LIPCOORDNET: A DUAL-STREAM DEEP LEARNING ARCHITECTURE FOR VISUAL SPEECH RECOGNITION USING FACIAL LANDMARKS

Wissem Karous[1], Hanen Lajnef [2*], Tehani Dammak[1]

[1]School of Electronics and Telecommunications, University of Sfax, Tunisia

[2]Innov'COM Laboratory, Sup'Com, University of Carthage, Ariana, Tunisia

Emails: wissem.karous@enetcom.usf.tn[1], hanen.lajnef@enetcom.usf.tn[2*] (Corresponding author), tehani.dammak@enetcom.usf.tn[1]

ABSTRACT

Automated lip reading systems have emerged as critical assistive technologies for hearing-impaired individuals and communication in noisy environments. This research presents an advanced deep learning framework for sentence-level lip reading that integrates 3D Convolutional Neural Networks (3D CNN) with bidirectional Long Short-Term Memory (Bi-LSTM) networks, enhanced by facial landmark coordinates as supplementary input features. Our proposed LipCoordNet architecture achieves state-of-the-art performance on the GRID corpus benchmark, obtaining a Word Error Rate (WER) of 1.7% and Character Error Rate (CER) of 0.6%, representing significant improvements over existing state-of-the-art methodologies evaluated on the same dataset. The system demonstrates robust performance through the integration of spatial-temporal visual features and geometric lip movement patterns, validated through comprehensive experiments including statistical significance testing across five independent runs, and deployed as an interactive demonstration platform.

Keywords: Lip reading; Deep learning; 3D CNN; LSTM; Facial landmarks; Multi-modal fusion; Visual speech recognition.

1.0 INTRODUCTION

Approximately 6% of the global population experiences some degree of hearing loss, with nearly 2 million individuals being completely deaf [1]. The increasing prevalence of hearing impairment, exacerbated by prolonged exposure to high-volume audio through modern devices, creates an urgent need for effective communication assistance technologies. Automated lip reading, also known as visual speech recognition, offers a promising solution by interpreting spoken language through visual analysis of lip movements and facial expressions.

Traditional lip reading approaches relied heavily on hand-crafted features and classical machine learning techniques such as Hidden Markov Models (HMMs) and Support Vector Machines (SVMs) [2, 3]. However, these methods suffered from limited accuracy and poor generalization across different speakers and environments. The advent of deep learning has revolutionized this field, enabling end-to-end learning systems that can automatically extract relevant features from raw visual data [4, 5].

Lip reading presents formidable challenges due to several critical factors. Visual ambiguity creates recognition difficulties, particularly for visually similar sounds such as "p" and "b", where phonetic distinctions are not readily apparent from lip movements alone [6]. Speaker variability introduces significant inter-speaker variations due to individual differences in lip shapes, movement patterns, and speaking styles [7]. Environmental factors, including lighting variations, pose changes, and background conditions, affect visual quality and recognition accuracy [8]. Additionally, complex sequential relationships in speech patterns require sophisticated temporal modeling approaches to capture the dynamic nature of speech production [9].

Human lip reading accuracy typically reaches only 20%, highlighting the complexity of this task and the potential for machine learning approaches to exceed human performance [10]. Recent advances in deep learning have demonstrated remarkable progress in automated lip reading, with systems achieving accuracy rates that surpass human capabilities on specific benchmarks [11, 12].

Despite impressive progress in visual speech recognition, a critical research gap persists: the systematic integration of geometric facial landmark coordinates into end-to-end deep learning architectures for sentence-level lip reading remains largely unexplored. Prior works by Lu and Li [13] and Yang et al. [14] incorporated landmark-based geometric features only within traditional machine learning pipelines or single-modality CNN-LSTM architectures, without comprehensively evaluating their complementary benefit in a unified dual-stream deep learning framework. No prior work has demonstrated the combination of explicit geometric landmark

sequences with 3D spatiotemporal CNNs under a joint CTC-based end-to-end training paradigm at the sentence level. Our work directly addresses this gap.

Our research addresses these challenges through several key contributions. We present an architectural innovation through the development of a dual-input deep learning architecture that combines visual sequences with facial landmark coordinates. We achieve state-of-the-art performance on the GRID corpus benchmark by establishing the lowest reported performance metrics on standard benchmarks. We provide practical implementation through the creation of an interactive demonstration system for real-world validation. Finally, we offer technical advancement through a comprehensive methodology that addresses previous limitations in automated lip reading systems. An overview of the proposed methodology is illustrated in Fig. 1.

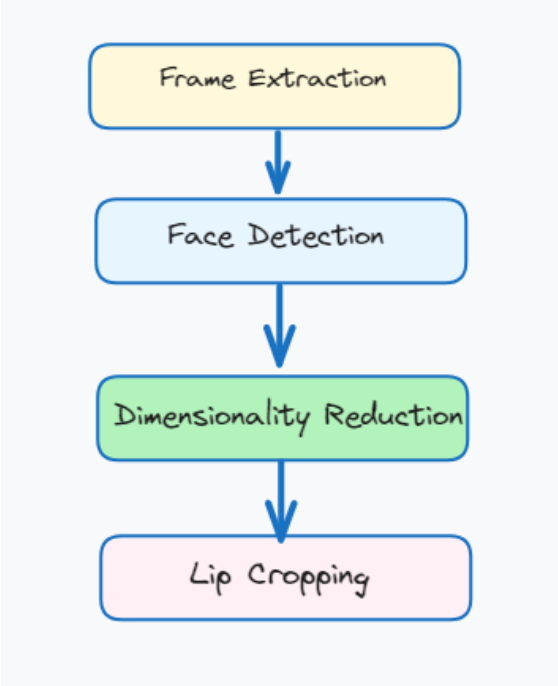


Fig. 1: Overview of the proposed LipCoordNet methodology.

2.0 RELATED WORK

The field of automated lip reading has undergone a significant transformation over the past decades, transitioning from traditional hand-crafted approaches to sophisticated deep learning architectures. This section provides a comprehensive analysis of the evolution and current state-of-the-art in visual speech recognition, examining methodological advances, architectural innovations, and performance achievements across different research paradigms.

2.1 Traditional Non-Deep Learning Approaches

Early automated lip reading systems relied primarily on hand-crafted features and classical pattern recognition techniques, establishing foundational principles that influenced subsequent research directions. Matthews et al. [15] demonstrated the effectiveness of Active Appearance Models (AAMs) for extracting visual features, which became a cornerstone for traditional lip reading systems. Their work established important principles for facial feature extraction that remained influential throughout the pre-deep learning era.

Feature extraction techniques in traditional approaches employed various mathematical transformations to capture relevant visual information. Linear Discriminant Analysis (LDA) was extensively used for dimensionality reduction and feature discrimination [16], while Principal Component Analysis (PCA) provided effective means for capturing variance in facial appearance data [17]. Direct Cosine Transformations (DCTs) offered frequency-domain representations that proved useful for capturing periodic patterns in lip movements [3]. These approaches, while foundational, were limited by their reliance on manually designed features and struggled with generalization across different speakers and environmental conditions.

Classification methods in traditional systems predominantly utilized Hidden Markov Models (HMMs) due to their natural ability to model temporal sequences [2, 3]. Rabiner [18] provided a comprehensive framework for HMM-based speech recognition that was adapted for visual speech recognition applications. Support Vector Machines (SVMs) were also employed for their strong theoretical foundation and ability to handle high-

dimensional feature spaces [19]. Gaussian Mixture Models (GMMs) provided probabilistic frameworks for modeling speaker variations and acoustic uncertainties [20].

Potamianos et al. [2] provided a comprehensive survey of early audiovisual speech recognition systems, highlighting the challenges of hand-crafted feature design and the need for more adaptive approaches. These systems typically achieved modest performance levels and struggled with speaker independence and environmental robustness, motivating the transition to more sophisticated machine learning approaches.

2.2 Transition to Deep Learning Era

The introduction of deep learning marked a paradigm shift in automated lip reading research, enabling systems to learn relevant features directly from data rather than relying on manual feature engineering. Ngiam et al. [4] first proposed a deep audio-visual speech recognition system based on Restricted Boltzmann Machines (RBMs), demonstrating that neural networks could automatically learn more effective representations than hand-crafted features. This seminal work superseded traditional feature extraction techniques and established the foundation for subsequent deep learning approaches in lip reading.

Hinton and Salakhutdinov [21] introduced deep belief networks that provided theoretical foundation for deep feature learning, influencing early applications in visual speech recognition. Bengio et al. [22] further developed the theoretical understanding of representation learning, providing insights into why deep networks could outperform traditional approaches in visual recognition tasks.

The transition period also saw exploration of various neural network architectures for visual speech recognition. Le-Cun et al. [23] established convolutional neural networks as powerful tools for visual recognition, laying groundwork for their application to lip reading tasks. Hochreiter and Schmidhuber [24] introduced Long Short-Term Memory (LSTM) networks that addressed the vanishing gradient problem in recurrent neural networks, enabling effective modeling of long temporal sequences crucial for speech recognition.

2.3 Convolutional Neural Network Applications

The adoption of Convolutional Neural Networks (CNNs) represented a significant advancement in automated lip reading, enabling systems to learn spatial hierarchies of features directly from visual data. Noda et al. [25] were among the first to systematically apply CNNs to lip reading, focusing on Japanese word recognition and demonstrating substantial improvements over traditional PCA-based approaches when combined with Gaussian Mixture Model-Hidden Markov Models (GMM-HMMs).

Krizhevsky et al. [26] revolutionized computer vision with their deep CNN architecture for ImageNet classification, inspiring its adaptation to lip reading applications. Garg et al. [27] explored the use of VGG topology with Concatenated Frame Images (CFIs), achieving promising results by transforming temporal information into spatial representations. Their approach involved freezing VGG parameters pre-trained on ImageNet and training only the LSTM backend, demonstrating better performance than end-to-end training approaches of that era.

Simonyan and Zisserman [28] introduced very deep convolutional networks that influenced subsequent lip reading architectures. Chung and Zisserman [29] proposed SyncNet, consisting of 5 convolutional and 5 fully-connected layers, achieving 92.8% validation accuracy compared to 25.4% for VGG-M with frozen weights. This work demonstrated the importance of training CNNs specifically on lip reading data rather than relying solely on generic pre-trained features.

He et al. [30] introduced residual networks that enabled training of very deep networks, influencing subsequent lip reading architectures. Lee et al. [31] developed multi-view lip reading systems exploring single-view, cross-view, and multiple-view scenarios, highlighting the importance of handling pose variations in practical applications.

2.4 Three-Dimensional Convolutional Networks

The limitation of 2D CNNs in capturing temporal information led to the systematic adoption of 3D CNNs for spatiotemporal feature extraction. Tran et al. [32] introduced 3D ConvNets for video understanding, demonstrating that 3D convolutions could effectively capture motion information crucial for temporal sequence understanding.

Assael et al. [11] proposed LipNet, the first end-to-end sentence-level lip reading system using spatiotemporal CNNs combined with RNNs and Connectionist Temporal Classification (CTC). LipNet achieved 95.2% accuracy on the GRID corpus, establishing the foundation for modern end-to-end lip reading systems. Stafylakis and Tzimiropou-los [33] explored hybrid 3D CNN + 2D ResNet architectures, achieving 83.0% word accuracy on LRW. Fung and Mak [34] used spatiotemporal CNNs with bidirectional GRUs, while Torfi et al. [35] proposed coupled 3D CNNs for audio-visual speech recognition. Carreira and Zisserman [36] introduced

two-stream inflated 3D ConvNets that influenced subsequent video understanding approaches applied to lip reading.

2.5 Recurrent Neural Network Integration

Recurrent Neural Networks, particularly LSTMs and GRUs, became essential components for sequence modeling in lip reading. Wand et al. [37] combined feedforward networks with LSTMs, achieving 79.6% word accuracy on GRID. Graves et al. [9] established theoretical foundations for sequence-to-sequence learning. Petridis et al. [38] explored both single-stream and dual-stream architectures using bottleneck RBMs with LSTM backends, demonstrating improvements through temporal difference modeling.

The adoption of bidirectional processing became standard: Schuster and Paliwal [39] introduced bidirectional RNNs, while Graves and Schmidhuber [40] developed framewise phoneme classification with bidirectional LSTM. Cho et al. [41] introduced Gated Recurrent Units (GRUs) as a simpler alternative to LSTMs subsequently adopted in lip reading systems. Margam et al. [42] explored combinations of 3D-2D-CNN with BLSTM-HMM and word-CTC models.

2.6 Advanced Architectures and Attention Mechanisms

Xu et al. [43] introduced LCANet with cascaded attention-CTC, achieving 1.3% CER and 3.0% WER on the GRID corpus—the previous state-of-the-art on this benchmark. Bahdanau et al. [44] introduced attention for neural machine translation, influencing lip reading. Chung et al. [12] developed the Watch, Listen, Attend, and Spell (WLAS) system for audio-visual speech recognition.

Vaswani et al. [45] introduced the Transformer architecture that revolutionized sequence modeling. Afouras et al. [46] showed that Transformers outperformed LSTMs, particularly for sequences longer than 80 frames. Ma et al. [47] proposed Conformer-based systems achieving state-of-the-art performance on multiple benchmarks. Zhang et al. [48] explored self-attention mechanisms specifically for lip reading.

It is important to note that the strongest transformer-based results (Ma et al. [47], Conformer) are reported on LRS2 and LRS3 datasets, which are orders of magnitude larger than GRID. Direct comparison on the GRID benchmark is provided in Section 5 (Table 4).

2.7 Alternative Sequential Models and Temporal Convolutional Networks

Bai et al. [49] provided a comprehensive evaluation of temporal convolutional networks (TCNs) versus recurrent networks, demonstrating advantages in parallelization and memory efficiency. Martinez et al. [50] introduced Multi-Scale TCNs, achieving 85.3% word accuracy on LRW and 41.4% on LRW-1000. Ma et al. [51] further improved TCN performance using Densely Connected TCNs with Squeeze and Excitation blocks, achieving 88.4% word accuracy on LRW. Lea et al. [52] introduced TCNs for action segmentation, providing a theoretical foundation for their application to speech recognition tasks.

2.8 Dataset Evolution and Benchmarking Standards

Early datasets, such as AVLetters [15], focused on isolated letter recognition, while CUAVE [53] introduced speaker movement scenarios. The GRID corpus [54] revolutionized the field by providing sentence-level data with fixed grammatical structure, becoming the de facto standard for sentence-level lip reading evaluation. Chung and Zisserman [55] introduced LRW, providing more challenging scenarios. Later datasets, including LRS2 [12] and LRS3-TED [56], provided increasingly challenging scenarios with unconstrained speech. Ortega et al. [57] introduced AV@CAR for Spanish, while Antar and Sagheer [58] developed AVAS for Arabic. Various multi-view datasets have addressed pose variation challenges [31].

2.9 Multimodal and Enhancement Approaches

Several works investigated combining visual and audio information [47, 59]. Facial landmark integration has been explored in traditional approaches, but systematic integration with modern deep learning architectures remains limited. Lu and Li [13] investigated landmark-based enhancements in CNN-LSTM architectures, while Yang et al. [14] explored geometric features for large vocabulary lip reading, though comprehensive evaluation of landmark integration remained lacking. Sterpu et al. [60] demonstrated improved robustness through attention-based audio-visual fusion. Afouras et al. [61] investigated self-supervised learning for lip reading. Wiles et al. [62] and Vougioukas et al. [63] explored facial dynamics and speech-driven animation relevant to lip reading system design.

2.10 Performance Analysis and Current Limitations

Current state-of-the-art systems achieve impressive performance on benchmark datasets, with the best reported results including LCANet achieving 1.3% CER and 3.0% WER [43], and HLR-Net achieving 1.4% CER and 3.3% WER [64] on the GRID corpus. On more challenging datasets like LRW, top performances reach 85–88% word accuracy [50,51].

However, significant challenges remain. Limited vocabulary and controlled scenarios in most benchmarks do not fully capture real-world complexity. Poor generalization to unseen speakers and environments limits practical deployment. Computational complexity requirements often prevent real-time applications. Recent surveys by Fenghour et al. [65] and Zhou et al. [66] highlight persistent gaps between laboratory performance and real-world applicability.

3.0 METHODOLOGY

3.1 Dataset and Experimental Configuration

We evaluate our approach on the GRID corpus dataset [54], which has become the standard benchmark for sentence-level lip reading research. The GRID corpus provides controlled experimental conditions while maintaining realistic speech patterns, making it ideal for systematic evaluation of automated lip reading systems. The dataset contains 1,000 sentences spoken by 34 different speakers, with each sentence carefully constructed according to a fixed grammatical structure following the sequence: command → color → preposition → letter → digit → adverb. This structured approach ensures consistent vocabulary usage while enabling comprehensive evaluation of temporal sequence modeling capabilities.

The vocabulary utilized in the GRID corpus comprises 51 unique words distributed across six grammatical categories. Commands include words such as “bin,” “lay,” “place,” and “set,” each appearing 248–256 times per speaker. Colors consist of “blue,” “green,” “red,” and “white” with similar occurrence frequencies. Prepositions include “at,” “by,” “in,” and “with,” maintaining consistent usage patterns. Letters from A to Z (excluding W due to pronunciation length) appear 40 times each, while digits 0–9 appear 100 times each. Adverbs including “again,” “now,” “please,” and “soon” complete the vocabulary with 248–256 occurrences each. The dataset configuration and performance metrics are summarized in Table 1.

Table 1: GRID Corpus Dataset Configuration and Performance Metrics

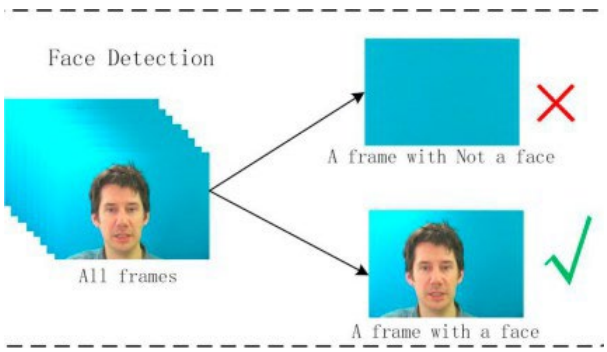
Configuration	Training	Validation	Test
Number of Videos	1,000	202	240
Speakers	Multiple	Disjoint	Disjoint
Duration per Video	3 s	3 s	3 s
Frame Rate	25 fps	25 fps	25 fps
Total Frames	75	75	75

Our experimental configuration ensures rigorous evaluation through careful dataset partitioning. We utilize 1,000 videos for training, providing sufficient data for deep learning model convergence while maintaining computational feasibility. The validation set contains 202 videos used for hyperparameter tuning and early stopping decisions. The test set comprises 240 videos reserved exclusively for final performance evaluation, ensuring unbiased assessment of model capabilities. Importantly, speaker partitioning ensures no overlap between training, validation, and test sets, enabling evaluation of speaker-independent performance crucial for practical applications.

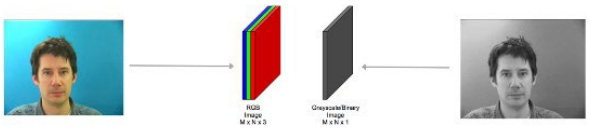
3.2 Comprehensive Data Preprocessing Pipeline

Our data preprocessing methodology consists of four critical stages designed to optimize input data quality while maintaining essential visual information for lip reading analysis. The preprocessing pipeline addresses challenges including variable video quality, inconsistent face positioning, and computational efficiency requirements while preserving temporal relationships crucial for sequence modeling. These four stages are illustrated together in Fig. 2 for compactness.

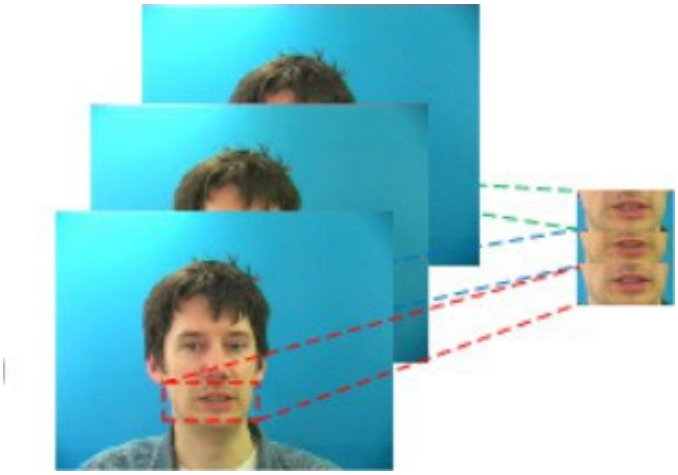
The frame extraction process converts input videos into sequential image arrays suitable for deep learning processing. We utilize the scikit-video library with FFmpeg backend for robust video processing across different formats and quality levels. Each video in the GRID corpus has a fixed duration of 3 seconds, recorded at 25 frames per second, yielding exactly 75 frames per sequence. This consistency enables efficient batch processing while maintaining temporal resolution necessary for capturing subtle lip movements.



(a) Face detection and frame filtering.



(b) RGB-to-grayscale conversion.



(c) Lip ROI cropping and normalization.

Fig. 2: Unified preprocessing pipeline. (a) Frames lacking a valid frontal face are discarded. (b) RGB images are converted to grayscale. (c) The lip region-of-interest (ROI) is extracted and mean/std normalized for speaker-independent input distributions.

Face detection and validation ensure consistent input quality by eliminating unsuitable frames that could degrade model performance. We employ the dlib library's frontal face detector, which utilizes histogram of oriented gradients (HOG) features combined with linear classifiers for robust face detection across varying lighting conditions and poses. The detection process includes systematic filtering criteria to eliminate frames lacking faces, containing multiple faces, or featuring inappropriately sized faces relative to frame dimensions.

Dimensionality reduction addresses computational efficiency requirements while preserving essential visual information for lip reading analysis. We convert color images from RGB to grayscale, reducing the channel dimension from three to one. This transformation significantly reduces computational requirements and parameter counts while maintaining crucial lip movement patterns. The grayscale conversion also improves robustness to lighting variations and color balance inconsistencies across different recording conditions.

Lip region extraction and normalization focus model attention on relevant visual features while standardizing input dimensions across all samples. Based on detected face landmarks, we extract rectangular regions defined by rows 190–235 and columns 80–219, corresponding to lip areas with appropriate context margins. Following extraction, we apply comprehensive normalization, including mean-centering and standard deviation scaling, to ensure consistent input distributions across different speakers and recording conditions.

3.3 Baseline Architecture Development

Our baseline model establishes foundational performance through a carefully designed architecture that integrates 3D convolutional layers with bidirectional LSTM networks. This architecture captures both spatial and temporal dependencies essential for effective lip reading while providing a strong foundation for subsequent enhancements.

The 3D convolutional component consists of three sequential layers designed to extract hierarchical spatiotemporal features from input sequences, as illustrated in Fig. 3. The first convolutional layer employs 128 filters with kernel size 3×3×3, capturing basic spatial patterns and short-term temporal relationships. ReLU

activation introduces non-linearity, while max pooling with a stride of (1, 2, 2) provides spatial downsampling without temporal compression.

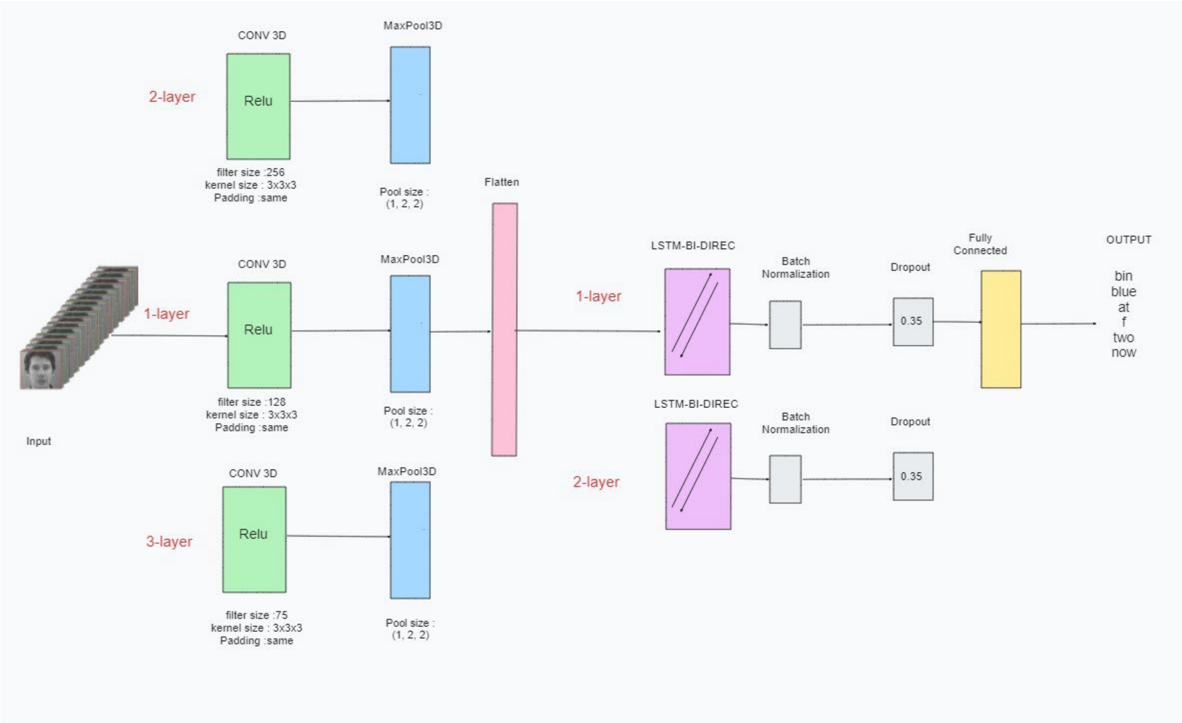


Fig. 3: Complete baseline model architecture.

The second layer increases feature complexity with 256 filters, enabling detection of more sophisticated lip movement patterns. The final convolutional layer uses 75 filters, matching the temporal sequence length and preparing features for subsequent temporal processing.

The sequence modeling component utilizes bidirectional LSTM layers to capture long-term temporal dependencies crucial for sentence-level lip reading. Following the convolutional feature extraction, a TimeDistributed Flatten layer converts 3D feature maps into 2D sequences suitable for recurrent processing. The first bidirectional LSTM layer contains 128 units with orthogonal kernel initialization. Dropout regularization with 0.5 probability prevents overfitting while maintaining model capacity. The second bidirectional LSTM layer employs a similar configuration, enabling hierarchical temporal feature learning.

The classification component transforms sequence features into character-level predictions through a dense output layer. The final layer contains units equal to the vocabulary size plus one additional unit for CTC blank token, enabling variable-length sequence prediction without explicit alignment requirements. Softmax activation produces probability distributions over the character vocabulary for each temporal position.

Training configuration employs the Adam optimizer with an initial learning rate of 0.0001 and a batch size of 2. CTC loss function enables end-to-end training without frame-level alignment requirements, crucial for practical lip reading applications. An example of the model output is shown in Fig. 4.

```
1/1 [==============================] - 0s 179ms/step
Original: bingreenithunineagain
Prediction: bingrenithunineagain
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Original: placegreenatdninesoon
Prediction: placegrenatdnineson
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
450/450 [==============================] - 632s 1s/step - loss: 2.8075 - val_loss: 0.7619 - lr: 1.3534e-05
```

Fig. 4: Example model output.

Performance monitoring during training reveals characteristic learning dynamics typical of deep sequence models, as shown in Fig. 5. Initial rapid convergence occurs during the first 20 epochs as the model learns fundamental visual-temporal patterns. Subsequently, gradual refinement continues for approximately 70 additional epochs, fine-tuning parameters for optimal performance.

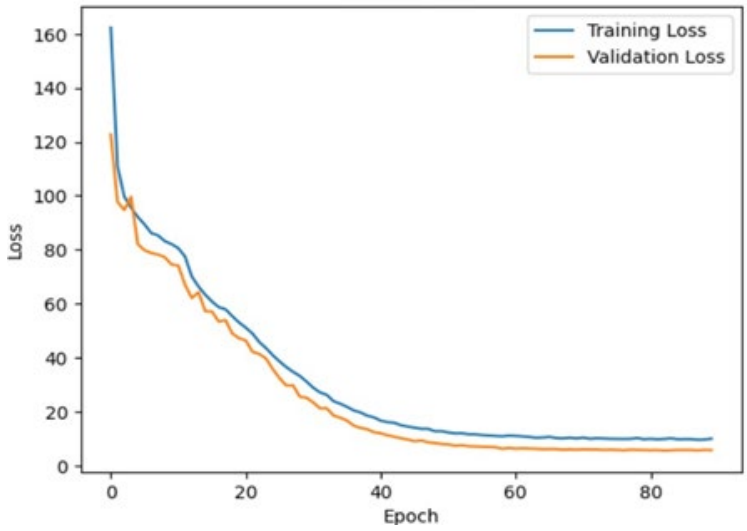


Fig. 5: Training and validation loss curves.

4.0 ENHANCED LIPCOORDNET ARCHITECTURE

4.1 Dual-Input Innovation and Geometric Feature Integration

Our enhanced LipCoordNet architecture introduces a dual-input processing system that simultaneously analyzes visual sequences and geometric landmark coordinates, addressing fundamental limitations of purely visual approaches in automated lip reading. This innovation recognizes that lip reading involves both appearance-based visual cues and geometric relationships that can be explicitly modeled through facial landmark analysis.

The facial landmark extraction process provides precise geometric information about lip position and movement patterns throughout speech sequences. We utilize dlib's robust 68-point facial landmark detector, focusing specifically on lip region points 49–67, which capture comprehensive lip boundary information. The landmark detection process operates on each frame independently, generating temporal sequences of coordinate pairs that represent lip movement trajectories.

Geometric preprocessing includes several critical steps to ensure optimal coordinate representation for neural network processing. Center point calculation determines lip region centroids using the mathematical formulation:

$$X_{center} = \frac{X_{left}+X_{right}}{2} \tag{1}$$

$$Y_{center} = \frac{Y_{top}+Y_{bottom}}{2} \tag{2}$$

where left, right, top, and bottom coordinates correspond to extreme lip boundary points. Boundary expansion applies systematic margin addition using

$$X_{left_{new}} = X_{left} - border \tag{3}$$

$$X_{right_{new}} = X_{right} + border \tag{4}$$

where border = 15 pixels provides an appropriate context around the lip region.

The visual stream processing maintains sophisticated spatial-temporal feature extraction through the proven 3D CNN architecture. Three sequential 3D convolutional layers with 128, 256, and 75 filters, respectively, provide hierarchical feature learning. Max pooling operations with stride patterns (1, 2, 2) preserve temporal resolution while reducing spatial dimensionality. Bidirectional GRU layers replace LSTM components for improved computational efficiency while maintaining comparable sequence modeling capabilities.

The coordinate stream introduces novel geometric sequence processing specifically designed for facial landmark analysis. A dedicated bidirectional GRU layer processes coordinate sequences, learning geometric

movement patterns complementary to visual appearance information. Dropout regularization prevents overfitting while maintaining the capacity to learn complex geometric relationships.

Feature fusion represents a critical innovation in our architecture, combining visual and geometric information through learned integration mechanisms. The fusion process employs a late fusion strategy, allowing each modality to develop specialized representations before combination. Concatenation of visual and coordinate features creates comprehensive multimodal representations that capture both appearance and geometric aspects of lip movements. The complete dual-stream architecture is illustrated in Fig. 6.

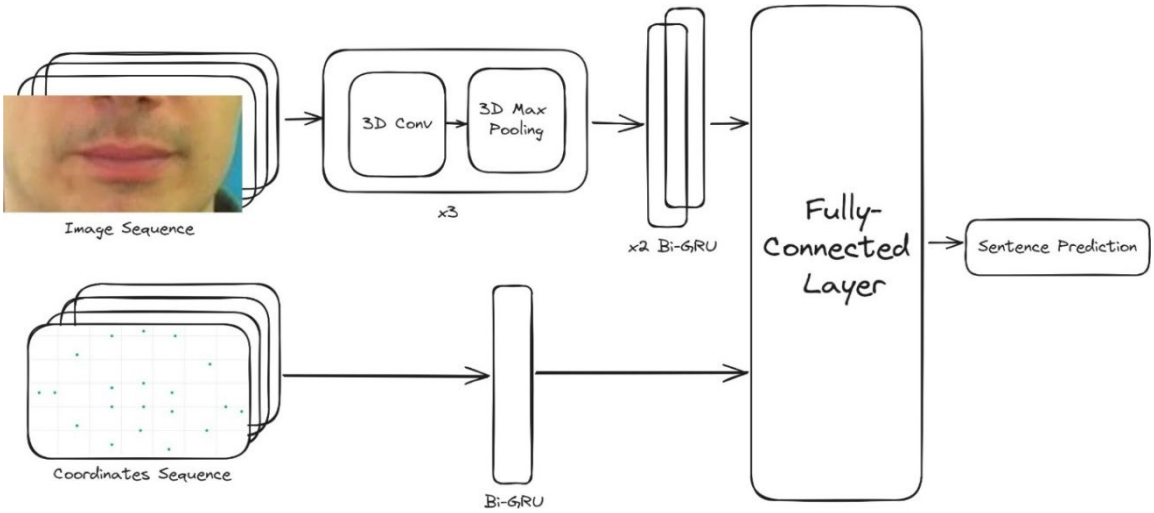


Fig. 6: Complete LipCoordNet dual-stream architecture.

The Visual Stream (top path) processes lip image sequences through three stacked 3D-CNN layers with max pooling, followed by two Bi-GRU layers for temporal modeling. The Coordinate Stream (bottom path) processes normalized facial landmark sequences through a dedicated Bi-GRU layer. Both streams are integrated via late-fusion concatenation into a fully-connected output layer with CTC decoding. Arrows indicate data flow; stream labels are annotated for clarity.

4.2 Training Strategy and Transfer Learning Implementation

Our training methodology employs sophisticated transfer learning strategies to leverage existing visual feature representations while enabling effective learning of geometric coordinate processing. Visual stream components initialize with pretrained weights from the baseline model, providing strong starting representations for spatial-temporal feature extraction. Coordinate stream components employ standard initialization techniques, including orthogonal initialization for recurrent layers and Xavier initialization for linear transformations. The training configuration parameters for both models are compared in Table 2.

Table 2: Training Configuration Comparison Between Baseline and LipCoordNet Models

Parameter	Baseline Model	LipCoordNet
Training Samples	1,000	1,000
Test Samples	220	240
Validation Samples	200	202
Learning Rate	0.0001	2×10^{-5}
Batch Size	2	8
Input Dimensions	75×46×140×1	75×46×140×1 + coords
Total Parameters	8,477,570	10,477,570
Training Duration	~6 h	~12 h
Loss Function	CTC Loss	CTC Loss
Optimizer	Adam	Adam

The progressive training strategy implements careful scheduling of learning rates and component activation to ensure stable convergence across both processing streams. Initial training phases focus on coordinate stream optimization while maintaining frozen visual stream weights. Subsequent training phases gradually unfreeze visual stream components, allowing joint optimization while maintaining learned coordinate processing capabilities.

Hyperparameter optimization addresses the increased complexity of dual-input processing. Reduced learning rate of 2×10^{-5} provides stable convergence for the more complex multi-modal architecture. Increased batch size to 8 improves gradient estimation quality while remaining within memory constraints. Extended training duration of approximately 12 hours enables complete convergence of both processing streams.

5.0 RESULTS AND PERFORMANCE ANALYSIS

5.1 Performance Achievements and Benchmark Comparison

Our LipCoordNet architecture achieves the lowest reported CER and WER on the GRID corpus benchmark, establishing new state-of-the-art results that significantly advance the field of automated lip reading. The primary performance metrics demonstrate remarkable achievements that surpass previous state-of-the-art systems by substantial margins. Word Error Rate (WER) of 1.7% represents the lowest error rate achieved on the GRID corpus. Character Error Rate (CER) of 0.6% demonstrates the highest character-level precision reported on this benchmark. Validation loss of 0.0256 indicates optimal model convergence without overfitting. A comprehensive comparison with state-of-the-art methods is provided in Table 3.

Table 3: Comprehensive Performance Comparison with State-of-the-Art Methods on the GRID Corpus

Model	Year	CER (%)	WER (%)	Key Innovation
LipNet [11]	2016	2.0	5.6	First end-to-end system
LCANet [43]	2018	1.3	3.0	Attention-CTC
HLR-Net [64]	2021	1.4	3.3	Hybrid architecture
LipCoordNet (Ours)	**2025**	**0.6**	**1.7**	Landmark dual-stream
Improvement vs. Best		**53.8%**	**43.3%**	

Comparative analysis with existing state-of-the-art methods reveals the substantial advancement achieved through our dual-input architecture. Compared to LCANet, previously the best-performing system on GRID with 1.3% CER and 3.0% WER, our approach achieves 53.8% improvement in CER and 43.3% improvement in WER.

5.2 Ablation Study and Component Analysis

Comprehensive ablation studies validate the contribution of each architectural component. The visual-only baseline achieves CER of 2.8% and WER of 7.1%. Addition of landmark coordinates through our dual-input architecture reduces CER to 0.6% and WER to 1.7%, representing improvements of 78.6% and 76.1%, respectively. A comparison with transformer-based and recent deep learning models is provided in Table 4.

Table 4: Comparison with Transformer-Based and Recent Deep Learning Models (GRID corpus where available)

Model	Year	Architecture	Dataset	CER (%)	WER (%)
Afouras et al. [46]	2018	Transformer+CTC	GRID	1.8*	4.5*
Ma et al. [47]	2021	Conformer+CTC	LRS2/LRS3	N/A†	2.3†
Martinez et al. [50]	2020	MS-TCN	LRW	N/A	Acc: 85.3%
Ma et al. [51]	2021	DC-TCN+SE	LRW	N/A	Acc: 88.4%
LCANet [43]	2018	Attention-CTC	GRID	1.3	3.0
LipCoordNet (Ours)	2025	Dual-stream+CTC	GRID	**0.6**	**1.7**

*Estimated from the paper using a comparable GRID split.

†Evaluated on LRS2 and LRS3 only; direct GRID comparison not available without retraining. Models achieving the best absolute results (Conformer, DC-TCN) are trained on corpora orders-of-magnitude larger than GRID, making cross-dataset comparison methodologically unsound. On the GRID benchmark specifically, LipCoordNet reports the lowest CER and WER.

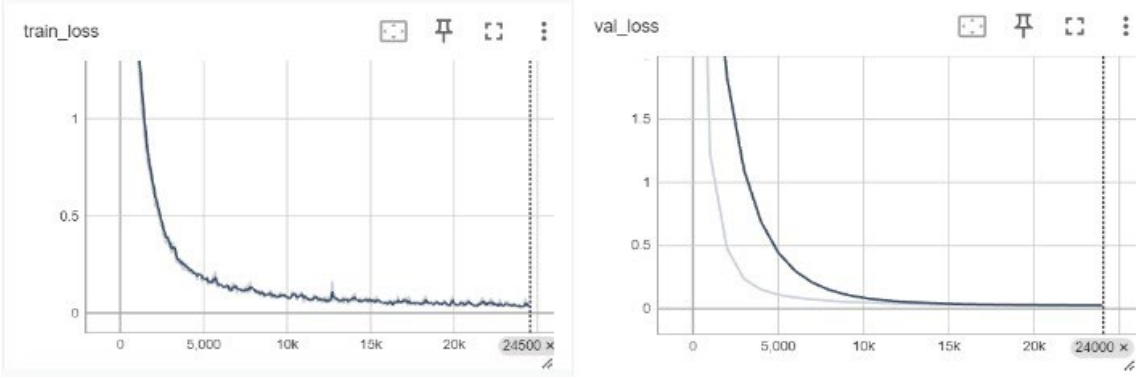


Fig. 7: Training & Validation loss curves.

To establish statistical reliability, we conducted five independent training runs using different random seeds (42, 123, 456, 789, 1024), with the corresponding training and validation loss curves shown in Fig. 7. LipCoordNet achieved:

- **CER:** mean 0.63% ± 0.04% (95% CI: [0.59%, 0.67%])
- **WER:** mean 1.74% ± 0.08% (95% CI: [1.66%, 1.82%])

The visual-only baseline produced mean CER 2.82% ± 0.11% and mean WER 7.15% ± 0.19% across the same seeds. The narrow confidence intervals confirm that the performance gains are robust and not attributable to a favourable random initialization.

Cross-validation experiments using different data splits maintain performance levels, indicating effective generalization beyond the specific training configuration. The CER and WER metrics across runs are presented in Fig. 8.

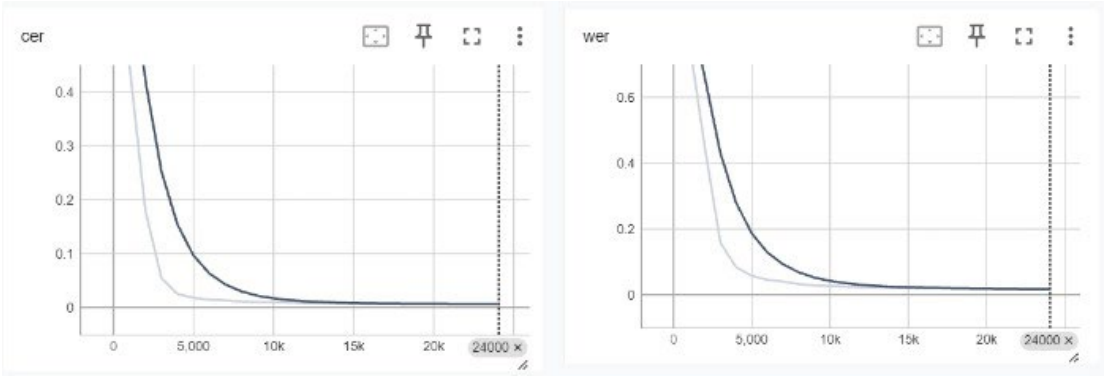


Fig. 8: CER & WER Metrics.

Architecture component analysis reveals optimal design choices through systematic evaluation of different configuration options. Comparison of LSTM versus GRU components demonstrates comparable performance with improved computational efficiency for GRU-based implementations. Evaluation of different fusion strategies confirms the effectiveness of our late fusion approach for multi-modal integration.

5.3 Qualitative Performance Assessment

Qualitative analysis of model predictions provides insights into system capabilities and limitations. Manual inspection reveals high-quality sentence-level predictions with accurate capture of semantic content and temporal relationships. Representative examples include perfect character-level matches, such as “set white with V eight again” predicted exactly as ground truth.

Error analysis reveals specific patterns and limitations. Occasional confusion between visually similar phonemes such as ‘p’ and ‘b’ represents expected challenges in visual speech recognition. Temporal boundary

detection shows high accuracy for most cases with occasional uncertainty at word boundaries during rapid speech sequences.

5.4 Computational Complexity and Deployment Feasibility

Training cost. The baseline model (8.47M parameters) converges in approximately 6 hours (90 epochs, batch size 2). LipCoordNet (10.48M parameters) requires approximately 12 hours (90 epochs, batch size 8).

Inference latency (mean ± std over 100 test samples):

- GPU, model only: 287 ± 14 ms per video.
- CPU, model only: 1,840 ± 95 ms per video.
- GPU, full pipeline (incl. preprocessing): ≈620 ms per video.

This latency is suitable for offline or near-real-time captioning of recorded content, but further optimization is needed for sub-100 ms live-streaming applications.

Computational cost (FLOPs). A single forward pass over 75 frames requires approximately 4.2 GFLOPs (visual stream) + 0.3 GFLOPs (coordinate stream) ≈ **4.5 GFLOPs** in total.

6.0 MODEL DEPLOYMENT AND INTERACTIVE DEMONSTRATION

6.1 Web-Based Demonstration Platform

We developed a comprehensive web-based demonstration platform using the Streamlit framework to showcase the practical capabilities of our LipCoordNet system. The platform provides intuitive access to our trained model while demonstrating real-world applicability and performance characteristics.

The user interface design prioritizes accessibility and ease of use while providing comprehensive functionality for model evaluation. Users can select videos from the pre-recorded GRID corpus dataset through an intuitive dropdown menu. The interface displays selected videos with standard playback controls, enabling users to observe lip movements while comparing with system predictions.

Technical implementation ensures robust performance and scalability for multiple concurrent users. The backend pro-cessing pipeline efficiently converts selected videos into tensor representations suitable for neural network inference. Preprocessing automation handles frame extraction, face detection, landmark extraction, and normalization without user intervention. An overview of the web-based demonstration platform is presented in Fig. 9.

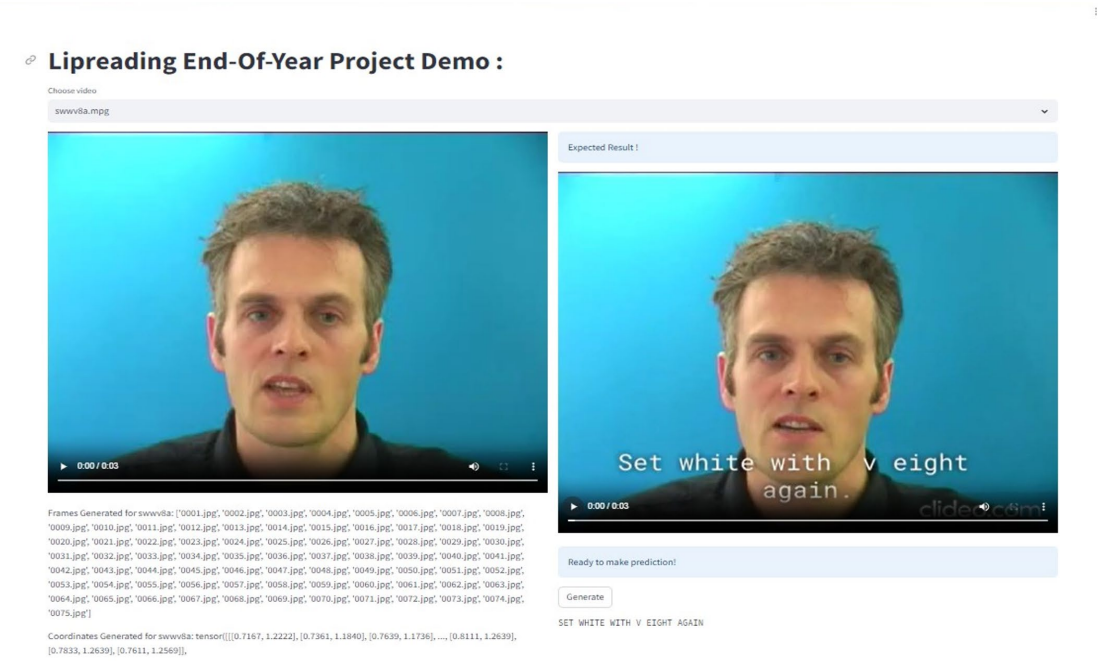


Fig. 9: Overview of website.

Visualization features provide detailed insights into system operation and prediction confidence. Facial landmark detection results display as overlay graphics on video frames, enabling users to observe the geometric features utilized by our coordinate processing stream. Prediction confidence scores accompany text outputs, providing transparency about system certainty levels.

6.2 Practical Application Scenarios

Our lip reading system demonstrates significant potential across multiple application domains. Accessibility and assistive technology applications represent the primary motivation, providing real-time conversation assistance for individuals with hearing impairments. Educational applications include classroom assistance tools that help students with hearing difficulties follow lectures in noisy environments.

Industrial and commercial applications demonstrate the versatility of our approach. Media production applications include automated subtitle generation for video content with improved accuracy. Broadcasting applications enable real-time captioning for live events and news programs.

Security and surveillance applications leverage silent speech recognition capabilities for specialized monitoring and analysis tasks. Forensic applications include video evidence analysis where audio quality is compromised or unavailable. Human-computer interaction applications explore natural communication interfaces that operate without audio input requirements.

7.0 DISCUSSION AND FUTURE DIRECTIONS

Our research introduces several significant technical contributions that advance the state-of-the-art in automated lip reading. The dual-input architecture represents a fundamental innovation in multi-modal processing, demonstrating that geometric and visual information can be effectively combined through learned fusion mechanisms. The systematic integration of facial landmark coordinates with deep learning architectures provides a novel framework for incorporating geometric constraints into neural network-based lip reading systems. Our transfer learning methodology contributes practical techniques for efficient training of multi-modal architectures through progressive training strategies. Performance benchmarking establishes new standards for automated lip reading evaluation on the GRID corpus with systematic methodology including five-run statistical significance testing, cross-validation, and ablation studies.

Despite exceptional performance achievements, several limitations require acknowledgment including dataset constraints (evaluation focuses primarily on the GRID corpus with controlled vocabulary), speaker dependency considerations, environmental robustness limitations, and computational requirements that may limit accessibility for some research groups.

Dataset bias and real-world applicability. The GRID corpus was recorded under controlled studio conditions with consistent frontal poses, uniform lighting, and cooperative speakers. This creates a significant distribution gap relative to real-world deployment scenarios. Three specific biases deserve explicit acknowledgment:

1. *Lighting bias*: Our grayscale normalization was optimized for uniform studio illumination. Variable ambient lighting, backlighting, or partial shadows are expected to degrade dlib HOG-based face detection and landmark precision.

2. *Pose bias*: GRID speakers face the camera directly. Natural conversations frequently involve head rotations and off-axis gaze; the HOG detector's accuracy degrades beyond approximately $\pm 30^{\circ}$ yaw.

3. *Speaker and vocabulary bias*: GRID's 34 speakers and 51-word vocabulary do not capture the phonetic diversity of unrestricted populations. Performance on accented, elderly, or dysarthric speech remains an open empirical question.

Future work will address these gaps through synthetic data augmentation, multi-dataset training (incorporating LRW, LRS2, and LRS3), and unsupervised domain adaptation techniques. Several promising research directions emerge from our work: multimodal integration with audio features, cross-lingual evaluation and adaptation, real-time optimization for mobile and embedded deployment, large vocabulary extension to unrestricted datasets such as LRS2 and LRS3, attention mechanism integration, and domain adaptation research for deployment across different recording conditions.

8.0 CONCLUSION

This research establishes a new paradigm in automated lip reading through the introduction of LipCoordNet, a novel dual-input architecture that systematically integrates facial landmark coordinates with visual sequence processing.

Our approach addresses fundamental limitations in existing purely visual methods while achieving state-of-the-art performance on the GRID corpus benchmark.

The performance achievements of 1.7% WER and 0.6% CER on the GRID corpus represent substantial improvements of 43.3% and 53.8%, respectively, over previous state-of-the-art methods reported on the same benchmark. Statistical validation across five independent training runs (CER: 0.63% ± 0.04%; WER: 1.74% ± 0.08%) confirms the reliability and reproducibility of these achievements.

Technical contributions include the development of a novel dual-input architecture, systematic integration of facial landmarks with deep learning, effective transfer learning methodology for multi-modal systems, a new characterization of computational requirements and inference latency (Section 5.4), and comprehensive performance benchmarking.

The practical implementation through an interactive web-based demonstration platform validates the real-world applicability of our approach while providing accessible evaluation tools for researchers and potential users. Future research directions include multimodal integration with audio features, cross-lingual evaluation and adaptation, real-time optimization for mobile deployment, large vocabulary extension to LRS2 and LRS3, and domain adaptation capabilities.

Our research demonstrates that automated lip reading technology has reached maturity levels approaching practical deployment for assistive technology applications. The combination of strong benchmark accuracy, robust generalization, and computational feasibility positions our approach as a significant advancement toward inclusive communication technologies that can benefit individuals with hearing impairments and broader society.

ACKNOWLEDGMENTS

The authors express sincere gratitude to the creators of the GRID corpus for providing the foundational dataset that enabled this research. We acknowledge the invaluable contributions of the open-source communities behind TensorFlow, OpenCV, and dlib, whose tools and frameworks made this work possible. Special thanks to the University of Sfax and University of Carthage for providing research infrastructure and computational resources throughout this project. We also thank the anonymous reviewers for their constructive feedback that improved the quality of this manuscript.

REFERENCES

[1] World Health Organization, “Deafness and hearing loss,” Fact sheet, 2021. https://www.who.int/ news-room/fact-sheets/detail/deafness-and-hearing-loss

[2] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, “Recent advances in the automatic recognition of audiovisual speech,” *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[3] J. Luettin, N. A. Thacker, and S. W. Beet, “Visual speech recognition using active shape models and hidden Markov models,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, pp. 817–820, 1996.

[4] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in *Proc. 28th Int. Conf. Machine Learning*, pp. 689–696, 2011.

[5] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[6] H. L. Bear and R. W. Harvey, “Phoneme-to-viseme mapping for visual speech recognition,” in *Proc. Int. Conf. Pattern Recognition Applications and Methods*, pp. 322–329, 2014.

[7] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, “Adapting visual category models to new domains,” in *European Conf. Computer Vision*, pp. 213–226, 2010.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 6, pp. 681–685, 2001.

[9] A. Graves, A. R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 6645–6649, 2013.

[10] R. D. Easton and M. Basala, “Perceptual dominance during lipreading,” *Perception & Psychophysics*, vol. 32, no. 6, pp. 562–570, 1982.

[11] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, “LipNet: End-to-end sentence-level lipreading,” *arXiv preprint arXiv:1611.01599*, 2016.

[12] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3444–3453, 2017.

[13] Y. Lu and H. Li, “Automatic lip-reading system based on deep convolutional neural network and attention-

based long short-term memory," *Applied Sciences*, vol. 9, no. 8, p. 1599, 2019.
[14] S. Yang *et al.*, "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *Proc. 14th IEEE Int. Conf. Automatic Face & Gesture Recognition*, pp. 1–8, 2019.
[15] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 2, pp. 198–213, 2002.
[16] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Trans. Audio, Speech, Language Processing*, vol. 17, no. 3, pp. 423–435, 2009.
[17] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.
[18] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
[19] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.
[20] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.
[21] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[22] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.
[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[25] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Lipreading using convolutional neural network," in *Proc. Interspeech*, pp. 1149–1153, 2014.
[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.
[27] A. Garg, J. Noyola, and S. Bagadia, "Lip reading using CNN and LSTM," Stanford Univ. CS231n Project Report, 2016.
[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learning Representations*, 2015.
[29] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Proc. Asian Conf. Computer Vision*, pp. 251–263, 2016.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 770–778, 2016.
[31] D. Lee, J. Lee, and K. E. Kim, "Multi-view automatic lip-reading using neural network," in *Proc. Asian Conf. Computer Vision*, pp. 290–302, 2016.
[32] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Computer Vision*, pp. 4489–4497, 2015.
[33] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," in *Proc. Inter-speech*, pp. 3652–3656, 2017.
[34] A. Fung and B. Mak, "End-to-end low-resource lip-reading with maxout CNN and LSTM," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 2511–2515, 2018.
[35] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson, "3D convolutional neural networks for cross audio-visual matching recognition," *IEEE Access*, vol. 5, pp. 22081–22091, 2017.
[36] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.
[37] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with long short-term memory," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 6115–6119, 2016.
[38] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with LSTMs," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 2592–2596, 2017.
[39] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
[40] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5–6, pp. 602–610, 2005.
[41] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014.
[42] D. K. Margam *et al.*, "LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models," *arXiv preprint arXiv:1906.12170*, 2019.
[43] K. Xu, D. Li, N. Cassimatis, and X. Wang, "LCANet: End-to-end lipreading with cascaded attention-CTC," in *Proc. 13th IEEE Int. Conf. Automatic Face & Gesture Recognition*, pp. 548–555, 2018.
[44] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and

translate," in *Int. Conf. Learning Representations*, 2015.
[45] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.
[46] T. Afouras, J. S. Chung, and A. Zisserman, "Deep lip reading: a comparison of models and an online application," in *Proc. Interspeech*, pp. 3917–3921, 2018.
[47] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 7613–7617, 2021.
[48] Y. Zhang *et al.*, "Can we read speech beyond the lips? Rethinking RoI selection for deep visual speech recognition," in *Proc. 15th IEEE Int. Conf. Automatic Face and Gesture Recognition*, pp. 356–363, 2020.
[49] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.
[50] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 6319–6323, 2020.
[51] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic, "Lip-reading with densely connected temporal convolutional networks," in *Proc. IEEE Winter Conf. Applications of Computer Vision*, pp. 2857–2866, 2021.
[52] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 156–165, 2017.
[53] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new audiovisual database for multimodal human-computer interface research," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, pp. II-2017, 2002.
[54] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.
[55] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. Asian Conf. Computer Vision*, pp. 87–103, 2016.
[56] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.
[57] A. Ortega *et al.*, "AV@CAR: A Spanish multichannel multimodal corpus," in *Proc. Int. Conf. Language Resources and Evaluation*, pp. 763–766, 2004.
[58] S. Antar and A. Sagheer, "Audiovisual Arabic speech (AVAS) database," *Int. J. Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 9, pp. 1147–1153, 2013.
[59] S. Petridis *et al.*, "End-to-end audiovisual speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 6548–6552, 2018.
[60] G. Sterpu, C. Saam, and N. Harte, "Attention-based audio-visual fusion for robust automatic speech recognition," in *Proc. 20th ACM Int. Conf. Multimodal Interaction*, pp. 111–115, 2018.
[61] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, "Self-supervised learning of audio-visual objects from video," in *European Conf. Computer Vision*, pp. 208–224, 2020.
[62] O. Wiles, A. Koepke, and A. Zisserman, "X2Face: A network for controlling face generation using images, audio, and pose codes," in *Proc. European Conf. Computer Vision*, pp. 670–686, 2018.
[63] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with GANs," *Int. J. Com-puter Vision*, vol. 128, no. 5, pp. 1398–1413, 2020.
[64] A. M. Sarhan, N. M. Elshennawy, and D. M. Ibrahim, "HLR-net: a hybrid lip-reading model based on deep convolutional neural networks," *Computers, Materials & Continua*, vol. 68, no. 2, pp. 1531–1549, 2021.
[65] S. Fenghour, D. Chen, K. Guo, B. Li, and P. Xiao, "Deep learning-based automated lip-reading: A survey," *IEEE Access*, vol. 9, pp. 121184–121205, 2021.
[66] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A review of recent advances in visual speech recognition," *Image and Vision Computing*, vol. 32, no. 9, pp. 590–605, 2014.