International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Audio-Visual Speech Reconstruction Using Hybrid Deep Learning with Conditional Random Fields and Intelligent Chasing Optimization

Aditya. N. Magdum[1*], S. B. Patil[2]

Submitted: 12/09/2023 **Revised:** 17/11/2023 **Accepted:** 27/11/2023

Abstract— With its many uses in virtual reality, education, training, and other domains, lip-to-speech (LTS) synchronization is an essential tool for creating lifelike face animations. However, existing approaches still struggle to create high-fidelity facial animations, particularly when faced with issues like lip jitter and unstable facial motions in continuous frame sequences. To improve LTS models' capacity to precisely reconstruct speech from visual data, this study develops a Hybrid Deep Learning model coupled with Conditional Random Field-based Intelligent Chasing Optimization (HDL-CRF-ICO). For the preprocessing stage, the model chooses 100 frames at random, and the Structured Similarity Index (SSIM) is used to identify keyframes. Similarity scores are computed by this index, and which frames are chosen for additional processing are determined by certain criteria. The model then makes use of sophisticated methods, such as AV features, which improve speech recognition by combining visual information from lip movements with audio inputs. By offering the optimum global solution, the ICO algorithm speeds up convergence, and by lowering the error value, it allows the model to produce precise results. Accordingly, the proposed model obtained the performance as Bilingual Evaluation Understudy (BLEU) scores of 0.48, Metric for Evaluation of Translation with Explicit Ordering (METEOR) scores of 0.30, Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores of 0.53, and Semantic Propositional Image Caption Evaluation (SPICE) scores of 24.9, as well as for K-Fold and METEOR with 0.31, SPICE with 25.7, BLEU with 0.49, and ROUGE with 0.54 for training percentage using Grid Audio-Visual Speech Corpus dataset.

Keywords - *Lip-to-Speech Synchronization, Hybrid Deep Learning, Conditional Random Field (CRF), Intelligent Chasing Optimization (ICO), Structured Similarity Index (SSIM), Audio-Visual (AV) Features, Speech Reconstruction*

I. INTRODUCTION

Visual speech recognition (VSR), also known as lip-reading, is a technology that aims to recognize and interpret speech from the movements of a person's lips and facial features. It involves using computer vision techniques to analyze the visual cues generated by the articulation of speech sounds and translating them into textual or auditory information. Visual speech recognition systems utilize video data, usually captured from a camera, and process the facial movements, lip shapes, and other visual cues exhibited during speech production. These visual cues are then matched with a database of phonetic or linguistic models to decipher the spoken words.

The technology often employs machine learning algorithms to improve its accuracy over time by training on large datasets. Speech recognition facilitates seamless communication between humans and machines. However, the effectiveness of such systems diminishes in the presence of background noise. To mitigate this, incorporating visual data derived from mouth movements and lip configurations can ameliorate the impact of acoustic disturbances, consequently enhancing the performance of speech recognition systems.

The primary focus of this study pertains to lip-reading techniques for enhancing speech recognition. The majority of related research has predominantly cantered on audio-visual speech recognition investigations. In VSR systems, the recognition of spoken words hinges on the analysis of visual signals produced during speech.

For effective speech recognition, pertinent visual information is gleaned from the mouth and lip area of the face. Consequently, lip-reading systems adopt various methodologies; some directly pinpoint the speaker's lips, while others first identify the face using prior knowledge and subsequently zero in on lip localization. Notably, achieving precise lip and face localization poses challenges due to factors like sensor quality, lighting conditions, background, lip dynamics, pose variations, shadows, facial expressions, scaling, rotations, and occlusions.

[1*] *Research Scholar, Department of Electronics Engineering, Shivaji University, Kolhapur, Maharashtra India, aditya.magdum112@gmail.com*
[2] *Professor, Department of Electronics & Telecommunication, JJMCOE, Jaysingpur. Kolhapur, Maharashtra India patilsb0908@gmail.com*

The practice of reconstructing spoken text from a speaker's lip movements in a soundless movie is known as LTS synchronization. This technique is especially useful when there is no audio for a number of reasons, such as inadequate recording gear, background noise, or transmission problems [1]. The new technology known as LTS generation is rapidly evolving and has advanced significantly. In addition to giving those who are deaf or speech-impaired a new way to communicate, it has a big impact on education. LTS creation, for instance, can assist kids with speech articulation and vocal expressiveness [2].

LTS technology is also necessary for in-person communication in daily life, particularly with the growth of video conferencing and virtual meetings. Among the difficulties is accurately recognizing conversations when voice signals are difficult to capture, such as in loud environments. Lip-to-speech technology can be helpful in a variety of settings, including loud events, busy shopping centers, and quiet video chats [3]. Consequently, the increasing demand for this technology underscores its potential to improve communication in circumstances when the voice is unclear or absent. An active research topic at the intersection of computer vision and speech processing is voice-driven lip synchronization.

The aim of this discipline is to generate face animations that accurately correspond to spoken text and facial images or videos [4]. It may be used in a variety of domains, including virtual reality, healthcare, digital entertainment, and distant learning. Despite the technology's early use in virtual presenters and intelligent customer service, there are still significant problems with lip-sync accuracy and naturalness [5] [6]. LTS synthesis, which involves predicting the relevant speech from a sequence of images of talking faces, is an essential stage in the development of LTS. This approach has a variety of applications, including dubbing silent movies, assisting patients who are mute, and restoring voice for video conferencing in loud environments [7]. Most currently available lip-reading and audio-visual speech recognition methods assume that the lips be visible and unhindered. However, in practice, this assumption is not accurate since the speaker's lips can easily be blocked by hands or microphones [8], which lowers performance. Due to the high sensitivity of contemporary audiovisual speech recognition systems to lip occlusion, error rates might occasionally increase [9]. In Lip to Speech or LTS synthesis, an encoder-decoder structure is employed. The encoder extracts the linguistic information and voice characteristics of a talking video, while the decoder converts the associated audio [7].

Current systems often have significant issues, despite advancements in deep learning techniques. One of the earliest lip-sync systems was Video Rewrite, which mapped phonemes to mouth shapes and blended them onto a target video. Though they offer more general solutions, modern methods like PC-AVS [10] and GCAVT [11] still have a lot of problems. These approaches often separate pose and emotion, but they lose the speaker's uniqueness, leading to uneven face boundaries and poor visual quality [12]. One of the several elements that contributes to the challenges in LTS creation is variation in pronunciation. Since multiple pronunciations of the same word might have different meanings, accurate LTS synthesis can be difficult. The HDL-CRF-ICO model is proposed to address the issues highlighted in the LTS synchronization section and generate text appropriately. This optimizes error reduction and effectively integrates advanced audio-visual elements to improve LTS synchronization. As a result, issues like lip occlusion and speech variability are successfully addressed, improving the accuracy of speech reconstruction from visual input.

The rest of this paper is constructed as follows: Section II provides an overview of the related work. The proposed approach for LTS synchronization is explained in Section III. And Section IV ultimately includes experimental results and performance assessments. In Section V, close the paper with the summary of the results.

II. RELATED WORK

By using data augmentation techniques to generate more data samples for classification model generation, He, Y., et al. [13] created improved AVSR models. Traditional methods were combined with more modern approaches, such as generative adversarial networks (GANs). After training the models using enhanced data from well-known datasets to validate their methods, they tested the models using the original data. Experimental results showed that the augmentation strategy and the proposed AVSR model improved performance in noisy datasets. However, advanced GAN models for visual and auditory modalities were not included in the AVSR model. For the first time in the Chinese LTS synthesis area, Yang, Q., et al. [2] presented the sentence-level lip-to-speech synthesis architecture FA-GAN.

The sophisticated speech-driven lip synchronization model VividWav2Lip was created by Liu, L., et al. [6]. A cross-attention mechanism for improved audio-visual feature fusion, an optimized network topology with Squeeze-and-Excitation (SE) residual blocks, and the inclusion of the Code Former facial restoration network for post-processing are its three primary contributions. Extensive experiments on a diverse dataset of different languages and face types showed promising results, with 85% of participants assessing the animations as more realistic than existing techniques. However, the model struggled to capture emotion-related facial movement characteristics since there were no advanced emotion networks like normalizing flows or vector quantization models.

The FastLips end-to-end neural AVTTS model was developed by Lenglet, M., et al. [14] using the FastSpeech2 architecture. They demonstrated that FastLips generated lip animations of superior quality as compared to the baseline AVTacotron2. The model highlighted the advantages of early differentiation between aural and visual modalities, which promotes more successful asynchronous actions, in order to predict lip motions. However, the model's inability to leverage the FastLips architecture for expressive audiovisual synthesis proved a limitation. Furthermore, visual cues like eye blinks and head nods were not included in the visual variance adapter.

Lu, J., et al. [15] conducted a novel task called automated voice-over (AVO), which aims to produce speech in real time using a quiet, pre-recorded film. Unlike traditional speech synthesis, AVO not only produces natural-sounding speech but also ensures perfect lip-speech synchronization. A logical approach to handle AVO is to condition speech rendering on the temporal evolution of the lip sequences in the video. However, the model's inability to use visual input limited its ability to achieve flawless lip-speech synchronization and fine-grained duration control.

Passos, L.A., et al. [16] combined Graph Neural Networks with canonical correlation analysis (CCA-GNN) to develop a novel multimodal self-supervised architecture for energy-efficient audio-visual (AV) speech augmentation. The technique was built on a state-of-the-art CCA-GNN that maximized the correlation between pairs of augmented views of the same input in order to learn representative embeddings and embellish disconnected features. Reducing duplicate information and eliminating augmentation-variant information while preserving augmentation-invariant data was the key idea. The lack of a physiologically realistic neuronal architecture and memory processes in the model hindered effective channel-to-channel and cross-channel interactions inside convolutional neural networks.

A novel audio-visual speech recognition architecture was presented by Li, J., et al. [17]. It combined audio and visual input utilizing unified cross-modal attention and temporal concatenation. The resulting sequence was then put into a unified Conformer encoder. They proposed an additional synchronization-aware loss optimization to increase the system's robustness in audiovisual out-of-synch conditions. A manual attention alignment method was also developed, which improved computation efficiency and identification accuracy. Through the resolution of synchronization issues, this technique significantly enhanced audiovisual speech recognition.

The focus of He, Y., et al. [18] was on applying a novel multimodal generative adversarial network (GAN)-based AVSR architecture for artificial intelligence in the Internet of Things (IoT). The study looked into both traditional and GAN-based techniques to increase the accuracy of AVSR classification. But when the AVSR architecture was applied to a range of IoT devices, significant problems including privacy security and the need for low-latency data processing surfaced. A number of issues needed to be fixed before the AVSR architecture could be successfully used in real-world IoT applications.

III. METHODOLOGY

The Proposed HDL-CRF-ICO architecture is used in the study to identify LTS synchronization. Both audio and video data from the Grid Audio-Visual Speech Corpus dataset are used as model inputs. The most pertinent frames are chosen from the movie to record significant lip movements in the first preprocessing step. To aid with more precise feature extraction, these frames are enhanced to lower noise and improve picture quality. ResNet-101 is used to detect complex patterns in lip movements, and frames are chosen to extract visual information.

Additionally, DLib's landmark identification technique is used to recognize facial landmarks, particularly those from the chin and lower face. Because they highlight important regions like the chin and the corners of the lips, these landmarks are essential for comprehending the facial dynamics of speech. The energy distribution in the audio stream is analyzed using statistical indicators like spectral centroid, spectral bandwidth, and spectral contrast on the audio side. This provides information about the speaker's traits and speech rhythm. The speaker's voice tone and characteristics may also be ascertained with the use of statistical descriptors such as mean, variance, skewness, and kurtosis. Vulture and Harris Hawks Optimization are used to improve the retrieved characteristics and concatenate them into a single vector in order to maximize performance.

A CNN-BiLSTM classifier, which generates logical captions by processing both audio and visual input, is the result of this development. Lastly, depending on the lip movements, the produced subtitles match the video material. In Figure 1, the suggested framework is schematically illustrated.

A total of 34,000 phrases from the Grid Audio-Visual Speech Corpus dataset was used as the input data. Thirty-four talkers, each giving 1,000 sentences, deliver these sentences. Frame selection and frame enhancement are the two steps in the pre-processing phase, which receives the raw input data. One hundred frames are chosen at equal intervals from every video as part of the frame selection procedure. Assume, *ff1, ff2,….ffr* [28] be the frames extracted from the videos and among which the selected frames (keyframes) are extracted using the similarity measure, namely SSIM and pretrained based similarity measure. Assume the two consecutive frames as *ff1* and *ff2.*

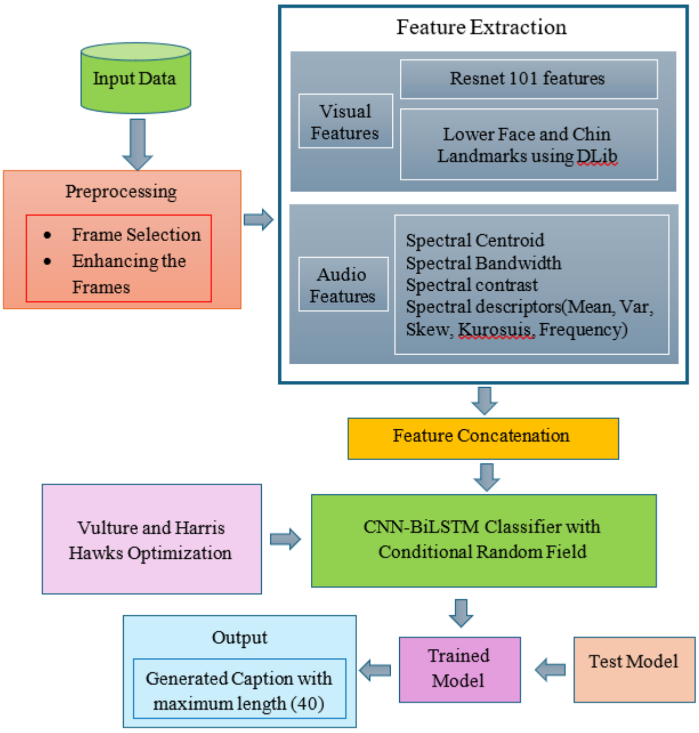

Input Data
Preprocessing
Frame Selection
Enhancing the Frames
Feature Extraction
Visual Features
Resnet 101 features
Lower Face and Chin Landmarks using DLib
Audio Features
Spectral Centroid
Spectral Bandwidth
Spectral contrast
Spectral descriptors(Mean, Var, Skew, Kurosuis, Frequency)
Feature Concatenation
Vulture and Harris Hawks Optimization
CNN-BiLSTM Classifier with Conditional Random Field
Output
Generated Caption with maximum length (40)
Trained Model
Test Model

Figure 1. Diagrammatic Illustration of the HDL-CRF-ICO Model

Between these two frames, *ff1* and *ff2*, calculate the SSIM value. Conversely, a similarity measure value based on pretrained data is calculated for the two successive frames and. The pretrained based similarity measure and the SSIM measure are also calculated by taking into account two consecutive frames for every frame that was recorded from the video. As a result, an average measure value is computed using the pretrained score value and SSIM. The frames with the average value are then chosen. Using these average values, the similarity must be checked once more; if it is greater than 1.5, the corresponding frames can be chosen as keyframes. This process is repeated until all of the keyframes have been chosen. As a result, the enhanced pre-processed output has a dimension of, which is mathematically denoted as:

$$H' = \left\{ \left((A_1, B_1), C_1 \right)', \ldots, \left((A_b, B_b), C_b \right)', \ldots, \left((A_a, B_a), C_a \right)' \right\} \qquad (1)$$

To improve the effectiveness of LTS systems, audio-visual (AV) capabilities—which integrate audio information with visual cues from lip movements—are crucial. While advanced models like ResNet-101, which can identify intricate patterns in lip dynamics, are employed to extract visual information, DLib is utilized to locate landmarks on the chin and lower face. To decode the energy distribution and characteristics of the speaker's voice, the audio side looks at features including spectral centroid, spectral bandwidth, spectral contrast, and other spectrum descriptors like mean, variance, skewness, kurtosis, and frequency.

The HDL-CRF-ICO framework efficiently processes audio-visual information and maximizes model performance to enable precise lip-to-speech synchronization. The model effectively synchronizes auditory and visual inputs by extracting features using Conv3D, BiLSTM, and CRF layers. The dimensions [Nx100x52x52x6] are the first parameters the video input layer uses to define the number of channels and batch size. Three Conv3D layers handle the input after that, with activation and max-pooling functions coming after each layer. These layers gradually reduce the input size [Nx100x2700] in order to concentrate on identifying patterns. The flattened layer is the result of processing the gathered attributes. BiLSTM and dropout layers receive the flattened video features as input, and by capturing both forward and backward temporal associations, they provide an output with dimensions [Nx256]. In parallel, audio features having an initial input dimension of [Nx395x1] are processed using BiLSTM layers. To prevent overfitting, a dropout layer is added to the AV features while keeping their size constant. The audio and video characteristics are concatenated, reshaped, and sent to the CRF layer for further processing to produce the output dimensions [Nx512]. This stage ensures optimal feature extraction and integration for synchronization. Dense layers then use the processed information to forecast the output class probabilities. The final output layer generates results with size [Nx40], which represent the anticipated classes for LTS synchronization. The HDL-CRF-ICO model's architecture is depicted in Figure 2.

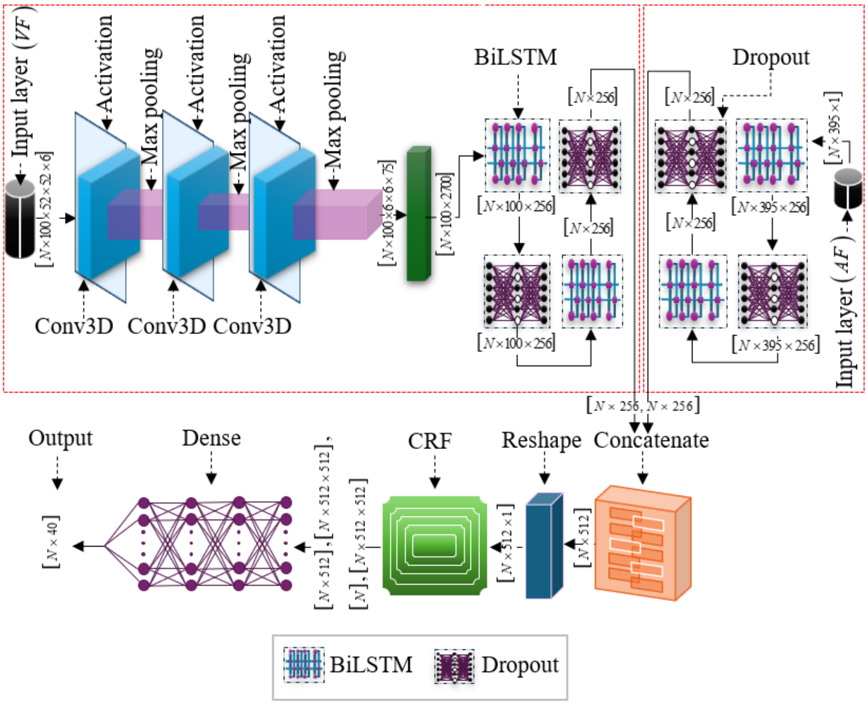


Figure 2. The HDL-CRF-ICO model's architecture

The suggested hybrid deep learning model detects lip-to-speech synchronization and generates text from video by using CNN [22] feature extraction and BI-LSTM training based on AV characteristics. The CNN-BiLSTM model uses a set of convolution layers, including a kernel, Rectified Linear Unit (ReLU), max pooling, and fully connected layers, to categorize the video frames. The convolution layer is the primary component in charge of learning the frame's features, extracting features from the source frame, and preserving the connection between pixels by applying small blocks of source data.

IV. EXPERIMENT RESULTS

The Grid Audio-Visual Speech Corpus dataset is subjected to a number of evaluation criteria in order to evaluate the synchronization process of LTS using the HDL-CRF-ICO model. These results are then succinctly described in the following sections. The Grid Corpus is a large audiovisual collection [27] designed for speech perception computational and behavioral studies. Its 34 talkers provide 1,000 sentences apiece, for a total of 34,000 phrases spoken. The phrases have a certain structure, such "put red at G9 now." High-quality wav audio at a 25 kHz sample rate is organized by talker in the audio 25k.zip file, while word-level temporal alignments for each talker are included in alignments.zip.

4.1 Performance Evaluation

The evaluation of the proposed HDL-CRF-ICO model is accessed using the matrices BLEU, METEOR, ROUGE, and SPICE.

BLEU: One of the most popular measures for assessing image captioning models is the Bilingual Evaluation Understudy (BLEU). Its first purpose was to evaluate translations automatically. The foundation of BLEU is the co-occurrence of n-gram analysis between the reference and the prediction.

METEOR: Some of the shortcomings of the BLEU statistic are attempted to be addressed by statistics for Evaluation of Translation with Explicit Ordering (METEOR). Along with finding matching words, METEOR calculates a harmonic mean using accuracy and recall values for unigram matches.

ROUGE: The amount of overlapping phrase sequences and word pairs in the generated captions is measured using Recall-Oriented Understudy for Gisting Evaluation (ROUGE). There are four variations of this metric: ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S.

SPICE: Instead of employing the n-gram technique, Semantic Propositional Image Caption Evaluation (SPICE) creates a scene graph from the reference and prediction sentences. The Stanford Scene Graph Parser serves as the basis for the pre-trained model of the SPICE.

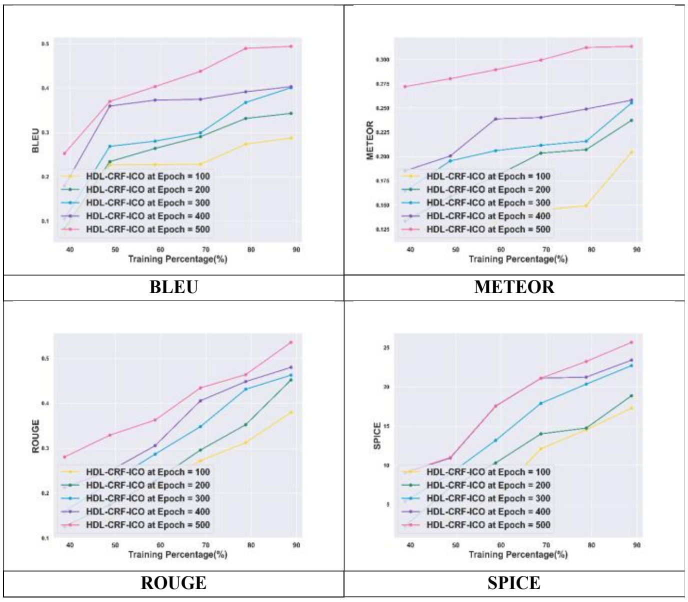
4.2 Performance Analysis

The Grid Audio-Visual Speech Corpus dataset is used for the performance analysis, which is based on the number of epochs on Training Percentage and K-Fold.

Using a number of metrics, including BLEU, METEOR, ROUGE, and SPICE, Figure 3 shows the performance of the proposed HDL-CRF-ICO model with respect to the training % on the Grid Audio-Visual Speech Corpus dataset. Through 100, 200, 300, 400, and 500 epochs, the model's performance is assessed with a constant training percentage of 90%. The BLEU score of the HDL-CRF-ICO model is 0.49 at epoch 100 and steadily declines across the epochs. Specifically, for epochs 200, 300, 400, and 500, the BLEU values are 0.29, 0.34, 0.40, and 0.04, respectively. The METEOR score is 0.20 at epoch 100 and increases to 0.24 at epoch 200. It remains at 0.26

in epochs 300 and 400 and rises to 0.31 in epoch 500. The ROUGE score, which is 0.38 in epoch 100, increases to 0.45 in epoch 200, 0.46 in epoch 300, 0.48 in epoch 400, and 0.54 in epoch 500. The corresponding SPICE scores for epochs 100, 200, 300, and 500 are 17.29, 18.87, 22.71, 23.40, and 25.68.

Figure 4 illustrates the metrics that are utilized to evaluate the performance of the proposed HDL-CRF-ICO model with K-Fold on the Grid Audio-Visual Speech Corpus dataset: BLEU, METEOR, ROUGE, and SPICE. A consistent 10-fold test is used to assess the model's performance over 100, 200, 300, 400, and 500 epochs. The HDL-CRF-ICO model's BLEU score starts at epoch 500 at 0.48 and gradually decreases across the epochs. Specifically, the BLEU values are 0.35, 0.42, 0.46, and 0.48 for epochs 100, 200, 300, and 400, respectively.



BLEU

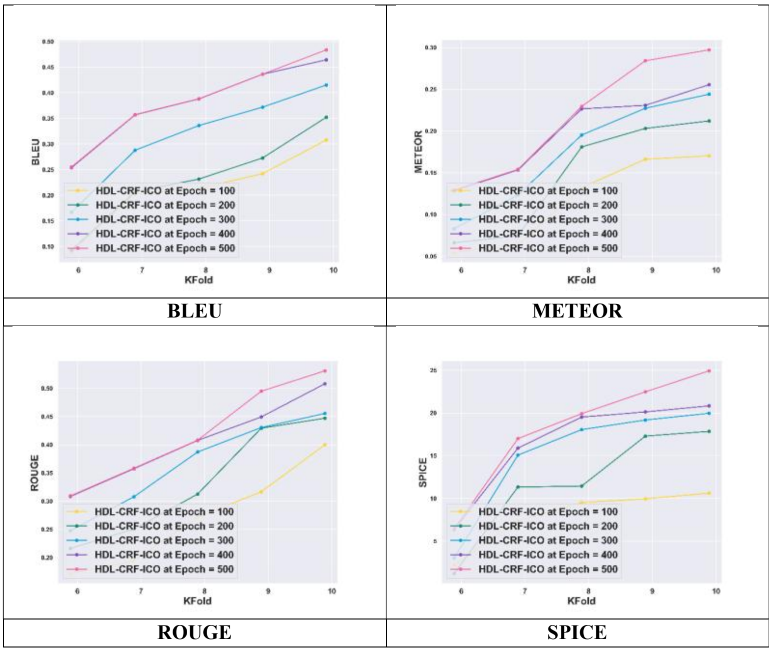
METEOR

ROUGE

SPICE

Figure 3. Performance analysis with training percentage using the Grid Audio-Visual Speech Corpus dataset

The METEOR score is 0.17 at epoch 100, 0.21 at epoch 200, 0.24 at epoch 300, 0.26 at epoch 400, and 0.30 at epoch 500. The rough score for Epoch 100 is 0.40, for Epoch 200 it is 0.45, for Epoch 300 it is 0.46, for Epoch 400 it is 0.51, and for Epoch 500 it is 0.53. The comparable SPICE scores for epochs 100, 200, 300, 400, and 500 are 10.62, 17.85, 19.97, 20.84, and 24.93, respectively. 90% of the data is used for training, with the remaining 10% going toward testing, according to the training percentage.The model performs better when more training data is used. The model continues to be trained for the designated number of epochs, with more training resulting in better performance. Epoch 100 also denotes that the model has been trained for 100 iterations. As a result, the findings show that the HDL-CRF-ICO model performs better as the number of epochs increases.



BLEU

METEOR

ROUGE

SPICE

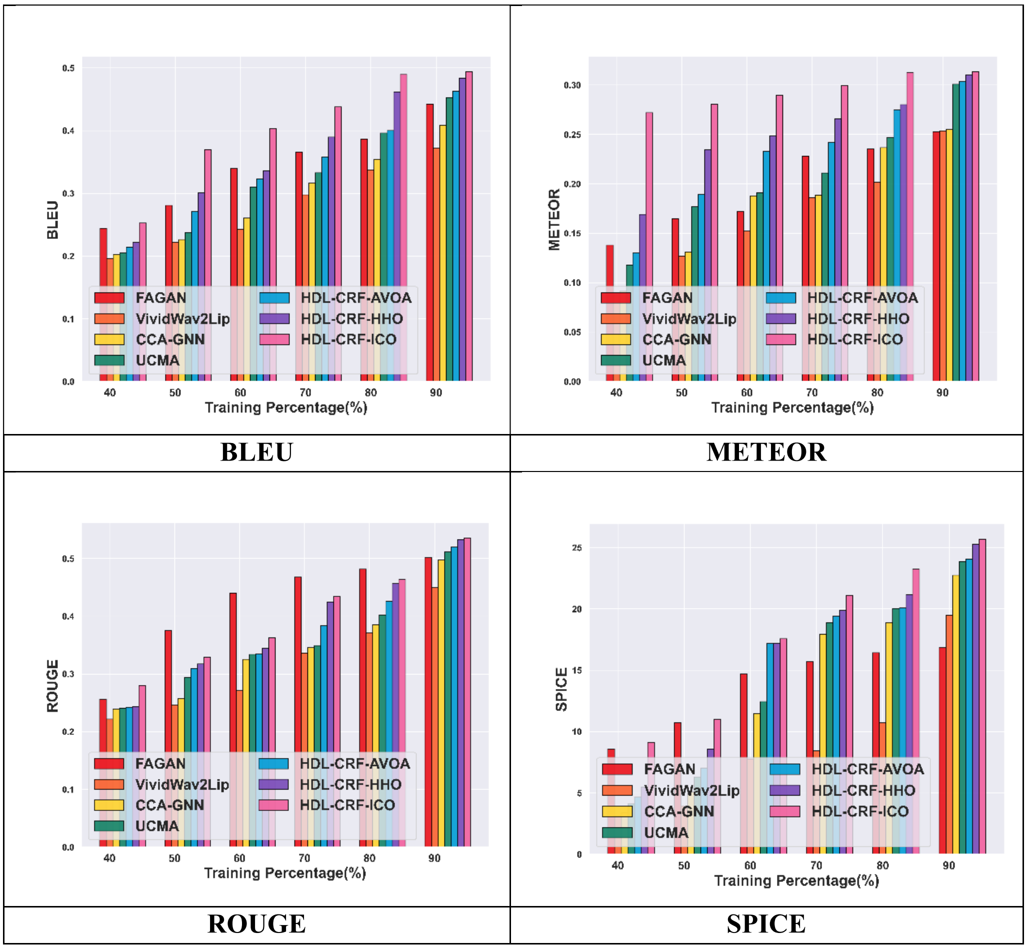
Figure 4. Performance analysis with K-Fold using the Grid Audio-Visual Speech Corpus dataset

4.3 Comparative Analysis

The Grid Audio-Visual Speech Corpus Dataset was utilized to evaluate the proposed HDL-CRF-ICO model, as shown in Figure 5. It's remarkable BLEU score of 0.49 outperformed several existing methods. Specifically, HDL-CRF-ICO performed better than HDL-CRF-AVOA (0.46), HDL-CRF-HHO (0.48), FAGAN (0.44), VividWav2Lip (0.37), CCA-GNN (0.41), and UCMA (0.45). With a METEOR score of 0.30, the HDL-CRF-ICO model performed noticeably better than the other method. It performed 19% better than FAGAN, VividWav2Lip, and CCA-GNN, 4% better than UCMA, 3% better than HDL-CRF-AVOA, and 1% better than HDL-CRF-HHO. HDL-CRF-AVOA scored 0.52, HDL-CRF-HHO scored 0.53, UCMA scored 0.51, VividWav2Lip scored 0.45, and FAGAN and CCA-GNN both scored 0.50. These were the results for the ROUGE measure. The HDL-CRF-ICO model had the highest ROUGE score of 0.54, outperforming all other methods. Finally, with a score of 25.7 on the SPICE test, the HDL-CRF-ICO model fared better than alternative strategies. The scores for HDL-CRF-AVOA, HDL-CRF-HHO, CCA-GNN, UCMA, and FAGAN were 24.1, 25.3, 22.7, 23.9, and 16.8, respectively. The score for VividWav2Lip was 19.5.

Figure 6 illustrates how the Grid Audio-Visual Speech Corpus Dataset was used to compare the suggested HDL-CRF-ICO model to existing techniques. The HDL-CRF-ICO model outperformed a number of current methods with an exceptional BLEU score of 0.48. It fared better than the HDL-CRF-AVOA and HDL-CRF-HHO models, which both received scores of 0.48; FAGAN, which received a score of 0.41; VividWav2Lip, which received a score of 0.36; CCA-GNN, which received a score of 0.39; and UCMA, which received a score of 0.45.

The HDL-CRF-ICO model significantly outperformed the other approaches, scoring 0.30 on the METEOR scale. Additionally, it outperformed UCMA by 19%, FAGAN by 12%, VividWav2Lip by 43%, CCA-GNN by 34%, HDL-CRF-AVOA by 3%, and HDL-CRF-HHO by 3%.



BLEU

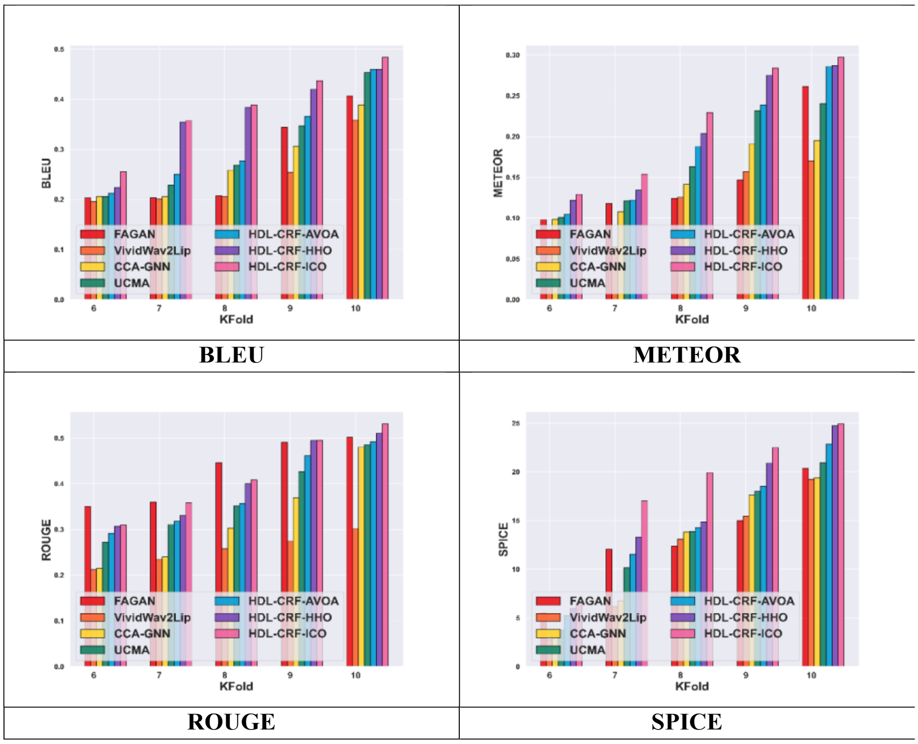
METEOR

ROUGE

SPICE

Figure 5. Comparative Analysis with training percentage using the Grid Audio-Visual Speech Corpus Dataset

The following were the outcomes of the optimum models and current methodologies for the ROUGE metric: HDL-CRF-HHO scored 0.51, UCMA and HDL-CRF-AVOA both scored 0.49, VividWav2Lip scored 0.48, FAGAN scored 0.50, and CCA-GNN scored 0.30. Outperforming all other models, the HDL-CRF-ICO model had the highest ROUGE score of 0.53. Lastly, the HDL-CRF-ICO model outperformed other approaches with a score of 24.9 on the SPICE measure. HDL-CRF-AVOA scored 22.9, HDL-CRF-HHO scored 24.7, UCMA scored 20.9, VividWav2Lip scored 19.2, CCA-GNN scored 19.4, and FAGAN scored 20.4.



BLEU

METEOR

ROUGE

SPICE

Figure 6. Comparative Analysis with K-Fold using the Grid Audio-Visual Speech Corpus Dataset

4.4 Comparative Result and Discussion

The HDL-CRF-ICO model's comparison with other current approaches is examined in this section. This study considers the following methods: CCA-GNN [16], FAGAN [2], VividWav2Lip [6], and UCMA [17]. Despite the advantages of each of these techniques, it is challenging to successfully coordinate lip motions due to their shortcomings. The FAGAN model has low attention accuracy and does not integrate facial expressions, normalizing fluxes, and vector quantization approaches.

Table 1. Comparative Discussion of the HDL-CRF-ICO model

Methods			FAGAN	VividWav2Lip	CCA-GNN	UCMA	HDL-CRF-AVOA	HDL-CRF-HHO	HDL-CRF-ICO
The Grid Audio-Visual Speech Corpus dataset	TP=90%	BLEU	0.44	0.37	0.41	0.45	0.46	0.48	**0.49**
		METEOR	0.25	0.25	0.25	0.30	0.30	0.31	**0.31**
		ROUGE	0.50	0.45	0.50	0.51	0.52	0.53	**0.54**
		SPICE	16.8	19.5	22.7	23.9	24.1	25.3	**25.7**
	K-Fold=10	BLEU	0.41	0.36	0.39	0.45	0.46	0.46	**0.48**
		METEOR	0.26	0.17	0.20	0.24	0.29	0.29	**0.30**
		ROUGE	0.50	0.30	0.48	0.49	0.49	0.51	**0.53**
		SPICE	20.4	19.2	19.4	20.9	22.9	24.7	**24.9**

The interactions among convolutional neural networks affect the performance of the CCA-GNN model, however the VividWav2Lip model is unable to manage this complexity. Moreover, the UCMA model requires a significant amount of computing work to train. The suggested HDL-CRF-ICO model solves these problems and outperforms the others on a variety of metrics for recognizing different lip movements. The suggested model outperforms the existing techniques, as seen in Table 1.

V. CONCLUSION

During the preprocessing stage, the suggested HDL-CRF-ICO model selects 100 randomly selected frames from each video. After that, keyframes are chosen for further processing based on SSIM-calculated similarity scores, with certain thresholds determining which frames are chosen. The method significantly improves lip-to-speech synchronization by combining AV feature extraction with Conditional Random Fields and Intelligent Chasing Optimization. This recently created method effectively addresses the shortcomings of existing systems, such as uneven facial boundaries and poor visual quality, by increasing the accuracy and efficiency of lip-to-speech synthesis. By analyzing both audio and visual data, the model yields better performance measures and offers a more detailed understanding of speech dynamics. For the TP, the HDL-CRF-ICO model showed remarkable performance with BLEU scores of 0.48, METEOR scores of 0.30, ROUGE scores of 0.53, and SPICE scores of 24.9, and for K-Fold and METEOR with 0.31, SPICE with 25.7, BLEU with 0.49, and ROUGE with 0.54. These findings show how beneficial it may be for communication, assistive technology, and multimedia content production. The HDL-CRF-ICO method improves synthesized speech quality and bridges communication barriers, paving the way for future research and development in lip-to-speech synchronization and related fields.

REFERENCES

[1] Niu, Z. and Mak, B., "On the Audio-visual Synchronization for Lip-to-Speech Synthesis", In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.7843-7852, 2023.

[2] Yang, Q., Bai, Y., Liu, F. and Zhang, W., "Integrated visual transformer and flash attention for lip-to-speech generation GAN", Scientific Reports, vol.14, no.1, pp.4525, 2024.

[3] Kim, M., Hong, J. and Ro, Y.M., "Lip-to-speech synthesis in the wild with multi-task learning", In ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.1-5, June 2023.

[4] Park, S.J., Kim, M., Choi, J. and Ro, Y.M., "Exploring Phonetic Context-Aware Lip-Sync for Talking Face Generation", In ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4325-4329, IEEE, April 2024.

[5] Wang, J., Qian, X., Zhang, M., Tan, R.T. and Li, H., "Seeing what you said: Talking face generation guided by a lip reading expert", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.14653-14662, 2023.

[6] Liu, L., Wang, J., Chen, S., and Li, Z., "VividWav2Lip: High-Fidelity Facial Animation Generation Based on Speech-Driven Lip Synchronization, Electronics, vol.13, no.18, pp.3657, 2024.

[7] Sheng, Z.Y., Ai, Y. and Ling, Z.H., "Zero-shot personalized lip-to-speech synthesis with face image based voice control", In ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.1-5, June 2023.

[8] Hong, J., Kim, M., Choi, J. and Ro, Y.M., "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.

[9] Wang, J., Pan, Z., Zhang, M., Tan, R.T. and Li, H., "Restoring Speaking Lips from Occlusion for Audio-Visual Speech Recognition", In Proceedings of the AAAI Conference on Artificial Intelligence, vol.38, no.17, pp.19144-19152, March 2024.

[10] Liang, B., Pan, Y., Guo, Z., Zhou, H., Hong, Z., Han, X., Han, J., Liu, J., Ding, E. and Wang, J., "Expressive talking head generation with granular audio-visual control", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.3387-3396, 2022.

[11] Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X. and Liu, Z., "Pose-controllable talking face generation by implicitly modularized audio-visual representation", In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.4176-4186, 2021.

[12] Mukhopadhyay, S., Suri, S., Gadde, R.T., and Shrivastava, A., "Diff2lip: Audio conditioned diffusion models for lip-synchronization", In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp.5292-5302, 2024.

[13] He, Y., Seng, K.P. and Ang, L.M., "Multimodal Sensor-Input Architecture with Deep Learning for Audio-Visual Speech Recognition in Wild. Sensors", vol.23, no.4, pp.1834, 2023.

[14] Lenglet, M., Perrotin, O. and Bailly, G., "FastLips: an End-to-End Audiovisual Text-to-Speech System with Lip Features Prediction for Virtual Avatars", In International Speech Communication Association ISCA, pp.3450-3454, September 2024.

[15] Lu, J., Sisman, B., Liu, R., Zhang, M. and Li, H., "Visualtts: Tts with accurate lip-speech synchronization for automatic voice over", In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.8032-8036, May 2022.

[16] Passos, L.A., Papa, J.P., Del Ser, J., Hussain, A. and Adeel, A., "Multimodal audio-visual information fusion using canonical-correlated graph neural network for energy-efficient speech enhancement", Information Fusion, vol.90, pp.1-11, 2023.

[17] Li, J., Li, C., Wu, Y. and Qian, Y., "Unified Cross-Modal Attention: Robust Audio-Visual Speech Recognition and Beyond", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.32, pp.1941-1953, 2024.

[18] He, Y., Seng, K.P. and Ang, L.M., "Generative adversarial networks (GANs) for audio-visual speech recognition in artificial intelligence IoT. Information, vol.14, no.10, pp.575, 2023.

[19] Kalshetty, R. and Parveen, A., "Abnormal event detection model using an improved ResNet101 in context aware surveillance system", Cognitive Computation and Systems, vol.5, no.2, pp.153-167, 2023.

[20] Lashkov, I., Kashevnik, A., Shilov, N., Parfenov, V. and Shabaev, A., "Driver dangerous state detection based on OpenCV & dlib libraries using mobile video processing", In IEEE International

Conference on Computational Science and Engineering and IEEE International Conference on Embedded and Ubiquitous Computing, pp.74-79, August 2019.

[21] Chen, L., Yao, X., Tan, C., He, W., Su, J., Weng, F., Chew, Y., Ng, N.P.H. and Moon, S.K., "In-situ crack and keyhole pore detection in laser directed energy deposition through acoustic signal and deep learning", Additive Manufacturing, vol.69, pp.103547, 2023.

[22] Spandana, S., Madhura, B., Sandhya, A., Manish, A., and Kumar, K.P., "A Hybrid CNN-BILSTM Model for Continuous Sign Language Recognition Using Iterative Training", International Journal of Engineering Science and Advanced Technology, vol.23, no.05, 2023.

[23] Murugaiyan, S. and Uyyala, S.R., "Aspect-based sentiment analysis of customer speech data using deep convolutional neural network and bilstm", Cognitive Computation, vol.15, no.3, pp.914-931, 2023.

[24] Peng, X., Cao, H., Prasad, R. and Natarajan, P., "Text extraction from video using conditional random fields", International Conference on Document Analysis and Recognition, IEEE, pp.1029-1033, September 2011.

[25] Shehab, M., Mashal, I., Momani, Z., Shambour, M.K.Y., AL-Badareen, A., Al-Dabet, S., Bataina, N., Alsoud, A.R. and Abualigah, L., "Harris hawks optimization algorithm: variants and applications", Archives of Computational Methods in Engineering, vol.29, no.7, pp.5579-5603, 2022.

[26] Abdollahzadeh, B., Gharehchopogh, F.S. and Mirjalili, S., "African vultures optimization algorithm: A new nature-inspired metaheuristic algorithm for global optimization problems", Computers & Industrial Engineering, vol.158, pp.107408, 2021.

[27] The Grid Audio-Visual Speech Corpus Dataset, "https://zenodo.org/records/3625687", on January 2025.

[28] Ganesan, P., Jagatheesaperumal, S.K., Gaftandzhieva, S. and Doneva, R., "Novel Cognitive Assisted Adaptive Frame Selection for Continuous Sign Language Recognition in Videos Using ConvLSTM", International Journal of Advanced Computer Science & Applications, vol.15, no.7, 2024.