Stability-Aware Preprocessing for Audio-Visual Speech Processing

Sara M. Sh
College of Engineering, Al-Iraqi University, Saba'a Abkar Complex
Baghdad, Iraq
saramohameed9a9@gmail.com
ORCID:0009-0000-6573-5831

Baraa M. Albaker
College of Engineering, Al-Iraqi University, Saba'a Abkar Complex
Baghdad, Iraq
https://orcid.org/0000-0002-6030-3121

***Abstract*—Audio-visual speech processing (AVSP) systems offer a compelling framework for robust speech analysis in adverse and unconstrained environments; however, their deployment in real-world conditions remains challenged by uncontrolled variability in audio and visual inputs. Factors such as environmental noise, illumination changes, spatial misalignment, and temporal inconsistency introduce instability that is frequently addressed through increasingly complex model architectures, while preprocessing is often treated as a secondary implementation step. This paper presents a unified, stability-aware preprocessing pipeline that reframes preprocessing as a primary design layer operating entirely prior to learning and inference. The proposed framework systematically conditions audio and visual streams through modality-specific normalization and explicit cross-modal consistency enforcement, yielding statistically coherent and temporally aligned representations. By constraining input variability and suppressing nuisance factors unrelated to speech content, the preprocessing stage emphasizes task-relevant structures and reduces the burden placed on downstream models, enabling lightweight yet effective learning behavior. The pipeline is designed to be architecture-agnostic and applicable across diverse audio-visual datasets, supporting consistent model behavior without reliance on dataset-specific preprocessing heuristics. Rather than pursuing task-level performance optimization, this work adopts an observational evaluation perspective to examine how principled preprocessing influences learning stability, convergence behavior, and robustness under realistic operating conditions. The findings position unified preprocessing as a foundational mechanism for building efficient, reproducible, and scalable AVSP systems suitable for real-world deployment.**

Keywords—audio-visual speech processing, stability-aware preprocessing, multimodal signal conditioning, representation stability, lightweight multimodal processing

I. INTRODUCTION

Audio-visual speech processing (AVSP) plays a crucial role in robust speech analysis under challenging real-world conditions involving noise, speaker overlap, reverberation, and channel variability. Despite advances in deep learning–based audio-only speech enhancement and separation, acoustic cues remain highly vulnerable to non-stationary noise, limiting robustness in adverse environments [1].

Audiovisual approaches address these limitations by incorporating visual speech information such as lip motion and facial dynamics, which provide complementary articulatory cues that remain reliable when acoustic signals are degraded. Prior work has consistently demonstrated that visual modalities significantly enhance performance in noisy and multi-speaker scenarios, motivating their widespread use in speech separation and recognition tasks [2], [3].

Recent AVSP research has largely emphasized architectural innovation, including deep clustering, transformer-based models, and diffusion-driven frameworks [4], [5]. Although these approaches achieve strong benchmark results, their robustness often degrades in practical settings due to sensitivity to audiovisual misalignment, data variability, and unstable optimization behavior [6]. These challenges are further intensified by large-scale in-the-wild datasets, where variability in illumination, pose, synchronization, and recording conditions can cause early-stage representation inconsistencies that propagate through deep models and impair convergence and cross-modal alignment [7].

Despite this, preprocessing is frequently treated as a secondary implementation detail. In practice, statistical conditioning, geometric normalization, and temporal consistency of audio and visual inputs critically shape feature distributions and training stability. Differences in amplitude scaling, illumination handling, spatial normalization, or feature equalization can substantially influence learning behavior, independent of model architecture [8].

Motivated by these observations, this work repositions preprocessing as a first-class design element in audiovisual speech processing. Rather than proposing a new architecture or targeting state-of-the-art performance, the focus is on a stability-aware preprocessing framework that enforces geometric, photometric, temporal, and statistical consistency across audio and visual streams. The aim is to establish well-conditioned multimodal representations that reduce model burden and promote reliable learning under realistic conditions.

The contributions of this work are threefold: (i) a unified preprocessing design for systematic audio–visual conditioning; (ii) an analysis of the impact of statistical conditioning and geometric normalization on cross-modal consistency and training stability; and (iii) practical, architecture-agnostic preprocessing guidelines suitable for lightweight AVSP systems in unconstrained environments. Experimental illustrations are used solely to analyze stability trends and representation consistency, without claiming superiority over existing methods.

II. RELATED LITERATURE

Audio-visual speech processing (AVSP) has been widely investigated as a robust alternative to audio-only speech systems, particularly in noisy and multi-speaker environments where acoustic cues become unreliable. By incorporating visual speech information such as lip motion and facial dynamics, audiovisual systems exploit complementary articulatory cues that remain informative even when the acoustic signal is degraded [2], [3].

Early AVSP research demonstrated that integrating visual information can significantly improve speech separation and enhancement performance compared with audio-only baselines. Subsequent studies introduced deep learning–

based audiovisual models that jointly learn audio and visual representations for speech separation and recognition tasks [4], [5]. These approaches showed notable performance improvements but also revealed a strong dependence on accurate audiovisual alignment and consistent feature extraction.

More recent work has focused on advanced modeling architectures, including transformer-based multimodal frameworks and diffusion-based generative models for audiovisual speech enhancement and separation [6]–[8]. While these methods achieve strong benchmark results, they often exhibit increased sensitivity to input variability, audiovisual misalignment, and distribution shifts, particularly in unconstrained real-world datasets.

Large-scale benchmarking efforts such as AV-SUPERB further highlight the variability of system performance across datasets and experimental setups, indicating that preprocessing and input conditioning can substantially influence learning outcomes [9]. Additional research has explored audiovisual synchronization and alignment strategies to improve cross-modal correspondence in multimodal speech systems [10], [11].

Despite these advances, preprocessing is typically embedded implicitly within complex modeling pipelines rather than treated as an explicit design component. As a result, the systematic design of stable and unified preprocessing strategies for audiovisual speech processing remains relatively underexplored. This observation motivates the present work, which investigates preprocessing as a foundational mechanism for stabilizing multimodal representations prior to downstream learning as shown in Table I.

TABLE I. REPRESENTATIVE AVSP STUDIES AND THEIR SENSITIVITY TO INPUT CONDITIONING

Ref.	Task Focus	Modeling Approach	Observed Sensitivity	Implication
[5]	AV Speech Separation	Cross-modal correspondence learning	Sensitive to feature distribution mismatch	Stable multimodal conditioning is required
[7]	AV Speaker Separation	Transformer-based architecture	Sensitive to audiovisual misalignment	Alignment-aware preprocessing is important
[6]	AV Speech Enhancement	End-to-end multimodal model	Sensitive to input variability	Robust normalization improves stability
[8]	Speech Enhancement	Diffusion-based generative model	Sensitive to distribution shifts	Reliable statistical conditioning is necessary
[9]	AV Benchmarking (AV-SUPERB)	Large-scale benchmark evaluation	Performance varies across preprocessing setups	Preprocessing affects reproducibility.

III. DESIGN PRINCIPLES OF UNIFIED PREPROCESSING

This section presents the design principles of the proposed unified preprocessing pipeline for audio-visual speech processing. The pipeline is formulated as a pre-learning stage that stabilizes and conditions multimodal inputs prior to any learning, fusion, or inference processes. Within this framework, preprocessing is treated as an explicit design layer that operates independently of downstream model architectures. As illustrated in Fig. 1, the pipeline applies modality-specific preprocessing to the audio and visual streams, followed by a cross-modal conditioning stage that enforces temporal alignment and statistical consistency. This design aims to produce well-conditioned multimodal representations before downstream processing, while learning architectures and evaluation procedures remain outside the scope of this work.

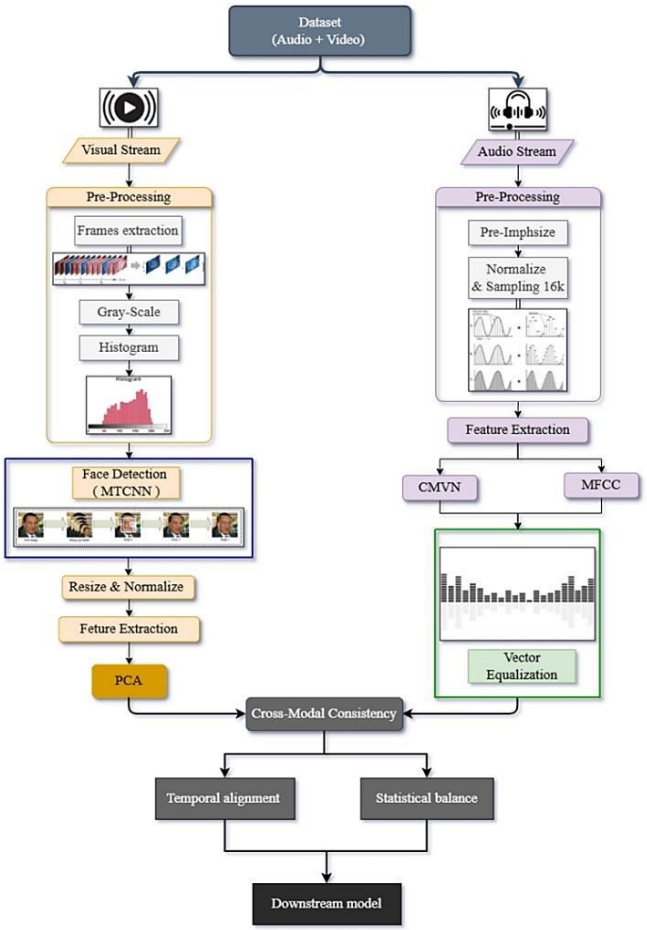

Fig. 1. Overview of the proposed unified preprocessing pipeline for audio-visual speech processing

A. *Preprocessing as a Stability Mechanism*

In many audio-visual speech processing systems, preprocessing is often treated as a peripheral step. However, the statistical properties of input representations strongly influence training stability and convergence behavior. Audio and visual signals captured in unconstrained environments exhibit substantial variability due to illumination changes, recording conditions, and speaker characteristics. Such variability introduces distribution shifts that increase optimization difficulty and place additional burden on downstream models.

From a design perspective, preprocessing functions as a mechanism for stabilizing input distributions prior to learning. By enforcing statistical normalization, spatial consistency, and temporal alignment, preprocessing reduces unnecessary variability and constrains the effective input space. Consequently, models can focus on task-relevant patterns rather than compensating for input inconsistencies. Importantly, the objective of preprocessing is not to directly improve task performance, but to establish stable and reproducible learning conditions across datasets and experimental settings.

B. *Visual Stream Preprocessing*

Visual speech signals captured in unconstrained environments are highly sensitive to variations in

illumination, facial pose, camera resolution, and background clutter. If left unregulated, these factors propagate into learned representations and degrade cross-modal correspondence in audiovisual systems. The proposed visual preprocessing pipeline therefore stabilizes visual inputs through photometric, geometric, and statistical conditioning prior to representation learning.

The process begins with frame extraction from the input video stream to form a temporally consistent sequence. As illustrated in Fig. 2, each frame is converted from RGB to grayscale to suppress chromatic variability while preserving structural facial information. Since visual speech cues primarily depend on lip motion and facial dynamics rather than color information, grayscale conversion reduces photometric complexity while maintaining linguistically relevant visual features [3].

Audio-visual speech processing (AVSP) has been widely investigated as a robust alternative to audio-only speech systems, particularly in noisy and multi-speaker environments where acoustic cues become unreliable. By incorporating visual speech information such as lip motion and facial dynamics, audiovisual systems exploit complementary articulatory cues that remain informative even when the acoustic signal is degraded [2], [3].

Early AVSP research demonstrated that integrating visual information can significantly improve speech separation and enhancement performance compared with audio-only baselines. Subsequent studies introduced deep learning–based audiovisual models that jointly learn audio and visual representations for speech separation and recognition tasks [4], [5]. These approaches showed notable performance improvements but also revealed a strong dependence on accurate audiovisual alignment and consistent feature extraction. More recent work has focused on advanced modeling architectures, including transformer-based multimodal frameworks and diffusion-based generative models for audiovisual speech enhancement and separation [6–8]. While these methods achieve strong benchmark results, they often exhibit increased sensitivity to input variability, audiovisual misalignment, and distribution shifts, particularly in unconstrained real-world datasets.

Large-scale benchmarking efforts such as AV-SUPERB further highlight the variability of system performance across datasets and experimental setups, indicating that preprocessing and input conditioning can substantially influence learning outcomes [9]. Additional research has explored audiovisual synchronization and alignment strategies to improve cross-modal correspondence in multimodal speech systems [10], [11].

Despite these advances, preprocessing is typically embedded implicitly within complex modeling pipelines rather than treated as an explicit design component. As a result, the systematic design of stable and unified preprocessing strategies for audiovisual speech processing remains relatively underexplored. This observation motivates the present work, which investigates preprocessing as a foundational mechanism for stabilizing multimodal representations prior to downstream learning. evaluation procedures remain outside the scope of this work.

Fig. 2. Visual photometric normalization and spatial localization.

To further stabilize intensity statistics, global histogram equalization is applied to grayscale frames, as illustrated in Fig. 2. This operation enforces consistent contrast characteristics across frames and recording conditions, thereby reducing photometric variability and promoting statistical uniformity across large-scale datasets [4].

Following photometric normalization, full-face detection is performed using a Multi-task Cascaded Convolutional Network (MTCNN) [12], as illustrated in Fig. 3. The cascaded architecture comprising the Proposal Network (P-Net), Refinement Network (R-Net), and Output Network (O-Net) progressively filters candidate regions and refines facial boundaries to localize the complete facial region along with landmark estimation.

Fig. 3. Multi-stage full-face detection using MTCNN for visual stream preprocessing.

Unlike approaches that focus solely on the mouth region, the proposed framework retains the entire detected face. This design choice improves robustness under real-world conditions where localized regions may be partially occluded by masks or hand movements. Preserving holistic facial motion and structural cues provides more stable visual information prior to representation learning.

After photometric and spatial normalization, the localized facial region undergoes statistical conditioning using Principal Component Analysis (PCA). PCA serves as a lightweight representation mechanism that reduces visual redundancy while preserving dominant facial structures relevant to visual speech [3]. As illustrated in Fig. 4, a small number of principal components captures most of the visual variance, enabling compact and stable feature representations.

Within the proposed framework, PCA is not intended to replace deep visual encoders but rather to regularize the visual feature space prior to audiovisual learning. By constraining feature variability, PCA contributes to stable training behavior while maintaining low computational complexity, consistent with the stability-aware design philosophy of the preprocessing pipeline.

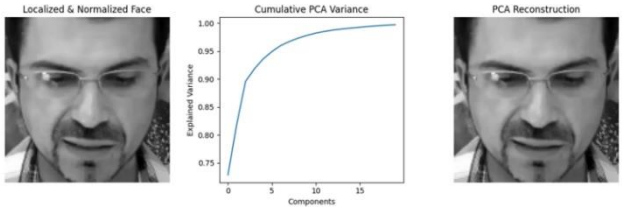

Fig. 4. Statistical conditioning using Principal Component Analysis.

In Fig. 4. Left: cumulative explained variance as a function of retained principal components. Right: reconstructed facial representation using the selected components, illustrating noise suppression and structural preservation.

A. Audio Stream Preprocessing

Speech signals recorded in unconstrained environments exhibit significant variability due to differences in recording

devices, speaker loudness, background noise, and sampling rates. If left unregulated, these factors propagate into extracted features and negatively affect learning stability. The proposed audio preprocessing pipeline, therefore conditions raw speech signals through a sequence of signal-level operations designed to enforce statistical consistency prior to learning.

The process begins with pre-emphasis filtering applied to the raw waveform, as illustrated in Fig. 5. This operation compensates for the natural spectral tilt of speech by amplifying higher-frequency components associated with articulation, producing a more balanced spectral distribution and reducing low-frequency dominance [1].

Following spectral conditioning, amplitude normalization is applied to standardize signal energy across utterances. By suppressing loudness-related variability, normalization prevents downstream models from learning correlations related to recording gain rather than speech content, thereby improving the stability of subsequent feature extraction stages [1].

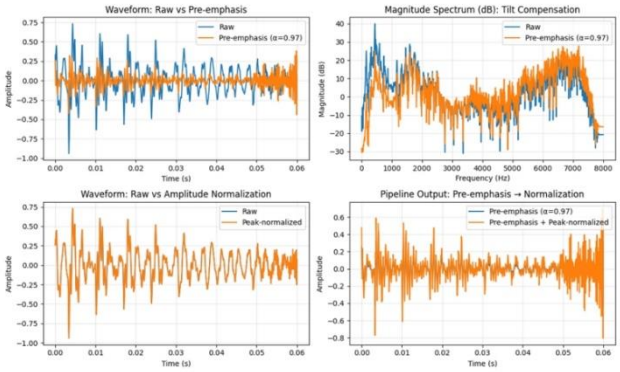

Fig. 5. Audio preprocessing effects of pre-emphasis filtering and amplitude normalization.

Fig. 5 illustrates the effects of pre-emphasis filtering and amplitude normalization on the speech waveform. Pre-emphasis amplifies higher-frequency components and compensates for the natural spectral tilt of speech, while amplitude normalization stabilizes signal energy across utterances. Together, these operations produce a waveform with reduced amplitude fluctuations and improved dynamic range stability prior to feature extraction. To ensure consistent time–frequency resolution, all audio signals are resampled to a uniform sampling rate before feature extraction. This step mitigates variability introduced by heterogeneous recording sources and supports reproducible feature representations. Following signal-level conditioning, Mel-Frequency Cepstral Coefficients (MFCCs) are extracted as the primary audio representation due to their compact and perceptually motivated description of speech spectra. Cepstral Mean and Variance Normalization (CMVN) is subsequently applied to enforce zero-mean and unit-variance statistics across cepstral dimensions, thereby reducing channel effects and inter-utterance variability [1]. Finally, vector-level equalization aligns feature distributions across samples, producing well-conditioned audio representations for downstream learning. Within the proposed framework, this step serves primarily as a stability-enhancing mechanism rather than a performance-optimizing transformation.

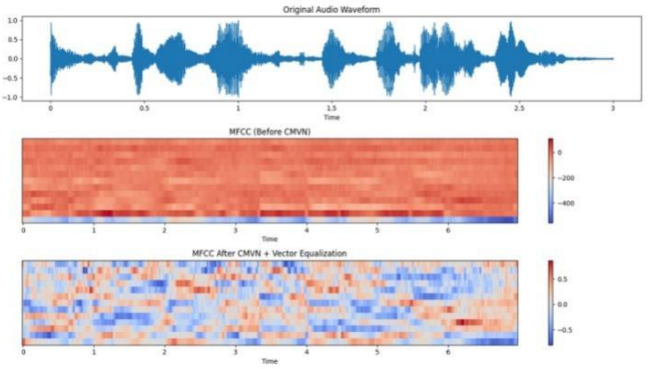

Fig. 6. Audio stream preprocessing stages. Top: original speech waveform illustrating raw amplitude variability.

Although modality-specific preprocessing is necessary, it is insufficient for stable audiovisual learning when audio and visual streams are combined. Even when each modality is independently normalized, inconsistencies across modalities may degrade correspondence learning and reduce the effectiveness of multimodal fusion as shown in Fig. 6.

Temporal alignment is a fundamental requirement for cross-modal consistency. Misalignment between audio and visual streams introduces ambiguity in correspondence learning and can negatively affect representation stability and downstream performance [10], [11].

In addition to temporal coherence, statistical balance between modalities is equally important. Differences in feature scale, variance, or dynamic range may bias fusion mechanisms toward a dominant modality, thereby limiting effective multimodal integration. Ensuring comparable statistical conditioning across audio and visual representations promotes balanced modality contribution and more stable learning behavior [4].

Within the proposed framework, cross-modal consistency is treated as an explicit preprocessing objective. By enforcing temporal alignment and statistical balance before fusion, the pipeline establishes well-conditioned multimodal representations that provide a stable foundation for subsequent learning stages.

TABLE II. SOURCES OF INSTABILITY AND CORRESPONDING PREPROCESSING STRATEGIES

Instability Source	Preprocessing Operation	Expected Effect	Ref
Illumination variability	Grayscale conversion	Reduced photometric variance	[3]
Contrast inconsistency	Histogram equalization	Stabilized frame statistics	[4]
Background clutter	Face localization (ROI)	Improved spatial consistency	[12]
Spectral variability	Pre-emphasis filtering	Balanced spectral distribution	[1]
Loudness variation	Amplitude normalization	Standardized signal energy	[3]
Audio-visual misalignment	Temporal synchronization	Stable multimodal correspondence	[10], [11]

Table II summarizes the relationship between common instability in real-world audiovisual data and the preprocessing operations used in the proposed pipeline to mitigate them. For clarity, Algorithm 1 outlines the unified preprocessing pipeline at a conceptual level, abstracting implementation details and downstream learning architectures.

Algorithm 1: Unified Preprocessing Pipeline (Design-Level Pseudocode)

```
Input:
   Raw audio stream A
   Raw visual stream V
Output:
   Conditioned audio representation A'
   Conditioned visual representation V'
Begin
   /* Visual Stream Conditioning */
   Extract frames from V
   Perform photometric conditioning
   Normalize global intensity statistics
   Localize facial region of interest
   Apply spatial normalization
   Reduce visual redundancy using a stability-oriented
    transformation
   Obtain conditioned visual representation V'
   /* Audio Stream Conditioning */
   Apply spectral conditioning to A
   Normalize signal amplitude
   Enforce uniform temporal resolution
   Extract statistically normalized spectral features
   Apply vector-level feature equalization
   Obtain conditioned audio representation A'
   /* Cross-Modal Consistency */
   Enforce temporal alignment between A' and V'
   Balance statistical properties across modalities
   Return A', V'
End
```

IV. Experimental Illustration

This section provides an experimental illustration to contextualize the behavior of the proposed unified preprocessing pipeline. The objective is not to establish quantitative performance gains, but rather to qualitatively examine how preprocessing influences learning stability under realistic audiovisual conditions.

A. Dataset Description

The illustration is conducted using a large-scale audio-visual speech dataset containing synchronized audio–video recordings captured under unconstrained conditions. Such datasets exhibit substantial variability in speaker identity, illumination, recording devices, background noise, and temporal synchronization. Prior benchmarking studies have shown that learning stability and representation consistency in audiovisual systems are highly sensitive to input conditioning and preprocessing choices [9].

B. Backbone Model

visual backbone is adopted as an illustrative probe. The architecture includes a minimal temporal modeling component followed by a compact multimodal integration stage. The model is intentionally simple and is not optimized for task-level performance. Instead, it serves only to examine how stabilized representations influence learning behavior, consistent with prior studies that analyze multimodal representation dynamics using simplified architectures [4].

C. Observational Training Behavior

When the representations produced by the proposed preprocessing pipeline are used as inputs, several qualitative training trends are observed. First, smoother convergence behavior appears during optimization, indicating improved input conditioning and reduced sensitivity to unstable gradient updates. Second, lower variance across training iterations emerges due to increased statistical consistency within and across modalities. Finally, improved stability is observed under challenging conditions such as audiovisual misalignment and low-quality visual inputs, which are known to affect multimodal learning performance [10], [11].

D. Scope Clarification

This illustration intentionally avoids reporting quantitative performance metrics or comparative evaluations. The objective is to highlight preprocessing-induced stability effects rather than to demonstrate state-of-the-art performance. As illustrated in Fig. 7, the proposed framework focuses on preprocessing, while modeling, fusion, and evaluation stages are treated as downstream components outside the scope of this work.

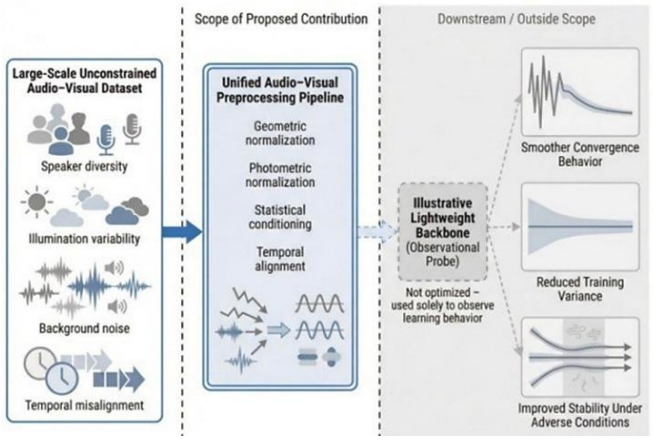

Fig. 7. Conceptual Experimental Illustration of the Unified Preprocessing Pipeline.

The figure qualitatively contextualizes how preprocessing stabilizes representations and influences downstream learning behavior under realistic audiovisual variability. The illustrative backbone model is employed solely as an observational probe, while all modeling, fusion, and evaluation components lie outside the scope of this work.

V. Discussion

This work positions unified preprocessing as a foundational design layer in audio-visual speech processing rather than a secondary implementation step. A key observation is that a significant portion of learning instability originates at the level of raw signal conditioning and cross-modal consistency before model optimization. Variability introduced by illumination changes, amplitude scaling, spatial misalignment, and temporal inconsistency propagates into learned representations and increases the burden on downstream models [4].

By stabilizing input distributions prior to learning, the proposed framework shifts part of the robustness responsibility from model architecture to principled input conditioning. Visual operations such as grayscale conversion and histogram equalization, together with CMVN-based audio normalization, constrain statistical variability and promote more stable optimization behavior, consistent with observations reported in previous audiovisual studies [4].

Because preprocessing is performed entirely before model introduction, the framework remains architecture-agnostic and can be applied across separation, enhancement, and recognition systems. This separation of preprocessing and modeling improves experimental clarity and

reproducibility, particularly in large-scale audiovisual benchmarks where preprocessing variability alone may produce inconsistent learning behavior [9].

The proposed design is particularly relevant in real-world conditions involving heterogeneous devices, illumination fluctuations, and imperfect synchronization. By enforcing spatial normalization, temporal alignment, and statistical balance, the pipeline produces multimodal representations that are more resilient to environmental variability and, therefore, more suitable for lightweight or resource-constrained deployments.

Nevertheless, preprocessing alone cannot resolve all challenges, particularly under severe occlusion, extreme noise, or missing modalities. Consequently, unified preprocessing should be considered a complementary design layer that enhances stability and efficiency while remaining compatible with downstream learning architectures.

VI. Conclusion

This paper has presented a unified preprocessing framework for audio-visual speech processing that treats preprocessing as a deliberate design decision rather than a secondary implementation step. By isolating preprocessing from downstream modeling, the framework highlights the importance of input conditioning and cross-modal consistency in shaping stable learning behavior prior to any optimization or inference stage. A key takeaway of this work is that a substantial portion of variability commonly observed in audio-visual systems can be addressed before introducing model-specific solutions. Through principled preprocessing of both audio and visual streams, the proposed framework establishes well-conditioned representations that simplify the learning problem and reduce unnecessary dependence on model capacity. The framework is intentionally architecture-agnostic, allowing it to be integrated seamlessly into diverse audio-visual pipelines. This separation of concerns supports reproducibility and clarity in system design, enabling researchers to reason more explicitly about the origins of instability in multimodal learning and to address them at the appropriate stage. Looking ahead, this work motivates further investigation into how preprocessing strategies interact with learning dynamics across different audio-visual tasks and deployment scenarios. While downstream modeling remains essential, the perspective advanced in this paper suggests that principled preprocessing can play a complementary and foundational role in building robust and efficient audio-visual speech processing systems.

References

[1] D. Wang, “Deep learning for non-stationary noise suppression,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1–12, 2020.

[2] C. Li and Y. Qian, “Listen, watch, and understand at the cocktail party,” in Proc. INTERSPEECH, 2020, pp. 1426–1430.

[3] G. Sterpu, J. Saam, and N. Harte, “How much does visual speech help in noise?,” Computer Speech & Language, vol. 62, p. 101089, 2020.

[4] D. Michelsanti, Z. Tan, S. Zhang, Y. Xu, and J. Jensen, “An overview of deep-learning-based audio-visual speech enhancement and separation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1368–1396, 2021.

[5] Y. Tang, P. Ma, and M. Pantic, “Audio-visual speech separation using cross-modal correspondence loss,” in Proc. ICASSP, 2021, pp. 6671–6675.

[6] Z. Zhu et al., “Real-time audio-visual end-to-end speechenhancement,” in Proc. ICASSP, 2023.

[7] V. A. Kalkhorani, A. Brendel, and E. A. P. Habets, “Time-domain transformer-based audio-visual speaker separation,” in Proc. INTERSPEECH, 2023.

[8] J. Richter, M. Tesch, and R. Haeb-Umbach, “Speech enhancement with diffusion-based models,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023.

[9] W.-C. Tseng et al., “AV-SUPERB: A multi-task benchmark for audio-visual speech processing,” in Proc. ICASSP, 2024.

[10] A. Afouras, S. Chung, and A. Zisserman, “Auto-AVSR: Audio-visual speech recognition with automatic synchronization,” in Proc. ICASSP, 2024.

[11] W. Wang, R. Gao, and K. Grauman, “Streaming audio-visual speech recognition with alignment regularization,” in Proc. INTERSPEECH, 2024.

[12] J. Zhao, Y. Cheng, Y. Xu, and W. Xiong, “Face detection and alignment in unconstrained environments: A comparative study of deep learning-based methods,” Neurocomputing, vol. 453, pp. 34–49, 2021.