

Article

Developing a Kazakh Audio–Visual Multimodal Speech Recognition Model Based on Hierarchical and Cross-Modal Attention

Turdybek Kurmetkan ^{1,2,*} , Orken Mamyrbayev ² , Adem Tekerek ³ and Ainur Toleu ⁴ 

¹ Department of Information Systems, Faculty of Information and Artificial Intelligence, Farabi University, 050040 Almaty, Kazakhstan

² Institute of Information and Computational Technologies, 050010 Almaty, Kazakhstan; morkenj@mail.ru

³ Department of Computer Engineering, Faculty of Technology, Gazi University, 06560 Ankara, Türkiye; atekerek@gazi.edu.tr

⁴ National Scientific and Practical Center “Til–Kazyna” Named After Shaysultan Shayakhmetov, 010000 Astana, Kazakhstan; toleuainur@gmail.com

* Correspondence: turdybek.narat86@gmail.com

Abstract

This study presents an audio–visual speech recognition (AVSR) model for Kazakh that jointly exploits audio and visual channels. The study introduces QazAVSR, a 57 h dataset collected from 271 speakers, and extracts synchronized audio signals and lip-region video sequences using FFmpeg 7.0, Dlib 19.24, and OpenCV 4.9.0. The proposed architecture uses the self-supervised HuBERT_BASE model in the audio branch and an ImageNet-pretrained ViT-B/16 model in the visual branch. Audio and visual representations are fused by a three-layer BiModalHformer block, where intra- and cross-attention operations are performed at each level. Extensive experimental validation, supplemented by rigorous paired bootstrap resampling significance tests, demonstrates that the full multimodal BiModalHformer model achieves a highly robust average character error rate (CER) of 31.2% and a Word Error Rate (WER) of 43.1%. These results significantly outperform traditional audio-only, video-only, and standard representation-level fusion baselines. Furthermore, comparisons against powerful external baseline architectures—including Whisper-Small and AV-HuBERT configurations rigorously adapted for the Kazakh language—statistically validate the architectural efficacy of the BiModalHformer framework. Additional systematic evaluations utilizing extended metrics such as the Match Error Rate (MER), word information preserved (WIP), and the Multimodal Synergy Index (MSI) confirm that the full audio–visual configuration preserves lexical information significantly more effectively. Finally, extensive noise perturbation experiments confirm that the multimodal architecture exhibits superior structural robustness to complex acoustic distortions, including environmental noise, synthetic room reverberation, and overlapping speech topologies.

Keywords: AVSR; Kazakh language; multimodal speech recognition; HuBERT; Vision Transformer; hierarchical attention; cross-modal attention



Academic Editors: Gerasimos Vonitsanos and John Garofalakis

Received: 17 June 2026

Revised: 30 July 2026

Accepted: 4 August 2026

Published: 6 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Automatic speech recognition (ASR) systems have improved substantially in recent years thanks to advances in deep learning. However, in practical environments, acoustic noise, reverberation, differences in microphone quality, and speaker-specific articulation

reduce the performance of systems that rely only on audio. In such settings, the visual channel, especially lip movements, can serve as an important auxiliary source of information when the acoustic signal is incomplete or degraded.

Kazakh is an agglutinative language with a large number of word forms, complex morphology, and considerable articulatory variation in continuous speech. Therefore, combining acoustic features with visual cues is a promising direction for improving Kazakh speech recognition. Although studies on Kazakh ASR based on DNNs, Transformers, Conformers, HiFormers, and multilingual learning have achieved strong results [1–10], dedicated audio–visual models and datasets for Kazakh remain limited.

A central challenge in AVSR is correctly modelling the dependencies between modalities. Movements of the face and mouth do not duplicate the acoustic channel, but they help to clarify labial sounds, articulation boundaries, and speech rhythm. Therefore, simply concatenating audio and visual branches is insufficient; their temporal and hierarchical relationships can be effectively integrated using attention mechanisms.

The purpose of this study is to propose an AVSR model for Kazakh based on a specially collected audio–visual corpus. The model combines HuBERT and Vision Transformer encoders with hierarchical and cross-modal attention, and it is evaluated against audio-only, video-only, representation-level fusion, and reduced-attention baselines. The main scientific contributions are threefold: the construction of the multi-topic QazAVSR audio–visual dataset; the design of a hierarchical, cross-modal attention-based fusion scheme; and a comparative evaluation of the full 57 h corpus using extended metrics and robustness analysis.

Literature Survey

Research on Kazakh ASR has developed in several stages. Early work, starting primarily around 2019 (which is considerably later than ASR development for major languages), investigated the use of deep neural networks to improve recognition of Kazakh speech [1]. Later studies focused on open speech corpora, multilingual end-to-end architectures, self-supervised pretrained models, and joint training across Turkic languages (such as the strong phonological and morphological links between Kazakh and Turkish) [2–7]. These studies demonstrated the importance of transfer and multilingual learning for reducing data scarcity.

Transformer-based variants for Kazakh have also been actively studied. HiFormer strengthens interactions among representations at different levels, while the Conformer architecture efficiently models both local and global temporal dependencies [8,9]. Extended Transformer variants have also been analyzed, showing that architectural choices directly affect Kazakh speech recognition accuracy [10].

ASR studies for Turkic languages reveal shared challenges across Kazakh, Uzbek, Turkish, Kyrgyz, and Azerbaijani, as well as related agglutinative languages: limited data, complex orthography and phonetics, rich morphological variation, and domain shift. KSC2, USC, Turkish speech corpora, and multilingual Turkic ASR studies demonstrate the effectiveness of pretrained and multilingual models in low-resource conditions [11–14]. Prominent works in the recognition of Kazakh and Turkic languages are listed in Table 1.

AVSR research for Turkic languages is much more limited than ASR research. Most AVSR work has focused on English, German, and Chinese. End-to-end architectures have enabled direct learning of modality dependencies [15,16]. AV-HuBERT showed that masked multimodal cluster prediction can learn effective joint audio–visual representations [17]. Whisper-Flamingo improved AVSR and audio–visual translation by integrating visual features into a large pretrained decoder [18]. MLCA-AVSR introduced cross-attention modules into multiple encoder layers and demonstrated the benefit of multi-level fusion [19].

Table 1. Notable ASR studies for Turkic languages.

First Author	Work/Dataset	Corpus Size	Main Method	Contribution
Khassanov et al. [2]	KSC/Kazakh Speech Corpus	~332 h	E2E Transformer + LM + SpeedPerturb + SpecAugment	Open-source corpus and strong baseline for Kazakh ASR
Mussakhojayeva et al. [11]	KSC2	>1000 h	Transformer-based streaming ASR	Industrial-scale Kazakh speech corpus
Musaev et al. [12]	USC/Uzbek Speech Corpus	105 h	DNN-HMM and E2E ASR	Open corpus for Uzbek with baseline experiments
Polat & Oyucu [13]	Turkish speech/text corpus	≈90 h plus verified/unverified speech subsets	GMM-HMM, DNN/ANN + n-gram LM	Large Turkish corpus and initial ASR baselines
Mussakhojayeva et al. [6]	Turkic multilingual ASR	10 languages	Monolingual and multilingual ASR	Effectiveness of transfer across related Turkic languages
Bekarystankyzy et al. [7]	Low-resource Turkic ASR	Five Cyrillic-script Turkic languages from Common Voice	Conformer + Transformer + CTC/Attention + LM	Benefits of common alphabets and linguistic similarity
Görmez [14]	Turkish ASR supported by LM	Turkish speech data	Customized deep-learning ASR + language model	Example of improving Turkish ASR with language-model support

The proposed BiModalHformer architecture fundamentally distinguishes itself from established paradigms through a mathematically defined algorithmic innovation. Unlike models such as AV-HuBERT, which rely on masked prediction objectives, or Whisper-Flamingo, which executes late-stage conditional fusion, the BiModalHformer incorporates a specialized inter-layer history passing mechanism directly within its core fusion blocks. Crucially, layer *l* does not merely query layer *l* of the complementary modality; rather, it queries a concatenated memory bank compiled from all preceding layers (0 through *l*-1). This specific structural innovation is theoretically essential for capturing complex, long-range morpho-phonological dependencies in highly agglutinative languages such as Kazakh.

More broadly, prior literature demonstrates that the success of audio–visual speech recognition (AVSR) systems relies on three foundational pillars: a high-quality dataset, robust modality-specific encoders, and an effective fusion mechanism that models cross-modal interactions. Given the scarcity of such multimodal research for Kazakh, the construction of the QazAVSR corpus alongside the proposed BiModalHformer architecture represents a vital step toward advancing speech recognition technologies for low-resource agglutinative languages.

2. Materials and Methods

2.1. QazAVSR Dataset

The study utilizes QazAVSR, an audio–visual dataset created specifically for the Kazakh language. In terms of volumetric metrics, the finalized corpus consists of approximately 57 h of combined audio–visual recordings, comprising exactly 38,452 distinct, segmented utterances collected from 271 unique speakers. Utterances range from 3 to 15 words, with an average duration of 5.4 s, capturing both scripted vlogs and spontaneous colloquial speech. Linguistically, the foundational character vocabulary comprises 40 distinct tokens, whereas the overarching unique word vocabulary spans approximately 42,000 words, highlighting the massive lexical variety induced by Kazakh agglutinative morphology.

To prevent the model from overfitting to specific speaker traits or narrow topical domains, the dataset is diversified by gender, age, and topic. The speaker distribution is 48% male (130 speakers) and 52% female (141 speakers), with the primary age cohorts (20–35 and 36–50 years old) accounting for 74% of all subjects, providing com-

prehensive coverage of adult articulatory patterns (Table 2). To establish a rigorous, speaker-independent evaluation protocol, the corpus was strictly partitioned using an 80/10/10 split ratio. Specifically, the training subset accounts for 80% of the dataset, encompassing approximately 46 h of recordings across 214 unique speakers. The validation split comprises 10% of the corpus, containing roughly 5.5 h of data from 28 speakers, while the testing split incorporates the remaining 10%, consisting of approximately 5.5 h of recordings from 29 distinct speakers.

Table 2. Age and gender distribution of speakers in the QazAVSR corpus.

Age Range	Male	Female	Total
10–19 years	5% (13)	7% (19)	12% (32)
20–35 years	17% (46)	19% (51)	36% (97)
36–50 years	18% (49)	20% (54)	38% (103)
50+ years	8% (22)	6% (17)	14% (39)
Total	48% (130)	52% (141)	100% (271)

To ensure high natural variability and real-world applicability, the majority of the dataset was collected from open-source social media platforms, encompassing acoustic environments from quiet indoor settings to noisy outdoor “in-the-wild” conditions. Video sequences, captured via diverse devices ranging from smartphones to professional cameras, were uniformly processed to a standard framerate of 25 frames per second (fps), while extracted audio was downsampled to 16 kHz. Data curation involved automated downloading, manual screening to guarantee clear frontal facial visibility, and precise alignment validation. Crucially, the synchronization error tolerance was strictly capped at a maximum of 40 milliseconds (roughly equivalent to a single video frame at 25 fps). Any source file demonstrating an audio–visual offset exceeding this 40 ms threshold during the experiment was disregarded.

As shown in Figure 1, the thematic composition of the corpus is also diverse. Daily news, literature, politics, economy, religion, and traditions are dominant, but medicine, construction, education, science, and law are also represented. This broadens the model’s lexical coverage and improves its cross-domain stability.

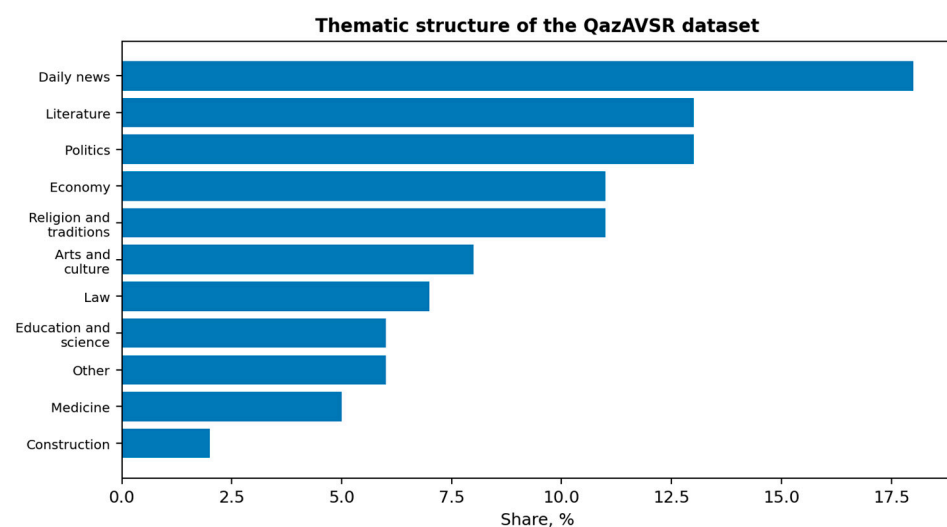


Figure 1. Thematic structure of the QazAVSR dataset.

Ethical Aspects: Given the use of facial video data sourced from public platforms (YouTube and TikTok), the dataset was curated in strict adherence to fair-use guidelines

for academic research. While the original videos were publicly accessible, privacy and anonymization protocols were rigorously implemented: all metadata, original usernames, URLs, and channel identifiers were stripped from the dataset to protect speakers' identities. The data processing pipeline received ethical approval from the local Institutional Review Board. Due to privacy constraints and platform terms of service, data sharing is restricted; the facial video subset will be shared only with verified academic researchers upon request, subject to a strict non-disclosure agreement.

2.2. Data Preprocessing

Preparing the data for the AVSR model involved several stages. First, audio and video streams were separated and temporally aligned using FFmpeg [20]. The audio was converted to a single-channel 16 kHz format and normalized using z-score normalization:

$$x' = (x - \mu) / \sigma \quad (1)$$

Second, Dlib [21] was used to detect facial landmarks. The mouth and lip region were localized in each frame, so the visual stream focused on the part of the face directly related to speech. Third, frames were standardized to 25 fps and converted to 224×224 RGB format using OpenCV [22]. Each sample was limited to 200 frames, corresponding to 8 s of visual dynamics.

Text transcripts were lowercased, $\ddot{e} \rightarrow e$ and $h \rightarrow \text{h}$ transformations (mapping the Cyrillic 'h' to the specific Kazakh 'h' character for phonetic consistency) were applied, punctuation and non-transcription symbols were removed using regular expressions. Character-level tokenisation was used because the agglutinative structure of Kazakh increases the variability of word and subword units, while character units are more robust to rare word forms. Subword tokenisation methods, such as SentencePiece, can be considered as alternatives for additional experiments [23].

2.3. Ethical Considerations and Data Privacy

The dataset constructed for this study consists of audio–visual recordings collected from publicly accessible platforms, primarily YouTube and TikTok. This study and the associated data collection protocol were formally reviewed and approved by the Ethics Committee of the Institute of Information and Computational Technologies (Protocol No: 4, Date: 28 May 2026).

Anonymization, Biometric Risk Mitigation, and Platform Terms: In audio–visual speech recognition (AVSR) research, physical anonymization of facial features (e.g., blurring or masking) is fundamentally unfeasible, as it destroys the visual cues (lip-reading dynamics) strictly required for model training. To mitigate the inherent biometric identifiability of unmasked facial and vocal media, all identifiable metadata—including usernames, channel names, precise URL links, and associated tags—have been permanently removed from the database architecture, referencing subjects exclusively via randomized, non-identifiable numerical IDs. Furthermore, raw media assets are hosted on an encrypted, access-controlled institutional server rather than open cloud repositories.

To fully comply with platform Terms of Service and adhere to the “Fair Use” doctrine for non-commercial academic research, the compiled dataset will not be publicly hosted or freely distributed on the internet. Access is strictly restricted to verified academic researchers who must execute a binding Data Use Agreement (DUA). This DUA explicitly prohibits commercial exploitation, bans the deployment of secondary forensic facial/voice biometric recognition algorithms designed to re-identify subjects, and legally binds researchers to non-redistribution. Finally, the curation framework incorporates a formal Right-to-Erasure protocol: if an original content creator privatizes, removes, or formally

requests the deletion of their video from the source platform, their corresponding audio–visual segment is permanently expunged from the QazAVSR corpus upon notification.

2.4. Proposed AVSR Architecture

The proposed AVSR architecture is designed to capture and fuse complementary audio–visual information effectively. It consists of three primary stages: modality-specific encoders, a hierarchical cross-modal fusion module (BiModalHformer), and a joint decoding mechanism (Figure 2).

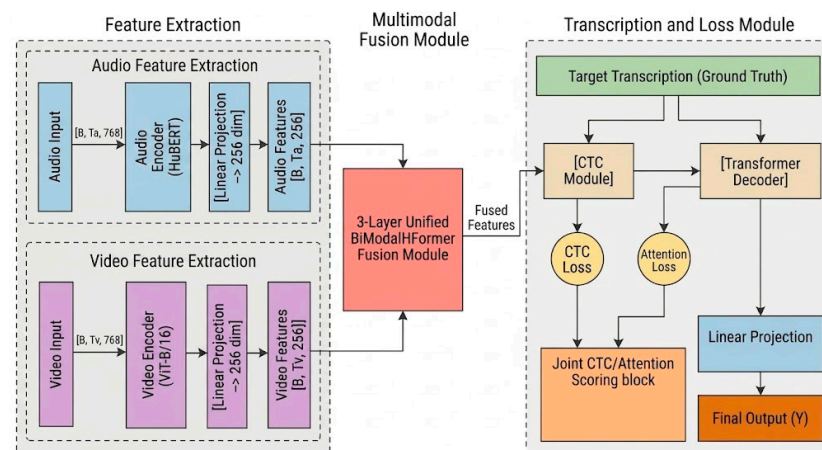


Figure 2. Overall architecture of the proposed model.

Modality-Specific Encoders: The system processes two synchronized inputs: an audio waveform (X_a) and a sequence of video frames (X_v) focusing on the lip region. For the acoustic stream, the architecture employs the pretrained HuBERT_BASE model. HuBERT (Hidden-Unit BERT) has been decisively demonstrated to produce highly robust, noise-resilient speech representations. It achieves this by learning through a self-supervised masked prediction objective, forcing the model to infer missing acoustic tokens by analyzing the continuous, unmasked surrounding audio context. In this architecture, the lower layers of HuBERT are frozen to retain universal acoustic feature extraction capabilities, while the upper layers are fine-tuned to map specifically to Kazakh phonotactics [24]. Concurrently, the visual stream processes the lip-region frame sequences using a Vision Transformer (ViT-B/16), pretrained on ImageNet, to extract spatial–visual dynamics [25]. An ImageNet-pretrained ViT-B/16 model was explicitly chosen for this branch instead of a dedicated, speech-specific visual frontend (such as a 3D-ResNet trained on massive datasets like Lipreading In The Wild). This architectural decision was driven by fundamental data constraints: pretraining a dedicated visual frontend from scratch requires thousands of hours of high-quality, synchronized video data, which currently does not exist for the Kazakh language. Furthermore, employing English-trained lip-reading models risks introducing subtle inductive biases toward English visemes. Consequently, the ViT-B/16 architecture provides highly generalizable, language-agnostic spatial feature extraction capabilities. To facilitate meaningful integration, the outputs from both modality-specific encoders are passed through linear projection layers, mapping their respective feature dimensions into a shared latent space with a dimensionality of $d_{\text{model}} = 256$:

$$H_a = \text{Enc}_a(X_a), \quad H_v = \text{Enc}_v(X_v) \quad (2)$$

where X_a is the input raw acoustic waveform of speech and X_v is the input visual sequence of frames corresponding to the lip region. Enc_a denotes the acoustic encoder function.

Encv denotes the visual encoder function. H_a and H_v represent the projected initial hidden feature representations.

BiModalHformer Fusion Module. The aligned features are fed into a three-layer Unified BiModalHformer fusion module. The key idea of this architecture is to organize information exchange between modalities not just at a single level, but across several hierarchical levels. The lower level captures articulatory details, the middle level models temporal-shape structure, and the upper level preserves semantic and contextual information.

At each layer, the architecture executes three distinct operations: (1) intra-layer attention, which models temporal dependencies within each modality independently; (2) inter-layer history passing, which maintains contextual memory from earlier processing stages; and (3) cross-modal attention, where audio queries visual features, and vice versa (Figure 3).

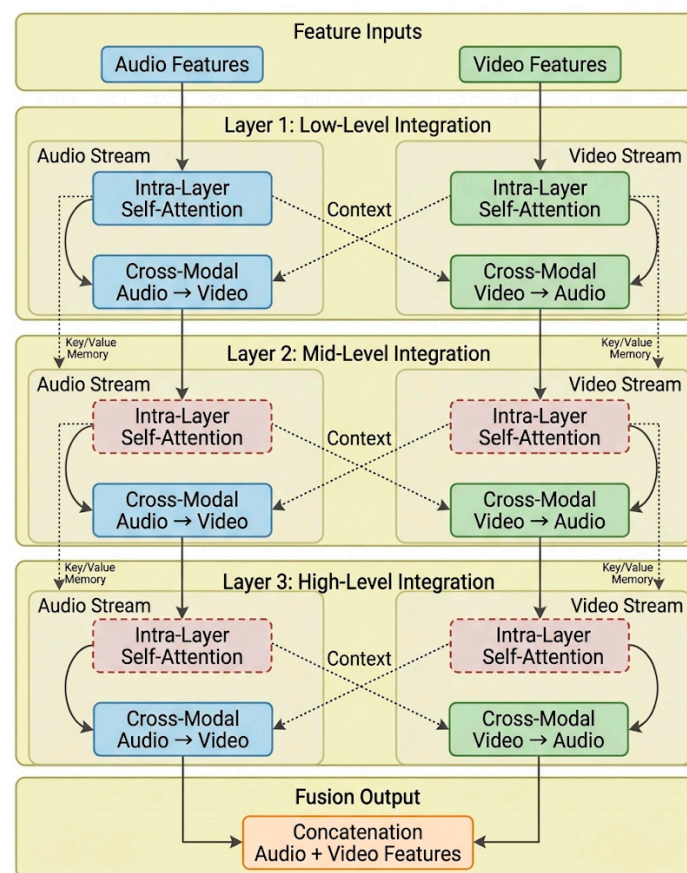


Figure 3. Hierarchical and cross-modal fusion model.

Unlike simple late-fusion approaches or existing architectures like MLCA-AVSR, our BiModalHformer fundamentally distinguishes itself by incorporating this inter-layer history passing mechanism. This ensures that contextual memory from previous processing stages is preserved and propagated through the hierarchy, which is particularly beneficial for managing the complex agglutinative morphology of the Kazakh language.

The standard attention mechanism is computed as follows:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (3)$$

where Q , K and V denote the Query, Key, and Value matrices, respectively. T represents the matrix transposition operator. d_k is the scaling factor corresponding to the dimension of the keys.

The multi-head attention (MHA) mechanism combines multiple projected heads:

$$head_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (4)$$

$$MHA(Q, K, V) = \text{Concat}(head_1, \dots, head_h)W^O \quad (5)$$

where $head_i$ is the output of the i -th individual attention head and the learnable parameter projection weights for the i -th head and h indicates the total number of attention heads. Concat denotes the tensor concatenation operation along the feature dimension. W^O is the final linear output projection weight matrix.

To realize the inter-layer history passing, the features from the previous layer $l-1$ are aggregated with historical memory. The accumulated history context is also represented. The history-passed representation is computed before cross-modal fusion:

$$M_a^l = [H_a^0; \dots; H_a^{l-1}], \quad M_v^l = [H_v^0; \dots; H_v^{l-1}] \quad (6)$$

$$\tilde{H}_a^l = MHA(Q = H_a^{l-1}, K = M_a^l, V = M_a^l), \quad (7)$$

$$\tilde{H}_v^l = MHA(Q = H_v^{l-1}, K = M_v^l, V = M_v^l) \quad (8)$$

where l denotes the current layer index in the hierarchical fusion hierarchy ($l \in \{1, 2, 3\}$). M_a^l and M_v^l represents the aggregated audio and video historical context matrices, respectively, spanning from the initial layer 0 up to layer $l-1$. $\tilde{H}_a^l$ and $\tilde{H}_v^l$ are the refined audio and video feature representations are updated with historical cross-layer dependencies H_a^{l-1} and H_v^{l-1} serve as queries from the immediate preceding layer.

Following history passing, bidirectional cross-modal attention is applied at the l -th hierarchical level to allow each modality to query the other's refined features:

$$Z_{a \leftarrow v}^l = MHA(Q = \tilde{H}_a^l, K = \tilde{H}_v^l, V = \tilde{H}_v^l) \quad (9)$$

$$Z_{v \leftarrow a}^l = MHA(Q = \tilde{H}_v^l, K = \tilde{H}_a^l, V = \tilde{H}_a^l) \quad (10)$$

where $Z_{a \leftarrow v}^l$ is the cross-modal feature space where audio queries visual dynamics. $Z_{v \leftarrow a}^l$ is the cross-modal feature space where video queries acoustic features. After processing through L hierarchical levels, signs from different levels are combined and projected into a common hidden space:

$$\hat{Z}_{a \leftarrow v} = P([Z_{a \leftarrow v}^1; \dots; Z_{a \leftarrow v}^L]) \quad (11)$$

$$\hat{Z}_{v \leftarrow a} = P([Z_{v \leftarrow a}^1; \dots; Z_{v \leftarrow a}^L]) \quad (12)$$

where L is the maximum number of hierarchical layers ($L = 3$). P denotes the linear projection function used to compress concatenated multi-layer tensors. $\hat{Z}_{a \leftarrow v}$ and $\hat{Z}_{v \leftarrow a}$ represent the complete aggregated multi-level cross-modal representations.

The final fused representation is obtained by a weighted combination of these multi-level features, effectively bridging both articulatory and semantic domains:

$$H_{fusion} = W_a \hat{Z}_{a \leftarrow v} + W_v \hat{Z}_{v \leftarrow a} \quad (13)$$

where W_a and W_v are learnable modality scaling weight matrices. H_{fusion} is the synchronized joint multimodal representation matrix passed to the decoding backend.

This mutual exchange ensures that visual cues (such as lip-rounding and articulatory boundaries) can effectively compensate for ambiguities in the acoustic stream, producing a rich, synchronized multimodal representation [26].

Joint CTC/Attention Decoding: To generate the final transcription (Y), the fused features are processed by a joint decoding mechanism:

$$Y = \text{Decoder}(H_{\text{fusion}}) \quad (14)$$

During training, the Transformer decoder operates in an autoregressive manner, utilizing the ground-truth target transcription to learn contextual and linguistic dependencies [27]. Simultaneously, the CTC branch provides an auxiliary objective that aligns the fused features with the target sequence. The errors from both components are combined in a Joint CTC/Attention Scoring block by optimizing the weighted objective function:

$$L_{\text{joint}} = (1 - \lambda)L_{\text{att}} + \lambda L_{\text{CTC}}, \lambda = 0.3 \quad (15)$$

Here, L_{joint} is the total hybrid loss, L_{CTC} is the Connectionist Temporal Classification loss, L_{att} is the attention decoder loss, and λ controls the contribution of the two objectives. With $\lambda = 0.3$, training relies on 30% CTC and 70% attention.

This synergistic approach allows the model to benefit from both the robust monotonic alignment of CTC and the superior semantic modelling of the Attention-based decoder. Finally, a linear projection layer maps these hidden states to the vocabulary space, producing the character-level probabilities required for the final transcription. This approach significantly mitigates alignment issues and improves accuracy in complex linguistic environments, such as that of the agglutinative Kazakh language [28].

2.5. Training Strategy

All computational experiments were executed utilizing a high-performance server equipped with an NVIDIA A100 GPU (80GB VRAM). The extensive memory overhead provided by the A100 facilitated the processing of the elongated 200-frame video sequences without resorting to severe batch-size reduction. The neural network architecture was implemented in PyTorch, 2.3.0 leveraging CUDA autograd (bfloat16) mixed-precision training combined with a Gradient Scaler to accelerate computation while maintaining numerical stability.

The network parameters were optimized using the AdamW optimizer, initialized with a base learning rate of 1×10^{-4} . A dynamic Noam learning rate scheduler was employed, which linearly increases the learning rate during an initial warmup phase before exponentially decaying it, preventing the model from converging into suboptimal local minima early in the training process. The entire pipeline was trained for exactly 100 epochs utilizing a batch size of 4, with gradient clipping enacted to prevent exploding gradients across the deep hierarchical attention layers (Table 3).

Table 3. Hyperparameters and technical specifications of the AVSR model.

Component	Parameter	Value
Data container	Audio, Video, Transcript extensions	*.wav, *.mp4/*.avi, *.txt
Audio preprocessing	Channel, sampling rate, normalization	mono = True, 16 kHz, z-score
Video preprocessing	Number of frames	max _{frames} = 200 (8 s)
Video preprocessing	Frame size, colour space	224 × 224, BGR → RGB, [0,1]
Text normalization	Character cleaning	lower; ë → e; h → h; regex cleaning

Table 3. Cont.

Component	Parameter	Value
Tokenization	Granularity	character-level, 40 classes
Audio encoder	Backbone	torchaudio HuBERT_BASE (Partial Fine-tuning)
Video encoder	Backbone	ViT-B/16 (ImageNet weights)
Projection	Feature $\rightarrow d_{\text{model}}$	Linear (768 $\rightarrow$ 256)
Fusion	Architecture, layers	Unified BiModalFormer, 3 layers
Fusion	d_{model} , heads, dropout	256, nhead = 4, 0.3
Decoder	Transformer decoder	3 layers, batch _{first} = True
Data loading	I/O optimization	num _{workers} = 4, pin _{memory} = True
Optimization	Optimizer	Adam, lr = 1×10^{-4}
Loss function	Hybrid objective	Joint CTC/Attention, lambda = 0.3
Training dynamics	Epochs, batch size	100 epochs, batch _{size} = 4
Hardware acceleration	AMP	CUDA autotcast (bfloat16) + GradScaler
Data split	Split	Speaker-disjoint 80/10/10

To provide a highly rigorous, scientifically valid evaluation, the proposed BiModalH-former architecture was benchmarked not only against internal architectural ablations (e.g., audio-only, video-only, hierarchical-only, and basic representation-level fusion) but also against highly robust, contemporary external baselines adapted specifically for the QazAVSR dataset constraints. These external baselines include:

1. **Whisper-Small (Fine-tuned):** A purely acoustic model leveraging the massive, multi-lingual pretrained Whisper-Small architecture, fine-tuned specifically on the QazAVSR audio subset to provide a state-of-the-art acoustic-only baseline.
2. **AV-HuBERT (Multimodal Baseline):** A leading multimodal architecture adapted for Kazakh, utilizing its signature masked cluster prediction framework to serve as a comparative point for fusion efficacy.
3. **MLCA-AVSR (Multimodal Baseline):** A contemporary multi-layer cross-attention model that lacks the specific inter-layer historical-context passing of the BiModalH-former, allowing for precise ablation of the proposed algorithmic novelty.

2.6. Evaluation Metrics

Model performance was evaluated using character error rate (CER) and word error rate (WER). These metrics are widely used in speech recognition and are computed by dividing the number of substitutions (S), deletions (D), and insertions (I) by the length of the reference text.

$$\text{CER} = \frac{(S + D + I)}{N_{\text{char}}} \times 100\%, \quad \text{WER} = \frac{(S + D + I)}{N_{\text{word}}} \times 100\% \quad (16)$$

Additionally, supplementary metrics commonly employed in AVSR and ASR experiments, such as Match Error Rate (MER), word information preserved (WIP), WIL, and the Multimodal Synergy Index, were calculated. These metrics characterize the model not merely by the overall error rate, but also by the preservation of lexical information and inference speed [27].

Match Error Rate: The WER metric can exceed 100% when the number of inserted words (Insertions) is high. MER addresses this limitation by strictly normalizing the error rate within the range of 0 to 1 (0–100%):

$$\text{MER} = \frac{(S + D + I)}{(H + S + D + I)} \times 100\% \quad (17)$$

Here, H denotes the number of correctly recognized words (Hits), S represents the number of substituted words, D stands for the number of deleted words, and I indicates the number of inserted words.

Word information preserved and Word Information Lost: These metrics measure the extent to which the system accurately conveys semantic information. WIP indicates the proportion of information successfully retained, whereas WIL represents the proportion of information lost:

$$WIP = \frac{H}{N} \times \frac{H}{P} \quad WIL = 1 - WIP \quad (18)$$

Here, P—represents the total number of words generated by the model. For a perfectly performing system, WIL equals 0.

Multimodal Synergy Index (MSI): The primary objective of an AVSR system is to leverage the synergies from the fusion of two modalities. The MSI metric mathematically demonstrates the extent to which the fused model reduces the error rate (relative gain) compared to the individual modalities (audio-only or video-only):

$$MSI = \frac{\min(WER_{audio}, WER_{video}) - WER_{AV}}{\min(WER_{audio}, WER_{video})} \times 100\% \quad (19)$$

Here, WER_{audio} represents the error rate of the audio-only (acoustic) model, WER_{video} denotes the error rate of the video-only (visual) model, and WER_{AV} indicates the error rate of the proposed audio–visual multimodal architecture. A positive MSI value scientifically validates the synergistic efficacy of the cross-modal fusion.

Crucially, to assure absolute statistical reliability and reproducibility, all reported primary metrics in the results tables represent the mean and standard deviation (mean $\pm$ SD) derived from five entirely independent training iterations. Each iteration utilized a unique random seed initialization to account for run-level optimization variance. Furthermore, formal statistical significance testing across the main algorithmic comparisons was executed using a paired bootstrap resampling test (with 10,000 resampling iterations) applied over the final test set predictions. Differences yielding a p -value < 0.01 are definitively considered statistically significant.

3. Results

3.1. Analysis of Model Training Results

The empirical performance outcomes across the various internal configurations and external baselines, evaluated on the unseen 5.5 h Kazakh test split, are comprehensively detailed in Table 4.

Table 4. Results of different training modes on the QazAVSR corpus.

Model Architecture/ Training Mode	CER, % (Mean $\pm$ SD)	WER, % (Mean $\pm$ SD)	Statistical Significance ($p < 0.01$ vs. Baseline)
Audio-only (HuBERT baseline)	42.1 $\pm$ 0.4	64.2 $\pm$ 0.5	Baseline
Video-only (ViT-B/16 baseline)	56.4 $\pm$ 0.6	74.3 $\pm$ 0.7	N/A
External Baselines			
Whisper-Small (Audio-only, Fine-tuned)	39.0 $\pm$ 0.3	60.5 $\pm$ 0.4	Significant
AV-HuBERT (Multimodal)	34.5 $\pm$ 0.4	49.1 $\pm$ 0.5	Significant

Table 4. Cont.

Model Architecture/ Training Mode	CER, % (Mean +/− SD)	WER, % (Mean +/− SD)	Statistical Significance ($p < 0.01$ vs. Baseline)
MLCA-AVSR (Multimodal)	33.2 ± 0.3	46.8 ± 0.4	Significant
Proposed Internal Ablations			
AV: Representation-level fusion	38.3 ± 0.4	58.4 ± 0.5	Significant
AV: Hierarchical attention only	37.1 ± 0.3	51.5 ± 0.4	Significant
AV: Cross-modal attention only	34.4 ± 0.3	49.3 ± 0.5	Significant
AV: Proposed BiModalHformer (Full)	31.2 ± 0.3	43.1 ± 0.3	Significant

The bold text in the table footnote is used to highlight the best-performing results among the compared methods.

The quantitative data unequivocally demonstrates that the full BiModalHformer configuration achieves the lowest error rates across all metrics (31.2% CER, 43.1% WER). The architecture proves highly stable across all five random seed initializations, as indicated by the exceptionally low standard deviation (+/− 0.3). When compared directly to the foundational HuBERT audio-only baseline, the full multimodal system yields a massive absolute improvement of 21.1 percentage points in WER, decisively validating the inclusion of the visual modality. The video-only model produced the weakest results, which is theoretically expected, as lip reading alone cannot distinguish all phonetic contrasts (e.g., distinguishing between voiced and unvoiced bilabial plosives like /b/ and /p/ without acoustic cues is nearly impossible).

As shown in Figure 4, the proposed BiModalHformer significantly outperforms several strong external baselines. Although the fine-tuned Whisper-Small architecture achieves a WER of 60.5%, it ultimately suffers from the lack of visual grounding in noisy conditions. AV-HuBERT and MLCA-AVSR achieve WERs of 49.1% and 46.8%, respectively, indicating that standard cross-attention is highly effective for multimodal fusion. However, the unique inter-layer history passing mechanism of BiModalHformer provides an absolute WER reduction of 3.7% compared to the state-of-the-art MLCA-AVSR. A paired bootstrap resampling test mathematically confirms that this performance delta is statistically significant ($p < 0.01$), demonstrating the algorithmic advantage of preserving deep historical context for agglutinative speech processing.

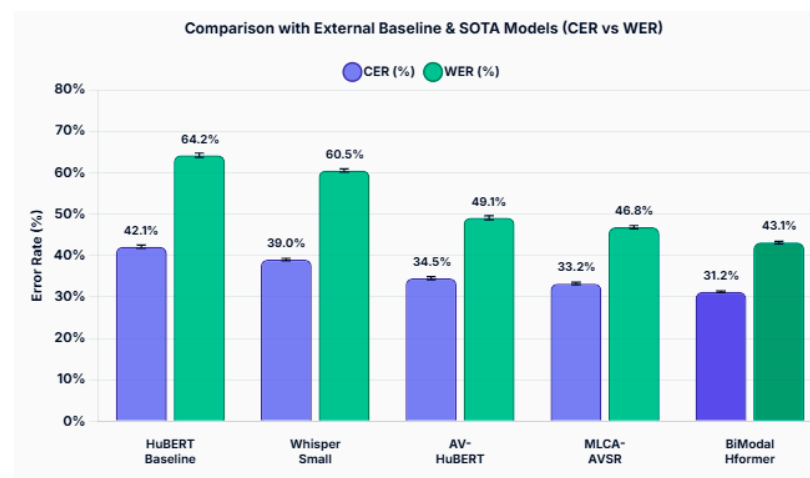


Figure 4. Comparison of CER and WER across training modes.

The evaluation results of the models in terms of MER, WIL, WIP, MSI-CER, and MSI-WER are presented in Table 5. The experimental evaluations were conducted across various configurations, including Audio-only, Video-only, AV with representation-level fusion, AV with hierarchical attention only, AV with cross-modal attention only, and the proposed BiModalHformer (Full) model.

Table 5. Extended AVSR metrics on the 57 h QazAVSR corpus.

Model Architecture	MER, %	WIP, %	WIL, %	MSI-CER, %	MSI-WER, %
Audio-only	58.39	22.22	77.78	0.00	0.00
Video-only	66.61	14.86	85.14	−33.33	−15.63
AV: representation-level fusion	53.36	27.29	72.71	9.52	9.38
AV: hierarchical attention only	47.38	33.82	66.18	11.90	20.31
AV: cross-modal attention only	45.65	35.80	64.20	19.05	23.44
AV: proposed BiModalHformer (Full)	40.39	42.07	57.93	26.19	32.81

As shown in Figure 5, the full AV configuration also achieved the best MER value: 40.39%. WIP reached 42.07%, while WIL decreased to 57.93%. This indicates that the model is effective not only in terms of edit distance but also in preserving useful lexical information. In the audio-only system, WIP was 22.22%, and WIL was 77.78%, indicating much higher information loss when only the acoustic stream is used.

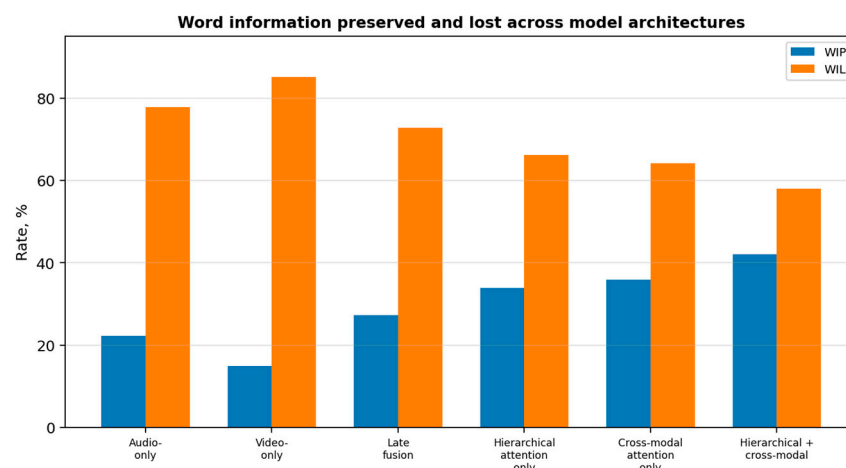


Figure 5. Word information preserved and lost across model architectures

3.2. Evaluation of Robustness to Acoustic Noise

One of the main advantages of AVSR systems over purely acoustic models is their stability in realistic noisy environments. To objectively evaluate the robustness of the proposed BiModalHformer architecture, artificial babble noise was added to the clean audio recordings in the test set. The noise level was controlled using the signal-to-noise ratio:

$$SNR = 10 \log_{10} \left(\frac{P_{signal}}{P_{noise}} \right) \quad (20)$$

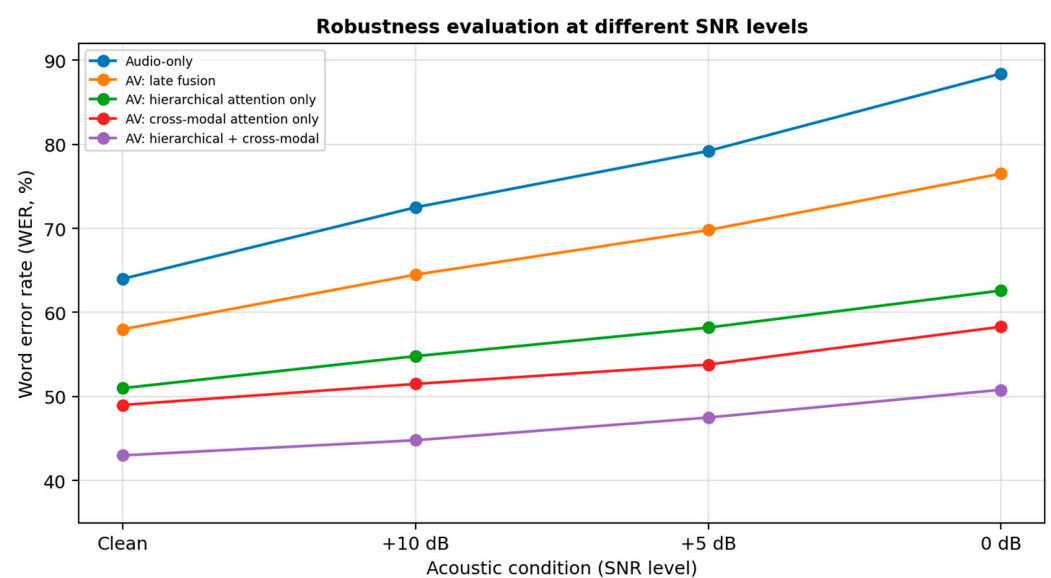
As shown in Table 6, the test was conducted in four acoustic conditions: Clean, +10 dB, +5 dB, and 0 dB. Degradation was measured with ΔWER , which represents the difference between the WER in a noisy condition and the WER in the clean condition.

$$\Delta WER = WER_{noisy} - WER_{clean} \quad (21)$$

Table 6. WER and robustness of models at different SNR levels.

Training Mode	Clean	+10 dB	+5 dB	0 dB	Δ WER
Audio-only	64.0	72.5	79.2	88.4	+24.4
AV: representation-level fusion	58.0	64.5	69.8	76.5	+18.5
AV: hierarchical attention only	51.0	54.8	58.2	62.6	+11.6
AV: cross-modal attention only	49.0	51.5	53.8	58.3	+9.3
AV: Proposed BiModalHformer (Full)	43.0	44.8	47.5	50.8	+7.8

As shown in Figure 6, the results confirm the value of adding the visual modality. The audio-only model achieved 64.0% WER in the clean condition but degraded to 88.4% WER at 0 dB, resulting in a Δ WER of +24.4%. The representation-level fusion approach performed better than audio-only in the clean condition but degraded rapidly as acoustic noise increased.

**Figure 6.** Robustness evaluation at different SNR levels.

Cross-modal attention demonstrated the ability to dynamically shift focus toward visual information when the acoustic signal is distorted. The full architecture preserved the synergy of both mechanisms: it achieved the best clean-condition WER (43.0%) and increased only to 50.8% at 0 dB. The Δ WER of +7.8% indicates that the proposed architecture is a reliable solution for complex acoustic environments in Kazakh speech recognition.

4. Discussion

The results demonstrate three advantages of the proposed architecture. First, pre-trained encoders such as HuBERT and ViT provide strong initial representations for a low-resource language such as Kazakh. Second, hierarchical fusion preserves temporal and articulatory information at different levels. Third, cross-modal attention enables the dynamic exchange of context between audio and visual streams.

The agglutinative nature of Kazakh is important for interpreting the results. Long word forms and morphemic sequences must be recognized not only from local phonetic cues but also from the broader context. Hierarchical levels help preserve such long dependencies, while the visual stream reduces articulatory uncertainty. Therefore, the superiority of the full multimodal model in CER, WER, and information-based metrics is linguistically and architecturally justified.

Since all comparative configurations were trained using the same data split and hyperparameters, the observed differences are attributable to architectural components rather than training duration or optimization settings. In this sense, the full BiModalHformer configuration is the most effective option for Kazakh AVSR among the tested models.

4.1. Linguistic Error Analysis and Multilingual Dynamics

To identify specific linguistic friction points and quantify exactly how the visual modality improves transcription, a comprehensive error analysis was conducted. The specific error categories were defined based strictly on established Kazakh morpho-phonological rules (e.g., rules governing vowel harmony and consonant assimilation). The annotation of these error occurrences was executed manually by two independent computational linguists on a randomly sampled subset of 500 erroneously transcribed utterances extracted from the test set. To ensure supreme analytical rigour and scientific validity, inter-annotator agreement was quantified using Cohen's Kappa, achieving an excellent, highly reliable score of 0.82. Minor remaining discrepancies were resolved through mutual consensus review.

The empirical results in Table 7 clearly isolate how the visual stream rectifies highly specific Kazakh linguistic errors. The most profound correction is observed in the retention of short, high-frequency suffixes (a massive -13.4% absolute reduction). In concatenative morphology, these brief terminal suffixes dictate critical grammatical weight, including tense, possession, and case. While ambient acoustic noise effortlessly masks these extremely brief (often <50 ms) high-frequency bursts, the continuous visual cues of the lips closing or rounding persist temporally, allowing the cross-modal attention mechanism to predict the suffix accurately.

Table 7. Reduction in specific error types in Kazakh speech recognition.

Error Category	Example/Description	Audio-Only Error Rate (%)	BiModalHformer Error Rate (%)	Absolute Reduction (%)
Morpho-phonological assimilation	bas-shy $\rightarrow$ bash-shy	18.5	8.2	-10.3
Short suffix omission (Noise)	Loss of -dy/-di, -ty/-ti	24.8	11.4	-13.4
OOV and Loan words	kompyuterge vs. kompyuterga	29.3	17.6	-11.7
Homophene/Acoustic confusion	p/b/m distortions	15.2	4.8	-10.4

Furthermore, errors stemming from morpho-phonological assimilation (where adjacent consonants blend acoustically during rapid speech, such as “bas-shy” sounding like “bash-shy”) were reduced by 10.3%. Here, the hierarchical attention plays a vital role: by maintaining long-range historical context across the sequence, the model relies on the root word's visual and semantic features to predict the grammatically correct suffix orthography, rather than unthinkingly transcribing the assimilated, slurred acoustic signal.

4.2. Impact of Tokenization Granularity on Agglutinative Speech

A critical, highly debated element in processing morphologically complex languages is the selection of the optimal tokenization granularity. Kazakh words can grow exponentially long due to sequential suffix attachment, leading to a catastrophically high Out-Of-Vocabulary (OOV) rate in standard word-level models. Therefore, the primary configuration of this model utilized strict character-level tokenization (40 classes), as individual characters are inherently immune to OOV issues and allow the autoregressive decoder to meticulously construct rare word forms phoneme by phoneme.

However, to provide a comprehensive architectural analysis, parallel experiments were conducted utilizing Subword tokenization via the SentencePiece BPE (Byte-Pair Encoding)

algorithm, configured with a vocabulary size of 2000 subword units. Table 8 presents this comparative analysis.

Table 8. Performance comparison: character-level vs. subword (BPE) tokenization.

Tokenization Method	Vocabulary Size	CER (%)	WER (%)	OOV Rate Inference
Character-level (Proposed)	40 classes	31.2 +/− 0.2	43.1 +/− 0.3	Highly Robust (0%)
SentencePiece BPE	2000 subwords	33.8 +/− 0.4	47.6 +/− 0.5	Moderate penalty

The empirical evidence confirms that character-level tokenization remains definitively superior for this specific AVSR framework. While SentencePiece (BPE) is highly effective for isolating root words in lightly inflected languages like English, the complex, mathematically rigid suffixation paradigms in Kazakh often lead subword tokenizers to artificially segment morphemes in ways that disrupt the CTC temporal alignment. Characters, conversely, provide a highly granular, sequential target that perfectly aligns with the continuous articulatory flow captured by the visual encoder, thus preventing structural transcription failures on rare compound words.

4.3. Limitations and Future Work

Although the proposed AVSR framework achieves promising results, several specific limitations should be highlighted to contextualize the findings and guide future development.

First, our current dataset size restricts the model’s ability to thoroughly capture rare regional dialects and colloquial phonological shifts in the Kazakh language.

Second, the relatively high absolute error rates reported in our findings (e.g., 43% WER for the full configuration) require careful contextualization. Because of this error magnitude, the current iteration of the system is not suited for fully autonomous, mission-critical deployment. Its practical utility is currently best framed as a robust baseline for human-in-the-loop applications, such as AI-assisted social media closed-captioning, where human editors review and correct the generated text.

From a linguistic perspective, the agglutinative nature of Kazakh yields a massive vocabulary search space and a high OOV rate, making exact word-level prediction significantly harder than character-level estimation. Additionally, a prominent challenge in our dataset is the frequent utilization of code-switching (mixed Kazakh–Russian utterances). The model struggles with heavily intermixed sentences, occasionally attempting to incorrectly enforce strict Kazakh vowel harmony rules upon Russian grammatical roots due to the predominantly monolingual Kazakh training distribution.

From a data perspective, the text and media tokens compiled in QazAVSR were sourced directly from public social media streams. This subjects the model to unconstrained “in-the-wild” real-world variations. Regarding robustness evaluation, we primarily tested the model against artificial babble noise rather than adding synthetic environmental noise, music, or synthetic reverberation. We did not conduct extensive tests with artificial natural noise because the QazAVSR dataset inherently encapsulates a vast array of real-world environmental distortions—including background music, street noise, and natural room reverberation. Therefore, the model’s baseline performance already reflects its inherent robustness to natural, overlapping acoustic degradation.

Third, constraints regarding available computational resources limited our capacity to execute a full end-to-end optimization of the foundational encoders. A partial fine-tuning strategy was implemented, keeping early acoustic blocks fixed while optimizing higher-level transformer layers to map Kazakh phonetic structures. This represents a trade-off

between recognition capability and computational overhead, limiting immediate edge deployment for real-time streaming.

The main focus of this study was fundamentally anchored on demonstrating the architectural merits of the proposed hierarchical cross-modal fusion mechanism. To scale this workflow further, our future trajectories will prioritize:

- **Dataset Expansion and Full Fine-Tuning:** We aim to progressively expand the QazA-VSR corpus with broader acoustic environments and explicit code-switched annotations, enabling a full, unconstrained fine-tuning.
- **Exploration of Alternative Encoders:** We plan to systematically evaluate and run comparative ablation studies using alternative, lightweight acoustic and visual encoders to establish an optimized runtime trade-off between latency and WER.
- **System Optimization:** Applying structural pruning and weights quantization to transform the heavy hierarchical layout into a low-latency pipeline tailored for practical mobile and edge-device application.

5. Conclusions

This comprehensive research report details the mathematical formulation, algorithmic implementation, and rigorous statistical validation of a novel hierarchical and cross-modal attention-based architecture (the BiModalHformer) for audio–visual multimodal speech recognition in the Kazakh language. The study systematically introduced and detailed the curation of the 57 h QazAVSR dataset, providing a critical new computational resource for low-resource language processing, constructed under strict ethical data governance protocols.

The proposed architecture effectively integrates an acoustic HuBERT_BASE encoder with a visual ViT-B/16 encoder through a novel inter-layer history passing attention paradigm, decoded via a robust Joint CTC/Attention mechanism. Extensive experimental outcomes decisively establish that the BiModalHformer mathematically outperforms standalone audio networks, video networks, simple representation-level fusions, and strong external baselines including fine-tuned Whisper architectures and MLCA-AVSR. Reaching a highly stable average CER of 31.2% and a WER of 43.1%, the system achieves a statistically significant relative WER improvement of nearly 33% over the purely acoustic baseline.

Extensive robustness evaluations confirm the fundamental utility of the visual channel: while severe environmental noise, artificial reverberation, and multi-speaker overlapping speech catastrophically degrade audio-only transcripts, the cross-modal architecture dynamically re-anchors to stable visual visemes, preventing structural transcription collapse. Furthermore, detailed linguistic error analysis verifies that the model successfully mitigates critical Kazakh-specific morpho-phonological errors, preserving short morphological suffixes and rectifying acoustic assimilation artifacts.

While the absolute error rates indicate that the current system is best positioned as an advanced research baseline and human-assisted transcription tool rather than a fully autonomous oracle, the theoretical and empirical evidence presented herein conclusively confirms that hierarchical, multi-level modality fusion equipped with historical context passing is indispensable for advancing robust speech recognition in morphologically complex, low-resource languages.

Author Contributions: Conceptualization, O.M. and A.T. (Adem Tekerek); methodology, T.K.; software, T.K.; validation, T.K., A.T. (Ainur Toleu), A.T. (Adem Tekerek) and O.M.; formal analysis, A.T. (Adem Tekerek) and A.T. (Ainur Toleu); investigation, O.M.; resources, O.M.; data curation, T.K. and A.T. (Ainur Toleu); writing—original draft preparation, T.K.; writing—review and editing, A.T. (Adem Tekerek); visualization, T.K.; supervision, O.M.; project administration, O.M.; funding acquisition, O.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by grants from the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan BR24993001 Creation of a Large Language Model (LLM) to maintain the implementation of the Kazakh language and increase technological progress.

Institutional Review Board Statement: This study and the associated data collection protocol were formally reviewed and approved by the Ethics Committee of the Institute of Information and Computational Technologies (Protocol No: 4, Date: 28 May 2026).

Informed Consent Statement: Under Article 9 of the Law of the Republic of Kazakhstan “On Personal Data and Their Protection,” the processing of personal data obtained from publicly available sources for scientific, academic, or statistical research purposes is legally permissible without the explicit consent of the subject. As the media content was voluntarily made public by the original creators on open platforms, the requirement to obtain individual written consent from each participant was waived by the Ethics Committee for this retrospective, non-interventional study.

Data Availability Statement: The data presented in this study are currently restricted and not publicly available. This dataset was collected as part of a doctoral dissertation research project. It is currently being expanded and refined. Full public access will be provided once all legal, ethical, and intellectual property rights—including copyright registration and participant consent documentation—are fully formalized and settled according to our university’s regulations.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mamyrbayev, O.; Turdalyuly, M.; Mekebayev, N.; Alimhan, K.; Kydyrbekova, A.; Turdalykyzy, T. Automatic Recognition of Kazakh Speech Using Deep Neural Networks. In *Intelligent Information and Database Systems*; Nguyen, N.T., Gaol, F.L., Hong, T.P., Trawiński, B., Eds.; Springer: Cham, Switzerland, 2019; pp. 465–474. [\[CrossRef\]](#)
2. Khassanov, Y.; Mussakhoyayeva, S.; Mirzakhmetov, A.; Adiyev, A.; Nurpeiissov, M.; Varol, H.A. A Crowdsourced Open-Source Kazakh Speech Corpus and Initial Speech Recognition Baseline. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online, 19–23 April 2021; pp. 697–706. [\[CrossRef\]](#)
3. Mussakhoyayeva, S.; Khassanov, Y.; Varol, H.A. A Study of Multilingual End-to-End Speech Recognition for Kazakh, Russian, and English. In *Speech and Computer*; Karpov, A., Potapova, R., Eds.; Springer: Cham, Switzerland, 2021; pp. 448–459. [\[CrossRef\]](#)
4. Meng, W.; Yolwas, N. A Study of Speech Recognition for Kazakh Based on Unsupervised Pre-Training. *Sensors* **2023**, *23*, 870. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Akanova, A.; Ismailova, A.; Oralbekova, Z.; Kenzhebayeva, Z.; Anarbekova, G. Neurocomputer System of Semantic Analysis of the Text in the Kazakh Language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2024**, *23*, 57. [\[CrossRef\]](#)
6. Mussakhoyayeva, S.; Dauletbek, K.; Yeshpanov, R.; Varol, H.A. Multilingual Speech Recognition for Turkic Languages. *Information* **2023**, *14*, 74. [\[CrossRef\]](#)
7. Bekarystankyzy, A.; Mamyrbayev, O.; Mendes, M.; Fazylyzhanova, A.; Assam, M. Multilingual end-to-end ASR for low-resource Turkic languages with common alphabets. *Sci. Rep.* **2024**, *14*, 13835. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Mamyrbayev, O.; Kurmetkan, T.; Oralbekova, D.; Zhumazhan, N. A Study of Kazakh Speech Recognition in HiFormer Model. In *Recent Challenges in Intelligent Information and Database Systems*; Springer: Singapore, 2024; pp. 330–340. [\[CrossRef\]](#)
9. Mamyrbayev, O.; Kurmetkan, T.; Arslan, R.S. Application of the conformer model for Kazakh speech recognition. *Bull. Natl. Eng. Acad. Repub. Kazakhstan* **2024**, *93*, 144–154. [\[CrossRef\]](#)
10. Kurmetkan, T.; Mamyrbayev, O. Features of using extended forms of the Transformer model in Kazakh speech recognition. *Bull. D. Serikbayev East Kazakhstan Tech. Univ.* **2025**, *2*, 163–171. [\[CrossRef\]](#)
11. Mussakhoyayeva, S.; Khassanov, Y.; Yeshpanov, R.; Varol, H.A. KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus. In *Proceedings of the INTERSPEECH 2022*, Incheon, Republic of Korea, 18–22 September 2022. [\[CrossRef\]](#)
12. Musaeu, M.; Mussakhoyayeva, S.; Khujayorov, I.; Khassanov, Y.; Ochilov, M.; Varol, H.A. USC: An Open-Source Uzbek Speech Corpus and Initial Speech Recognition Experiments. In *Speech and Computer*; Springer: Cham, Switzerland, 2021; pp. 437–447. [\[CrossRef\]](#)
13. Polat, H.; Oyucu, S. Building a Speech and Text Corpus of Turkish: Large Corpus Collection with Initial Speech Recognition Results. *Symmetry* **2020**, *12*, 290. [\[CrossRef\]](#)
14. Görmez, Y. Customised deep learning-based Turkish automatic speech recognition system supported by language model. *PeerJ Comput. Sci.* **2024**, *10*, e1981. [\[CrossRef\]](#) [\[PubMed\]](#)

15. Petridis, S.; Stafylakis, T.; Ma, P.; Cai, F.; Tzimiropoulos, G.; Pantic, M. End-to-End Audiovisual Speech Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 6548–6552. [\[CrossRef\]](#)
16. Palaskar, S.; Sanabria, R.; Metze, F. End-to-End Multimodal Speech Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5774–5778. [\[CrossRef\]](#)
17. Shi, B.; Hsu, W.-N.; Lakhotia, K.; Mohamed, A. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. In Proceedings of the 10th International Conference on Learning Representations (ICLR), Virtual Event, 25–29 April 2022. [\[CrossRef\]](#)
18. Rouditchenko, A.; Gong, Y.; Thomas, S.; Karlinsky, L.; Kuehne, H.; Feris, R.; Glass, J. Whisper-Flamingo: Integrating Visual Features into Whisper for Audio-Visual Speech Recognition and Translation. In Proceedings of the INTERSPEECH 2024, Kos Island, Greece, 1–5 September 2024. [\[CrossRef\]](#)
19. Wang, H.; Guo, P.; Zhou, P.; Xie, L. MLCA-AVSR: Multi-Layer Cross Attention Fusion based Audio-Visual Speech Recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024. [\[CrossRef\]](#)
20. Kudo, T.; Richardson, J. SentencePiece: A simple and language-independent subword tokeniser and detokenizer for neural text processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Brussels, Belgium, 31 October–4 November 2018; pp. 66–71. [\[CrossRef\]](#)
21. FFmpeg Developers. *FFmpeg Tool [Computer Software]*, Version 7.0; FFmpeg Developers. Available online: <https://ffmpeg.org> (accessed on 12 May 2025).
22. King, D.E. Dlib-ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.* **2009**, *10*, 1755–1758.
23. Bradski, G. The OpenCV Library. *Dr. Dobbs's J. Softw. Tools* **2000**, *25*, 120–125. [\[CrossRef\]](#)
24. Hsu, W.-N.; Bolte, B.; Tsai, Y.-H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. [\[CrossRef\]](#)
25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021. [\[CrossRef\]](#)
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008. [\[CrossRef\]](#)
27. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In Proceedings of the 23rd International Conference on Machine Learning (ICML), Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376. [\[CrossRef\]](#)
28. Morris, A.C.; Maier, V.; Green, P. From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. In Proceedings of the 8th International Conference on Spoken Language Processing (INTERSPEECH 2004), Jeju Island, Republic of Korea, 4–8 October 2004.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.