SBC Computing Reviews, 2026, 05:1
ISSN: 2966-3938 • doi: 10.5753/rocs.2026.6237

Review

A Systematic Review on Audio-Visual Speech-To-Speech Translation Models

Alexandre de Godoy Pereira [Instituto de Pesquisa Tecnológica (IPT), São Paulo | *alexandre.pereira@ensino.ipt.br*]
Renato Cordeiro Ferreira [Instituto de Matemática e Estatística, Universidade de São Paulo (IME–USP) | *renatocf@ime.usp.br*]
Alfredo Goldman [Instituto de Matemática e Estatística, Universidade de São Paulo (IME–USP) | *gold@ime.usp.br*]

Instituto de Pesquisa Tecnológica (IPT), São Paulo – SP, Brazil.

Abstract. This systematic literature review summarizes the evolution of Audio-Visual Speech-to-Speech Translation (AV-S2ST), focusing on the shift from cascaded systems to direct end-to-end models. Key advancements utilize self-supervised learning (SSL) (e.g. AV-HuBERT) and discrete units (e.g. TransFace) to overcome data scarcity and enable textless multilingual translation. Current research tackles challenges in noise robustness, lip-synchrony, isometric translation, and speaker preservation. Techniques like cross-modal distillation show promise, driving progress towards robust, real-time AV-S2ST with enhanced zero-shot capabilities.

Keywords: Audio-Visual Speech-to-Speech Translation, Systematic Literature Review, Self-Supervised Learning, AV-HuBERT, Discrete Units, Multimodal Machine Translation

Received: DD Month YYYY • **Accepted:** DD Month YYYY • **Published:** DD Month YYYY

1 Introduction

1.1 Background and Motivation

The recent growth of multimedia platforms requires advances in cross-lingual communication technologies, driven by the increasing demand for seamless interaction in online meetings, global content dissemination, and virtual education [Cheng *et al*., 2024; Choi *et al*., 2024]. Audio-only Speech-to-Speech Translation (S2ST) systems have demonstrated significant progress, particularly with end-to-end models [Di Gangi *et al*., 2019; Inaguma *et al*., 2019] and techniques like pseudo-labeling to address data limitations [Dong *et al*., 2022]. Nevertheless, they exhibit performance degradation in noisy acoustic environments [Huang *et al*., 2023; Han *et al*., 2024]. Furthermore, audio-only modalities do not convey the full communicative context present in visual speech cues, such as lip articulation, which is crucial for perceptual realism and intelligibility, in particular when audio is compromised [Choi *et al*., 2024; Goncalves *et al*., 2025]. Audio-Visual Speech-to-Speech Translation (AV-S2ST) addresses these limitations by integrating visual information [Choi *et al*., 2024; Huang *et al*., 2023; Cheng *et al*., 2023; Han *et al*., 2024], with the aim of improving noise robustness [Huang *et al*., 2023; Han *et al*., 2024] and providing a more concrete translation that includes synchronized visual speech output [Choi *et al*., 2024; Cheng *et al*., 2024; Goncalves *et al*., 2025]. This capability is crucial for applications demanding high fidelity and realism [Choi *et al*., 2024], including virtual conferencing, media dubbing [Cheng *et al*., 2024], and accessibility tools.

1.2 Challenges in AV-S2ST

The development of effective AV-S2ST systems faces significant obstacles. Conventional approaches often employ cascaded systems that integrate Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), Text-to-Speech (TTS), and Talking Face Generation (TFG) modules [Choi *et al*., 2024; Cheng *et al*., 2024; Huang *et al*., 2023; Goncalves *et al*., 2025]. However, these pipelines are susceptible to error propagation [Choi *et al*., 2024; Huang *et al*., 2023; Cheng *et al*., 2024], increased latency [Choi *et al*., 2024; Cheng *et al*., 2024], and inflexibility, particularly for unwritten languages where intermediate textual representations are not available [Choi *et al*., 2024; Huang *et al*., 2023].

Consequently, the field is moving towards direct end-to-end models. This paradigm shift, while promising, introduces different research challenges.

1. **Data Scarcity**: Parallel audio-visual translation data are considerably scarcer than their audio-only or text counterparts [Choi *et al*., 2024; Huang *et al*., 2023; Han *et al*., 2024], posing a major obstacle to training data-intensive end-to-end models [Choi *et al*., 2024; Huang *et al*., 2023; Cheng *et al*., 2024].
2. **Lip-Synchrony**: Achieving precise temporal alignment between the translated audio and generated lip movements (lip-synchrony) is key for tangible quality but technically demanding and often requires dedicated mechanisms for visual data processing.
3. **Isometric Translation**: Ensuring the duration of the translated AV output matches the source target reference is critical for applications like dubbing to avoid artifacts from frame repetition or truncation. This requires effective duration modeling, often termed isometric translation.
4. **Speaker Preservation**: Maintaining the vocal timbre and facial identity of the source speaker in the translated output is often desirable for preserving authenticity.
5. **Multilingual Generalization and Robustness**: Developing models that perform robustly across many language pairs, handle low-resource languages, and maintain accuracy under various noise conditions remains a challenge.

1.3 Context and Scope of the Review

Although the domain of direct audio-only Speech-to-Speech Translation (S2ST) has been addressed in prior survey literature [Sarim *et al*., 2025; Gupta *et al*., 2024], the sub-field of Audio-Visual S2ST (AV-S2ST) has undergone a distinct and accelerated evolution. In around 3 years, referencing the 2022-2024 timeframe of the core papers, we have witnessed rapid advancements in AV-S2ST, driven by new methodologies aimed at mitigating the preceding challenges. This concentrated progress necessitates a systematic literature review to synthesize the current state of the art.

Key innovations include the adoption of self-supervised learning (SSL) frameworks, notably AV-HuBERT and its variants, for learning unified audio-visual representations, and the utilization of discrete units for textless translation. Other techniques such as cross-modal distillation, adapted target forcing mechanisms for multimodal inputs, dedicated lip-synchrony constraints, and bounded duration predictors are actively being researched.

2 Background and Theoretical Approaches

The advancement of AV-S2ST is built upon theoretical foundations and model architectures originating from speech processing, multimodal learning, and machine translation. This section details the pivotal approaches referenced in the selected literature, focusing on the shift towards direct translation and the enabling role of self-supervised learning and discrete representations.

2.1 From Cascaded Pipelines to Direct End-to-End Systems

Initial efforts in S2ST, and subsequently AV-S2ST, often adopted cascaded architectures. These typically involve discrete stages: ASR converts source speech to text, NMT translates the text, TTS synthesizes target language audio, and potentially TFG creates visual speech [Choi *et al*., 2024; Cheng *et al*., 2024; Goncalves *et al*., 2025]. While leveraging specialized modules, this approach introduces significant latency due to sequential processing [Choi *et al*., 2024; Cheng *et al*., 2024] and suffers from error propagation, where errors in earlier stages (e.g., ASR) negatively impact downstream modules [Choi *et al*., 2024; Huang *et al*., 2023; Cheng *et al*., 2024]. Furthermore, cascaded systems often require textual representations, limiting their applicability to languages without established writing systems [Choi *et al*., 2024; Huang *et al*., 2023]. A comparison of these two architectures is illustrated in Figure 1.

To avoid these issues, the academy has increasingly explored direct end-to-end models [Choi *et al*., 2024; Huang *et al*., 2023; Cheng *et al*., 2024; Han *et al*., 2024; Dong *et al*., 2022]. These models aim to map the source modality (audio or audio-visual speech) directly to the target modality (target language audio-visual speech) within a single, jointly optimized neural network [Choi *et al*., 2024; Huang *et al*., 2023]. Pioneer works like Translatotron demonstrated the feasibility of direct S2ST [Dong *et al*., 2022], often employing multi-task learning strategies where auxiliary tasks like ASR or ST guide the primary translation task. Models like AV-TranSpeech [Huang *et al*., 2023] and AV2AV [Choi *et al*., 2024] explicitly adopt this direct, multimodal paradigm. While promising for reducing latency and potential errors [Choi *et al*., 2024; Cheng *et al*., 2024], direct models intensify the need for large parallel datasets [Huang *et al*., 2023; Choi *et al*., 2024] and require sophisticated architectures to handle the complex cross-modal and cross-lingual mapping [Han *et al*., 2024; Di Gangi *et al*., 2019; Inaguma *et al*., 2019].

2.2 Self-Supervised Learning for Speech and Audio-Visual Representations

SSL has become a cornerstone for advancing speech processing tasks, particularly in addressing data scarcity. These models learn powerful representations from vast amounts of unlabeled data, which can then be fine-tuned for downstream tasks.

In the audio domain, models like HuBERT (Hidden Unit BERT) have proven highly effective [Huang *et al*., 2023; Cheng *et al*., 2024; Han *et al*., 2024; Choi *et al*., 2024]. HuBERT employs a masked prediction approach similar to BERT for text [Huang *et al*., 2023; Han *et al*., 2024]. It first generates target labels (discrete hidden units) for masked audio segments using an offline clustering step (e.g., k-means) on features like Mel-Frequency Cepstral Coefficients (MFCCs) or learned representations from a previous iteration [Han *et al*., 2024; Choi *et al*., 2024]. The model is then trained to predict these discrete units for the masked portions of the input audio sequence [Han *et al*., 2024]. This forces the model to learn rich contextualized acoustic and linguistic representations. Wav2vec 2.0 is another influential SSL model, often using a contrastive loss objective on masked segments [Huang *et al*., 2023; Dong *et al*., 2022]. Pre-trained SSL encoders like these are frequently used as powerful feature extractors in direct S2ST and AV-S2ST systems [Huang *et al*., 2023; Dong *et al*., 2022].

Extending SSL to the multimodal domain, AV-HuBERT adapts the HuBERT pre-training objective for synchronized audio and visual (lip movement) inputs [Huang *et al*., 2023; Han *et al*., 2024; Cheng *et al*., 2023; Choi *et al*., 2024]. AV-HuBERT typically uses separate front-end networks (e.g., CNNs for audio, ResNets for visual) to extract features, fuses them (e.g., element-wise addition or concatenation), and feeds the result into a Transformer backbone [Huang *et al*., 2023; Han *et al*., 2024; Choi *et al*., 2024]. In summary, it applies masking independently to both audio and visual streams and predicts shared, discrete units derived from audio features or iteratively refined multimodal features [Han *et al*., 2024; Choi *et al*., 2024]. To encourage true multimodal learning and prevent over-reliance on one modality, modality dropout is often employed during pre-training, randomly masking out entire streams [Choi *et al*., 2024]. The goal is to learn unified audio-visual representations that are robust and informative regardless of whether the input is audio-only, visual-only, or both [Choi *et al*., 2024]. XLAVS-R [Han *et al*., 2024] explicitly builds upon an audio-only SSL model (XLS-R) and injects visual modality during continued AV pre-training using an AV-HuBERT-like objective. The core mechanisms of both HuBERT and AV-HuBERT are depicted in Figure 2.

Model Name	Input Modality(ies)	Primary Pre-training Objective	Typical Use Case in Reviewed Papers
HuBERT Cheng *et al*. [2023]; Huang *et al*. [2023]; Choi *et al*. [2024]; Han *et al*. [2024]; Cheng *et al*. [2024]	Audio	Masked Prediction (predicting clustered hidden units)	- Generating discrete units for target speech (textless approach) - Base model for extensions (e.g., AV-HuBERT)
Wav2Vec 2.0 Huang *et al*. [2023]; Dong *et al*. [2022]	Audio	Contrastive Prediction (over masked features or quantized units)	- Pre-trained audio encoder for S2ST models (often as baseline or component)
AV-HuBERT Choi *et al*. [2024]; Huang *et al*. [2023]; Han *et al*. [2024]; Cheng *et al*. [2023]	Audio-Visual	Masked Prediction (predicting shared units, often audio-derived or iteratively refined)	- Learning unified AV representations - Foundation for AV-S2ST & Audio-Visual Speech Recognition (AVSR) - Modality-agnostic feature extraction
XLS-R Han *et al*. [2024]	Audio (Multilingual)	Contrastive / Quantized Prediction (similar to Wav2Vec 2.0)	- Large-scale multilingual audio pre-training - Base model for XLAVS-R audio-visual adaptation

Table 1. Summary of Key SSL Models Referenced in Reviewed Papers

2.3 Discrete Units as Intermediate Representation

A key technique enabled by SSL models like HuBERT and AV-HuBERT is the use of discrete units [Choi *et al*., 2024; Cheng *et al*., 2023; Huang *et al*., 2023; Han *et al*., 2024]. By applying a quantization method (typically k-means clustering) [Choi *et al*., 2024; Han *et al*., 2024] to the learned representations (e.g., from an intermediate Transformer layer), the continuous speech signal can be mapped to a sequence of discrete tokens, similar to words or subwords in text. These discrete units serve as effective intermediate targets for textless speech processing tasks [Choi *et al*., 2024; Huang *et al*., 2023; Cheng *et al*., 2024]. In the context of S2ST/AV-S2ST, the translation model learns to map source speech features (or units) to target language discrete units [Choi *et al*., 2024; Cheng *et al*., 2024; Huang *et al*., 2023]. A separate unit-based vocoder (or AV synthesizer) then converts these target units back into the target waveform and visual speech [Choi *et al*., 2024; Cheng *et al*., 2024; Huang *et al*., 2023]. This approach avoids direct regression onto complex spectrograms or waveforms, potentially simplifying the learning task [Choi *et al*., 2024; Cheng *et al*., 2024], and crucially, it allows leveraging large audio-only parallel data (like A2A datasets) for training AV-S2ST systems by extracting units from the audio stream while utilizing the modality-agnostic properties of models like AV-HuBERT [Choi *et al*., 2024]. Models like TransFace [Cheng *et al*., 2024] and AV2AV [Choi *et al*., 2024] are explicitly built upon this unit-based translation architecture. A comparison between the main SSL models and their input modalities is shown in Table 1, discriminating their primary pre-training objectives and what they are mainly used for within the reviewed literature.

3 SLR - Methodology

This systematic literature review was conducted following a defined protocol based on established guidelines for systematic literature reviews [Kitchenham and Charters, 2007]. The methodology ensured a structured and reproducible approach for identifying, evaluating, and synthesizing research focused on end-to-end multilingual Audio-Visual Speech-to-Speech Translation (AV-S2ST) systems. To facilitate the management of this complex process, enhance transparency, and ensure adherence to the protocol, the web-based tool Parsifal[1] was utilized throughout the review stages. The use of dedicated tools like Parsifal is recommended in SLR methodologies to support tasks such as study screening, data extraction, and quality assessment, thereby reducing manual burden and potential bias [Shintaku *et al*., 2016; Felizardo *et al*., 2012; Marshall and Brereton, 2015]. This structured tool support is particularly beneficial for reviews assessing complex architectures and generalization across studies [Inaguma *et al*., 2019]. The article selection criteria and steps are shown in Figure 3.

3.1 Scope Definition and Research Questions

The review's scope was centered on recent advancements in direct AV-S2ST. The primary research question guiding this review was:

SLRQ1: *How can multilingual audio-visual speech-to-speech translation (AV-S2ST) models effectively integrate audio-visual modalities to improve translation accuracy and robustness across multiple languages, while ensuring precise lip synchronization, even considering the scarcity of datasets and noisy environments?*

[1] See https://parsif.al/ for more information on Parsifal.

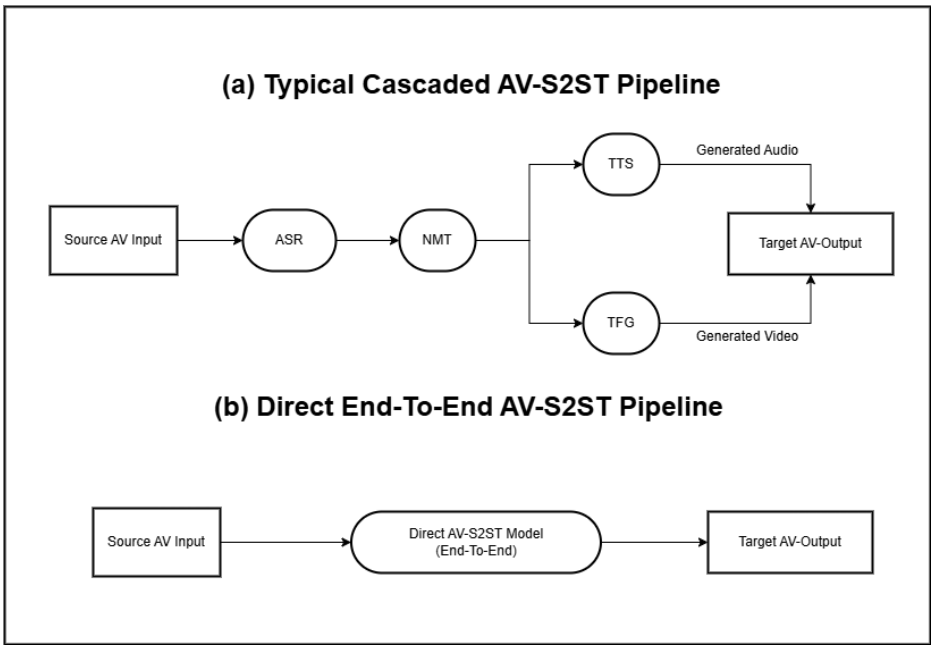

Figure 1. Comparison of AV-S2ST Architectures

Secondary questions addressed techniques for handling data scarcity and noise, and the specific evaluation metrics suitable for AV-S2ST:

SLRQ2: *Which techniques improve the translation of AV-S2ST models given of a lack of large-scale multilingual audiovisual datasets, including noisy environments, and how does visual information contribute to robustness?*

SLRQ3: *Which metrics are used to evaluate AV-S2ST models, and how do they balance translation accuracy (BLEU) with audiovisual synchronization (LSE-D, MOS)?*

Parsifal was used to document these questions and the associated PICOC (Population, Intervention, Comparison, Outcome, and Context) framework elements during the planning phase.

3.2 Study Selection Criteria

The selection of literature was governed by the following specific inclusion criteria:

I.1. Does the article analyze how visual information contributes to the intelligibility and perceptual quality of the translated speech?

I.2. Does the article propose new methods or improvements for end-to-end AV-S2ST models?

I.3. Does the article address specific challenges in translating multiple languages, such as cross-lingual interference or zero-shot learning?

I.4. Does the article address data limitations and test methodologies to improve performance in low-resource scenarios?

I.5. Does the article investigate how different audiovisual fusion strategies affect translation accuracy and naturalness?

I.6. Does the article propose or evaluate new metrics to balance translation fidelity and audiovisual synchronization?

The following exclusion criteria were also applied consistently during the screening phases within the tool:

E.1. Does the article emphasize only one language or not discuss strategies for multilingual translation?

E.2. Does the article focus on speech translation but without considering audiovisual integration?

E.3. Does the article ignore important perceptual aspects, such as the realism of lip synchronization or combined audiovisual quality?

E.4. Does the article not test methods that can be generalized to AV-S2ST models or not discuss implications for practical cases?

E.5. Does the article deal only with speech recognition or synthesis, without addressing translation between languages?

E.6. Does the article use standard translation or synthesis metrics without adapting the evaluation for an audiovisual context?

3.3 Search Strategy and Selection Process

The literature search utilized major scientific databases (ACM Digital Library, Elsevier ScienceDirect, IEEE Xplore) and ArXiv, for more recent articles, as presented in Table 2. Search results were imported into Parsifal for screening. The search covered publications in English from 2019 to 2025.

To ensure comprehensive coverage of the relevant literature within the defined scope, a systematic search strategy was implemented using a combination of primary keywords

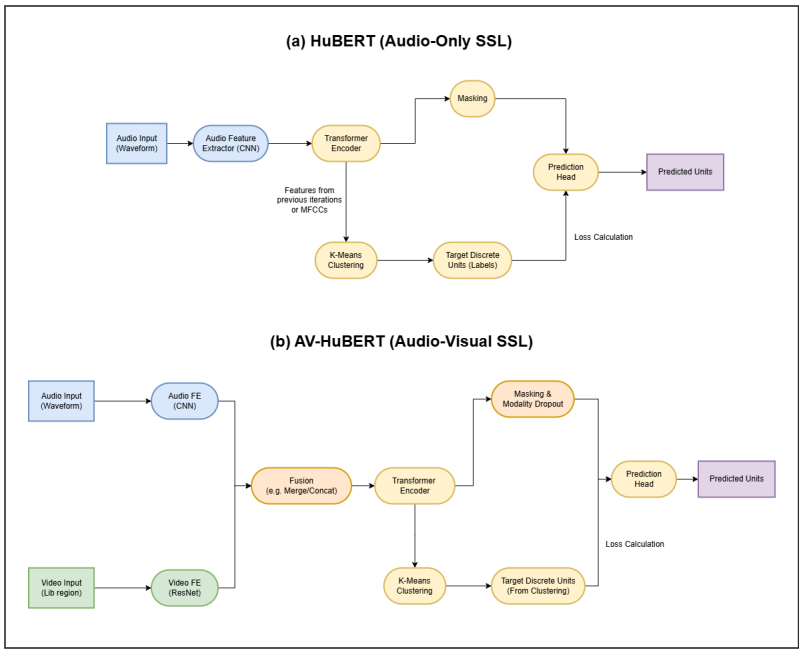

Figure 2. Core Mechanisms of HuBERT and AV-HuBERT

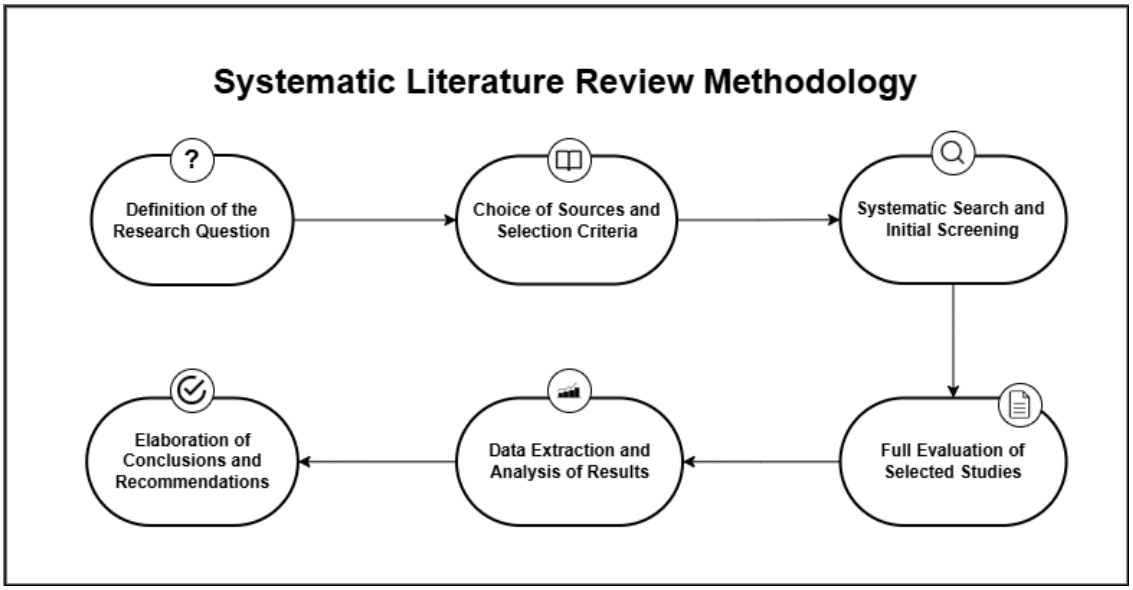

Figure 3. Step-by-Step Approach to Articles Selection and Analysis

Sources	URLs
ACM Digital Library	https://dl.acm.org/
Elsevier ScienceDirect	https://www.sciencedirect.com/
IEEE Xplore	https://ieeexplore.ieee.org/
ArXiv	https://arxiv.org/

Table 2. Database/Search Engines

and structured search strings derived from the research questions and core concepts of AV-S2ST, multilinguality, multimodality, and specific technical challenges. Figure 4 outlines the principal keywords used in English, along with an example of the Boolean search logic employed across the targeted databases. The search string is sorted to the most recently included, based on relevant articles selected. The search strings were applied to the title, abstract, and keyword fields of the publications within the selected databases.

An initial pool of 180 articles was selected. The selection made is represented in Figure 5, with the stages: Title and abstract screening, reducing to 80; abstract and conclusion review, reducing to 40; and full-text assessment, reducing to 20. The final set included six primary studies focusing directly on AV-S2ST: [Choi *et al.*, 2024], [Goncalves *et al.*, 2025], [Huang *et al.*, 2023], [Cheng *et al.*, 2023], [Cheng *et al.*, 2024], and [Han *et al.*, 2024]. Additionally, three foun-

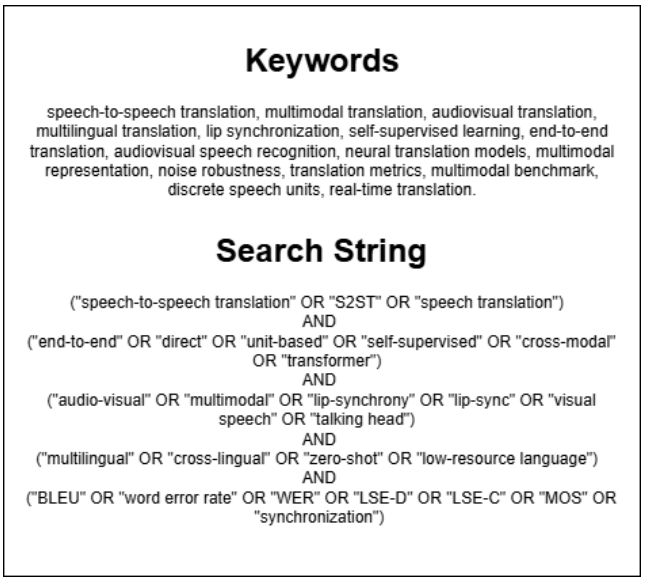

Figure 4. Keywords And Search String

dational audio-only S2ST studies were included to provide essential context on direct translation paradigms, multilingual strategies, and data augmentation techniques that influenced the AV-S2ST field: Di Gangi *et al*. [2019], Inaguma *et al*. [2019], Dong *et al*. [2022].

3.4 Data Extraction and Synthesis

A predefined data extraction form guided the systematic collection of information from the nine selected studies. Extracted data included model architectural techniques, datasets, metrics, results, challenges, and limitations. The extracted data facilitated a qualitative synthesis using thematic analysis to compare approaches, identify trends, and consolidate findings related to the research questions, as presented in Section 4.

4 Results and Analysis

This section synthesizes the primary findings from the systematic review of the nine final selected articles. The primary studies selected for this review are summarized in Table 3, which outlines their core contributions.

The results highlight key architectural trends, representation learning strategies, methods for addressing core challenges such as data scarcity, synchronization, and noise robustness, and the performance evaluation landscape within the reviewed literature.

Analysis of the selected studies reveals convergence towards specific techniques and architectural choices for direct AV-S2ST. We can have a clearer look at the objectives of each model and its accepted inputs and processed outputs in Table 4.

4.1 End-to-End Frameworks and Unit-Based Translation

The reviewed literature strongly favors end-to-end sequence-to-sequence architectures, moving away from traditional cascaded systems. These models often utilize Transformer or Conformer backbones [Di Gangi *et al*., 2019; Huang *et al*., 2023; Dong *et al*., 2022]. A central technique enabling textless translation is the use of discrete units derived from SSL models, primarily HuBERT or its multilingual variants like mHuBERT [Huang *et al*., 2023; Cheng *et al*., 2024; Choi *et al*., 2024]. These units serve as intermediate targets, simplifying the translation process. For instance, AV2AV [Choi *et al*., 2024] employs a unit-to-unit translation framework operating on units from a pre-trained mAV-HuBERT. Similarly, TransFace [Cheng *et al*., 2024] uses mHuBERT units to feed its parallel audio-visual synthesizer (Unit2Lip). This unit-based paradigm is foundational to several of the state-of-the-art direct models reviewed [Choi *et al*., 2024; Cheng *et al*., 2024; Huang *et al*., 2023].

4.2 Multimodal Fusion and Representation Learning

Integrating audio and visual streams effectively is the key technique for cutting-edge models. The principles of AV-HuBERT, focusing on masked prediction over fused modalities, are also mostly influential [Huang *et al*., 2023; Cheng *et al*., 2023; Han *et al*., 2024; Choi *et al*., 2024]. Key strategies identified include:

- **Unified AV Representations:** Learning modality-agnostic features, often through pre-training with modality dropout, allows models like AV2AV [Choi *et al*., 2024] to be trained using audio-only derived units but perform inference with AV inputs.
- **Cross-Modal Learning & Distillation:** Techniques like stream mixup [Cheng *et al*., 2023] interpolate AV features to bridge the modality gap, while AV-TranSpeech [Huang *et al*., 2023] uses cross-modal distillation, transferring knowledge from audio-only S2ST models to improve AV-S2ST performance, particularly beneficial when AV data is scarce.
- **Audio-Enhanced AV Pre-training:** XLAVS-R [Han *et al*., 2024] demonstrates the effectiveness of leveraging massive audio-only multilingual SSL pre-training (using XLS-R) before injecting visual components and continuing with AV self-supervised learning.

The techniques and approaches taken for each model analyzed in this study are summarized in Table 5.

4.3 Multilingual Strategies

Addressing translation across multiple languages is basis strategy for several studies:

- **Large-Scale Cross-Lingual Models:** XLAVS-R [Han *et al*., 2024] explicitly targets cross-lingual AV representation learning for over 100 languages, achieving strong zero-shot transfer capabilities.
- **Target Forcing Adaptations:** For direct speech input, simply prepending a language token is less effective than in NMT. Di Gangi *et al*. [2019] explored modifications like merging learnable language embeddings with the entire input feature sequence to provide stronger language conditioning.
- **Shared Architectures:** Using shared encoders/decoders for multiple languages, as explored in early multilingual S2ST work [Di Gangi *et al*., 2019; Inaguma *et al*., 2019], promotes knowledge transfer and can improve performance on lower-resource language pairs,

Year	Title	Authors	Proposes new AV-S2ST methods?	Studies audiovisual fusion?	Addresses multilingual challenges?	Proposes or evaluates metrics?	Analyzes visual contribution?	Addresses data limitations?
2023	AV-TranSpeech: Audio-Visual Robust Speech-to-Speech Translation	Huang *et al*. [2023]	×	×	×	×	×	×
2023	MixSpeech: Cross-Modality Self-Learning with Audio-Visual Stream Mixup for Visual Speech Translation and Recognition	Cheng *et al*. [2023]	×	×	×	×	×	×
2024	XLAVS-R: Cross-Lingual Audio-Visual Speech Representation Learning for Noise-Robust Speech Perception	Han *et al*. [2024]	×	×	×	×	×	×
2024	AV2AV: Direct Audio-Visual Speech to Audio-Visual Speech Translation with Unified Audio-Visual Speech Representation	Choi *et al*. [2024]	×	×			×	×
2024	TransFace: Unit-Based Audio-Visual Speech Synthesizer for Talking Head Translation	Cheng *et al*. [2024]	×	×			×	×
2025	Improving Lip-synchrony in Direct Audio-Visual Speech-to-Speech Translation	Goncalves *et al*. [2025]	×	×		×	×	

Table 3. Analysis of Audio-Visual Speech Translation Methods

Model	Objective	Input	Output
AV-TranSpeech Huang *et al*. [2023]	Robust Audio-Visual Speech-to-Speech Translation (AV-S2ST) without intermediate text, focusing on noise robustness.	Audio-Visual	Audio
MixSpeech Cheng *et al*. [2023]	Cross-modality self-learning (Audio → Visual) for Visual Speech Translation (V2T) and Recognition (VSR).	Audio & Visual & Audio-Visual	Text
XLAVS-R Han *et al*. [2024]	Cross-lingual AV speech representation learning for noise-robust AVSR and AVS2TT in multiple languages (>100).	Audio-Visual	Text
AV2AV Choi *et al*. [2024]	Direct Audio-Visual Speech to Audio-Visual Speech Translation (AV2AV) with unified representation.	Audio-Visual	Audio-Visual
TransFace (Unit2Lip) Cheng *et al*. [2024]	Direct Talking Head Translation (AV-S2ST) using discrete units and parallel AV synthesis for speed and isochrony.	Audio & Audio-Visual	Audio-Visual
Lip-Sync Loss Integration Goncalves *et al*. [2025]	Improving lip-synchrony in Direct Audio-Visual Speech-to-Speech Translation (AVS2S) without degrading quality.	Audio-Visual	Audio

Table 4. Overview of Models and Objectives from the Articles

Model	SSL Base Model / Approach	Uses Discrete Units?	Fusion Method / Modality Handling	Multilingual Approach
AV-TranSpeech Huang *et al*. [2023]	AV-HuBERT (Features/Init)	Yes	Cross-Modal Distillation (from Audio-only S2ST)	No (Focus: Bilingual)
MixSpeech Cheng *et al*. [2023]	AV-HuBERT (Encoder Features)	No (Direct V2T/VSR)	Audio-Visual Stream Mixup with Curriculum Learning	Yes (Evaluated on AVMuST-TED)
XLAVS-R Han *et al*. [2024]	XLS-R (Audio-only) followed by AV adaptation (AV-HuBERT objective)	Yes (Targets for Pre-training)	Visual Modality Injection; Learned Audio Features; Single-round AV training	Yes (>100 Languages; Cross-lingual Repr.)
AV2AV Choi *et al*. [2024]	mAV-HuBERT (Custom Multilingual AV Pre-training)	Yes	Unified AV Representation (Implicit Fusion in mAV-HuBERT)	Yes (Many-to-Many Unit Translation)
TransFace Cheng *et al*. [2024]	HuBERT / mHuBERT (for Unit Generation)	Yes	N/A (Focus on Unit-based AV *Synthesis*)	No (Focus: Synthesis)
Improving Lip-Synchrony Goncalves *et al*. [2025]	N/A (Focus on Loss Integration)	No	Early Fusion (Audio-Visual Concat))	No (Focus: Sync)

Table 5. Summary of Architectural and Representation Learning Techniques

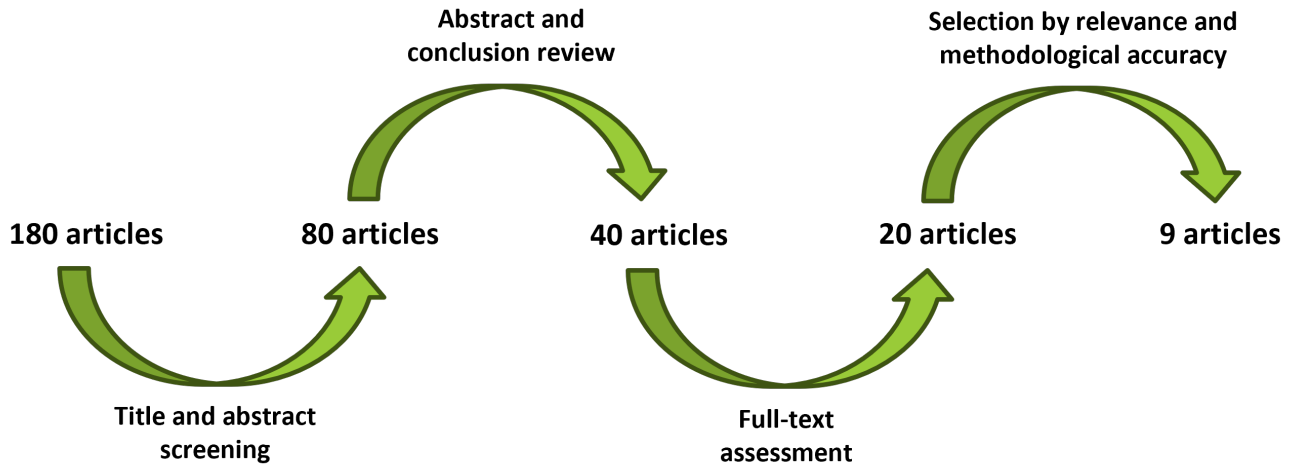

Figure 5. Articles Selection

though managing inter-language interference is a challenge [Inaguma *et al*., 2019].

4.4 Lip Synchrony and Duration Control

Achieving high perceptual quality in the generated audiovisual output requires precise control over two critical temporal aspects: lip synchrony and translation duration . Ensuring that the synthesized lip movements align perfectly with the translated audio (lip-synchrony) and that the output video's length matches the source (isometric translation) is essential for realism. The reviewed literature presents several targeted architectural and training strategies to address these challenges directly.

- **Explicit Lip-Sync Loss:** To directly optimize visual alignment, Goncalves *et al*. [2025] successfully integrated a lip-synchrony loss function based on SyncNet into the training loop, specifically targeting the duration predictor module. This significantly reduced lip-sync error (LSE-D by 9.2%) without compromising translation quality or audio naturalness.
- **Isometric Translation:** TransFace [Cheng *et al*., 2024] addresses the need for output video duration to match the input (critical for dubbing) by introducing a Bounded Duration Predictor. This module normalizes predicted unit durations to precisely match a target length, achieving 100% length compliance in their experiments.
- **Parallel Synthesis Architecture:** Models generating audio and video streams concurrently from a shared representation (like the discrete units used by AV2AV [Choi *et al*., 2024] and TransFace [Cheng *et al*., 2024]) inherently facilitate synchronization, contrasting with cascaded approaches where timing mismatches can accumulate.

4.5 Data Scarcity

Handling data scarcity for AV models is a great obstacle to effective training, and studies have diverse approaches to overcome these barriers. In the table below (Table 6) we can quickly identify the main datasets used by each model (and referred article) and key reported results.

- **Pseudo-Labeling:** Dong *et al*. [2022] provided a systematic study demonstrating that large volumes of pseudo-labeled data (generated via ASR->MT->TTS) can significantly improve direct S2ST when used for pre-training and specialized fine-tuning strategies like mixed-tuning or prompt-tuning (yielding up to +11.6 BLEU).
- **Leveraging Audio-Only Corpora:** Multiple studies effectively utilize abundant audio-only parallel data (A2A) or SSL corpora. AV2AV [Choi *et al*., 2024] trains its core unit translator using units derived solely from audio inputs leveraging mAV-HuBERT's modality robustness. XLAVS-R [Han *et al*., 2024] bases its strong performance on extensive audio-only pretraining before AV adaptation. AV-TranSpeech [Huang *et al*., 2023] employs cross-modal distillation specifically to transfer knowledge from audio-only trained models.

4.6 Noise Robustness

The integration of visual information is consistently shown to improve robustness against acoustic noise.

- **AV-TranSpeech**: Huang *et al*. [2023] explicitly focuses on robustness, demonstrating substantial BLEU score improvements (e.g. +7 BLEU gain reported) compared to audio-only systems under various noise levels (babble, music, speech) and SNRs.
- **XLAVS-R**: Han *et al*. [2024] highlights noise robustness as a key outcome of its cross-lingual AV representations, reporting significant WER reductions (up to 18.5%) and BLEU gains (up to 4.7) on noisy AV test conditions within the MuAViC benchmark.
- **MixSpeech**: Cheng *et al*. [2023] reports enhanced robustness and BLEU score improvements (+1.4 to +4.2) in noisy scenarios on the AVMuST-TED dataset.

5 Performance Evaluation Overview

The evaluation methodologies for AV-S2ST, as detailed in Table 6, are inherently multifaceted, requiring a combination of metrics to assess translation quality, audiovisual synchronization, perceptual realism, and robustness. Translation performance is predominantly measured using BLEU scores,

where models like TransFace report scores as high as 61.93 on the LRS3-T dataset, and techniques like pseudo-labeling demonstrate significant gains of over 11 BLEU points against baselines.

Beyond translation accuracy, lip-synchrony is critically evaluated using LSE-D and LSE-C, with targeted loss functions showing marked improvements, such as a 9.2% reduction in distance error. The perceptual quality and realism of the generated video are further quantified by Mean Opinion Scores (MOS) and Fréchet Inception Distance (FID). Furthermore, a significant focus is placed on noise robustness, where audio-visual models consistently demonstrate superiority over audio-only systems, with one study showing a BLEU score of 23.9 versus 0.1 in a -10dB noise condition. This comprehensive evaluation approach underscores the field's focus not only on linguistic accuracy but also on the critical audiovisual aspects that define the user experience.

5.1 Datasets

Commonly used corpora include LRS2/3 [Afouras *et al.*, 2018a,b], VoxCeleb2 [Chung *et al.*, 2018], MuAViC [Anwar *et al.*, 2023], LRS3-T [Huang *et al.*, 2023], AVMUST-TED [Cheng *et al.*, 2023], TedEn2Zh [Liu *et al.*, 2019], CVSS-C [Jia *et al.*, 2022], alongside SSL corpora like AVSpeech [Ephrat *et al.*, 2018], mTEDx [Salesky *et al.*, 2021], and VoxPopuli [Wang *et al.*, 2021]. Noise data sets such as MUSAN [Snyder *et al.*, 2015] are frequently used for robustness tests.

5.2 Metrics

A combination of metrics is typically reported in the researched papers:

- Translation Quality: **ASR-BLEU** [Papineni *et al.*, 2002; Post, 2018] is the standard metric, measuring quality by comparing the ASR transcription of the translated audio against reference text. It is calculated as $BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$, where BP is a brevity penalty and p_n is the modified n-gram precision.
- Audiovisual Synchronization: Lip-sync error is measured using **LSE-D (Distance)**, the Euclidean distance between audio and lip movement embeddings, and **LSE-C (Confidence)** [Prajwal *et al.*, 2020], the confidence score from a pretrained lip-sync expert model. For both metrics, lower values indicate better synchronization.
- Quality: **MOS** for perceptual evaluation (naturalness, image fidelity, overall quality).
- Isometry: **Length Ratio (LR)** and **Length Compliance (LC)** [Cheng *et al.*, 2024].

The combined use of these diverse metrics provides a holistic evaluation of AV-S2ST systems, reflecting the multifaceted nature of the task. While ASR-BLEU remains the standard for assessing linguistic accuracy, the widespread adoption of specialized metrics such as LSE-D and LSE-C for synchronization, MOS for perceptual quality, and LC for isometry is particularly revealing. This comprehensive evaluation framework demonstrates the field's maturity; progress is no longer measured by translation fidelity alone, but also by how well models address the specific multimodal challenges that define high-quality audio-visual translation. These results, therefore, underscore the significant advancements in direct AV-S2ST, driven by SSL and unit-based methods that are increasingly validated against this full suite of targeted metrics.

6 State of Art and Limitations

The analysis of the selected studies also several limitations and challenges, which in turn suggest important directions for future research. These limitations are brought together in Table 7.

Some of the observed limitations revolves around data availability and quality. The significant scarcity of large-scale, parallel AV-S2ST corpora remains a major impediment, often compelling researchers to rely on audio-only datasets combined with modality-agnostic representations [Choi *et al.*, 2024], synthesized data, or pseudo-labeling techniques [Dong *et al.*, 2022]. While strategies like leveraging extensive audio-only pre-training [Han *et al.*, 2024] or cross-modal distillation [Huang *et al.*, 2023] show promise, the performance ceiling may still be constrained by the lack of genuine parallel AV data. Furthermore, the AV data available for pre-training covers fewer languages compared to audio-only resources [Han *et al.*, 2024; Choi *et al.*, 2024], and the quality of data generated by external pipelines (ASR, MT, TTS) for pseudo-labeling can introduce noise or bias [Dong *et al.*, 2022].

Evaluation methodologies also present challenges. The prevalent use of ASR-BLEU for assessing translation quality is heavily dependent on the accuracy of the ASR system, which can be unreliable for synthesized speech or underrepresented languages, potentially masking true translation performance [Huang *et al.*, 2023; Han *et al.*, 2024; Cheng *et al.*, 2023]. Moreover, evaluation is frequently confined to language pairs involving English (e.g., X-En, En-X), limiting insights into performance across a wider linguistic spectrum [Huang *et al.*, 2023; Han *et al.*, 2024; Di Gangi *et al.*, 2019; Inaguma *et al.*, 2019].

Regarding modeling, scaling end-to-end multilingual models to surround a vast number of languages effectively while managing computational resources (GPU memory, training time) remains a significant hurdle [Han *et al.*, 2024; Choi *et al.*, 2024; Di Gangi *et al.*, 2019; Inaguma *et al.*, 2019]. Ensuring model robustness not just against controlled noise types but also across diverse real-world acoustic and visual conditions, different domains, and speaker variations necessitates further research [Huang *et al.*, 2023; Han *et al.*, 2024; Cheng *et al.*, 2023].

In addition, critical aspects of output quality, specifically synchrony and temporal consistency, present ongoing challenges. While techniques like explicit sync losses [Goncalves *et al.*, 2025] and bounded duration predictors [Cheng *et al.*, 2024] address these directly, achieving a perfect balance between precise lip-synchrony, strict isometry (duration matching), translation fidelity, and natural speech prosody is complex [Goncalves *et al.*, 2025; Cheng *et al.*, 2024]. The visual realism of generated faces, particularly in zero-shot speaker scenarios, also warrants continued im-

Article Title (Abbreviated) / Key Model	Main Datasets Used	Key Reported Results (Examples)
AV2AV: Direct Audio-Visual Speech to Audio-Visual Speech Translation Choi *et al*. [2024]	LRS2/3, VoxCeleb2, mTEDx, AVSpeech (for mAV-HuBERT pre-train); VoxPopuli, mTEDx (for A2A train); MuAViC, LRS3-T (for Fine-tune/Eval.)	AV2AV (MuAViC X-En): 26.57 BLEU (Es-En), 31.27 BLEU (Fr-En); AV2A (LRS3-T En-Es 200h): 50.9 BLEU; AV-Renderer (LRS3): 7.91 LSE-C, 6.65 LSE-D, 3.18 FID
AV-TranSpeech: Audio-Visual Robust Speech-to-Speech Translation Huang *et al*. [2023]	LRS3-T (Eval.), CVSS-C (S2ST Train), MUSAN (Noise)	AV-S2ST (LRS3-T En-Es): 45.2 BLEU / 3.82 MOS (Clean); Substantial improvement vs. A-only in noise (e.g., 23.9 vs 0.1 BLEU at -10dB); Low-Resource (10h AV + CoVoST): +7.6 BLEU (avg.)
Improving Lip-Synchrony in Direct AV S2ST Goncalves *et al*. [2025]	LRS3 (Train/Eval.)	Avg. 4 pairs: 10.67 LSE-D (9.2% improvement), 2.72 LSE-C; No degradation in PESQ, BLASER, ASR-BLEU.
TransFace: Unit-Based Audio-Visual Speech Synthesizer Cheng *et al*. [2024]	LRS2 (Unit2Lip Train), LRS3-T (Translation Eval.)	Unit2Lip (LRS2): 7.19 LSE-C / 7.34 LSE-D (vs. original audio), 5.14 FID, x4.35 speedup vs. U2S+Wav2Lip; TransFace (LRS3-T): 61.93 BLEU (Es-En), 47.55 BLEU (Fr-En); Bounded: 100% LC
Leveraging Pseudo-labeled Data to Improve Direct S2ST Dong *et al*. [2022]	Fisher Spanish, TedEn2Zh (Primary); MLS.Es, Giga-Sub (Secondary/PTL)	Fisher (Es-En): 43.6 BLEU (+11.6 vs baseline); TedEn2Zh (En-Zh): 20.8 BLEU (+9.6 vs baseline); MOS: +0.19 (Fisher), +0.61 (TedEn2Zh) with PTL.
XLAVS-R: Cross-Lingual AV Speech Representation Learning Han *et al*. [2024]	MuAViC (AV Pre-train/Fine-tune/Eval.); ~436K hrs Audio-Only (A Pre-train); FLEURS (OOD Eval.); MUSAN (Noise)	AVSR (Avg. 9 lang., Noisy, AV): 37.3% WER (XLAVS-R 2B), up to 18.5% WER impr. vs baseline; AVS2TT (Avg. 6 X-En, Noisy, AV): 18.7 BLEU (XLAVS-R 2B), up to 4.7 BLEU impr.
MixSpeech: Cross-Modality Self-Learning Cheng *et al*. [2023]	AVMUST-TED (Proposed), LRS2, LRS3, CMLR	V2T (AVMUST-TED): +1.4 to +4.2 BLEU (avg.); VSR (LRS2): 25.5% WER; VSR (LRS3): 28.0% WER; VSR (CMLR): 11.1% WER.
One-to-Many Multilingual End-to-End Speech Translation Di Gangi *et al*. [2019]	MuST-C	SLT (MuST-C, M-Pre+ASR): Impr. > 1 BLEU in Nl, It, Pt vs baseline; Comparable perf. in De, Es, Fr; Langdetect: >95% correct lang.
Multilingual End-to-End Speech Translation Inaguma *et al*. [2019]	MuST-C	Comparable performance to cascade in high-resource settings; Improved performance in low-resource settings.

Table 6. Comparison of Main Datasets Used and Key Reported Results (Examples) from Reviewed Studies

provement [Choi *et al*., 2024; Cheng *et al*., 2024].

These systems often depend heavily on the quality of the underlying pre-trained SSL models [Choi *et al*., 2024; Huang *et al*., 2023; Cheng *et al*., 2024; Han *et al*., 2024; Cheng *et al*., 2023], and ethical considerations regarding facial data privacy and the potential for misuse (e.g., deepfakes) must be continually addressed [Goncalves *et al*., 2025].

7 Future Research

The limitations outlined above naturally guide future research efforts in the field. A critical need exists for the creation and expansion of diverse, large-scale parallel AV-S2ST datasets, covering a wider range of domains and languages (including low-resource ones such as specific variants of Portuguese). Continued exploration of data augmentation, mining techniques, and refinement of pseudo-labeling strategies [Dong *et al*., 2022] will remain vital in the interim. Establishing language-specific benchmarks, for example, for different Portuguese accents under various noise conditions, would also facilitate more targeted evaluation and development.

Further research is needed to refine synchronization and duration control mechanisms. This includes developing models that can dynamically balance isometry with natural speech rhythm (addressing limitations of methods like those in Cheng *et al*. [2024]) and enhancing lip-sync loss functions or architectural components to work robustly across speakers and languages without degrading translation quality [Goncalves *et al*., 2025]. Fine-tuning synchrony models in specific accents is also a viable direction.

Improvements in modeling should focus on enhancing multilingual scalability and efficiency (addressing challenges noted by Han *et al*. [2024]; Di Gangi *et al*. [2019]; Inaguma *et al*. [2019]), improving zero-shot cross-lingual performance, and increasing robustness to real-world acoustic and visual variability beyond benchmark noise conditions [Huang *et al*., 2023; Han *et al*., 2024]. Investigating innovative architectures that better leverage the complementary nature of AV information is ongoing.

Developing more reliable evaluation metrics, particularly automatic metrics for AV synchrony and perceptual quality that are less dependent on ASR [Huang *et al*., 2023; Han *et al*., 2024], is essential to push progress. Enhancing the visual realism of synthesized talking heads, especially preserving speaker expressiveness in zero-shot scenarios [Choi *et al*., 2024; Cheng *et al*., 2024], remains important. Lastly, continuous attention must be paid to ethical considerations, developing safeguards against misuse, and ensuring data privacy as these powerful generation technologies mature.

8 Discussion

This section revisits the secondary research questions posed in Section 3.1 to provide direct answers based on the findings synthesized from the selected literature. By consolidating the results, we demonstrate how the systematic literature review fulfilled its objectives in clarifying the state of the art in AV-S2ST.

Limitation Category	Description of Limitation
Data	
	• Lack of large parallel datasets for AV2AV or AV-S2ST, necessitating the use of A2A, synthesized, or pseudo-labeled data [Choi *et al*., 2024; Dong *et al*., 2022]. • Quality of data generated by external pipelines (ASR, MT, TTS) for pseudo-labeling or evaluation can limit final performance [Dong *et al*., 2022]. • AV data for pre-training is less abundant and covers fewer languages than audio-only data [Han *et al*., 2024; Choi *et al*., 2024].
Evaluation	
	• Dependency on ASR quality for calculating BLEU in speech translation tasks, which can be imprecise, especially for synthesized speech or non-English languages [Huang *et al*., 2023; Han *et al*., 2024]. • Evaluation frequently limited to translations to/from English (X-En, En-X) due to data and ASR tool restrictions [Huang *et al*., 2023; Han *et al*., 2024; Di Gangi *et al*., 2019].
Modeling	
	• Multilingual models can face difficulty scaling to many languages effectively [Han *et al*., 2024; Di Gangi *et al*., 2019]. • High computational and GPU memory costs for training and running large AV models, especially Transformers [Choi *et al*., 2024; Han *et al*., 2024]. • Robustness to diverse noise types and generalization to varied domains require more investigation [Huang *et al*., 2023; Han *et al*., 2024; Cheng *et al*., 2023].
Synchronization, Isotropy, Quality	
	• Forced isotropy (to avoid visual artifacts) can lead to unnatural speech rhythm (accelerated/slowed) if translation length differs significantly [Cheng *et al*., 2024]. • Balancing lip-sync optimization with translation quality and speech naturalness remains a challenge [Goncalves *et al*., 2025]. • Visual generation quality (realism of lips/face), especially in zero-shot scenarios, can still be improved [Choi *et al*., 2024; Cheng *et al*., 2024].
Dependencies and Methods	
	• Performance is dependent on the quality of representations/units from pre-trained SSL models. • Impact of external textual data on lip translation (V2T) not explored in some works focused on direct AV interaction [Cheng *et al*., 2023].
Ethics and Use	
	• Privacy concerns due to the use of visual data (faces). • Potential for misuse of technology to create deepfakes or generate non-consensual content [Goncalves *et al*., 2025].

Table 7. Summary of Identified Limitations

SLRQ2: Which techniques improve the translation of AV-S2ST models given a lack of large-scale multilingual audiovisual datasets, including noisy environments, and how does visual information contribute to robustness?

The review identified two primary strategies to mitigate data scarcity. The first is the effective use of pseudo-labeling, where large volumes of unlabeled data are used to pre-train models, yielding significant improvements in BLEU scores. The second, and more prevalent, strategy is leveraging large, existing audio-only corpora. This is achieved through techniques like cross-modal distillation, where knowledge from robust audio-only models is transferred to the audio-visual model , and through the use of powerful pre-trained audio SSL models as a foundation before adapting them with visual data. Regarding robustness, the integration of visual information consistently proves to be a key factor. Studies explicitly demonstrate that the visual stream (lip movements) provides complementary information that makes the models more resilient to acoustic noise . This is quantitatively supported by significant gains in BLEU scores and reductions in WER in noisy conditions when compared to their audio-only counterparts, confirming that visual cues are integral to achieving robustness.

SLRQ3: Which metrics are used to evaluate AV-S2ST models, and how do they balance translation accuracy (BLEU) with audiovisual synchronization (LSE-D, MOS)?

The evaluation of AV-S2ST models is inherently multifaceted, requiring a suite of metrics that go beyond simple translation accuracy. While ASR-BLEU remains the standard for assessing linguistic quality, it is almost always complemented by specialized metrics to address the multimodal nature of the task. The balance between accuracy and synchronization is achieved by jointly reporting on:

- **Audiovisual Synchronization:** Measured by LSE-D (Distance) and LSE-C (Confidence), which quantify the alignment between lip movements and the generated audio.
- **Perceptual Quality:** Assessed through Mean Opinion Scores (MOS), where human evaluators rate aspects like naturalness and overall quality.
- **Temporal Consistency:** Evaluated using metrics like Length Compliance (LC) to ensure the output duration matches the source, a critical factor for applications like dubbing.

This combined evaluation framework demonstrates that the field prioritizes not only what is being said (translation accuracy) but also how it is being presented (synchronization and realism).

9 Threats to Validity

This section addresses the potential threats to the validity of our study. Acknowledging these limitations is crucial for contextualizing the findings and ensuring transparency in the research process. The discussion is structured around four primary categories of validity threats: construct, internal, external, and conclusion validity, with a description of the mitigation strategies employed for each.

9.1 Construct Validity

Construct validity refers to the alignment between the theoretical constructs of the research questions and the specific operational measures used, primarily the search strategy. A potential threat in this regard is the formulation of a search string that may not fully capture all relevant literature, potentially omitting pertinent studies due to keyword selection. To mitigate this threat, the search query was systematically developed based on an initial set of keywords derived from primary articles in the AV-S2ST field. This initial string was then iteratively refined throughout the screening process to enhance its comprehensiveness. Furthermore, a wide range of synonyms and related terms was used within boolean logic to maximize coverage across different terminologies used in the literature, as detailed in Figure 4.

9.2 Internal Validity

Internal validity concerns potential biases introduced by the researchers during the review execution, particularly in the study selection and data extraction phases. A primary threat is the risk of subjective judgment leading to inconsistent application of the selection criteria. This threat was mitigated by establishing strict and explicit inclusion and exclusion criteria prior to the screening process. The entire review workflow was managed and documented using the Parsifal web-based tool, which enforces a systematic and traceable process, thus minimizing the potential for researcher bias and ensuring consistent adherence to the predefined protocol.

9.3 External Validity

External validity refers to the generalization of the conclusions drawn from this review to the broader field of AV-S2ST. The main threat to external validity lies in the scope of our search, which was limited to publications from 2019 to 2025 and a specific set of scientific databases. Consequently, our findings may not be fully generalizable to research published outside this time frame or in other venues. To mitigate this, our search strategy was designed to be comprehensive within its defined scope, including not only major peer-reviewed databases like IEEE Xplore and the ACM Digital Library but also the ArXiv pre-print server. This inclusion is critical for capturing the most recent and cutting-edge research in a rapidly advancing field like artificial intelligence.

9.4 Conclusion Validity

Conclusion validity addresses the degree to which the conclusions drawn are reasonable and supported by the data extracted. In a qualitative synthesis, such as this review, a potential threat arises from the possibility of misinterpreting the findings of the primary studies. To address this, a standardized data extraction form was utilized to ensure that information was collected consistently between all selected articles. The synthesis process itself was based on thematic analysis, focusing on identifying recurring patterns, techniques, and challenges across studies, rather than drawing conclu-

sions from a single source. The structured presentation of these findings in tables (e.g., Table 4, Table 5) further contributes to the transparency and reliability of the conclusions presented in this review.

10 Conclusion

This systematic literature review analyzed the current landscape of direct, end-to-end Audio-Visual Speech-to-Speech Translation (AV-S2ST) models, synthesizing architectural innovations and core techniques from studies published between 2019 and 2025. The findings reveal a decisive trend away from traditional cascaded pipelines towards unified, direct systems. This paradigm shift is critically enabled by two synergistic advancements: the use of self-supervised learning (SSL) models, such as AV-HuBERT, to learn robust multimodal representations, and the adoption of discrete units to facilitate textless translation .

The review highlights that the field is maturing beyond simple proof-of-concept, with researchers now developing targeted solutions for the core challenges of AV-S2ST, including dedicated mechanisms for improving lip-synchrony, ensuring isometric translation, and enhancing noise robustness. A key contribution of this synthesis is the confirmation that the visual stream is not merely supplementary but is integral to achieving higher-quality and more resilient speech translation.

While significant progress has been made, the continued evolution of AV-S2ST will depend on addressing the persistent challenges of data scarcity and the development of more comprehensive evaluation frameworks. Overcoming these hurdles will be crucial for transitioning these powerful technologies from research to robust, real-world deployment, ultimately enhancing the dissemination of audio-visual content across linguistic barriers.

Declarations

Authors' Contributions

Alexandre de Godoy Pereira was the primary author and main contributor to the writing, data collection, and analysis of this systematic review. Renato Cordeiro Ferreira and Alfredo Goldman provided critical supervision, conceptual guidance, and substantial reviews of the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of research artifacts

The data and other materials created and/or used in this literature review (search protocol, selection criteria, and extraction forms originally managed in Parsifal) are available in the project's GitHub repository: https://github.com/alegodoyp/A-Systematic-Review-on-Audio-Visual-Speech-To-Speech-Translation-Models.

Use of Artificial Intelligence

The authors used generative AI tools in the development of this work. Claude (Anthropic) was used to suggest keywords, to review the manuscript, and to assist with the formatting/typesetting of this paper in LaTeX. Gemini 3 (Google) was used to support the analysis of the selected articles and to review the text. All AI-assisted output was reviewed, verified, and edited by the authors, who take full responsibility for the entire content of this paper. These tools are not listed as authors.

Further relevant information

This is a secondary study (a systematic literature review) that does not involve human or animal subjects; therefore, ethics-committee approval was not applicable.

References

Afouras, T., Chung, J. S., and Zisserman, A. (2018a). Deep audio-visual speech recognition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 653–668. DOI: 10.1109/tpami.2018.2889052.

Afouras, T., Chung, J. S., and Zisserman, A. (2018b). LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*. DOI: 10.48550/arXiv.1809.00496.

Anwar, M., Han, H., Pino, J., Carpuat, M., Shi, B., and Wang, C. (2023). MuAViC: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation. *arXiv preprint arXiv:2305.15814*. DOI: 10.48550/arXiv.2303.00628.

Cheng, X., Huang, R., Li, L., Wang, Z., Jin, T., Yin, A., Feiyang, C., Duan, X., Huai, B., and Zhao, Z. (2024). TransFace: Unit-based audio-visual speech synthesizer for talking head translation. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9973–9986, Bangkok, Thailand. Association for Computational Linguistics. DOI: 10.18653/v1/2024.findings-acl.593.

Cheng, X., Li, L., Jin, T., Huang, R., Lin, W., Wang, Z., Liu, H., Wang, Y., Yin, A., and Zhao, Z. (2023). Mixspeech: Cross-modality self-learning with audio-visual stream mixup for visual speech translation and recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15689–15699. DOI: 10.1109/iccv51070.2023.01442.

Choi, J., Park, S.-J., Kim, M., and Ro, Y. M. (2024). AV2AV: Direct audio-visual speech to audio-visual speech translation with unified audio-visual speech representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.48550/arXiv.2312.02512.

Chung, J. S., Nagrani, A., and Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. In *Interspeech 2018*, pages 1086–1090. DOI: 10.21437/interspeech.2018-1929.

Di Gangi, M. A., Negri, M., and Turchi, M. (2019). One-to-many multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 585–592. IEEE. DOI: 10.1109/asru46091.2019.9004003.

Dong, Q., Yue, F., Ko, T., Wang, M., Bai, Q., and Zhang, Y. (2022). Leveraging Pseudo-labeled Data to Improve Direct Speech-to-Speech Translation. In *Interspeech 2022*, pages 1781–1785. DOI: 10.21437/Interspeech.2022-10011.

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., and Rubinstein, M. (2018). Looking to listen at the cocktail party: A speaker-

independent audio-visual model for speech separation. In *ACM SIGGRAPH 2018 Papers*, pages 1–11. DOI: 10.1145/3197517.3201357.

Felizardo, K. R., Mendes, E., Fagan, M., and Nakagawa, E. Y. (2012). Systematic literature review using parsifal. In *Proceedings of the 8th international conference on the quality of information and communications technology*, pages 363–368. Book.

Goncalves, L., Mathur, P., Niu, X., Lavania, C., Houston, B., Vishnubhotla, S., Sun, L., and Ferritto, A. (2025). Improving lip-synchrony in direct audio-visual speech-to-speech translation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. DOI: 10.1109/ICASSP49660.2025.10890684.

Gupta, M., Dutta, M., and Maurya, C. K. (2024). Direct Speech-to-Speech Neural Machine Translation: A Survey. *arXiv preprint arXiv:2401.14453*. DOI: 10.1016/j.specom.2025.103317.

Han, H., Anwar, M., Pino, J., Hsu, W.-N., Carpuat, M., Shi, B., and Wang, C. (2024). XLAVS-R: Cross-lingual audio-visual speech representation learning for noise-robust speech perception. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12896–12911, Bangkok, Thailand. Association for Computational Linguistics. DOI: 10.18653/v1/2024.acl-long.697.

Huang, R., Liu, H., Cheng, X., Ren, Y., Li, L., Ye, Z., He, J., Zhang, L., Liu, J., Yin, X., and Zhao, Z. (2023). AV-TranSpeech: Audio-visual robust speech-to-speech translation. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8590–8604, Toronto, Canada. Association for Computational Linguistics. DOI: 10.18653/v1/2023.acl-long.479.

Inaguma, H., Kiyono, S., Duh, K., Karita, S., Yalta, N., Hayashi, T., and Watanabe, S. (2019). Multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 570–577. IEEE. DOI: 10.1109/asru46091.2019.9003832.

Jia, Y., Zen, H., Rivera, J., Tran, T., Chiu, C.-K., Wu, Y., Wang, C.-I., Chen, N., Pino, J., and Wang, Y. (2022). Cvss: A large-scale multilingual speech-to-speech translation corpus. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 310–317. IEEE. DOI: 10.48550/arXiv.2201.03713.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Available at:`https://docs.edtechhub.org/lib/EDAG684W`.

Liu, Y., Luan, F., Wang, J., Wang, C., Xiao, T., and Zhu, J. (2019). End-to-end speech translation with knowledge distillation. In *Interspeech 2019*, pages 2020–2024. DOI: 10.21437/interspeech.2019-2582.

Marshall, C. and Brereton, P. (2015). Systematic literature reviews in software engineering–a tertiary study. *Information and Software Technology*, 64:1–13. DOI: 10.1016/j.infsof.2010.03.006.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. DOI: 10.3115/1073083.1073135.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics. Available at:`https://aclanthology.org/W18-6319`.

Prajwal, K., Mukhopadhyay, R., Namboodiri, V. P., and Jawahar, C. (2020). A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492. DOI: 10.1145/3394171.3413532.

Salesky, E., Lo, J., Miculicich, L., Chen, B., Sokolov, A., Williams, A., Lane, I., Pino, J., Cherry, C., and Neubig, G. (2021). mTEDx: A multilingual transcribed and translated speech corpus of TEDx talks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3711–3725, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2021.naacl-main.295.

Sarim, M., Shakeel, S., Javed, L., Jamaluddin, and Nadeem, M. (2025). Direct Speech to Speech Translation: A Review. *arXiv preprint arXiv:2503.04799*. DOI: 10.48550/arxiv.2503.04799.

Shintaku, K., de Souza, B., Durelli, V., Durelli, R., and Nakagawa, E. (2016). Assisting systematic literature review with a web-based text mining tool. In *2016 42th Latin American Computing Conference (CLEI)*, pages 1–10. IEEE. Book.

Snyder, D., Chen, G., and Povey, D. (2015). Musan: A music, speech, and noise corpus. In *arXiv preprint arXiv:1510.08484*. DOI: 10.48550/arxiv.1510.08484.

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., and Dupoux, E. (2021). VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 955–969, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2021.acl-long.75.