2024 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE) | 979-8-3503-7156-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/AMATHE61652.2024.10582171

Enhancing Robustness in Audio Visual Speech Recognition: A preprocessing approach with Transformer and CTC Loss

Bharath N V Ithal *
Dept. of Computer Science
PES University
Bengaluru, India
https://orcid.org/0009-0003-0010-5312

Lagan T G *
Dept. of Computer Science
PES University
Bengaluru, India
https://orcid.org/0009-0008-5975-2457

Rhea Sudheer *
Dept. of Computer Science
PES University
Bengaluru, India
https://orcid.org/0009-0002-0745-1448

Swathi Rupali N V *
Dept. of Computer Science
PES University
Bengaluru, India
https://orcid.org/0009-0007-5553-667X

Dr. Mamatha H R
Dept. of Computer Science
PES University
Bengaluru, India
https://orcid.org/0000-0002-5409-1329

Abstract—**Audio Visual Speech Recognition (AVSR) is a promising technology for speech recognition that is more robust to noise and other challenging conditions than traditional Audio Speech Recognition (ASR). AVSR systems work by fusing audio and visual features using machine learning algorithms. One of the key challenges in AVSR is developing effective fusion methods. AVSR systems are currently being used in many different applications, having made great strides in recent years despite these obstacles. In this paper, we introduce a preprocessing technique applied on a Transformer model with Connectionist Temporal Classification (CTC) loss that increases the robustness of the AVSR model against noise. We preprocess the training audio clips by adding noise to it, resulting in significant reductions in word error rates and a more efficient training process. The project significantly advances in the field of audio-visual speech recognition and highlights the potential for practical applications, particularly in noisy and challenging environments.**

Index Terms—**AVSR, TM-CTC, preprocessing, Transformers, CTC Loss, LRS2 Dataset**

I. Introduction

In the evolving landscape of human-computer interaction, an increasing amount of research is being done on the fusion of audio and visual data. Audio-Visual Speech Recognition (AVSR) stands at the crossroads of computer vision, speech processing, and machine learning, promising a revolutionary way to enhance the capabilities of automated systems, from voice assistants to lip-reading technology and more. AVSR, as a technology, holds great potential for bridging gaps in communication and human-computer interaction by leveraging both auditory and visual cues to transcribe spoken language.

The main challenge faced by AVSR is deciphering both the acoustic aspects of speech simultaneously. Automatic Speech Recognition (ASR) systems rely on audio data but they still struggle in noisy environments or when speech is not clear. On the other hand visual information such as lip movements, facial expressions and gestures can provide context to clarify speech ambiguity. Hence AVSR aims to create a synergy between these two modalities in an effort to raise the precision and dependability of voice recognition systems.

The significance of AVSR goes beyond its role in enhancing speech recognition technology .This technology has implications across domains, such as accessibility, surveillance, human robot interaction and more. Being able to understand language through cues even in difficult audio conditions can remove obstacles for people with hearing impairments. It can also enhance security systems and enable natural interactions between humans and computers.

In the subsequent sections of this paper, we will delve into the foundations of AVSR, the methodologies and technologies involved, and the current state of research in this domain. Furthermore, we will explore potential applications and the future prospects of AVSR, emphasizing the transformative potential it holds in redefining how we interact with machines and each other through enhanced audio-visual communication. We shall explore AVSR's foundations, associated techniques, technologies, and the state of research in this field in the sections that follow. We will also discuss future directions and possible uses for AVSR, highlighting its revolutionary potential to redefine human-machine and interpersonal interactions through improved audio-visual communication.

II. Related works

In an effort to better understand how to represent and merge the audio-visual modalities, the majority of AVSR research till date has concentrated on innovative architectures and supervised learning techniques. A Transformer-based [1] AVSR system with sequence-to-sequence loss is proposed by TM-seq2seq [2]. A hybrid seq2seq/CTC loss [3] AVSR system based on RNNs is proposed by Hyb-RNN [4]. For the AVSR task, RNN-T [5] uses a recurrent neural network transducer [6]. An RNN-based multimodal speech recognition and audio enhancement system is constructed by EG-seq2seq [7]. LF-MMI TDNN [8] proposes a hybrid audio-visual speech separation and recognition system based on TDNN. Using hybrid seq2seq/CTC loss, the audio-visual streams are independently encoded and then concatenated for decoding in the Hyb-Conformer's [9] Conformer-based [10] AVSR system. The LRS3 and LRS2 datasets have both yielded the super-vised learning SOTA for this system. MoCo+wav2vec [11] enhances AVSR performance with self-supervised pre-trained audio/visual front-ends it has achieved the SOTA on the LRS2 dataset. However, without specific interactions to capture their deep correlations, many investigations only concatenate the auditory and visual characteristics for multimodal fusion. The latest iteration of u-HuBERT [12] builds upon the recently proposed AV-HuBERT [13] by capturing contextual correlations between audio-visual data through self-supervised learning. The SOTA has been achieved on the LRS3 dataset by using this combined multimodal and unimodal pre-training framework.

Speech recognition for audiovisual media ”in the wild” [14] [15] refers to unfiltered speech in the open environment. Sutskever et al. [16] employed neural networks to solve a sequence-to-sequence problem for the first time in accordance with Bahdanau et al. [17] and Luong et al. [18]. Enhancements were created using recently developed attention mechanisms. Vaswani et al. [19] developed a transformer network to find global relationships between inputs and outputs based on an attention technique. Learning several translating strategies improves overall performance, particularly for low-resource languages, claims Johnson et al. [20]. Deep networks have been used by recurrent neural networks (RNNs); these were first reported in [21] [22]. While these designs are subject-dependent and yield natural output, they require rebuilding and retraining techniques in order to adjust to new faces.

[23] employs CNNs to convert audio data into a three-dimensional mesh of particular speakers. The CNN method estimates mesh points in three dimensions and expresses dynamics using comment threads. A CNN based on Mel-frequency Cepstral coefficients (MFCCs), developed by Chung et al. [24], may generate subject-independent clips with a simple image and audio data. This technique includes an L1 loss on the image, which makes it hazy and necessitates further deblurring. Besides these, deviation from the training clip is discouraged by pixel loss, which keeps the system from evoking real emotions and results in basically frozen faces except for the lips.

This paper [25] proposes a method to improve speech recognition accuracy by improving the accuracy of lip reading in noisy environments. Using a one-to-many mapping relationship model between lips and speech, the authors first enable the lip reading model to take into account the articulations that are represented by the input lip movements. Subsequently, they create an accurate visual feature extraction lip-to-audio one-to-many mapping model as the encoding component of the lip recognition model. For the fusion of cross-modal features, they also create a joint cross-fusion model with multiple attention weights that are computed based on the attention mechanism, which is effective in obtaining correlations between various modalities and also within the current state.

Two different goal functions for their closely related models are combined to jointly train a two-stage network and optimise a new multi-objective loss function. Various techniques are employed in the field of speech and sound research, such as speech separation and dereverberation [26] [27], denoising and dereverberation [28] [29], noise suppression and sound event detection [30], and voice activity detection [31]. These studies combine a noise-cancelling model for input audio in the first stage with a self-implementing model in the second stage to achieve stable performance even under noisy environments. Merged an ASR model with an acoustic echo cancellation (AEC) model [34]. Since their goal was to recognise input speech, they considered how to improve the accuracy of recognition when improved speech was given to an ASR model. They trained the AEC model to improve the pre-trained ASR model's accuracy in speech recognition, which led to the model's resilience against echo-induced distortion. [7] combined the AVSE and AVSR models. They first used the AVSE model to process visual cues like the speaker's lip movements in order to separate target speech from background noise. The AVSR model was then trained to utilise both the visual and augmented audio information to increase the recognition accuracy. [35] [36] integrated an ASR model with an active speaker detection (ASD) model. [35] combined and trained the models using multi-task loss, resulting in an improvement in ASR performance and ASD accuracy. [36] attached the visual context attention model to address the label ambiguity for multi-talker modelling. Although [34] constructed two-stage models to enhance speech recognition performance, they achieved this by modifying pre-trained ASR and AVSE models, respectively. Unlike [34] [4], they offer a jointly trained model that uses a novel loss function to optimise the AVSE and AVSR models by combining two different goal functions. An AVSR method for overlapped speech with interfering speakers in the real world was proposed by Yu et al. [8]. Their approach used time-delay neural networks (TDNNs) equipped with a discriminative criterion for lattice-free MMI (LF–MMI) (referred to as an LF–MMI TDNN system). Two AVSR architectures were put forth by the authors: an end-to-end architecture and a hybrid architecture. In their experiments using the LRS2 dataset, they discovered that the hybrid AVSR architecture performed better than the end-to-end AVSR architecture. The authors demonstrated how their proposed AVSR architecture can have a word-error rate (WER) as low as about 29% less than a baseline ASR architecture that solely uses audio. Afouras et al.'s [2] strategy for AVSR in the wild makes use of transformer-based models and deep learning methods. The architectures for AVSR and two transformer (TM) models proposed by the authors were an encoder-decoder-attention-structure TM architecture and a self-attention TM-stack architecture. By stacking multiple self-attention and feedforward layer stacks, the self-attention transformer stack—also referred to as the TM–CTC architecture—produced the CTC loss's posterior probabilities.The experimental results on the LRS2-BBC dataset demonstrated that the suggested TM AVSR architectures could provide a lower WER of 8.2% in comparison to the WER of 10.1% provided by an audio-only baseline AVSR architecture. An alternative method for AVSR field research is provided by Son Chung et al. [37]. The author's AVSR architecture, also referred to WLAS network model is described in this work. It has the ability to describe speech transcription. It is possible to configure attention models in their WLAS architecture for audio input, visual input, and audio-visual input. Their experimental results showed that their proposed AVSR architecture produced a lesser WER of 23.8% and 3.0% for the Lip-Reading in the Wild (LRW) and GRID datasets, respectively, compared to alternative methods.

The goal of the methodology presented in the paper [38] is to use multimodal sensor-input architecture and deep learning to improve natural language audio and video stream's audio-visual speech recognition (AVSR) performance. Fusion-based AVSR approaches, Speech Recognition (SR) structure by speech modality, and SR structure by facial modality are the three main sections of the methodology in [2]. The authors propose employing an LSTM network for sequence modelling subsequent to the use of a CNN for the purpose of extracting characteristics from speech audio signals [2]. A CNN is used in the SR architecture by facial modality to extract features from facial images, and an LSTM network is used for sequence modelling [2]. The authors suggest using data augmentation techniques, such as random image cropping and flipping and adding noise to audio signals, to enhance the performance of the AVSR models. By using these methods, the training data can be made larger and more diverse, which can strengthen the models' ability to withstand changes in the input signals. Using a multimodal fusion network, the fusion-based AVSR technique combines the outputs of the facial and speech modalities. The model can selectively pay attention to various modalities depending on their relevance to the task by using a gated fusion mechanism, as suggested by the authors. The LRS2 dataset, which includes natural language audio and video streams in unrestricted settings, is used to assess the suggested methodology [8]. The authors show that their fusion-based approach outperforms various other models in terms of word error rate by comparing the performance of their models to several cutting-edge AVSR models.

To enhance the performance of AVSR in natural language audio and video streams, they present a methodology that leverages deep learning and multimodal sensor-input architecture. The methodology comprises data augmentation techniques, fusion-based AVSR approaches with a gated fusion mechanism, and SR architecture by speech and facial modalities. The LRS2 dataset is used to assess the methodology, and the findings show improved accuracy when compared to the most recent models [8] [2].

III. Datasets

In this study, we investigate a carefully chosen range of notable AVSR datasets, each presenting a different mix of difficulties, possibilities, and real-world uses. These datasets are vital resources that support the development of AVSR technology and the accomplishment of numerous objectives, ranging from word recognition to full sentence comprehension and beyond.
Outlining the several benchmark datasets in brief:

A. GRID Corpus

Description: The grid corpus dataset is an Audio-visual dataset which is vividly used in the field of Speech perception and generation. It consists of around 1000 facial recordings. The dataset also has annotations present along the video.
Key Features: This dataset is widely used for building and evaluating AVSR systems due to its high-quality recordings and diverse speakers.

B. LRW (Lip Reading in the Wild)

Description: LRW is a well-known dataset created for lip-reading tasks involving a large vocabulary. It has over five hundred video clips showing people speaking words and sentences in a controlled setting.
Key Features: LRW is a vital tool for developing and testing deep learning models for visual speech recognition because of its excellent recordings and distinct articulation.

C. LRW-1000

Description: The LRW-1000 dataset is a notably larger and more comprehensive version of the LRW dataset, with a higher number of speakers and vocabulary. Spoken by hundreds of speakers, it includes a thousand distinct words.
Key Features: This dataset can be used for a variety of difficult and demanding lip-reading tasks because it has a larger vocabulary and a more diverse range of speakers.

D. LRS (Lip Reading Sentences)

Description: The LRS dataset is a collection of audio and visual recordings at the sentence level. It was one of the first datasets made with the intention of recognising visual speech at the sentence level, whereas the earlier datasets were primarily concerned with word recognition.
Key Features: LRS gives researchers in the AVSR field basic data that they can use to build and evaluate lip-reading models.

E. LRS2 (Lip Reading Sentences 2)

Description: The creators of LRS2-BBC gathered thousands of hours of exchanges between speakers of words and phrases in addition to the accompanying facetracks from a variety of BBC shows.
Key Features: LRS2 offers a more realistic portrayal of

real-world scenarios and lip-reading challenges because of its diverse and unconstrained data.

F. LRS3 (Lip Reading Sentences 3)

Description: A multi-modal dataset for audio-visual and visual speech recognition. It has word alignment boundaries and matching subtitles for over 400 hours of TED and TEDx videos, in addition to face tracks from those videos.
Key Features: LRS3 is the largest available dataset for AVSR, setting it as the current benchmark for AVSR.

TABLE I
DATASET INFORMATION

Dataset	Source	Utterances	Language	No. hrs	Speaker Info
GRID	—	33,000	English	27.5	51
LRW-1000	Broadcast News	718k	Chinese	—	2,000
LRS-2	BBC News Program	118k	English	246h	1,000+
LRS-3	Ted Talks	165k	English	475h	1,000+

Our selection of the LRS-2 dataset for our model is motivated by its relatively underexplored nature in comparison to the extensively studied LRS-3 dataset in existing research. A comprehensive literature survey revealed a predominant trend where the majority of models demonstrated superior performance on the LRS-3 dataset as opposed to the LRS-2 dataset. This observation underscores the potential for further investigation and exploration of the unique challenges and characteristics presented by the LRS-2 dataset in the context of our research objectives.

IV. PRE-PROCESSING

Preprocessing: Gathering Data Files: A list of video files (.mp4) in the dataset is generated by traversing the data directory. This list serves as the basis for preprocessing each video sample.

Preprocessing Each Sample: For each video sample, the preprocessing script executes the following tasks:

A. Extracting Audio

The audio track is extracted from the video file using the FFmpeg utility and saved as a separate WAV file. This step separates the audio content from the video for subsequent audio processing.

B. Resizing and Cropping Frames

Frames from the video are processed individually. Each frame is converted to grayscale, resized to 224x224 pixels, and then cropped to retain the central region of 112x112 pixels. This operation produces a sequence of Region of Interest (ROI) frames.

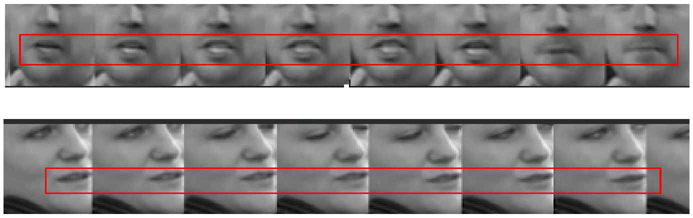

Fig. 1. Resized and cropped video frames post pre-processing

C. Normalizing Frames and Extracting Visual Features

By deducting the mean (normMean) and dividing by the standard deviation (normStd), the grayscale ROI frames are normalised. Subsequently, these normalized frames are input into the Visual Frontend model (vf) to extract visual features. The resulting features are saved in a .npy file. These visual features contain critical visual information for audio-visual speech recognition.

D. Generating a Noise File

To facilitate background noise modeling during audio processing, a 1-hour noise file is generated. This noise file is created by aggregating audio samples from 20 random video files. The length of these audio clips matches the duration of the shortest audio sample among the 20 randomly selected files.

E. Generating preval.txt for Pretrain Set Split

A preval.txt file is created to manage the split between the pre training set for the audio-visual model. This split ensures a clear division between the training and validation sets, which is a vital aspect of training the audio-visual speech recognition model.

V. ARCHITECTURE

A. Audio features

We use spectral magnitudes in 321 dimensions that we extracted from audio data. These magnitudes are computed with a 40-millisecond window and a 10-millisecond gap, at a sample rate of 16 kHz. The audio data is matched with the video frames, which are recorded at a rate of 25 frames per second(or a 40-millisecond interval per frame), by mapping each video frame to four acoustic feature frames. This means that, as is typical in stable CTC training, we cluster the audio features into sets of four. This shortens the input sequence length and guarantees that the audio and video data have the same temporal scale for efficient joint processing.

B. Visual frontend

A vital part of the visual processing pipeline, the spatio-temporal visual front-end is built to extract temporal as well as spatial information from the input data. The architecture is broken down into a number of layers that are optimized for efficient feature extraction. First, a 3D CNN layer is used, which uses 64 filters with 5 × 7 × 7 dimensions and a stride pattern of [1, 2, 2]. The temporal (T), height (H), and width

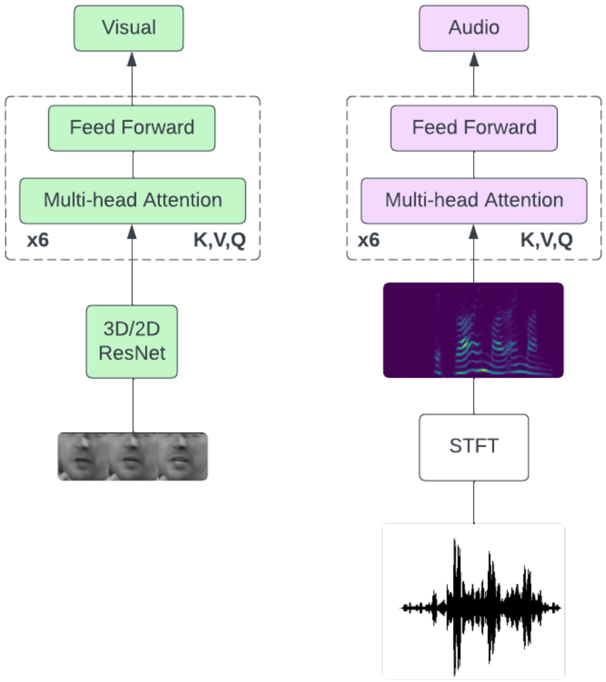

Fig. 2. Common encoder

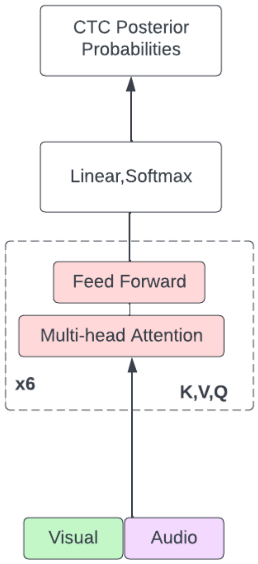

Fig. 3. TM-CTC

(W) dimensions are successfully downsampled, but the number of channels is increased to 64. The dimensions are then further reduced to four times their initial values by preserving the 64 channels in a 3D Max Pooling layer with a stride of [1, 2, 2].

C. Common encoder

An integral part of AVSR systems, the encoder extracts key characteristics from each modality by processing audio and video inputs independently. While CNNs are used to process the video input, short-time Fourier transform (STFT) is Mel-frequency cepstral coefficients (MFCCs) are obtained by applying the short-time Fourier transform (STFT) to the audio input. The AVSR model training and recognition process is based on the fusion of these collected features into a single feature vector, usually through the use of a concatenation layer. Better accuracy and simplicity of use are two benefits of the common encoder, however subtle interactions between audio and visual inputs might not be properly captured. In addition, the design makes use of multi-head self-attention layers to improve learning, and distinct encoders for every modality to guarantee thorough feature extraction. The architecture of the same is shown in figure 2.

D. Model Architecture

The TM-CTC model, designed for Audio-Visual Speech Recognition (AVSR), combines visual and aural cues to improve speech recognition precision. The model commences by extracting audio features, often utilizing Mel-Frequency Cepstral Coefficients (MFCCs) or spectrograms, and visual features derived from video frames, capturing lip movements and facial expressions. These features are subsequently embedded into a common feature space to align audio and visual modalities. The crux of the model resides in the Transformer encoder, comprising multiple layers equipped with self-attention mechanisms. This Transformer architecture efficiently captures temporal dependencies and contextual information in the combined audio-visual feature vector.

Following feature extraction and embedding, the model predicts a sequence of tokens, often phonemes or words, by calculating posterior probabilities for each token. To align the actual sequence of tokens with the one predicted, the Connectionist Temporal Classification (CTC) loss function is used. The CTC loss is essential in handling insertions, deletions, and substitutions that may occur during recognition.

The final output word sequence, embodying recognized spoken words, is generated using the CTC algorithm. Notably, the TM-CTC model's design facilitates a seamless integration of both audio and visual data, enhancing speech recognition accuracy in various scenarios, particularly those where audio-only information may lack essential context. Additionally, the model undergoes training and fine-tuning, where it learns to align audio and visual cues with the actual spoken words, ensuring optimal performance. This comprehensive architecture encapsulates the core components of the TM-CTC model, which offers a versatile solution for AVSR applications across diverse domains.

VI. Result and Discussion

In the evaluation of our model, the Word Error Rate (WER) emerged as a pivotal metric, providing insight into the accuracy of the transcribed content By taking into account the cumulative effects of insertions, deletions, and substitutions in relation to a reference transcript, WER determines the percentage of words that are erroneously recognized. The essence of this metric is expressed as in

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

is the total number of words in the reference transcript and I , D, and S are the insertions, deletions, and substitutions, respectively. WER enables a thorough evaluation of the system's ability to reliably translate spoken language into written text, providing insights into possible areas for enhancement and optimization.

Likewise, the Character Error Rate (CER) is a critical indicator in our assessment model that provides a detailed analysis of transcription accuracy down to the character level. CER handles substitutions, insertions, and deletions just like WER does, however it works with individual characters instead of words. The CER computation is expressed by the

$$CER = \frac{I + D + S}{C}$$

where C is the total number of characters in the reference transcript and I, D, and S are insertions, deletions, and substitutions, respectively. Through its emphasis on the complexities of character-level correctness, CER offers insightful information about certain nuances and difficulties in the transcription process, which helps us improve and fine-tune our audio-visual speech recognition system.

TABLE II
Word Error Rate (WER) for different modalities in clean and noisy conditions.

Type	Clean (Greedy)	Clean (Beam + ext LM)	Noisy (Greedy)	Noisy (Beam + ext LM)
AO	0.117	0.089	0.646	0.54
VO	0.616	0.53	0.616	0.53
AV	0.042	0.035	0.108	0.071

TABLE III
Character Error Rate (CER) for different modalities in clean and noisy conditions.

Type	Clean (Greedy)	Clean (Beam + ext LM)	Noisy (Greedy)	Noisy (Beam + ext LM)
AO	0.047	0.042	0.382	0.284
VO	0.358	0.223	0.358	0.223
AV	0.029	0.023	0.041	0.034

In this research investigation, we systematically evaluated the model performance of proposed AVSR model, employing two distinct search strategies: Greedy and Beam search, the latter augmented with an external language model (ext LM). The assessments were conducted across varied environmental conditions, encompassing both pristine and noisy settings, to comprehensively ascertain the model's adaptability and robustness. Across clean conditions, as detailed in Table 2, the Beam search with ext LM consistently outperformed the Greedy approach across all speech modalities—Audio-Only (AO), Visual-Only (VO), and Audio-Visual (AV). Noteworthy improvements were observed, with AO experiencing a substantial decrease in Word Error Rate (WER) from 0.117 to 0.089, VO exhibiting a significant improvement from 0.616 to 0.5555, and AV displaying a commendable reduction in WER from 0.042 to 0.035.

Under the challenging circumstances of noisy environments, as presented in Table 2, the Beam search with ext LM exhibited remarkable resilience, consistently surpassing the performance of the Greedy approach. This resilience was particularly pronounced in the VO and AV modalities, showcasing the model's robust capacity to discern and interpret audio-visual cues amidst noise. Furthermore, the comprehensive assessment of Character Error Rate (CER) presented in Table 3 reinforced the superiority of the Beam search with ext LM, substantiating its potential to enhance the precision and reliability of audio - visual speech recognition systems in real-world scenarios.

It is noteworthy that the Audio-Visual (AV) modality emerged as a standout performer, achieving the lowest WER and CER in both clean and noisy conditions, underscoring the model's exceptional capacity to seamlessly integrate audio and visual information for superior accuracy and efficacy. These findings, detailed in Tables 2 and 3, illuminate the significance of advanced search strategies, particularly in challenging acoustic environments, and highlight the adaptability of our model to diverse audio-visual inputs. In the examination of the presented results, as shown in Table 4, a nuanced understanding emerges regarding the Word Error Rate (WER) and Character Error Rate (CER) across diverse models in the domain of speech and visual processing. The ”Lip-subword correlation (LRW)” method, underscored by its striking CER of 0.2458, distinctly excels in character-level accuracy. In addition, the ”Cross-Modal Global Interaction and Local Alignment” and ”Visual Corruption Modelling and Reliability Scoring” approaches demonstrate remarkable effectiveness, as demonstrated by their excellent WER values of 0.1036 and 0.1336, respectively, demonstrating their competence in word-level recognition tasks.

TABLE IV
Comparitive analysis of proposed model with existing models

Method for Audio Visual	WER	CER
Lip-subword correlation (LRW)	-	0.2458
Cross-Modal Global Interaction and Local Alignment	0.1036	-
Visual Corruption Modeling and Reliability Scoring	0.1336	-
UniVPM (LRS-3)	0.267	-
Overlapped Speech (LRS-2)	0.049	-
Proposed model	0.071	0.034

In contrast, challenges become apparent with the ”UniVPM (LRS-3)” and ”Overlapped Speech (LRS-2)” models, as reflected in higher WER values of 0.267 and 0.049, respectively.

These instances suggest potential areas for improvement in the context of these specific models. On a positive note, the our proposed model emerges as a standout performer, showcasing a well-balanced performance with a WER of 0.071 and a relatively low CER of 0.034. This implies that the proposed model strikes an optimal equilibrium between word and character-level accuracy, positioning it as a promising candidate for further exploration and application in the nuanced landscape of audio-visual speech recognition tasks.

The proposed noise generation method introduced several novel features that enhanced its applicability in audio research and technology development. The method dynamically composed a 1-hour noise file by randomly selecting variable-length audio samples from different files, fostering adaptability to diverse acoustic environments. By efficiently utilizing dataset samples without duplication, the method ensured the richness required for effective model training, addressing scenarios with limited dataset sizes. Additionally, the adaptive adjustment of noise duration aligned with specific duration requirements, enhancing the method's flexibility. The incorporation of normalization and scaling steps underscored a commitment to signal compatibility, crucial for seamless integration into various audio-related tasks. Furthermore, the method's simulation of non-uniform background noise, achieved through the randomness in sample selection, mirrored the natural variability found in real-world acoustic environments. In summary, these innovative characteristics collectively positioned the method as a valuable tool for realistic background noise simulation in audio-related applications.

VII. Conclusion and future work

In conclusion, our research underscores the promising trajectory of Audio-Visual Speech Recognition (AVSR) technology, offering enhanced robustness in challenging conditions compared to traditional Automatic Speech Recognition systems. Fusion of audio and visual features through machine learning algorithms, while presenting challenges, has seen significant advancements. Our contribution to this domain involves the introduction of a preprocessing technique applied to a Transformer model with CTC loss, effectively enhancing the AVSR model's resilience against noise. By introducing controlled noise during the training phase, we observed substantial reductions in word error rates, contributing to a more efficient training process. This project represents a noteworthy stride forward in the realm of audio-visual speech recognition, demonstrating its potential for practical applications, particularly in noisy and demanding environments.

Moving forward, our goal is to improve audio-visual feature fusion techniques, enhancing our AVSR model's adaptability across diverse real-world scenarios. Expanding datasets, integrating advanced machine learning techniques, and exploring applications in assistive technologies are key priorities. Our commitment lies in further advancing AVSR, addressing evolving challenges, and extending its impact in practical domains.

VIII. Acknowledgement

This project and research represent a collaborative effort, and it is a privilege to have worked alongside numerous individuals whose contributions have been instrumental in its success. We would like to convey our gratitude to the authors of the referenced works, whose contributions formed the foundation of our research and allowed for a comprehensive comparison of various approaches.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin". Attention is all you need." Advances in neural information processing systems, 30, 2017.

[2] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. "Deep audio-visual speech recognition". IEEE transactions on pattern analysis and machine intelligence, 2018

[3] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. "Hybrid ctc/attention architecture for end-to-end speech recognition". IEEE Journal of Selected Topics in Signal Processing, 11(8):1240–1253, 2017.

[4] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. "Audio-visual speech recognition with a hybrid ctc/attention architecture". In 2018 IEEE Spoken Language Technology Workshop (SLT), pages 513–520. IEEE, 2018.

[5] Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. "Recurrent neural network transducer for audio-visual speech recognition". In 2019 IEEE automatic speech recognition and understanding workshop (ASRU), pages 905–912. IEEE, 2019.

[6] Alex Graves. "Sequence transduction with recurrent neural networks". arXiv preprint arXiv:1211.3711, 2012

[7] Bo Xu, Cheng Lu, Yandong Guo, and Jacob Wang. "Discriminative multi-modality speech recognition". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1443314442, 2020

[8] Jianwei Yu, Shi-Xiong Zhang, Jian Wu, Shahram Ghorbani, Bo Wu, Shiyin Kang, Shansong Liu, Xunying Liu, Helen Meng, and Dong Yu. "Audio-visual recognition of overlapped speech for the lrs2 dataset". In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6984–6988. IEEE, 2020.

[9] Pingchuan Ma, Stavros Petridis, and Maja Pantic. "End-to-end audio-visual speech recognition with conformers." In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7613–7617. IEEE, 2021

[10] Anmol Gulati, James Qin, Chiu Chung-Cheng, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. "Conformer: Convolution-augmented transformer for speech recognition". In Interspeech, pages 5036–5040, 2020.

[11] Xichen Pan, Peiyu Chen, Yichen Gong, Helong Zhou, Xinbing Wang, and Zhouhan Lin. "Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition". In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4491–4503, Dublin, Ireland, May 2022. Association for Computational Linguistics

[12] Wei-Ning Hsu and Bowen Shi. uhubert: "Unified mixed-modal speech pretraining and zero shot transfer to unlabeled modality". In Advances in Neural Information Processing Systems, 2022.

[13] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. "Learning audio-visual speech representation by masked multimodal cluster prediction. In International Conference on Learning Representations", 2022

[14] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2017. :Deep voice 3: Scaling text-tospeech with convolutional sequence learning". arXiv preprint arXiv:1710.07654 (2017)

[15] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. "Natural tts synthesis by conditioning wavenet

on mel spectrogram predictions". In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 4779–4783.

[16] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. "Sequence to sequence learning with neural networks". In Advances in neural information processing systems. 3104– 3112

[17] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. "Neural machine translation by jointly learning to align and translate". arXiv preprint arXiv:1409.0473 (2014)

[18] Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025 (2015).

[19] Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025 (2015).

[20] Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat et al. "Google's multilingual neural machine translation system: Enabling zero-shot translation." Transactions of the Association for Computational Linguistics 5 (2017): 339-351.

[21] Fan, Bo, Lijuan Wang, Frank K. Soong, and Lei Xie. "Photo-real talking head with deep bidirectional LSTM." In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4884-4888. IEEE, 2015.

[22] Suwajanakorn, Supasorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. "Synthesizing obama: learning lip sync from audio." ACM Transactions on Graphics (ToG) 36, no. 4 (2017): 1-13.

[23] Karras, Tero, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. "Audio-driven facial animation by joint end-to-end learning of pose and emotion." ACM Transactions on Graphics (TOG) 36, no. 4 (2017): 1-12.

[24] Chung, Joon Son, Amir Jamaludin, and Andrew Zisserman. "You said that?." arXiv preprint arXiv:1705.02966 (2017).

[25] Li, Dengshi, Yu Gao, Chenyi Zhu, Qianrui Wang, and Ruoxi Wang. "Improving Speech Recognition Performance in Noisy Environments by Enhancing Lip Reading Accuracy." Sensors 23, no. 4 (2023): 2053.

[26] Tan, Ke, Yong Xu, Shi-Xiong Zhang, Meng Yu, and Dong Yu. "Audio-visual speech separation and dereverberation with a two-stage multimodal network." IEEE Journal of Selected Topics in Signal Processing 14, no. 3 (2020): 542-553.

[27] Togami, Masahito. "Joint training of deep neural networks for multi-channel dereverberation and speech source separation." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3032-3036. IEEE, 2020.

[28] Zhao, Yan, Zhong-Qiu Wang, and DeLiang Wang. "Two-stage deep learning for noisy-reverberant speech enhancement." IEEE/ACM transactions on audio, speech, and language processing 27, no. 1 (2018): 53-62.

[29] Fan, Cunhang, Jianhua Tao, Bin Liu, Jiangyan Yi, and Zhengqi Wen. "Joint Training for Simultaneous Speech Denoising and Dereverberation with Deep Embedding Representations." In INTERSPEECH, pp. 4536-4540. 2020.

[30] Son, Jin-Young, and Joon-Hyuk Chang. "Attention-based joint training of noise suppression and sound event detection for noise-robust classification." Sensors 21, no. 20 (2021): 6718.

[31] Tan, Xu, and Xiao-Lei Zhang. "Speech enhancement aided end-to-end multi-task learning for voice activity detection." In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6823-6827. IEEE, 2021.

[32] Jung, Youngmoon, Younggwan Kim, Yeunju Choi, and Hoirin Kim. "Joint Learning Using Denoising Variational Autoencoders for Voice Activity Detection." In Interspeech, pp. 1210-1214. 2018.

[33] Xu, Tianjiao, Hui Zhang, and Xueliang Zhang. "Joint training ResCNN-based voice activity detection with speech enhancement." In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1157-1162. IEEE, 2019.

[34] Howard, Nathan, Alex Park, Turaj Zakizadeh Shabestary, Alexander Gruenstein, and Rohit Prabhavalkar. "A neural acoustic echo canceller optimized using an automatic speech recognizer and large scale synthetic data." In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7128-7132. IEEE, 2021.

[35] Braga, Otavio, and Olivier Siohan. "Best of both worlds: Multi-task audio-visual automatic speech recognition and active speaker detection." In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6047-6051. IEEE, 2022.

[36] Rose, Richard, and Olivier Siohan. "End-to-end multi-talker audio-visual ASR using an active speaker attention module." arXiv preprint arXiv:2204.00652 (2022).

[37] Son Chung, J., A. Senior, O. Vinyals, and A. Zisserman. "Lip Reading Sentences in the Wild." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, July 21–26, 2017, 6447–6456.

[38] He, Yibo, Kah Phooi Seng, and Li Minn Ang. 2023. "Multimodal Sensor-Input Architecture with Deep Learning for Audio-Visual Speech Recognition in Wild." Sensors 23, no. 4 (2023): 1834. https://doi.org/10.3390/s23041834

[39] Hong, Joanna, Minsu Kim, Jeongsoo Choi, and Yong Man Ro. "Watch or Listen: Robust Audio-Visual Speech Recognition with Visual Corruption Modeling and Reliability Scoring." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18783-18794. 2023.

[40] Hu, Yuchen, Ruizhe Li, Chen Chen, Heqing Zou, Qiushi Zhu, and Eng Siong Chng. "Cross-Modal Global Interaction and Local Alignment for Audio-Visual Speech Recognition." arXiv preprint arXiv:2305.09212 (2023).

[41] Hu, Yuchen, Ruizhe Li, Chen Chen, Chengwei Qin, Qiushi Zhu, and Eng Siong Chng. "Hearing Lips in Noise: Universal Viseme-Phoneme Mapping and Transfer for Robust Audio-Visual Speech Recognition." arXiv preprint arXiv:2306.10563 (2023).