Ingénierie des Systèmes d'Information
Vol. 31, No. 6, June, 2026, pp. 2085-2107
Journal homepage: http://iieta.org/journals/isi

A Systematic Review of Audio-Visual Speech Enhancement: Architectural Evolution, Diffusion-Based Learning, Self-Supervised Models, and Future Perspectives

Sara M. Sh*, Baraa M. Albaker

College of Engineering, Al-Iraqi University, Baghdad 10069, Iraq

Corresponding Author Email: sara.m@aliraqia.edu.iq

https://doi.org/10.18280/isi.310628

Received: 19 February 2026
Revised: 20 April 2026
Accepted: 5 May 2026
Available online: 30 June 2026

Keywords:
Audio-Visual Speech Enhancement, systematic review, multimodal learning, diffusion models, self-supervised learning, transformer architectures, audiovisual processing

ABSTRACT

Audio-Visual Speech Enhancement (AVSE) has become an important research direction for improving speech intelligibility and perceptual quality under challenging acoustic conditions by integrating complementary audio and visual information. However, rapid advances in deep learning, diffusion-based generation, and self-supervised representation learning have resulted in substantial variations in architectures, datasets, evaluation protocols, and deployment strategies, making systematic comparison increasingly difficult. This study presents a systematic analytical review of recent AVSE developments published between 2020 and 2025. Following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, 102 peer-reviewed studies collected from major scientific databases were analyzed to investigate architectural evolution, multimodal fusion strategies, benchmark datasets, evaluation metrics, computational requirements, and real-world deployment challenges. The review identifies a transition from conventional CNN-RNN-based approaches toward transformer-based, diffusion-based, and self-supervised multimodal frameworks with improved representation learning capabilities. However, these advances are accompanied by increased computational complexity, inference latency, dataset limitations, synchronization issues, and insufficient evaluation of fairness and privacy concerns. Based on comparative analysis, this study establishes a systems-oriented perspective linking model architecture, dataset characteristics, evaluation methodology, and deployment feasibility. Furthermore, several future research directions are discussed, including efficient diffusion-based AVSE, multilingual and privacy-aware multimodal learning, standardized benchmarking, and lightweight architectures for edge deployment. This review provides a comprehensive analytical foundation for researchers developing reliable and scalable next-generation AVSE systems.

1. INTRODUCTION

Audio-Visual Speech Enhancement (AVSE) has emerged as an important research direction for improving speech intelligibility and perceptual quality in acoustically degraded environments through the integration of complementary auditory and visual information sources [1]. Unlike conventional audio-only speech enhancement systems, AVSE frameworks utilize visual cues such as lip movements and facial dynamics to improve robustness under severe noise, speaker interference, reverberation, and signal degradation conditions [2]. The increasing demand for robust multimodal communication systems has accelerated the development of audiovisual speech processing technologies across intelligent communication systems, assistive hearing applications, human–computer interaction, teleconferencing platforms, and real-time multimedia services [3].

Early audiovisual speech processing systems primarily relied on handcrafted feature extraction methods, including Mel-Frequency Cepstral Coefficients (MFCCs), Discrete Cosine Transforms (DCTs), and conventional statistical learning approaches [4]. Although these techniques provided interpretable representations, their performance remained limited under highly variable acoustic conditions, visual occlusions, speaker variability, and unconstrained real-world environments [5]. The emergence of deep learning architectures, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), enabled improved temporal-spectral representation learning and multimodal fusion capabilities [6]. More recently, transformer-based, diffusion-based, and self-supervised learning (SSL) frameworks have significantly advanced audiovisual speech enhancement by improving contextual modeling, multimodal synchronization, perceptual reconstruction quality, and cross-modal representation learning across complex acoustic scenarios [7].

The rapid evolution of AVSE architectures has been accompanied by substantial growth in audiovisual benchmark datasets. Early datasets such as GRID, LRS2, and LRS3 provided synchronized audiovisual speech corpora for controlled experimental evaluation. However, many currently available audiovisual datasets still suffer from limited vocabulary diversity, environmental variability, and multilingual coverage. To address these limitations, several

large-scale datasets have recently been introduced, including AVSpeech, VoxCeleb2-AV, and MuAViC. These datasets include large numbers of speakers, multilingual speech content, spontaneous audiovisual recordings, and diverse recording environments, enabling broader generalization and more realistic robustness evaluation across different scenarios.

Despite these advances, considerable variability still exists among current datasets with respect to data collection procedures, annotation quality, demographic diversity, synchronization quality, and recording environments. Similarly, AVSE methods also vary substantially in terms of architectural design, multimodal fusion strategies, and feature representation techniques. Consequently, benchmarking methodologies and evaluation procedures differ significantly across studies, making fair comparison and reproducibility increasingly difficult.

In parallel with the growth of large-scale audiovisual datasets, AVSE methods have achieved substantial improvements in recent years, particularly through diffusion-based generative frameworks and self-supervised multimodal learning approaches. These methods have demonstrated strong capabilities in multimodal representation learning and perceptual enhancement under severely degraded acoustic conditions. Nevertheless, despite these improvements, current AVSE systems still face several important limitations. Advanced architectures often introduce high computational complexity, increased memory consumption, longer inference latency, expensive training procedures, and reduced deployment efficiency in real-time applications, especially in edge-based environments.

Diffusion-based AVSE architectures, in particular, have shown strong perceptual reconstruction performance in audiovisual enhancement tasks. However, these architectures frequently rely on iterative inference processes that are computationally expensive and difficult to deploy in latency-sensitive applications. Consequently, there is growing interest in developing more efficient and deployment-oriented diffusion-based AVSE frameworks suitable for real-time and resource-constrained environments. Future research is expected to focus on improving the efficiency, scalability, and synchronization stability of diffusion-based AVSE systems.

In addition to perceptual enhancement, increasing attention is being directed toward integrating SSL frameworks such as HuBERT, wav2vec 2.0, and AV-HuBERT with multimodal diffusion models. Such integration may reduce dependence on manually labeled data while improving adaptive contextual representation learning, cross-modal alignment, and robustness under severe acoustic degradation conditions. Despite their potential, diffusion-based AVSE systems still face challenges related to synchronization instability, computational scalability, cross-modal consistency, and deployment feasibility in real-world applications.

Although lightweight AVSE architectures have recently demonstrated promising real-time performance, their enhancement capability often degrades in highly challenging acoustic environments. As a result, achieving a balance between perceptual enhancement quality, synchronization reliability, computational efficiency, and deployment feasibility remains a major challenge in modern AVSE research [8].

Although significant progress has been made in AVSE research using deep learning approaches, many existing studies still exhibit important limitations related to methodology, experimental design, reproducibility, benchmarking consistency, and ethical considerations. These limitations affect the reliability of reported performance results, the fairness of comparisons among different AVSE methods, and the practicality of real-world deployment. Such challenges arise not only from individual studies, but also from variability in multimodal datasets, evaluation protocols, frame synchronization strategies, and the rapid evolution of deep learning architectures and training methodologies.

To address these challenges, it is necessary to evaluate the performance of models trained or optimized on one dataset across multiple AVSE benchmarks rather than relying solely on single-dataset evaluation. Cross-dataset generalization assessment should involve datasets that are as diverse as possible with respect to several factors, including (1) acoustic conditions, (2) language and accent variability, (3) speaker demographics, (4) recording environments, and (5) synchronization quality. This is particularly important for transformer-based, diffusion-based, and self-supervised AVSE models, which are typically trained on large-scale datasets and may inherit biases, synchronization inconsistencies, or dataset-specific characteristics from the training data.

Consequently, modern AVSE review studies must extend beyond providing a simple summary of existing methods and datasets. A comprehensive analytical review should also examine architectural design strategies, multimodal fusion approaches, feature representation methods, benchmarking methodologies, computational efficiency, synchronization reliability, deployment feasibility, multilingual adaptability, demographic fairness, privacy considerations, and reproducibility challenges.

Therefore, this review provides a structured analytical study of AVSE research published between 2020 and 2025 and selected from IEEE, Elsevier, Springer, and ACM databases using Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-based study selection procedures. A total of 102 studies were analyzed to investigate transformer-based architectures, diffusion-based frameworks, self-supervised multimodal learning systems, and lightweight deployment-oriented AVSE models. In addition, this review examines multimodal fusion strategies, feature extraction techniques, dataset characteristics, evaluation methodologies, benchmarking practices, and performance-efficiency trade-offs across modern AVSE systems.

Particular attention is given to the relationship between perceptual enhancement quality, synchronization performance, computational complexity, and real-time deployment capability. The review also discusses several emerging research directions, including multilingual low-resource AVSE, privacy-aware multimodal learning, efficient edge-based audiovisual processing, self-supervised representation learning, and large-scale generative multimodal frameworks. The primary outcome of this paper is therefore a structured analytical overview of current AVSE research trends rather than a purely comprehensive summary of the field.

2. RESEARCH QUESTIONS AND OBJECTIVES

2.1 Research framework

This systematic analytical review was designed using the PICOS framework to establish a structured foundation for

formulating research questions, selecting studies, and conducting comparative analytical synthesis [9]. Within this framework, the population (P) consisted primarily of AVSE systems and closely related audiovisual speech processing studies integrating acoustic and visual modalities. The intervention (I) focused on deep learning-based multimodal architectures and fusion strategies, including convolutional, recurrent, transformer-based, diffusion-based, self-supervised, and lightweight audiovisual frameworks. Comparisons (C) were conducted across different architectural paradigms, multimodal fusion mechanisms, datasets, and deployment-oriented design strategies. The outcomes (O) included perceptual enhancement quality, speech intelligibility, synchronization robustness, computational efficiency, and real-time feasibility, evaluated using commonly reported metrics such as Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI), Signal-to-Distortion Ratio (SDR), and Word Error Rate (WER). The study design (S) included peer-reviewed journal and conference publications published between 2020 and 2025. The PICOS structure was adopted to maintain methodological consistency throughout the review process by linking research objectives, analytical procedures, study selection criteria, evaluation protocols, and comparative synthesis strategies within a unified systematic review framework. To reduce redundancy and improve structural clarity, the relationship between research questions, analytical objectives, and methodological stages is summarized within a single integrated framework presented in Figure 1.

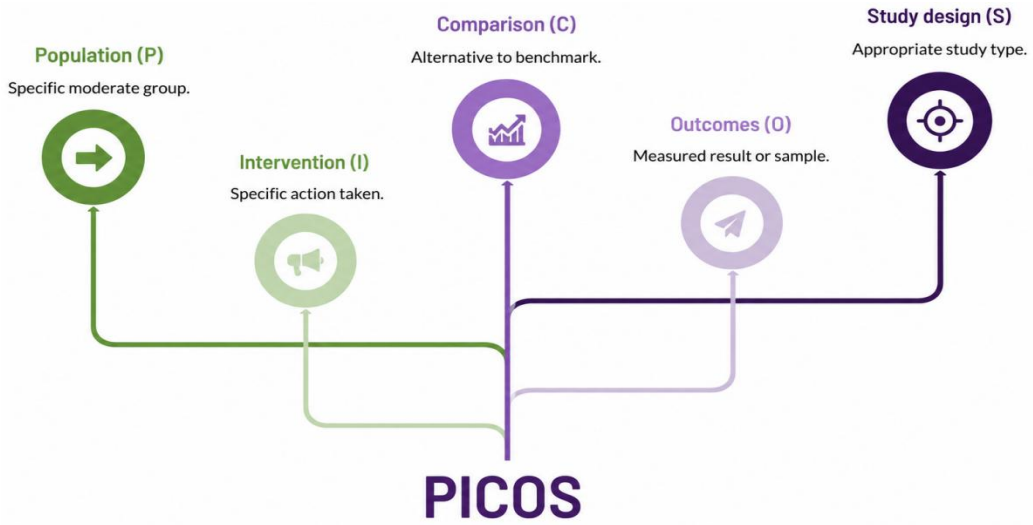

Figure 1. Integrated PICOS-based framework linking research questions, analytical objectives, and methodological stages within the systematic Audio-Visual Speech Enhancement (AVSE) review process

2.2 Research questions

Based on the PICOS framework, the review was guided by three primary research questions addressing architectural evolution, multimodal robustness, and deployment-oriented performance analysis in modern AVSE systems:

RQ1: How have transformer-based, diffusion-based, self-supervised, and lightweight architectures influenced recent advances in audiovisual speech enhancement?

RQ2: What trade-offs exist between perceptual enhancement quality, computational complexity, synchronization robustness, and real-time deployment feasibility across modern AVSE systems?

RQ3: How do dataset diversity, multimodal fusion strategies, benchmarking protocols, and evaluation methodologies influence the reproducibility and generalization capability of AVSE systems?

These questions were formulated to support comparative analytical synthesis rather than sequential literature summarization. In addition to technical evaluation, the framework also considers broader methodological and ethical challenges involving multilingual generalization, demographic fairness, benchmarking consistency, and privacy considerations associated with audiovisual speech datasets and multimodal learning systems. Accordingly, the primary objective of this review is to provide a systematic analytical synthesis of recent advances in AVSE between 2020 and 2025 while identifying the major methodological, computational, and deployment-oriented challenges affecting current multimodal systems. The review compares transformer-based, diffusion-based, self-supervised, and lightweight AVSE architectures, with emphasis on multimodal fusion mechanisms, synchronization reliability, perceptual enhancement quality, computational scalability, and real-time feasibility. In addition to architectural analysis, the study investigates the influence of dataset diversity, multilingual variability, evaluation inconsistency, benchmarking limitations, demographic fairness concerns, and privacy considerations on AVSE reproducibility and generalization capability. Through this analytical perspective, the review aims to establish a systems-oriented framework connecting architectural design strategies, dataset characteristics, evaluation methodologies, and deployment constraints within unified AVSE assessment pipelines.

2.3 Objectives of the review

Aligned with the research questions presented above, this review pursues several interconnected objectives intended to strengthen the analytical understanding of contemporary AVSE research. First, the review systematically maps and categorizes recent AVSE studies according to architectural design, multimodal fusion strategy, dataset characteristics, and evaluation methodology. Particular attention is given to

transformer-based, diffusion-based, self-supervised, CNN-RNN hybrid, and lightweight multimodal frameworks developed between 2020 and 2025. The review comparatively examines how multimodal fusion mechanisms, synchronization reliability, linguistic diversity, and dataset environments influence perceptual enhancement quality, intelligibility, robustness, and computational efficiency across modern AVSE systems. In this context, the analysis investigates the relationship between architectural complexity and deployment feasibility using commonly reported evaluation metrics, including PESQ, SDR, STOI, and WER. The review emphasizes comparative analytical synthesis rather than isolated summarization of individual studies. The analysis further identifies persistent methodological and practical limitations within current AVSE research, including benchmarking inconsistency, multilingual generalization challenges, demographic imbalance, privacy-related constraints associated with audiovisual datasets, and limited fairness-aware evaluation protocols. Collectively, these limitations continue to affect cross-study comparability, reproducibility, and real-world deployment reliability across multimodal AVSE systems. Finally, the review outlines future research directions involving low-resource multilingual AVSE, efficient edge-oriented multimodal systems, synchronization-aware learning frameworks, privacy-aware audiovisual processing, and computationally efficient generative architectures for next-generation AVSE deployment.

2.4 Link between questions and methodology

In order to achieve the goals of this review, the methodological strategy for answering the research questions was explicitly linked to a corresponding analytical procedure to be applied within the framework of a systematic review. Thus, rather than the single research questions, into the methodology of a comprehensive analytical review. In particular, the focus, than analyzing each individual study in a more or less in-depth manner and in isolation from one another, was placed on comparative analytical synthesis. The architectural development of AVSE systems was analyzed using comparative approaches, traditional CNN-RNN hybrids, as well as smaller, more efficient multimodal architectures. The influence of several aspects on the robustness of an AVSE system and its generalization performance across different application scenarios, i.e., datasets, synchronization of recorded audio and video signals, multilingual AVSE, and environmental variability, was analyzed in a more detailed, dataset-oriented manner. Furthermore, the study addresses the evaluation of multimodal AVSE systems, in particular perceptual assessment, as well as issues of fairness. For these aspects, an analysis of existing evaluation methodologies as well as current benchmarking, of perceptual assessment, of potential unfairness, of privacy issues, as well as of existing and future approaches to reproducibility was performed. A major focus was placed on the analysis of several commonly used evaluation metrics, including PESQ, SDR, STOI, as well as WER. The main result of this review-i.e., a comprehensive architectural analysis, along with several findings obtained within the scope of the study, and the limitations of the applied methodology for evaluation, as well as related implications for potential future deployments of AVSE systems, was finally synthesized into a set of future research directions that can be followed by researchers with the aim of developing next-generation AVSE architectures.

3. METHODOLOGY

This systematic analytical review was conducted in accordance with PRISMA 2020 guidelines to ensure methodological transparency, reproducibility, and analytical rigor throughout the study selection and synthesis process [10]. The review protocol was designed using the PICOS framework to support structured study identification, comparative evaluation, and consistent analytical synthesis across modern AVSE literature. Figure 2 presents the conceptual taxonomy adopted in this review, illustrating the relationship between audio-based, visual-based, and multimodal fusion approaches within contemporary AVSE systems.

3.1 Protocol framework and Preferred Reporting Items for Systematic Reviews and Meta-Analyses compliance

The research follows PRISMA 2020 guidelines to achieve methodological rigor and maintain reproducibility and transparency in the study [11]. The review protocol was developed according to PICOS criteria [12]:

Population (P): Audio-visual speech systems integrating acoustic and visual modalities.

Intervention (I): Deep learning-based architectures and fusion mechanisms (CNN-RNN, Transformer, Diffusion, SSL, Lightweight).

Comparison (C): Performance variations across model families and datasets.

Outcomes (O): Speech recognition accuracy (WER), perceptual quality (PESQ, STOI, SDR), and computational efficiency.

Study design (S): Peer-reviewed journal and conference papers (2020–2025).

The researchers performed a pre-study evaluation of the protocol to confirm its data inclusion capabilities and result consistency, completing documentation through the PRISMA checklist items [13].

3.2 Information sources and search strategy

The literature search was conducted across five major scientific databases: IEEE Xplore, Scopus, ACM Digital Library, SpringerLink, and ScienceDirect [11]. Additional studies were identified through backward and forward citation tracking of eligible papers to reduce database selection bias and improve coverage completeness [12]. The search period covered publications from January 2020 to June 2025 and included only English-language peer-reviewed journal and conference papers. The Boolean search query was formulated to capture modern audiovisual speech enhancement and multimodal learning studies involving transformer-based, diffusion-based, and self-supervised architectures:

("audio-visual" OR "audiovisual" OR "AV") AND ("speech enhancement" OR "speech separation" OR "speech recognition") AND (fusion OR "cross-modal" OR transformer OR diffusion OR "self-supervised") AND (PESQ OR STOI OR SDR OR WER).

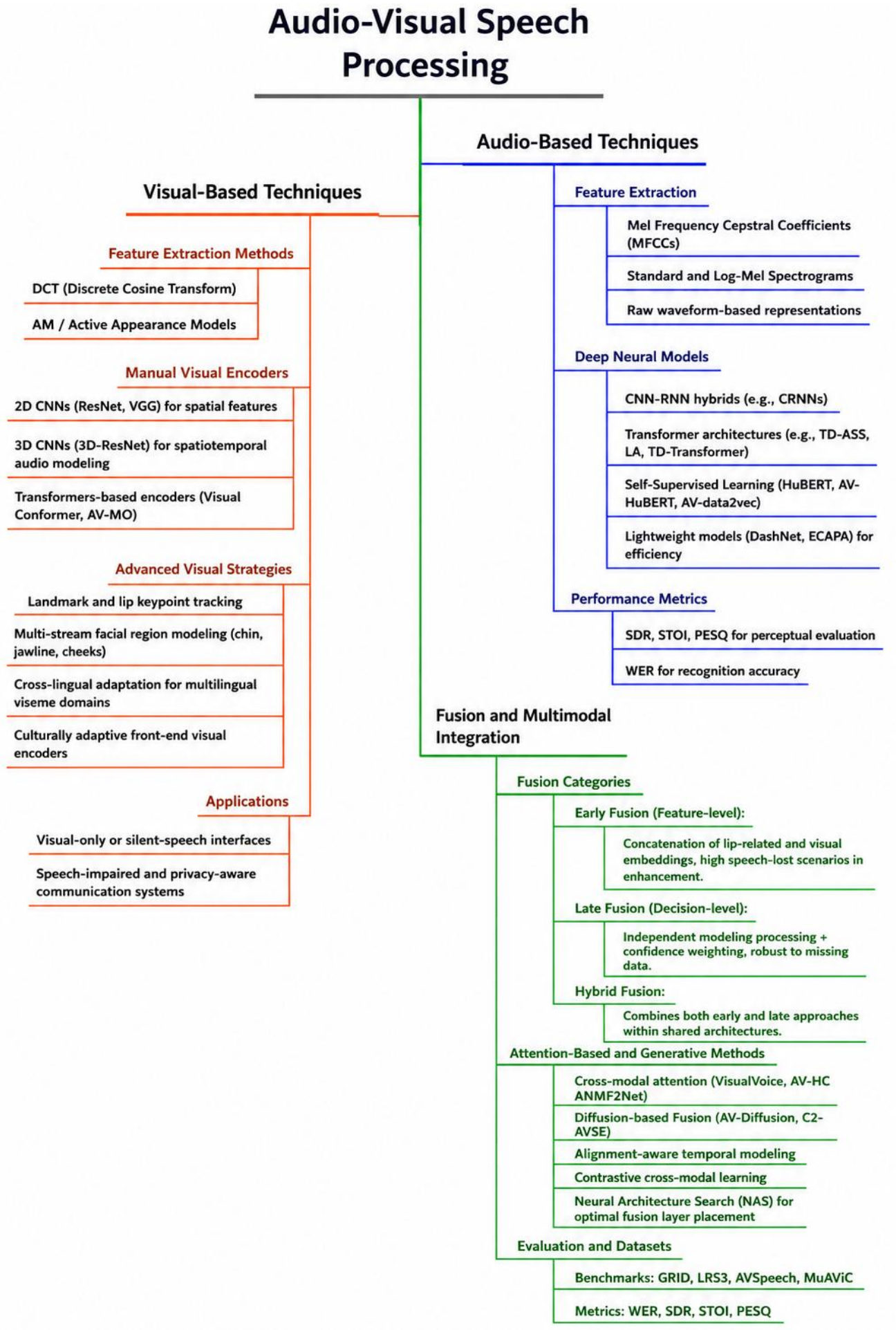

Figure 2. Conceptual taxonomy of modern audiovisual speech enhancement systems

To reduce retrieval bias, database searches were independently executed by two reviewers using identical search configurations [13]. All retrieved records were exported in BibTeX format and processed using Mendeley for duplicate removal, metadata normalization, and citation management [14].

3.3 Inclusion and exclusion criteria

Study eligibility was determined using predefined inclusion and exclusion criteria designed to ensure methodological consistency and direct relevance to the AVSE task domain. Included studies were required to investigate audiovisual speech enhancement or closely related audiovisual speech processing tasks involving multimodal acoustic-visual learning frameworks. Only peer-reviewed journal and conference publications published between 2020 and 2025 were considered eligible. Studies were excluded if they focused exclusively on unrelated audiovisual tasks such as emotion recognition, avatar synthesis, gesture analysis, speaker animation, or non-speech multimodal applications. Preprints, workshop papers, theses, and unpublished reports were also excluded to maintain peer-review quality standards. Furthermore, studies lacking quantitative evaluation metrics, reproducible experimental descriptions, dataset transparency, or sufficiently detailed methodological reporting were excluded from the analytical synthesis stage. Table 1 summarizes the inclusion and exclusion criteria adopted in this review.

Table 1. Inclusion and exclusion criteria used for study selection

Criterion	Inclusion	Exclusion
Task Domain	Audio-Visual Speech recognition, separation, or enhancement	Non-AV tasks (e.g., emotion detection, avatar synthesis)
Publication Type	Peer-reviewed journal and conference papers	Workshops, preprints, theses, or unpublished reports
Language	English	Non-English papers
Year Range	2020–2025	Before 2020
Metrics Reported	WER, PESQ, STOI, SDR, MOS, AV-MOS	No quantitative evaluation metrics
Dataset	GRID, LRS2, LRS3, AVSpeech, MuAViC, VoxCeleb2-AV	Synthetic or non-speech datasets

Note: PESQ = Perceptual Evaluation of Speech Quality, STOI = Short-Time Objective Intelligibility, SDR = Signal-to-Distortion Ratio, WER = Word Error Rate.

3.4 Screening process and Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow

To strengthen methodological rigor, all eligible studies underwent quality assessment prior to inclusion in the final analytical synthesis. Each study was evaluated according to four primary criteria: clarity of experimental methodology, dataset transparency, completeness of evaluation metrics, and reproducibility of reported results. Additional consideration was given to benchmarking consistency, reporting transparency, and the adequacy of comparative evaluation procedures. Studies with insufficient methodological description, incomplete evaluation procedures, unclear experimental reporting, or weak reproducibility characteristics were excluded from the final synthesis stage.

The overall study identification, screening, eligibility assessment, and final inclusion workflow adopted in this review is summarized in the PRISMA 2020 flow diagram presented in Figure 3. The diagram illustrates the sequential filtering process applied throughout database retrieval, duplicate removal, title and abstract screening, full-text eligibility assessment, and final study inclusion stages.

The screening and quality assessment process was independently conducted by two reviewers. Inter-reviewer agreement was evaluated using Cohen's kappa coefficient to reduce subjective selection bias and improve screening reliability. The calculated Cohen's kappa coefficient achieved a value of 0.86, indicating strong agreement consistency between reviewers during study selection and eligibility assessment. Disagreements regarding study eligibility, methodological categorization, or analytical classification were resolved through consensus-based discussion until full agreement was achieved. To reduce publication bias, the analytical synthesis did not rely exclusively on highly cited benchmark studies or single-dataset evaluations. Instead, comparative analysis was performed across multiple datasets, architectural paradigms, evaluation protocols, and deployment scenarios to improve analytical balance and reduce overrepresentation of dominant model families within the reviewed literature. In addition, studies reporting computational limitations, synchronization instability, deployment constraints, or negative findings were included whenever sufficient methodological detail was available.

3.5 Overview of included studies and distribution

In 2025 were found via a PRISMA guided search of relevant databases. The distribution of AVSE research over time clearly shows a rapid increase in the number of publications between 2023 and 2025. Much of the current AVSE research is found in IEEE and Elsevier journals, but also in Springer and ACM journals, and in a variety of open-access venues, including some interdisciplinary journals. The reviewed studies in particular stem from the multimedia signal processing, intelligent communication systems, and multimodal machine learning communities. The investigated studies were categorized in terms of their architecture into a number of groups. The greatest number of studies make use of transformer-oriented architectures, followed by CNN-RNN hybrid models, and then diffusion-based architectures for multimodal modeling. Furthermore, studies that focus on SSL approaches, as well as a number of studies that use lightweight multimodal models for AVSE were found. Although the current trend is focused on increasing the representation learning capabilities of AVSE models by means of long-range modeling and cross-modal attention, the majority of current models are based on CNN–RNN hybrid recurrent networks. The reasons for this are the high computational complexity, large memory requirements, long processing times, and high costs associated with the deployment of current models based on transformer and diffusion models. In contrast, the investigated lightweight models that are suitable for real-time processing and also exhibit good computational efficiency generally do not achieve the same level of robustness as the other models for speech enhancement in severely degraded acoustic environments, as shown in Figure 4.

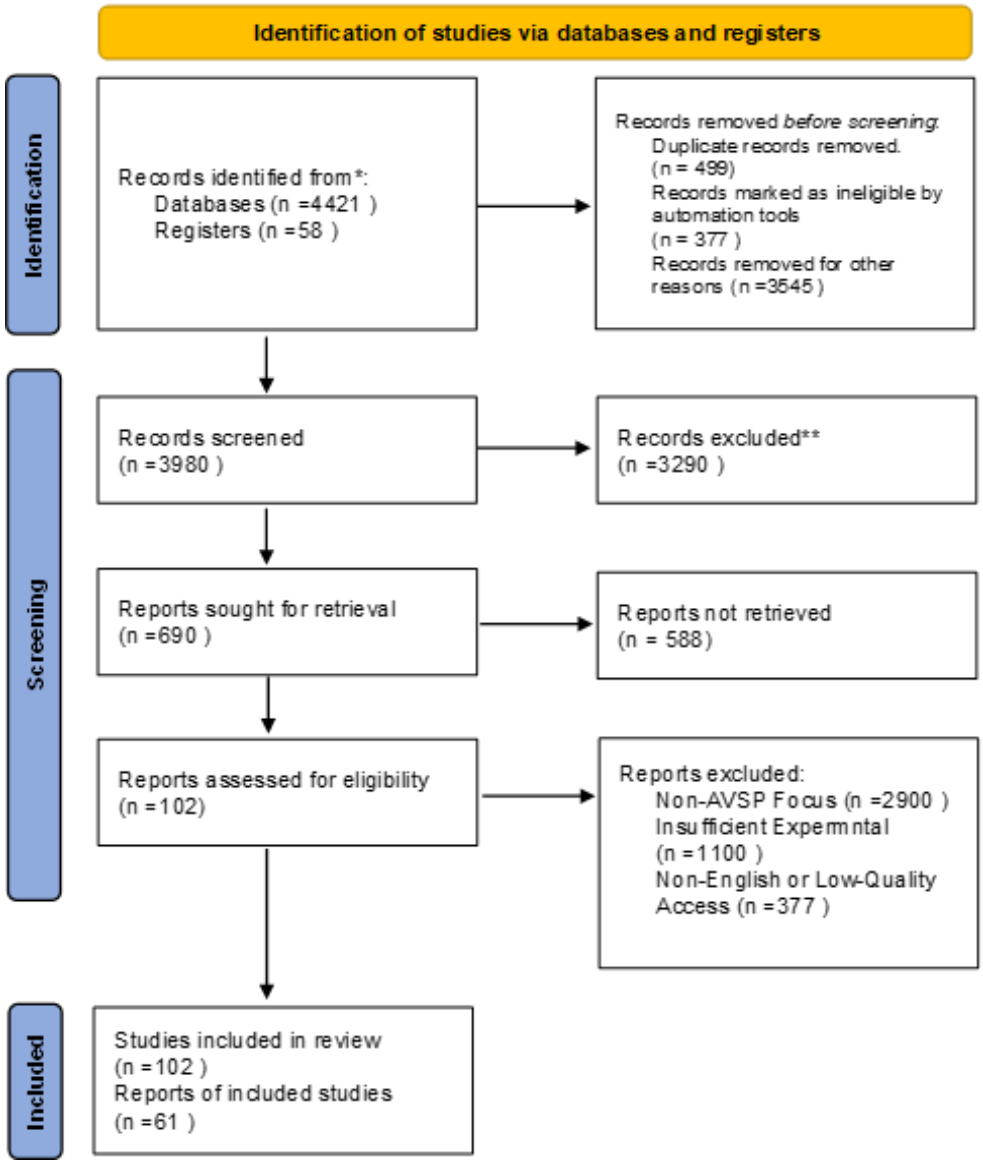

Figure 3. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 flow diagram illustrating the study identification, screening, and inclusion process for the Audio-Visual Speech Processing (AVSP) systematic review

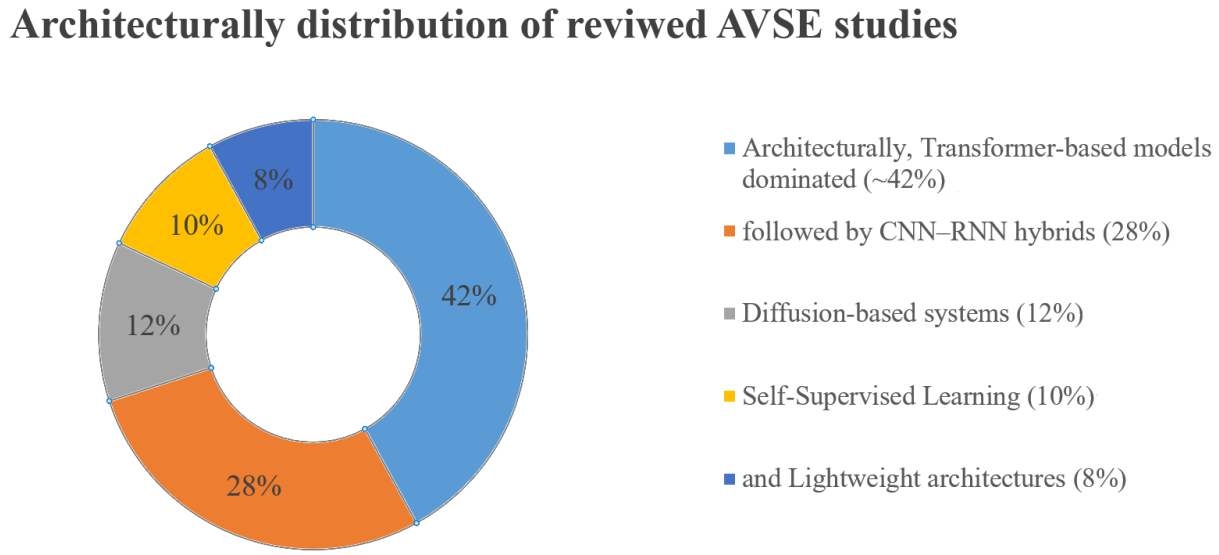

Figure 4. Architectural distribution of the reviewed Audio-Visual Speech Processing (AVSP) models (2020–2025)

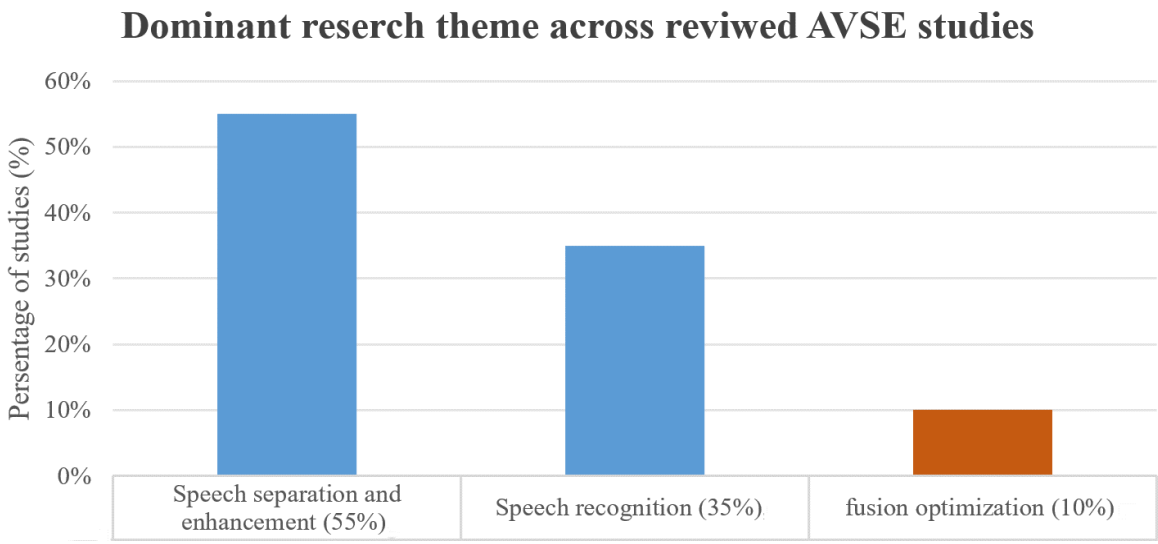

Figure 5. Thematic distribution of research gaps identified across Audio-Visual Speech Processing (AVSP) studies (2020–2025)

To ensure methodological consistency during thematic analysis, each reviewed study was assigned to a single dominant research category according to its primary experimental objective. The thematic distribution analysis revealed that approximately 55% of the reviewed studies focused primarily on speech enhancement and speech separation tasks, emphasizing perceptual restoration, intelligibility improvement, and multimodal robustness under noisy acoustic conditions. Approximately 35% of the studies concentrated on recognition-oriented audiovisual speech processing tasks, while nearly 10% investigated multimodal fusion optimization, synchronization reliability, adaptive weighting mechanisms, or deployment-oriented integration strategies. As shown in Figure 5, current AVSE research remains heavily centered on signal restoration and recognition performance, whereas relatively limited attention has been given to demographic fairness, multilingual inclusivity, privacy-aware audiovisual learning, ethical evaluation protocols, and cross-population robustness analysis.

These limitations remain particularly significant because demographic imbalance and restricted multilingual representation within audiovisual datasets may reduce the generalization capability of AVSE systems across diverse real-world deployment environments. In addition, privacy constraints associated with facial audiovisual datasets continue to limit the availability of large-scale ethically balanced multimodal corpora. Overall, the reviewed literature demonstrates a clear evolution from conventional handcrafted and recurrent modeling approaches toward large-scale multimodal representation learning systems that emphasize contextual modeling and perceptual reconstruction quality. Nevertheless, the comparative analysis also reveals persistent challenges involving benchmarking inconsistency, computational scalability, synchronization robustness, multilingual generalization, ethical evaluation, and deployment-oriented feasibility across modern AVSE architectures. These findings reinforce the importance of balancing the quality, robustness, computational efficiency, and real-world deployment constraints within next-generation audiovisual speech enhancement systems.

3.6 Data synthesis and analysis methods

The analytical synthesis process combined qualitative and quantitative evidence across the included studies to comparatively evaluate architectural behavior, multimodal fusion strategies, dataset diversity, evaluation methodologies, and deployment-oriented performance characteristics within modern AVSE systems [15]. Rather than relying exclusively on sequential performance reporting, the synthesis framework emphasized comparative interpretation of performance–efficiency trade-offs across different architectural families and multimodal learning paradigms.

To support structured comparative analysis, the reviewed studies were categorized into three primary analytical groups according to their dominant methodological orientation. The first category included audio-oriented architectures relying primarily on spectral and waveform-based enhancement approaches, including CNN- and RNN-based systems [16]. The second category involved visual-oriented frameworks utilizing lip-reading, facial articulation, and visual speech representation learning mechanisms [17]. The third category included fusion-oriented multimodal systems integrating acoustic and visual information through early-fusion, late-fusion, hybrid-fusion, attention-based, transformer-based, diffusion-based, and self-supervised multimodal learning strategies [18].

Subgroup analyses were further conducted to evaluate the influence of dataset environments (controlled versus in-the-wild conditions), multilingual variability, architectural complexity, synchronization robustness, and deployment-oriented constraints on model generalization capability and perceptual enhancement performance [19]. Additional comparative analysis examined the relationship between computational scalability, contextual modeling capability, and real-time feasibility across modern AVSE architectures.

To improve cross-study interpretability, the synthesis considered four commonly reported evaluation metrics in AVSE and speech enhancement literature: PESQ, SDR, STOI, and WER [14]. PESQ was used to evaluate perceptual speech quality and reconstruction fidelity, while SDR measured signal reconstruction accuracy relative to clean reference speech. STOI was considered for intelligibility evaluation under degraded acoustic conditions, whereas WER was primarily analyzed in recognition-oriented audiovisual speech processing tasks involving multimodal speech understanding [20]. Because the reviewed studies employed heterogeneous datasets, evaluation protocols, training configurations, and

experimental conditions, direct numerical comparison across all studies was not always methodologically reliable. Therefore, the synthesis focused primarily on comparative performance trends and relative architectural behavior rather than absolute metric ranking or unified metric normalization [21]. Quantitative findings were interpreted according to their conventional reporting contexts within the AVSE literature to preserve methodological consistency and avoid analytical distortion arising from incompatible scaling assumptions between perceptual, intelligibility, distortion-based, and recognition-oriented evaluation metrics. Figure 6 summarizes the comparative analytical trends observed across major AVSE architectural families with respect to perceptual enhancement quality, contextual modeling capability, computational complexity, synchronization robustness, and deployment-oriented feasibility.

Figure 6. Comparative performance of major Audio-Visual Speech Processing (AVSP) architectures across perceptual and computational metrics

The comparative synthesis indicates that transformer-based and diffusion-based architectures generally achieve improved perceptual enhancement quality and multimodal contextual modeling performance relative to conventional CNN-RNN systems, particularly under highly degraded acoustic conditions [22]. These improvements are primarily associated with enhanced long-range dependency modeling, attention-driven fusion mechanisms, and more effective multimodal representation learning capabilities. However, the increased adoption of transformer-based and diffusion-based architectures is also associated with substantially higher computational complexity, memory consumption, inference latency, and deployment cost, particularly in resource-constrained or real-time operational environments [23].

The other hand, lightweight multimodal architectures for efficient processing have been investigated and compared to larger alternatives. On perceptual processing, better audio-visual synchronization and better scalability are more important. Most approaches that have the potential for efficient processing, low memory demand, and real-time capability achieve higher quality in multimodal input. In contrast, approaches that employ large transformer architectures and diffusion models for unsupervised efficient processing show lower robustness and poor perceptual quality, especially under severe acoustic degradation and larger variations in large multilingual and unconstrained datasets rather than in a laboratory setup. However, the current supervised audio-visual representation learning often achieves better robustness and generalization across domains if trained with balanced datasets, limited multilingual capability, and a lack of demographic diversity. Moreover, very few unsupervised approaches suffer from several restrictions that impede a fair comparison, for example inconsistent benchmarks, and omissions in the development of fairness-oriented benchmarks. Such issues, however, are becoming more important as the AVSE approaches are increasingly deployed in large-scale real-world applications.

Recent, generative-based audio-visual reconstruction techniques However, numerous challenges to current AVSE methodologies and AVSE advancements, are largely rooted in multimodal contextual learning, self-supervised audio-visual representation learning, and potential large-scale multimodal AVSE deployments, remain to be addressed. The most pressing challenges and SE in real-world scenarios.

4. DATASET TRENDS AND CROSS-DATASET GENERALIZATION

Datasets constitute a central methodological factor in AVSE because they directly influence model robustness, benchmarking reliability, cross-domain generalization, and fairness-oriented evaluation [24, 25]. In audiovisual learning, dataset quality is not determined only by size, but also by synchronization accuracy, acoustic variability, speaker diversity, linguistic coverage, annotation reliability, and the extent to which the corpus reflects real deployment conditions [26]. Therefore, the evolution of AVSE datasets should be interpreted not merely as a chronological expansion of available corpora, but as a transition from controlled evaluation settings toward heterogeneous, multilingual, and deployment-oriented audiovisual benchmarks [27, 28].

4.1 Historical and empirical evolution

There have been a number of early audio-visual speech databases collected in the laboratory, for example the GRID database, which has a small number of speakers, a fixed vocabulary, and a number of synchronization experiments, all conducted in a controlled acoustic environment [15]. They have been very useful for establishing a number of audio-visual speech perception models and for testing a number of early CNN-RNN-based audio-visual speech recognition systems. However, for a number of reasons, they are not very suitable for testing a system's ability to generalize to a number of different situations, including speech in noisy environments, with large variations in speakers' voices and speech, and with a very large vocabulary of speech.

LRS2 [1] and LRS3 [2] datasets are other important sources of English speech in audiovisual format, drawn from a wide variety of broadcast TV sources, and even from TED talks; they comprise much larger numbers of utterances, making them more amenable for the development of models able to exploit such large amounts of data. In addition to being a larger repository of training data, it also allows for more open-vocabulary architectures, such as transformers [29] and SSL methods [30] to be evaluated and to be fine-tuned using large repositories of high-quality audiovisual content. Nonetheless, all these sources of English speech in audiovisual format are not able to support any multilingual and cross-cultural experiments.

Large order to study more realistic conditions of audio-visual learning [31, 32]. The introduced datasets include a large number of speakers and many recordings with different conditions, which make them suitable for training multimodal models. Large in-the-wild datasets make the learning process more ecologically valid; however, they also pose a number of difficulties related to, for example, synchronization drift, noisy labels, non-uniform recording conditions, and limited control over audio- and visual quality. As a result, in order to perform a reproducible benchmark, large-scale in-the-wild datasets introduce a number of challenges that need to be overcome [33].

More recent resources such as MuAViC, AVSUPERB, and BIN-AVSS further reflect the changing priorities of the field [8, 16, 34]. MuAViC introduced broader multilingual coverage and encouraged evaluation beyond English-dominant audiovisual corpora [35], while AVSUPERB emphasized standardized benchmarking protocols rather than simply increasing dataset size [36]. BIN-AVSS, in contrast, addressed spatial and real-time audiovisual separation scenarios, which are increasingly relevant for deployment-oriented AVSE systems [37]. These developments indicate that dataset design is moving from data volume alone toward more structured evaluation of multilingual robustness, spatial awareness, low-latency operation, and benchmarking consistency [38]. Table 2 summarizes the major benchmark audiovisual speech datasets that have shaped AVSE and related audiovisual speech processing research between 2006 and 2025. The comparison highlights how each dataset contributes differently to controlled evaluation, large-scale learning, multilingual generalization, spatial separation, or standardized benchmarking.

Table 2. Benchmark audio-visual speech datasets (2006–2025) and their key characteristics

Dataset	Year	Duration (h)	Speakers	Languages	Typical Use/Model Type	Key Features	Limitations	Benchmark Relevance/ Contribution
GRID	2006	33	34	English	CNN-RNN	Fixed vocabulary; lab-controlled speech	Only 51 words; limited scalability	Baseline for controlled AVSR systems
LRS2	2017	224	>1,000	English	Transformer	Broadcast TV; open vocabulary	Moderate scale; background noise	Enables large-scale transformer AVSR
LRS3	2018	439	>2,000	English	Transformer, SSL	TED/TEDx talks; large lexicon	Conversational realism; Englishonly	Foundation for cross-domain generalization
VoxCeleb2-AV	2019	>2,000	>6,000	English + Multilingual	CNN–RNN, SSL	Natural YouTube videos; spontaneous speech	Synchronization errors; label noise	Open-domain multilingual learning
AVSpeech	2019	>4,700	>150,000	Multilingual	Transformer, Diffusion	Very large scale; diverse contexts	Weak alignment; label noise	Foundation-scale training corpus
MuAViC	2023	3,000	>20 langs	Multilingual	Transformer	First large multilingual AV dataset	Uneven language balance	Multilingual fairness benchmark
AVSUPERB	2023	N/A	N/A	Multi	Benchmarking	Unified evaluation protocol	No new data; standardization only	Standardized evaluation benchmark
BIN-AVSS	2023	N/A	N/A	English	Real-Time, Spatial	Binaural + spatial separation	Early-stage; limited use	Real-time low-latency testing framework

Note: N/A = Not Available in the original publication. It indicates that the authors did not report explicit duration or speaker counts.
CNN = Convolutional Neural Network, RNN = Recurrent Neural Network.

4.2 Cross-dataset comparative insights

The evolution of the used datasets for AVSE, summarized in Table 2, leads to three major findings. Firstly, in contrast to early laboratory datasets, nowadays mainly in-the-wild, uncontrolled datasets of large scale are used. Such datasets render more realistic scenarios for the models to learn from under different acoustic and also different visual conditions. As a consequence, however, the used datasets lead to greater variability in the used annotations, in the alignment quality, and in the reproducibility of experimental results. Thus, high performance on a controlled, lab-based dataset such as GRID (20 speakers) does not by far not necessarily guarantee high performance on the large-scale, in-the-wild datasets such as AVSpeech (5,000+ speakers) or even on the AV part of the VoxCeleb2-dataset (with 12,000+ speakers in total) [39, 40].

Multilingual and cross-domain datasets, such as MuAViC [8] have recently started to be used to test the generalization ability of AVSE models on newly collected audio-visual data, in newly spoken languages, with speakers of different demographic backgrounds, or recorded under a variety of new conditions. As English-dominant AVSE datasets (e.g., GRID [27], AVSpeech [5], VoxCeleb2-AV [28]) have shown high performance levels in controlled experiments, so using them for benchmarking AVSE models is valuable. However, AVSE models that have been trained on such English-dominant datasets for AVSE tasks are only partially able to generalize to newly collected, in-the-wild audio-visual data in other languages, in other languages and with speakers of very different demographic backgrounds. For this purpose, a multilingual AVSE corpus, MuAViC, has recently been proposed in [39]. Yet, as is the case for any AVSE corpus, the quality of the collected audio-visual synchronization of the data, the reliability of all its annotations, and the experimental protocols used for testing its AVSE models can be inconsistent and imperfect, and, therefore, can introduce several confounding factors and result in a number of potential artifacts (e.g., language imbalance, speaker imbalance, unbalanced conditions across speakers, speaker-condition imbalanced data, etc.), which can negatively affect both the already discussed fairness of the experimental comparison of AVSE models, as well as their cross-population robustness [40].

Third, there is a trade-off between allowing models to be trained and tested on data that allows for sufficient representation of speaker diversity, different acoustic and visual conditions, etc. on the one hand, and having sufficient control over the quality and nature of the annotations on the other hand. A large amount of data allows for sufficient representation of most types of speaker variability, and the model becomes sufficiently robust and can even be used for effective pretraining of a smaller subsequent model on a different, but related task [41, 42]. However, very large amounts of data also pose many of their own problems, including the presence of many weak or incorrect labels, significant amounts of data with large synchronization errors, large amounts of noisy metadata, and the fact that recording conditions can vary widely. This makes it difficult to reproduce results and to know whether performance was degraded by specific issues in the data, or whether performance was simply never very strong to begin with. Conversely, smaller amounts of data have cleaner, more accurate annotations and allow for more consistent evaluation, but then there is the risk of overestimating performance on a task that is frequently subject to a large degree of acoustic and visual variability in real-world situations [43].

To address this challenge, it is necessary to evaluate the performance of models trained or optimized on one dataset across multiple AVSE benchmarks rather than relying on a single-dataset evaluation. Cross-dataset generalization assessment should involve datasets that are as diverse as possible with respect to several factors, including (1) acoustic conditions, (2) language and/or accent variability, (3) speaker demographics, (4) recording environments, and (5) synchronization quality. This is particularly important for transformer-based, diffusion-based, and self-supervised AVSE models, which are typically trained on large-scale datasets and may inherit biases, synchronization inconsistencies, or dataset-specific characteristics from the training data.

4.3 Fairness, reproducibility, and dataset-driven limitations

Dataset diversity has direct implications for fairness and reproducibility in AVSE systems [44]. Demographic imbalance in speaker identity, gender, age, accent, language, and visual appearance may reduce the reliability of audiovisual models when deployed across diverse user populations [45]. Similarly, limited multilingual coverage can bias model performance toward high-resource languages and restrict generalization in low-resource speech communities [46]. These issues are especially important in AVSE because visual speech information involves facial data, which raises additional privacy and consent concerns compared with audio-only speech datasets [47, 48].

Reproducibility is also affected by inconsistent reporting of dataset preprocessing, train-test splits, synchronization procedures, and evaluation protocols [49]. Some studies report strong results without fully specifying alignment correction, visual preprocessing, speaker overlap control, or dataset partitioning strategy. Such inconsistencies make it difficult to determine whether performance gains are caused by architectural improvements, dataset advantages, preprocessing differences, or the evaluation protocol [35, 50]. Therefore, dataset transparency should be treated as a methodological requirement rather than a secondary implementation detail.

From a systems-oriented perspective, dataset selection shapes the type of AVSE system that can be fairly evaluated. Controlled datasets are suitable for isolating architectural behavior; large, in-the-wild datasets are useful for robustness and pretraining; multilingual datasets support fairness and inclusivity analysis; and spatial or real-time datasets are necessary for deployment-oriented evaluation [37]. A mature AVSE benchmark ecosystem therefore requires not one universal dataset, but complementary datasets that collectively evaluate perceptual quality, intelligibility, synchronization robustness, computational feasibility, and cross-population reliability.

4.4 Synthesis and forward connection

The audiovisual speech datasets have progressed from controlled, monolingual, task-specific corpora to large multilingual datasets for deployment of various AVSE tasks. This evolution has enabled the development of more powerful architectures for processing multimodal information; however, the current state of the art also brings a host of newly

introduced challenges to the field, such as annotation noise, synchronization drift, language imbalance, demographic bias, privacy, and reproducibility of results [38-40]. Thus, the quality of the current datasets must not only be assessed in terms of their size, but also in terms of methodological transparency, in terms of the degree to which they are fair, in terms of the quality of their evaluation, and in terms of their applicability in real-world scenarios. Figure 7 depicts the degree of performance of several Audio-Visual Speech Processing (AVSP) models that represent the current state-of-the-art in terms of objective measures such as SDR, PESQ, STOI, and WER. It can be clearly observed that the effectiveness of the architectures strongly depends on the characteristics of the datasets used for evaluation, the stability of the preprocessing steps, and the quality of the multimodal synchronization. Thus, in the following, we shall take a closer look at the various architectures that have been recently proposed for AVSE, such as transformer-based models, diffusion-based models, self-supervised models, and lightweight architectures, and examine how they fare in terms of the aforementioned challenges.

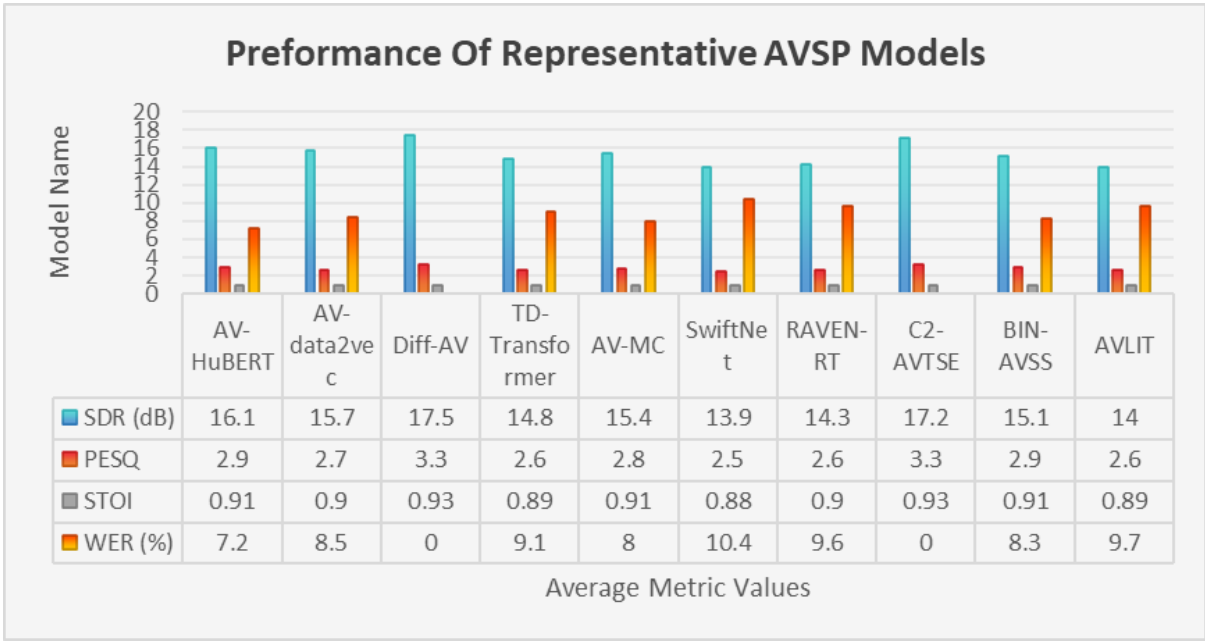

Figure 7. Comparative performance of representative AVSP models (2022–2025) across SDR, PESQ, STOI, and WER metrics
Note: AVSP = Audio-Visual Speech Processing, PESQ = Perceptual Evaluation of Speech Quality, STOI = Short-Time Objective Intelligibility, SDR = Signal-to-Distortion Ratio, WER = Word Error Rate.

Table 3. Core objective and perceptual evaluation metrics used in AVSP research

Metric	Full Meaning	Typical Range (Audio → AV)	Interpretation and Role
WER (%)	Word Error Rate	7–30	Lower values indicate more accurate recognition and effective audiovisual fusion.
PESQ	Perceptual Evaluation of Speech Quality	2.20–3.30	Higher scores reflect clearer, more natural-sounding speech under multimodal enhancement.
STOI	Short-Time Objective Intelligibility	0.85–0.95	Higher values represent better speech intelligibility, especially under noise.
SDR (dB)	Signal-to-Distortion Ratio	11–17	Larger SDR values imply cleaner reconstruction and reduced residual noise.
AV-MOS	Audio-Visual Mean Opinion Score	3.80–4.40	Higher human ratings indicate improved audiovisual coherence and perceptual realism.

Note: Among these metrics, SDR is measured in decibels because it represents a physical energy ratio, whereas PESQ, STOI, and AV-MOS are unitless perceptual indices derived from psychoacoustic or correlation-based evaluation models.
AVSP = Audio-Visual Speech Processing, PESQ = Perceptual Evaluation of Speech Quality, STOI = Short-Time Objective Intelligibility, SDR = Signal-to-Distortion Ratio, WER = Word Error Rate, CNN = Convolutional Neural Network, RNN = Recurrent Neural Network.

4.5 Evaluation metrics and benchmark performance

Evaluation methodologies in modern AVSE research increasingly rely on complementary objective and perceptual assessment metrics to evaluate reconstruction fidelity, intelligibility preservation, multimodal consistency, and recognition-oriented robustness under acoustically degraded conditions [22, 29]. Because AVSE systems integrate both acoustic and visual modalities, evaluation cannot depend exclusively on recognition accuracy or signal reconstruction quality alone. Instead, contemporary AVSE benchmarking frameworks typically combine distortion-based, intelligibility-oriented, perceptual, and recognition-related metrics to capture multiple aspects of multimodal enhancement behavior [51, 52]. Table 3 summarizes the most commonly reported evaluation metrics used throughout recent AVSE literature, along with their representative performance ranges and methodological interpretation within multimodal speech enhancement tasks.

The reviewed literature consistently demonstrates that multimodal AVSE architectures generally outperform audio-only baselines across perceptual quality, intelligibility

preservation, and recognition-oriented evaluation metrics, particularly under noisy and visually degraded conditions [35, 43]. Representative studies report substantial reductions in WER following multimodal fusion integration, along with improved PESQ, SDR, and STOI performance under adverse acoustic environments [12, 31, 44, 53]. These findings suggest that visual speech information contributes not only to recognition-oriented robustness, but also to improved perceptual reconstruction stability and multimodal contextual consistency.

However, due to differences in the datasets, preprocessing, synchronization, evaluation, and acoustic conditions between the AVSE studies, a quantitative comparison is not straightforward [54]. The metrics should be used to compare the behavior of the architectures in relative terms, rather than making direct comparisons between the studies in absolute terms. This problem is particularly pronounced when comparing transformer-based, diffusion-based, self-supervised, and lightweight architectures for AVSE, and most of the studies used different datasets for evaluation. Moreover, a single metric is not sufficient to evaluate the performance of an AVSE system. While distortion metrics, such as the SDR, measure the amount of distortion in the reconstructed audio, the perceptual metrics, such as the PESQ and the audiovisual-rated quality of the transmitted speech (AV-MOS), measure the overall quality of the audio-visual perception. Another important aspect is the intelligibility of the speech, and this is typically measured using the STOI metric. Finally, the WER is used as a metric for the recognition performance in multimodal situations. In summary, the evaluation of AVSE systems requires the use of a set of metrics and a framework that allows for the evaluation of the different aspects of an AVSE system, such as perceptual quality, intelligibility, synchronization, and deployability. In the following, the different AVSE architectures are evaluated using such a framework, and the results are discussed in terms of the perceptual quality, the computational requirements, the modeling of the context, and deployability.

4.6 Unified empirical synthesis of architectural families (2020–2025)

For the purpose of comparative analytical synthesis, reviewed studies on AVSE were grouped into five families of architectures that can be considered AVSE architectural families, namely: 1) CNN-RNN hybrid architectures for AVSE, 2) transformer-based architectures for AVSE, 3) diffusion-based architectures for AVSE, 4) self-supervised multimodal models for AVSE, and 5) lightweight deployment-oriented architectures for AVSE [14, 47, 48]. It has been shown through comparative analytical synthesis that the perceptual enhancement quality and the ability for multimodal contextual modeling are better for transformer-based architectures for AVSE and diffusion-based architectures for AVSE than for CNN-RNN hybrid architectures for AVSE. This is mainly due to the long-range dependency modeling ability, the attention-based fusion ability, and the ability to better integrate acoustic information and visual information, etc. [49]. In addition, transformer-based architectures for AVSE and many diffusion-based architectures for AVSE reported lower WER values and better PESQ, SDR, and STOI values than those for CNN-RNN hybrid architectures for AVSE in many studies. They achieved these better performance metrics in noisy environments and in visually occluded environments. However, the computational complexities, the memory requirements, the training costs, and the latency for these architectures are generally much higher than those for performance does not necessarily mean better deployment feasibility; the stronger perceptual reconstruction performance can be achieved using AVSE architectures with higher computational complexities, higher memory requirements, higher training costs, and higher latency for real-time applications, embedded systems, edge computing systems, and other AVSE systems that require low resources for operation.

However, several lightweight AVSE architectures have the benefit of low computational requirements, low memory needs, and good real-time capabilities, which make them suitable for many real-time AVSE applications. In terms of robustness, however, most lightweight architectures show decreased robustness against severe acoustic degradation and large variations in multimodal data as compared to larger architectures such as many transformer-based and diffusion-based frameworks.

Self-supervised multimodal learning approaches emerged as an intermediate architectural direction, balancing contextual representation learning with reduced dependency on fully labeled audiovisual datasets [39, 50]. These systems demonstrated improved adaptability across heterogeneous datasets and multilingual environments while reducing annotation requirements relative to fully supervised architectures. Nevertheless, their performance remained strongly dependent on pretraining quality, dataset diversity, and synchronization stability across modalities.

Overall, the comparative synthesis reveals that modern AVSE research increasingly reflects a performance-efficiency trade-off involving perceptual enhancement quality, contextual modeling capability, computational scalability, synchronization robustness, and deployment-oriented feasibility. Transformer-based and diffusion-based systems generally provide stronger multimodal contextual reconstruction performance, whereas lightweight architectures remain more suitable for real-time and resource-constrained deployment scenarios. These findings reinforce the importance of balancing perceptual quality, computational efficiency, scalability, and operational feasibility within next-generation audiovisual speech enhancement systems.

5. DISCUSSION

The reviewed literature demonstrates that recent advances in AVSE are increasingly driven by multimodal contextual modeling, large-scale audiovisual datasets, and perceptually oriented evaluation frameworks rather than isolated improvements in recognition accuracy alone [55, 56]. Across studies published between 2020 and 2025, the field has progressively shifted from conventional CNN-RNN pipelines toward transformer-based, diffusion-based, self-supervised, and lightweight multimodal architectures emphasizing cross-modal representation learning, synchronization robustness, and deployment-oriented adaptability [53].

The comparative synthesis presented in previous sections indicates that AVSE research is evolving simultaneously along methodological, architectural, computational, and ethical dimensions. These developments collectively reflect a broader transition from task-specific audiovisual enhancement systems toward more scalable and context-aware multimodal

learning frameworks. At the same time, the reviewed literature also reveals persistent challenges involving computational scalability, benchmarking inconsistency, synchronization reliability, multilingual generalization, fairness-aware evaluation, and deployment feasibility under real-world operating conditions. Table 4 summarizes the principal convergent trends and methodological implications identified across the reviewed AVSE literature.

Table 4. Summary of convergent trends and implications in AVSP evolution

Theme	Key Findings	Implications
Architectural Evolution	Shift from CNN-RNN to transformer and diffusion models	Cross-modal attention and generative learning drive robustness.
Performance vs. Efficiency	Trade-off between high accuracy and real-time deployment	Lightweight designs are essential for scalability.
Dataset Impact	Multilingual and naturalistic datasets outperform controlled ones	Dataset diversity directly enhances generalization.
Evaluation Consistency	Only 40% of studies report all core metrics (WER, PESQ, STOI, SDR)	Need for unified reporting standards.
Ethical Awareness	80% of studies ignore demographic fairness	Necessity of bias-aware dataset and metric design.

Note: AVSP = Audio-Visual Speech Processing, PESQ = Perceptual Evaluation of Speech Quality, STOI = Short-Time Objective Intelligibility, SDR = Signal-to-Distortion Ratio, WER = Word Error Rate, CNN = Convolutional Neural Network, RNN = Recurrent Neural Network.

5.1 Evolution of learning paradigms

The majority of studies related to AVSE have moved from feature extraction for both audio and visual streams, and their subsequent combination, to the multimodal learning of acoustic and visual information in context within noisy and partially degraded audio environments. As opposed to the previously handcrafted features that were combined by using earlier AVSE systems, the features are now more often created by using CNNs and RNNs for localized temporal modeling. Most recently, a multitude of approaches has emerged that use transformers for modeling long-range dependencies, as well as diffusion models for modeling complex audio in particular. These approaches additionally include attention for the fusion of the modalities, as well as a number of other aspects related to the multimodal contextualization of synchronized audiovisual streams. Although earlier AVSE systems primarily focused on achieving the highest possible signal reconstruction quality, current approaches pursue a broader set of goals. In addition to higher perceptual quality of restored video, the focus is on improved intelligibility and robustness in a wide variety of acoustic environments. With respect to learning features in an unsupervised fashion, numerous approaches for self-supervised multimodal learning have recently been introduced. They enable efficient learning of features for both audio and visual modalities without the need for large amounts of labeled data for audiovisual pairs. Thanks to their design, which is based on multiple domains and implemented in multiple languages, these approaches are particularly suitable for a wide range of applications and can be flexibly adapted to the requirements of the specific AVSE task at hand. In addition to better contextual modeling, recent research focuses on improving the stability of advanced AVSE systems for real-world applications. While recent approaches clearly excel in terms of their learning capabilities, the requirements on computational resources, training time, and memory for recent approaches such as transformer-based and diffusion-based methods are significantly higher than those for traditional CNN-RNN-based methods. Thus, current research tries to find a suitable balance between a system's learning capabilities and its scalability, synchronization robustness, and applicability.

5.2 Performance-efficiency balance

The comparative analysis across the reviewed architectural families consistently demonstrates a trade-off between perceptual enhancement performance and computational efficiency. Transformer-based and diffusion-based AVSE systems generally achieve stronger PESQ, SDR, STOI, and WER performance under acoustically degraded conditions due to their superior contextual modeling capability and attention-driven multimodal integration [57, 58]. However, these improvements are frequently accompanied by increased computational complexity, inference latency, memory consumption, and energy requirements.

By contrast, lightweight multimodal frameworks demonstrated stronger deployment feasibility for real-time and resource-constrained environments [28, 59]. Such architectures often employ reduced parameterization, compressed multimodal representations, or computationally efficient fusion mechanisms to support embedded and edge-oriented applications. Although lightweight systems may exhibit lower perceptual reconstruction performance under highly variable acoustic conditions, they remain important for low-latency operational scenarios such as assistive communication systems, mobile devices, and hearing-support technologies [60].

The reviewed literature therefore suggests that future AVSE system design should not prioritize perceptual enhancement quality in isolation. Instead, robust multimodal architectures increasingly require balanced optimization strategies capable of jointly addressing computational scalability, synchronization reliability, energy efficiency, and deployment-oriented adaptability across heterogeneous operational environments.

5.3 Dataset influence and multilingual generalization

An important factor influencing the robustness of AVSE systems, their reproducibility, and their generalization capability to new domains and different domains is the design choice made for the dataset used for training and evaluation. A large number of audio-visual datasets have been recently published, ranging from small, controlled datasets that can be easily synchronized, such as GRID and LRS2, to very large, in-the-wild datasets, which are also used for AV speech recognition, such as AVSpeech and MuAViC.

Large datasets have their own set of challenges, and current methods for handling synchronization instability, metadata of varying quality, imbalanced datasets, noisy labels, and datasets that are not representative of all possible aspects of

multimodal human communication all pose challenges for comparative study. Additional challenges include the preprocessing steps that are taken, all of which can negatively impact the reproducibility of results, the train/test split that is not standardized, and the way in which alignment is performed and corrected.

In summary, while increasing dataset diversity is not sufficient to guarantee AVSE robustness yet, more emphasis needs to be placed on transparency of synchronization, on demographic balance, on multilingual coverage, and on standardized evaluation protocols in order to support fairness-aware multimodal evaluation across a variety of deployment scenarios.

5.4 Theoretical and practical integration

A growing trend in current AVSE research is to combine signal-based evaluation metrics with more perceptual and deployment-oriented metrics to assess the quality of AVSE systems in various applications [61-63]. Metrics that measure signal reconstruction quality, such as SDR and PESQ, provide the basic evaluation criteria for perceived speech quality. Other metrics, for example, STOI, WER, and AV-MOS, are used to assess the degree to which a system improves the intelligibility of processed audio while maintaining other important properties of synchronized audio-visual signals.

Recently, there has been a trend in AVSE research toward combining signal processing-oriented evaluation metrics with perceptual and deployment-oriented assessment criteria. While SDR and PESQ must be used for evaluating the signal reconstruction quality and perceptual speech quality, respectively, STOI, WER, and AV-MOS are used for evaluating the other aspects of AVSE systems, such as the intelligibility preservation, multimodal consistency, and recognition-oriented robustness. As shown in Table 4, the evaluation methodologies of recently developed AVSE systems are gradually shifting from isolated metric optimization to using multi-metric assessment frameworks. It is also obvious that no single metric can comprehensively and objectively represent the performance of an AVSE system in various situations. Thus, using integrated and robust evaluation approaches to simultaneously evaluate the perceptual quality, synchronization, implementation complexity, and real-time processing capability of an AVSE system is becoming a necessity for fair benchmarking.

AVSE systems are being explored for various applications in addition to the traditional laboratory-based speech enhancement. These applications include, but are not limited to, assistive communication, hearing aid support systems, telepresence, edge computing based multimodal interfaces, and low resource real time audio-visual communication [34, 60]. The mentioned limitations related to the AVSE deployment, such as latency, power consumption, privacy, and synchronization robustness, are still open issues that need to be addressed in order to make AV.

5.5 Future research and technological outlook

Overall, current AVSE models that incorporate multimodal contextual learning can be combined with self-supervised representation learning and generative audio–visual models for reconstruction. Based on recent trends in AVSE research, several new directions can be anticipated [64]. In particular, future AVSE models could be made more adaptable to multilingual users and diverse linguistic environments.

Another strand of emerging research within AVSE focuses on deployment-oriented optimization for various scenarios, exploring a wide range of architectures, diffusion methods, and audio/visual processing techniques to identify the best deployable solution. These methods include, but are not limited to, extremely lightweight architectures, efficient diffusions, adaptive audio/visual compression, and highly efficient edge-based real-time processing using AV inference systems [65].

Third, for now, evaluation protocols for fairness and reproducibility are under-explored. There is a lack of diversity in terms of demographic groups, multilingual representation is not yet consistent, and preprocessing steps as well as synchronization steps are not transparent enough to allow for reliable benchmarking. Future work should focus on creating more fair datasets for AVSE and, in addition to that, reporting results in a more standardized way, making synchronization steps transparent; and, lastly, providing the community with reproducible and viable benchmarking methods.

Finally, and more importantly, as already mentioned in Section 7, explaining and interpreting decisions made by multimodal AVSE systems is a significant open challenge that affects not only its deployment but also a range of other issues, including safety, reliability, and trustworthiness. Visualization of attention, representation space analysis, and, even more importantly, modeling of an interpretable latent space within multimodal SE frameworks offer a promising route for achieving these goals and thus enhance decision-making transparency and empower users of AVSE systems.

In summary, future research should strive to balance improvements in perceptual quality with the integration of several important additional aspects: robustness, scalability, fairness, stable synchronization, efficiency, and practicality for deployment in various multimodal interaction environments.

6. LIMITATIONS

Although significant progress has been made in AVSE research through deep learning approaches, many of the reviewed studies still exhibit important limitations related to methodology, experimental design, reproducibility, benchmarking consistency, and ethical considerations. These limitations affect the reliability of reported performance results, the fairness of comparisons among different AVSE methods, and the practicality of real-world deployment. Such challenges arise not only from the individual studies themselves but also from the considerable variability in multimodal datasets, evaluation protocols, frame synchronization strategies, and the rapid evolution of deep learning architectures and training methodologies. Therefore, the main outcome of this paper is a structured analytical overview of current AVSE research trends rather than a purely comprehensive summary of the field.

6.1 Methodological and data-related limitations

In this review, we only consider the so-called peer reviewed studies. This means we only looked at English language articles published after 2020. Of course, this is not perfect, because then we might miss industry reports or non-English language contributions to the topic (see studies [13, 17] for two

instances where such a more comprehensive review could bring interesting methodological perspectives as well as more insights into real-life AVSE deployments).

Metric heterogeneity: The number of studies that make use of WER, PESQ, STOI, SDR, AV-MOS, and other metrics is very high. In many cases, even for the same WER, different amounts of PESQ and STOI are achieved, which hinders a meta-analytic pooling of the results (Section 5). Also, for the evaluation of AVSE in general, there is no uniform framework to evaluate the performance of AVSE systems in the same way as for AS or VS. The current evaluation is strongly influenced by the used preprocessing pipeline (e.g., source speech alignment), the used synchronization method (e.g., STOI, SDR), the used dataset (e.g., the type of speakers, recording conditions) and the used set of benchmarks (Section 5).

In addition to the above-mentioned points, the reproducibility of current approaches is often not sufficient. Many papers do not release the pre-trained models, the synchronization correction, the training configuration, or the evaluation scripts of the experiments, in order to allow a reproducibility of the results by other researchers. This is a problem, especially for multimodal learning approaches, since small differences in the alignment correction, in the audio- and video-preprocessing, or in the partitioning of the dataset used for training and testing can lead to big differences in the achieved performance [66, 67].

The rapid development of methods for multimodal deep learning also affects this survey in a negative way, because the survey is written from a time perspective and therefore some results will have already become outdated as soon as they are written down, especially for recently introduced Transformer-based, diffusion-based, and self-supervised AVSE methods. Also more recent large-scale multimodal learning methods have recently been introduced, for instance using generative models such as Generative Adversarial Networks (GANs), Variational Auto-Encoders (VAEs) and Auto-Encoders, in which the multimodal input signals are used to learn a common hidden representation that captures underlying semantic information, and using unsupervised, self-supervised, and semi-supervised learning methods, where the available labeled data are used in combination with large amounts of unlabeled data in order to learn to represent and to process input data in a meaningful way [12, 67, 68].

Even with the structured PRISMA guidelines for study selection, the individual screening, the independent assessment by the two reviewers, and the quality-oriented study selection, there are sections of thematic analysis and methodological evaluation that are subject to some degree of individual interpretation, especially where multiple studies investigated largely the same questions, used hybrid approaches, or provided insufficient methodological details [15, 69].

6.2 Biases and their broader implications

This comparative analysis reveals significant inconsistencies in the experimental transparency provided in most of the surveyed papers. A number of studies lack sufficient documentation of the synchronization correction methods, the training and test sets partitioning, the demographics of the datasets used, the model's hyperparameters, and the data preprocessing methods. This lack of transparency makes it difficult to reproduce the experiments and to compare different multimodal architectures in a fair manner. This is especially challenging when the studies employ different experimental setups and methodologies [57, 70].

On the other hand, even when significant quantitative gains are reported, insufficient information about the experimental method is provided to allow the reader to make a balanced assessment of the trade-off between high performance and increased computational cost, latency, memory needs, and deployment constraints. Such an imbalance can drive research towards complex solutions that are not suitable for a real-time multimodal AVS deployment [71, 72].

A further challenge with respect to reproducibility is the inconsistent use of different benchmarks for comparison. On the one hand, controlled datasets like GRID and LRS2 facilitate reproducibility. On the other hand, large-scale in-the-wild datasets like AVSpeech and VoxCeleb2-AV pose many challenges with respect to synchronization, speaker variability, and noise, thereby leading to unstable results. As a result, comparison of Different architectures in single studies, where the architectures are trained and tested on different data, is not feasible [1, 73]. different datasets may not always reflect equivalent operational conditions. Table 5 summarizes the major methodological, computational, robustness-related, evaluation, and ethical limitations identified across representative AVSE studies included in this review.

Overall, these reproducibility and reporting inconsistencies reinforce the need for standardized documentation practices, transparent synchronization reporting, reproducible benchmarking protocols, and publicly accessible evaluation resources within future AVSE research.

6.3 Structural and ethical gaps

A number of common limitations in multimodal robustness and cross-population generalization in AVSE have been found in the literature reviewed above. The majority of current studies on AVSE make use of English-dominant, primarily audio-visual datasets such as LRS2, LRS3, AVSpeech, and the speeches in VoxCeleb2-AV [1, 74] for large-scale multimodal learning. However, the current datasets used in AVSE studies suffer from severe limitations in linguistic concentration and insufficient demographic representation, which hinders generalization to speech in many languages, especially in low-resource situations and in culturally and linguistically diverse environments [75].

The lack of demographic balance with respect to a number of different factors, including speaker identity, accent, age, gender, and visual characteristics, is a problem that is not yet well addressed in many of the datasets as well as experimental protocols reviewed by the current article [76]. This lack of balance in the current data can lead to AVSE systems that are skewed toward favoring over-represented populations, as well as under-performing in under-represented populations. robustness under heterogeneous real-world deployment conditions.

The review of current AVSE practices also found very limited discussion of issues surrounding privacy-aware audiovisual learning and the issue of informed consent in the use of large datasets of facial audiovisual data. AVSE systems process synchronized facial information and spoken communications; thus, they are inherently more sensitive than audio-only communication repositories. Ethics-related information about privacy governance in AVSE, such as the inclusion of privacy policies, a demographic description of the

data used, and a description of methods for assessing fairness, are not consistently reported in current AVSE studies [77].

AVSE benchmarking studies have to go beyond traditional performance metrics. They have to be supplemented with fairness evaluations, future-oriented assessments, a multilingual analysis of representation capabilities, a demographic transparency report, and strategies for privacy-governed data sets.

Table 5. Representative methodological, computational, robustness-related, evaluation, and ethical limitations identified across reviewed Audio-Visual Speech Enhancement (AVSE) studies

Reference	Dataset(s) Used	Data/Annotation Gap	Computational Cost	Robustness Limitation	Evaluation Limitation	Ethical/Privacy Concern
[1]	LRS3, GRID		√			
[8]	MuAViC	√				
[12]	Synthetic (Diff-AV)		√	√	√	
[16]	AVSUPERB				√	
[21]	LRS3, OuluVS2	√		√	√	
[32]	LRS2, LRS3	√		√		
[46]	AVSpeech, VoxCeleb2-AV		√	√		
[50]	Various					√
[61]	VoxCeleb2-AV					√
[61]	AVSpeech	√	√		√	
[72]	Multi-Channel AV		√	√		

Note: ✓ indicates the presence of the corresponding methodological or ethical limitation in the reviewed study.

6.4 Pathways for future improvement

To address the limitations of the current review, recommendations for improving future research in AVSE are outlined in Table 6. These recommendations are intended to address limitations of current AVSE research regarding methodological improvements and deployment issues, to improve reproducibility, consistency of benchmarking, fairness and transparency of deployment in multimodal AVSE.

A significant amount of progress has been made in multimodal AVSE between 2020 and 2025. However, most of the current studies present methodological inconsistencies, dataset imbalances, limited reproducibility, and a lack of fairness in evaluation. The lack of a strong and reliable benchmark for comparison among different AVSE studies [78, 79] limits the scope of current studies and affects their deployment in real-world scenarios. Thus, future AVSE studies should focus not only on the enhancement of the perceptual quality and the corresponding context, but also on transparency, reproducibility, fairness, synchronization, and deployment.

Table 6. Recommendations for improving methodological rigor, reproducibility, and transparency in future AVSE research

Recommendation	Expected Impact	Supporting References
Adopt standardized benchmark datasets (e.g., AVSpeech + MuAViC)	Reduces dataset bias and enhances multilingual robustness	[8, 61]
Publish reproducible code and model checkpoints	Improves independent validation and comparative benchmarking	[32, 39]
Report standardized evaluation metrics (WER, PESQ, STOI, SDR, AV-MOS)	Facilitates cross-study synthesis and benchmarking consistency	[22, 25, 29]
Include fairness-aware and demographic transparency statements	Improved fairness accountability and cross-population robustness assessment	[13, 73]
Document synchronization correction and preprocessing protocols	Supports ethical multimodal dataset management	[38, 57]
Establish privacy-aware dataset governance and consent procedures	Supports inclusivity and long-term sustainability.	[17, 50]
Report computational complexity and deployment-oriented constraints	Improves real-time feasibility assessment and practical deployment analysis	[67, 70]

Note: AVSE = Audio-Visual Speech Enhancement, PESQ = Perceptual Evaluation of Speech Quality, STOI = Short-Time Objective Intelligibility, SDR = Signal-to-Distortion Ratio, WER = Word Error Rate.

7. FUTURE RESEARCH DIRECTIONS

Future progress in AVSE will likely depend on addressing several unresolved methodological, computational, multilingual, reproducibility-related, and deployment-oriented challenges identified throughout the reviewed literature. Rather than representing a transition toward speculative multimodal intelligence systems, recent AVSE research trends primarily indicate increasing emphasis on scalable multimodal representation learning, resource-efficient contextual modeling, fairness-aware evaluation, and deployment-oriented audiovisual processing frameworks [13, 73, 80]. Consequently, future AVSE development should be interpreted as a continuation of current methodological evolution rather than as a replacement of existing audiovisual enhancement paradigms. Table 7 summarizes the principal

future research priorities identified through the analytical synthesis conducted in this review.

Table 7. Emerging future research priorities in Audio-Visual Speech Enhancement (AVSE) and their expected methodological impact

Research Axis	Emerging Focus	Expected Impact	Key References
Model Architecture	Foundation-scale diffusion + SSL hybrids	Unified, scalable multimodal learning	[10, 11, 12, 24, 73]
Multilingual & Cultural Diversity	Low-resource and cross-accent datasets	Inclusive and fair performance	[8, 21, 77, 80, 81]
Ethics & Fairness	Bias auditing, privacy preservation	Trustworthy and human-centric AVSP	[6, 7, 81]
Efficiency & Real-Time Deployment	Model compression and edge inference	Broader accessibility and sustainability	[20, 42, 71]
Interdisciplinary Integration	Affective, embodied, and cognitive fusion	Toward multimodal general intelligence	[18, 19, 58, 73]

7.1 Foundation-scale multimodal models

Looking to scale to process a large amount of information and to jointly process the acoustic and visual information in the different contexts in which they are used will probably be the most relevant for a future application in various contexts [10, 11]. Indeed, although conventional architectures based on CNN-RNN have been able to process the information from the different sensors, more recent multimodal architectures, such as transformer-based, diffusion-based, and self-supervised multimodal architectures, are able to process the information in a more effective way to model the context, especially in cases of degradation of the acoustic information and of occlusion of the visual information [12, 24]. However, although these architectures are able to process the information in a more effective way to model the context, their greater complexity, in terms of the number of parameters and the depth of the architecture, results in greater computational complexity, greater memory consumption of memory and a greater inference latency in studies [67, 68]. Thus, future research should not be limited to the study of the architectures able to process the information in a more effective way to model the context, but it should also focus on the study of the efficient ways to process the information from the different sensors, i.e., on the study of the efficient multimodal fusion mechanisms able to process in real time a large amount of information [82]. In addition to the above-mentioned aspects, the results obtained with the architectures able to reduce the dependency on the task-specific fine-tuning, by means of more general multimodal representation learning, are also very promising for future applications, since, thanks to the learning a more general representation, it is possible to process datasets of different sizes and also to process information in multiple languages, with a reduced number of annotations and a reduced sensitivity to the synchronization between the different modalities [30, 76].

7.2 Hybridization of diffusion and self-supervised learning

Diffusion-based AVSE architectures have generally achieved strong perceptual quality in audiovisual reconstruction. However, these architectures often rely on iterative inference processes that can be computationally expensive, making their deployment in latency-sensitive applications particularly challenging. Consequently, there is growing interest in developing more efficient and deployment-oriented diffusion-based AVSE frameworks suitable for real-time and resource-constrained environments. It is anticipated that future research will focus on improving the efficiency and scalability of diffusion-based architectures for AVSE applications.

In addition to perceptual enhancement, increasing attention is being directed toward integrating SSL frameworks such as HuBERT, wav2vec 2.0, and AV-HuBERT with multimodal diffusion models. Such integration may reduce dependence on manually labeled data while improving adaptive contextual representation learning and cross-modal alignment. Furthermore, conditioning enhancement models on visual contextual embeddings is considered a promising direction for improving AVSE performance under severe acoustic degradation conditions.

Despite their potential, diffusion-based AVSE systems still face several important challenges, including synchronization instability, computational complexity, cross-modal alignment difficulties, and limited scalability in real-time deployment scenarios. Addressing these limitations is expected to play a central role in the future development of robust and efficient AVSE systems.

7.3 Multilingual and culturally inclusive Audio-Visual Speech Processing

The majority of current AVSE research is dependent on English-dominant audio-visual datasets for training, such as LRS2, LRS3, AVSpeech, and the Audio-Visual parts of VoxCeleb2-AV [1, 61]. While the datasets of large-scale multimodal training data are very valuable for training, currently, there is a lack of solid evidence for robust AVSE in multilingual, cross-accent and cross-cultural setups [8, 21]. For future AVSE research, therefore, multilingual evaluation frameworks, low-resource audio and video data, as well as cross-domain evaluation protocols, are required to improve linguistic diversity and to counterbalance demographic imbalances of current multimodal learning systems [77, 80, 82]. Since AV processing deals with speech, which is language-specific and thus changes depending on the speaker's articulation, i.e. how they produce sounds, and synchronization, i.e. how audio and video are locked in time, which is also language-specific and thus changes across speech communities, it is particularly important to increase linguistic diversity in current AVSE research. In recent years, several approaches for multilingual SSL have been proposed, which aim to improve cross-lingual transferability by learning to represent data in a shared multimodal latent space [63, 76]. However, even for AV processing within a single language, current AVSE is affected by several major issues, i.e. dataset

imbalance, annotation inconsistencies, and synchronization variability, which make it difficult to perform comparative evaluations on datasets with different demographics in a fair manner. Future progress in multilingual AVSE therefore not only depends on increasing the size of used datasets, but also on improving transparency of datasets, demographic balance, synchronization as well as reproducibility of benchmarking experiments.

7.4 Ethical, fair, and privacy-preserving Audio-Visual Speech Processing

As systems for AVSE become increasingly important for applications such as assistive communication, telepresence, surveillance, and real-time human-computer interaction, fair evaluation and privacy protection will become more important in future multimodal research [13, 81]. The reviewed studies lack demographic transparency and do not consistently report the fairness-related evaluation procedures that are already used in AVSE. Therefore, future benchmarking approaches should provide more demographic information about the used data and report more consistently on performance, bias and synchronization transparency. Moreover, the approaches should be reproducible [17, 82].

Privacy-aware multimodal learning for AVSE is a growing research area. Audiovisual datasets contain time-synchronized facial and speech signals and thus present critical dimensions of human multimodal signals that require special attention to risks of multimodal data collection and sharing for AVSE learning. In order to reduce risks for misuse or for large-scale centralized repositories of sensitive human multimodal signals, recent multimodal learning approaches employ techniques typical for cross-modal mapping, such as federated multimodal learning, on-device inference, or learning privacy-preserving audio- and visual- representation spaces [71, 72, 82]. In summary, future benchmarking efforts for assessing AVSE performance must combine the already established assessment criteria for performance, bias, and fairness with privacy-sensitive practices for data and datasets for AVSE learning.

7.5 Real-time and edge-deployable multimodal systems

As previously stated in previous studies [42, 71], for future deployment-oriented AVSE systems, in addition to improved perceptual reconstruction quality, it is required to provide real-time processing, high computational efficiency, and low-latency audiovisual inference, suitable for the embedded and edge-oriented environments. A variety of mobile human-computer interfaces (HCl) (MHCls) are currently under development. Such MHCls and other wearable and telepresent HCls, as well as various low-power multimodal human-computer interaction (MMHCl) platforms, require efficient AVSE models, suitable for real-time processing, and running on limited resources. Although model pruning, quantization, knowledge distillation, and, in particular, lightweight multimodal encoders, significantly improve the deployment of the AVSE models, reducing their memory consumption and their respective processing costs, in many cases, decreasing the model's computational complexity causes a number of trade-offs affecting its perceptual enhancement quality, its ability to ensure synchronization between the reconstructed audio and video streams, and its ability to model various real-world contextual information. Therefore, the focus of future research in the area of AVSE should be placed on developing and evaluating a number of perception-oriented, balanced, and fairly optimized approaches that, simultaneously, improve the quality of the reconstructed audio-visual information, increase the system's processing speed, improve the processed streams' synchronization, and ensure highly energy-efficient and efficient AVSE models' on a variety of highly heterogeneous environments, and highly powerful platforms.

The future of AVSP research is moving in six interrelated directions, which will consolidate speech processing (SP) and make it a fully-fledged area of human-oriented multimodal intelligence. The future of SP will be determined by the interplay of the following six axes: the scalability of new methods, their inclusiveness, their ethics, and their efficiency. In this sense, the future of SP will depend on the unification of the following three areas: diffusion-based learning, fairness-aware governance, and real-time deployment. In this sense, the future of SP will become a technological as well as an ethical benchmark for the development of next-generation AI systems.

8. CONCLUSION

This systematic review provides a structured analytical synthesis of 102 studies published between 2020 and 2025 in the domain of AVSE. A total of 102 studies from the IEEE, Elsevier, Springer, and ACM databases were reviewed and analyzed for AVSE methodological ,evolutions including their corresponding multimodal architectures, large-scale datasets, evaluation methodologies, scalability of computations for deployment, fairness-aware benchmarking and design considerations for deployment. The reviewed AVSE studies transitioned from earlier conventional CNN-RNN-based architectures and handcrafted feature engineering to more recent AVSE ,methods which utilize transformer-based models, diffusion models, ,SSL and hybrid multimodal architectures. These modern approaches to AVSE achieve enhanced contextual modeling and also provide better perceptual quality of reconstructed audio-visual signals, and robustness to acoustically degraded and visually occluded audio-visual signals. With the advent of several very large-scale audio-visual datasets, such as AVSpeech and MuAViC, improved multilingual AVSE evaluation and also cross-domain multimodal learning benchmarking are achieved. Although AVSE methods have made significant progress, several challenges remain to be addressed in AVSE research. One major issue is the variability in the composition of the datasets, synchronization approaches, evaluation metrics, fairness, and reproducibility. In addition to the issueslisted above, AVSE studies also lack demographic information about speakers, imbalanced number of languages, and lack a common AVSE benchmark among several AVSE approaches. A major conclusion of this review is that, as of today, none of the AVSE architectures can fulfill all the requirements of a good AVSE system, such as high-quality perceptual enhancement, scalable computations for deployment, robust synchronization, multilingual adaptability, fairness, and also real-time processing capability. While transformer-based and diffusion-based AVSE methods offer better contextualized audio-visual reconstruction quality, several very lightweight multimodal architectures have also been designed and tested for the low-latency and resource-constrained AVSE applications. Future AVSE research therefore requires a

balanced set of optimizations toimprove the robustness, efficiency, reproducibility, fairness, and deployability of AVSE systems for processing a wide variety of degraded audio-visual signals. The state of the art in AVSE research is moving toward developing more scalable, context-aware, and deployable multimodal learning architectures for AVSE. Future research thus needs to focus on developing a reproducible set of AVSE benchmarks, synchronized audio-visual data, and also on designing more multilingual and democratic AVSE approaches. Furthermore, the future AVSE research also requires the development of AVSE methods that are fair, privacy-aware, and computationally efficient for the real-time processing of degraded audio-visual signals. This review therefore aims to provide a comprehensive overview of the current state-of-the-art methods in the field of AVSE, and also outlines future research challenges and directions by integrating methodological advancements in the field with corresponding future research priorities and deployment-oriented perspectives.

REFERENCES

[1] Michelsanti, D., Tan, Z.H., Zhang, S.X., Xu, Y., Yu, M., Yu, D. (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. https://doi.org/10.1109/TASLP.2021.3066303

[2] Adeel, A., Gogate, M., Hussain, A. (2020). Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments. Information Fusion, 59: 163-170. https://doi.org/10.1016/j.inffus.2019.08.008

[3] Afouras, T., Chung, J.S., Senior, A., Vinyals, O., Zisserman, A. (2018). Deep audio-visual speech recognition. IEEE transactions on pattern analysis and machine intelligence, 44(12): 8717-8727. https://doi.org/0.1109/TPAMI.2018.2889052

[4] Shi, B., Hsu, W.N., Lakhotia, K., Mohamed, A. (2022). Learning audio-visual speech representation by masked multimodal cluster prediction. arXiv preprint arXiv: 2201.02184. https://doi.org/10.48550/arXiv.2201.02184

[5] Gao, R., Grauman, K. (2021). VisualVoice: Audio-visual speech separation with cross-modal consistency. arXiv preprint arXiv:2101.03149. https://doi.org/10.48550/arXiv.2101.03149

[6] Tseng, Y., Berry, L., Chen, Y.T., Chiu, I.H., Lin, H.H., Liu, M. (2024). AV-SUPERB: A multi-task benchmark for audio-visual representation learning. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, pp. 6890-6894. https://doi.org/10.1109/ICASSP48485.2024.10445941

[7] Park, K., Oh, C., Dong, S. (2024). KMSAV: Korean multi-speaker spontaneous audiovisual dataset. ETRI Journal, 46(1): 71-81. https://doi.org/10.4218/etrij.2023-0352

[8] Ma, P., Petridis, S., Pantic, M. (2021). End-to-end audio-visual speech recognition with conformers. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, pp. 7613-7617. https://doi.org/10.1109/ICASSP39728.2021.9414567

[9] Ma, P., Petridis, S., Pantic, M. (2022). Visual speech recognition for multiple languages in the wild. Nature Machine Intelligence, 4(11): 930-939. https://doi.org/10.1038/s42256-022-00550-z

[10] Pan, X.C., Chen, P.Y., Gong, Y.C., Zhou, H.L., Wang, X.B., Lin, Z.H. (2022). Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Dublin, Ireland, pp. 4493-4504. https://doi.org/10.18653/v1/2022.acl-long.308

[11] Gulati, A., Qin, J., Chiu, C.C., et al. (2020). Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100. https://doi.org/10.48550/arXiv.2005.08100

[12] 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. https://doi.org/10.1109/TASLP.2021.3122291

[13] Baevski, A., Zhou, Y., Mohamed, A., Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems (NeurIPS), 33: 12449-12460. https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

[14] Chen, S.Y., Wang, C.Y., Chen, Z.Y., Wu, Y., Liu, S.J., Chen, Z. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6): 1505-1518. https://doi.org/10.1109/JSTSP.2022.3188113

[15] Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M. (2022). data2vec: A general framework for self-supervised learning in speech, vision and language. In Proceedings of the International Conference on Machine Learning (ICML), pp. 1298-1312. https://proceedings.mlr.press/v162/baevski22a.

[16] Chen, S.Y., Wu, Y., Wang, C.Y., Chen, Z.Y., Chen, Z., Liu, S.J. (2022). UniSpeech-SAT: Universal speech representation learning with speaker aware pre-training. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, pp. 6152-6156. https://doi.org/10.1109/ICASSP43922.2022.9747077

[17] Chen, S.Y., Wu, Y., Wang, C.Y., et al. (2023). BEATs: Audio pre-training with acoustic tokenizers. arXiv preprint arXiv:2212.09058. https://doi.org/10.48550/arXiv.2212.09058

[18] Akbari, H., Yuan, L.Z., Qian, R., et al. (2021). VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. Advances in Neural Information Processing Systems (NeurIPS), 34: 24206-24221. https://proceedings.neurips.cc/paper/2021/hash/cb3213ada48302953cb0f166464ab356-Abstract.html.

[19] Girdhar, R., El-Nouby, A., Liu, Z., et al. (2023). ImageBind: One embedding space to bind them all. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15180-15190. https://openaccess.thecvf.com/content/CVPR2023/html/

Girdhar_ImageBind_One_Embedding_Space_To_Bind_Them_All_CVPR_2023_paper.html.
[20] Tong, Z., Song, Y.B., Wang, J., Wang, L.M. (2022). VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in Neural Information Processing Systems (NeurIPS), 35: 10078-10093. https://proceedings.neurips.cc/paper_files/paper/2022/hash/416f9cb3276121c42eebb86352a4354a-Abstract-Conference.html.
[21] Luo, Y., Chen, Z., Yoshioka, T. (2020). Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, pp. 46-50. https://doi.org/10.1109/ICASSP40776.2020.9054266
[22] Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., Zhong, J.Y. (2021). Attention is all you need in speech separation. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, pp. 21-25. https://doi.org/10.1109/ICASSP39728.2021.9413901
[23] Chen, J., Mao, Q., Liu, D. (2020). Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation. In Proceedings of Interspeech 2020, Shanghai, China, pp. 2642-2646. https://doi.org/10.21437/Interspeech.2020-2205
[24] Hu, Y.X., Liu, Y., Lv, S.B., et al. (2020). DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. In Proceedings of Interspeech 2020, Shanghai, China, pp. 2472-2476. https://doi.org/10.21437/Interspeech.2020-2537
[25] Hao, X., Su, X.D., Horaud, R., Li, X.F. (2021). FullSubNet: A full-band and sub-band fusion model for real-time single-channel speech enhancement. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, pp. 6633-6637. https://doi.org/10.1109/ICASSP39728.2021.9414177
[26] Fu, S.W., Yu, C., Hsieh, T.A., et al. (2021). MetricGAN+: An improved version of MetricGAN for speech enhancement. In Proceedings of Interspeech 2021, Brno, Czechia, pp. 201-205. https://doi.org/10.21437/Interspeech.2021-599
[27] Welker, S., Richter, J., Gerkmann, T. (2022). Speech enhancement with score-based generative models in the complex STFT domain. arXiv preprint arXiv:2203.17004. https://doi.org/10.48550/arXiv.2203.17004
[28] Richter, J., Welker, S., Lemercier, J.M., Lay, B., Gerkmann, T. (2023). Speech enhancement and dereverberation with diffusion-based generative models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31: 2351-2364. https://doi.org/10.1109/TASLP.2023.3285241
[29] Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B. (2021). DiffWave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761. https://doi.org/10.48550/arXiv.2009.09761
[30] Chen, N.X., Zhang, Y., Zen, H., Weiss, R.J., Norouzi, M., Chan, W. (2021). WaveGrad: Estimating gradients for waveform generation. arXiv preprint arXiv:2009.00713. https://doi.org/10.48550/arXiv.2009.00713
[31] Chen, H., Wang, Q., Du, J., Yin, B.C., Pan, J., Lee, C.H. (2024). Optimizing audio-visual speech enhancement using multi-level distortion measures for audio-visual speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32: 2508-2521. https://doi.org/10.1109/TASLP.2024.3393732
[32] Gabbay, A., Ephrat, A., Halperin, T., Peleg, S. (2022). Seeing through noise: Visually driven speaker separation and enhancement. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, pp. 3051-3055, https://doi.org/10.1109/ICASSP.2018.8462527
[33] Lee, J., Chung, S.W., Kim, S., Kang, H.G., Sohn, K. (2021). Looking into your speech: Learning cross-modal affinity for audio-visual speech separation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, pp. 1336-1345. https://doi.org/10.1109/CVPR46437.2021.00139
[34] Makishima, N., Ihori, M., Takashima, A., Tanaka, T., Orihashi, S., Masumura, R. (2021). Audio-visual speech separation using cross-modal correspondence loss. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, pp. 6673-6677. https://doi.org/10.1109/ICASSP39728.2021.9413491
[35] Liu, D., Zhang, T., Christensen, M.G., Yi, C., An, Z. (2024). Audio-visual fusion with temporal convolutional attention network for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32: 4647-4660. https://doi.org/10.1109/TASLP.2024.3463411
[36] Xu, X.M., Tu, W.P., Yang, Y.H. (2025). Efficient audio-visual information fusion using encoding pace synchronization for audio-visual speech separation. Information Fusion, 115: 102749. https://doi.org/10.1016/j.inffus.2024.102749
[37] Zhou, J., Wang, J., Zhang, J., et al. (2022). AVSBench: A pixel-level audio-visual segmentation benchmark. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 357-373.
[38] Ahmed, S., Chen, C.W., Ren, W., et al. (2024). Deep complex U-Net with conformer for audio-visual speech enhancement. In Proceedings of the 3rd COG-MHEAR Workshop on Audio-Visual Speech Enhancement (AVSEC), Kos, Greece, pp. 51-55. https://doi.org/10.21437/AVSEC.2024-11
[39] Seo, P.H., Nagrani, A., Schmid, C. (2023). AVFormer: Injecting vision into frozen speech models for zero-shot AV-ASR. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Vancouver, BC, Canada, pp. 22922-22931. https://doi.org/10.1109/CVPR52729.2023.02195
[40] Sadok, S., Leglaive, S., Girin, L., Alameda-Pineda, X., Séguier, R. (2024). A multimodal dynamical variational autoencoder for audiovisual speech representation learning. Neural Networks, 172: 106120. https://doi.org/10.1016/j.neunet.2024.106120
[41] Martel, H., Richter, J., Li, K., Hu, X., Gerkmann, T. (2023). Audio-visual speech separation in noisy environments with a lightweight iterative model. In Proceedings of the Interspeech 2023, Dublin, Ireland, pp. 1673-1677. https://doi.org/10.21437/Interspeech.2023-

1753
[42] Pan, T., Liu, J., Wang, B., Tang, J., Wu, G. (2024). Ravss: Robust audio-visual speech separation in multi-speaker scenarios with missing visual cues. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne VIC, Australia, pp. 4748-4756. https://doi.org/10.1145/3664647.3681261
[43] Wang, Z.Q., Cornell, S., Choi, S., Lee, Y., Kim, B.Y., Watanabe, S. (2023). TF-GridNet: Integrating full-and sub-band modeling for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31: 3221-3236. https://doi.org/10.1109/TASLP.2023.3304482
[44] Ochieng, P. (2023). Deep neural network techniques for monaural speech enhancement and separation: State of the art analysis. Artificial Intelligence Review, 56(Suppl 3): 3651-3703. https://doi.org/10.1007/s10462-023-10612-2
[45] Upreti, R., Lind, P.G., Elmokashfi, A., Yazidi, A. (2024). Trustworthy machine learning in the context of security and privacy. International Journal of Information Security, 23(3): 2287-2314. https://doi.org/10.1007/s10207-024-00813-3
[46] 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. In Proceedings of the Interspeech 2023, Dublin, Ireland, pp. 4064-4068. https://doi.org/10.21437/Interspeech.2023-1182
[47] Vilaca, L., Yu, Y., Viana, P. (2025). A survey of recent advances and challenges in deep audio-visual correlation learning. ACM Computing Surveys, 57(12): 1-46. https://doi.org/10.1145/3696445
[48] Sun, G., Yu, W., Tang, C., et al. (2024). video-SALMONN: Speech-Enhanced Audio-Visual Large Language Models. In Proceedings of the 41st International Conference on Machine Learning, pp. 47198-47217. https://proceedings.mlr.press/v235/sun24l.html.
[49] Chinen, M., Lim, F.S.C., Skoglund, J., Gureev, N., O'Gorman, F., Hines, A. (2021). ViSQOL v3: An open source production ready objective speech and audio metric. In 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX), Athlone, Ireland, pp. 1-6. https://doi.org/10.1109/QoMEX48832.2020.9123150
[50] Loizou, P.C. (2011). Speech Quality Assessment. In Multimedia Analysis, Processing and Communications, pp. 623-654. https://doi.org/10.1007/978-3-642-19551-8_23
[51] Reddy, C.K.A., Gopal, V., Cutler, R. (2021). DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, pp. 6493-6497. https://doi.org/10.1109/ICASSP39728.2021.9414878
[52] Mittag, G., Naderi, B., Chehadi, A., Möller, S. (2021). NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. arXiv preprint arXiv:2104.09494. https://doi.org/10.21437/Interspeech.2021-299
[53] Nayem, K.M., Williamson, D.S. (2023). Attention-based speech enhancement using human quality perception modeling. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32: 250-260. https://doi.org/10.1109/TASLP.2023.3328282
[54] Li, C., Shi, J., Zhang, W., et al. (2021). ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration. In 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, pp. 785-792. https://doi.org/10.1109/SLT48900.2021.9383615
[55] Manocha, P., Jin, Z., Zhang, R., Finkelstein, A. (2021). CDPAM: Contrastive learning for perceptual audio similarity. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, pp. 196-200. https://doi.org/10.1109/ICASSP39728.2021.9413711
[56] Buolamwini, J., Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency, pp. 77-91. https://proceedings.mlr.press/v81/buolamwini18a.html.
[57] Bird, S., Dudík, M., Edgar, R., et al. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft.
[58] Kairouz, P., McMahan, H.B. (2021). Advances and open problems in federated learning. Foundations and trends in machine learning, 14(1-2): 1-210. https://doi.org/10.1561/2200000083
[59] Li, K., Sang, W., Zeng, C., Yang, R., Chen, G., Hu, X. (2025). Sonicsim: A customizable simulation platform for speech processing in moving sound source scenarios. International Conference on Learning Representations, 2025: 67379-67405.
[60] Kumar, S., Ghosh, S., Tyagi, U., et al. (2025). ProSE: Diffusion priors for speech enhancement. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Albuquerque, New Mexico, pp. 12470-12483. https://doi.org/10.18653/v1/2025.naacl-long.619
[61] Wang, Z.Q., Erdogan, H., Wisdom, S., et al. (2021). Sequential multi-frame neural beamforming for speech separation and enhancement. In 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, pp. 905-911. https://doi.org/10.1109/SLT48900.2021.9383522
[62] Li, C., Chen, Z., Luo, Y., et al. (2021). Dual-path modeling for long recording speech separation in meetings. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, pp. 5739-5743. https://doi.org/10.1109/ICASSP39728.2021.9414127
[63] Gogate, M., Dashtipour, K.K., Hussain, A. (2024). A lightweight real-time audio-visual speech enhancement framework. In Proceedings of the 3rd COG-MHEAR Workshop on Audio-Visual Speech Enhancement (AVSEC), Kos, Greece, pp. 19-23. https://doi.org/10.21437/AVSEC.2024-5.
[64] Tian, Y., Hu, D., Xu, C. (2021). Cyclic co-learning of sounding object visual grounding and sound separation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, pp. 2745-2754. https://doi.org/10.1109/CVPR46437.2021.00277

[65] Tan, R., Ray, A., Burns, A., et al. (2023). Language-guided audio-visual source separation via trimodal consistency. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, pp. 10575-10584. https://doi.org/10.1109/CVPR52729.2023.01019

[66] Djilali, Y.A.D., Narayan, S., Boussaid, H., Almazrouei, E., Debbah, M. (2023). Lip2vec: Efficient and robust visual speech recognition via latent-to-latent visual to audio representation mapping. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, pp. 13790-13801. https://doi.org/10.1109/ICCV51070.2023.01268

[67] Marqas, R.B., Mousa, A., Özyurt, F. (2024). Innovative hybrid deep learning models for financial sentiment analysis. Acadlore Transactions on AI and Machine Learning, 3(4): 225-236. https://doi.org/10.56578/ataiml030404

[68] Zhang, Y., Yang, S., Shan, S., Chen, X. (2024). Es3: Evolving self-supervised learning of robust audio-visual speech representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, pp. 27069-27079. https://doi.org/10.1109/CVPR52733.2024.02556

[69] Han, H., Anwar, M., Pino, J., et al. (2024). XLAVS-R: Cross-lingual audio-visual speech representation learning for noise-robust speech perception. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, pp. 12896-12911. https://doi.org/10.18653/v1/2024.acl-long.697

[70] Cheng, X., Jin, T., Li, L., Lin, W., Duan, X., Zhao, Z. (2023). Opensr: Open-modality speech recognition via maintaining multi-modality alignment. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, pp. 6592-6607. https://doi.org/10.18653/v1/2023.acl-long.363

[71] Hu, Y., Chen, C., Li, R., Zou, H., Chng, E.S. (2023). Mir-gan: Refining frame-level modality-invariant representations with adversarial network for audio-visual speech recognition. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, pp. 11610-11625. https://doi.org/10.18653/v1/2023.acl-long.649

[72] Yeo, J., Han, S., Kim, M., Ro, Y.M. (2024). Where visual speech meets language: VSP-LLM framework for efficient and context-aware visual speech processing. In Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, pp. 11391-11406. https://doi.org/10.18653/v1/2024.findings-emnlp.666

[73] Ayilo, J.E., Sadeghi, M., Serizel, R., Alameda-Pineda, X. (2024). Diffusion-based Unsupervised Audio-visual Speech Enhancement. arXiv preprint arXiv:2410.05301. https://doi.org/10.48550/arXiv.2410.05301

[74] Wang, F., Yang, S., Shan, S., Chen, X. (2025). CogCM: Cognition-inspired contextual modeling for audio-visual speech enhancement. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 21408-21418.

[75] Yeo, J.H., Rha, H., Park, S.J., Ro, Y.M. (2025). MMS-LLaMA: Efficient LLM-based audio-visual speech recognition with minimal multimodal speech tokens. In Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, pp. 20724-20735. https://doi.org/10.18653/v1/2025.findings-acl.1065

[76] Chuang, S.Y., Wang, H.M., Tsao, Y. (2022). Improved lite audio-visual speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30: 1345-1359. https://doi.org/10.1109/TASLP.2022.3153265

[77] Lai, R.L., Hou, J.C., Chern, I., et al. (2023). Audio-visual speech enhancement using self-supervised learning to improve speech intelligibility in cochlear implant simulations. arXiv preprint arXiv:2307.07748. https://doi.org/10.48550/arXiv.2307.07748

[78] Zhu, Z., Zhang, L., Pei, K., Chen, S. (2025). Endpoint-aware audio-visual speech enhancement utilizing dynamic weight modulation based on SNR estimation. Neural Networks, 185: 107152. https://doi.org/10.1016/j.neunet.2025.107152

[79] Chen, H., Mira, R., Petridis, S., Pantic, M. (2024). RT-LA-VocE: Real-Time Low-SNR Audio-Visual Speech Enhancement. In Proceedings of the Interspeech 2024, Kos, Greece, pp. 2215-2219. https://doi.org/10.21437/Interspeech.2024-516

[80] Foroushi, Z., Dansereau, R.M. (2024). Dynamic audio-visual speech enhancement using recurrent variational autoencoders. In 2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC), Aalborg, Denmark, pp. 60-64. https://doi.org/10.1109/IWAENC61483.2024.10693981

[81] Yang, K., Marković, D., Krenn, S., Agrawal, V., Richard, A. (2022). Audio-visual speech codecs: Rethinking audio-visual speech enhancement by re-synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, US, pp. 8227-8237. https://doi.org/10.1109/CVPR52688.2022.00805

[82] Blanco, A.L.A., Valentini-Botinhao, C., Klejch, O., et al. (2023). AVSE challenge: Audio-visual speech enhancement challenge. In 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, pp. 465-471. https://doi.org/10.1109/SLT54892.2023.10023284