AI

MDPI

Article

Dual-Stream Former: A Dual-Branch Transformer Architecture for Visual Speech Recognition

Sanghun Jeon [1,*], Jieun Lee [2] and Yong-Ju Lee [1]

[1] Electronics and Telecommunications Research Institute, Daejeon 34129, Republic of Korea; yongju@etri.re.kr
[2] Korea Culture Technology Institute, Gwangju 61005, Republic of Korea; jieunlee@gist.ac.kr
* Correspondence: jeon0123@etri.re.kr; Tel.: +82-10-9294-1843

Abstract

This study proposes Dual-Stream Former, a novel architecture that integrates a Video Swin Transformer and Conformer designed to address the challenges of visual speech recognition (VSR). The model captures spatiotemporal dependencies, achieving a state-of-the-art character error rate (CER) of 3.46%, surpassing traditional convolutional neural network (CNN)-based models, such as 3D-CNN + DenseNet-121 (CER: 5.31%), and transformer-based alternatives, such as vision transformers (CER: 4.05%). The Video Swin Transformer captures multiscale spatial representations with high computational efficiency, whereas the Conformer back-end enhances temporal modeling across diverse phoneme categories. Evaluation of a high-resolution dataset comprising 740,000 utterances across 185 classes highlighted the effectiveness of the model in addressing visually confusing phonemes, such as diphthongs (/ai/, /au/) and labio-dental sounds (/f/, /v/). Dual-Stream Former achieved phoneme recognition error rates of 10.39% for diphthongs and 9.25% for labiodental sounds, surpassing those of CNN-based architectures by more than 6%. Although the model's large parameter count (168.6 M) poses resource challenges, its hierarchical design ensures scalability. Future work will explore lightweight adaptations and multimodal extensions to increase deployment feasibility. These findings underscore the transformative potential of Dual-Stream Former for advancing VSR applications such as silent communication and assistive technologies by achieving unparalleled precision and robustness in diverse settings.

Keywords: speech recognition; Video Swin Transformer; conformer; spatiotemporal feature extraction

Academic Editor: Sharifu Ura

Received: 28 July 2025
Revised: 29 August 2025
Accepted: 2 September 2025
Published: 9 September 2025

Citation: Jeon, S.; Lee, J.; Lee, Y.-J. Dual-Stream Former: A Dual-Branch Transformer Architecture for Visual Speech Recognition. *AI* **2025**, *6*, 222. https://doi.org/10.3390/ai6090222

1. Introduction

Visual speech recognition (VSR), commonly referred to as lipreading, is a sophisticated task that involves interpreting spoken language through visual cues such as the movement of the lips, tongue, and other facial regions without relying on audio input. This field has gained considerable attention in recent years owing to its wide range of practical applications [1]. In noisy environments, where audio-based speech recognition systems struggle to perform effectively, VSR provides a robust alternative by extracting meaningful information solely from visual data [1–4]. In addition, it serves as a crucial tool for assisting people with hearing loss by providing improved accessibility and communication capabilities [1,5]. Furthermore, VSR plays a pivotal role in advancing human–computer interactions, enabling more intuitive and inclusive interfaces for applications such as silent dictation, speech recognition in public spaces, and biometric authentication [1,4,6,7].

Despite these advancements and their promising potential, VSR systems still face significant challenges that hinder their widespread adoption and effectiveness [1]. One of the primary difficulties is the efficient capture and processing of spatiotemporal features, which are essential aspects of visual speech. Unlike audio-based systems that rely on sequential sound patterns, VSR must account for spatial variations in lip shapes and movements as well as temporal dependencies across frames in a video sequence [1]. This dual complexity often requires sophisticated models and substantial computational resources, making real-time performance challenging [3,8,9].

Existing VSR models often struggle to effectively model spatiotemporal dependencies. Traditional convolutional neural network (CNN)-based approaches, such as 3D CNN, capture local patterns well but fall short in representing long-range temporal relationships and require significant computational resources. These limitations highlight the need for more efficient architectures that can capture both local and global contexts in visual speech data. Moreover, ensuring robust performance across diverse settings, such as varying lighting conditions, speaker variability, and differing head poses, remains an unresolved issue [5–7]. These challenges highlight the need for more advanced architectures and innovative approaches to improve VSR performance and usability [2,3,5,7,10].

Building on this foundation, recent approaches have focused on improving VSR performance using advanced spatiotemporal modeling techniques [2,3,11,12]. Among traditional methods, 3D convolutional neural networks (3D CNNs) have been widely employed for capturing spatiotemporal features because of their ability to model both spatial and temporal information simultaneously [13–15]. However, despite their effectiveness, 3D CNN-based approaches have several critical limitations that hinder their scalability and performance in real-world applications. 3D CNNs extend 2D convolution operations to include the temporal axis, thereby enabling joint spatial and temporal feature extraction. This approach effectively captures localized spatiotemporal patterns, making 3D CNNs a popular choice for action recognition and gesture-based tasks [8,14,15]. Video-based tasks, such as action recognition, video captioning, and lipreading, require the efficient modeling of spatial and temporal dependencies. Historically, three-dimensional (3D) CNNs have been the backbone of video analysis owing to their ability to extend convolutional paradigms across temporal dimensions. However, they face critical limitations in computational efficiency, representation flexibility, and global dependency modeling [1,8,14,15]. Transformers, which were first introduced in natural language processing, have demonstrated superior performance in vision tasks by leveraging self-attention mechanisms. These mechanisms provide an intrinsic capability to model global spatiotemporal relationships, addressing several shortcomings of CNN-based approaches [8,9,14–16]. This study explored the architectural differences, performance tradeoffs, and application scenarios of 3D CNNs and transformer models when processing video data.

Transformers represent a significant shift in video analysis by employing self-attention mechanisms that allow for flexible nonlocal feature aggregation. Unlike 3D CNNs, which rely on hierarchical convolutional filters, the transformer computes pairwise dependencies across all tokens, thereby providing a holistic understanding of spatiotemporal data. For example, the Vision Transformer (ViT) modifies its architecture for image processing by segmenting the input into patches and encoding positional relationships. Extensions, such as the Video Swin Transformer, further optimize this framework by introducing a hierarchical structure and localized attention windows, making it more computationally efficient [16–18]. Transformers excel at modeling long-range dependencies, which is a critical advantage in tasks that require temporal coherence over extended video sequences. Additionally, their shape-based bias aligns more closely with human perception, as demonstrated in studies comparing human error consistency with model behavior [17].

The primary objective of this study is to develop an advanced and efficient VSR model that leverages only visual information for lipreading tasks. The model is designed to address key challenges in VSR, such as computational efficiency, temporal dependency modeling, and recognition accuracy, through innovative integration of the Video Swin Transformer, Conformer, and a transformer-based decoder.

Efficient Spatiotemporal Feature Extraction: The proposed model employs a Video Swin Transformer at the front-end to extract spatiotemporal features efficiently. Its hierarchical architecture and shifted-window attention mechanism enable localized attention computation, which makes it particularly suitable for lipreading datasets with small spatial dimensions.

Robust Temporal Dependency Modeling: A Conformer-based encoder is integrated into the back-end to capture both short- and long-term temporal dependencies. This component enhances the ability of the model to represent the sequential nature of speech data with high temporal precision.

Accurate and Context-Aware Decoding: To improve transcription accuracy, a transformer-based decoder is incorporated. By leveraging self-attention mechanisms, it effectively captures the global context and models long-range dependencies within an output sequence.

Therefore, the proposed Dual-Stream Former model significantly advances VSR by achieving a balance between computational efficiency and recognition performance. By integrating a Video Swin Transformer [16], the model achieves lower computational complexity. This is made possible by the use of a hierarchical structure and shifted-window attention mechanism, which are well suited for processing small input sizes typically found in lipreading datasets. It enhances spatiotemporal representations with robust feature extraction and precise temporal dependency modeling facilitated by the Conformer back-end [5]. Additionally, the attention-based architecture of the model aligns closely with human perception, as demonstrated in previous studies that compared human error patterns with those of transformer-based models [17]. This alignment contributes to improved robustness and scalability across diverse datasets and real-world scenarios. These advancements have positioned the model as a cutting-edge solution for achieving efficiency and accuracy in lipreading tasks.

Contributions and Novelty. We introduce a dual-branch transformer for visual-only speech recognition that integrates a Video Swin Transformer front-end with a Conformer back-end and a transformer decoder with Connectionist Temporal Classification (CTC)/Attention. We evaluate a new high-resolution dataset of 740,000 utterances across 185 classes from 40 speakers with a fixed 70/30 within-subject split (Sections 4.1–4.3). Beyond aggregate CER, we provide phoneme-group analyses and confusion matrices, showing the largest gains for visually confusable categories such as diphthongs and labiodentals (Section 5.3). Under identical conditions, our model achieves the lowest CER (3.46%), converges faster than baselines, and shows participant-wise statistical significance (Section 5.4;). Together, these results clarify when and why hierarchical video transformers coupled with Conformer modeling are effective for visual-only speech recognition.

2. Related Work

2.1. CNN-Based Architectures

The CNN has emerged as one of the most commonly used architectures for deep learning. Various CNN-based architectures, such as VGG [19], ResNet [20], MobileNet [21], EfficientNet [22], and DenseNet [23], have been widely adopted in VSR tasks because of their effectiveness in learning visual representations from lip movements. These models offer different tradeoffs in terms of depth, computational cost, and feature-extraction capac-

ity, making them suitable for various VSR settings. However, despite their widespread use, these architectures are inherently limited in modeling long-range temporal dependencies and often struggle with speaker variability and pose changes, which are critical for robust VSR performance.

Chung et al. proposed the first end-to-end deep visual representation learning the method for word-level VSR [24]. Their study employed a VGG-M backbone network to evaluate various image-sequence inputs (Multiple Towers vs. Early Fusion) and temporal fusion techniques (2D CNNs vs. 3D CNNs), highlighting the strengths and limitations of each approach. The experimental results revealed that 2D CNNs significantly outperform 3D CNNs. However, the conclusions drawn from the study lack robustness because they are based on limited ablation studies and short-term word-level datasets, restricting their generalizability to continuous speech scenarios.

Building upon this, subsequent models have attempted to address these limitations by incorporating deeper temporal modeling and architectural modifications. In 2017, Assael et al. [2] introduced LipNet, the first end-to-end sentence-level VSR model. LipNet extracts visual features using a three-layer Spatiotemporal CNN (STCNN, also known as 3D CNN). The experimental results support the intuition that extracting spatiotemporal features with an STCNN is superior to aggregating only spatial features. Considering that 3D CNNs are better at capturing the dynamics of the mouth region and 2D CNNs are more efficient in terms of time and memory, Stafylakis and Tzimiropoulos [25] proposed a combination of 3D and 2D CNNs for visual feature extraction. Their proposed visual backbone network integrates a shallow 3D CNN and 2D ResNet. The 3D CNN layer captures the short-term temporal dynamics of lip movements. Although this hybrid architecture improved local feature modeling and computational efficiency, it remained constrained in its ability to capture global temporal structures and generalize to speakers with varying styles and accents.

Owing to its remarkable performance, numerous VSR models [11,26–29] have adopted it as a backbone network for visual feature extraction. Recently, Feng et al. [30] enhanced this architecture by integrating a Squeeze-and-Extract [31] module to improve local feature discrimination. However, these enhancements do not fully address the limitations of CNN-based models in capturing long-range dependencies or in providing robustness under challenging real-world conditions. In addition to VGG and ResNet, researchers have employed other prominent 2D CNN architectures, including DenseNet [23], ShuffleNet [27], and MobileNet [21]. However, these models also rely on local receptive fields and exhibit similar constraints in scalability and context modeling, further motivating the exploration of transformer-based alternatives to VSR.

These cumulative limitations have prompted a shift toward attention-based models for VSR. As illustrated in Figure 1a, 3D CNNs extend traditional 2D convolutions by incorporating the temporal dimension, thereby enabling simultaneous extraction of spatial and temporal features. This layerwise convolutional processing facilitates localized spatiotemporal learning across frames. However, 3D CNNs are fundamentally limited by their fixed receptive fields, which hinders the modeling of long-range temporal dependencies. In addition, their computational demands increase significantly because of the large number of parameters introduced by added temporal dimensions [8,14,15]. These limitations have motivated the transition toward attention-based models, such as transformers, which can more effectively capture the global context in spatiotemporal data.

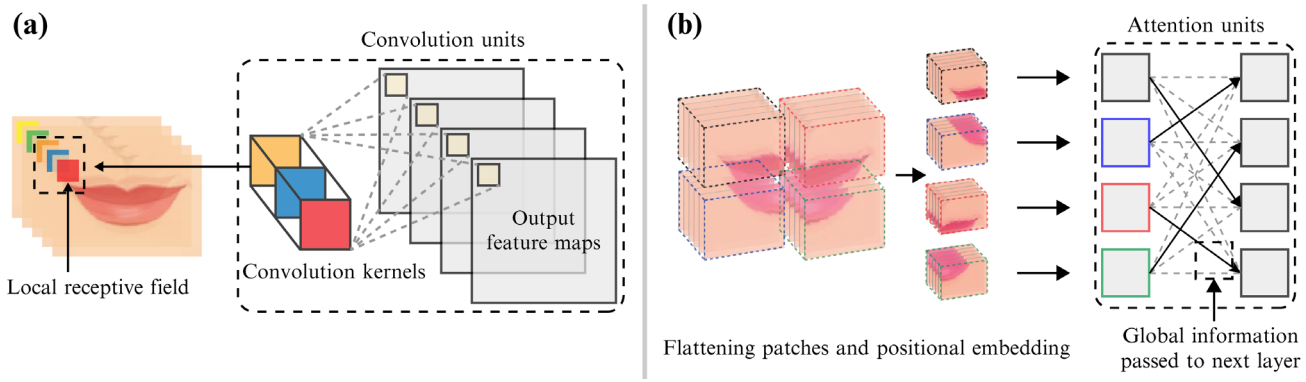

Figure 1. Visualization of the front-end processing pipeline: (**a**) convolutional units for local spatiotemporal feature extraction; (**b**) transformer-based attention mechanisms for capturing global dependencies in visual inputs.

Although transformer-based models address some of the key limitations of 3D CNNs, such as fixed receptive fields and insufficient global-context modeling, they still face challenges in terms of computational efficiency and simultaneous capture of local and global spatiotemporal features. Existing models often rely on flat attention mechanisms that lack inductive bias, leading to slower convergence and increased resource demand. To overcome these issues, we propose Dual-Stream Former, a novel architecture that integrates a Video Swin Transformer for efficient hierarchical spatial modeling with a Conformer for robust temporal dependency extraction.

As illustrated in Figure 1b, transformer-based architectures emphasize self-attention mechanisms for feature extraction. The input video data are divided into patches, and positional embeddings are applied to retain spatial and temporal order. The architecture uses multihead self-attention layers to compute the global dependencies between all patches, enabling a holistic understanding of the video sequence. Unlike 3D CNNs, transformers dynamically allocate attention weights by focusing on salient features and suppressing irrelevant information. This global processing enhances the ability of the model to capture long-range spatiotemporal relationships while being more computationally efficient, especially in hierarchical designs such as Video Swin Transformers [14–17].

2.2. Transformer-Based Architectures

Motivated by the remarkable success of transformer architectures in the NLP domain, researchers have recently begun to apply transformers to computer vision tasks [17]. Transformers have demonstrated potential as viable alternatives to CNNs. Prajwal et al. [32] introduced an end-to-end visual transformer-based pooling mechanism that learns to track and aggregate lip-movement representations. The proposed visual backbone network minimizes dependency on complex preprocessing and enhances the robustness of visual representations. The ablation study results clearly indicate that the visual transformer-based pooling mechanism significantly increases the VSR performance. Although transformers offer strong capabilities for modeling global spatiotemporal dependencies, they often require large amounts of data and computational resources and lack strong local inductive biases that make CNNs efficient for lower-level feature extraction. By contrast, CNNs are computationally efficient and effective in modeling local spatial patterns but struggle with long-range temporal modeling and global-context integration. The complementary nature of CNNs and transformers highlights the potential benefits of integrating both approaches. CNNs can serve as lightweight local feature extractors, whereas transformers provide flexible and powerful global-context models.

Beyond the core literature on visual speech recognition, it is also important to consider cross-disciplinary advances in multimodal interaction and wearable technologies. For example, the Funabot-Sleeve study [33] introduced a wearable device that employs McKibben artificial muscles to deliver haptic sensations. Although this work addresses a different modality, it exemplifies the broader trend of integrating vision, haptics, and human–machine interfaces. Such cross-domain perspectives highlight the potential relevance of our proposed framework for applications in assistive technology and embodied AI, where visual speech recognition could be combined with haptic or multimodal feedback to increase accessibility and robustness.

To leverage the strengths of both architectures, we propose Dual-Stream Former, a hybrid model that combines the hierarchical efficiency of the Video Swin Transformer with the temporal modeling capabilities of the Conformer. This integration aims to achieve a balanced solution for efficient and accurate VSR.

We will now discuss positioning with respect to the recent VSR literature. Prior CNN-based and flat transformer front-ends emphasize either local patterns or global attention, leaving a gap in jointly modeling hierarchical spatial structures with sequence-level temporal dependencies (Sections 2.1 and 2.2). Our approach explicitly fuses a hierarchical Video Swin visual encoder with a Conformer and CTC/Attention decoding in one end-to-end pipeline and quantifies the benefits at both the aggregate and phoneme-group levels, with participant-wise statistics.

3. Proposed Architecture

We propose Dual-Stream Former, a novel VSR architecture designed to integrate powerful spatiotemporal representation learning with efficient sequence modeling. The proposed model consists of three key components: a Video Swin Transformer-based front-end, a Conformer encoder, and a transformer decoder. This modular design captures localized visual patterns, models temporal dependencies, and generates sequences with high accuracy and robustness.

At a high level, the Video Swin Transformer extracts hierarchical spatiotemporal features using 3D shifted-window attention. These features are processed by the Conformer, which integrates convolutional operations with self-attention to model both local and global temporal relationships. Finally, a transformer decoder with a hybrid Connectionist Temporal Classification (CTC)/Attention mechanism generates the output sequences, enabling effective sequence transduction. The detailed structure is shown in Figure 2 and summarized in Table 1. Supplementary Equations (S1)–(S4) provide the mathematical formulations and implementation details.

The proposed architecture is designed to excel in isolated word-level lipreading tasks, achieving superior accuracy while maintaining computational efficiency. Table 1 summarizes the full Dual-Stream Former pipeline, including a 3D convolutional front-end, four Swin-B stages for hierarchical feature extraction, a 12-layer Conformer encoder for temporal modeling, and a 6-layer transformer decoder. This configuration integrates hierarchical spatial processing, precise temporal alignment, and advanced decoding in a unified system. As shown in Figure 2, this design enables effective sequence generation and demonstrates robustness across diverse datasets for both the Korean and English language tasks.

3.1. Video Swin Transformer (See Supplementary Equation (S1) for Mathematical Details)

The front-end of Dual-Stream Former leverages the Video Swin Transformer [16] to extract dense spatiotemporal representations from the input video frames. It extends the 2D Swin Transformer to 3D by employing non-overlapping 3D patch partitioning and shifted-window-based multihead self-attention. This design balances global-context modeling and

computational efficiency by restricting attention computation to local windows while still enabling cross-window interaction through a shifted mechanism.

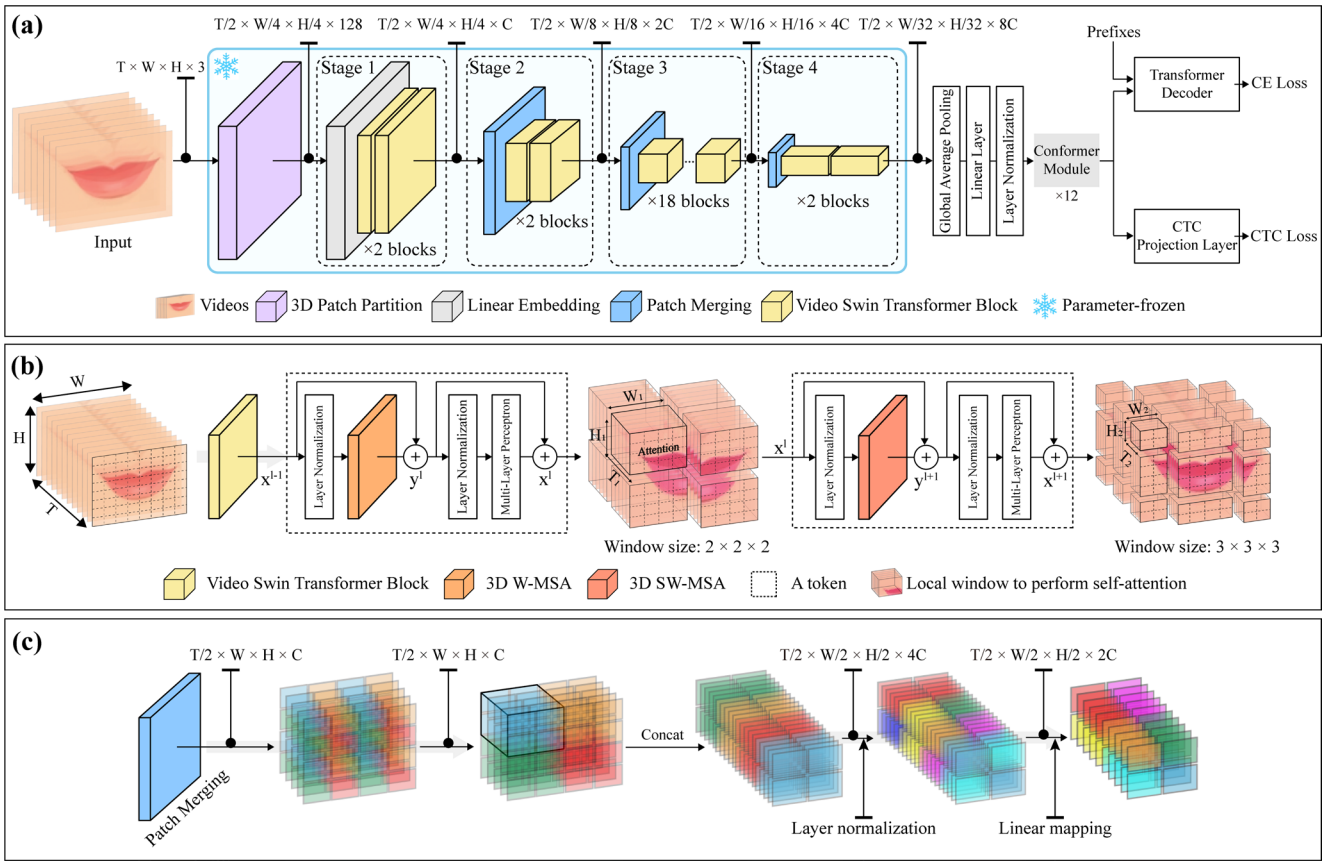

Figure 2. Proposed Dual-Stream Former pipeline for visual speech recognition: (**a**) stage-wise processing of visual input via 3D patch partitioning, Swin Transformer blocks, and patch merging, followed by back-end Conformer and decoder; (**b**) internal structure of Swin Transformer blocks highlighting multi-head attention mechanisms; (**c**) patch merging and linear mapping operations.

Table 1. Architecture of our proposed Visual Speech Recognition model, which integrates a Swin-B-based front-end, Conformer-based temporal encoder, and a transformer decoder. These components work in tandem to extract spatiotemporal features, model temporal dependencies, and generate output sequences using a hybrid Connectionist Temporal Classification (CTC)/Attention mechanism.

Stage	Layer		Output Shape
Input	Image sequences		T × W × H × 3
Front-end	Conv3D: 4 × 4 × 4, 96 filters, stride 4 × 4 × 4		T/2 × W/4 × H/4 × 128
	Stage1	Swin-B × 2 Blocks, W: 7 × 7, P: 4 × 4, H: 4	T/2 × W/4 × H/4 × 128
	Stage2	Swin-B × 2 Blocks, W: 7 × 7, P: 4 × 4, H: 4	T/2 × W/8 × H/8 × 256
	Stage3	Swin-B × 18 Blocks, W: 7 × 7, P: 4 × 4, H: 4	T/2 × W/16 × H/16 × 512
	Stage4	Swin-B × 2 Blocks, W: 7 × 7, P: 4 × 4, H: 4	T/2 × W/32 × H/32 × 1024
	Global Average Pooling		T/2 × 1024
Back-end	Conformer encoder × 12 Blocks		T/2 × 512
Decoder	Transformer decoder × 6 Blocks		T/2 × V
-	CTC/Attention hybrid mechanism		-

The architecture follows a four-stage hierarchy in which the spatial resolution is progressively reduced through patch merging while the temporal resolution remains unchanged. Each stage consists of Swin-B blocks with multihead attention, feedforward layers, and layer normalization wrapped in residual connections. These operations facilitate the extraction of both fine-grained and high-level motion features, which are critical for lipreading tasks.

3.2. Conformer Encoder (See Supplementary Equation (S2) for Mathematical Details)

The intermediate encoding module is based on the Conformer [5] architecture, which has shown superior performance in modeling temporal sequences. The Conformer combines multi-head self-attention with convolutional modules to effectively capture both long-term dependencies and short-term local variations within a visual speech signal.

The overall structure of the Conformer encoder block is illustrated in Figure 3, where (a) and (d) correspond to the feedforward modules, (b) denotes the convolutional module, and (c) represents the self-attention module. This arrangement enables the balanced modeling of both local and global temporal features.

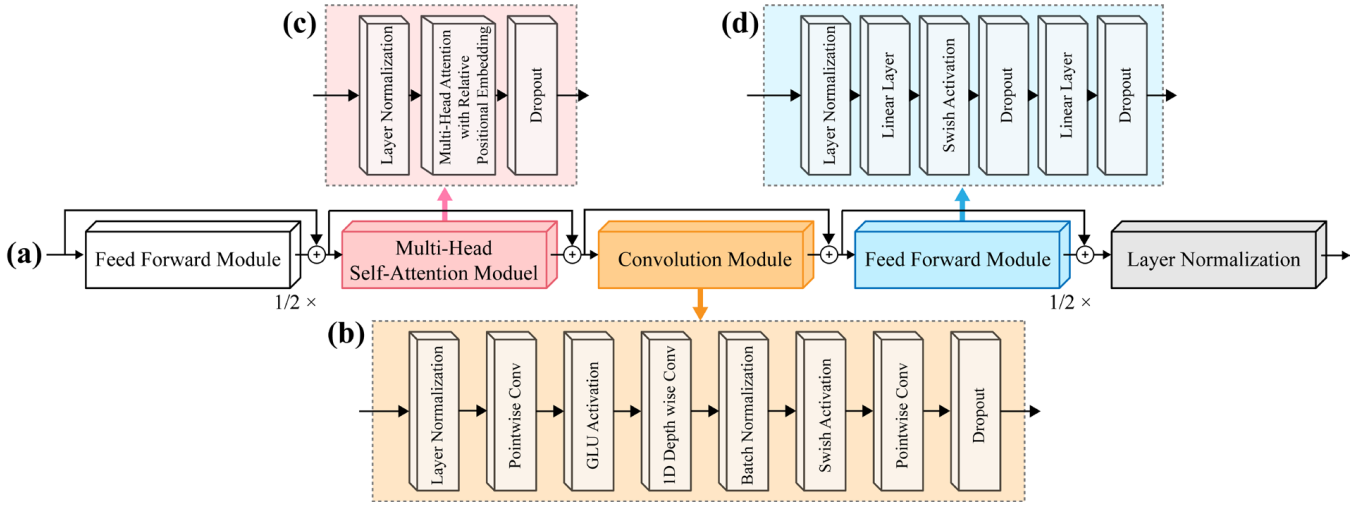

Figure 3. Overview of the Conformer encoder layer structure: (**a**) feedforward module, (**b**) convolutional module for local feature extraction, (**c**) multi-head self-attention for global dependency modeling, and (**d**) final feedforward module with normalization and residual connections.

Each encoder block includes two feedforward layers, a self-attention module, and a depthwise separable convolution module, all of which are equipped with residual connections and layer normalization. This structure enhances the temporal modeling capacity of the system and ensures smooth integration between the Swin-based visual features and the transformer decoder.

3.3. Transformer Decoder (See Supplementary Equation (S3) for Mathematical Details)

The final stage of Dual-Stream Former employs a transformer decoder to generate a target character sequence. Each decoder layer comprises masked multihead self-attention (to ensure autoregressive decoding), cross-attention (to incorporate encoder outputs), and feedforward networks. All the submodules are connected through residual paths and normalization layers.

To improve alignment and recognition accuracy, a hybrid decoding mechanism combining CTC [34] and attention-based loss was used during training (see Supplementary Equation (S4) for details). This hybrid approach helps guide the decoder in learning monotonic alignments while preserving its flexibility in handling complex sequences. Dur-

ing inference, the joint beam search fuses the CTC and attention outputs to generate the final transcription.

Further details of each module's internal mechanisms, including attention computations, patch merging, and positional encoding, are presented in Supplementary Equations (S1)–(S4).

4. Experiment

4.1. Dataset

To develop a robust and diverse dataset for VSR, data collection was conducted in a controlled experimental environment. This study included 40 participants (20 males and 20 females), with an average age of 29.8 years. The participants ranged from 22 to 38 years, and no subject older than 40 years was included. The cohort primarily consisted of individuals of Asian, Egyptian, and Pakistani descent, all of whom were proficient in English at a near-native level. Younger adults were intentionally recruited because the subsequent application stage of this research involved human–computer interaction (HCI) experiments, and participants accustomed to IT devices were less likely to encounter difficulties during experimental procedures. As a reason for indemnification, in accordance with Article 13 of the Enforcement Rules of "Research on Human Subjects," a study that does not collect and record personal identification information is exempt from deliberation. This study only used simple contact measurement equipment or observation equipment that do not cause any physical changes. Only human voice and lip information were collected and used; thus, there was no personal identification information. Therefore, ethics approval was not required for indemnification reasons. The participants provided written and oral informed consent.

The recordings were captured using a custom-designed head-mounted device to ensure consistent alignment and resolution across all sessions, as illustrated in Figure 4. The Logitech C920 HD PRO webcam (Figure 4a) mounted on the head-mounted device was positioned 490 mm from the participant's face. This setup ensured consistent frontal-pose capture and minimized variability owing to head movements (Figure 4b,d,e). Videos were recorded at a resolution of 1920 × 1080 pixels (FHD) and frame rate of 30 frames per second (fps). Each participant pronounced 185 independent word classes, including letters, numbers, emotions, and daily vocabulary. Each keyword was recorded for 3 s and repeated 100 times, resulting in 740,000 utterances. In total, the dataset used for modeling comprised 740,000 video utterances (616.67 h) across 185 classes from 40 speakers (20 females, 20 males). The recording environment was optimized to minimize external influences, such as light.

Although the capture rig included a microphone for synchronization (Figure 4), only visual signals were used in this study, and audio tracks were excluded from all training, evaluation, and distribution. Recordings were pseudonymized with randomized identifiers, and no names or personal identifiers were collected. All recordings were performed openly with explicit written consent, prohibiting any covert use.

Our dataset represents a significant advancement over existing public datasets [35–40] in the field of VSR. As shown in Tables 2 and S1–S3, earlier datasets, such as Tulips1 [35] and M2VTS [36], are limited in both scale and diversity, with a small number of speakers and utterances. For example, Tulips1 contained only four classes with a total of 96 utterances, whereas M2VTS included 10 classes with 1000 utterances. Similarly, datasets such as AVLetters [37] and AVLetters2 [38], which focus on letters, offer limited utterance counts and diversity, with resolutions ranging from 376 × 288 to 1920 × 1080 pixels.

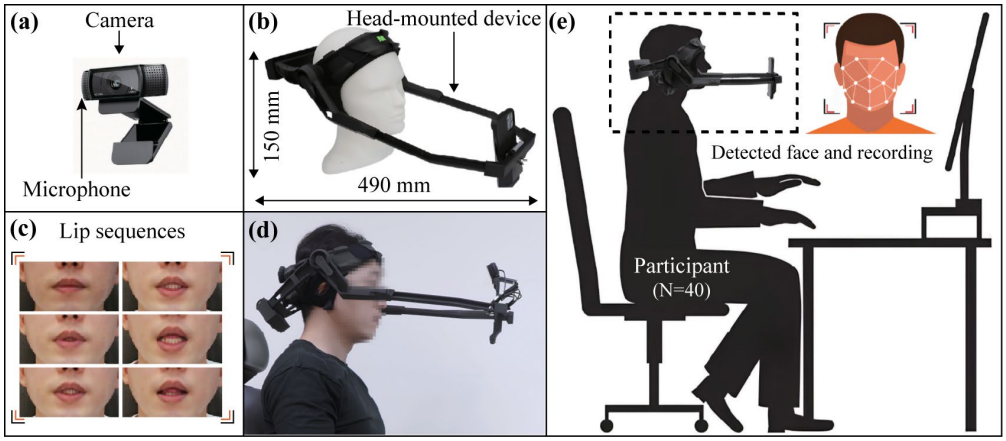

Figure 4. Experimental setup for dataset collection: (**a**) camera and microphone for video and audio recording, (**b**) head-mounted device ensuring consistent distance and alignment, (**c**) extracted lip region sequences, (**d**) detailed view of the head-mounted device, and (**e**) participant seated during the recording process.

Table 2. Comprehensive breakdown of the dataset by topic, including the number of classes, total recording duration, dataset proportion, and total utterances for each category (Table S2 provides the detailed information).

Topic	Classes	Hours	Ratio	Utterances
Alphabet	26	86.67	14.05	104,000
Number	10	33.33	5.41	40,000
System	20	66.67	10.81	80,000
Emotion	27	90.00	14.59	108,000
Daily	25	83.33	13.51	100,000
Food	25	83.33	13.51	100,000
Work	17	56.67	9.19	68,000
Education	25	66.67	10.81	80,000
Finance	15	50.00	8.11	60,000
Total	185	616.67	100	740,000

By contrast, our dataset introduced 40 speakers, each of whom recorded 185 independent classes comprising letters, digits, and words, resulting in a total of 740,000 utterances. This uniform coverage ensured consistent representation across participants and far surpassed the scale of all the prior datasets. Furthermore, the dataset maintained a high resolution of 1920 × 1080 pixels, enabling superior visual quality for detailed lip motion analysis. The consistency of the frontal pose across all recordings further strengthened the suitability of the dataset for the VSR tasks.

Compared with AVDigits (2018) [40], which includes pose variations (0°, 45°, and 90°), our dataset provides focused and standardized frontal-pose alignment, reducing variability and improving the reliability of model training. With the highest number of classes and resolution among the public datasets, our dataset establishes a new standard for VSR research, facilitating the creation of robust models that can effectively address diverse linguistic and visual challenges.

4.2. Data Preprocessing

The visual data preprocessing pipeline in our study, inspired by previous work [3], employs a Dlib linear classifier using a histogram of oriented gradient features to extract 68 facial landmarks and localize the lip region [41,42]. The cropped regions were resized to 96 × 96 pixels, converted into grayscale, and normalized. We selected the

96 × 96 resolution to preserve the essential lip contours for visualization while minimizing the computational load, following findings from prior work [12]. During training, random cropping to 88 × 88 pixels and horizontal flipping (p = 0.5) were applied to improve the generalizability. These augmentations simulate natural variations in head movements and speaker orientation, which are common in real-world lipreading scenarios. For validation, center cropping and normalization were performed to ensure consistency. In addition, affine transformations were incorporated to correct the scale and positional shifts caused by lighting or participant variability, thereby aligning lip features more accurately across samples [12]. These steps increase the robustness of visual inputs for speech recognition and reduce the risk of overfitting.

4.3. Training and Implementation

The proposed model was trained on datasets split into training and validation sets at a 7:3 ratio, with data collected 100 times per class. To mitigate potential issues related to participants' focus and pronunciation degradation during repeated trials, specific data segments were reserved exclusively for validation. Specifically, the early (21–30), middle (51–60), and late (81–90) portions were allocated to the validation set, whereas the training set consisted of the remaining subsets, including groups 1–9, 10–20, 31–40, 41–50, 61–70, 71–80, and 91–100. Importantly, this split was applied to each participant to ensure that both the training and validation sets included data from all the participants. This within-subject stratification strategy mitigated the overfitting risks associated with reduced pronunciation quality in repetitive data while preserving participant-dependent consistency. Accordingly, 518,000 utterances were used for training and 222,000 utterances for validation (70/30 split applied per participant and per class).

The experiment was conducted under the same conditions as those described previously [43], with the key difference being that the model was trained entirely from scratch using a newly collected dataset without any pretrained weights. All modules, including the front-end, were initialized randomly. The backend was configured with the hyperparameters $e = 12$, $d_{ff} = 2048$, $d_k = 256$, and $d_v = 256$, where e denotes the number of Conformer blocks. In the visual-only configuration, four attention heads ($n_{head} = 4$) were used, and each depthwise convolutional layer had a kernel size of 31. The transformer decoder comprised six self-attention blocks with hyperparameter settings for the feedforward and self-attention modules consistent with the encoder. Training was conducted using the Adam optimizer $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$ with a mini-batch size of eight. The learning rate increased linearly over the first 25,000 steps to a peak of 0.0004, followed by decay according to the inverse square root of the step count. Training was performed over 30 epochs for a total of 212.40 h, using a single NVIDIA A100 GPU. The experiments were implemented in PyTorch 1.12.0 with CUDA 11.8, ensuring compatibility with the hardware and efficient GPU utilization.

For model selection, we tuned hyperparameters on the fixed validation set described above by monitoring the validation CER; no k-fold cross-validation was used because of the computational cost. The final hyperparameters reported in this subsection were adopted for all the main experiments.

4.4. Evaluation Metrics

In the field of VSR, character error rate (CER) is a key metric used to evaluate the accuracy of speech recognition systems. CER quantifies the percentage of characters that are incorrectly predicted compared with the ground truth, with lower values indicating better performance. This is calculated as follows:

$$\mathrm{CER}(\%) = \frac{\mathrm{S} + \mathrm{D} + \mathrm{I}}{\mathrm{N}} \times 10$$

where S is the number of substitutions, D the number of deletions, I the number of insertions, and N the total number of characters in the ground truth. This metric provides a comprehensive measure of a model's transcription accuracy by capturing different error types and is especially effective for benchmarking performance under noisy or visually ambiguous conditions. Its standardized formulation ensures consistent and comparable evaluations across different VSR architectures.

In addition to reporting the CER, we conducted statistical significance testing using paired t tests to determine whether the observed performance differences between the models were due to chance. To further quantify the magnitude of these differences, we computed Cohen's d, a standardized effect size measure defined as the difference between two means divided by their pooled standard deviations. Effect sizes were interpreted using the following conventional thresholds: small (d = 0.2), medium (d = 0.5), large (d = 0.8), and very large (d > 1.2). This dual approach, which combines statistical significance with effect size analysis, enabled us to assess not only the reliability but also the practical relevance of the performance improvements observed in transformer-based architectures over their CNN-based counterparts.

4.5. Hyperparameter Tuning and Validation Protocol

We tuned the hyperparameters using the fixed hold-out validation set described in Section 4.3 (7:3 split with within-subject stratification of utterance indices). Model selection was validation-driven: we monitored the validation CER and training stability and retained settings that yielded a lower CER without incurring a prohibitive training time. The final configuration used in all the main experiments is as follows (see also Section 4.3): Conformer blocks $e = 12$, feed-forward dimension $d_{ff} = 2048$, key/value dimensions $d_k = d_v = 256$, number of attention heads $n_{head} = 4$, depthwise convolutional kernel size = 31, and a transformer decoder with six blocks. Optimization used Adam ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$) with a linear warm-up for the first 25 k steps up to a peak learning rate of 4×10^{-4}, followed by inverse square-root decay and a mini-batch size of eight for 30 epochs.

We did not employ k-fold cross-validation for hyperparameter tuning owing to the computational cost of full end-to-end training on our large-scale, high-resolution video dataset. To reduce split-specific bias, we adopted the above within-subject stratification and report participant-wise statistical tests in Section 5.4, which consistently support the superiority of the proposed model. Future work will consider k-fold or nested cross-validation and leave-one-speaker-out protocols to further assess generalizability.

5. Performance Evaluation

5.1. Comparison of Character Error Rate

To evaluate the performance of various model architectures for processing input data consisting of 30 frames of 3 s video, we compared CNN-based and transformer-based visual front-end structures (Table 3). This comparison highlights the differences in their ability to extract spatial and temporal features, with transformer-based architectures consistently demonstrating superior performance in tasks requiring comprehensive spatiotemporal modeling. Importantly, a paired t test conducted on the CER results across multiple participants confirmed that the observed improvements were statistically significant ($p < 0.001$), thereby reinforcing the robustness of the transformer-based front-ends over their CNN-based counterparts.

Table 3. Evaluation of various visual front-end models based on their parameters, epoch time, training duration, and CER in the context of visual speech recognition (Table S3 provides the detailed information).

Visual Front-End	Encoder + Decoder	Total Params (M)	Epoch Time (h)	Training Time (h)	CER (%)
3D-CNN	Conformer + Transformer + CTC/Attention	113.6	4.77	143.11	10.71
3D-CNN + VGG-16		218.6	9.18	275.39	7.70
3D-CNN + ResNet-18		91.78	3.85	115.62	5.46
3D-CNN + DenseNet-121		109.2	4.59	137.57	5.31
3D-CNN + MobileNet-224		84.8	3.56	106.83	5.96
3D-CNN + EfficientNet-B0		85.9	3.61	108.22	6.02
ViT-Base		166.6	7.00	209.88	4.05
(Our) Dual-Stream Former		168.6	7.08	212.40	3.46

Our architecture achieved the highest performance, with a CER of 3.46%. Dual-Stream Former excels in modeling spatial and temporal relationships simultaneously by leveraging a hierarchical structure that is optimized for video tasks. When combined with a Conformer for fine-grained temporal dependency modeling and a transformer Decoder with CTC/Attention for precise sequence decoding, this architecture provides a robust framework for VSR. The transformer-based visual front-end method is particularly effective for capturing global and local patterns across video sequences, making it the most suitable choice for this task.

The second-best performing model, ViT [18], achieved a CER of 4.05%. The Vision Transformer (ViT) processes spatial information by dividing the input frames into patches and training global spatial representations. Although its temporal modeling relies primarily on the former and transformer components, the ViT-based visual front-end effectively handles spatial feature extraction, particularly for high-resolution video data. This synergy makes the ViT architecture a strong alternative for tasks that emphasize spatial representation. Among the CNN-based architectures, DenseNet-121 [23] ranks third in terms of performance, achieving a CER of 5.31%. DenseNet-121′s densely connected layers promote feature reuse, and the 3D-CNN facilitates direct temporal modeling, highlighting the strength of the CNN in localized pattern recognition. However, it falls short of transformer-based models in capturing comprehensive spatiotemporal dependencies, indicating its reliance on the downstream components for temporal learning.

Other CNN-based architectures provide additional tradeoffs between computational efficiency and modeling complexity. ResNet-18 [20] achieves a CER of 5.46% and benefits from its lightweight and straightforward design, which helps reduce the risk of overfitting. However, its simplicity limits its ability to model complex spatial and temporal patterns. Similarly, MobileNet-224 [21] achieved a CER of 5.96%, offering computational efficiency suitable for resource-constrained environments, albeit at the cost of reduced spatial and temporal modeling capacities. EfficientNet-B0 [22] reported a CER of 6.02%, leveraging EfficientNet's scaling principles to balance model complexity and performance, although its ability to model temporal dependencies remains limited.

The VGG-16 [19] architecture achieved a CER of 7.70%. Although VGG-16 ensures stable spatial feature learning, its dated design and inefficiency in modeling temporal relationships result in suboptimal performance for video-based tasks. The baseline 3D-CNN, which lacked a dedicated front-end for advanced spatial feature learning, performed

the worst, with a CER of 10.71%. This highlights the necessity of robust visual front-end designs to achieve competitive performance.

These results highlight the complementary strengths and limitations of CNN and transformer-based visual front-end architectures. Although CNN-based models offer efficient localized feature extraction that is suitable for resource-constrained environments, they struggle to capture long-range spatiotemporal dependencies. By contrast, transformer-based models, including the Video Swin Transformer and ViT, effectively model the global context and temporal dynamics, making them more suitable for complex video-based tasks such as VSR. Overall, transformer-based front-ends demonstrate a clear advantage in scenarios that require holistic sequence understanding.

5.2. Convergence Speed and Performance

We compared the convergence speed and performance of the CNN-based (3D-CNN + DenseNet-121) and transformer-based (ViT and Video Swin Transformer) models (Figure 5). Among the tested models, the proposed Dual-Stream Former achieved the fastest convergence and lowest CER of 3.46%, despite having the second-largest parameter count (168.6 M). This result highlights that performance and convergence are influenced primarily by architectural efficiency rather than by parameter size alone. Specifically, Dual-Stream Former benefits from a local window-based self-attention mechanism that focuses on the computation of localized spatiotemporal regions, effectively reducing complexity while preserving essential features. Furthermore, hierarchical representation learning progressively integrates fine-grained local features into higher-level global representations to achieve a balance between efficiency and expressive capacity. These design characteristics enable Dual-Stream Former to be rapidly and stably optimized, making it particularly effective for high-dimensional video data. To assess generalizability, we compared the training and validation CERs across epochs for all the models; the trajectories exhibited closely aligned convergence without a widening training–validation gap, indicating limited overfitting under our protocol (Figure 5).

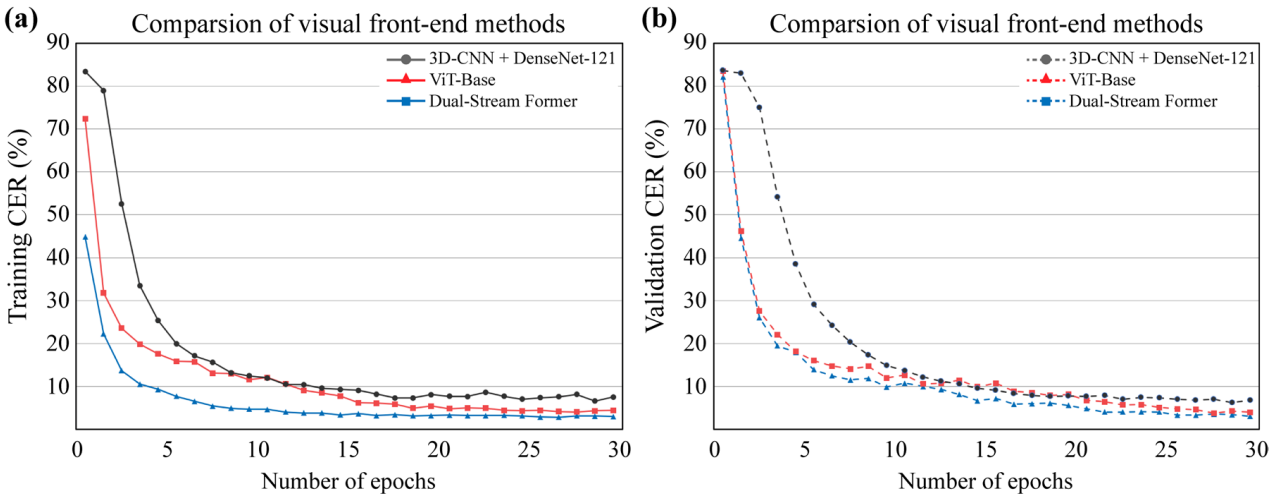

Figure 5. Comparison of training (**a**) and validation (**b**) CERs across epochs for different visual front-end architectures.

By contrast, ViT, with a parameter count of 166.6 M, converges at a moderate speed. ViT employs a global self-attention mechanism that effectively captures long-range dependencies across the entire spatiotemporal input. However, this global attention introduces considerable computational overhead because its complexity scales quadratically with the input size. This poses optimization challenges, particularly in high-dimensional video

tasks with lengthy token sequences. Despite these limitations, ViT achieves a competitive CER of 4.05%, demonstrating a strong capacity for spatiotemporal representation after convergence. Although powerful, its flat token structure lacks the hierarchical inductive biases inherent to the Video Swin Transformer, contributing to a slower initial optimization.

Although the CNN-based (3D-CNN + DenseNet-121) model had the smallest parameter count (109.2 M) among the three architectures, it exhibited the slowest convergence. The reliance on 3D convolutions for spatiotemporal processing imposes significant computational demands because these operations require the simultaneous processing of both spatial and temporal dimensions. Furthermore, the dense-connectivity pattern of DenseNet, which is effective for feature reuse and gradient flow in small datasets, is inefficient for large-scale video data. The deep and tightly connected structure of DenseNet results in slower gradient propagation and optimization challenges, particularly in the initial stages of training. Nevertheless, this architecture achieved a CER of 5.31%, demonstrating its ability to effectively model the local spatiotemporal features.

These findings highlight that convergence speed is not dictated solely by parameter count but rather by architectural design and its ability to efficiently capture spatiotemporal dependencies in video data. The Video Swin Transformer, with its hierarchical and localized attention mechanisms, has emerged as the most effective visual front-end, achieving an optimal balance between rapid convergence and high accuracy. This makes it particularly well suited for demanding VSR tasks, where both efficiency and performance are critical.

5.3. Comparison of Phonemes and Confusion Matrices

VSR faces significant challenges because of the inherent difficulty in distinguishing visually similar phonemes, which are often grouped as visemes [2]. Many phonemes, such as bilabial (/p/, /b/, /m/), alveolar (/t/, /d/, /n/), and labiodental (/f/, /v/) sounds, share highly similar or nearly indistinguishable lip shapes, making them difficult to differentiate without auditory cues. These categories are characterized by minimal or subtle lip movements that provide limited visual distinctions. Additionally, diphthongs (e.g., /ai/, /au/, /ei/, /ou/), which involve continuous transitions between two vowel sounds, further complicate recognition as they require the capture of precise, temporally dynamic articulatory patterns. Such phoneme types exhibit varying levels of visual discriminability in the VSR, posing significant obstacles to accurate lip reading.

5.3.1. Bilabial Sounds

Produced by closing and opening the lips, these sounds offer relatively clear visual cues, yet their lip shapes are nearly identical, making them challenging to differentiate (Table 4a).

5.3.2. Alveolar Sounds

Pronounced by the tongue contacting the alveolar ridge (e.g., /t/, /d/, /n/), these sounds involve minimal lip movement, which further complicates visual recognition (Table 4b).

5.3.3. Labiodental Sounds

Formed by the upper teeth touching the lower lip (e.g., /f/, /v/), these sounds provide subtle visual cues, making them among the most challenging to distinguish (Table 4c).

5.3.4. Diphthongs

Characterized by continuous lip and jaw movements when two vowels are blended within a single syllable, diphthongs present the greatest difficulty owing to their dynamic and varied visual patterns (Table 4d).

Table 4. A table of visually confusable phonemes categorized into bilabial, alveolar, labiodental, and diphthongs, illustrating overlapping visual characteristics of lip movements.

(a)	(b)	(c)	(d)
Bilabial	**Alveolar**	**Labiodental**	**Diphthongs**
Pat	Tip	Fan (/f/)	Buy (aI)
Bat	Dip	Van (/v/)	Bow (aU)
Mat	Nip	Fit (/f/)	Lie (aI)
Pan	Tan	ViT (/v/)	Low (oU)
Ban	Dan	Fear (/f/)	Tie (aI)
Man	Nan	Veer (/v/)	Tow (oU)
Map	Tap	Fizz (/f/)	Cry (aI)
Cap	Dap	Vis (/v/)	Cow (aU)
Lap	Nap	Fail (/f/)	Pie (aI)
Nap	Net	Veil (/v/)	Pow (aU)

We conducted a recognition study using 30 pairs of words that were particularly challenging in VSR. Among the examined phoneme categories, diphthongs presented the highest CER, with the CNN-based model recording 16.60% and the transformer-based model achieving a substantially lower CER of 10.39%, representing an absolute improvement of 6.21%. This indicates the superior ability of the transformer model to capture continuous and overlapping articulatory patterns, which require fine-grained temporal modeling. Similarly, labiodental sounds such as /f/ and /v/ showed a notable CER reduction from 14.08% (CNN) to 8.55% (transformer), a 5.53% improvement, reflecting the benefit of enhanced spatiotemporal attention in distinguishing subtle lip–teeth movements. For alveolar phonemes (/t/, /d/, and/n/), which often involve minimal lip movement and are visually ambiguous, the transformer-based model reduced the CER from 10.64% to 6.82%, indicating a 3.82% absolute gain. Finally, bilabial phonemes (/p/, /b/, /m/), although highly similar in terms of lip shape, also improved, with the CER decreasing from 9.49% (CNN) to 6.28% (transformer), a 3.21% reduction, suggesting an improved resolution of subtle visual cues during lip closure and release.

In this study, we compared the performance of CNN (3D-CNN + DenseNet-121)- and transformer (Video Swin Transformer)-based visual front-end structures in VSR, focusing on their ability to handle visually confusable phoneme groups (Table 5): bilabial, alveolar, labiodental, and diphthongs. The experimental results, including confusion matrices and quantitative metrics, indicate that the transformer-based approach consistently outperformed the CNN-based method across all the phoneme categories.

Table 5. Performance of CNN (3D-CNN + DenseNet-121)- and transformer (Video Swin Transformer)-based visual front-end structures.

Phonemes	CNN	Transformer
CER (%)	5.31	3.46
Bilabial (%)	9.44	5.26
Alveolar (%)	11.24	6.39
Labiodental (%)	15.7	9.25
Diphthong (%)	16.6	10.39

Transformers excel at capturing long-range temporal dependencies, which are critical for recognizing dynamic phonemes, such as diphthongs. Diphthongs (Figure 6g,h) involve continuous transitions between two vowels accompanied by complex and overlapping lip

and jaw movements. The transformer-based structure achieved a significantly lower CER of 10.39% compared with 16.60% for the CNN-based structure. This performance gain can be attributed to the self-attention mechanism of transformers, which enables the effective modeling of temporal variations and subtle visual cues across frames.

Labiodental sounds (Figure 6e,f) (/f/, /v/), characterized by subtle interactions between the upper teeth and lower lip, pose a significant challenge in VSR owing to limited visual distinctiveness. The transformer-based approach demonstrated a marked improvement in recognizing labiodental phonemes (9.25% CER) compared to CNNs (15.70%). This is likely because of the transformer's ability to focus on small, localized features while simultaneously considering global contextual information, thereby reducing confusion between similar phonemes.

For alveolar sounds (Figure 6c,d) (/t/, /d/, /n/), which involve minimal lip movement and are articulated primarily through tongue placement, the transformer outperformed the CNNs (6.39% vs. 11.24%). This result highlights the ability of the transformer to extract subtle visual features associated with minimal articulatory changes, thereby overcoming the limitations of CNNs, which often rely on prominent spatial features.

Bilabial sounds (Figure 6a,b) (/p/, /b/, /m/), which presented clear visual cues owing to lip closure and opening, were recognized with high accuracy by both models. However, the transformer-based structure achieved a marginally lower accuracy (5.26%) than the CNNs (9.44%), demonstrating its superior ability to generalize, even for visually distinct phonemes. Quantitative analysis revealed a significant performance advantage of the transformer-based structure over the CNN-based structure, with an overall CER improvement of approximately 2% (3.46% vs. 5.31%). This improvement underscores the transformer's superior ability to model both the spatial and temporal aspects of visual speech data, leading to more robust recognition across diverse phoneme categories.

The confusion matrices (Figure 6) indicate that the transformer-based models significantly reduced misclassification, particularly among the visually confusable phoneme pairs. Notable improvements were observed in distinguishing /f/ from /v/ (labiodental), /ai/ from /au/ (diphthongs), and /t/ from /d/ (alveolar), whereas the CNN-based models presented higher confusion rates. The self-attention mechanism enables the transformer to capture subtle articulation differences over time, which are often overlooked by CNNs because of their limited temporal receptive fields.

Ablation considerations are as follows. Although the primary focus of this study is on the overall effectiveness of the Dual-Stream Former relative to CNN- and flat transformer-based baselines, we note that preliminary internal experiments provide indirect evidence of the contribution of individual modules. For example, replacing the Conformer with a standard transformer encoder led to a noticeable increase in CER, and removing the hierarchical Swin front-end in favor of a flat CNN degraded recognition performance, particularly on visually confusable phoneme groups such as diphthongs and labiodentals. These observations are consistent with our confusion matrix analysis (Figure 6), where improvements are concentrated in categories that benefit from both fine-grained temporal modeling and hierarchical spatial encoding. A full ablation study disentangling the contribution of each component was not feasible within the current study's scope because of the computational cost of retraining multiple large-scale video transformer variants. Nevertheless, we plan to conduct systematic ablation experiments in future work to more rigorously quantify the relative impact of each component.

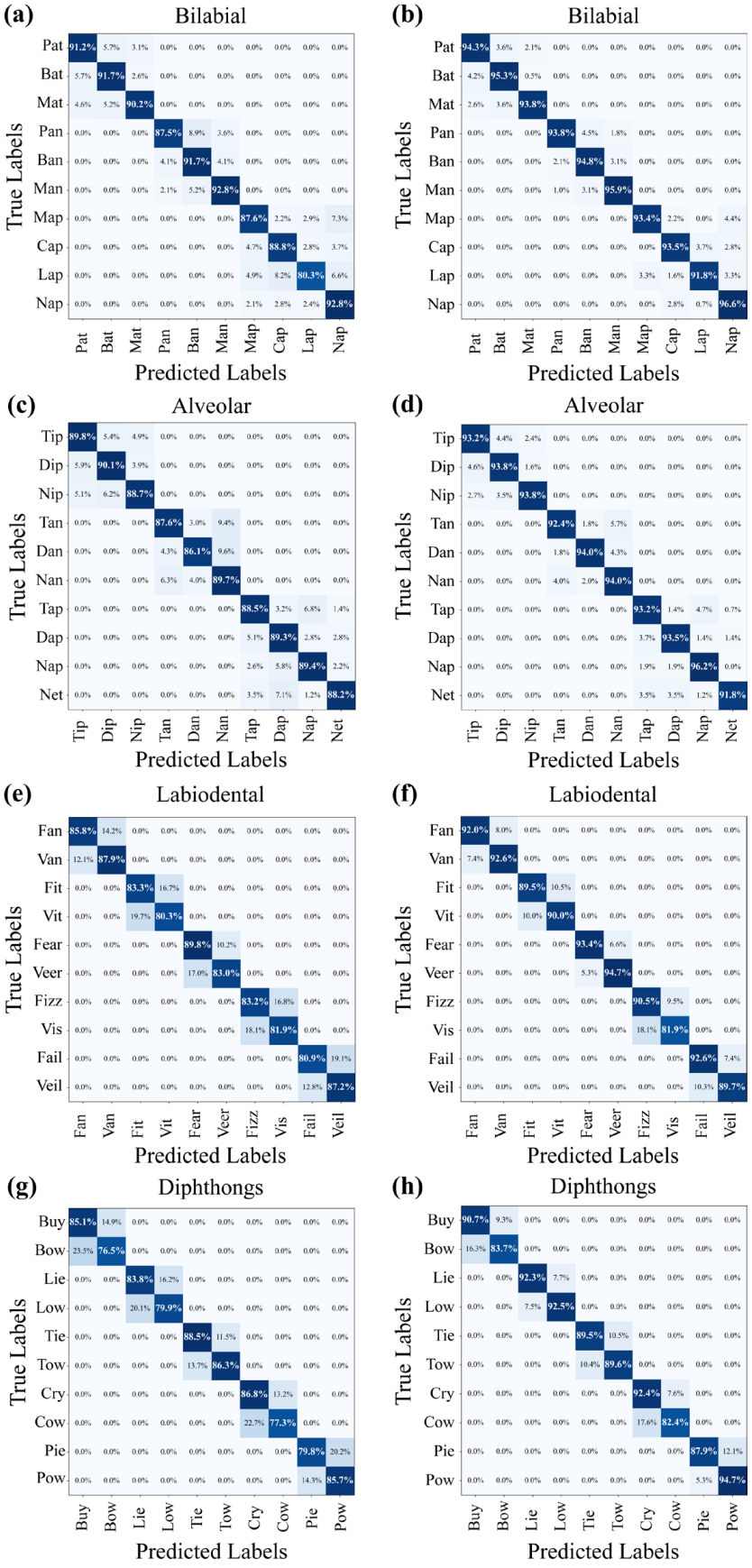

Figure 6. Confusion matrices for visually confusable phonemes across four categories: bilabial, alveolar, labiodental, and diphthongs, evaluated using two visual front-ends. Subfigures (**a**,**c**,**e**,**g**) depict the results from the CNN-based visual front-end, whereas (**b**,**d**,**f**,**h**) represent the results from the transformer-based visual front-end.

5.4. Statistical Analysis

To further validate the performance of the evaluated visual front-end architectures, we conducted a statistical analysis using paired *t* tests based on the CERs of four models: 3D CNN, 3D CNN with DenseNet-121, ViT, and Video Swin Transformer (Figure 7). The evaluation involved ten participants (mean age: 27.9 years), all of whom were proficient in IT but who were not involved in dataset construction. Each participant randomly selected ten words from Table 2 and repeated them ten times, yielding 100 evaluations per participant. To mitigate phoneme imbalance, the word pool was designed to cover a representative distribution of major phoneme types, ensuring that no categories were over- or under-represented in the test set.

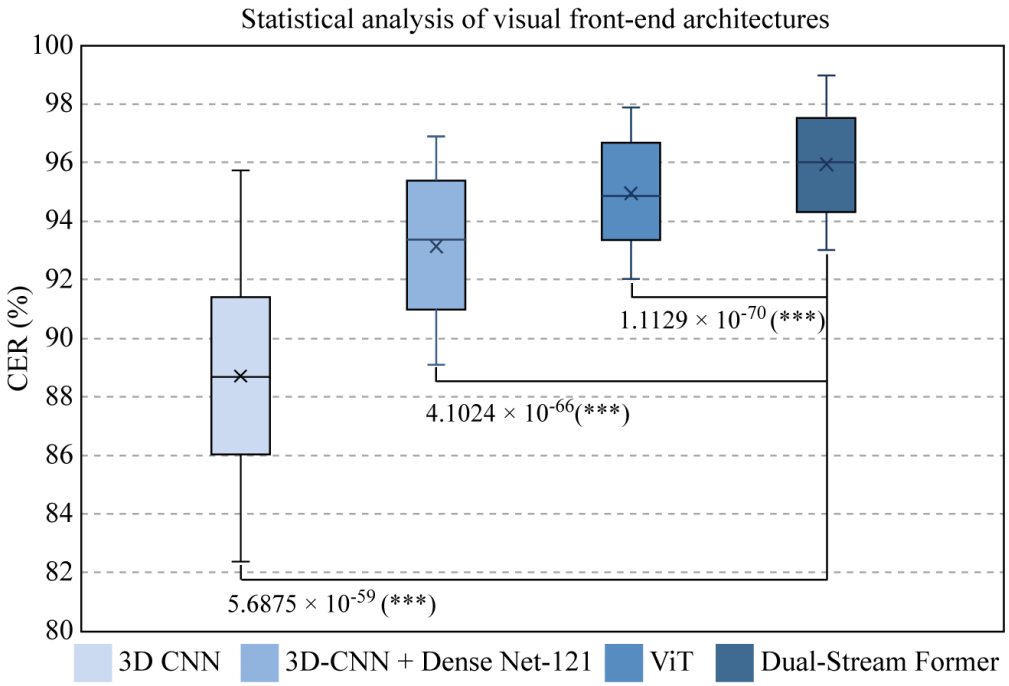

Figure 7. Statistical comparison of CER across four visual front-end architectures: 3D CNN, 3D CNN with DenseNet-121, ViT, and Dual-Stream Former. Box plots represent CER distributions across 10 participants, with mean (×) and median (line) indicated. Paired *t* tests revealed statistically significant improvements of Dual-Stream Former over all other models, with corresponding *p* values and effect sizes (Cohen's d): 3D CNN ($p = 5.6875 \times 10^{-59}$, d = 2.41), 3D CNN + DenseNet-121 ($p = 4.1024 \times 10^{-66}$, d = 2.12), and ViT ($p = 1.1129 \times 10^{-70}$, d = 1.89). These results highlight the superior accuracy and stability of attention-based architectures for visual speech recognition. (* indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$).

Among the architectures, the Video Swin Transformer achieved the lowest average CER (4.06%), with a standard deviation (SD) of 1.79%, indicating both strong accuracy and stable performance across participants. By contrast, the 3D CNN yielded the highest average CER (11.29%) with the greatest variability (SD = 3.76%), reflecting less consistent results.

Statistical comparisons using paired *t* tests revealed that the transformer-based models (ViT and Dual-Stream Former) significantly outperformed their CNN-based counterparts. As illustrated in Figure 7, Dual-Stream Former demonstrated the lowest CER among all the models, with statistically significant differences compared with 3D CNN ($p = 5.6875 \times 10^{-59}$, Cohen's d = 2.41), 3D CNN + DenseNet-121 ($p = 4.1024 \times 10^{-66}$, d = 2.12), and ViT ($p = 1.1129 \times 10^{-70}$, d = 1.89). These effect sizes indicate a large-to-very large practical significance, emphasizing the substantial impact of the Dual-Stream Former architecture. Moreover, the narrow interquartile range of the box plots underscores its

consistency and robustness across participants, highlighting the architectural advantages of localized self-attention and hierarchical modeling in VSR.

In addition to the overall CER, we analyzed the phoneme-level error distributions. Dual-Stream Former substantially reduced the number of errors in visually confusable categories such as diphthongs and labiodentals compared to CNN- and transformer-based baselines (Table 5; Figure 6), providing more fine-grained linguistic evidence of its effectiveness.

6. Discussion

This study investigated the effectiveness of the proposed Dual-Stream Former architecture for VSR, demonstrating notable performance improvements over both CNN-based- and transformer-based baselines. Dual-Stream Former achieved a CER of 3.46%, outperforming 3D-CNN + DenseNet-121 (CER: 5.31%) and Vision Transformer (CER: 4.05%). To assess robustness beyond statistical significance, we computed 95% confidence intervals (95% CIs) for CER using participant-wise bootstrap resampling with 1000 iterations. Dual-Stream Former achieved a CER of 4.06% (95% CI: [3.70%, 4.41]), confirming that its superiority is consistent across speakers. The full CI results for all the models are provided in the Supporting Information (Table S4). These improvements can be attributed to the synergy between the Video Swin Transformer's localized hierarchical attention mechanism and the Conformer's precise temporal modeling, which enables the network to capture both global and fine-grained spatiotemporal features.

The model also excelled at recognizing visually confusing phonemes. Diphthongs, known for their continuous transitions and articulatory complexity, were recognized with a CER of 10.39%, which is substantially lower than the 16.60% observed in CNN-based models. Similarly, for labiodental phonemes such as /f/ and /v/, the transformer-based approach achieved a CER of 9.25% compared with 15.70% with CNNs. These results highlight the ability of the transformer to model subtle and overlapping visual cues that CNNs typically struggle to distinguish.

Another strength of Dual-Stream Former is its efficient training behavior. Despite having 168.6 million parameters, second only among the evaluated models, it converged faster than the other models. This efficiency is partly due to its localized window-based attention, which limits the computational overhead by focusing on the most informative spatiotemporal regions, a mechanism shown to significantly reduce complexity and accelerate convergence in other domains [44]. In addition, the hierarchical structure facilitates the progressive abstraction of features, contributing to both generalizability and speed.

Within the controlled frontal-pose setting evaluated in this study, the dataset's scale (740,000 utterances; 616.67 h), balanced speaker composition (20 females/20 males; 40 total), and lexical diversity (185 classes across nine topical categories; Table 2) are adequate to support generalization across speakers and lexical categories. Nevertheless, broader generalizability to in-the-wild conditions—e.g., spontaneous speech, lighting variation, occlusions, and larger head motion—should be established in future work; we plan to conduct cross-dataset evaluations and additional data collection under more diverse capture settings.

With regard to novelty and added value, the proposed dual-branch Video Swin + Conformer architecture attained the lowest CER (3.46%) with faster convergence (Figure 5) and yielded disproportionate gains for visually confusable phoneme groups—especially diphthongs and labiodentals (Table 5; Figure 6). Evaluating on a large-scale, a high-resolution dataset using a fixed within-subject validation split (Sections 4.1–4.3; Table 2) and reporting participant-wise tests (Section 5.4; Figure 7) provide reproducible evidence that hierarchical video attention, paired with sequence-level Conformer modeling, improves

visual-only speech recognition. This clarifies the conditions under which attention-based visual front-ends outperform CNNs and flat transformers.

However, this study has several limitations. First, the high parameter counts and computational demands of the model may hinder the deployment of resource-constrained devices. Second, although extensive, the dataset was collected under controlled and repetitive conditions, which potentially limits its generalizability to real-world scenarios. Factors such as spontaneous speech, lighting variations, occlusions, and head movements have not been sufficiently addressed. These aspects can influence the performance in real-time or with in-the-wild settings and may increase the risk of overfitting. Finally, although the model showed promising results in English, its performance in other languages and multilingual datasets remains unexplored.

Regarding the choice of hybrid CTC/Attention, our architecture combines CTC and attention to exploit their complementary strengths. CTC provides efficient frame-level alignment but has limited language modeling capacity, whereas pure attention-based models capture richer dependencies but often suffer from unstable alignments and slower convergence. The hybrid approach integrates both advantages, resulting in more stable training and improved accuracy. This balance has been consistently reported in prior speech recognition research and is confirmed in our experiments, indicating that the hybrid method is particularly well suited for visual speech recognition.

Future studies should focus on developing lightweight Dual-Stream Former variants for real-time application and mobile deployment. Techniques such as attention pruning, sparse computation, and adaptive inference can reduce latency without sacrificing performance. Furthermore, extending the model to multilingual VSR introduces new challenges, including language-specific visual overlap and pronunciation patterns. Exploring audio-visual integration may also enhance robustness under conditions where visual input is insufficient, such as in low-light environments or during rapid articulation.

Model compression techniques—including pruning, quantization, and knowledge distillation—are promising approaches for reducing the 168.6 M parameters of our model. Although the present study focuses on establishing a full-capacity baseline, systematic exploration of lightweight variants remains a key direction for future research to enable deployment on mobile and embedded platforms.

Ethical considerations are as follows. Although visual speech recognition enables accessibility and user-facing applications, it could also raise concerns of surveillance or non-consensual use. Our study is explicitly scoped to consent-based, user-facing contexts. We stress that the deployment of lipreading systems must avoid covert applications, and safeguards such as ROI-only processing, pseudonymization, and restricted access are essential for responsible use.

A further limitation is that we did not adopt k-fold or nested cross-validation for hyperparameter tuning and model selection; instead, we used a fixed, within-subject validation split for computational feasibility. Future work will evaluate speaker-aware k-fold (or leave-one-speaker-out) cross-validation to assess split sensitivity more fully.

Another limitation concerns the demographic composition of our dataset. All participants were young adults (mean age = 29.8 years, range 22–38 years), and no subject older than 40 years was included. The cohort primarily consisted of Asian, Egyptian, and Pakistani participants, all of whom were proficient in English at a near-native level. Although this recruitment strategy facilitated smoother execution of HCI-oriented experiments—because younger adults are typically more familiar with IT devices—it also restricts the generalizability of our findings. Future work should therefore expand recruitment to older adults and more diverse populations to ensure robustness, fairness, and ecological validity.

In addition, this study did not include a full ablation analysis to isolate the contributions of individual architectural components (e.g., Video Swin front-end, Conformer back-end, or dual-branch design). Although indirect evidence from confusion matrices and preliminary observations suggest that each component contributes meaningfully to the observed performance gains, a systematic ablation study was not feasible within the current study's scope owing to the computational cost of retraining multiple large-scale video transformer variants. Future work should conduct structured ablation experiments to rigorously quantify the relative impact of each module.

7. Conclusions

This study introduced Dual-Stream Former, a hybrid transformer-based architecture that integrates a Video Swin Transformer with a Conformer module, to address the key challenges in VSR. The model achieved a CER of 3.46%, outperforming both CNN-based and prior transformer-based alternatives in terms of accuracy, consistency, and convergence speed.

Dual-Stream Former demonstrated particularly strong performance in recognizing challenging phoneme types, such as diphthongs and labiodentals, owing to its capacity for model overlapping and fine-grained spatiotemporal features. Although the large dataset and high-resolution video recordings contributed to its robustness, the model architecture played a critical role in achieving these results.

Although this model has significant computational requirements, its performance underscores the feasibility of using transformer-based VSR systems in high-precision applications. With further optimization and adaptation, Dual-Stream Former has the potential to serve a wide range of real-world applications, including silent communication tools, accessibility technologies for hearing-impaired individuals, and privacy-aware interfaces in public environments.

In summary, Dual-Stream Former has strong potential for advancing the field of VSR. Ongoing efforts to improve efficiency, generalizability, and multimodal integration are essential for realizing full applicability in diverse and dynamic speech recognition scenarios.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ai6090222/s1, Supplementary Equation (S1). Video Swin Transformer; Supplementary Equation (S2). Conformer Encoder; Supplementary Equation (S3). Transformer Decoder; and Supplementary Equation (S4). CTC/Attention Hybrid Loss; Table S1. Comparison of our dataset with existing word-level audio–visual speech datasets across key metrics such as tasks, speakers, and resolution; Table S2. Overview of categorized keywords used for visual speech recognition dataset creation; Table S3. Performance comparison of visual front-end architectures and encode–decoder combinations based on Average CER (%); Table S4. CER (%) and 95% confidence intervals (CIs) across models, computed via participant-wise boot-strap resampling (1000 iterations).

Author Contributions: S.J.: writing—original draft; data curation; investigation; validation; software supervision; visualization; conceptualization. J.L.: writing—review and editing; methodology. Y.-J.L.: funding acquisition; project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government, Ministry of Science and ICT (MSIT), grant number 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration; and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government, Ministry of Science and ICT (MSIT), grant number RS-2022-00187238, Development of Large Korean Language Model Technology for Efficient Pretraining.

Institutional Review Board Statement: In accordance with Article 13 of the Enforcement Rules of Research on Human Subjects, studies that do not collect or record personally identifiable information and involve one or more of the following conditions are exempt from IRB deliberation. Specifically, this includes studies that utilize only simple contact measurement devices or observation equipment that do not induce any physical changes. In this study, only human voice and lip-movement data were collected using non-invasive observation equipment, and no personally identifiable information was recorded. Therefore, the study qualifies for exemption from IRB review, and no IRB registration number is issued based on the stated indemnification criteria.

Informed Consent Statement: Each participant provided written and oral informed consent.

Data Availability Statement: The original contributions presented in this study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

VSR	Visual speech recognition
CER	Character error rate
CNN	Convolutional neural network
ViT	Vision transformer
STCNN	Spatiotemporal CNN
CTC	Connectionist temporal classification
SD	Standard deviation

References

1. Sheng, C.; Kuang, G.; Bai, L.; Hou, C.; Guo, Y.; Xu, X.; Pietikäinen, M.; Liu, L. Deep learning for visual speech analysis: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 6001–6022. [CrossRef] [PubMed]
2. Assael, Y.M.; Shillingford, B.; Whiteson, S.; De Freitas, N. LipNet: End-to-end sentence-level lipreading. *arXiv* **2016**, arXiv:1611.01599. [CrossRef]
3. Jeon, S.; Lee, J.; Yeo, D.; Lee, Y.; Kim, S. Multimodal audiovisual speech recognition architecture using a three-feature multi-fusion method for noise-robust systems. *ETRI J.* **2024**, *46*, 22–34. [CrossRef]
4. Radha, N.; Shahina, A.; Khan, A.N. A survey on visual speech recognition approaches. In Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 25–27 March 2021; pp. 934–939. [CrossRef]
5. Chang, O.; Liao, H.; Serdyuk, D.; Shah, A.; Siohan, O. Conformer is all you need for visual speech recognition. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 10136–10140. [CrossRef]
6. Park, Y.-H.; Park, R.-H.; Park, H.-M. Swinlip: An efficient visual speech encoder for lip reading using Swin transformer. *SSRN* **2024**. [CrossRef]
7. Ma, P.; Petridis, S.; Pantic, M. Visual speech recognition for multiple languages in the wild. *Nat. Mach. Intell.* **2022**, *4*, 930–939. [CrossRef]
8. Tuli, S.; Dasgupta, I.; Grant, E.; Griffiths, T.L. Are Convolutional Neural Networks or Transformers more like human vision? *arXiv* **2021**, arXiv:2105.07197. [CrossRef]
9. Bajgoti, A.; Gupta, R.; Balaji, P.; Dwivedi, R.; Siwach, M.; Gupta, D. Swinanomaly: Real-time video anomaly detection using video Swin transformer and sort. *IEEE Access* **2023**, *11*, 111093–111105. [CrossRef]
10. Arakane, T.; Saitoh, T. Efficient DNN model for word lip-reading. *Algorithms* **2023**, *16*, 269. [CrossRef]
11. Petridis, S.; Stafylakis, T.; Ma, P.; Cai, F.; Tzimiropoulos, G.; Pantic, M. End-to-end audiovisual speech recognition. *arXiv* **2018**, arXiv:1802.06424. [CrossRef]
12. Jeon, S.; Elsharkawy, A.; Kim, M.S. Lipreading architecture based on multiple convolutional neural networks for sentence-level visual speech recognition. *Sensors* **2021**, *22*, 72. [CrossRef]
13. Ji, S.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231. [CrossRef]

14. Jiang, M.; Khorram, S.; Fuxin, L. Comparing the decision-making mechanisms by transformers and CNNs via explanation methods. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 9546–9555. [CrossRef]
15. Lahoud, J.; Cao, J.; Khan, F.S.; Cholakkal, H.; Anwer, R.M.; Khan, S.; Yang, M.H. 3D vision with transformers: A survey. *arXiv* **2022**, arXiv:2208.04309. [CrossRef]
16. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video Swin transformer. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 3192–3201. [CrossRef]
17. Selva, J.; Johansen, A.S.; Escalera, S.; Nasrollahi, K.; Moeslund, T.B.; Clapés, A. Video transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12922–12943. [CrossRef] [PubMed]
18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2021**, arXiv:2010.11929. [CrossRef]
19. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2015**, arXiv:1409.1556. [CrossRef]
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
21. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861. [CrossRef]
22. Tan, M.; Le, Q.V. Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019.
23. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]
24. Son Chung, J.; Senior, A.; Vinyals, O.; Zisserman, A. Lip reading sentences in the wild. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3444–3453. [CrossRef]
25. Stafylakis, T.; Tzimiropoulos, G. Combining residual networks with LSTMs for lipreading. *arXiv* **2017**, arXiv:1703.04105. [CrossRef]
26. Martinez, B.; Ma, P.; Petridis, S.; Pantic, M. Lipreading using temporal convolutional networks. *arXiv* **2020**, arXiv:2001.08702. [CrossRef]
27. Ma, P.; Martinez, B.; Petridis, S.; Pantic, M. Towards practical lipreading with distilled and efficient models. *arXiv* **2021**, arXiv:2007.06504. [CrossRef]
28. Zhang, X.; Cheng, F.; Shilin, W. Spatio-temporal fusion based convolutional sequence learning for lip reading. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 713–722. [CrossRef]
29. Sheng, C.; Pietikäinen, M.; Tian, Q.; Liu, L. Cross-modal self-supervised learning for lip reading: When contrastive earning meets adversarial training. In *MM '21: Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event China, 20–24 October 2021*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 2456–2464. [CrossRef]
30. Feng, D.; Yang, S.; Shan, S.; Chen, X. Learn an effective lip reading model without pains. *arXiv* **2020**, arXiv:2011.07557. [CrossRef]
31. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
32. Prajwal, K.R.; Afouras, T.; Zisserman, A. Sub-word level lip reading with visual attention. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5152–5162. [CrossRef]
33. Peng, Y.; Sakai, Y.; Funabora, Y.; Yokoe, K.; Aoyama, T.; Doki, S. Funabot-sleeve: A wearable device employing McKibben artificial muscles for haptic sensation in the forearm. *IEEE Robot. Autom. Lett.* **2025**, *10*, 1944–1951. [CrossRef]
34. Watanabe, S.; Hori, T.; Kim, S.; Hershey, J.R.; Hayashi, T. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1240–1253. [CrossRef]
35. Movellan, J.R. Visual speech recognition with stochastic networks. In *Advances in Neural Information Processing Systems 7 (NIPS 1994)*; MIT Press: Cambridge, MA, USA, 1994; Volume 7.
36. Pigeon, S.; Vandendorpe, L. The M2VTS multimodal face database (Release 1.00). In *Lecture Notes in Computer Science, Proceedings of the Audio- and Video-Based Biometric Person Authentication, Crans-Montana, Switzerland, 12–14 March 1997*; Bigün, J., Chollet, G., Borgefors, G., Eds.; Springer: Berlin/Heidelberg, Germany, 1997; Volume 1206, pp. 403–409. [CrossRef]
37. Matthews, I.; Cootes, T.F.; Bangham, J.A.; Cox, S.; Harvey, R. Extraction of visual features for lipreading. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 198–213. [CrossRef]

38. Cox, S.J.; Harvey, R.W.; Lan, Y.; Newman, J.L.; Theobald, B.J. The challenge of multispeaker lip-reading. In Proceedings of the Auditory-Visual Speech Processing 2008 (AVSP 2008), Tangalooma, Australia, 26–29 September 2008; pp. 179–184.
39. Rekik, A.; Ben-Hamadou, A.; Mahdi, W. An adaptive approach for lip-reading using image and depth data. *Multimed. Tools Appl.* **2016**, *75*, 8609–8636. [CrossRef]
40. Petridis, S.; Shen, J.; Cetin, D.; Pantic, M. Visual-only recognition of normal, whispered and silent speech. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 6219–6223. [CrossRef]
41. Suh, S.; Park, S.; Jeong, Y.; Lee, T. *Designing Acoustic Scene Classification Models with CNN Variants*; DCASE Community: Washington, DC, USA, 2020.
42. Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; Pantic, M. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops, Sydney, Australia, 2–8 December 2013; pp. 397–403. [CrossRef]
43. Ma, P.; Petridis, S.; Pantic, M. End-to-end audio-visual speech recognition with conformers. *arXiv* **2021**, arXiv:2102.06657. [CrossRef]
44. Geng, S.; Zhai, S.; Li, C. Swin transformer based transfer learning model for predicting porous media permeability from 2D images. *Comput. Geotech.* **2024**, *168*, 106177. [CrossRef]