Audio-Visual Synchronization Analysis for Deepfake Detection: A Comprehensive Review

Halil Ibrahim Dursunoglu
Western Michigan University, Kalamazoo, MI 49006 USA
halilibrahim.dursunoglu@wmich.edu

ABSTRACT

The development of deepfake production, which has produced audio-visual frauds of unprecedented realism, has been fueled by recent innovations such as lip synchronizing, neural face reenactment, and synthetic speech. As forensic evidence, low-level visual artifacts and evident audio distortions are becoming less trustworthy, and the identification is becoming more reliant on the fundamentals of human speech, which is multimodal and has strong temporal, articulatory, and semantic links between lip movements and acoustic phonemes. Consequently, audio-visual integration offers a fair forensic indicator that is conceptually sound and mostly consistent across recording settings, languages, and speakers. The study explores the application of synchronicity-based methods to identify deepfakes in physiological coherence modeling, zero-shot semantic consistency analysis, transformer-based multimodal fusion, supervised and self-supervised deep learning, and traditional signal processing. We utilize syncnet-style lip-audio matching, AVTS and AV-HuBERT self-supervision, ASR-VSR semantic matching, cross-modal focus networks, and historical and current research in addition to novel lip-sync-specific sensors. genuine datasets, evaluation methods, robustness, real-world applications, and generalization patterns Logistical challenges also need to be taken into account.

Keywords — Audiovisual speech, Audio-visual synchronization, Deepfake detection, Lip-sync analysis, Multimodal forensics

I. INTRODUCTION

The recent advancement of technology in terms of generative modeling has led to the creation of very advanced models that can produce talking-head movies, including lip-sync, facial expressions, and speech simulation. As a result of the development in neural networks, diffusion models, and transformers, we can now simulate facial expression and mood, as well as generate voices based on text and lips based on sounds. The capability to do so using only traditional relics creates a very big threat to biometric security and media credibility, which is why the capabilities need to be used with great care. Recent research shows that deepfakes not only affect one but both audio and video data, and even that a modality by itself can be unreliable. [1, 2].

Multimodal is the production of sound and facial expressions in human speech. For instance, linguistic restrictions like bilabial plosives demand the mouth to close, linguistic constraints like the mouth opening width relates to the quantity of acoustic energy, and linguistic limitations like prosodic stress links to the head/face movement serve to combine into a single system. Reliable interaction with the surroundings is ensured by these limits. Contrariwise, generative pipelines ought to have a clear goal for producing or altering audio/video data to maximize perceptual realism but not any necessary biomechanics. Small contradictions accumulate in every generation process even if all approach seem reasonable independently.

As a technique for researching synchronicity, SyncNet is a lip-audio system that may be likened to embedding learning [3]. Future better performance may come from other models like Wav2Lip, hence demonstrating the contextual nature of synchronization and model choice [4]. Self-supervised models like AVTS and AV-HuBERT [5,6] demonstrated that large networks have the capacity to learn to identify cross-modal correspondence within unlabelled video data, hence generating universal representations. The aforementioned developments are the cause of replacing semantic and temporal feature detectors for artifact-level detector.

Other than artifacts, survey and benchmark works highlight the importance of synchronization-based signals for being more resilient to generator evolution and processing. The reason is that while artifacts may be easily removed using encoder-redecoder approaches, synchronization constraints are less vulnerable to this type of attack due to the fact that synchronization constraints rely on the structure of human speech production [1, 7, 8]. Experiments have shown that detectors developed for one class of generators do not necessarily perform well on another set of generators [9–11]. As a result, more emphasis is put on using transferable constraints, e.g., alignment, and semantic agreement.

In the context of ICASSP/ICIP, we review synchronization-based deepfake detectors combining signal processing, temporal models, and semantic alignment in a single framework.

II. DEEPFAKE GENERATION AND SYNCHRONIZATION FAILURE MODES

The reliability of audiovisual synchronization for use as forensic evidence is apparent from an analysis of how audiovisual content is created by modern deepfakes. While such deepfakes create highly realistic faces and lifelike voice synthesis, their design and goals place inherent constraints on them which often cause synchronization problems.

Deepfake systems can be categorized into visual synthesis, audio synthesis, and cross-modal conditioning. While most techniques maximize realism, identity maintenance, or perceptual accuracy, none explicitly enforce physical constraints on

articulation. Therefore, although they capture coordination dynamics at a macro level, deepfakes often fall short in replicating phoneme–viseme transition dynamics and other fine timing phenomena [1, 2]. **Facial manipulation.** Models based on autoencoders and GANs for face-swapping seek to alter identity while maintaining facial expressions, without any particular emphasis on mouth movement. Therefore, the mouth may suffer from geometric or temporal distortion with respect to the speech despite generating visually realistic faces [2, 7]. Reenactment methods like Face2Face place more emphasis on overall facial motion transfer, not lip sync per se; thus, there might be subtle discrepancies between lip movements and the spoken words [12]. **Neural lip synchronization.** Audio-conditioned lip-sync generators like Wav2Lip synthesize mouth motion using speech features [4].

TThey might turn out to be very precise on an individual frame-by-frame basis; however, these techniques will always be limited by the many-to-one correspondence between phonemes and visemes, reliance on context, and lack of capability to model coarticulation over an extended time interval. Experimental research has demonstrated that lip motion in some cases can turn out to be excessively smooth or delayed compared to the dynamics of the actual natural speech [4, 13, 14]. **Voice synthesis and voice cloning.** Models for synthesizing speech from the text and technologies used for cloning voice also can generate audio tracks, but they might also contain discrepancies concerning duration, stress, and pauses in relation to the actual movements of the mouth. This will happen no matter whether the synthesized voice or lip motion itself sounds realistic enough [5, 15, 16]. Multimodal audio-visual generation systems based on the diffusion technique are even more likely to have shared representations between audio and visual domains; however, the target here is realism, not biomechanical accuracy, hence mismatches will be inevitable [1, 14, 17]. **Bias in data sets.** The training dataset worsens the problem. Big web corpora frequently include post-production effects like dubbing, editing, and compression which generate benign mismatching. While training on these types of datasets, machine learning systems may pick up the approximate synchronizations and spread slight time errors. Nevertheless, spoken language follows strict physical coordination between the speech production mechanism and facial expressions, giving rise to consistent alignment features that cannot be easily faked. The continuous mismatches render synchronization-based techniques robust forensic tools [5, 6, 18].

III. RELATED WORK AND BACKGROUND

Three generations can be traced in research on deepfakes. In the first generation, the effort concentrated on observable visual artifacts and inconsistency cues in the spatial domain of early face-swapping GANs, such as boundary artifacts, color mismatches, and unusual local textures. Such cues worked well with early generators but did not generalize to artifacts of images processed using compression, resizing, or re-encoding procedures prevalent in social media. The second generation of solutions relied on deep classification approaches trained on large datasets, which indirectly learned to distinguish real from fake signals based on cues of realism. A common problem of these approaches was that the model trained on specific generator artifacts would perform poorly when evaluated against an unseen generator, even of the same type of manipulation. Generalization across datasets has been identified by survey papers as a core problem in media forensics [9–11].

Synchronisation of audio-visual elements is another example whereby evaluation is done of whether the dynamic alignment among mouth movement, sound onset timing, and semantics matches the anticipated synchronization of natural speech dynamics. It is vital to remember that there could be minor variations between these variables in naturally recorded video snippets brought on by post-processing effects including editing, lag, dubbing, or even by processing through device pipelines. This implies that detectors ought to learn how to operate with scores instead of using a threshold to an offset estimate. Learning synchronisation really came before efforts to identify deepfakes. Studies using embedding techniques such SyncNet have demonstrated that synchronization may be learnt in the natural environment. Learning synchronization as a pretext assignment and providing transferable multimodal embeddings, self-supervised synchronisation learning takes the form of AVTS. Furthermore, multimodal pre-training frameworks like AV-HuBERT have demonstrated the feasibility of generating multimodal representations for audio-visual speech that do not depend on any labels for deepfakes. These representations stand for the previous information that determines what is regarded as "natural" audio-visual speech, therefore making them unevitable by artifact removal citekorbar2018,shi2022.

IV. PROBLEM FORMULATION AND THREAT MODEL

This review considers deepfake detection in the setting of *audiovisual* speech, where the observed sample consists of a video stream $V = \{v_t\}_{t=1}^{T}$ (frames) and an audio stream $A = \{a_t\}_{t=1}^{T}$ sampled over the same time interval. The key hypothesis behind synchronization-based forensics is that natural speech exhibits constrained cross-modal relationships arising from shared physical causes, whereas manipulated content violates these constraints in measurable ways.

A. Synchronization as a Constraint Satisfaction Problem

At a conceptual level, the task of the detector can be understood as determining whether the input signal conforms to a number of constraints, namely: (i) **temporal alignment** (there is a consistent time difference between the audio and mouth signals), (ii) **articulatory feasibility** (the mouth configurations and motions are realistic given the inferred phonemes), (iii) **semantically matching** (the audio semantics match those inferred from the mouth motions), and (iv) **behavioral consistency** (speech timing is aligned with facial micro-behaviors like blinking). Different synchronization approaches rely on different constraints.

B. Adversary Capabilities

A spectrum of attacker capabilities is considered:

TABLE 1: Threats, Typical Manipulations, and Synchronization Evidence

Threat Type	Common Manipulation	Synchronization Evidence	Robust Detector Cues
Audio replacement	Voice cloning, dubbing, sentence splicing	Semantic mismatch (ASR vs VSR), prosody–motion inconsistency	Semantic consistency + long-window fusion
Video replacement	Face swap, reenactment, mouth edit	Offset instability, viseme/phoneme inconsistency	Alignment scoring + articulatory representations
Joint generation	Talking-head generation with TTS	Subtle long-range drift, coarticulation errors	Transformer fusion + self-supervised AV priors
Post-processing	Re-encoding, time-stretch, frame interpolation	Benign lag, blurred lip ROI, jitter	Calibration + multi-cue fusion, robustness stress tests

- **Audio replacement:** Replace or modify audio while leaving the original video mostly unchanged (dubbing, voice cloning).
- **Video replacement:** Replace face or mouth motion while keeping original audio (face swap, reenactment, lip edit).
- **Joint generation:** Generate both audio and video from scratch or from shared latent variables (talking-head diffusion, conditional generators).
- **Post-processing:** Apply compression, re-encoding, time stretching, pitch shifting, frame interpolation, and latency shifts to evade detectors.

The hardest attacker to deal with is joint synthesis with post-processing. But the best joint systems also have to faithfully synthesize intricate structure across multiple time scales, and not just mimic realistic perceptions. The types of attackers, as well as the synchronization patterns associated with each, are listed in Table 1.

C. Benign Mismatch vs. Malicious Mismatch

Practical implementation needs to differentiate between bad mismatch and good misalignment. The following can be causes of good misalignment: platform delay, dubbing in a genuine video, camera pipeline delay, and postproduction. Thus, an approach that gives the output as calibrated scores (even better, along with an explanation such as what part of the sequence is misaligned) would perform better compared to the approach "anything with delay $> \tau$ is fake."

D. Deployment Scenarios

Synchronization forensics finds use in multiple deployment contexts, namely, **(i) media authentication** (journalism and debunking), **(ii) social media moderation** (compressed videos on social platforms), **(iii) biometric security** (spoofing and impostor detection), and **(iv) legal and forensic investigations**, which demand interpretability in findings. These applications present requirements for latency, reliability, and interpretability, thus influencing which synchronization signatures can be employed.

V. AUDIO-VISUAL SYNCHRONIZATION IN HUMAN SPEECH

It is important to note that the sound production and movement of the mouth are two processes that cannot be separated since they occur simultaneously when an individual speaks. Sound production relies on the neuromuscular control system that causes the movements of the lips, jaw, tongue, and facial muscles. Because of this, the timing and type of sound and the movements involved in its production are closely associated. The close association between sound production and lip movements makes it possible to identify small variations between the two. One such relationship involves the phoneme-viseme relationship. Phonemes refer to units of sound production in speech, while visemes refer to the mouth shapes required to produce specific phonemes. It is challenging to lip-read because some phonemes may produce similar mouth shapes, but temporal context solves this challenge because the sounds preceding and succeeding other sounds define the path the mouth will follow during coarticulation [4, 13, 14]. These larger timing patterns prove harder to recreate while using generation that focuses only on mouth ROI rendering. Learning from self-supervised audiovisual data, such as AVTS and AV-HuBERT, proves that multi-scale timing structure is indeed learnable and effective at identifying whether data is real or manipulated [5, 6, 18]. The last characteristic of synchronization is meaning. For the content in the video to be real, its auditory information must correspond to what is seen visually in the mouths of people speaking in it. Some approaches use this by comparing ASR and VSR/visemes; significant discrepancies may mean that data was manipulated [15, 16].

TABLE 2: Taxonomy of Synchronization-Based Deepfake Detection Methods

Method Family	Signal Level	Training Need	Typical Strength
Lip–Audio Alignment	Temporal	Real AV	Offset sensitivity
Self-Supervised AV	Representation	Real AV	Generalization
Transformer Fusion	Cross-modal	Labeled AV	Long-range modeling
Semantic Consistency	Linguistic	No fake data	Generator-agnostic
Physiological	Behavioral	Real video	Explainability

VI. SYNCHRONIZATION-BASED DEEPFAKE DETECTION

A. Taxonomy and Design Dimensions

Different synchronization-based detector approaches depend on what is considered the most convincing evidence of synchronization. Temporal alignment approaches are based on signal-level modeling and try to determine whether there is temporal alignment between the voice and mouth signals by using similarity scores or alignment networks that have been learned. Representation-based approaches learn to create an embedding space where the aligned audio and visual speech will form a manifold, after which they measure divergence. Fusion-based approaches use cross-attention to model synchronization by directly capturing long-range inconsistencies over larger contexts.The taxonomy of the major methods used for deepfake detection on the basis of synchronization is provided in Table 2. The following features served as a basis for classification: volume of evidence; necessity of supervision; and benefits. Taxonomic classification allows gaining knowledge about the peculiarities of intermodality synchronization used in various models. The major criteria used during the process of taxonomic classification are as follows: time period under assessment (number of frames, words, or sentences); necessity for appropriate tracking; presence of synthetic data during the process of training the model; simplicity of model implementation; and interpretability. Cross-family comparison of the algorithms described above is done with the help of Table 3. In particular, the following parameters are used for the comparison: evidence types, advantages, and disadvantages. [2, 14, 17].

B. Temporal Alignment and Offset Estimation

The first approaches to synchronization involved hand-engineered representations (visual mouth landmarks, optical flow, and MFCCs) along with alignment algorithms like cross-correlation and dynamic time warping. Though these were shown to work well, they were affected by noise, compression artifacts, and the inaccuracies of mouth tracking. Contemporary alignment approaches learn embeddings of the auditory and visual channels of speech and determine their similarity based on temporal offset; natural speech is characterized by a narrow peak of high similarity at the correct offset value, whereas tampered content may exhibit a wide and/or unstable peak [3].

C. Supervised Temporal Modeling

The supervised deep learning techniques involve training of CNN/LSTM/TCN model for synchronization anomaly detection based on the use of labeled real or fake data. These models learn temporal patterns associated with the lip and voice synchronization. The real-time approaches are characterized by an emphasis on temporal smoothing and effective feature extraction [13]. The LIPINC-V2 technique emphasizes short and long-term consistency in lip syncing based on vision transformers.

D. Self-Supervised Synchronization Learning and Transfer

Synchronicity learning based on self-supervision assumes that alignment is a pretext task. AVTS trains an audio-visual classifier to classify aligned and non-aligned audio-visual data from real videos [5]. AV-HuBERT uses masked pre-training to learn the joint audio-visual speech representation modeling local articulation as well as global properties [6]. Such representations are interesting to use for detection tasks since they are learned using natural audio-visual speech. The AV-Lip-Sync+ method uses the learned representations for lip-sync deepfake detection [18].

E. Transformer-Based Fusion and Cross-Attention

Transformer-based model fusion has been another significant recent development. The model DF-TransFusion uses cross-attention between lip and audio streams combined with facial stream attention for learning fine-grained correspondences [17]. On the other hand, the model Lips are Lying (shortened as LipFD) employs an audiovisual global transformer in order to identify inconsistencies and long-term correlations even in the presence of consistent shorter streams [14].

F. Semantic Consistency and Zero-Shot Detection

Consistency in semantics is achieved at the content level. One such design employs ASR to extract audio content and VSR/lip-reading to extract visual content, and then determines the discrepancy score between the two (e.g., edit distance, confidence score difference). Such designs can easily generalize across different types of fakes and detect inconsistency independent of how the fakes were made. “Lost in Translation” shows that even if synchronization looks fine at the low level, content-based inconsistencies may still be detectable [16].

G. Physiological and Behavioral Coherence

Physiological coherence techniques rely on non-verbal indicators associated with speech communication such as blink synchronization and facial dynamics. Blink-based forensic techniques have established the inability of most synthetic techniques to simulate natural eye movements [19]. The TrueSync technique is based on blink rate modeling along with lip synchronization due to the understanding that talking is a full-face activity [20].

TABLE 3: Comparison of Synchronization-Based Detector Families

Family	Primary Evidence	Strengths	Common Failure Modes	Practical Notes
Temporal alignment / offset	Peak similarity vs. temporal shift	Interpretable; localizes mismatch	Sensitive to tracking/compression; benign lag can trigger alarms	Benefits from calibration
Self-supervised AV representations	Deviation from natural AV manifold	Strong transfer/generalization	Pretraining domain mismatch; may miss certain dubbing cases	Strong backbone for fusion
Transformer fusion	Cross-attention consistency over time	Captures long-range dependencies	Heavy compute; may overfit dataset bias	Strong when compute allows
Semantic consistency (ASR vs VSR)	Transcript disagreement / confidence gap	Generator-agnostic; robust to high visual realism	VSR uncertainty (low-res, occlusion); language coverage	Excellent complement to alignment
Physiological coherence	Blink / facial rhythm patterns	Adds interpretability; whole-face signal	High person variability; environment effects	Best as secondary evidence

H. Score Calibration and Decision Fusion in Practice

The main limitation of evidence-based synchronisation approaches is the inability to compare the raw information, which includes alignment similarity, embedding distance, transcript inconsistencies and so on, between several videos since they will vary greatly depending on their quality, frame rates, movement clarity and background noise, even in case of real videos. Hence, any implementation of evidence-based approach should include a calibration step, where raw information will be transformed into confidence measures, similar to probabilities. For example, temperature scaling, Platt scaling, isotonic regression and so on can be used as calibration algorithms that will transform raw information into probabilities using some validation sets in realistic conditions. As well as calibration procedures, fusion techniques play a crucial role for such systems. Sometimes, videos may become slightly synchronised yet misleading. Alternatively, fake videos can have poor semantic coherency or synchronisation across all frames although having excellent alignment locally. In order to make use of both types of information, one can try developing a multiple cues fusion algorithm that incorporates (i) alignment confidence, (ii) transformer consistency, (iii) ASR-VSR semantic mismatch and maybe even (iv) physiological evidence, either in learning-based way or confidence-gating way.

VII. DATASETS AND EVALUATION

The creation of synchronization-based deepfake detection techniques requires that datasets contain plausible alterations as well as appropriate knowledge regarding audio-visual synchronicity. The popular benchmarks for audiovisual and face manipulation detection are listed in Table 4, along with their modalities, scale, and standard uses. In contrast to traditional deepfake detection approaches, which solely depend on images, synchronization-based techniques must be trained using datasets in which synchronicity between facial movement and speech has either been maintained or disrupted [7, 8, 21].

TABLE 4: Major Datasets Used in Audio–Visual Deepfake Detection

Dataset	Modalities	Scale	Notes
FaceForensics++	V	Med	Multi-method benchmark [7]
DFDC	V	Large	Diversity and scale [8, 22]
Celeb-DF	V	Med	Higher realism stress test [23]
FakeAVCeleb	AV	Med	Audio+video fakes [24]
DeepFakeTIMIT	AV	Small	Controlled timing [25]

A. Benchmark Protocols and Generalization

The evaluation strategy should also be influenced by deployment concerns. Beyond data splits, the modern approach to evaluation requires dataset-wise, generator-wise, and compression-wise evaluations [7]. DFDC may provide variation and a large-scale of examples; however, splitting plays a crucial role in preventing any leakage and biases against the source [8]. On the basis of multiple papers, the lack of proper splitting leads to learning dataset features rather than the synchronization itself [9, 10]. What matters during implementation is whether it is an unintentional or intended discrepancy. It goes without saying that real videos may suffer from AV desynchronization due to video editing, voice dubbing, or some pipeline issues, and hence the detector should properly react to such confidence intervals. The main question is whether the desynchronization detection capability is combined with a low-confidence error caused by lip-reading.

B. Lip-Sync-Focused and Short-Video Benchmarks

Problems related to lip sync at the millisecond scale can be identified through lip-sync benchmarking datasets. Using the

TABLE 5: Evaluation Metrics and Robustness Factors for AV Synchronization Detectors

Item	Description
AUC / EER / F1	Standard detector performance measures
Offset stability	Variance of predicted lag across windows
Peak sharpness	Concentration of alignment confidence near best lag
Transcript mismatch	ASR–VSR disagreement (token distance/confidence gap)
Compression stress	Performance vs. codec/bitrate re-encoding
Frame-rate stress	Performance under downsampling and jitter
Occlusion stress	Performance under mouth ROI occlusion/pose
Cross-dataset	Generalization to unseen benchmarks/generators

specific dataset for benchmarking lip-sync, it would become easier for an individual to identify whether he/she is sensitive to temporal mismatch and contextual dependency [13, 14]. Perception of mouth movement could get impacted by the existence of data sets that consist of short videos, which have gone through rigorous coding and come with different frame rates. It is due to the fact that they do not rely solely on pixel-level motion [15, 18].

C. *Metrics Beyond Accuracy*

In addition to accuracy, AUC, and F1, synchronizaton-based techniques also benefit from being able to use interpretable diagnostic measures. Alignment-based techniques may provide the maximum similarity peak, peak sharpness, and offset consistency. Representation-based techniques may use modal dissonance distances, while semantics-based techniques could provide the percentage of disagreements among transcripts and the confidence difference score. Table 5 lists commonly reported metrics specific to synchronization-based measures.

VIII. ROBUSTNESS, EVASION, AND COUNTERMEASURES

As detectors become more sophisticated, the attackers will continue to evolve their methods. Issues relating to multiscale analysis, evasion techniques, and the defense measures utilized against such attacks are all discussed in this part.

A. *Evasion Strategies*

(1) Offset spoofing: The attacker might purposely manipulate global AV asynchrony in accordance with the actual dissemination of delay. Even if it could reproduce the consistent phoneme-level synchrony and coarticulation found in natural human communication, it might yet be able to deceive detectors based on lags. **(2) Local patch alignment:** During the process of generating voice samples, the system will allow drift in large chunks, although it will be able to correctly synchronize small chunks, leading to realistic synchronization of the local regions without being completely synchronized throughout the whole audio clip. **(3) Semantic smoothing:** Even if voice cloning using lip movements based on transcripts can minimize discrepancies between ASR and VSR, there might still be issues regarding the authenticity of the articulation and rhythm dissimilarities. **(4) Post-processing attacks:** Low-quality detection of the mouth area and synchronization errors can result from video frame manipulation through time stretching, interpolation, blurring, and heavy compression.

B. *Countermeasures*

Robust detectors incorporate: **multi-scale modeling** (short + long temporal windows), **uncertainty-aware scoring** (down-weight unreliable modalities), **stress-test evaluation** (compression, frame-rate, occlusion), and **ensemble fusion** of heterogeneous evidence sources. A practical strategy is to treat synchronization forensics as a hypothesis test with confidence intervals: when mouth ROI quality is low, the system can report "insufficient evidence" rather than forcing a high-confidence decision.

C. *Why Semantics Helps Under Generator Evolution*

Semantic agreement offers another orthogonal dimension of evidence beyond just timing. Although a generator can achieve improvements at the lower level of lip alignment, ensuring that the content is aligned between audio and visually generated speech even when the camera view changes, when occlusions occur, and when the signals are compressed is difficult.

IX. DEPLOYMENT AND SYSTEMS CONSIDERATIONS

Experimentation settings have often been used for the assessment of synchronization detection algorithms; yet, such processes are always conducted within particular operational limitations. This part considers a number of technical concerns that affect significantly the performance of the algorithm in practice.

A. *Latency and Throughput*

It is essential to ensure low latency in case of a real-time evaluation of a synchronization platform and content moderation. Despite being able to produce acceptable results, large audio-visual transformers require a huge amount of computational resources. Among popular methods there are: (i) ROI encoding with small mouth windows, (ii) downsampling without loss of phoneme segmentation, (iii) audio feature caching, and (iv) knowledge distillation from multimodal fusion to smaller student networks.

B. *Tracking Reliability and Failure Handling*

Face detection and ROI extraction for mouths constitute essential components in any synchronization pipeline. Nonetheless, problems of motion blur, incorrect illumination, occlusion, or face misalignment may cause inaccuracies in predictions by the algorithms discussed earlier. The resolution for such challenges implies identifying the reason behind tracking failure and taking one of the following steps: (a) use additional cues (semantic analysis only) or (b) deliver inaccurate predictions.

C. Human-Interpretable Outputs

Some samples of explainable output expected with regard to workflows of journalism, law, and security are: Mismatch heatmaps of ROI, the curve of confidence of alignment changes, and token-level ASR-VSR disagreement maps. They will facilitate auditing and help with decreasing resistance towards implementation.

D. Ethical and Fairness Considerations

The process of synchronization depends on the language in question, its dialect, and the manner of speech. In cases where datasets are overloaded with voices recorded in studios in English, there may be a negative effect on the performance of detectors for out-of-domain speakers and under-resourced languages. Multilingual evaluations should consider different language groups and recording conditions in terms of language-specific failure rate and video quality binning.

X. CASE STUDIES AND FAILURE ANALYSIS

However, failures are understandable and avoidable. The next section looks at the examples of outstanding failures and the reasons why the system accounting for various features and uncertainty measures is needed. Table 6 gives the description of several common examples of failures of synchronization-based detectors and suggestions.

A. Case 1: Benign Dubbing and Legitimate Re-voicing

News segments, documentaries, and foreign media also rely heavily on dubbing or narration. In this case, the sound will deliberately not correspond to the speech that can be seen in the video, thus activating semantic consistency checks. Pairs of ASR-VSR that differ just from semantic consistency issues might yield a false positive result. The system should be able to distinguish dubbing as a distinct category of "benign mismatch," for example, with: (i) identical face identity and dynamics, consistent lip movements, and different content; (ii) consistent mismatch throughout the clip (as opposed to local editing); and (iii) metadata. A possible solution in real-world pipelines would be a two-stage process where we check if there is visible speech in the clip at all, and only then apply consistency checks.

B. Case 2: Heavy Compression and Low Frame Rate Short Videos

A low resolution, high compression, and a low frame rate have been witnessed in most of the short video sharing sites. This could pose some problems on the face region of interest and thus, would lead to difficulties in aligning the faces accurately, making the videos less authentic. In such a case, it would be quite noisy aligning individual frames with different offsets in different windows without having any peak and alignment disparities. There exist several solutions to solve this problem, namely (i) a larger time window (word/sentence level), (ii) using self-supervised features that would reduce the sensitivity to small motions in the face regions, and (iii) performing some verification where signals from VSR/ROI are not used when the mouth is absent.

C. Case 3: Local Audio Splicing and Partial Manipulation

The first category of attack would be partial manipulation, whereby certain elements within the speech undergo changes regarding their meanings but the rest of the video is untouched. Coarse lag detector may prove inadequate for this particular attack due to the preservation of timing during the local splicing technique. The issue with the local splicing technique is that it results in local anomalies since there are strange prosody changes, although the face maintains its rhythm due to change of meaning.

D. Case 4: Generator Overfitting and 'Too Perfect" Local Alignment

Whereas modern generator technologies can successfully forge synchronization in the lip movement by means of short windows, the forgery can still be successful in bypassing detectors because short window alignment can serve as a basis for detecting the forgery. The ability to achieve perfect synchronization in the short window while there may still be discrepancies in the long window in terms of co-articulation, rhythm, and semantic synchronization is among the reasons why forgery detection is so difficult.

E. Practical Takeaway

For all these applications, the basic idea is that no one cue suffices by itself. The properties of the most successful systems in terms of reliability and accuracy include: (i) multiple cue combination, (ii) analysis of the reliabilities of the modalities involved, and (iii) explanation of the reasons for the mismatch.

XI. IMPLEMENTATION PIPELINE AND REPRODUCIBILITY GUIDELINES

The tasks of comparative evaluation and reproducibility studies can only be performed when synchronization-based deepfake detection algorithms are described and analyzed using a modular pipeline approach. Considering existing forensic tools and audiovisual machine learning techniques, the subsequent section explains the elements of the pipeline for a synchronization-based approach to detecting deepfakes. Table 7 gives an overview of the entire pipeline architecture and potential problems with deepfake detection using synchronization.

A. Reference Processing Pipeline

The typical detector using the synchronization technique can be described by the following stages: (i) face detection and tracking, (ii) mouth ROI detection and normalization, (iii) audio processing and feature extraction, (iv) cross-modal coding, (v) synchronization score estimation, and (vi) decision making. Small modifications at some stage could greatly affect its efficiency, e.g., noisy face tracking will produce noisy visual features, and bad audio resampling will generate wrong phoneme timings. The interrelation between AV feature extraction, synchronization scoring, semantic verification, and decision making are depicted in Figure 1. The used detector and tracker, frame rate after pre-processing, ROI dimensions, audio sampling rate, as well as window length, step and padding should be stated in order to make the test repeatable. Even minor variations in syn-

TABLE 6: Common Failure Modes in Synchronization-Based Deepfake Detection and Mitigations

Scenario	Why It Fails	Mitigation Strategy	Best Evidence
Legitimate dubbing	ASR–VSR mismatch is expected	Detect 'visible speech" first; treat as benign mismatch class; rely on non-semantic cues	Alignment stability + representation
Heavy compression / low fps	Mouth ROI degraded; tracking noisy	Window-level aggregation; uncertainty weighting; fallback to audio-only or semantic-only if ROI unreliable	Long-window fusion
Partial audio splicing	Global lag unchanged; manipulation localized	Boundary localization via offset variance + semantic heatmaps; long-context modeling	Semantic + attention
Occlusion / pose	VSR unreliable, landmark jitter	Reliability gating; multi-view or robust face tracking; abstain when evidence weak	Calibration + gating
Strong local lip-sync generators	Short windows look perfect	Multi-scale checks; sentence-level drift analysis; prosody-motion coherence	Transformer fusion

chronization score calculations can lead to differences in the benchmark ranking results.

B. Windowing and Temporal Resolution Choices

The synchronization can be computed at the level of frames, phonemes, words, or even whole sentences. The shorter the window size, the better the resolution; however, such measures are also susceptible to noise and data loss due to compression. The longer the window, the more stable the measure, but it loses resolution. An optimal solution would be a fusion of multiple window sizes.

C. Training and Validation Protocols

For avoiding leakage of identity, identity-disjoint splitting is very necessary in case of supervised training. Always mention cross-dataset validation where it is applicable. Information regarding pretraining datasets and whether these pretraining datasets were derived from evaluation datasets should always be provided during pretraining of self-supervised backbones.

D. Open Resources and Benchmark Hygiene

It is strongly advised to provide the source code, pretrained weights, and preprocessing steps for achieving reproducibility. Variance reporting, publication of splits, and seed fixing are the best practices for benchmarking. Since synchronization algorithms heavily depend on monitoring jitter and temporal alignments, it is suggested to report the mean and standard deviations.

XII. RESEARCH ROADMAP TOWARD 10-YEAR ROBUSTNESS

To remain effective under rapidly evolving generators, synchronization-based detection should move toward robust priors, strict evaluation, and deployable uncertainty-aware sys-

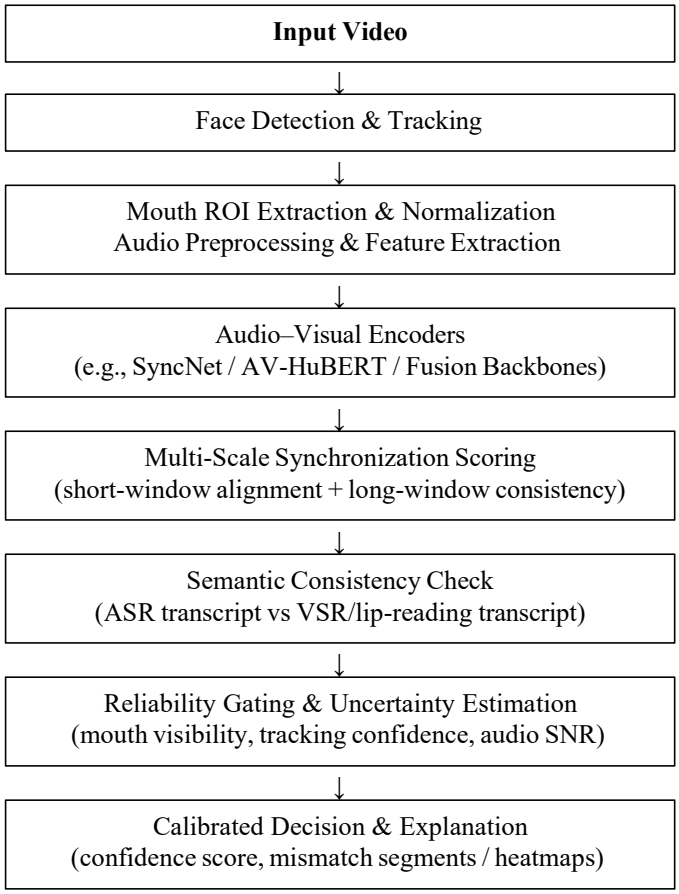

Fig. 1: Reference modular workflow for synchronization-based deepfake detection.

TABLE 7: End-to-End Components in Synchronization-Based Deepfake Detection Pipelines

Stage	Typical Methods	Key Risks	Best Practice
Face detection/tracking	CNN or transformer detectors	Drift, missed faces, pose failure	Report tracker and failure rate
Mouth ROI extraction	Landmark or mesh-based crop	Landmark jitter, blur	Temporal smoothing, quality checks
Audio features	MFCC, log-mel, learned encoders	Resampling mismatch	Fix sampling rate, publish params
AV encoding	SyncNet, AV-HuBERT, fusion encoders	Domain shift	State pretraining corpus
Sync scoring	Offset curves, embedding distance	Threshold sensitivity	Calibrated scoring
Decision fusion	Late fusion / gating	Overconfidence	Reliability-weighted fusion

tems.

A. Unified Multimodal Priors

Using many examples of actual speech videos for multimodal pre-training could serve as an excellent prior for real audiovisual speech. The objective is to create cross-modal data distribution models for every language, under any visual conditions, and without regard for speaking styles, rather than maximizing accuracy regardless of expense. Therefore, detectors need to be considered "violations of constraints" instead of fingerprint detectors.

B. Standardized Stress Tests

(i) compression bands, (ii) variation in frame rates, (iii) occlusions, (iv) honest dubbing, and (v) generator drifts are all examined using standardized benchmark datasets. For claims of out-of-distribution performance, accuracy reports on the dataset are inadequate.

C. Uncertainty-Aware Fusion

Any effective detector would take reliability cues from every modality (e.g., mouth visibility, tracker reliability, audio SNR) into account and properly integrate them. The detector must either avoid making quick binary decisions when there is not enough reliable input data or report confidently low accuracy.

D. Explainability as a First-Class Output

Explanations are ideal for synchronization detectors because they can provide visualizations such as offset curves, mismatch maps, and even disagreements between automatic speech recognition and visual speech recognition at the level of individual tokens.

XIII. CHALLENGES AND FUTURE DIRECTIONS

There are still several challenges to overcome prior to actual deployment, notwithstanding the remarkable development achieved thus far. First, there is generalization. This challenge involves cases when detectors operate successfully within training data and generators, but fail to do so when applied to other synthetic and/or environmental models. This type of challenge has already been acknowledged in several benchmarks and surveys [8–11]. At present, the informal benchmark for stress testing is generalization in cross-datasets. In specific, datasets obtained via unknown generators and compressed images, algorithms which work efficiently in one dataset tend to face major issues in another one [7, 8]. Due to their higher level of representations, algorithms based on self-supervised learning and semantic similarity demonstrate better results [10, 28, 11]. The monitoring process should include occlusions, movements, and low-light conditions. It is specifically challenging due to its influence on detectors using ROIs for tracking the mouth. Social media videos pose an additional challenge due to the inconsistency in frame rate and compression. Algorithms that rely on semantic agreements or self-supervised robust features may work better due to their higher abstraction. [6, 16, 18]. A further limitation is the cost of computing. Methods using transformers for fusion are reliable yet costly; in real-time scenarios, it is necessary to employ an encoding technique or compress/distill the model. On the other hand, the adversary continuously evolves; as synchronization techniques capitalize on the synchronization aspect of audio and video streams for identifying deepfakes, the generator learns to synchronize its lip movements to imitate biological phenomena. Various types of signals like synchronization, consistency, and behavior might have to be considered for a robust detector [1, 17, 20]. In terms of forensics and public policy, the interpretability of the detector is important. The synchronization technique holds promise as it offers an interpretation framework of the results (in case ASR and VSR words differ or synchronization fails).

A. Open Research Problems

Other open issues for research include: **(i) Cross-language alignment priors:** Most of the available training sets and benchmarks are in English. It is important to have cross-language priors that can be adaptable despite the differences in viseme configurations and speech rhythm. **(ii) Differentiating**

dubbing from real dubbing: Dubbing leads to audio-video mismatch when it is done properly. It is crucial to detect malicious audio manipulation without detecting legitimate dubbing as a false positive. **(iii) Robust detection despite high compression rates:** Social media platforms can eliminate the ROI of the mouth in the preprocessing phase. Approaches relying on abstraction have proven effective, but they might not be the best. **(iv) Integration with probabilistic calibration:** To integrate different modalities that contain uncertainty, there is a need for a robust framework. **(v) Forensic identification that is comprehensible:** It is important that the subject can be identified by an outsider for authentication purposes.

XIV. CONCLUSION AND BROADER IMPLICATIONS

Analysis of audio-visual synchronization is arguably the most progressive and technically grounded approach in detecting deepfakes. While artifact detection uses weaknesses of the model to find inconsistencies, synchronization is based on cross-modal limitations, such as the close connection between speech production components, including voice, articulation movement, facial expression, and semantics, that is rooted in human physiology. Although current deep learning models can generate increasingly realistic images and voices, the relationships between these elements provide a more solid foundation for a detection process due to the identical physical/neural principles behind their interactions. Recent advances in the field were discussed in the article under review in several research directions regarding different synchronization signals: temporal alignment, representational-level cross-modal consistency, cross-modal transformers, semantic agreement of automatic speech recognition and visual speech recognition systems, and physiological cues. As per the findings of the survey, the development of the area was logical: first, initial alignment techniques showed their feasibility, then increasingly sophisticated supervised temporal deep models emerged, and finally, self-supervised and multi-modal transformer-based models were developed. One takeaway from this survey is that forensic synchronization should be implemented through multi-constraint models instead of simple score methods. Overall, multi-constraint models that consider factors such as short-term alignment, long-term consistency, semantics, and modality reliability estimation tend to outperform one-dimensional metrics regarding compression tolerance, occlusion, and adaptation to different domains. Thus, for a viable forensic synchronization system, proper calibration and uncertainty handling are essential but not only technical challenges. Nevertheless, several limitations should be mentioned. Compression, a low frame rate, mouth occlusion, and the multi-language nature of visemes may negatively impact synchronization. It is necessary to establish whether the discrepancy between modalities stems from intentional modification or simply reflects inter-modal inconsistencies due to dubbing and voice-over modifications. The available benchmarking efforts exhibit a preference toward certain languages and data collection procedures. Currently, there is no universal stress test for forensic synchronization.The systemic perspective underscores the significance of having well-functioning pipelines, precise monitors, and meaningful results. In instances that require justifications, synchronizers are particularly useful since they will make evident the points at which synchronization was not achieved, at which words do not match across modalities, and the sections of the speech that were highly uncertain in their synchronization. These are essential in contexts such as forensics, content moderation, and journalism, among others, since justifications will be required for the actions taken as part of the processes involving the AI-driven functions. Going forward, the key focus should be on using large multimodal foundation models with forensic considerations.With multilingual data, thorough evaluations on different data sets, uncertainty-aware integration, and robustness metrics for standardization, this method is capable of generating detectors that will maintain their effectiveness even as the generators become more advanced. Finally, synchronization-based detection of deepfakes is an idea of multimodal verification, a universal principle, and not a technology alone. If the methods developed for deepfake detection involve cross-modal speech synchronization rather than abnormality recognition, this field will move toward more robust detection techniques that are also more interpretable and transferable and resilient against generator updates.

REFERENCES

REFERENCES

[1] M. F. Hashmi, S. A. Shahzad, C.-W. Lin, Y. Tsao, and H.-M. Wang, “A comprehensive survey of audio-visual deepfake generation and detection techniques,” arXiv:2411.07650, 2024.

[2] Z. Zhang and S. Li, “Joint audio-visual deepfake detection,” in *Proc. IEEE ICCV Workshops*, 2021.

[3] J. S. Chung and A. Zisserman, “Out of time: Automated lip-sync in the wild,” in *Proc. Asian Conf. Computer Vision (ACCV)*, 2016.

[4] K. R. Prajwal, R. Mukhopadhyay, V. Namboodiri, and C. V. Jawahar, “Wav2Lip: Accurately lip-syncing videos in the wild,” in *Proc. ACM Multimedia*, 2020.

[5] B. Korbar, D. Tran, and L. Torresani, “Cooperative learning of audio and video models from self-supervised synchronization,” in *Proc. NeurIPS*, 2018.

[6] B. Shi et al., “AV-HuBERT: Self-supervised learning of audio-visual speech representation,” in *Proc. NeurIPS*, 2022.

[7] A. Rössler et al., “FaceForensics++: Learning to detect manipulated facial images,” in *Proc. IEEE ICCV*, 2019.

[8] B. Dolhansky et al., “The DeepFake Detection Challenge (DFDC) dataset,” arXiv:2006.07397, 2020.

[9] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[10] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys*, vol. 54, no. 1, 2021.

[11] R. Tolosana et al., "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.

[12] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE CVPR*, 2016.

[13] M. Datta, Y. Jia, and S. Lyu, "Detecting lip-syncing deepfakes using vision temporal transformers (LIPINC-V2)," arXiv:2504.01470, 2025.

[14] X. Liu et al., "Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes," arXiv:2401.15668, 2024.

[15] Y. Li et al., "Zero-shot fake video detection via ASR–VSR semantic consistency," arXiv:2406.07854, 2024.

[16] M. Bohác˘ek and H. Farid, "Lost in translation: Lip-sync deepfake detection from audio-video mismatch," in *Proc. IEEE/CVF CVPR Workshops*, 2024.

[17] R. Kharel, W. Cai, and K. Radecka, "DF-TransFusion: Multimodal deepfake detection via lip-audio cross-attention and facial self-attention," arXiv:2309.06511, 2023.

[18] S. Khan et al., "AV-Lip-Sync+: Leveraging AV-HuBERT for lip-sync deepfake detection," arXiv:2311.02733, 2023.

[19] Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI-created fake videos by detecting eye blinking," in *Proc. IEEE WIFS*, 2018.

[20] A. A. El-Taj et al., "TrueSync: Deepfake detection based on visual lip-sync match and blink rate," ResearchGate Preprint, 2025.

[21] M. Javed et al., "Audio–visual synchronization and lip movement analysis for real-time deepfake detection," *International Journal of Computational Intelligence Systems*, 2025.

[22] B. Dolhansky et al., "The DeepFake Detection Challenge dataset," arXiv:1910.08854, 2019.

[23] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF CVPR*, 2020.

[24] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," in *Proc. NeurIPS Datasets and Benchmarks Track*, 2021.

[25] P. Korshunov and S. Marcel, "DeepFakeTIMIT: A dataset for deepfake detection," Idiap Research Institute, 2018.

[26] R. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other—Audio-visual dissonance-based deepfake detection and localization," in *Proc. ACM Multimedia*, 2020.