ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume 29, Number 1, 2026, 53–64
doi: 10.59277/ROMJIST.2026.1.05

Multimodal Visual Speech Recognition for Under-Resource Languages via Cross-Modal Learning and Large Language Models

Ruxandra TAPU[1,2,*], Bogdan MOCANU[1,2], and Ionut-Cosmin CHIVA[2]

[1]SAMOVAR, Télécom SudParis, IP Paris, France
[2]National University of Science and Technology POLITEHNICA Bucharest, Romania
Email: ruxandra.tapu@upb.ro*, bogdan.mocanu@upb.ro, ionut_cosmin.chiva@upb.ro
* Corresponding author

Abstract. This paper introduces a unified approach to multilingual visual speech recognition (VSR) that combines cross-modal phonetic modeling with large-scale language decoding to enable robust generalization across low-resource and previously unseen languages. The architecture within the approach includes a Cross-Modal Transcriber that encodes synchronized audio-visual speech inputs into a language-agnostic phoneme space via a fine-grained cross-attention mechanism. To bridge perception and language understanding, two decoding pathways are explored: (1) a modular configuration that maps phonetic sequences to text using a pretrained large language model (LLM), and (2) an end-to-end formulation in which fused visual features are projected into the LLM's embedding space via a lightweight adapter for direct transcription. Experimental evaluations on the mTEDx multilingual corpus show that the architecture surpasses state-of-the-art VSR models, achieving up to a 6% absolute improvement in WER across Latin-derived languages.

Key-words: Cross-modal attention; large language models; multilingual learning; visual speech recognition.

1. Introduction

Visual Speech Recognition (VSR), also known as *lip reading*, is the task of inferring spoken language from a visual input (*i.e.*, speaker's articulatory movements and facial expressions), particularly the lips. As a modality-agnostic alternative to audio-based ASR, VSR excels when audio is noisy or unavailable, such as in crowded spaces, surveillance contexts, or communication with hearing- or speech-impaired individuals [1]. Despite its practical importance, VSR presents distinct and substantial challenges.

A primary challenge in VSR comes from the inherent ambiguity of visual speech: many phonemes (*e.g.*, /p/, /b/, /m/) produce highly similar lip movements and are visually indistinguishable. This issue becomes more severe in continuous speech, where rapid coarticulation, facial expressions, and head pose variations blur the temporal and spatial boundaries between phonemes and words. Unlike audio, which provides prosodic cues for disambiguation, the visual modality offers far less explicit segmental information. VSR systems must also handle wide variations in speakers, lighting, and viewpoints, while processing high-dimensional video data that increases computational demands and requires robust, efficient architectures. While English VSR has progressed quickly thanks to large annotated datasets, advances in non-English and low-resource languages remain limited. This raises the question of whether future gains should rely on larger corpora or on methods that enable stronger cross-lingual generalization. We argue that real progress requires both expanded resources and improved architectures tailored to multilingual VSR. To this end, this paper proposes a scalable cross-modal approach that injects fused audio-visual representations into a pretrained Large Language Model through lightweight adapter tuning. Designed for low-resource scenarios, the approach combines lip-motion cues and mouth-region dynamics with linguistic priors in a unified pipeline, enabling accurate lip reading with minimal target-language supervision.

The main contributions of ths paper are as follows: (1) *Cross-modal architecture for language-agnostic representation learning.* This paper proposes an end-to-end approach that captures fine-grained temporal dependencies between audio and visual modalities using a cross-attention mechanism. Unlike prior VSR systems that rely on unimodal encoders or simple feature concatenation, the model jointly encodes synchronized audio-visual input into pronunciation-aware, language-agnostic embeddings. At inference, it runs in a visual-only mode, leveraging the multimodal alignment learned during training to achieve zero-shot recognition in unseen languages. (2) *Integration of pretrained LLMs with adapter-based conditioning.* The VSR pipeline is extended by embedding fused audio-visual features directly into a pretrained large language model. Unlike prior approaches where language models act as loosely coupled decoders, the proposed design incorporates lightweight adapter modules that permit efficient fine-tuning while preserving pretrained parameters. This integration grounds perceptual features in a semantic space, improving disambiguation of visually indistinct phonemes and enabling multilingual decoding without retraining the full model. (3) *The Romanian in-the-wild Visual Speech Recognition (RoVSR) database*. A 165-hour automatically curated audiovisual dataset sourced from YouTube is introduced, representing the largest visual speech corpus to date for a previously unrepresented language. Transcriptions are generated using WhisperX [2], enabling scalable, low-cost data collection with minimal human intervention. To scale lip reading across languages, a multilingual training paradigm is adopted that jointly optimizes the model across multiple Latin-script languages. Furthermore, the pretrained model is effectively fine-tuned on Romanian, achieving strong performance without requiring extensive task-specific retraining.

The paper is organized as follows: Section 2 reviews related work, Section 3 details the proposed cross-modal approach, Section 4 describes the experimental methodology and results, and Section 5 concludes the paper and outlines future directions of research.

2. Related Work

With the emergence of deep learning, visual speech recognition shifted from handcrafted features to end-to-end architectures capable of learning spatial and temporal patterns directly from data. Early models employed recurrent networks (*e.g.*, Bi-GRUs, LSTMs [3]), later replaced by

Transformer-based architectures [4] and, more recently, Conformers [5], which better balance local and global dependencies. To improve generalization, research has leveraged self-supervised learning [6], knowledge distillation from ASR [7], and large-scale multimodal pretraining (*e.g.*, AV-HuBERT [8]). More recently, large language models have been integrated into VSR [9] to enhance semantic disambiguation of visually similar phonemes.

Over the past five years, VSR has seen remarkable progress, particularly in English, driven by the availability of large-scale video-text corpora such as LRS3 [10]. Benchmark performance has improved dramatically, with WER on LRS3 reduced from over 60% to 26.9% using AV-HuBERT [8], and further down to 12.8% in later models [11]. Despite these advances, VSR remains highly biased toward English due to severe imbalances in training resources. For example, the state-of-the-art multilingual VSR system in [12] uses 331 hours of Spanish data, just 25% of what was available for English as early as 2018, and less than 5% of current English resources. This disparity highlights a broader trend: while English datasets continue to grow, non-English video-text resources remain scarce, limiting generalization and hindering multilingual VSR development. The bottleneck arises primarily from the high cost of accurate transcription, especially across diverse linguistic settings requiring international annotation efforts.

To address this, prior work [13] has explored multilingual and transfer learning approaches, leveraging representations learned from high-resource languages (*e.g.*, English) and adapting them to low-resource targets via fine-tuning or language-specific modules. Despite progress, current lip-reading systems still face key limitations. Most research focuses on high-resource languages, particularly English, leaving underrepresented languages largely unexplored. As a result, cross-lingual generalization and zero-shot performance remain poorly understood. Moreover, while pretrained language models have improved decoding, they are often used as separate modules rather than being fully integrated into the VSR pipeline, limiting joint optimization and contextual learning. Achieving robust lip reading in diverse, multilingual, and low-resource settings requires not only scalable data strategies but also architectures capable of unifying perceptual and linguistic signals within a complete framework.

The challenges of multilingual visual speech recognition are addressed with a cross-modal approach designed for robust generalization to low-resource and unseen languages. Unlike prior work relying on language-specific decoding [9] or large labeled datasets [12], authors' unified approach learns pronunciation-aware representations from synchronized audio-visual input and conditions a pretrained language model directly on these features. Through cross-modal attention and adapter-based fine-tuning, the architecture enables direct interaction between visual input and linguistic context for semantically informed decoding. At the core of the system lies a language-agnostic transcriber that removes the need for grapheme-level supervision and enables zero-shot transfer by aligning visual speech patterns with phonetic and semantic representations. The approach is evaluated on Romanian, a Latin-derived language unseen during training, showing effective adaptation with minimal labeled data and strong scalability to multilingual settings.

3. Proposed Approach

The proposed approach (Fig. 1) is a modular yet tightly integrated architecture for visual speech recognition. It comprises three complementary components: (1) a Language-Agnostic Transcriber via Cross-Modal Fusion (purple modules), (2) a Large Language Model (LLM) employed as a neural decoder (green modules), and (3) a Multilingual Adaptation and Transfer mechanism for low-resource languages, which builds on all modules in the diagram.

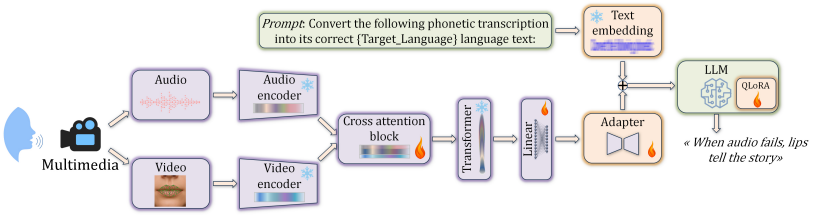

Fig. 1. The proposed end-to-end system architecture designed for VSR.

3.1. Language-agnostic transcriber via cross-modal fusion

To enable robust cross-lingual generalization, the proposed approach employs a pronunciation centric output space defined over phonemic units, rather than language-specific graphemes. This design choice decouples the decoding process from orthographic conventions and writing system constraints, providing a linguistically grounded but language-agnostic representation of spoken content. By modeling speech at the phonetic level, the system benefits from a consistent and transferable supervision signal that captures articulatory structure across diverse languages. This abstraction facilitates the alignment of visual and linguistic modalities in a manner that is invariant to spelling or script, thereby eliminating the need for language-specific output vocabularies. Motivated by prior work highlighting the cross-lingual robustness of phoneme-based modeling [14], this representation is adopted to unify the output space across languages and enable generalization to low-resource languages.

The Cross-Modal Transcriber is a core element of the proposed architecture designed to predict language-agnostic phonetic transcriptions from audio-visual speech input (shown in purple in Fig. 1). The objective is to capture pronunciation-level information that is invariant to orthographic systems and specific language identities, thereby enabling generalization across multilingual settings. The model consists of an audio encoder E_{audio}, a video encoder E_{video}, a cross-attention module for modality fusion, a transformer encoder T for temporal modeling, and a linear projection head for outputting phoneme probabilities. The proposed approach incorporates audio input exclusively during training as an auxiliary supervisory signal. This design enables the model to learn visually grounded phonetic representations by aligning visual features with the articulatory structure captured in the acoustic modality. Crucially, only the visual modality is used during inference, ensuring that the system remains applicable in lip-reading scenarios. By exploiting audio-visual alignment during training and decoupling audio at test time, the model achieves improved generalization and robustness for real-world VSR applications.

Given a training tuple (a, v, r), where $a \in \mathbb{R}^{F \times T_a}$ denotes the log-mel spectrogram features extracted from the audio waveform using F frequency bins over T_a time steps, and $v \in \mathbb{R}^{H \times W \times C \times T_v}$ represents a sequence of T_v lip-region video frames with spatial dimensions height H, width W, and C color channels, and where r is the corresponding target phoneme sequence, the model first computes modality-specific embeddings using dedicated audio and visual encoders. The audio encoder produces a sequence of acoustic embeddings $e_{\text{audio}} = E_{\text{audio}}(a) \in \mathbb{R}^{T \times d}$, while the visual encoder outputs a corresponding sequence of visual features $e_{\text{video}} = E_{\text{video}}(v) \in \mathbb{R}^{T \times d}$, where T denotes the number of temporally aligned frames and d represents the dimensionality of the shared embedding space. Temporal alignment across modalities is ensured through preprocessing steps and subsampling operations applied to both audio and video streams.

To standardize the temporal resolution across modalities, all video sequences are resampled to a fixed frame rate of 25 Hz, while audio waveforms are downsampled to 16 kHz, ensuring consistent frame-to-frame correspondence in the synchronized audio-visual stream. Facial region detection is performed using S3FD [15], followed by frame-wise landmark extraction to isolate the orofacial region. The extracted mouth region of interest (ROI) is then cropped and resized to a spatial resolution of 96×96 pixels, providing a consistent visual input across samples. To enhance data variability and promote generalization to out-of-distribution conditions, spatiotemporal augmentation strategies are applied during training. These include deterministic horizontal flipping, applied consistently across all frames within a clip to preserve temporal coherence, and random spatial cropping to 88×88 pixels, introducing mild perturbations in spatial localization while retaining articulatory information. Transcriptions are preprocessed using the Fairseq [3] normalization pipeline, which standardizes punctuation, case formatting, and spacing, facilitating multilingual alignment and compatibility with the phoneme-based vocabulary used during training. The visual encoder employs a ResNet-18 backbone, preceded by a 3D convolution to capture short-term spatiotemporal dynamics in the lip region across consecutive frames. In parallel, the audio branch uses a lightweight linear encoder to project log-mel spectrogram features into a shared embedding space. Rather than relying on early fusion techniques such as naive concatenation, which fail to capture modality-specific structure and interactions, a cross-modal attention mechanism is implemented to explicitly model dependencies between audio and visual feature streams. This module aligns temporally synchronized embeddings from each modality and performs feature-wise conditioning, allowing one modality to inform the representation of the other through learned relevance mappings. The attention process enriches unimodal representations with complementary cues, enhancing robustness to noise and modality-specific ambiguities. Cross-modal dependencies between audio and visual streams are captured by computing a learned similarity matrix $M \in \mathbb{R}^{T \times T}$.

This matrix is obtained by projecting the modality-specific embeddings into a shared latent space and computing pairwise interactions according to $M = e_{\text{audio}}\, W\, e_{\text{video}}^{\top}$, where $W \in \mathbb{R}^{d \times d}$ is a trainable projection matrix. Each element m_{ij} of M quantifies the alignment strength between the i-th temporal feature from the audio stream and the j-th temporal feature from the visual stream. To convert these similarity scores into an attention distribution, each row of M is normalized using a softmax operation: $\mathrm{Att}_{\text{audio}} = \mathrm{softmax}(M)$ and $\mathrm{Att}_{\text{video}} = \mathrm{softmax}(M^{\top})$, where $\mathrm{Att}_{\text{audio}}$ and $\mathrm{Att}_{\text{video}}$ denote the attention weight matrices obtained by applying a row-wise softmax normalization to the similarity matrix and its transpose, respectively. Each element of these matrices represents the normalized alignment strength between pairs of temporal features across modalities. The resulting attention weights are then used to compute cross-attended representations for each modality, specifically $\tilde{e}_{\text{audio}} = \mathrm{Att}_{\text{audio}}\, e_{\text{audio}}$ and $\tilde{e}_{\text{video}} = \mathrm{Att}_{\text{video}}\, e_{\text{video}}$.

To retain the discriminative content of the original modality-specific features, a residual refinement step is applied using non-linear transformations: $\hat{e}_{\text{audio}} = \tanh(e_{\text{audio}} + \tilde{e}_{\text{audio}})$ and $\hat{e}_{\text{video}} = \tanh(e_{\text{video}} + \tilde{e}_{\text{video}})$. This formulation ensures that the cross-attended features are both modality-enhanced and modality-preserving, maintaining core signal fidelity while leveraging inter-modal cues for downstream transcription. Finally, the fused multimodal representation is obtained by concatenating the modality-enhanced embeddings along the feature dimension: $e_{\text{fused}} = [\,\hat{e}_{\text{audio}} \,\|\, \hat{e}_{\text{video}}\,] \in \mathbb{R}^{T \times 2d}$, where $\|$ denotes the concatenation operation. Following the cross-modal attention stage, the fused representation is passed through a linear projection layer $W_{\text{proj}} \in \mathbb{R}^{2d \times d}$ to reduce its dimensionality and align it with the expected input space of the subsequent transformer encoder ($\mathbb{R}^{T \times d}$). This operation preserves cross-modal semantic con-

tent while minimizing computational overhead. The resulting sequence is then processed by a transformer encoder T, which captures both local and global temporal dependencies through self-attention mechanisms. The transformer's output is then passed through a fully connected layer that maps each time step to a distribution over a fixed phonetic vocabulary. This vocabulary consists of language-independent phoneme units along with a special blank token required for alignment-free decoding. The cross-attended audio-visual representation is processed by a deep transformer stack comprising 24 self-attention layers. Each transformer block is parameterized with a hidden dimensionality of 1024, a feed-forward subnetwork of size 4096, and multi-head attention with 16 parallel heads. To enable training without the need for explicit frame-to-label alignment, the Connectionist Temporal Classification (CTC) loss [16] is employed.

3.2. Integrating pretrained language models for semantically grounded visual speech transcription

A hierarchical architecture is proposed that decomposes the visual speech recognition task into two functionally distinct stages: *phonetic encoding* and *language-conditioned decoding*. In the first stage, a Cross-Modal Transcriber processes time-aligned audio-visual input via a cross-attention mechanism to produce a sequence of phoneme-level tokens that encode pronunciation-specific information. This latent representation is deliberately designed to be *language-agnostic*, decoupling the perceptual modality from orthographic constraints. In the second stage, a pretrained LLM is employed as a neural decoder that maps the phonetic token sequence to well-formed text in the target language (highlighted in green in Fig. 1). This stage leverages the LLM's *semantic priors* and *contextual reasoning* to resolve phoneme ambiguities and enforce syntactic coherence, enabling robust transcription even in low-resource language settings.

To facilitate language-specific decoding, a lightweight natural language prompt is prepended to the phonetic transcription, explicitly instructing the Large Language Model to generate text in the desired target language (*e.g., "Convert the following phonetic transcription into its correct {Target Language} form:"*). This explicit conditioning guides the LLM toward the desired output language and format, ensuring consistent decoding across multilingual settings. This prompt-based conditioning enables the LLM to perform *semantic disambiguation*, *syntactic reconstruction*, and *orthographic adaptation* without any task-specific fine-tuning. Operating entirely in inference mode, the LLM leverages its pretrained multilingual representations to resolve pronunciation ambiguities and generate coherent, well-formed output across a wide range of languages. The proposed system architecture facilitates robust generalization and enables scalable transfer to previously unseen or under-resourced languages by separating perceptual encoding from language-specific decoding.

3.3. Multilingual adaptation and transfer for low-resource language

Recent advances in multimodal language modeling have demonstrated that directly conditioning pretrained LLMs on continuous speech representations (rather than intermediate symbolic abstractions such as phonemes) can yield superior contextual grounding, more fluent transcription, and enhanced disambiguation. Motivated by these findings, a unified cross-modal approach is proposed, which performs end-to-end modeling of perceptual and linguistic signals within a shared semantic space. Rather than first decoding phonetic units, authors' approach directly maps high-level fused audio-visual features to language model inputs, effectively collapsing the visual speech recognition and text generation pipeline into a single stage.

To achieve this, the output of the Cross-Modal Transcriber is leveraged, which encodes synchronized articulatory dynamics across modalities as feature sequences $e_{(a\text{-}v)} \in \mathbb{R}^{T \times d}$. Given the possible mismatch in temporal resolution and feature dimensionality between the multimodal encoder and the LLM, a lightweight adapter module is introduced that bridges these representational spaces (corresponding to the orange modules in Fig. 1). Specifically, the adapter first applies temporal compression via stride-1D convolution to produce a shorter sequence with $T' < T$, followed by a linear projection to match the LLM's token embedding space: $\hat{e}_{(a\text{-}v)} = \mathrm{Proj}(\mathrm{Compressed}(e_{(a\text{-}v)})) \in \mathbb{R}^{T' \times d_{\mathrm{LLM}}}$, where d_{LLM} is the LLM's token embedding dimension and $\mathrm{Proj}(\cdot)$ is a linear projection that maps the compressed feature sequence into the token embedding space of the LLM.

The transformed representation $\hat{e}_{(a\text{-}v)}$ is then prepended with a templated instruction prompt *(e.g., "Convert the following phonetic transcription into its correct {Target Language} language text:")*, which is first passed through a text embedding layer to obtain a dense feature representation (shown in orange in Fig. 1). The resulting sequence is fed into a pretrained LLM operating in frozen inference mode, leveraging its multilingual priors for semantic reconstruction, syntactic formatting, and script adaptation. To enable efficient fine-tuning in low-resource settings, the QLoRA paradigm [17] is adopted, updating only the adapter and LoRA (Low-Rank Adaptation) modules while keeping the base model parameters unchanged. In addition, the cross-attention mechanism in the Transcriber is optimized to refine modality alignment. This design is validated on Romanian, a previously unseen and underrepresented language, demonstrating both zero-shot generalization and rapid task adaptation with minimal supervision.

4. Experimental Evaluation

This section details the experimental protocol used to evaluate the proposed approach. The datasets employed for training and evaluation, the architectural components of the system, and the training procedures adopted for both zero-shot inference and supervised fine-tuning are described. In addition, implementation details, hyperparameter configurations, and the evaluation metrics used to quantify model performance are specified.

Dataset: Training and evaluation of the visual speech recognition approach are conducted using the Multilingual TEDx (mTEDx) corpus [18], a large-scale multilingual dataset originally curated for automatic speech recognition and speech translation tasks. Despite its initial design for audio-centric applications, mTEDx is well-suited for VSR due to its structured format, high-fidelity recordings, and controlled single-speaker settings. Each instance in the dataset comprises synchronized video and audio streams, along with human-verified transcriptions, making it an ideal resource for cross-lingual VSR research. To support low-resource language modeling, the RoVSR (Romanian Visual Speech Recognition) dataset is introduced as a Romanian audiovisual corpus constructed through a semi-automated pipeline. The dataset comprises 96502 segments (approximately 165 hours) sourced from YouTube, totaling over 1.14 million words and 80303 phrases. The image sequences collectively contain more than 14.6 million frames, with segment durations ranging from 3 seconds to 15 seconds, and a median duration of 6 seconds. All segments contain a single visible speaker, recorded under diverse, in-the-wild conditions. Preprocessing involves face detection, tracking and cropping using S3FD [15], followed by audiovisual synchronization filtering via SyncNet [19], retaining only clips with alignment offsets within ± 10 frames. Transcriptions are generated with WhisperX [2], enabling accurate word-level alignment without manual annotation, as validated in prior work [9]. The dataset is partitioned into train (86281 segments) and test (10221 segments), with no speaker overlap. Manual corrections on

the test set ensure evaluation reliability.

Implementation details: For linguistic decoding, the LLaMA 3.2B [20] pretrained language model is employed in two modes: (1) frozen inference, where only the Cross-Modal Transcriber is trained while the LLM remains fixed; and (2) instruction-conditioned fine-tuning using QLoRA [17], which injects trainable low-rank adaptation weights into selected transformer layers, allowing for efficient domain adaptation while keeping the base LLM parameters unchanged. All experiments are conducted on a computer setup comprising $2\times$ NVIDIA RTX 6000 Ada GPUs. Training is carried out for 40 epochs, using the AdamW optimizer with a cosine learning rate schedule and 5-epoch warm-up. The learning rate peaks at 4×10^{-4}. To manage memory and batch sizing, the total number of frames per batch is capped at 1800 frames, and gradient accumulation is set to 8 to simulate larger batch sizes.

Quantitative experimental results: To evaluate the effectiveness of the proposed architecture, two standard metrics are used: Word Error Rate (WER) and Character Error Rate (CER). Experiments are conducted on all Latin-script languages included in the mTEDx corpus (*i.e.,* French, Italian, Spanish and Portuguese). Training is performed in a multimodal setting, where both audio and visual streams are available to facilitate the learning of temporally aligned and semantically enriched cross-modal representations. During inference, however, a real-world visual speech recognition scenario is simulated by disabling the audio stream entirely, thus enforcing a strict visual-only decoding condition. Table 1 reports the lip-reading performance of authors' approach under two distinct architectural configurations, both initially trained exclusively on the multilingual mTEDx corpus: (1) *Two-Stage Phoneme-to-Text Pipeline (PtoT):* A Cross-Modal Transcriber uses cross-attention to generate phoneme-level tokens from time-aligned visual (and optionally audio) input. A pretrained LLM then maps the phonetic sequence to coherent text in the target language, making this setup suitable for studying language transfer and phoneme-to-text disambiguation. (2) *End-to-End Visual-to-Text Model (VtoT):* This approach removes the symbolic transcription step. Fused visual features are projected directly into the LLM's token embedding space via a lightweight adapter that performs temporal compression and dimensional alignment. The LLM then decodes continuous visual representations directly into natural language, capturing both articulation and linguistic context in a single stage.

Table 1. Lip reading performance evaluation under two configurations: (1) phoneme-to-text decoding via an LLM (PtoT), and (2) end-to-end decoding with direct visual feature integration (VtoT)

Methods	**Portuguese**		**Spanish**		**French**		**Italian**	
	WER↓	CER↓	WER↓	CER↓	WER↓	CER↓	WER↓	CER↓
PtoT	53.41	30.05	49.82	28.77	64.12	41.25	53.49	32.47
VtoT	45.23	26.21	44.18	22.48	54.47	35.45	48.85	27.67

The results in Table 1 reveal a consistent and statistically supported performance advantage of the end-to-end visual-to-text configuration over the phoneme-to-text approach across all evaluated languages, with averages over repeated trials confirming robustness to LLM-induced stochastic variability. This improvement can be attributed to several architectural and representational factors. First, bypassing intermediate symbolic representations (*i.e.,* phoneme tokens) allows the VtoT system to preserve richer contextual and prosodic information embedded in the continuous audio-visual features. This reduces information loss that typically occurs during phoneme-level abstraction and mitigates phoneme-to-grapheme ambiguities, which can vary

significantly across languages. Second, the direct integration of fused visual features into the LLM's embedding space enables the model to exploit the full expressive power of the LLM for syntactic reconstruction and semantic disambiguation. Finally, the visual-only inference setup leverages the VtoT model's ability to learn a unified perception-to-language mapping, avoiding intermediate decoding stages that may introduce compounding errors, particularly when dealing with visually ambiguous or fine-grained articulatory cues.

Comparative experimental results: Tabel 2 presents a comparative performance analysis on the mTEDx dataset against state-of-the-art methods [13], [21], [9], using a consistent evaluation protocol. Since some baselines report only Word Error Rate, the comparison is restricted to WER in those cases to ensure a fair and consistent evaluation.

Table 2. Performance comparison of lip-reading systems across Latin-derived languages

Methods	**Portuguese**		**Spanish**		**French**		**Italian**	
	WER↓	CER↓	WER↓	CER↓	WER↓	CER↓	WER↓	CER↓
Kim *et al.* [13]	58.57	37.68	56.96	32.28	64.92	42.69	59.74	33.30
Ma *et al.* [21]	61.50	–	56.30	–	66.20	–	57.40	–
Yeo *et al.* [9]	47.89	29.74	45.71	25.90	58.30	37.85	51.79	29.83
Ours (*VtoT*)	45.23	26.21	44.18	22.48	54.47	35.45	48.85	27.67

The results in Table 2 show that authors' approach achieves superior performance across all evaluated Latin-derived languages, outperforming prior state-of-the-art approaches in both WER and CER. The consistent improvement reflects the effectiveness of authors' unified cross-modal architecture, which uses audio-visual features and language model integration to enhance decoding accuracy. Unlike earlier approaches that depend on audio supervision or perform only shallow modality fusion, authors' approach leverages fine-grained cross-attention and strong language modeling to better handle pronunciation variability and visually ambiguous cues.

Monolingual vs. multilingual: Table 3 compares monolingual and multilingual training strategies for visual speech recognition in Latin-derived languages. In the monolingual setup, a separate model is trained per language using only language-specific data, leading to higher WERs due to limited training diversity. In contrast, the multilingual model is trained jointly on all languages, enabling the learning of language-agnostic phonetic patterns and more robust representations through cross-lingual sharing. This unified approach yields substantially lower WERs across all languages, showing the advantages of parameter sharing in low-data settings.

Evaluation of the LLM stochastic variability To systematically address the stochastic variability introduced by the prompting mechanisms of LLMs, three independent experimental trials were conducted under identical settings. This repetition ensured that the reported results were not influenced by random variations across inference runs caused by the probabilistic nature of LLM decoding (*e.g.*, temperature-based sampling and token-level randomness). The performance values reported in Tables 1–3 represent the average results computed across these three trials, while Table 4 presents the individual trial outcomes along with their mean and standard deviation.

Table 3. Impact of multilingual training on lip reading accuracy

Methods	**Metric**	**Portuguese**	**Spanish**	**French**	**Italian**
Monolingual	WER↓	57.24	50.77	65.62	58.42
Multilingual	WER↓	45.23	44.18	54.47	48.85

The observed standard deviations (ranging from 0.25 to 0.96 WER points, corresponding to less than 2% absolute variation relative to the mean) indicate stable and consistent performance across independent runs. These findings suggest that the results are reproducible under the described experimental conditions.

Table 4. WER when performing repeated experimental trials

Experiment	Portuguese	Spanish	French	Italian
1	44.89	43.98	53.87	48.63
2	46.31	45.18	55.21	49.12
3	44.49	43.39	54.33	48.80
Mean	45.23	44.18	54.47	48.85
Standard deviation	0.96	0.91	0.68	0.25

Evaluation on low-resource language: A targeted evaluation of the visual speech recognition approach is performed on Romanian, a previously unseen low-resource language, under two systematically distinct experimental conditions: zero-shot inference and supervised fine-tuning. These scenarios are designed to assess both the cross-lingual generalization capabilities and the adaptability of the proposed architecture. Importantly, this evaluation leverages the RoVSR benchmark, which to authors' knowledge represents the largest audiovisual dataset for Romanian VSR and constitutes the first systematic resource of its kind. Since no prior VSR models have been exposed to Romanian in training, this benchmark provides a unique opportunity to assess how architectures generalize to genuinely underrepresented languages.

In the zero-shot setting, the model is trained solely on the multilingual mTEDx corpus, with no exposure to Romanian speech, and is directly evaluated on the RoVSR benchmark. This setting proves the effectiveness of the phoneme-level, language-agnostic latent space induced by the Cross-Modal Transcriber, and its ability to support robust transfer to typologically similar, but unseen, languages. Notably, since the Romanian phoneme inventory shares substantial overlap with other Latin-derived languages, this configuration isolates the model's reliance on learned articulatory patterns rather than language-specific lexical or orthographic information. In the supervised adaptation scenario, the approach is fine-tuned using the RoVSR dataset. The cross-modal attention layers, temporal adapters, and LoRA-injected parameters in the language model are selectively fine-tuned using low-resource QLoRA [17] training, while the core backbone remains frozen. This lightweight fine-tuning allows the system to refine modality alignment and lexical decoding without overfitting or requiring full-model retraining.

Table 5 presents the performance of authors' proposed lip-reading approach on Romanian, evaluated under two conditions: zero-shot inference, where the model has never seen Romanian data during training, and supervised fine-tuning, where the model is adapted using the RoVSR dataset.

Table 5. Performance comparison on RoVSR under zero-shot and fine-tuned settings

Methods	Romanian	
	WER↓	CER↓
Zero-shot inference on RoVSR	81.23	50.18
Supervised fine-tuning on RoVSR	59.14	38.21

Supervised fine-tuning on Romanian results in significant improvements relative to the zero-

shot condition, in both WER and CER, demonstrating the model's capacity to adapt its internal representations to the phonetic and orthographic regularities of the target language. These gains are attributable to the synergistic effect of two key factors: (1) a language-agnostic phoneme-level interface that facilitates cross-lingual transfer by decoupling pronunciation modeling from grapheme-specific constraints, and (2) multilingual pretraining across Latin-derived languages, which provides a strong cross-modal prior and promotes parameter generalization across related linguistic domains. The fine-tuned model exhibits enhanced sensitivity to subtle visual cues specific to Romanian articulation, while preserving the global structural priors acquired during multilingual training. This contrasts with language-specific VSR systems, which are often overfitted to small datasets and lack the representational flexibility required for efficient adaptation. By comparison, authors' unified architecture adapts quickly to new languages through lightweight fine-tuning, without retraining or altering the underlying model.

5. Conclusions

This paper introduced a modular and extensible approach to visual speech recognition (VSR) that addresses the challenges of cross-lingual transfer and generalization to low-resource languages. The approach follows a two-stage design: a Cross-Modal Transcriber that encodes synchronized audio-visual signals into a phoneme-level representation, and a pretrained Large Language Model (LLM) that decodes this representation into fluent, language-specific text. An end-to-end variant embeds fused audio-visual features directly into the LLM via a lightweight adapter, enabling joint perceptual-linguistic modeling without intermediate transcriptions.

Comprehensive experiments on the mTEDx corpus across multiple Latin-script languages show that the proposed system outperforms prior methods under visual-only inference, with the visual-to-text (VtoT) configuration achieving the best results. Multilingual training further improves performance through parameter sharing and reduces overfitting in low-data regimes. To assess transferability, the 165-hour Romanian in-the-wild dataset RoVSR was introduced. Zero-shot evaluation confirmed strong cross-lingual generalization, while lightweight fine-tuning yielded significant improvements, highlighting the approach's adaptability. Future work will extend the approach to non-Latin scripts, improve robustness to speaker and pose variation, and optimize real-time inference for deployment.

Acknowledgements. This work was supported by two grants of the Ministry of Research, Innovation and Digitization, CCCDI – UEFISCDI, project number PN-IV-P6-6.3-SOL-2024-2-0238 and PN-IV-P6-6.3-SOL-2024-0049, within PNCDI IV.

References

[1] Y.-A. DAHOU, S. NARAYAN, H. BOUSSAID, E. ALMAZROUEI and M. DEBBAH, *Lip2Vec: Efficient and robust visual speech recognition via latent-to-latent visual to audio representation mapping*, Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023, pp. 1–10.

[2] M. BAIN, J. HUH, T. HAN and A. ZISSERMAN, *WhisperX: Time-accurate speech transcription of long-form audio*, Proceedings of Interspeech, Dublin, Ireland, 2023, pp. 1–5.

[3] Y.-M. ASSAEL, B. SHILLINGFORD, S. WHITESON and N. DE FREITAS, *LipNet: Sentence-level lipreading*, arXiv:1611.01599, 2016.

[4] A. VASWANI, N. SHAZEER, N. PARMAR, and I. POLOSUKHIN, *Attention is all you need*, Proceedings of Advances in Neural Information Processing Systems, Long Beach, CA, USA, 2017, vol. 30, pp. 5998–6008.

[5] P. MA, S. PETRIDIS and M. PANTIC, *End-to-end audio-visual speech recognition with conformers*, Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, Canada, 2021, pp. 7613–7617.

[6] H. HAN, M. ANWAR, J. PINO, W.-N. HSU, M. CARPUAT, B. SHI and C. WANG, *XLAVS-R: Cross-lingual audio-visual speech representation learning for noise-robust speech perception*, arXiv:2403.14402, 2024.

[7] A. ROUDITCHENKO, H. KUEHNE, R. FERIS and J. GLASS, *Whisper-Flamingo: Integrating visual features into Whisper for audio-visual speech recognition and translation*, arXiv:2406.10082, 2024.

[8] B. SHI, W.-N. HSU, K. LAKHOTIA and A. MOHAMED, *Learning audio-visual speech representation by masked multimodal cluster prediction*, arXiv:2209.15326, 2022.

[9] J.-H. YEO, M. KIM, S. WATANABE and Y.-M. RO, *Visual speech recognition for languages with limited labeled data using automatic labels from Whisper*, Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Seoul, South Korea, 2024, pp. 10471–10475.

[10] T. AFOURAS, J.-S. CHUNG and A. ZISSERMAN, *LRS3-TED: A large-scale dataset for visual speech recognition*, arXiv:1809.00496, 2018.

[11] O. CHANG, H. LIAO, D. SERDYUK, A. SHAHY and O. SIOHAN, *Conformer is all you need for visual speech recognition*, Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Seoul, South Korea, 2024, pp. 10136–10140.

[12] K.-R. PRAJWAL, S. HEGDE and A. ZISSERMAN, *Scaling multilingual visual speech recognition*, Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Hyderabad, India, 2025, pp. 1–5.

[13] M. KIM, J. PARK, D. KIM, Y. LEE and G. KIM, *Lip reading for low-resource languages by learning and combining general speech knowledge and language-specific knowledge*, Proceedings of IEEE/CVF International Conference on Computer Vision, Paris, France, 2023, pp. 1–10.

[14] J. ZHAO, V. PRATAP and M. AULI, *Scaling a simple approach to zero-shot speech recognition*, arXiv:2407.17852, 2024.

[15] S. ZHANG, X. ZHU, Z. LEI, H. SHI, X. WANG and S.-Z. LI, *S3FD: Single shot scale-invariant face detector*, Proceedings of IEEE International Conference on Computer Vision, Venice, Italy, 2017, pp. 192–201.

[16] A. GRAVES, S. FERNÁNDEZ, F. GOMEZ and J. SCHMIDHUBER, *Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks*, Proceedings of International Conference on Machine Learning, Pittsburgh, PA, USA, 2006, pp. 369–376.

[17] T. DETTMERS, A. PAGNONI, A. HOLTZMAN and L. ZETTLEMOYER, *QLoRA: Efficient finetuning of quantized LLMs*, arXiv:2305.14314, 2023.

[18] E. SALESKY, M. WIESNER, J. BREMERMAN, R. CATTONI, M. NEGRI, M. TURCHI, D.-W. OARD and M. POST, *The Multilingual TEDx corpus for speech recognition and translation*, Proceedings of Interspeech, Shanghai, China, 2021, pp. 3655–3659.

[19] J.-S. CHUNG and A. ZISSERMAN, *Out of time: Automated lip sync in the wild*, Workshop on Multi-view Lip-reading, Asian Conference on Computer Vision, Taipei, Taiwan, 2016, pp. 1–7.

[20] A. GRATTAFIORI, G. ROCHE, T. WOLF, T. SCIAKY, L. MOU et al., *The LLaMA 3 herd of models*, arXiv:2407.21783, 2024.

[21] P. MA, S. PETRIDIS and M. PANTIC, *Visual speech recognition for multiple languages in the wild*, Nature Machine Intelligence **4**(11), 2022, pp. 930–939.