Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

Vol. 7, No. 2, June, 2026

A Multimodal Preprocessing Pipeline for Robust Audio-Visual Speech Separation and Recognition

Sara M. Sh [1*], Baraa M. Albaker [1]

[1] *College of Engineering, Al-Iraqia University, Saba'a Abkar Complex, Baghdad, Iraq.*

Received 23 February 2026; Revised 11 May 2026; Accepted 14 May 2026; Published 01 June 2026

Abstract

Audiovisual speech separation aims to improve speech intelligibility in challenging real-world environments, such as noisy meetings and multi-speaker acoustic scenes. However, many existing approaches rely on computationally intensive architectures or unstable multimodal representations, limiting robustness and practical deployment. This study proposes a multimodal audiovisual speech separation framework based on a structured preprocessing pipeline and a lightweight hybrid deep learning architecture. The proposed method enforces geometric, photometric, temporal, and statistical consistency across audio and visual streams. The visual pathway employs grayscale conversion, histogram equalization, face detection, spatial normalization, and eigenface-based PCA feature extraction to obtain stable articulatory representations, while the audio pathway incorporates pre-emphasis filtering, normalization, resampling, and MFCC-based feature extraction with vector-level equalization. The resulting representations are processed using a hybrid Conv1D–LSTM–GRU architecture for efficient temporal modeling. Experimental evaluation on the AVSpeech dataset achieved an average SDR of 19.33 dB, SIR of 15.72 dB, and PESQ of 4.22. The proposed HAVS-Net architecture contains approximately 227K trainable parameters while maintaining robust performance and efficient computational behavior under diverse real-world conditions.

Keywords: Audiovisual Speech Separation; Feature Normalization; MFCC; PCA; Lightweight Deep Learning; AVSpeech Dataset.

1. Introduction

Real-world communication systems increasingly rely on robust speech processing capabilities to support applications such as online meetings, remote collaboration platforms, and multimedia content captured in unconstrained environments [1, 2]. In these scenarios, speech signals are frequently degraded by background noise, speaker overlap, and dynamic acoustic interference, making accurate speech separation a challenging task [3]. The absence of controlled acoustic conditions further intensifies this challenge, particularly in multi-speaker environments where non-stationary noise and competing speech sources significantly reduce signal clarity and intelligibility [4]. A representative example of such conditions includes online meetings or crowded conference environments, where overlapping speakers, background conversations, and environmental noise significantly degrade speech intelligibility and complicate speaker separation tasks.

Traditional audio-only speech separation approaches are inherently limited under such conditions, as they rely solely on acoustic cues that become unreliable in the presence of severe interference [5]. As a result, their performance deteriorates in real-world environments, where complex auditory scenes introduce ambiguity in signal reconstruction [6]. To address these limitations, audiovisual speech separation has emerged as a promising alternative, leveraging visual cues such as lip movements and facial expressions to improve separation performance and robustness when acoustic signals are corrupted [7].

* Corresponding author: sara.m@aliraqia.edu.iq

https://doi.org/10.28991/HIJ-2026-07-02-023

Early audiovisual approaches relied on handcrafted feature representations, including Mel-Frequency Cepstral Coefficients (MFCC) for audio and basic visual descriptors [8]. While MFCC provides a compact and perceptually meaningful representation of the speech spectrum, its effectiveness is limited in highly non-stationary noise conditions [9]. Similarly, statistical techniques such as Principal Component Analysis (PCA) have been employed as eigenface-based visual feature extraction methods for capturing stable articulatory representations [10]. In addition to dimensionality reduction, PCA preserves the most informative facial variance components while reducing redundancy and computational complexity. Although PCA is inherently limited to linear transformations, sensitivity to severe illumination variation remains a common challenge across many visual feature extraction approaches [11].

Recent advances in deep learning, particularly convolutional and recurrent neural networks, have significantly improved speech separation performance by enabling effective temporal modeling [12]. More recently, transformer-based and diffusion-based models have achieved state-of-the-art results by capturing long-range dependencies and complex cross-modal interactions [13]. However, these approaches typically require substantial computational resources and large-scale labeled datasets, and their performance does not necessarily generalize well across heterogeneous real-world conditions [14].

Despite these advancements, a critical challenge remains insufficiently addressed in the literature: the stability and consistency of multimodal feature representations [11]. In many existing works, preprocessing is treated as a secondary step rather than a fundamental component of system design. Consequently, variations in visual alignment, illumination conditions, temporal synchronization, and feature scaling introduce instability in multimodal representations, negatively affecting model convergence and cross-modal fusion efficiency [15].

Motivated by these limitations, this study proposes a multimodal audiovisual speech separation framework based on a structured, stability-driven preprocessing pipeline combined with a lightweight HAVS-Net hybrid deep learning architecture containing approximately 227K trainable parameters. The proposed approach enforces geometric, photometric, temporal, and statistical consistency across audio and visual streams prior to model learning. By emphasizing representation stability, computational efficiency, and generalization-oriented design, the proposed framework aims to achieve robust speech separation performance under realistic and diverse audiovisual conditions.

2. Literature Review

Audio-visual speech separation has attracted significant attention as an effective solution for speech enhancement in complex acoustic environments, particularly under conditions involving noise, reverberation, and overlapping speakers [1]. Early studies have demonstrated that audio-only speech separation systems suffer from substantial performance degradation in such scenarios, primarily due to their reliance on acoustic cues that become unreliable under severe interference [5]. These limitations motivated the integration of visual information, such as lip movements and facial dynamics, to improve robustness and separation accuracy [16].

Initial audiovisual approaches relied on handcrafted feature representations, where Mel-Frequency Cepstral Coefficients (MFCC) were combined with basic visual descriptors [8]. While these methods provided compact and interpretable representations, their performance remained sensitive to environmental variations, including illumination changes and facial misalignment. Subsequent developments introduced learning-based approaches, such as deep clustering and convolutional neural networks, which improved speaker discrimination and temporal modeling capabilities. However, these methods still depended heavily on stable input conditions and controlled datasets, limiting their effectiveness in real-world scenarios [17].

More recent approaches have leveraged advanced deep learning architectures, including transformer-based and diffusion-based models, to capture long-range temporal dependencies and complex cross-modal interactions [18]. These models have demonstrated strong performance in speech separation tasks, particularly under low signal-to-noise ratio conditions. Nevertheless, their practical deployment is constrained by high computational complexity, large training data requirements, and sensitivity to audiovisual synchronization errors [19].

To address computational challenges, lightweight and efficient audiovisual architectures have been proposed. These models aim to reduce latency and enable real-time processing; however, they often suffer from reduced modeling capacity and remain vulnerable to preprocessing inconsistencies and visual noise [20]. In addition, several studies have shown that even small misalignments between audio and visual streams can significantly degrade model convergence and separation performance, highlighting the importance of temporal consistency and stable feature representation [21].

A comparative summary of representative audiovisual speech separation methods is presented in Table 1. The table highlights a clear trade-off in existing approaches: high-complexity models achieve strong performance but require substantial computational resources, whereas lightweight models improve efficiency at the cost of robustness and fine-grained modeling capability. Furthermore, many methods exhibit sensitivity to preprocessing variations, such as misalignment, illumination changes, and feature scaling inconsistencies.

Table 1. Comparative analysis of representative audiovisual speech separation methods

Study	Model Type	Audio Representation	Visual Representation	Dataset	Strengths	Limitations	Complexity
[22]	AV Codec	Spectrogram	CNN-based	AVSpeech	High perceptual quality	High computational cost	Very High
[23]	Diffusion	Spectrogram	CNN-based	AVSpeech	Strong reconstruction at low SNR	High latency (iterative inference)	Very High
[24]	Diffusion	Log-Mel	CNN-based	LRS3	Robust enhancement performance	Expensive inference process	Very High
[25]	Transformer	Spectrogram	Attention-based	LRS3	Captures long-range dependencies	Requires large-scale data	Very High
[26]	Attention-based	Spectrogram	Attention-based	VoxCeleb2	Effective multimodal fusion	Sensitive to misalignment	High
[27]	Lightweight AV	Spectrogram	Light CNN	AVSpeech	Real-time capability	Limited fine-grained modeling	Low
[20]	Efficient AV	Spectrogram	CNN-based	LRS2	Low latency	Reduced performance in complex noise	Low
[28]	Waveform AV	Raw waveform	CNN-based	AVSpeech	End-to-end processing	Sensitive to preprocessing instability	High
[29]	AV-HuBERT	Learned embeddings	Transformer-based	LRS3	Strong representation learning	Heavy pre-trained model	Very High

Following the comparison in Table 1, it becomes evident that existing research primarily focuses on improving model architecture, often at the expense of computational efficiency and robustness. High-performance models achieve strong results through increased complexity, while efficient models sacrifice accuracy and stability under challenging conditions. More importantly, the majority of existing approaches treat preprocessing as a secondary component, despite its critical influence on feature consistency, model convergence, and cross-modal fusion.

While recent studies have demonstrated the effectiveness of CNN-based visual feature extractors in capturing complex spatial representations [30], such approaches typically require large-scale training data and introduce significant computational overhead. In addition, robust face localization remains essential in unconstrained audiovisual datasets, particularly under pose variation and partial profile-face conditions. In contrast, the present work adopts Principal Component Analysis (PCA) as a stable and computationally efficient feature extraction method. The choice of PCA is motivated by its ability to provide stable and compact representations without introducing additional trainable parameters [31].

In addition, PCA performs explicit dimensionality reduction by projecting high-dimensional facial data into a compact subspace, thereby reducing redundancy and noise while preserving the most informative variance components. Unlike deep visual encoders, PCA operates directly on normalized facial inputs, making it less sensitive to overfitting and more stable under moderate illumination and alignment variations [11]. Although PCA is inherently limited to linear transformations and may not fully capture highly nonlinear articulatory patterns, this limitation represents an intentional design trade-off in favor of stability, efficiency, and generalization. This design aligns with the core objective of the proposed framework, which emphasizes representation stability and computational efficiency. Future work will explore the integration of deep visual feature extractors, such as CNN-based encoders, to further investigate the trade-off between representational capacity and computational complexity.

This observation reveals a key gap in the literature: the lack of a structured and stability-oriented preprocessing strategy that explicitly addresses multimodal variability before model learning. Variations in geometric alignment, illumination conditions, temporal synchronization, and feature scaling introduce instability in multimodal representations, which negatively impacts both performance and generalization [1, 19].

Motivated by this gap, the present study proposes a multimodal audiovisual speech separation framework that prioritizes preprocessing as a fundamental design component. By enforcing geometric, photometric, temporal, and statistical consistency across audio and visual streams, the proposed approach aims to enhance robustness, reduce computational complexity, and support generalization across diverse real-world conditions.

3. Methodology

The research methodology depends on a systematic audiovisual processing system that generates stable, synchronized multimodal data through its efficient computational methods. The research process from start to finish appears in Figure 1, which shows all stages from data preparation to feature extraction, fusion, and evaluation. 1. The framework shows the design decisions that determine how audio and visual data undergo processing before they enter the proposed hybrid model system. The following subsections explain all components of the pipeline through detailed descriptions, starting with dataset selection, followed by visual and audio preprocessing, then feature extraction methods, and finally the speech separation and recognition fusion approach.

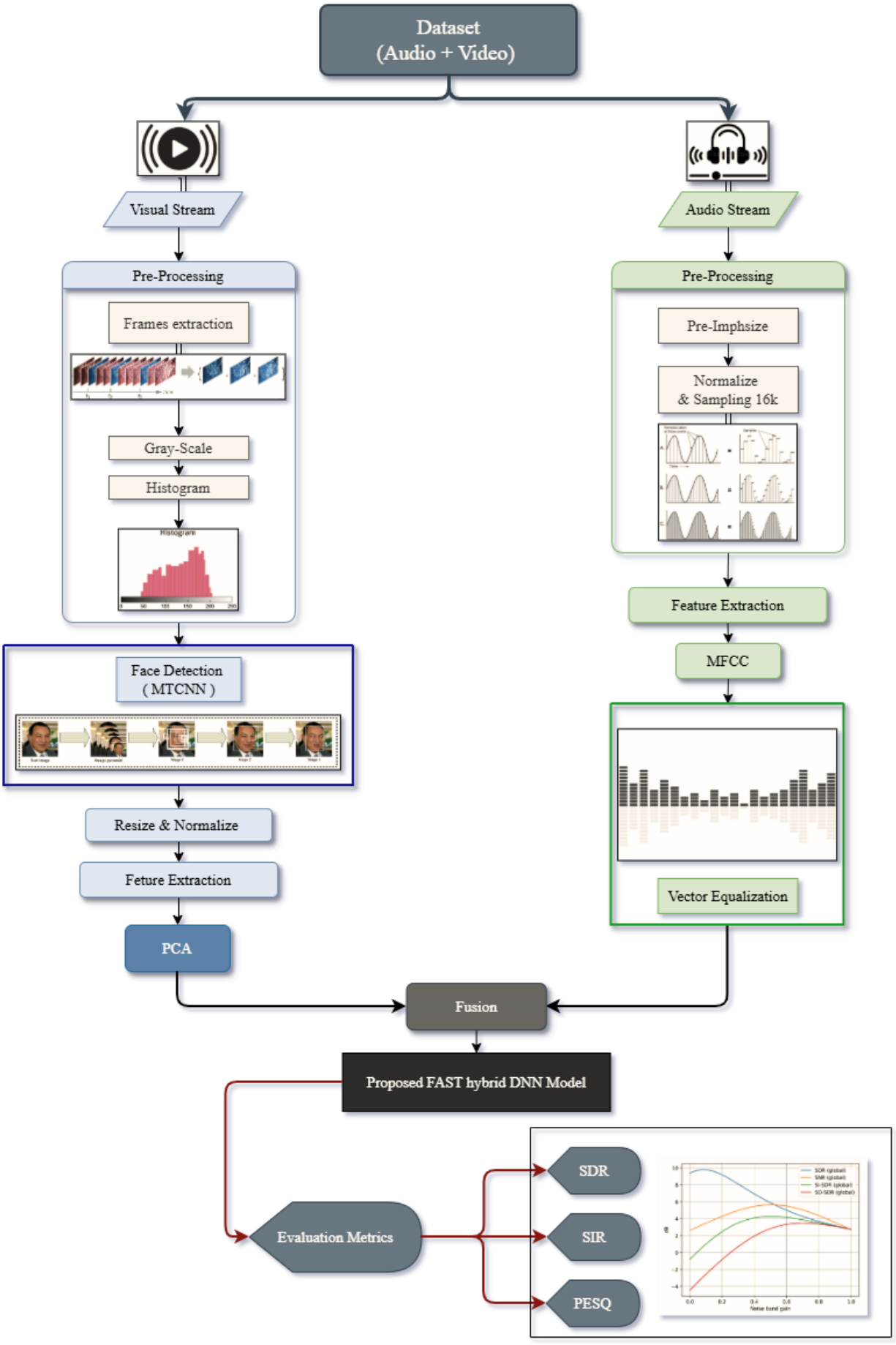


Figure 1. Overview of the proposed audiovisual processing pathway, illustrating the parallel visual and audio preprocessing stages, feature extraction, multimodal fusion, and evaluation metrics

3.1. Dataset

This study adopts a systematic audiovisual processing framework designed to produce stable and synchronized multimodal representations under realistic operating conditions. The overall processing pipeline, illustrated in Figure 1, reflects the key design choices governing the transformation of audio and visual data prior to their integration within the proposed hybrid model [32]. The AVSpeech dataset is selected as the primary data source due to its large-scale, in-the-wild characteristics, which make it particularly suitable for evaluating audiovisual speech separation under realistic conditions. Unlike controlled datasets, AVSpeech contains a wide range of variations in speakers, recording environments, background noise, illumination conditions, and audio-visual synchronization. This diversity provides a challenging testbed that better reflects real-world scenarios, such as online meetings and multimedia communication systems [16]. The use of a large-scale dataset with high variability is aligned with the core objective of this work, which focuses on improving robustness through structured preprocessing and stable feature representation. By operating on data that inherently contains significant variability, the proposed framework is evaluated under conditions that emphasize generalization and practical applicability rather than performance under constrained settings [33]. Furthermore, the emphasis of this study is not on dataset-specific optimization, but on developing a preprocessing-driven framework that enhances representation stability across heterogeneous inputs [11]. This design choice supports the goal of achieving consistent performance under diverse real-world conditions. Future work will extend the evaluation to additional datasets in order to further assess cross-dataset generalization and validate the robustness of the proposed approach across different domains [23].

3.2. Visual Stream

The visual stream provides structured articulatory signals that improve the quality of acoustic information. Real-world video data collection produces video information that contains multiple issues affecting its statistical stability, including frame jitter, illumination changes, compression problems, and detector system instability [26]. The visual pipeline of this study uses geometric and photometric stabilization methods before feature embedding, according to the diagram shown in Figure 1.

3.2.1. Preprocessing

- **Grayscale Conversion**

The AVSpeech dataset together with uncontrolled audiovisual recordings, contains color data that shows high sensitivity to lighting conditions and camera configurations, so researchers need to eliminate non-linguistic variations for articulatory modeling. Research on audiovisual speech processing has demonstrated that luminance-based representations work well for detecting facial speech movements because they provide better stability and require less processing power than other methods [11]. The conversion of RGB frames into one grayscale channel serves two purposes, which include dimension reduction for stability and visual representation stabilization. The standard linear luminance formulation enables grayscale conversion through the following process in Equation 1:

$$I_{gray}=0.30R+0.59G+0.1B \tag{1}$$

where, Igray represents the resulting grayscale intensity, while R, G, B denote the red, green, and blue channel intensities of the original RGB image, respectively.

The operation maintains linear complexity while requiring minimal computational resources, which makes it appropriate for large audiovisual systems [30]. The process of grayscale conversion removes color-dependent noise while maintaining facial structural details, which results in better photometric consistency and produces.

- **Histogram Equalization**

The process of grayscale conversion reduces chromatic variations, but the uneven lighting in uncontrolled video recordings continues to affect the contrast levels of essential facial features. Research on audiovisual speech processing shows that global intensity normalization enables better identification of articulatory regions when lighting conditions change [34]. The solution to this problem involves using histogram equalization on grayscale face images to create balanced intensity distributions, which improve overall image contrast [31]. The cumulative distribution function (CDF) needs the following mathematical operation to function in Equation 2:

$$CDF(X)=\sum_{i=1}^{x} h(i) \tag{2}$$

where, $CDF(X)$ denotes the cumulative distribution function up to gray level x, h(i) represents the histogram frequency (number of pixels) at intensity level i, and i is the intensity index.

The function in Equation 2 accepts x values, which represent gray levels, and hi represents the number of histogram entries at intensity level i. The formula transforms pixel intensity values through this process.

$$T[pixel] = round\left((CDF(X)\text{-}CDF_{min})\times(L\text{-}1)\right) \quad (3)$$

where, T(x) represents the transformed pixel intensity value, CDF_{min} is the minimum non-zero value of the cumulative distribution function, N denotes the total number of pixels in the image, and L is the total number of possible gray levels.

The process in Equation 3 extends the range of low-contrast regions, resulting in improved visual clarity and more distinct lip shape representation, which is critical for accurate articulatory modeling [15]. This pipeline applies histogram equalization to the entire detected facial region rather than individual subregions. Global equalization ensures consistent illumination across the face while preventing local artifacts that could negatively affect PCA-based feature extraction and other dimensionality reduction techniques [1]. Previous studies have shown that this approach improves visual consistency and supports stable feature learning in large-scale audiovisual datasets captured under real-world conditions.

3.2.2. Face Detection

The process of extracting audiovisual features needs face localization to be both precise and to maintain its position over time because any small change in bounding-box position will affect lip region detection and create visual stream alignment problems. Such variability frequently includes partial side-view facial poses and moderate head rotations, which increase the difficulty of maintaining stable visual alignment in unconstrained audiovisual datasets. Research conducted earlier showed that facial localization instability during the early stages produces negative effects that negatively impact both convergence performance and audiovisual learning system generalization [33]. The proposed preprocessing pipeline uses MTCNN as its face detection framework due to its robustness under real-world pose variation and partial profile-face conditions, in addition to its computational efficiency [10]. The MTCNN system operates through a three-stage hierarchical cascade, which improves face candidate detection through successive stages while keeping processing times fast.

- ***Proposal Network (P-Net)***

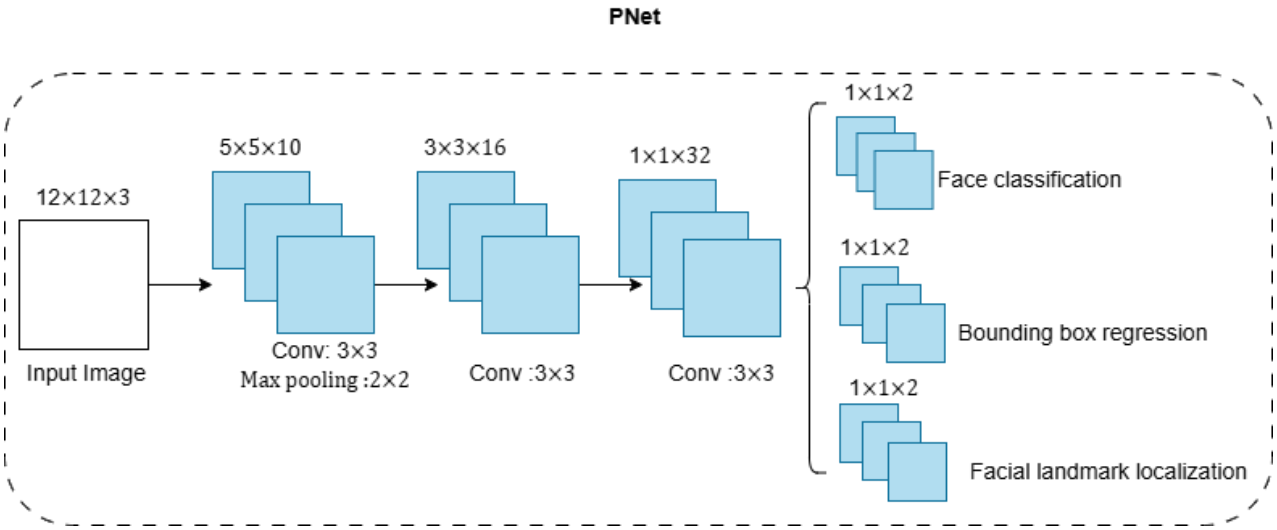
The P-Net functions as the initial detection stage, which examines input video frames at different scales to produce basic face detection results while using minimal processing resources. The system generates candidates at high speed while quickly eliminating non-face areas to handle large video datasets effectively (see Figure 2(a)) [31].

- ***Refine Network (R-Net)***

The R-Net takes candidate regions from the P-Net to perform additional filtering which removes incorrect predictions while improving the accuracy of box position predictions. The system achieves better location precision during this middle phase, which maintains its operational speed by minimizing the position variation between successive video frames (see Figure 2(b)) [34].

- ***Output Network (O-Net)***

The O-Net operates as the final detection stage, which performs exact face recognition and determines the box location. The system produces face regions that remain stable over time while maintaining exact spatial alignment and all necessary articulatory elements required for visual feature extraction during the following stage (see Figure 2(c)) [11].



(a)

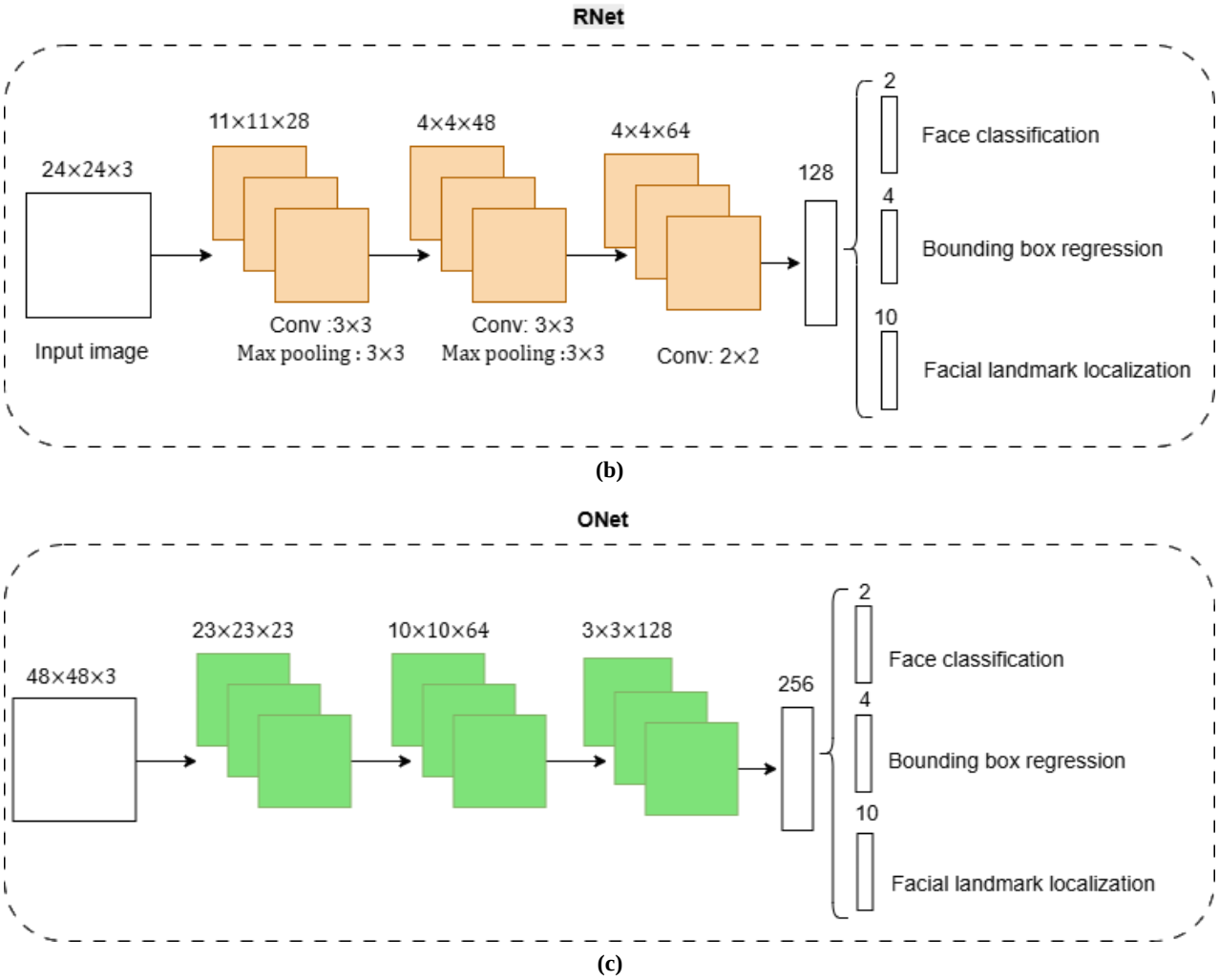


Figure 2. MTCNN Layers (a) Proposal Network (P-Net): coarse multi-scale face proposal generation and early non-face rejection; (b) Refine Network (R-Net): candidate refinement and false-positive suppression for improved spatial stability; (c) Output Network (O-Net): final face classification and high-precision bounding-box regression.

To further enhance robustness under real-world conditions, the proposed framework incorporates simple handling mechanisms for potential face detection failures. In cases where MTCNN fails to detect a valid facial region due to occlusion, motion blur, or illumination variations, the corresponding frame is skipped to avoid introducing noisy or misaligned visual features [11]. In addition, temporal consistency is implicitly preserved through the sequential nature of the input data, allowing the model to rely on neighboring frames for stable feature representation. This temporal smoothing effect reduces the impact of occasional detection failures without requiring explicit post-processing [15].

A lightweight fallback strategy is also considered, where previously detected face regions can be reused during short-term detection gaps. This ensures continuity in the visual stream and prevents abrupt spatial inconsistencies that may affect downstream feature extraction [2].

MTCNN achieves an effective balance between localization accuracy and computational efficiency through its progressive cascade-based refinement process. The process of enforcing face localization at the same position throughout all frames helps to minimize spatial instability while maintaining the lip region structure. Research shows that systems that begin with exact audiovisual signal alignment will produce better results when processing unregulated audiovisual speech signals [34].

3.2.3. Resize and Normalize

The AVSpeech recordings contain faces that show wide variations in size, camera position, and recording environment because they were recorded in real-life situations [16]. The requirement for geometric standardization exists because various spatial representations occur from the same articulatory movements, which makes it difficult to extract visual features and reduces the accuracy of cross-modal alignment [11]. The system uses a standard method to reduce face size to a specific dimension before it starts processing facial features for visual speech analysis. Let I(x,y) denote the grayscale face image cropped from a video frame [31]. The spatial normalization process requires bilinear interpolation to transform I into a fixed dimension of H×W:

$$I_r(u,v)=I\big((u/H)H_0,(v/W)W_0\big) \quad (4)$$

where $I_r(u,v)$ denotes the resized image at spatial location (u, v), $I(\cdot)$ represents the original image, H_0 and W_0 are the original image height and width, respectively, and H and W denote the target height and width. The variables $u \in [1, H]$ and $v \in [1, W]$ represent the spatial coordinates in the resized image.

The original image dimensions in Equation 4 are defined by, H_0 and W_0, while the resized coordinates (u,v) span the target spatial domain. The interpolation process ensures that all input frames are transformed into a fixed resolution, enabling consistent geometric alignment across samples. This spatial normalization preserves the structural characteristics of facial regions while introducing only minimal scale-dependent distortions [34].

Furthermore, spatial normalization facilitates subsequent intensity normalization, which reduces variations in global brightness and ensures uniform feature distributions. This consistency is particularly important for downstream feature extraction methods, such as PCA, where stable input representations directly influence the reliability of learned features [2].

$$I_n=(I_r-\mu)/\sigma \quad (5)$$

where, I_n denotes the normalized image, I_r represents the resized image, μ is the mean intensity value computed over the facial region, and σ is the corresponding standard deviation.

Given the resized image I_r, pixel intensities are normalized using Equation 5 to reduce variations caused by illumination and exposure conditions. This normalization process removes global intensity shifts and scales the data to a consistent range, enabling more stable feature representation. By combining spatial normalization with intensity normalization, the proposed preprocessing pipeline ensures both geometric and photometric consistency across input samples. This low-level conditioning significantly reduces variability unrelated to speech articulation, allowing the model to focus on relevant facial motion patterns. As a result, the framework achieves more stable optimization behavior and improved generalization performance, particularly in unconstrained real-world environments [1].

3.2.4. Visual Feature Extraction

The visual stream processes two PCA-based representations, which are normalized through geometric and photometric transformations. The Principal Component Analysis (PCA) technique produces compact and stable eigenface-based feature representations that capture essential articulatory patterns while suppressing redundant visual variability. The traditional encoding methods create efficient computational models that produce stable statistical results, which work best for large real-world datasets, including AVSpeech [20]. The proposed method maintains structural consistency through its use of predefined transformation mechanisms instead of learning new encoders, which also simplifies the model before it combines different data types.

3.2.4.1. Principal Component Analysis (PCA)

Following geometric and photometric normalization, the visual stream is encoded using a compact representation based on Principal Component Analysis (PCA). The researchers chose PCA as their main design approach because they followed recent audiovisual speech processing research that focused on developing stable preprocessing methods and robust systems instead of using complex learned visual encoders [10].

Unlike classification-oriented techniques such as Linear Discriminant Analysis (LDA), which require explicit class discrimination, PCA is more suitable for the proposed audiovisual speech processing framework because the objective focuses on stable feature representation rather than category-based classification. In this work, PCA is primarily employed as an eigenface-based feature extraction method that captures dominant facial articulation patterns while maintaining computational efficiency and representation stability.

The audiovisual datasets without restrictions show wide variations in their lighting conditions and their subjects' facial appearances, body positions, and recording video quality. The current deep visual encoding systems develop specific biases that make them more responsive to unstable visual input when the image quality remains unpredictable [33]. The projection method of PCA generates a fixed low-complexity transformation that extracts vital articulation patterns from speech data by eliminating unimportant appearance details to achieve improved statistical results among different speakers and recording environments [30].

Let the training set consist of M vectorized in Equation 6:

$$x_1,x_2,\ldots,x_M, \quad x_n \in R^D \quad (6)$$

where, M denotes the total number of training samples, x_n represents the vectorized face image, and D is the dimensionality of the vectorized input. The mean face vector is computed as:

$$\mu=\frac{1}{M}\sum_{n=1}^{M} x_n \quad (7)$$

where, μ is the mean face vector computed over the training set.

$$\tilde{x_n}=x_n-\mu$$
$$C=(1/M)\text{sum}_{n=1}^{M}\tilde{x}_n\tilde{x}_n^T \quad (8)$$

where, μ denotes the mean face vector computed over the training set, and $\tilde{x}_n$ represents the mean-centered sample obtained by subtracting the mean vector from the original sample x_n. The covariance matrix C is estimated using these mean-centered samples.

Eigen-decomposition of C in Equation 8 yields an ordered set of orthogonal eigenvectors:

$$Cu_i=\lambda_i u_i, \quad \lambda_1\geq\lambda_2\geq\cdots \quad (9)$$

where, λ_i represents the i-th eigenvalue, and u_i is the corresponding eigenvector.

In Equation 9 The leading eigenvectors define a low-dimensional subspace that captures the most significant variations in facial appearance. These components encode essential articulatory patterns such as lip motion and jaw dynamics, while reducing the influence of moderate illumination variation and minor detection inconsistencies [1].

Each normalized face image is projected onto the first K principal components as:

$$c_n=U_K^T(x_n-\mu) \quad (10)$$

where, U_k contains the first K principal eigenvectors, and $\mathbf{c}_n \in \mathbb{R}^K$ is the resulting low-dimensional embedding with $K \ll D$.

This projection produces a compact and stable visual representation that preserves discriminative articulatory information while reducing dimensionality. Without introducing additional trainable visual encoding networks [20]. analysis of PCA coefficients reveals their time-dependent patterns, which show how they change over time. The discrete Fourier transform of a single principal component sequence c[n] in Equation 10 over a window of length N produces frequency-domain characteristics.

$$C[k]=\sum_{n=0}^{N-1}c[n]e^{-j2\pi kn/N} \quad (11)$$

where, c[n] represents the temporal sequence of PCA coefficients, N is the length of the analysis window, and k denotes the frequency index.

This transformation extracts frequency-domain characteristics of articulatory motion, capturing rhythmic speech patterns such as periodic lip movements. The use of PCA coefficients in the frequency domain enables robust modeling of temporal dynamics while reducing sensitivity to frame-level noise and visual variability. Operating within the PCA subspace enables joint spatial–temporal modeling without introducing additional trainable parameters. This contributes to improved stability, efficient multimodal fusion, and enhanced generalization performance in real-world audiovisual speech processing systems [35]. This trade-off prioritizes robustness and computational efficiency over highly complex nonlinear visual embedding strategies, making the proposed framework more suitable for real-world audiovisual conditions.

3.3. Audio Stream Preprocessing

The audio stream shows how speech sounds when spoken, but recordings before processing contain background noises, unstable volume levels, and transmission problems, which hide the speech spectrum. As illustrated in Figure 1, the objective of the audio pipeline is to produce a clean, normalized, and statistically consistent signal that remains tightly aligned with the visual stream. The system achieves its objective through waveform-level preprocessing which enables MFCC-based stable spectral feature extraction and vector equalization for cross-recording variability reduction [36].

3.3.1. Pre-Processing

- ***Pre-Emphasis Filtering***

The energy content of speech signals becomes lower when frequencies increase because of how the vocal tract functions and how microphones detect sound [37]. To compensate for this attenuation, a first-order pre-emphasis filter is applied:

$$y[n]=x[n]-\alpha x[n-1], \quad 0.95 \leq \alpha \leq 0.97 \quad (12)$$

where, x[n] denotes the input speech signal at time index n, y[n] represents the pre-emphasized output signal, and α is the pre-emphasis coefficient controlling the contribution of the previous sample x[n−1].

The coefficient α controls the degree of high-frequency emphasis in the speech signal. Values in the range of 0.95 to 0.97 are commonly adopted in contemporary audiovisual and multimodal speech processing systems, as they provide an effective balance between spectral flattening and noise robustness. While higher values of α increase the amplification

of high-frequency components, they may also introduce sensitivity to noise, particularly in large-scale datasets with diverse acoustic conditions. Therefore, selecting an appropriate coefficient is essential to ensure stable feature extraction and reliable downstream processing [21].

- ***Amplitude Normalization and Resampling***

The AVSpeech system records audio signals with wide variations in loudness levels and sampling rates because it operates without any specific recording restrictions [33]. The process of amplitude-induced bias suppression in spectral features needs waveform x[n] normalization to reach a particular dynamic range according to this mathematical equation (see Equation 13).

$$\tilde{x}[n]=\frac{x[n]}{\max(|x[n]|)} \tag{13}$$

where, x_n denotes the normalized speech signal, x_n (before normalization) represents the original signal amplitude at sample index n, and $\max_k|x_k|$ is the maximum absolute amplitude value across the entire signal.

The normalization process ensures that all following spectral representations show phonetic information instead of recording gain levels [1]. All signals are then resampled to a common sampling frequency

$$x16k(n) = R(xn, fs \rightarrow 16kHz) \tag{14}$$

where, $x16k(n)$) denotes the resampled speech signal at 16 kHz, xn is the normalized input signal, fs is the original sampling frequency, and $R(\cdot)$ represents the resampling operation.

The normalization process ensures that the resulting spectral representations emphasize phonetic information rather than variations caused by recording gain. By scaling the signal amplitude to a fixed range, the preprocessing stage reduces variability across samples and improves the consistency of subsequent feature extraction.

Following amplitude normalization, all audio signals are resampled to a common sampling frequency of 16 kHz, as defined in Equation 14. This resampling step standardizes the temporal resolution of the input signals while preserving the underlying speech dynamics. As a result, both audio and visual streams can be more effectively aligned in time.

The combination of amplitude normalization and resampling establishes a standardized acoustic representation, which enhances feature consistency and improves cross-modal synchronization. This preprocessing step is particularly important for multimodal systems, where stable temporal alignment directly impacts the effectiveness of audio-visual fusion and overall model performance [38].

3.3.2. Mel-Frequency Cepstral Coefficients (MFCC)

MFCCs use a restricted set of data to create speech spectral envelope representations which makes them the preferred front-end selection for modern audiovisual speech systems because they provide both perceptual benefits and noise immunity at moderate noise levels. In Equation 15, the preprocessing operation results in waveform normalization, which allows the creation of short overlapping frames that have their spectral energies transformed into Mel scale frequencies to match human hearing sensitivity [8].

The Mel frequency corresponding to a linear frequency f is defined as:

$$\text{mel(f)}=2595\log_{10}\left(1+\frac{f}{700}\right) \tag{15}$$

where, $m(f)$ denotes the Mel-frequency corresponding to the linear frequency f (in Hz).

This transformation maps linear frequency to the perceptual Mel scale, which approximates the nonlinear frequency response of the human auditory system. After logarithmic compression of the Mel filterbank energies, the cepstral coefficients are obtained using the discrete cosine transform [14]:

$$c_k=\sum_{n=1}^{B}\log(E_n)\cos\left(\frac{\pi k}{B}\left(n-\frac{1}{2}\right)\right) \tag{16}$$

where, c_k represents the k-th cepstral coefficient, E_n denotes the energy of the n-th Mel filterbank, and B is the total number of Mel filters.

The resulting MFCCs capture the coarse spectral envelope of the speech signal while suppressing fine harmonic details. This representation aligns well with visual articulatory cues, such as lip movements and jaw dynamics, making it suitable for audiovisual speech processing tasks. To further improve robustness, cepstral mean and variance normalization (CMVN) is applied to the MFCC vectors [4]. CMVN standardizes the statistical distribution of cepstral coefficients by removing long-term channel effects and reducing inter-recording variability. This normalization enhances the stability of feature representations across different speakers and recording conditions.

As a result, the MFCC features exhibit consistent patterns that improve model generalization and robustness in large-scale datasets such as AVSpeech. Previous studies have demonstrated that CMVN significantly enhances system

performance by mitigating variability across diverse acoustic environments and enabling more reliable multimodal learning [39]. An example of the resulting MFCC representation, along with the corresponding waveform and Mel-spectrogram, is shown in Figure 3.

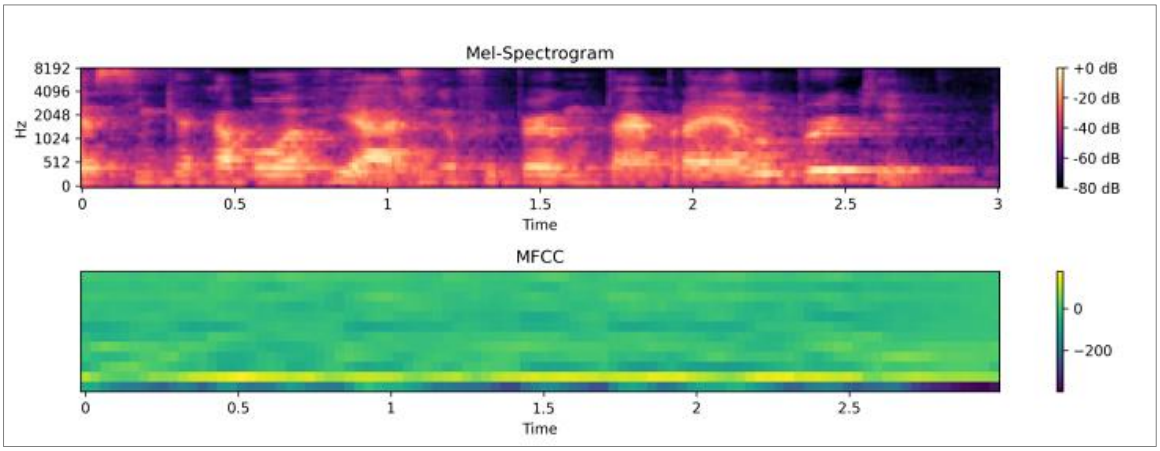


Figure 3. Mel-spectrogram and MFCC representations obtained from the proposed MFCC extraction pipeline

3.3.4. Vector Equalization

The raw MFCC features show wide variations between different recordings because of how well the recording channels function, how loudly the speakers speak, and what background noises exist. The different levels of energy in the data can create bias during multimodal learning because the fusion process becomes controlled by the most powerful frames and frequency bands. The solution to this problem in Equation 16 requires vector-level equalization, which preserves statistical audio feature consistency while maintaining its time-based connection to visual data [4]. For MFCC features, cepstral mean and variance normalization (CMVN) is applied on a t basis [2]. Let ct∈RD denote the MFCC vector at time frame. Normalization is defined as:

$$\hat{c}_t=(c_t-\mu)/\sigma \quad (17)$$

where, $\hat{c}_t$ denotes the normalized MFCC vector, c_t is the original MFCC vector at time frame t, and μ and σ represent the mean and standard deviation computed over all frames within the same utterance.

The parameters μ and σ are estimated using all frames within a given utterance, ensuring that normalization is consistent at the sequence level. This operation effectively reduces channel variability and temporal drift in spectral characteristics, which are common in real-world audiovisual recordings. By standardizing the MFCC feature space, CMVN enhances the robustness of the extracted features and improves consistency across different speakers and recording conditions. This is particularly important for large-scale datasets such as AVSpeech, where variations in recording environments, devices, and speaking styles can introduce significant variability. Consequently, CMVN plays a critical role in enabling stable feature learning and improving generalization performance in multimodal speech processing systems [32].

3.4. Proposed HAVS-Net Model

The HAVS-Net system uses a hybrid deep learning framework that takes advantage of the processed audio data resulting from the Section 3.3 preprocessing system. The network employs convolutional and recurrent mechanisms to identify particular spectral patterns and prolonged temporal relationships that appear in speech data. The design choice represents the non-stationary characteristics of speech because speech contains important information that spreads across different time periods.

3.4.1. Architectural Design Rationale

The process of speech separation and recognition needs to combine models that analyze both brief spectral changes and extended time-based patterns. Research studies about audiovisual speech processing from recent times show that successful systems require both local pattern detection and sequential relationship analysis to achieve reliable results in operational environments [12]. The extraction of localized temporal and spectral patterns functions well in convolutional layers, but these layers fail to detect long-term temporal relationships [13]. The recurrent layers of the model effectively capture long-term temporal patterns, but they lose their ability to detect detailed local patterns when used independently. The system HAVS-Net solves this problem through its design, which combines convolutional and recurrent layers to perform different temporal analyses at various scales within a single framework. The design of HAVS-Net starts with a recurrent layer that creates a worldwide temporal framework for processing the entire input sequence before convolutional processing takes place [9]. The initial temporal conditioning process enables the following convolutional filters to process information that has been made aware of context, which results in better local feature extraction with enhanced temporal consistency and coherence [6].

3.4.2. Layer Composition and Hierarchical Temporal Modeling

The network uses a hierarchical temporal structure that alternates between LSTM and Conv1D and GRU layers [6]. The first stages of the model use powerful LSTM units that detect extended patterns in the input data because this ability enables both speaker identity preservation and speech segment overlap recognition [5]. Convolutional layers are introduced after temporal conditioning to extract discriminative local temporal features within this broader context. The network depth requires a step-by-step reduction of convolutional filters, which creates a compression stage that extracts vital temporal data while eliminating unimportant information [40]. The additional recurrent modeling in these layers enhances the existing representations through multiple recurrent modeling approaches [5]. The model includes GRU layers at intermediate stages to achieve a balance between information representation and processing speed because these layers perform well in time series analysis with their reduced parameter count compared to LSTM units. The compression-refinement approach in HAVS-Net allows the model to detect both large-scale time patterns and detailed spectral details while preventing overcomplication of the model structure [20].

3.4.3. Activation Functions and Pooling Strategy

The network uses Leaky ReLU activations to maintain gradient flow and protect negative activation information, which speech-derived features tend to lose. The selection enhances training stability when working with deep temporal stacks, and it helps prevent vanishing gradient effects (see Figure 4).

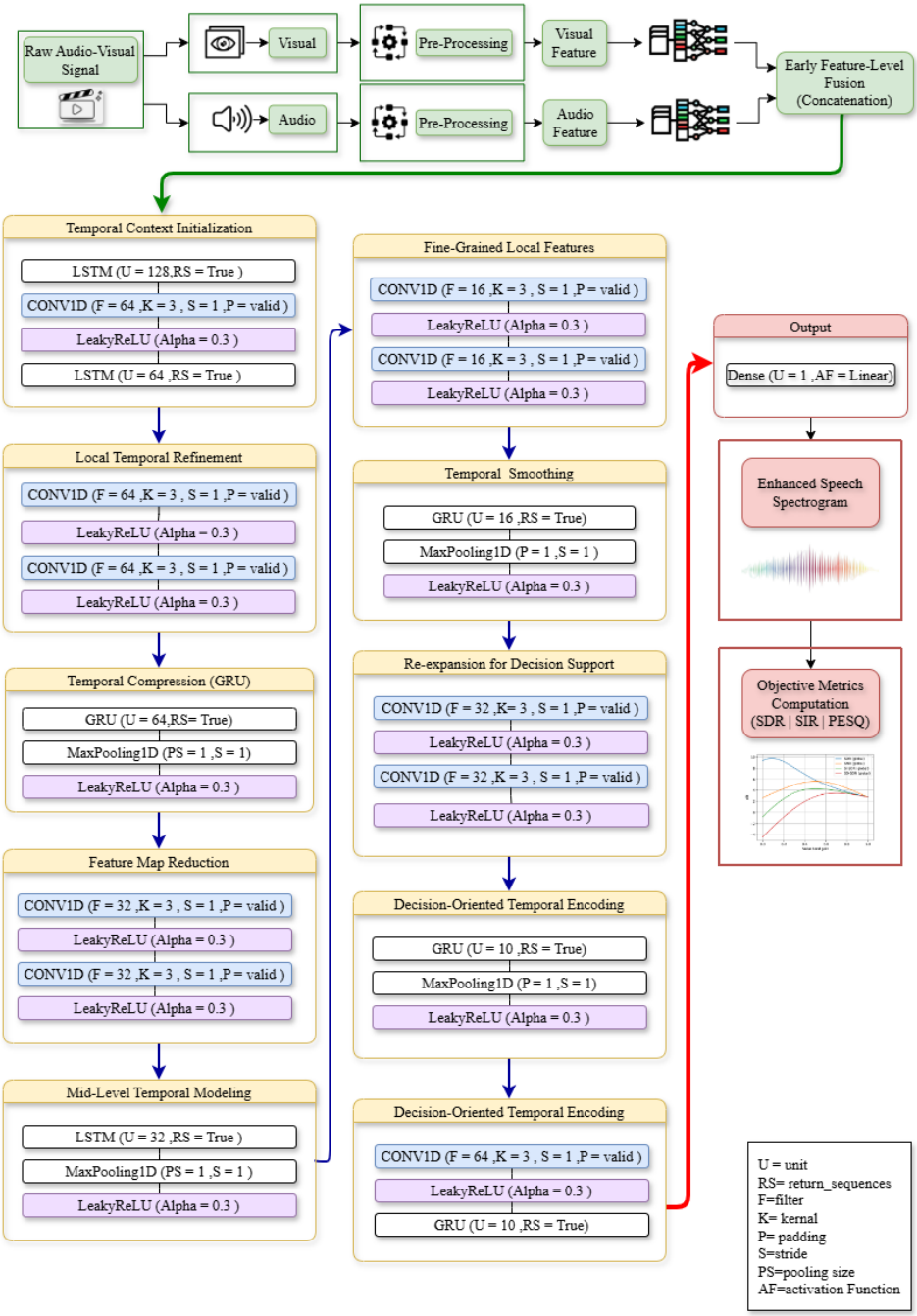


Figure 4. Overall architecture of the proposed audio–visual speech separation framework

The research incorporates max-pooling layers, which function as structural regularizers instead of using them for temporal downsampling. The model maintains its time-based precision because it controls pooling size to prevent the loss of frame-by-frame alignment, which is essential for audiovisual synchronization. The process of pooling enables feature hierarchy reorganization while maintaining visual stream temporal relationships (see Table 2).

Table 2. Layer-wise functional specification of the proposed HAVS-Net architecture

Block	Functional Role	Core Operation	Output Dimension	Trainable Parameters
Input Temporal Modeling Block	Global temporal context initialization	LSTM	(30, 128)	66,560
Temporal Feature Extraction Block I	Local temporal pattern extraction	Conv1D + LeakyReLU	(28, 64)	24,640
Temporal Context Refinement Block	Long-range dependency refinement	LSTM	(28, 64)	33,024
Local–Temporal Refinement Block I	Joint local and gated temporal modeling	Conv1D + GRU	(24, 64)	49,472
Intermediate Feature Extraction Block	Temporal abstraction and dimensionality reduction	Conv1D	(20, 32)	9,280
Hierarchical Temporal Modeling Block	Mid-level temporal structure modeling	LSTM	(20, 32)	8,320
Fine-Grained Temporal Refinement Block	Compact temporal feature refinement	Conv1D + GRU	(16, 16)	3,920
High-Level Feature Extraction Block	High-level temporal representation learning	Conv1D	(12, 32)	4,672
Compact Temporal Representation Block	Temporal compression and stabilization	GRU	(12, 10)	1,290
Final Temporal Refinement Block	Final gated temporal adjustment	Conv1D + GRU	(10, 64)	26,752
Output Mapping Layer	Frame-level speaker isolation mapping	Fully Connected (Linear)	(10, 1)	65
	Total number of parameters			**227,995**

3.4.4. Model Complexity and Practical Considerations

The HAVS-Net architecture contains 47 layers that use 227k trainable parameters to achieve strong temporal modeling abilities through its lightweight design that surpasses transformer-based audiovisual models. The model operates at high speed for large datasets such as AVSpeech because it uses 1D convolutions and limited pooling and gated recurrent units. The system operates on audio features that have undergone both preprocessing and equalization according to its designed architecture. The HAVS-Net system dedicates its processing power to discovering essential time-based and frequency-based connections because it operates under the assumption that all input data remain constant. The close relationship between data preparation and model construction methods produces better results that maintain their performance in different recording environments.

3.4.4.1. Model Advantages

- **Lightweight architecture**

The HAVS-Net model operates with fewer than 230k trainable parameters, which results in lower memory usage and faster computation than transformer-based and diffusion-based audiovisual models. The design features a straightforward architectural structure that lacks intricate elements. The model lacks attention mechanisms, multi-branch fusion blocks, and high-dimensional visual encoders because it employs sequential Conv1D and recurrent layers to perform efficient temporal modeling.

- ***Hierarchical Temporal Modeling***

The LSTM and GRU units operate independently to develop models that monitor both extended speaker voice patterns and brief articulatory changes to achieve successful audiovisual speech separation.

- ***Balanced Recurrent Design***

The model uses LSTM layers to analyze data that require long-term temporal memory storage, but it uses GRU layers for parameter-efficient temporal refinement, which enhances training stability and decreases the risk.

- ***Preprocessing-Aware Modeling***

The model learns speech-related temporal and spectral patterns because it assumes that input features follow a statistical distribution that ignores amplitude differences and channel effects.

- ***Improved Generalization***

The model HAVS-Net achieves generalization across different recording conditions through its dependence on stable normalized representations that do not require extensive data or model growth.

- ***Practical Deployability***

The model features a compact sequential design that enables efficient processing of large data sets during real-world operations that require both efficiency, reproducibility, and robustness.

4. Experimental Setup and Implementation Details

All experiments were conducted under a controlled and fully reproducible environment to evaluate the proposed audiovisual speech separation framework in terms of both separation performance and computational efficiency. The experimental setup used equipment that replicated real field environments instead of using laboratory-specific instruments. The system performed training and evaluation operations on a regular consumer laptop that ran CPU-based operations instead of using GPU acceleration. The system contained an Intel® Core™ i7-10870H CPU @ 2.20 GHz processor together with 16 GB of RAM. The proposed architecture shows its lightweight design because it does not contain dedicated GPU hardware, which enables it to perform audiovisual separation tasks at a competitive level using minimal computational resources. The total training time for the complete model was approximately 48,225 s (13.4 h) for 100 training epochs. The system operated through Python, which served as its programming language for execution. The backend system received explicit verification during runtime initialization to verify its numerical stability and preserve stable layer operations throughout training and evaluation procedures. The HAVS-Net model was trained using a fixed and consistent configuration across all experiments (see Table 3).

Table 3. Training configuration and hyperparameter settings

Parameter	Value
Number of Epochs	100
Optimizer	Adam
Learning Rate	10-3
Loss Function	Mean Squared Error (MSE)
Activation Function	Leaky ReLU (all convolutional and recurrent layers)
Feature Normalization	Min–Max normalization (applied prior to training)
Patch Size	64

The experimental setup unites all components to achieve stable convergence, controlled optimization behavior, and equal testing conditions for all results presented. The proposed audiovisual speech separation framework demonstrates generalization and reproducibility through its use of normalized input features and its compact training configuration, which operates under actual system conditions.

4.1. Evaluation Protocol

The system performance evaluation used established objective metrics that followed standardized methods to achieve reproducible results for comparison with previous audiovisual speech separation research [41]:

- ***The Signal-to-Distortion Ratio (SDR)***

Measures the total distortion that exists between the extracted speech signal and its original clean version because it combines all types of remaining interference, signal degradation, and artifacts. Classical SDR, as defined in the BSS_eval framework, was employed and is given by:

$$SDR=10\log_10\left(||s(n)||^2\right)-10\log_10\left(||s(n)-\hat{s}(n)||^2\right) \quad (11)$$

In Equation 11 where s(n) denotes the clean reference speech signal, $\hat{s}(n)$ represents the estimated speech signal produced by the proposed model, and n is the discrete-time sample index. The operator |.| represents the Euclidean norm, which equals signal energy. The reconstruction quality improves when SDR values increase because the overall distortion level decreases [42].

- ***Signal-to-Interference Ratio (SIR)***

Measures how well the system removes background speech elements from the main audio signal [41]. Following the BSS_eval formulation, SIR can be expressed in conceptual form as:

$$SIR=10\log_10\left(||s(n)||^2\right)-10\log_10\left(||i(n)||^2\right) \quad (12)$$

The model in Equation 12 processes input speech signals n through two separate paths, which produce output speech signals s(n) and i(n). The strength of interference suppression and source separation ability increases with higher SIR values.

- ***Perceptual Evaluation of Speech Quality (PESQ)***

The wideband PESQ metric operates at 16 kHz sampling rate to evaluate speech quality through ITU-T Recommendation P.862.2. The PESQ system uses an objective method to predict human listening quality through its simulation of how listeners perceive signal degradation [4]. The PESQ score exists in a basic conceptual structure that can be expressed as:

$$PESQ \approx a_0+a_1D_{ind}+a_2A_{ind} \quad (13)$$

In Equation 13, where D_{ind} denotes the disturbance indicator; A_{ind} represents the asymmetry indicator; a_0, a_1, and, a_2 are fixed parameters defined by the ITU-T standard and are internally computed within the PESQ algorithm. The evaluation of speech quality through PESQ produces results that show better perceived speech quality when the scores reach higher values. The authors used validated, publicly available implementations from previous audiovisual speech separation research to calculate all their metrics. The system calculated metric values for each test utterance before it combined them to generate the results that appear in the final report [43].

5. Results and Discussion

This section presents the quantitative evaluation and convergence behavior of the proposed audiovisual speech separation framework under the experimental protocol described in Section 3.

5.1. Performance Evolution Across Training Epochs

The evaluation was conducted using 435 utterances from the test set, where each separated signal was compared with its corresponding clean reference. Valid metric values were obtained for 433 utterances due to metric-specific constraints.

Figure 5. Performance evolution of SDR, SIR, and PESQ across training epochs under normalized feature configuration. This progressive improvement reflects the model's ability to learn increasingly accurate representations of the target speech signal. In particular, the early-stage increase in SIR suggests that the model first learns to suppress interfering sources by leveraging multimodal cues. This is followed by improvements in SDR, indicating enhanced reconstruction fidelity of the target speech signal. The gradual increase in PESQ further indicates that perceptual quality improves at later stages of training, as the model refines fine-grained spectral and temporal details.

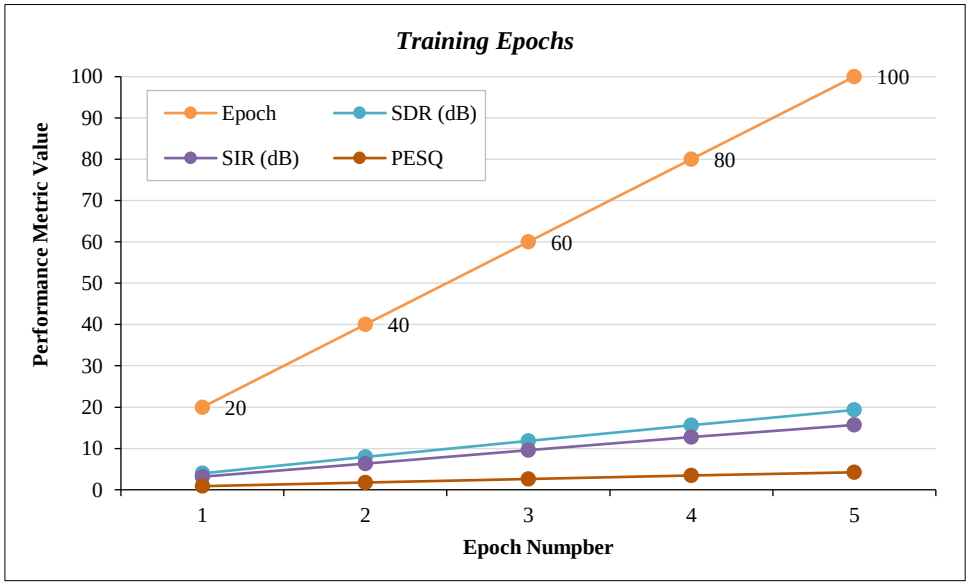


Figure 5. Performance evolution of SDR, SIR, and PESQ across training epochs under normalized feature configuration

The observed convergence behavior also highlights the contribution of the proposed stability-driven preprocessing pipeline, which provides consistent multimodal feature representations and improves optimization stability during training. From a practical perspective, these results indicate that the proposed framework can maintain reliable speech separation performance under realistic and acoustically challenging audiovisual conditions.

5.2. Quantitative Performance Summary

The numerical results supporting these observations are presented in Table 4.

Table 4. Quantitative performance evolution across training epochs under normalized feature configuration

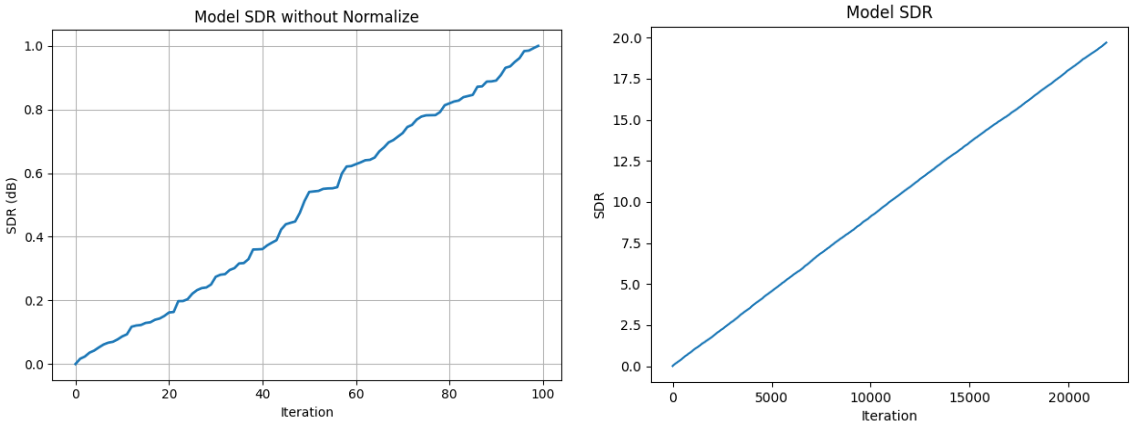
Epoch	SDR (dB)	SIR (dB)	PESQ
20	3.97	3.17	0.87
40	7.94	6.36	1.74
60	11.86	9.55	2.58
80	15.64	12.73	3.43
100	19.32	15.69	4.23

The results demonstrate a substantial improvement across all evaluation metrics as training progresses. The SDR improvement from 3.97 dB to 19.32 dB indicates a significant reduction in distortion and improved signal reconstruction. Similarly, the increase in SIR from 3.17 dB to 15.69 dB reflects the model's enhanced capability to suppress interfering speech sources. The PESQ improvement from 0.87 to 4.23 highlights a considerable enhancement in perceptual speech quality, indicating that the reconstructed signals are both intelligible and acoustically natural.

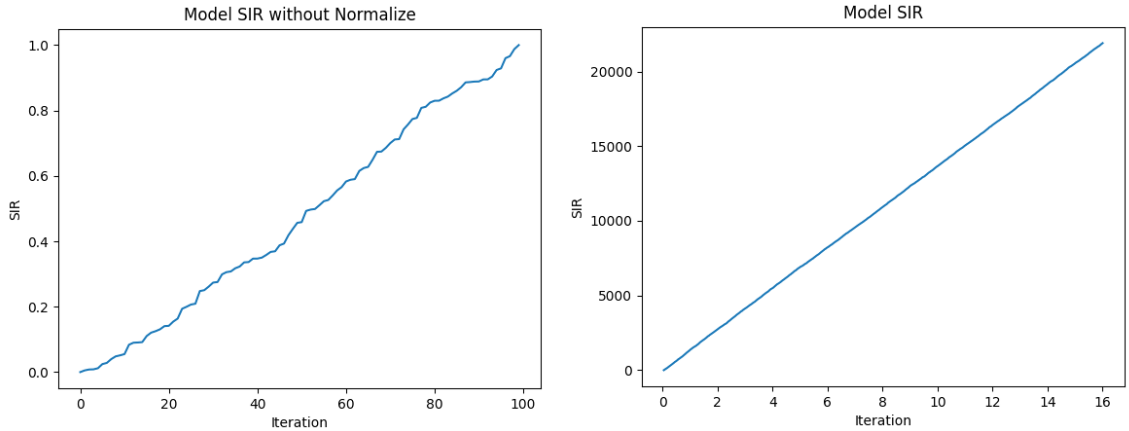
5.3. Effect of Feature Normalization

The impact of feature normalization is illustrated in Figure 6, which compares the convergence behavior of SDR, SIR, and PESQ under normalized and non-normalized feature configurations. Under non-normalized conditions, although all metrics exhibit a generally increasing trend, convergence remains significantly slower, and the achieved performance values are substantially lower throughout training. This behavior indicates that unstable feature distributions negatively affect optimization stability and limit the model's ability to learn reliable multimodal representations.

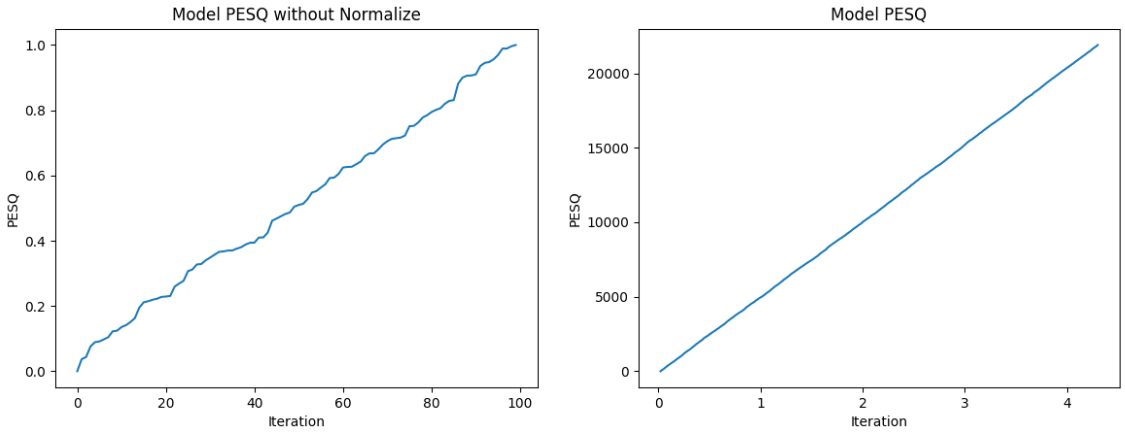
In contrast, normalized features lead to faster convergence and higher saturation levels across all evaluation metrics. This improvement can be attributed to enhanced feature conditioning, which stabilizes the statistical distribution of both audio and visual representations prior to multimodal fusion. As a result, the model can more effectively learn cross-modal relationships and achieve improved optimization performance. The faster improvement of SDR and SIR compared to PESQ suggests that interference suppression and distortion reduction are learned during earlier training stages, whereas perceptual quality refinement requires more accurate temporal–spectral reconstruction and therefore emerges later in the learning process.



a) SDR



b) SIR



c) PESQ

Figure 6. Iteration-wise convergence of (a) SDR, (b) SIR, and (c) PESQ under normalized and non-normalized feature configurations

These results further demonstrate the importance of stable preprocessing and feature normalization in supporting robust convergence behavior and improving speech separation performance under realistic audiovisual conditions.

5.4. Discussion

The experimental results confirm that the proposed audiovisual speech separation framework achieves robust and stable performance across multiple evaluation metrics [1]. The improvement in SDR demonstrates that the model effectively reconstructs the target speech signal with reduced distortion. This can be directly linked to the structured preprocessing pipeline, which ensures consistent and well-conditioned input representations, enabling the model to learn reliable mappings between noisy and clean speech signals [6]. The observed increase in SIR highlights the effectiveness of multimodal fusion in suppressing interfering sources. By incorporating visual cues such as lip movements and facial dynamics, the model gains additional discriminative information that reduces ambiguity in multi-speaker scenarios. This is particularly important in real-world environments where acoustic cues alone are insufficient for accurate separation. The high PESQ scores further indicate that the reconstructed speech signals are perceptually natural and intelligible [21]. This improvement can be attributed to the combination of MFCC-based spectral representations and feature normalization, which preserve both global spectral structure and fine-grained temporal details. As a result, the output signals exhibit reduced artifacts and improved clarity from a perceptual standpoint. From a system-level perspective, these results highlight the importance of representation stability over architectural complexity. Unlike transformer-based and diffusion-based models that rely on large-scale parameterization, the proposed framework achieves competitive performance through structured preprocessing and efficient hybrid temporal modeling. To further evaluate the efficiency of the proposed framework, Table 5 presents a detailed breakdown of the computational and memory complexity of the HAVS-Net components.

Table 5. Computational and Memory Complexity Analysis of HAVS-Net Components

Component	Type	Complexity	Impact
Conv1D Layers	Computational	$O(N \cdot K \cdot C)$	Low (local temporal feature extraction)
LSTM Layers	Computational	$O(T \cdot H^2)$	Medium (long-range temporal modeling)
GRU Layers	Computational	$O(T \cdot H^2)$	Medium (parameter-efficient sequence modeling)
Fully Connected Layers	Computational	$O(N)$	Low
Audio–Visual Feature Fusion	Computational	$O(N)$	Low
Parameter Storage	Memory	227,995 parameters	Low
Activation Storage	Memory	$O(T \cdot H)$	Medium
Training Time	Time	Depends on epochs and batch size	Medium
Inference Time	Time	CPU-based, near real-time capable	Low

The analysis shows that the proposed model maintains a low overall computational footprint, with most operations exhibiting linear or near-linear complexity. The relatively small parameter size further confirms the lightweight design of the framework, enabling efficient deployment in resource-constrained environments [21].

Importantly, it is necessary to distinguish between preprocessing complexity and model complexity. The preprocessing pipeline is executed as an offline stage and does not contribute to the runtime computational cost during inference. In contrast, the HAVS-Net architecture operates as a lightweight model with a low parameter count and efficient temporal operations. Although the preprocessing stage involves multiple steps, these operations are computationally inexpensive and primarily consist of linear transformations. Therefore, the overall system achieves a balance between robustness and efficiency, without introducing significant computational overhead during deployment. A comparison with representative audiovisual speech separation methods is provided in Table 6.

Table 6. Comparison with representative audiovisual speech separation methods

Method	Architecture	Dataset	SDR (dB)	SIR (dB)	PESQ	Complexity
Wu et al. [40]	CNN–RNN	AVSpeech	15.1	12.3	3.2	Medium
Rahimi et al. [43]	Transformer	LRS3	18.5	—	4.1	High
Kalkhorani et al. [13]	Transformer	AVSpeech	17.9	14.8	3.9	High
Diffusion-based AV [44]	Diffusion	AVSpeech	19	—	4.2	Very High
HAVS-Net	CNN–LSTM–GRU	AVSpeech	19.32	15.69	4.23	Low

The comparison indicates that HAVS-Net achieves competitive performance across all evaluation metrics while maintaining significantly lower computational complexity. Unlike transformer-based and diffusion-based models, which rely on large-scale architectures, the proposed framework achieves similar or better performance through structured preprocessing and efficient temporal modeling [44]. This demonstrates that performance improvements can be achieved and competitive performance through representation stability rather than increased model complexity. These results further confirm that performance gains are primarily driven by representation stability and preprocessing consistency rather than increased architectural complexity. This highlights the effectiveness of the proposed design in achieving robust and generalizable speech separation under realistic conditions [2].

6. Conclusion

This study presented a multimodal audiovisual speech separation framework that emphasizes structured preprocessing and feature normalization as fundamental components for robust multimodal learning. Unlike conventional approaches that prioritize increasing model complexity, the proposed method adopts a stability-driven design philosophy, where preprocessing plays a central role in ensuring consistent and well-conditioned feature representations across both audio and visual streams. By enforcing geometric, photometric, temporal, and statistical consistency, the framework produces aligned multimodal features that facilitate efficient temporal modeling using a lightweight hybrid architecture. The experimental results demonstrate that the proposed approach achieves stable convergence behavior and competitive separation performance under realistic conditions. Specifically, the model attained an average SDR of 19.32 dB, SIR of 15.69 dB, and PESQ of 4.23, indicating effective interference suppression, accurate signal reconstruction, and high perceptual speech quality.

These results confirm that performance improvements can be achieved through representation stability rather than increased architectural complexity. Furthermore, the analysis highlights that feature normalization plays a critical role in accelerating convergence and improving overall model reliability. From a system-level perspective, the proposed framework provides an effective balance between performance and computational efficiency, making it suitable for real-world applications such as online meetings, multimedia communication systems, and noisy audiovisual environments. The lightweight design enables practical deployment in resource-constrained settings without sacrificing separation quality. Future work will focus on extending the framework to cross-dataset and multilingual scenarios in order to further evaluate its generalization capability. In addition, integrating advanced visual feature representations may provide further improvements in capturing complex articulatory patterns under challenging real-world conditions.

7. Declarations

7.1. Author Contributions

Conceptualization, S.M.Sh. and B.M.A.; methodology, S.M.Sh.; software, S.M.Sh.; validation, S.M.Sh. and B.M.A.; formal analysis, S.M.Sh.; investigation, S.M.Sh.; resources, B.M.A.; data curation, S.M.Sh.; writing—original draft preparation, S.M.Sh.; writing—review and editing, B.M.A.; visualization, S.M.Sh.; supervision, B.M.A.; project administration, B.M.A. All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

7.4. Institutional Review Board Statement

Not applicable.

7.5. Informed Consent Statement

Not applicable.

7.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8. References

[1] Michelsanti, D., Tan, Z. H., Zhang, S. X., Xu, Y., Yu, M., Yu, D., & Jensen, J. (2021). An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation. IEEE/ACM Transactions on Audio Speech and Language Processing, 29, 1368–1396. doi:10.1109/TASLP.2021.3066303.

[2] Du, J., Jin, Z., Yang, P., Liu, J., Li, Z., Liu, X., & Li, M. (2025). Audio-Visual Speech Enhancement in Complex Scenarios with Separation and Dereverberation Joint Modeling. arXiv Preprint, arXiv:2510.26825. doi:10.48550/arXiv.2510.26825.

[3] Rahimi, A. (2025). Restoring Degraded Multi-Speaker Speech through Separation and Enhancement. 4th Cogmhear Audio-Visual Speech Enhancement Challenge (AVSEC), ISCA, 1–5. doi:10.21437/avsec.2025-1.

[4] Wang, X., Guo, B., Huo, X., Zhang, Y., & Tao, J. (2024). Speech Enhancement Techniques Based on Microphone Arrays and Deep Learning. 2024 IEEE 8th International Conference on Vision, Image and Signal Processing, ICVISP 2024, 1–4. doi:10.1109/ICVISP64524.2024.10959537.

[5] Luo, Y., Chen, Z., & Yoshioka, T. (2020). Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2020-May, 46–50. doi:10.1109/ICASSP40776.2020.9054266.

[6] Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., & Xie, L. (2020). DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2020-October, 2472–2476. doi:10.21437/Interspeech.2020-2537.

[7] Gao, R., & Grauman, K. (2021). VisualVoice: Audio-visual speech separation with cross-modal consistency. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 15490–15500. doi:10.1109/CVPR46437.2021.01524.

[8] Lopez-Olvera, J. A., Perez-Meana, H. M., Garcia-Rios, E., & Escamilla-Hernandez, E. (2026). Leveraging MFCC and Mel-Spectrogram Representations for Deep Learning-Based Speech Recognition. Engineering Proceedings, 123(1), 22. doi:10.3390/engproc2026123022.

[9] Sach, M., Franzen, J., Defraene, B., Fluyt, K., Strake, M., Tirry, W., & Fingscheidt, T. (2023). EffCRN: An Efficient Convolutional Recurrent Network for High-Performance Speech Enhancement. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2023-August, 2508–2512. doi:10.21437/Interspeech.2023-799.

[10] Naser, O. A., Mumtazah, S., Samsudin, K., Hanafi, M., Shafie, S. M. B., & Zamri, N. Z. (2025). Comparative Analysis of MTCNN and Haar Cascades for Face Detection in Images with Variation in Yaw Poses and Facial Occlusions. Journal of Communications Software and Systems, 21(1), 109–119. doi:10.24138/jcomss-2024-0084.

[11] Vilaça, L., Yu, Y., & Viana, P. (2025). A Survey of Recent Advances and Challenges in Deep Audio-Visual Correlation Learning. ACM Computing Surveys, 57(12), 1–299 46. doi:10.1145/3696445.

[12] Radfar, M., Barnwal, R., Swaminathan, R. V., Chang, F. J., Strimel, G. P., Susanj, N., & Mouchtaris, A. (2022). ConvRNN-T: Convolutional Augmented Recurrent Neural Network Transducers for Streaming Speech Recognition. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2022-September, 4431–4435. doi:10.21437/Interspeech.2022-10844.

[13] Kalkhorani, V. A., Kumar, A., Tan, K., Xu, B., & Wang, D. L. (2023). Time-domain Transformer-based Audiovisual Speaker Separation. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2023-August, 3472–3476. doi:10.21437/Interspeech.2023-2098.

[14] Baevski, A., Babu, A., Hsu, W. N., & Auli, M. (2023). Efficient Self-supervised Learning with Contextualized Target Representations for Vision, Speech and Language. Proceedings of Machine Learning Research, 202, 1416–1429.

[15] Hu, Y., Li, R., Chen, C., Zou, H., Zhu, Q., & Chng, E. S. (2023). Cross-Modal Global Interaction and Local Alignment for Audio-Visual Speech Recognition. IJCAI International Joint Conference on Artificial Intelligence, 2023-August, 5076–5084. doi:10.24963/ijcai.2023/564.

[16] Li, C., & Qian, Y. (2020). Listen, Watch and Understand at the Cocktail Party: Audio-Visual-Contextual Speech Separation. Interspeech, 1426-1430.

[17] Jin, Z., Yang, Y., Shi, M., Kang, W., Yang, X., Yao, Z., Kuang, F., Guo, L., Meng, L., Lin, L., Xu, Y., Zhang, S. X., & Povey, D. (2024). LibriheavyMix: A 20,000-Hour Dataset for Single-Channel Reverberant Multi-Talker Speech Separation, ASR and Speaker Diarization. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 702–706. doi:10.21437/Interspeech.2024-90.

[18] Yemini, Y., Ellinson, Y., Ben-Ari, R., Gannot, S., & Fetaya, E. (2026). SSNAPS: Audio-Visual Separation of Speech and Background Noise with Diffusion Inverse Sampling. arXiv Preprint, arXiv:2602.01394. doi:10.48550/arXiv.2602.01394.

[19] Lee, S., Jung, C., Jang, Y., Kim, J., & Chung, J. S. (2024). Seeing Through the Conversation: Audio-Visual Speech Separation Based on Diffusion Model. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 12632–12636. doi:10.1109/ICASSP48485.2024.10447679.

[20] Gogate, M., Dashtipour, K. K., & Hussain, A. (2024). A Lightweight Real-time Audio-Visual Speech Enhancement Framework. In 3rd COG-MHEAR Workshop on Audio-Visual Speech Enhancement (AVSEC), ISCA, 19–23. doi:10.21437/avsec.2024-5.

[21] López-Espejo, I., Joglekar, A., Peinado, A. M., & Jensen, J. (2024). On Speech Pre-emphasis as a Simple and Inexpensive Method to Boost Speech Enhancement. IberSPEECH 2024, ISCA, 96–100. doi:10.21437/iberspeech.2024-20.

[22] Yang, W., Li, P., Yang, W., Liu, Y., He, Y., Petrosian, O., & Davydenko, A. (2023). Research on Robust Audio-Visual Speech Recognition Algorithms. Mathematics, 11(7), 1733. doi:10.3390/math11071733.

[23] Richter, J., Frintrop, S., & Gerkmann, T. (2023). Audio-Visual Speech Enhancement with Score-Based Generative Models. Speech Communication - 15th ITG Conference, 275–279. doi:10.30420/456164054.

[24] Chen, C. W., Hou, J. C., Tsao, Y., Chen, J. C., & Chien, S. Y. (2024). DAVSE: A diffusion-based generative approach for audio-visual speech enhancement. 3rd COG-MHEAR Workshop on Audio-Visual Speech Enhancement (AVSEC), 1 September 2024, Kos, Greece.

[25] Wahab, F., Saleem, N., Hussain, A., Rizwan, M., & Hossen, M. B. (2024). Multi-Model Dual-Transformer Network for Audio-Visual Speech Enhancement. In 3rd COG-MHEAR Workshop on Audio-Visual Speech Enhancement (AVSEC), ISCA, 1–5. doi:10.21437/avsec.2024-1.

[26] Chen, J., Wang, Z., Tuo, D., Wu, Z., Kang, S., & Meng, H. (2022). FullSubNet+: Channel Attention Fullsubnet with Complex Spectrograms for Speech Enhancement. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2022-May, 7857–7861. doi:10.1109/ICASSP43922.2022.9747888.

[27] Zhu, Q. S., Zhou, L., Zhang, J., Liu, S. J., Hu, Y. C., & Dai, L. R. (2023). Robust Data2VEC: Noise-Robust Speech Representation Learning for ASR by Combining Regression and Improved Contrastive Learning. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, 2023-June, 1–5. doi:10.1109/ICASSP49357.2023.10095373.

[28] Tiwari, S., Mentel, G., Si Mohammed, K., Rehman, M. Z., & Lewandowska, A. (2024). Unveiling the role of natural resources, energy transition and environmental policy stringency for sustainable environmental development: Evidence from BRIC +1. Resources Policy, 96, 105204. doi:10.1016/j.resourpol.2024.105204.

[29] Lian, J., Baevski, A., Hsu, W. N., & Auli, M. (2023, December). Av-data2vec: Self-supervised learning of audio-visual speech representations with contextualized target representations. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 1-8. doi:10.1109/ASRU57964.2023.10389642.

[30] Pala, A. K., Mallik, S., Tripathy, M., Sahoo, R. R., Swain, R., & Dash, D. K. (2026). Deep Learning Based Face Recognition System with Modified MTCNN and FaceNet. Computing, Communication and Intelligence, 77-82.

[31] Zhang, N., Luo, J., & Gao, W. (2020). Research on face detection technology based on MTCNN. Proceedings - 2020 International Conference on Computer Network, Electronic and Automation, ICCNEA 2020, 154–158. doi:10.1109/ICCNEA50255.2020.00040.

[32] Xu, X., Tu, W., Yang, Y., Li, J., Zhang, Y., & Chen, H. (2026). Contribution-aware Dynamic Multi-modal Balance for Audio-Visual Speech Separation. IEEE Transactions on Multimedia, 1–13. doi:10.1109/tmm.2026.3654399.

[33] Anwar, M., Shi, B., Goswami, V., Hsu, W. N., Pino, J., & Wang, C. (2023). MuAViC: A Multilingual Audio-Visual Corpus for Robust Speech Recognition and Robust Speech-to-Text Translation. Proceedings of the Annual Conference of the International

Speech Communication Association, INTERSPEECH, 2023-August, 4064–4068. doi:10.21437/Interspeech.2023-2279.

[34] Sheng, C., Kuang, G., Bai, L., Hou, C., Guo, Y., Xu, X., Pietikainen, M., & Liu, L. (2024). Deep Learning for Visual Speech Analysis: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(9), 6001–6022. doi:10.1109/TPAMI.2024.3376710.

[35] Zhang, S., Shankar, S., Nguyen, T., Fanelli, A., & Fiterau, M. (2025). Audio-Visual Speech Separation via Bottleneck Iterative Network. arXiv Preprint, arXiv:2507.07270. doi:10.48550/arXiv.2507.07270.

[36] Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 2020-December, 12449–12460.

[37] Wang, C., & Liu, F. (2025). Ghost Module-Enhanced MTCNN: A Lightweight Cascade Framework for High-Accuracy Face Detection in Edge-Deployable Scenarios. IEEE Access, 13, 107694–107709. doi:10.1109/ACCESS.2025.3581428.

[38] Zhang, X., Ren, X., Zheng, X., Chen, L., Zhang, C., Guo, L., & Yu, B. (2021). Low-delay speech enhancement using perceptually motivated target and loss. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2, 826–830. doi:10.21437/Interspeech.2021-1410.

[39] Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., & Rubinstein, M. (2018). Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. ACM Transactions on Graphics, 37(4), 3201357. doi:10.1145/3197517.3201357.

[40] Wu, Y., Li, C., & Qian, Y. (2023). Light-Weight Visualvoice: Neural Network Quantization on Audio Visual Speech Separation. ICASSPW 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing Workshops, Proceedings, 1–5. doi:10.1109/ICASSPW59220.2023.10193263.

[41] Roux, J. Le, Wisdom, S., Erdogan, H., & Hershey, J. R. (2019). SDR - Half-baked or Well Done? ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, 2019-May, 626–630. doi:10.1109/ICASSP.2019.8683855.

[42] Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Transactions on Audio Speech and Language Processing, 29, 3451–3460. doi:10.1109/TASLP.2021.3122291.

[43] Rahimi, A., Afouras, T., & Zisserman, A. (2022). Reading to Listen at the Cocktail Party: Multi-Modal Speech Separation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2022-June, 10483–10492. doi:10.1109/CVPR52688.2022.01024.

[44] Zhang, Z., Li, X., Li, Y., Dong, Y., Wang, D., & Xiong, S. (2021). Neural noise embedding for end-to-end speech enhancement with conditional layer normalization. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2021-June, 7113–7117. doi:10.1109/ICASSP39728.2021.9413931.