This section provides insights into the diverse methods employed for lip reading recognition.

Jonathan Noyola, Sameep Bagadia, and Amit Garg[7] propose multiple ways for recognizing words and sentences only from video data, without depending on audio signals. In our work, we deploy a pre-trained VGGNet primarily intended for detecting human faces of celebrities obtained from IMDB and Google Images. We study numerous approaches for employing this network to process picture sequences. The VGGNet is trained on picture concatenations produced from many frames within each sequence. Furthermore, they incorporate LSTMs to collect temporal information. Given the limited success of LSTM models, ascribed to several variables, the concatenated visual model using nearest-neighbor interpolation exhibits performance, obtaining a validation accuracy of 76%.

Bo Xu, Cheng Lu, Yandong Guo, and Jacob Wang[8] proposed a two-stage model for speech recognition aimed at enhancing audio speech recognition (ASR) through the incorporation of visual information, ultimately advancing multi-modality speech recognition (MSR). In the initial phase, their focus is on elevating ASR by leveraging visual cues from lip movements to distinguish target speech from background noise, thereby enhancing the model's ability to accurately "listen." Subsequently, in the second phase, the authors integrate the audio modality with the visual modality to improve speech recognition by implementing a dedicated MSR sub-network. To boost the model's performance, several key enhancements are introduced. They employ a P3D-based visual front-end to extract distinctive features. Additionally, they enhance the temporal convolution block by replacing the 1D ResNet with TCN, which is better suited for temporal tasks. To validate their approach, extensive experiments are conducted on the LRS3-TED and LRW datasets. Their two-phase model, termed audio-enhanced multi-modality speech recognition (AE-MSR), consistently outperforms existing state-of-the-art methods by a substantial margin, affirming the efficacy of AE-MSR in improving speech recognition in multi-modal settings.

Pingchuan Ma, Stavros Petridis[9] have introduced a novel approach to Video Super-Resolution (VSR) that surpasses the performance of conventional systems trained on publicly available datasets by a significant margin. Their solution involves implementing a VSR model with auxiliary tasks, simultaneously addressing VSR while predicting auditory and visual representations. The researchers demonstrated the effectiveness of VSR model in various other languages such as Spanish, Mandarin, Italian, French, and Portuguese. Importantly, they highlighted the model's enhanced performance with the expansion of training sets, incorporating unlabeled data consisting of automatically generated transcriptions or movies in multiple languages. This finding suggests that recent advancements in the field may be attributed to the increase in dataset volumes rather than inherent improvements in models. Additionally, the researchers delved into the challenges faced by VSR systems and proposed solutions. They actively engaged in a meaningful discourse on ethical issues associated with such systems. This extensive research provides a comprehensive understanding of the landscape, offering valuable insights for the ongoing dialogue in the field.

Jie Shen, in their work [10], proposed a comprehensive investigation aimed at enhancing the efficacy of isolated word lip-reading. They emphasized the need for integrating various training methodologies and temporal models, as previous publications have primarily focused on individual approaches without exploring the potential synergies between different tactics and their respective impacts. In their study, Shen systematically evaluated multiple strategies for improving lip-reading performance. The findings indicated that Time Masking emerged as the most crucial augmentation technique, while mixup and DC-TCN stood out as the most effective temporal models. Combining all these methods resulted in an impressive 93.4% accuracy. Furthermore, there is potential for further enhancement by

incorporating pre-training on additional databases. This approach not only preserves the original order and structure of the information but also ensures the elimination of plagiarism.

Peiyu Chen, Yichen Gong, Helong Zhou[11] have addressed the challenge of training Transformer-based models for audio-visual speech recognition (AVSR), a task that typically demands a significant amount of labeled data. Recognizing the costliness of acquiring aligned and annotated multimodal data, the authors propose an innovative strategy. Their approach involves utilizing unimodal self-supervised learning to enhance multimodal AVSR. The authors specifically train independent audio and visual front-ends on extensive unimodal datasets. Subsequently, they integrate components from both front-ends into a comprehensive multimodal framework. This multimodal system is designed to learn and recognize concurrent audio-visual input by combining CTC (Connectionist Temporal Classification) and seq2seq decoding algorithms. The integration of components acquired through unimodal self-supervised learning is highlighted as a crucial strategy. The outcomes of their experiments demonstrate the effectiveness of this methodology. Remarkably, even without the need for an external language model, their proposed model achieves state-of-the-art performance on both word-level and sentence-level challenges. Notably, their model significantly surpasses previous techniques on the Lip Reading Sentences 2 (LRS2) dataset, showing a relative improvement of 30%.

Two important new findings are presented by Karan Shrestha[12] in their research. The first step in Chung and Zisserman automated system pipeline for lip-reading is to collect video data from TV broadcasts. Secondly, the model is trained utilizing the obtained dataset by means of a CNN architecture, as mentioned in the article. Using sequences of lip movements, the CNN is trained to identify individual words. Incorporating temporal fusion architectures, the article evaluates the efficacy of the CNN implementation and delves into several data input approaches. In this regard, the article presents the elasticity-3 (Early Fusion with 3D Convolution) design that I have included in my project. As an added bonus, the research dives into temporal sequence analysis using models like RNNs and Hidden Markov Models. Unfortunately, the accuracy of forecasts is affected by these models' shortcomings when it comes to responding to visual motion. Conversely, CNN accuracy while dealing with shifting themes is a matter of contention. A top-1 accuracy of 65.4% is shown by the CNN model that was used in the research. This means that the model successfully predicts the word label 65.4% of the time.

Navin Kumar Mudaliar, Kavita Hegde, Anand Ramesh, [13] present an innovative approach to visual voice recognition, emphasizing breakthroughs in machine learning enabling robots to comprehend human speech through visual cues. They address a significant limitation in traditional lip-reading, where professionals can only deduce 3-4% of spoken words. The authors propose a solution leveraging advanced deep learning algorithms and detail their methodology, employing a ResNet architecture with 3D convolution layers as the encoder and GRU as the decoder. Their method utilizes the entire video sequence for word-level classification, achieving an impressive 90% accuracy on the BBC dataset and 88% on a custom video dataset. Additionally, the authors suggest extending their technique to include short phrases or sentences, highlighting its potential for broader applications in visual speech detection.

Chris Payyappilly, Edwin J.C[14] proposed a method for lip reading in their research, focusing on the analysis of lip form and movement to identify and predict speech patterns, ultimately converting spoken language into text. The authors highlight the significance of lip reading, particularly for individuals with hearing challenges or in situations where auditory information is unavailable, providing an alternative means of understanding spoken language without the need to learn a new language. The researchers advocate for the use of computerized lip reading services, which rely on image processing techniques for detection and classification, finding applications across various sectors. The paper acknowledges challenges associated with lip reading, such as coarticulation and homophones. To address these challenges, the authors recommend the adoption of deep learning methods, like LSTM networks in combination with facial feature extraction to enhance the lip reading process. Additionally, the article emphasizes the importance of incorporating color imaging along with depth sensing to improve the accuracy of the lip reading classifier. Moreover, the researchers propose the integration of a Facial Expression Recognition algorithm capable of distinguishing facial expressions. This approach aids in recognizing specific facial features and monitoring their movements. These combined techniques are presented

as potential solutions to enhance the effectiveness of computerized lip reading systems and improve their performance in identifying and interpreting speech based on lip movements.

Brais Martinez, Pingchuan Ma, Stavros Petridis [15] have recommended several modifications and enhancements to the current state-of-the-art model for the identification of solitary words in real-world scenarios. These proposed changes involve replacing BGRU layers with Temporal Convolutional Networks (TCN) to improve the model's architecture. Additionally, they suggest simplifying the training process to facilitate one-stage training and addressing the issue of the current model's poor generalization to variations in sequence length by introducing variable-length augmentation. The authors present their findings on two extensive datasets, LRW and LRW1000, for word recognition in English and Mandarin respectively. They demonstrate that the proposed model yields an absolute improvement of 1.2% and 3.2% on these datasets, establishing it as the new state-of-the-art performance in this domain.

In another scholarly work by Peratham Wiriyathammabhum [16], a groundbreaking deep learning architecture is proposed for word-level lip reading. The methodology involves the implementation of an innovative deep neural network named SpotFast, derived from the Slow-Fast networks conventionally employed in action identification. The SpotFast architecture incorporates a temporal window designated as the "spot pathway," alongside utilizing all frames as the "fast pathway." Furthermore, the model incorporates memory-augmented lateral transformers to capture sequential information for effective categorization. The efficacy of this model is evaluated using the LRW dataset, and the results demonstrate its superiority over several state-of-the-art models. Notably, the inclusion of memory-augmented lateral transformers contributes to a significant 3.7% enhancement in the performance of the SpotFast networks.

A novel self-supervised representation learning framework named Audio-Visual Hidden Unit BERT (AV-HuBERT) is proposed by Shi et al.[17], aiming at enhancing audio-visual speech processing. AV-HuBERT leverages video recordings of speech to extract meaningful representations from both lip movements and corresponding audio signals. The methodology involves masking multi-stream video inputs and predicting hidden units in a multimodal manner, leading to exceptional results demonstrated on a comprehensive lip-reading benchmark. Notably, AV-HuBERT surpasses the previous state-of-the-art approach with only 30 hours of labeled data, a remarkable improvement considering the vast reduction in required training data. Moreover, when integrated with self-training, AV-HuBERT significantly reduces the Word Error Rate in lip-reading. The benefits of the learned audio-visual representations extend to audio-only speech recognition, showcasing a remarkable 40% relative decrease in WER compared to the current state-of-the-art performance. This innovative approach not only outperforms traditional methodologies but also demonstrates efficiency in resource utilization, making it a promising advancement in the field of audio-visual speech processing.

In their work, Yannis M Asseal, Shimon Whiteson, Nando de Freitas, and Brendan Shillingford [18] introduce a novel approach to address the challenge of lipreading, a process involving the interpretation of text by observing the movements of a speaker's lips. Traditional lipreading methods followed a two-stage procedure, wherein visual characteristics were either designed or learned first, followed by prediction generation. However, contemporary deep lip reading techniques, exemplified by Wand et al. (2016) and Chung & Zisserman (2016a), have been designed to be end-to-end trainable, streamlining the entire process. A significant contribution comes in the form of "LipNet," a model specifically designed to translate a dynamic sequence of video frames into text, enabling sentence-level sequence prediction. LipNet achieves this through the utilization of spatiotemporal convolutions and temporal classification loss. Notably, the model is trained in an end-to-end fashion, concurrently learning spatiotemporal visual characteristics and a sequence model. LipNet stands out as the first end-to-end sentence-level lipreading model to possess this dual capability, representing a substantial advancement in the field. The research yields impressive results, demonstrating that LipNet attains a remarkable accuracy of 95.2% in sentence-level lipreading on the GRID dataset. This performance exceeds that of experienced human lip readers and marks a significant improvement over the models, which were limited to word-level categorization with an accuracy of 86.4%. This breakthrough signifies a substantial enhancement in lipreading technology by enabling more precise sentence-level predictions through end-to-end training.

In an alternative study, Themos Stafylakis and Georgios Tzimiropoulos propose a distinctive deep learning architecture for word-level voice recognition [19]. Their model integrates spatiotemporal convolutional networks, residual networks, and LSTM networks. The evaluation is conducted using the challenging Lipreading In-The-Wild benchmark, comprising 1.28-second video snippets from BBC TV broadcasts, with a focus on 500 target words. Notably, the suggested network achieves a word accuracy of 83.0%, demonstrating a significant absolute increase of 6.8% beyond current state-of-the-art methods. Remarkably, this improvement is achieved without relying on knowledge of word boundaries during testing or training.

Pingchuan Ma, Brais Martinez, Stavros Petridis [20] delve into the advancements in lipreading, particularly driven by the resurgence of neural networks. The authors highlight recent endeavors aimed at enhancing lip reading performance through architectural modifications and generalization strategies. However, they underscore a substantial disparity between these methodologies and the practical requirements for implementing lipreading in real-world applications. To bridge this gap, the paper introduces several innovations. In the first phase, the researchers significantly elevate the state-of-the-art performance on LRW and LRW-1000 datasets, achieving accuracy rates of 88.5% and 46.6%, respectively, through the application of self-distillation. Moving on to architectural enhancements, the article suggests the incorporation of a DS-TCN head, which effectively reduces computational costs while maintaining efficiency. The second innovation involves the introduction of the DS-TCN head, contributing to a substantial reduction in computing costs while preserving efficiency. Lastly, the authors demonstrate the utility of information distillation in restoring efficiency for lightweight models, resulting in models with varying accuracies. Importantly, the most promising lightweight models achieve performance levels comparable to the existing models while concurrently reducing computational costs. These breakthroughs hold the potential to facilitate the practical deployment of lipreading models across diverse applications.

III. METHODOLOGY:

1. **Architecture Overview**:

The architectural design of the model involves employing a 3D CNN (C3D)-P3D network for visual feature extraction, a departure from the commonly utilized C3D plus 2D ResNet approach in lipreading papers. The C3D-P3D combination enhances the generation of potent visual spatio-temporal representations, with C3D effectively capturing such features in videos, and P3D strategically replacing specific C3D layers to address challenges.

For audio feature extraction, we utilize the Short Time Fourier Transform (STFT) method, decomposing audio signals into constituent frequencies over time. This results in a critical visual representation for understanding non-stationary signals like speech and music, where frequency components dynamically change over time. Subsequently, visual and audio features undergo separate processing through temporal convolution blocks (TCN or 1D ResNet).

The EleAtt-GRU encoder independently encodes the enhanced audio magnitude and visual features (revisited from the C3D-P3D network). It then decodes the audio and visual contexts separately, concatenating the decoded context vectors and directing them to a final decoder layer to produce character probabilities. Our investigation explores the impact of three distinct temporal models: Bidirectional Gated Recurrent Units (BGRUs), Multi-Scale Temporal Convolutional Networks (MS-TCNs), and Dense Connection Temporal Convolutional Networks (DC-TCNs).

To enhance generalization and prevent overfitting, we incorporate data augmentation techniques, including random cropping of the mouth Region of Interest (ROI), horizontal frame flipping, mixup (combining input sequences and targets), and time masking (inspired by

SpecAugment). These techniques collectively contribute to an increased overall dataset size. Additionally, we integrate two supplementary techniques: a word boundary indicator, represented by a binary vector matching the length of video frames, the target word is set to 1 or 0, and a self-distillation technique using a teacher model that provides "soft targets" (probabilistic predictions) alongside conventional "hard targets" (one-hot encoded ground truth labels). These soft targets capture richer information about the data distribution, guiding the student model to learn more effective representations.

2. **Data Augmentation**:

Random Cropping: Random cropping includes taking random patches of varied sizes and places from an image or video frame. Specifically, during training, an 88x88 patch is randomly clipped for the mouth Region of Interest (ROI), ensuring it maintains within acceptable ranges of variability.

Flipping: Frames of a movie are randomly flipped with a chance of 0.5. This approach is typically used along with random cropping to add diversity. For activities requiring spatial connections, labels could need change after flipping to ensure consistency.
Mixup: Mixup involves picking two input data samples and their matching labels randomly. A random number λ, produced using a Beta distribution, is picked between 0 and 1. The two input samples are then mixed using the λ function, bringing extra variances within the training process.

Time Masking: Time masking involves randomly masking N consecutive frames, where N is sampled from a uniform distribution between 0 and Nmax. In this process, every masked frame is substituted with the mean frame of the sequence. This technique is inspired by SpecAugment, which is commonly employed in Automatic Speech Recognition (ASR) applications.

3. **Word Boundary Indicator**:

In accordance with established conventions, the temporal model is augmented with word boundary indicators as an additional input. These indicators are binary vectors of a length equal to the number of frames in the input video. Elements corresponding to frames featuring the target word are assigned a value of 1, while the remaining elements are set to 0. This vector is combined with the frame-wise visual data from the encoder and collectively inputted into the temporal model.

4. **Self-Distillation**:

Self-distillation is a method that entails the sequential training of models with akin architectures through a distillation process. The protocol begins with the training of a teacher network, followed by the training of a student model employing the similar architecture with guidance from the teacher network. This iterative process persists until no further improvements are discerned. The teacher network incorporates additional supervisory signals, integrating inter-class similarity information into the comprehensive loss function. This loss function is constituted by a weighted amalgamation of Cross-Entropy loss.

5. **Experimental Setup**:

-Datasets: The model employs the AV digits dataset, primarily applied in visual speech recognition. This dataset comprises recordings of individuals enunciating the digits ranging from 0 to 9, presented as brief phrases in English. These recordings encompass three perspectives, collected from 53 participants. The dataset furnishes a substantial collection, featuring more than 500 videos and 75,000 verbal expressions.

-Pre-Processing: In the initial phase of pre-processing, The audio undergoes STFT to generate the magnitude spectrogram from the waveform. Concurrently, the Mel-Scale Filter calculates mel-scale magnitude features using 80 mel-frequency bins within the range of 0 to 8kHz, emphasizing frequencies pertinent to human auditory perception. Visual pre-processing involves cropping the image with a focus on the mouth region as the Region of Interest (ROI). Subsequently, the C3D-P3D Network is employed to extract visual characteristics, capturing
spatio-temporal information. The 3D CNN (C3D) layer assimilates spatial and temporal features, and a combination of audio and visual features is performed along the channel dimension, yielding a multi-modal representation. This comprehensive pre-processing methodology lays the foundation for subsequent analysis and interpretation in the research paper intended for publication.

-Training Details: The initial stage involves the pre-training of the visual front-end (C3D-P3D) through a word-level classification network designed for lip reading. This is achieved by utilizing the Avdigits dataset, which encompasses 500 word classes. Input for the visual front-end consists of image frames. The back-end consists of one dense layer, and the training for word classification tasks employs cross-entropy loss.Both original and mixed-noise audio data are employed to maintain video characteristics and magnitude spectrograms. The dense layer in the back-end is replaced with two Bi-LSTM layers, succeeded by linear and SoftMax layers. The training process utilizes the Adam W optimizer and incorporates a cosine annealing learning rate strategy. This approach is intended for a research paper publication, ensuring adherence to proper academic standards and avoiding plagiarism.

-Temporal Models:The model incorporates a variety of temporal models, including MS-TCN, to effectively capture temporal information across different scales. This integration enhances the model's ability to represent long-range dependencies, surpassing the capabilities of traditional CNNs. DC-TCN utilizes dilated causal convolutions, where the dilation factor expands the receptive field without increasing the kernel size. This design choice enables the model to capture extended long-range temporal dependencies while maintaining causality. In comparison, MS-TCN builds upon the vanilla TCN by introducing multiple branches, each with distinct kernel size. The output features from each branch are then concatenated to seamlessly amalgamate information across various temporal scales. This sophisticated approach allows the model to potentially handle lengthier input sequences, presenting an advancement over conventional CNNs. This innovation is crucial for accommodating the complexities of temporal data and holds significant promise for applications requiring the analysis of extended temporal patterns.

IV. RESEARCH GAP:

Numerous studies in the realm of lipreading concentrate on the extraction of words or phonemes from lip information, typically derived from audio signals. However, the applicability of such models is hindered by the absence of real-world conditions, such as background noise, occlusions, and variations in lighting, limiting their generalizability. Addressing this limitation, the proposed model emphasizes a multimodal approach by integrating both audio and visual cues.

This innovative fusion not only yields significantly higher accuracy compared to single modality approaches but also proves particularly effective in challenging environments characterized by noise or ambiguous visual information. The model capitalizes on the complementary nature of visual and auditory cues to enhance robust speech recognition, employing a combination of different Convolutional Neural Networks (CNNs).

One major hurdle frequently encountered in research papers is the arduous and resource-intensive task of collecting and annotating large, diverse datasets for lip reading. The prevailing focus on isolated words or small vocabularies in existing datasets poses a challenge, impeding the performance of models in real-world conversations with diverse vocabulary and varying sentence structures. In contrast to many approaches utilizing Long Short-Term Memory (LSTM) networks, the proposed model opts for Temporal Convolutional Networks (TCNs). Unlike LSTMs, TCNs leverage convolutional layers, enabling parallel computation across the entire sequence simultaneously, thereby significantly accelerating both training and inference speed and enhancing the model's ability to capture long-range dependencies

V. CONCLUSION :

In conclusion, this work systematically explored the landscape of lip-reading, offering insights gleaned from an exhaustive study on the LRW dataset. Through rigorous examination of data augmentation and temporal models, we demonstrated a nuanced strategy for achieving state-of-the-art performance. Time Masking emerged as pivotal, closely trailed by mixup, underlining their significance in enhancing model robustness. Noteworthy, DC-TCNs exhibited superior

performance compared to MS-TCNs and BGRUs, providing a principled choice for temporal modeling. The integration of self-distillation and word boundary indicators substantially augmented classification accuracy, complemented by a discernible improvement through pre-training. Crucially, our models exhibited a marked enhancement in classifying challenging words. This systematic inquiry furnishes valuable insights, contributing to the refinement and sophistication of contemporary lip-reading paradigms.

VI. REFERENCES:

[1] J. S. Chung and A. Zisserman, "Lip Reading in the Wild," in Proceedings of the 13th Asian Conference on Computer Vision (ACCV), vol. 10112, 2016, pp. 87–103.

[2] T. Stafylakis and G. Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading," in
Proceedings of the 18th Annual Conference of International Speech Communication Association (INTERSPEECH), vol. 9, 2017, pp. 3652–3656.

[3] B. Martinez, P. Ma, S. Petridis, et al., "Lipreading Using Temporal Convolutional Networks," in Proceedings of the 45th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6319–6323.

[4] D. Feng, S. Yang, S. Shan, et al., "Learn an Effective Lip Reading Model without Pains," CoRR,
vol. abs/2011.07557, 2020.

[5] P. Ma, B. Martinez, S. Petridis, et al., "Towards Practical Lipreading with Distilled and Efficient Models," in Proceedings of the 46th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 7608–7612.

[6] Y. Zhang, S. Yang, J. Xiao, et al., "Can We Read Speech Beyond the Lips? Rethinking RoI Selection for Deep Visual Speech Recognition," in Proceedings of the 15th IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2020.

[7] T. Stafylakis, M. H. Khan, and G. Tzimiropoulos, "Pushing the Boundaries of Audiovisual Word Recognition using Residual Networks and LSTMs," Computer Vision and Image Understanding, vol. 176–177, pp. 22–32, 2018.

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141.

[9] S. Petridis, T. Stafylakis, P. Ma, et al., "End-to-end audiovisual speech recognition," in Proceedings of the 43rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 6548–6552.

[10] P. Ma, Y. Wang, J. Shen, et al., "Lip-reading with Densely Connected Temporal Convolutional Networks," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2857–2866.

[11] D. S. Park, W. Chan, Y. Zhang, et al., "SpecAugment: A Simple Data Augmentation Method for AutomaticSpeech Recognition," in Proceedings of the 20th Annual Conference of International Speech Communication Association (INTERSPEECH), 2019, pp. 2613–2617.

[12] P. Ma, R. Mira, S. Petridis, et al., "LiRA: LearningVisual Speech Representations from Audio Through Self-Supervision," in Proceedings of the 22nd Annual Conference of International Speech Communication Association (INTERSPEECH), 2021, pp. 3011–3015.

[13] T. Furlanello, Z. C. Lipton, M. Tschannen, et al.,"Born-again neural networks," in Proceedings of the 35th International Conference on Machine Learning (ICML), 2018, pp. 1602–1611.

[14] J. Deng, J. Guo, E. Ververas, et al., "RetinaFace: Single-Shot Multi-Level Face Localisation in the
Wild," in Proceedings of the 33rd IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2020, pp. 5203–5212.

[15] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D Face

Alignment problem? (and a dataset of 230,000 3D facial landmarks),” in Proceedings of the 16th IEEE/CVF International Conference on Computer Vision (ICCV), 2017.

[16] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” in Proceedings of the 7th International Conference on Learning Representations
(ICLR), 2019.

[17] T. Afouras, J. S. Chung, and A. Zisserman, “LRS3-TED: a large-scale dataset for visual speech recognition,” CoRR, vol. abs/1809.00496, 2018.

[18] J. S. Chung, A. Senior, O. Vinyals, et al., “Lip reading sentences in the wild,” in Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 3444–3453.

[19] A. Ephrat, I. Mosseri, O. Lang, et al., “Looking to listen at the cocktail party: speaker-independent audiovisual model for speech separation,” ACM Transactions on Graphics, vol. 37, no. 4, 112:1–112:11, 2018.

[20] P. Ma, S. Petridis, and M. Pantic, “Visual speech recognition for multiple languages in the wild,” CoRR, vol. abs/2202.13084, 202