DOI: 10.37943/24VEWA2615

Aru Ukenova
Manager of the Online Learning Department, Researcher, Faculty of Information Technologies
ukenovaaru07@gmail.com, orcid.org/0000-0002-2797-672X
L.N. Gumilyov Eurasian National University, Kazakhstan

Gulmira Bekmanova
PhD, Professor, Faculty of Information Technologies
gulmirara@gmail.com, orcid.org/0000-0001-8554-7627
L.N. Gumilyov Eurasian National University, Kazakhstan

Banu Yergesh
PhD, Deputy Director, Department of Digital Development
b.yergesh@gmail.com, orcid.org/0000-0002-8967-2625
L.N. Gumilyov Eurasian National University, Kazakhstan

Mamyr Altaibek
Master of Information Security, Programmer, Faculty of Information Technologies
mameralt@outlook.com, orcid.org/0009-0002-8219-0751
L.N. Gumilyov Eurasian National University, Kazakhstan

INTERFACE DESIGN OF AN INTELLIGENT INTERACTIVE LEARNING SYSTEM

Abstract: This study presents the design, implementation, and evaluation of an Intelligent Interactive Learning System that employs multimodal interaction to improve the adaptability, accessibility, and engagement of digital education. Conventional e-learning platforms typically rely on static and text-based resources, which restrict personalization and reduce learner motivation. The proposed system integrates natural language processing, speech synthesis, and avatar-based interfaces to deliver lectures through synchronized speech, gestures, and facial expressions. The system automatically processes uploaded lecture scripts and slide presentations, segmenting and aligning them to generate interactive video lectures. A novel contribution of this work is the incorporation of customized Kazakh-language support, implemented through intonation modeling, dependency parsing, and gesture mapping to enhance inclusivity for underrepresented linguistic communities. The system performance was evaluated using Facebook's variational inference text-to-speech model. Experimental results demonstrate real-time capability, with an average latency of 25.5 ms, throughput exceeding 4,200 characters per second, and low computational resource requirements. These findings confirm the suitability of the system for deployment in resource-constrained environments without compromising speech quality or responsiveness. Compared with conventional tutoring and static e-learning platforms, the system additionally provides automated assessment generation, multimodal feedback, and accessibility functions such as subtitles and adjustable playback controls. The study contributes a scalable model for intelligent, avatar-based learning that integrates speech synthesis, real-time interaction, and cultural-linguistic inclusivity. Future work will focus on extending personalization through adaptive learner modeling, incorporating affective computing for emotion-sensitive interaction, and enabling interoperability with established learning management systems.

Received: 16.06.2025 Accepted: 20.10.2025 Published: 30.10.2025

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

Keywords: multimodal interfaces; speech synthesis; avatar-based system; gestures; face mimics; Kazakh language.

Introduction

The swift advancement of digital technology has revolutionized the educational landscape, fostering the creation of e-learning systems that emphasize accessibility, flexibility, and personalization. Contemporary methodologies increasingly utilize artificial intelligence, natural language processing, and multimodal interaction to facilitate adaptive and interactive learning experiences.

Nonetheless, notwithstanding considerable advancements, contemporary intelligent learning systems encounter substantial constraints. Numerous solutions prioritize automation and scalability yet lack a robust pedagogical foundation. Recommendation algorithms and adaptive models demonstrate potential for enhancing student engagement; yet they often neglect inclusion for minority languages and cultural situations [1]. Likewise, although avatar-based interfaces can augment engagement via speech, gesture, and visual signals, the majority of commercial solutions are primarily focused on corporate or entertainment applications rather than educational ones [2]. Speech systems crucial for real-time engagement have difficulties in achieving a balance between naturalness, responsiveness, and efficiency, especially in resource-limited settings.

This study provides a triple scientific contribution. Initially, it introduces a unified multimodal synchronization framework that integrates linguistic parsing, prosodic modeling, and gesture timing into a single optimization-driven pipeline specifically adapted for the Kazakh language. This approach closes a gap in existing intelligent learning systems by enabling precise cross-modal coordination between speech and nonverbal behavior in underrepresented linguistic contexts. Second, the study formulates a mathematical model that defines the relationship between syntactic salience, clause type, and prosodic modulation, ensuring natural timing, expressiveness, and real-time stability of avatar-driven instruction. Third, it provides empirical validation of this model within the proposed intelligent interactive learning system (IILS), demonstrating low-latency, resource-efficient performance on commodity hardware while maintaining high intelligibility and synchronization accuracy. Collectively, these contributions establish the first scientifically grounded framework for culturally and linguistically adaptive avatar-based education in the Kazakh language.

Literature review

Recent studies illustrate various methodologies for the design and implementation of intelligent learning systems that incorporate machine learning, logical inference, and semantic modeling. Research, including [3] and [4], examines the development of personalized training models and recommendation systems aimed at enhancing learner engagement and educational outcomes. For example, [3] discusses frameworks for blended and lifelong learning, whereas [4] focuses on adaptive recommendation techniques that improve online education experiences.

Another research [5], [6] offers systematic reviews of AI-driven tutoring environments and adaptive educational systems. This research examines the role of machine learning algorithms and intelligent inference mechanisms in facilitating personalized instruction and promoting sustainable education. Furthermore, a notable area of research [7], [8] investigates the use of cloud computing platforms for scalable e-learning solutions. The studies illustrate that distributed architecture improves accessibility, performance, and collaboration within intelligent educational environments.

Innovations in [9] and [10] investigate interactive AI-driven teaching systems, with a focus on language and engineering education. The works highlight semantic modeling, simulation, and real-time feedback for learners to enhance educational interactions effectively and engagingly. The system presented in [11] utilizes intelligent query optimization and recommendation algorithms to improve course selection and navigation in online lectures. These approaches illustrate the combination of machine learning with semantic search and personalization methods. References [12] and [13] offer critical conceptual and technological insights into AI, emphasizing its implications, fundamental technologies, and its transformative impact on e-learning and data-driven education.

Analyses such as [14] and [15] examine the technical, ethical, and pedagogical challenges associated with the implementation of AI-enabled personalized learning environments, providing strategies to address these limitations. Empirical studies [16], [17] investigate the effects of AI, including recommendation algorithms and conversational agents, on learner motivation, engagement, and academic performance. These studies illustrate the substantial impact of adaptive systems on student outcomes.

In addition, studies like [18] and [19] offer multidisciplinary perspectives on the opportunities and implications of AI for educational policy, organizational strategy, and research agendas, establishing a broader context for intelligent learning technologies. The emergence of generative AI is discussed in sources [20], [21], [22], and [23], which examine AI-driven content generation, ethical considerations, and the educational potential of large language models like ChatGPT. This research evaluates the potential of generative systems to improve teaching, learning, and user engagement.

Systematic reviews such as [24] examine chatbot-based educational interfaces, assessing their design, interaction models, and efficacy in enhancing student learning and communication. Hybrid recommendation methods [25] that integrate machine learning with clustering and semantic reasoning are presented to enhance content delivery and learner adaptation in e-learning platforms.

Studies [26], [27], [28] examine the integration of IoT within e-learning ecosystems, emphasizing the contributions of sensor networks, edge devices, and fog computing to real-time data processing, adaptive content delivery, and improved learner experiences. Moreover, this research [29] investigates the relationship between intelligent systems and cognitive skill development, highlighting the role of data-driven instructional design in enhancing problem-solving and critical thinking abilities.

Research in [30] and [31] examines learning management platforms, technology acceptance frameworks, and infrastructure systems that support the implementation of intelligent and IoT-enabled educational solutions. These systems are capable of analyzing student behavior, predicting their needs and offering personalized recommendations in real time [32]. According to research [33], the intelligent learning system is one of the most promising areas in the field of digital pedagogy.

Much attention is paid to the development of multimodal interfaces that provide interaction with educational content through various channels of perception - visual, auditory and kinesthetic [34], [35]. This is especially important in the context of inclusive education and when teaching skills that require practical mastery. For example, the introduction of avatars with speech and gestural capabilities improves cognitive perception of the material and promotes the formation of an emotional connection with the system [36], [37], [38].

Notwithstanding this advancement, numerous constraints remain in the architecture of intelligent learning systems. Numerous current methodologies prioritize algorithmic optimization, scalability, or infrastructural efficiency, although they exhibit a deficiency in robust educational foundations and inclusivity for marginalized language populations. Present adaptive

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

and recommendation-based approaches are predominantly designed for globally prevalent languages, resulting in considerable deficiencies in accessibility for learners in minority language environments.

Speech synthesis technologies, crucial for natural and responsive interaction, continue to encounter difficulties in attaining low latency, high throughput, and effective deployment in resource-limited settings. These deficiencies limit the scalability and efficacy of intelligent tutoring systems across many educational contexts. Consequently, there is a distinct necessity for an intelligent interactive learning system that incorporates real-time speech synthesis, natural language processing, and multimodal avatar-based interaction, while expressly promoting cultural and linguistic diversity. Rectifying these deficiencies will promote the evolution of flexible, scalable, and pedagogically sound learning environments.

The aim and objectives of the study

This study aims to design, implement, and evaluate an intelligent interactive learning system that incorporates real-time speech synthesis, natural language processing, and avatar-based multimodal interaction to facilitate adaptive, personalized, and inclusive digital education, focusing specifically on underrepresented linguistic contexts like Kazakh.

To accomplish this goal, the study seeks to fulfill the following objectives:

1. To develop a modular system architecture that integrates natural language processing, text-to-speech synthesis, and avatar-based interaction for real-time educational applications.
2. To achieve cultural and linguistic inclusivity by developing and integrating support for the Kazakh language into the IILS, providing phonetic precision and culturally appropriate gestures.
3. To do a comprehensive assessment of the speech synthesis model regarding latency, real-time factor, throughput, and resource use, to ascertain its appropriateness for educational contexts, including resource-limited scenarios.
4. To evaluate the proposed system against current avatar-based platforms, emphasizing differences in instructional focus, adaptability, scalability, and cultural-linguistic congruence.
5. To evaluate the educational potential of the IILS by examining its ability to provide adaptive, interactive, and learner-centered experiences through multimodal feedback and tailored replies.

Materials and methods

Architecture of the intelligent system

The architecture of the IILS is designed to provide automated, adaptive, and engaging educational experiences. It features a modular and layered structure that allows for the seamless integration of multimedia content delivery, real-time user interaction, natural language processing, and speech synthesis [37]. Each component plays a distinct role in delivering lecture content, analyzing user input, generating feedback, and providing personalized responses. Architecture supports scalability, maintainability, and flexibility, enabling customization for various educational domains and learning contexts. The system integrates user interface components, content management tools, and a speech synthesis engine, all working in concert to create a smooth and immersive learning environment without relying on a formal knowledge base or ontological model.

The system employs an intelligent learning framework designed to deliver fully automated lectures through a virtual avatar. It supports both textual and visual materials, enabling interactive learning by engaging students with questions and responding to their inquiries. The system architecture (Fig.1) consists of several interconnected modules that work together to provide a personalized and immersive educational experience.

DOI: 10.37943/24VEWA2615

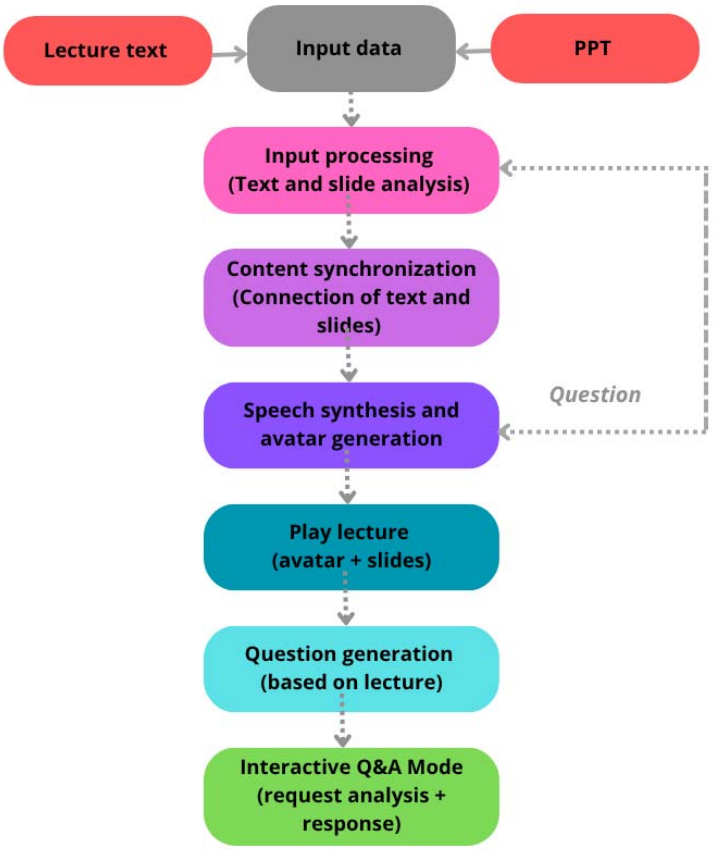

Figure 1. Architecture of the IILS

Educators are able to upload lecture scripts along with slide presentations (in a pdf format). The system processes the input by segmenting the lecture into coherent sections and aligning each with the corresponding slide content to ensure a seamless presentation. A customizable virtual avatar acts as the instructor, delivering the lecture using speech synthesis and animated gestures while displaying synchronized slides. The avatar's voice, pace, and visual appearance can be tailored to suit various educational contexts or user preferences.

Upon completion of each topic, the system automatically generates formative assessment items—including multiple choice, open-ended, and matching questions—using natural language processing techniques. To support real-time student interaction, particularly during the question-answering phase, the system leverages natural language processing (NLP) capabilities to interpret student queries submitted via voice or text. It analyzes the question content, matches it with relevant lecture material, and formulates an appropriate response. The answer is then delivered both visually and audibly via the avatar.

By combining semantic reasoning with multimodal interaction, the system replicates the dynamics of a real classroom environment, offering individualized feedback and enhancing learner engagement through adaptive and interactive learning support.

Multimodal Interfaces

A central component of the IILS is its multimodal interface, which enables learners to engage with educational content through multiple perceptual channels. Multimodality combines auditory, visual, and textual modes of information delivery, thereby enhancing cognitive processing, supporting diverse learning preferences, and improving accessibility for students with sensory impairments. Prior research highlights that multimodal systems increase learner engagement and foster deeper comprehension by simulating real-world communication pat-

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

terns [38]. This multimodal presentation aligns with the cognitive theory of multimedia learning (CTML), which posits that processing information via dual channels (verbal and visual) enhances understanding while reducing cognitive load in accordance with cognitive load theory (CLT) [39].

Extending CTML into AI-enhanced learning contexts, recent research proposes integrating intelligent adaptation into multimedia frameworks to better manage cognitive load and personalize learning experiences. In the IILS, auditory interaction is realized through a high-performance TTS engine, which generates natural and responsive voice output. This is complemented by visual components, including a virtual avatar capable of synchronized lip movements, facial expressions, and gestures, as well as dynamically rendered slide presentations [40]. Textual support is provided through subtitles, on-screen questions, and answer prompts, ensuring clarity and reinforcing spoken content. Together, these modalities replicate the richness of face-to-face classroom instruction [41].

The multimodal design is further enhanced by adaptive synchronization mechanisms that align speech with avatar animations, gestures, and visual cues. This coordination improves the perception of temporal coherence and facilitates the establishment of an emotional connection between the learner and the system. Moreover, multimodality supports inclusivity by allowing users to customize their mode of interaction: for example, by enabling subtitles for hearing-impaired learners or adjusting gesture intensity for reduced visual load [42]. By integrating auditory, visual, and textual modalities into a cohesive interface, the IILS provides an immersive and flexible learning experience [1]. This design not only increases engagement but also contributes to the personalization and accessibility of digital education, addressing the needs of heterogeneous learner populations.

Integration of components into a unified intelligent system

1. Linguistic analysis of the Kazakh language

Integration of individual modules—linguistic analysis of the Kazakh language, speech synthesis, lip synchronization, and animation control—into the architecture of a unified intelligent educational system is implemented. Each component performs specific tasks, yet their integration ensures complete and natural interaction between the user and the avatar.

This work extends the previous research [2] by integrating sentence-level gesture mapping and prosodic modeling into a unified multimodal synchronization framework specifically adapted for the Kazakh language. Unlike prior implementations that focused on separate gesture–intonation alignment, the present system achieves cross-modal coordination through dependency-based linguistic parsing, sentiment-weighted prosody control, and adaptive pause modeling, constituting a novel methodological contribution to Kazakh-language intelligent tutoring systems.

The system input consists of text provided by the instructor, which is processed by the linguistic analysis module for the Kazakh language. This stage forms a structural representation of the text, required for subsequent modules—speech synthesis and animation. The model is based on the dependency parsing approach [43], [44], specialized for Kazakh language processing [45], and capable of handling its characteristic syntactic and grammatical structures.

During the development of the intelligent model for Kazakh language processing, a dedicated function (*analyze_kazakh_sentence)* was implemented for automated analysis of user input. This function applies to the *Stanza* library and includes stages such as tokenization, morphological analysis, syntactic parsing, and classification of sentence constituents [46]. The processing is based on a dependency model that effectively interprets Kazakh's grammatical and syntactic features, including its free word order and rich morphology.

Each sentence component undergoes morphological analysis, yielding information such as case, number, person, aspect, tense, and other features. Based on syntactic relations (*deprel*),

DOI: 10.37943/24VEWA2615

words are classified according to their functional role in the sentence—subject (*sub*), predicate (*pre*), object (*obj*), adverbial modifier (*advmod*), attribute (*att*), and others. This classification plays a crucial role in constructing an accurate text understanding model, as it not only defines the sentence structure but also allows the educational system to adapt its behavior, such as providing feedback and visualization through an animated avatar.

The use of this function ensures a high level of contextual awareness and forms an integral part of the system's interactive analysis of user input in Kazakh. The model analyzes each sentence (Fig. 2), highlighting its structural components (sentence constituents). Such classification helps the model understand the functional organization of a sentence, which is essential for syntactic precision and accurate contextual interpretation.

In Kazakh sentence structure, each utterance typically consists of several main components. The *subject* usually denotes the actor or entity performing the action — for example, *"Мен"* ("I") in the sentence *"Мен кітапты оқимын."* ("I read a book"). The *predicate* expresses the main action or state — in this example, *"оқимын"* ("read"). The *object* is usually expressed by a noun or pronoun group, such as *"кітап"* ("book"). Adverbial and attributive words add extra information or characteristics, refining the meaning and context — for instance, in *"Мен кітапты бүгін оқимын."* ("I read the book today"), *the adverbial modifier "бүгін"* ("today") specifies the time of the action.

Such a structure — comprising subject, predicate, object, attribute, and adverbial modifier— is fundamental for understanding the logic and meaning of Kazakh sentences. Recognizing these elements and their interactions allows the model to perform deeper text analysis and provide accurate, context-sensitive feedback to learners.

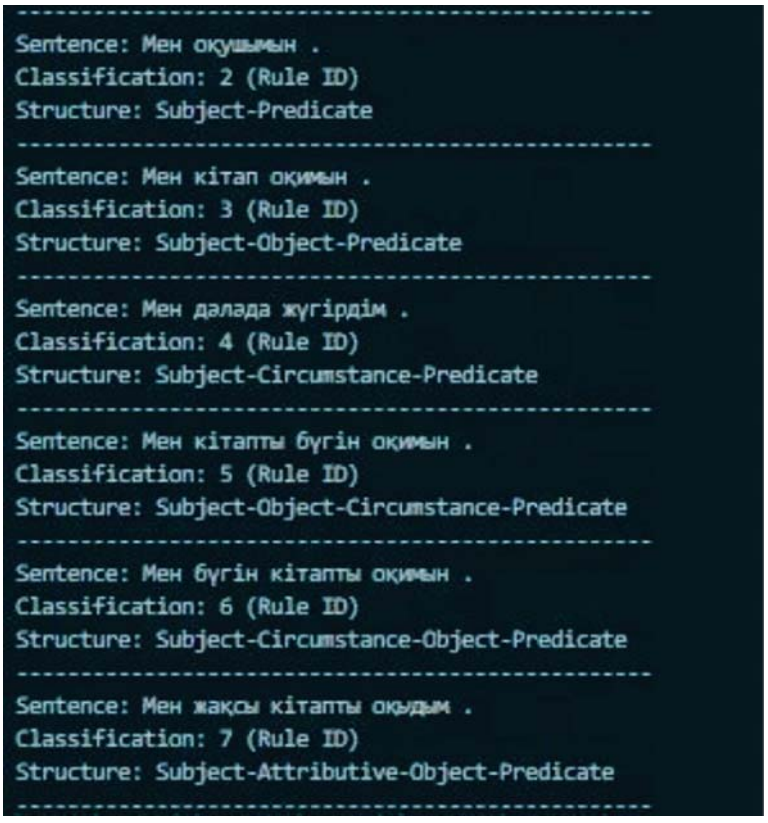

Figure 2. Classification of sentence structures in the Stanza library

This sentence-structure-based approach ensures that the model processes text according to the grammatical and syntactic norms of the Kazakh language, thereby facilitating effective

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

communication. The intelligent model not only understands the content of each utterance but also adapts the avatar's responses according to the structure and communicative intent of the learner's text.

2. Speech Synthesis

Speech synthesis and intonation modeling were incorporated to enhance the realism and expressiveness of the avatar-based educational system. The speech synthesis component is implemented using the Facebook MMS TTS model [47], which supports multilingual text-to-speech conversion, including Kazakh. This model was chosen for its ability to generate natural-sounding speech while preserving correct pronunciation and prosody.

The speech synthesis process includes several core stages. To ensure clarity and accurate segmentation, the input text is preprocessed using the Kazakh dependency model, involving tokenization, normalization, and punctuation correction [48]. This linguistic analysis serves as a neural pipeline for natural language processing (NLP), providing syntactic and semantic analysis crucial for speech synthesis and gesture synchronization [43].

Sentence tokenization and syntactic parsing identify individual words and sentence boundaries, while POS-tagging and dependency parsing determine grammatical roles such as subject, predicate, and object. Furthermore, semantic role labeling identifies key components—subjects, predicates, and adverbials—enabling dynamic adjustment of pitch and movement.

For example, in the sentence "Ұстаз әдемілеп оқыды." ("The teacher read beautifully."):

Ұстаз → Subject

әдемілеп → Adverbial modifier

оқыды → Predicate

This structured information is then used to generate an intonational template for speech synthesis, resulting in natural and contextually appropriate delivery.

In the avatar-based interactive system, the speech synthesis component relies on the Facebook MMS TTS platform, which supports Kazakh speech generation with prosodic features. Its primary function is to convert lecture text into audio output with natural intonation aligned with the semantic content and calculated sentiment score.

Table 1 presents the baseline pause-duration values determined for different punctuation marks. These values were empirically determined based on prior research [2] and further corroborated by current studies on pause behavior in reading and spontaneous speech. For example, [49] found that pause durations at commas (within sentences) and periods (between sentences) significantly affect listener perceptions of naturalness and rate, identifying optimal parameters for inter-sentence pauses. Moreover, the study [50] demonstrated that pause durations systematically increase with higher-level text structure (e.g., sentence end vs clause boundary) and presence of punctuation marks.

Table 1. Baseline pause duration

Punctuation mark	Average pause (ms)
Comma (,), semicolon (;) and colon (:)	300
Period (.)	350
Exclamation mark (!)	400
Question mark (?)	450
Dash (–)	380

Hence, the values in Table 1 reflect both the findings from our earlier work [2] and the empirical evidence from recent literature showing that sentence-final punctuation warrants

DOI: 10.37943/24VEWA2615

longer pauses than intra-sentential boundaries. These established norms underpin our design choice to assign ~300 ms to comma/semicolon/colon boundaries, ~350 ms for period (sentence end), ~400 ms for exclamation, ~450 ms for question mark, and ~380 ms for dash (—), thereby aligning pause durations with prosodic and syntactic salience.

After linguistic analysis, the system generates speech through Facebook MMS TTS with controlled intonation. The sentence is sent to MMS TTS, which produces an initial WAV file. The generated speech preserves the original Kazakh sentence structure but requires pitch adjustment for natural intonation.

To precisely control intonation, the fundamental frequency (F_0) of the synthesized speech (F_{TTS}) is extracted using the *YIN algorithm* [51], chosen for its high accuracy compared to traditional autocorrelation-based methods. YIN minimizes pitch detection errors via parabolic interpolation and performs well even in noisy conditions, making it reliable for Kazakh speech synthesis. Moreover, it performs well even under noisy conditions, which makes it reliable for Kazakh speech synthesis. Compared to complex deep learning–based methods [52], YIN offers a good balance between accuracy and computational efficiency. While many pitch detection algorithms [53], [54] are optimized for music, YIN is specifically tuned for human speech, making it ideal for pitch and intonation correction.

The YIN function calculates the periodicity of a speech waveform using the following formula:

$$d(\tau) = \sum_{t=1}^{N} [s(t) - s(t+\tau)]^2, \quad (1)$$

where $d(\tau)$ measures how much the signal changes for different time delays (τ), $s(t)$ is the speech signal waveform, and N is the number of discrete samples. By finding the minimum value of $d(\tau)$, the algorithm determines the most stable periodicity corresponding to the fundamental frequency $\mathrm{F_{TTS}}$.

Once F_{TTS} is obtained, the PSOLA (Pitch-synchronous overlap and add) method [55] is applied to ensure that the pitch of each word remains within a predefined range:

$$F_{cor} = max\ (F_{min}, min\ (F_{TTS}, F_{max})) \quad (2)$$

Here, F_{TTS} is the initial TTS-generated pitch, while F_{max} and F_{min} define allowable frequency bounds for each word. Adjustments (Fig.3) produce smooth and natural-sounding synthesized speech, subsequently used for motion synchronization.

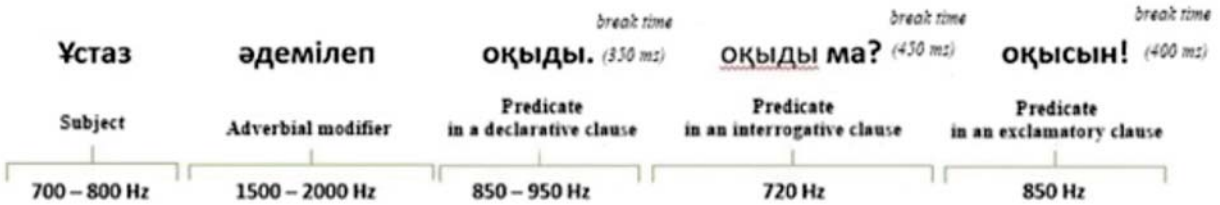

Figure 3. Pitch contour representation (in Hz) for example sentences

Thus, the architectural design (Prosody Controller + SSML) and algorithmic implementation (YIN + PSOLA) work jointly: the first manages high-level prosody control, while the second performs precise acoustic corrections. Together, they enable realistic sound generation and synchronization of speech with nonverbal avatar expressions. To formalize the prosody control mechanism for Kazakh speech synthesis, we can express the adjustment of prosodic parameters — such as pitch, speech rate, and pause duration — as a linear combination of syntactic and clause-type features:

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

$$P_i = P_0 + \alpha_{syn} \cdot S_i + \alpha_{clause} \cdot C_i, \quad (3)$$

where P_i is the adjusted prosodic parameter for the i-th segment (pitch in Hz, speech rate in syllables/sec); P_0 is the baseline prosodic value (context-specific or default); S_i represents the syntactic salience of the segment; C_i denotes clause type weighting (e.g., declarative, interrogative, exclamatory); α_{syn} and α_{clause} are empirically tuned coefficients controlling the influence of each factor to maintain naturalness.

To extend the prosody adjustment mechanism beyond speech generation, this model operates in close connection with the multimodal synchronization controller described in Section 4.3.4 Specifically, the prosodic parameter P_i not only defines the acoustic properties of synthesized speech but also directly modulates the timing and intensity of corresponding avatar gestures. The linguistic variables S_i (syntactic salience) and C_i (clause type) guide both intonation and nonverbal expressiveness — for example, interrogative clauses or highly salient words yield greater pitch variation and stronger gesture amplitude. This linkage ensures that synchronization encompasses not only timing but also expressive coherence, allowing the avatar to convey semantic and emotional nuance in real time.

In the context of Kazakh speech synthesis, previous studies have explored the impact of syntactic structure on pause duration. For instance, research has shown that the placement and duration of pauses in Kazakh speech can be influenced by syntactic boundaries and clause types, which are essential for natural-sounding speech [56]. To support the development of such models, the KazakhTTS2 corpus provides a substantial dataset for training text-to-speech systems in Kazakh, encompassing a variety of speakers and topics [57]. By integrating these linguistic features into the mechanism, the system can produce more natural and contextually appropriate speech outputs.

3. Lip Synchronization via Wav2Lip

Lip synchronization is a crucial component in creating a realistic and engaging avatar for interactive learning. To ensure accurate synchronization, the *Wav2Lip* model [58] is employed, aligning the avatar's lip movements with speech generated by the MMS TTS model. This model uses deep learning techniques to map audio waveforms to corresponding mouth movements, producing visually smooth and natural articulation.

Wav2Lip is designed to accept audio generated by the TTS system and synchronize it with the avatar's mouth movements in real time. The model was trained on a large dataset of paired video and audio samples, enabling it to learn how human lips move when pronouncing various sounds.

After the TTS system generates speech, Wav2Lip receives the corresponding audio file. The algorithm analyzes the sound and adjusts the avatar's mouth movements according to phonetic features, ensuring accurate articulation of each phoneme. For example, when producing the Kazakh sound *"S"*, the avatar's lips open slightly, while for *"P"*, they close tightly and then open, precisely mimicking phonetic characteristics.

Such audio-visual synchronization ensures full alignment between lip movement and speech, resulting in realistic and lifelike interaction. If synchronization were inaccurate, the avatar would appear robotic or desynchronized, undermining the sense of natural communication.

A key feature of Wav2Lip is its ability to synchronize mouth movements in real time, enabling smooth and responsive interaction. As MMS TTS generates speech, Wav2Lip processes the audio stream simultaneously, ensuring continuous alignment between sound and visual motion. This is critical for interactive systems where the avatar must react immediately to user input.

Synchronization occurs with imperceptible latency, maintaining natural and seamless interaction. Moreover, Wav2Lip is computationally efficient, allowing simultaneous speech gener-

DOI: 10.37943/24VEWA2615

ation and synchronization. This enhances immersion and engagement — key aspects of interactive educational systems.

The model adapts to natural rhythms of speech, accurately adjusting facial and lip movements, enhancing realism. As a result, the avatar not only delivers expressive and accurate speech via TTS but also synchronizes lip movements perfectly with sound. This combination of natural speech generation and flawless synchronization provide learners with an engaging and authentic educational experience.

4. System Integration

The key contributions of this work are threefold. First, a custom *(analyze_kazakh_sentence)* function for automated linguistic analysis of Kazakh was developed, incorporating tokenization, morphological parsing, syntactic dependency extraction, and sentence constituent classification. Second, we designed an adjustment module of the prosodic parameter that dynamically modulates pitch, speaking rate, and pause duration based on syntactic structure and sentiment features. Third, these components are fully integrated into a unified multimodal synchronization framework that combines speech synthesis, lip synchronization, and avatar animation, enabling real-time, expressive, and context-aware interactions. Collectively, these contributions provide a novel methodological approach for Kazakh-language intelligent tutoring systems, extending previous work [2] and offering tools specifically adapted to the linguistic and prosodic characteristics of Kazakh.

To achieve flawless and natural synchronization between voice and gesture animation, the integration is developed based on specific principles and mathematical formulas. The process relies on the following fundamental concepts.

First, when temporally aligning speech and gestures, the duration of a voice fragment, obtained using the frequency correction coefficient (F_{cor}), must correspond to the duration of the associated gesture animation. This ensures synchronization of movements with speech, creating a sense of natural and smooth interaction.

In mathematical form, this is expressed as follows:

$$T_G = T_S \tag{4}$$

where T_S — the corrected duration of speech;T_G — the duration of the gesture animation corresponding to the sentence structure.

If $T_G \neq T_S$ (in the case of pre-calculated animations), the playback speed of the gesture S_G is adjusted as follows:

$$S_G = T_S / T_G \tag{5}$$

where $S_G > 1$ means that the animation must be played faster to match the speech; $S_G < 1$ indicates that the animation speed should be reduced to ensure proper synchronization. Such adjustment guarantees synchrony of gestures with speech while maintaining a natural temporal relationship between these elements.

At the end of a sentence, a pause T_B is introduced, providing a natural break before the next utterance:

$$T_T = T_S + T_B \tag{6}$$

where T_B is determined depending on the type of sentence. To preserve visual consistency, the final frame of the gesture animation is held for the duration of T_B, which helps avoid unnecessary or unnatural transitions into an idle state.

To maintain a natural flow between movements and to prevent static or unnatural poses during speech pauses, pre-calculated blinking overlays are applied. The probability that a blink will occur during the pause T_B is modeled as follows:

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

$$B_B = f(T_B) \tag{7}$$

where B_B — the probability of a blink occurring during the pause; T_B — the duration of the speech pause.

Since blinks are pre-calculated and implemented as MP4 overlays, they can be easily integrated into the animation process without the need for real-time rendering. This approach ensures natural facial activity of the avatar during speech pauses, eliminates the perception of discreteness, and enhances the realism of interaction.

The objective of the multimodal synchronization controller is to minimize discrepancies between speech and gesture durations while preserving smooth transitions of prosody (pitch and rate) across consecutive segments. This is expressed as an optimization problem:

$$E = min_{s_i,b_i}\left|\sum_{i=1}^{N}(T_{S,i} - s_i T_{G,i})^2 + \lambda \sum_{i=1}^{N}(T_{S,i} - T_{S,i-1})^2\right| \tag{8}$$

where the first term enforces temporal alignment between speech and gesture, and the second term (weighted by λ) ensures temporal smoothness, preventing abrupt changes between segments. While the general quadratic optimization form is standard in signal synchronization problems, its application here is novel: it defines the first unified timing–prosody controller for Kazakh-language avatar-based intelligent tutoring systems.

$$s_i = T_{S,i} / T_{G,i} \tag{9}$$

and linear interpolation of intermediate durations. This ensures real-time performance with computational complexity $O(N)$.

The system solves this optimization locally for each sentence. Because the terms are quadratic, the solution is obtained efficiently using direct computation of

The avatar-based learning system is developed step by step, with each stage using advanced technologies to create an interactive and emotionally responsive avatar. The system integrates NLP, speech synthesis, synchronization of gestures and facial expressions, as well as lip synchronization. Technically, the integration is carried out in a Python 3.10 environment, which serves as the orchestrator between all system components. Through APIs and modules, the following are connected:

1. Stanza — for syntactic parsing and extraction of structural units [46];
2. MMS TTS — for speech synthesis with sentiment parameters [47];
3. Wav2Lip — for lip synchronization with audio output [58];
4. Avatar animation modules — for visual expressivity [2].

Thus, the systemic integration approach unifies the semantic, acoustic, and visual levels into a single coherent model. As a result, a lecturer-avatar is created, who does not simply reproduce the lecture text but delivers it in a form that closely resembles live, emotionally rich communication between a teacher and students.

Development of an interactive content and interface model based on an avatar

In the process of creating the IILS [2], special attention is paid to developing a model of interactive content and interface centered around a virtual avatar. The avatar functions as a digital instructor, capable not only of delivering educational material but also adapting to the context of interaction, maintaining the learner's attention through dynamic visual and verbal presentation.

The interactive content model is based on the ontological structure of the course, ensuring logical connections between concepts, modules, and tasks. The content includes text blocks, synthesized voiceovers, video components featuring the avatar, and animated elements (gestures, facial expressions, lip movements) that enhance information perception. Each content

DOI: 10.37943/24VEWA2615

element is accompanied by metadata that defines its placement within the learning trajectory, including learning objectives, expected skills, and corresponding difficulty levels.

The system implements two main methods for generating educational video content with the avatar. The first method allows the user to input text material, from which a video featuring the avatar is automatically generated. The avatar delivers the content verbally. A window will appear on the screen to preview the interface (Fig.2). From the menu on the right, users can select

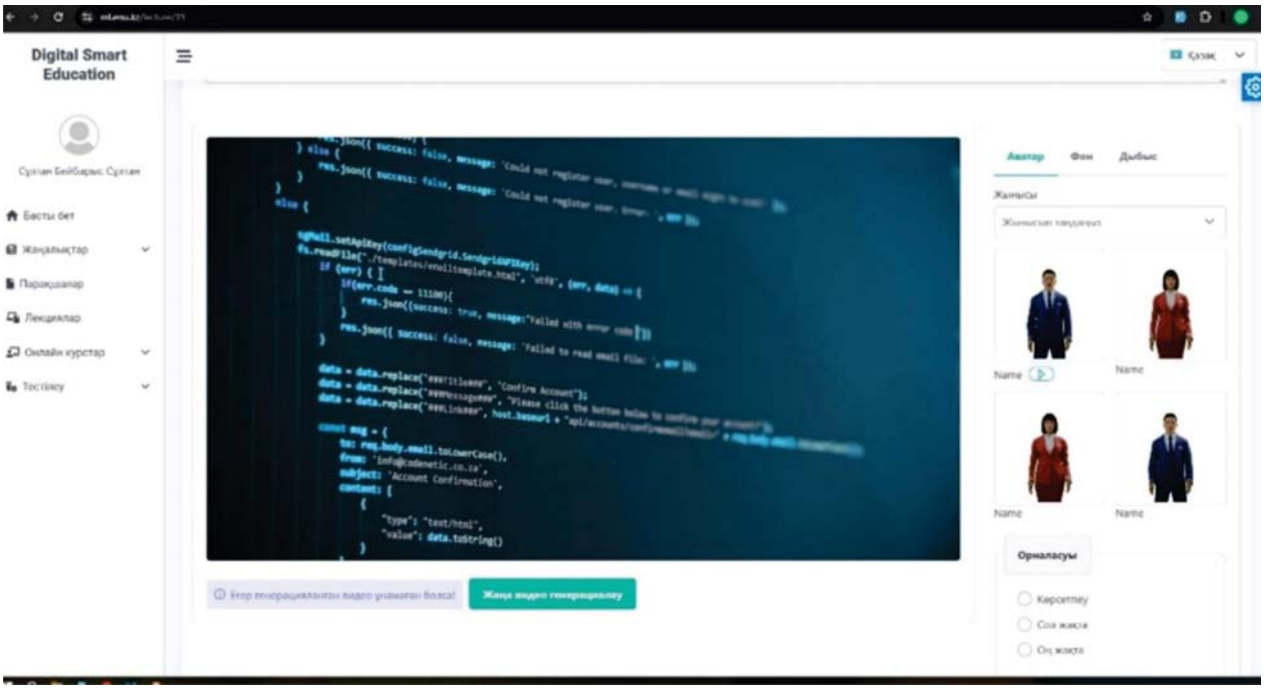

Figure 2. Interface of the Digital Smart Education system

the avatar type, interface background, and sound template. Users can also choose the avatar's position—placing it on the left or right side of the screen or removing it entirely. The speech is synchronized with lip movements, gestures, and facial expressions, and is accompanied by subtitles—making the material both accessible and expressive materials. This approach enables quick creation of adaptive educational videos on any topic.

The second method is designed to work with existing educational presentations (PDF format). After uploading a presentation, the system converts each slide into a video segment featuring the avatar reading the slide text in a chosen template. Additionally, quiz questions and their corresponding answers are automatically generated based on the lecture materials, and a built-in response model enables interactive knowledge assessment.

The interface is developed based on the principles of human-centered design, multimodality, and adaptability. Visually, it is organized to focus user attention on the avatar, which is placed at the center of the screen and can be scaled or repositioned across different devices. Navigation elements (menus, progress indicators, control panels) are arranged compactly and intuitively, avoiding visual clutter. The interface includes subtitles, mode switching such as “explanation,” “question,” “pause for reflection”, and options to control speech speed and toggle gestures and facial expressions. Special emphasis is placed on accessibility: the system supports users with hearing or visual impairments, with adaptive fonts and color schemes implemented.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

Content and interface are closely integrated with the system's technical components. Speech synthesis is implemented using MMS TTS [47], while lip movements of the avatar are synchronized using Wav2Lip [58]. Avatars were modeled and animated using Blender 3.6 [2], allowing detailed facial expressions and gesture control. To support Kazakh-language input, a dependency model for sentence structure was developed using Stanza [46], and a sentence-based gesture mapping system was designed [2] to ensure that avatar gestures align naturally with spoken content. Furthermore, a custom intonational model for the Kazakh language was developed to improve the naturalness and expressiveness of speech delivery. After the text is processed by NLP modules and synthesized, the avatar receives synchronized control signals for gesture, expression, and intonation rendering. This creates the effect of live interaction with a human instructor. When user feedback is received (answers to questions), the system dynamically adjusts both visual and speech components. Moreover, Python 3.10 serves as the core programming language, coordinating and automating the interaction of all components in real time [2]. Its rich ecosystem of scientific libraries and system monitoring tools supports reliable system execution and enables effective performance tracking throughout the development and deployment process. Thus, the avatar-based interactive content and interface model ensures deep integration of educational materials with audiovisual representation, fostering immersion, emotional engagement, and improved knowledge retention.

Results

Speech synthesis is critical in the IILS, since it directly influences the quality of interaction between the user and the virtual avatar. To achieve naturalness, responsiveness, and real-time processing, the system incorporates the variational inference text-to-speech (VITS) model from Facebook's massively multilingual speech (MMS TTS) framework. This model was chosen based on a prior internal comparison study with alternative architectures—details of which are currently under peer review — and was found to deliver superior performance across key technical criteria.

To determine its suitability for real-time educational applications, we conducted an internal performance evaluation of the VITS model. The focus was on four critical metrics: real-time factor (RTF), latency, throughput, and resource consumption. These metrics were selected due to their relevance in real-world deployment scenarios where the system must operate efficiently and responsively on a range of hardware configurations, including resource-constrained environments. RTF measures the speed of audio generation relative to the input text, with values below 1 indicating real-time capability. Latency is important for maintaining a fluid interactive experience, especially when learners ask questions or receive immediate feedback. Throughput indicates how much text the system can process per second, which is vital for scalability and smooth lecture delivery. Resource consumption — measured via CPU, GPU, and memory usage — helps assess the system's feasibility for widespread deployment without requiring high-end infrastructure.

The evaluation was carried out using a pre-trained VITS model in a controlled environment that simulated real-time application conditions. The results, summarized in Table 1, confirm that the model is both efficient and reliable, making it well-suited for integration into an adaptive and immersive educational system like the IILS.

DOI: 10.37943/24VEWA2615

Table 2. Performance evaluation of VITS model (Facebook MMS TTS)

Metric	Value	Description
RTF	0.004 ± 0.003	Measures generation speed relative to audio length (RTF < 1 = real-time capable)
Latency	25.51 ± 9.73 milliseconds	Time from text input to audio output (lower = more responsive)
Throughput	4274.0±3003.7 characters per second	Number of characters processed per second (higher = faster synthesis)
Peak GPU memory usage	150.68 MB	Maximum GPU memory used during inference
Peak CPU memory usage	1751.03 MB	Maximum RAM usage during synthesis
CPU utilization	1.1%	Percentage of CPU resources used during inference

These parameters demonstrate the model's high efficiency and responsiveness, making it suitable for deployment in systems with limited computational resources [59]. To support this evaluation, various Python-based tools were employed: *PyTorch* was used to run the VITS model. Moreover, *psutil* and *pynvml* collected system-level performance metrics. In addition, *glob*, *numpy*, and *scipy.io.wavfile* handled batch file processing and numerical analysis. Finally, *time* and *tqdm* monitored latency and inference speed during testing.

Additionally, the evaluation was part of a broader internal comparison study that included the Tacotron model [47]. Although the full findings are currently under peer review, the results demonstrated that VITS consistently outperformed Tacotron across all major metrics. These findings directly informed the decision to adopt the VITS model in the final system configuration.

Overall, the performance evaluation of the VITS model confirms its reliability and effectiveness for integration into the avatar-based learning environment, ensuring fast, natural, and computationally efficient speech synthesis necessary for immersive and adaptive education.

While avatar-based platforms such as Synthesia [60] and D-ID [61] have demonstrated substantial progress in AI-driven video generation and interactive media, their primary orientation lies outside of educational contexts. In contrast, the proposed IILS has been specifically designed with pedagogical considerations and language learning in mind. A systematic comparison highlights several key distinctions (Table 3).

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

Table 3. Comparative analysis of AI-driven avatar platforms

Dimension	Synthesia	D-ID	Proposed IILS
Pedagogical orientation	Corporate training, explainer videos.	Customer support, content automation.	Designed for language learning and intelligent tutoring.
Learning theory alignment	Presentation-focused; no explicit pedagogical grounding.	Limited interactivity; not grounded in learning theory.	Dual-channel and embodied learning integration; adaptive feedback.
Real-time adaptivity	Pre-rendered content.	Semi-real-time interaction.	Full real-time synchronization with speech, gesture, and emotion.
Cultural/linguistic relevance	Generic multilingual support; weak phonetic alignment.	Translation-driven; lacks culture-specific nuance.	Native Kazakh phonetics, culturally aligned gestures.
Scalability & accessibility	Cloud-dependent, subscription-based.	Cloud-first, high resource demand.	Hybrid deployment (local + cloud), adaptable to low-resource environments.
Ethical safeguards	GDPR-compliant, but vulnerable to misuse.	Security protocols, potential deepfake repurposing.	Controlled educational context, institutional data privacy policies.

Synthesia and D-ID primarily operate as media production platforms designed for marketing, customer support, and corporate training [62]. Their architectures are optimized for scalable content generation but lack mechanisms for pedagogical adaptation. In contrast, the proposed IILS is developed as an intelligent tutoring environment, embedding multimodal synchronization and adaptive dialogue to promote deeper learner engagement.

Commercial avatar systems rarely integrate established principles of multimedia learning or cognitive load theory. Their focus remains on achieving visual realism without explicit consideration of instructional design [63]. By contrast, IILS applies dual-channel processing and embodied learning theory, coordinating gestures, speech, and visual cues to minimize extraneous load and enhance semantic encoding.

Moreover, Synthesia and D-ID generally rely on pre-rendered or semi-automated outputs. While these outputs are visually polished, they restrict interactivity and hinder real-time responsiveness [2]. IILS introduces adaptive synchronization and sentiment-driven modulation, enabling conversational flow that dynamically responds to learner input and affective states. Although commercial platforms claim multilingual support across hundreds of languages, they often lack phonetic precision and cultural nuance. For instance, Kazakh is either unsupported or inaccurately represented in existing systems. IILS directly addresses this gap by incorporating Kazakh-specific phonetic adjustments and culturally relevant gestures, thereby ensuring linguistic authenticity and contextual appropriateness.

Both Synthesia and D-ID are cloud-based and computationally intensive, limiting their usability in low-resource educational contexts. IILS, however, supports hybrid deployment, including institutional server installation, which enhances accessibility for schools and universities with restricted infrastructure [64]. While commercial systems adopt GDPR and SOC 2 compliance frameworks, their general-purpose design creates risks of misuse, including deepfake generation and manipulative media. IILS mitigates these risks by constraining its application to education, implementing localized data privacy policies, and ensuring transparent and pedagogically aligned learner–system interactions.

The combined findings indicate that while commercial platforms such as Synthesia and D-ID achieve visual realism and scalability, they fall short in addressing real-time adaptivity,

DOI: 10.37943/24VEWA2615

pedagogical grounding, and cultural-linguistic precision. By contrast, the IILS bridges these gaps by integrating high-performing speech synthesis with instructional design principles, positioning it as a novel contribution to intelligent tutoring in underrepresented languages such as Kazakh.

Discussion

The development of the IILS demonstrates the potential of combining advanced digital technologies—such as speech synthesis, NLP, and multimodal user interfaces—to enhance the quality and accessibility of education. While the technical implementation and architectural components of the system have been described in greater detail in our previous work [2], this article focuses on system evaluation, user interaction, and the effectiveness of the integrated components in supporting adaptive, real-time learning.

One of the system's core components is the VITS speech synthesis model from Facebook's MMS TTS framework. The evaluation revealed that the model performs exceptionally well in real-time conditions, achieving the RTF of 0.004 and low latency. These metrics suggest that VITS provides responsive and natural-sounding speech with minimal computational overhead, making it highly suitable for deployment in educational platforms with limited resources. Compared to alternative models such as Tacotron, VITS showed consistently better performance in terms of speed, responsiveness, and resource efficiency.

The system architecture—with its modular, layered design — supports dynamic lecture presentation, real-time question answering, and individualized feedback. Notably, instead of relying on formal ontology, the system uses rule-based segmentation of lecture content and semantic matching techniques to generate context-aware responses. This approach maintains topic coherence and enables relevant, lecture-based answers to learner inquiries. Similarly, assessment items such as multiple-choice and open-ended questions are generated automatically from lecture text and slide content using NLP tools, facilitating personalized formative evaluation without requiring manual authoring.

The user interface of the IILS contributes significantly to learner engagement and accessibility. Developed using Blender and Wav2Lip, the avatar is capable of delivering speech with synchronized lip movements, facial expressions, and gestures. These features enhance the emotional and cognitive engagement of learners. The interface is designed to be intuitive and adaptable, with options for customizing the avatar's appearance, voice, background, and layout. Accessibility considerations, such as subtitles, playback controls, and high-contrast modes, further support diverse learner needs, including those with sensory impairments.

A key contribution of this project is the inclusion of Kazakh-language support. This was achieved through a customized intonation model, sentence parsing using Stanza, and synchronized gesture mapping. By addressing the linguistic and cultural needs of underrepresented communities, the IILS extends the reach of intelligent learning technologies and supports equitable access to digital education.

While the current system offers a strong foundation for interactive and personalized learning, several areas for future improvement remain. For example, personalization is primarily reactive and based on direct user input. Integrating advanced learner modeling and behavioral analytics could allow for more proactive and individualized support. In addition, the system's affective capabilities — limited to voice tone and avatar gestures — could be enhanced with emotion detection technologies to foster more empathetic and responsive interaction. Further integration with learning management systems would also enable more comprehensive tracking of learner progress and outcomes.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

Conclusion

This study has introduced an intelligent interactive learning system that demonstrates the feasibility of scalable, avatar-based education. The system integrates rule-based content processing, real-time speech synthesis, and multimodal interaction to support dynamic, inclusive, and pedagogically grounded learning experiences. The evaluation results suggest that the proposed approach is not only technically viable but also pedagogically promising, particularly in the domains of language learning, intelligent tutoring, and digital education. A distinctive contribution of the IILS lies in its explicit support for underrepresented languages such as Kazakh, achieved through phonetic precision and culturally relevant multimodal cues. This feature highlights the system's potential to advance equity in educational technology by extending access to linguistic communities that are frequently overlooked by mainstream platforms.

Nevertheless, several avenues for further development remain. Future work should investigate enhanced personalization through adaptive learner modeling, improved emotional interactivity via affective computing, and stronger interoperability with existing learning management systems. Addressing these directions will be essential for scaling the system across diverse contexts while maintaining responsiveness, cultural sensitivity, and pedagogical effectiveness. In sum, while the present findings underscore the potential of IILS as a novel contribution to multimodal intelligent tutoring, further empirical validation and longitudinal studies will be necessary to fully establish its educational impact.

Acknowledgment

This research is funded by the Scientific Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP23489504).

References

[1] Ukenova, A., & Bekmanova, G. (2023). A review of intelligent interactive learning methods. Frontiers in Computer Science, 5, 1141649. https://doi.org/10.3389/fcomp.2023.1141649

[2] Ukenova, A., Bekmanova, G., Zaki, N., Kikimbayev, M., & Altaibek, M. (2025). Assessment and Improvement of Avatar-Based Learning System: From Linguistic Structure Alignment to Sentiment-Driven Expressions. Sensors, 25(6), 1921. https://doi.org/10.3390/s25061921

[3] Bekmanova, G., Ongarbayev, Y., Somzhurek, B., & Mukatayev, N. (2021). Personalized training model for organizing blended and lifelong distance learning courses and its effectiveness in Higher Education. *Journal of Computing in Higher Education, 33*(3), 668-683. https://doi.org/10.1007/s12528-021-09282-2

[4] Gm, D., Goudar, R. H., Kulkarni, A. A., Rathod, V. N., & Hukkeri, G. S. (2024). A digital recommendation system for personalized learning to enhance online education: A review. *IEEE Access*, *12*, 34019-34041.

[5] Lin, C. C., Huang, A. Y., & Lu, O. H. (2023). Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. *Smart Learning Environments*, *10*(1), 41. https://doi.org/10.1186/s40561-023-00260-y

[6] Dong, J., Mohd Rum, S. N., Kasmiran, K. A., Mohd Aris, T. N., & Mohamed, R. (2022). Artificial intelligence in adaptive and intelligent educational system: a review. *Future Internet*, *14*(9), 245. https://doi.org/10.3390/fi14090245

[7] Alshwaier, A., Youssef, A., & Emam, A. (2012). A new trend for e-learning in KSA using educational clouds. *Advanced Computing*, *3*(1), 81.

[8] Wu, W., & Plakhtii, A. (2021). E-learning based on cloud computing. *International Journal of Emerging Technologies in Learning (IJET)*, *16*(10), 4-17.

[9] Shi, H., & Shi, C. (2022). Intelligent interactive English teaching system for engineering education. *Advances in Multimedia*, *2022*(1), 4676776. https://doi.org/10.1155/2022/4676776

[10] Li, X. (2022). Intelligent interactive english teaching discrete data modeling and simulation. *Scientific Programming*, *2022*(1), 3807762. https://doi.org/10.1155/2022/3807762

[11] Rafiq, M. S., Jianshe, X., Arif, M., & Barra, P. (2021). Intelligent query optimization and course recommendation during online lectures in E-learning system. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 10375-10394. https://doi.org/10.1007/s12652-020-02834-x
[12] Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business horizons*, *62*(1), 15-25. https://doi.org/10.1016/j.bushor.2018.08.004
[13] Gao, P., Li, J., & Liu, S. (2021). An introduction to key technology in artificial intelligence and big data driven e-learning and e-education. *Mobile Networks and Applications*, *26*(5), 2123-2126. https://doi.org/10.1007/s11036-021-01777-7
[14] Murtaza, M., Ahmed, Y., Shamsi, J. A., Sherwani, F., & Usman, M. (2022). AI-based personalized e-learning systems: Issues, challenges, and solutions. *IEEE access*, *10*, 81323-81342. doi: 10.1109/ACCESS.2022.3193938
[15] Ouyang, F., Zheng, L., & Jiao, P. (2022). Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. *Education and Information Technologies*, *27*(6), 7893-7925. https://doi.org/10.1007/s10639-022-10925-9
[16] Huang, A. Y., Lu, O. H., & Yang, S. J. (2023). Effects of artificial Intelligence–Enabled personalized recommendations on learners' learning engagement, motivation, and outcomes in a flipped classroom. *Computers & Education*, *194*, 104684. https://doi.org/10.1016/j.compedu.2022.104684
[17] Ali, J. K. M., Shamsan, M. A. A., Hezam, T. A., & Mohammed, A. A. (2023). Impact of ChatGPT on learning motivation: teachers and students' voices. *Journal of English Studies in Arabia Felix*, *2*(1), 41-49. https://doi.org/10.56540/jesaf.v2i1.51
[18] Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International journal of information management*, *57*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002
[19] Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). Opinion Paper:"So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International journal of information management*, *71*, 102642.
[20] Bag, S., Srivastava, G., Bashir, M. M. A., Kumari, S., Giannakis, M., & Chowdhury, A. H. (2022). Journey of customers in this digital era: Understanding the role of artificial intelligence technologies in user engagement and conversion. *Benchmarking: An International Journal*, *29*(7), 2074-2098. https://doi.org/10.1108/BIJ-07-2021-0415
[21] Alhazmi, A. K., Alhammadi, F., Zain, A. A., Kaed, E., & Ahmed, B. (2023). AI's role and application in education: Systematic review. *Intelligent Sustainable Systems: Selected Papers of WorldS4 2022, Volume 1*, 1-14. https://doi.org/10.1007/978-981-19-7660-5_1
[22] Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52-62. https://doi.org/10.61969/jai.1337500
[23] Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121-154. https://doi.org/10.1016/j.iotcps.2023.04.003
[24] Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, *28*(1), 973-1018. https://doi.org/10.1007/s10639-022-11177-3
[25] Bhaskaran, S., Marappan, R., & Santhi, B. (2021). Design and analysis of a cluster-based intelligent hybrid recommendation system for e-learning applications. *Mathematics*, *9*(2), 197. https://doi.org/10.3390/math9020197
[26] Soni, V. D. (2019). IOT connected with e-learning. *International Journal on Integrated Education*, *2*(5), 273-277.
[27] Kumar, K., & Al-Besher, A. (2022). IoT enabled e-learning system for higher education. *Measurement: Sensors*, *24*, 100480. https://doi.org/10.1016/j.measen.2022.100480

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

[28] Farahani, B., Firouzi, F., Chang, V., Badaroglu, M., Constant, N., & Mankodiya, K. (2018). Towards fog-driven IoT eHealth: Promises and challenges of IoT in medicine and healthcare. *Future generation computer systems*, *78*, 659-676. https://doi.org/10.1016/j.future.2017.04.036

[29] Tanty, H., Fernando, C., Valencia, J., & Justin, V. (2022). Critical thinking and problem solving among students. *Business Economic, Communication, and Social Sciences Journal (BECOSS)*, *4*(3), 173-180. https://doi.org/10.21512/becossjournal.v4i3.8633

[30] Al-Nuaimi, M. N., & Al-Emran, M. (2021). Learning management systems and technology acceptance models: A systematic review. *Education and information technologies*, *26*(5), 5499-5533. https://doi.org/10.1007/s10639-021-10513-3

[31] Deepika, M., Kavitha, M., Chakravarthy, N. K., Rao, J. S., Reddy, D. M., & Chandra, B. M. (2021, January). A critical study on campus energy monitoring system and role of IoT. In *2021 International Conference on Sustainable Energy and Future Electric Transportation (SEFET)* (pp. 1-6). IEEE.

[32] Veluvali, P., & Surisetti, J. (2022). Learning management system for greater learner engagement in higher education—A review. *Higher Education for the Future*, *9*(1), 107-121. https://doi.org/10.1177/23476311211049855

[33] Alrakhawi, H. A., Jamiat, N., & Abu-Naser, S. S. (2023). Intelligent tutoring systems in education: a systematic review of usage, tools, effects and evaluation. *Journal of Theoretical and Applied Information Technology*, *101*(4), 1205-1226.

[34] Chao, Z., Qing, S., & Mingwen, T. (2023). A study of multimodal intelligent adaptive learning system and its pattern of promoting learners' online learning engagement. *Psychol Res*, *13*(5), 202-6. doi:10.17265/2159–5542/2023.05.002.

[35] Al Omoush, M. H., Salih, S. E., Kishore, S., & Mehigan, T. (2023, November). Interactive multimodal learning: towards using pedagogical agents for inclusive education. In *2023 IEEE International Humanitarian Technology Conference (IHTC)* (pp. 1-7). IEEE. doi: 10.1109/IHTC58960.2023.10508848.

[36] Jung, M., Lim, Y., Kim, S., Jang, J. Y., Shin, S., & Lee, K. H. (2022, October). An emotion-based Korean multimodal empathetic dialogue system. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI* (pp. 16-22). https://aclanthology.org/2022.cai-1.3/

[37] Bekmanova, G., Ukenova, A., Omarbekova, A., Zakirova, A., & Kantureyeva, M. (2024, July). Features of the interface of system for solving social problems. In *2024 8th International Conference on Computer, Software and Modeling (ICCSM)* (pp. 5-13). IEEE.

[38] Azofeifa J. D. et al. Systematic review of multimodal human–computer interaction //Informatics. – MDPI, 2022. – Т. 9. – №. 1. – С. 13.

[39] Li, W., Yu, J., Zhang, Z., & Liu, X. (2022). *Dual coding or cognitive load? Exploring the effect of multimodal input on English as a foreign language learners' vocabulary learning*. Frontiers in Psychology, 13, 834706.

[40] Vasilaki, E., & Mavrogianni, A. (2025). *Extending Cognitive Load Theory: The CLAM Framework for Biometric, Adaptive, and Ethical Learning*. Psychology International, 7(2), 40.

[41] AlShaikh, R., Al-Malki, N., & Almasre, M. (2024). *The implementation of the cognitive theory of multimedia learning in the design and evaluation of an AI educational video assistant utilizing large language models*. Heliyon, 10(3).

[42] Worsley, M., Barel, D., Davison, L., Large, T., & Mwiti, T. (2018, June). *Multimodal interfaces for inclusive learning. In International Conference on Artificial Intelligence in Education* (pp. 389-393). Cham: Springer International Publishing.

[43] Ermakova, T., Fabian, B., Golimblevskaia, E., & Henke, M. (2023). *A comparison of commercial sentiment analysis services*. SN Computer Science, 4(5), 477.

[44] Yergesh, B., Bekmanova, G., & Sharipbay, A. (2019, February). *Sentiment analysis of Kazakh text and their polarity*. In Web Intelligence (Vol. 17, No. 1, pp. 9-15). Sage UK: London, England: SAGE Publications. https://doi.org/10.3233/WEB-190396

[45] Huang, X. (2020, February). *Construction and application of online course teaching in intelligent learning environment*. In The International Conference on Cyber Security Intelligence and Analytics (pp. 702-709). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-43306-2_99

[46] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). *Stanza: A Python natural language processing toolkit for many human languages*. arXiv preprint arXiv:2003.07082. https://doi.org/10.48550/arXiv.2003.07082
[47] Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., ... & Auli, M. (2024). *Scaling speech technology to 1,000+ languages*. Journal of Machine Learning Research, 25(97), 1-52. https://doi.org/10.48550/arXiv.2305.13516
[48] Kumar, S., Soni, N., & Maurya, A. K. (2025). *Multi-model review classification based on sentiments analysis. In Intelligent Computing and Communication Techniques* (pp. 73-80). CRC Press. https://doi.org/10.3389/fpsyg.2022.778018
[49] Šturm, P., & Volín, J. (2023). *Occurrence and duration of pauses in relation to speech tempo and structural organization in two speech genres. Languages*, 8(1), 23. https://doi.org/10.3390/languages8010023
[50] Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., ... & Liu, T. Y. (2024). *Naturalspeech: End-to-end text-to-speech synthesis with human-level quality*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(6), 4234-4245.
[51] Jeong, C. Y., Song, Y., Shin, S., & Kim, M. (2025). *Efficient pitch-estimation network for edge devices*. ETRI Journal, 47(1), 112-122.
[52] Brata, I. P. B. W., & Darmawan, I. D. M. B. A. (2021). *Comparative study of pitch detection algorithm to detect traditional Balinese music tones with various raw materials*. In Journal of Physics: Conference Series (Vol. 1722, No. 1, p. 012071). IOP Publishing.
[53] Huang, J., Benetos, E., & Ewert, S. (2022, May). *Improving lyrics alignment through joint pitch detection*. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 451-455). IEEE.
[54] Gujarathi, P., & Patil, S.R. (2021). *Review on unit selection-based concatenation approach in text to speech synthesis system*. In Cybernetics, Cognition and Machine Learning Applications: Proceedings of ICCCMLA 2020 (pp. 191-202). Singapore: Springer Singapore.
[55] Kaliyev, A., Rybin, S. V., Matveev, Y. N., Kaziyeva, N., & Burambayeva, N. (2018, June). *Modeling pause for the synthesis of Kazakh speech*. In Proceedings of the Fourth International Conference on Engineering & MIS 2018 (pp. 1-4).
[56] Mussakhojayeva, S., Janaliyeva, A., Mirzakhmetov, A., Khassanov, Y., & Varol, H. A. (2021). *KazakhTTS: An open-source Kazakh text-to-speech synthesis dataset*. arXiv preprint arXiv:2104.08459.
[57] Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2020, October). *A lip sync expert is all you need for speech to lip generation in the wild*. In Proceedings of the 28th ACM international conference on multimedia (pp. 484-492). https://doi.org/10.1145/3394171.3413532
[58] Liu, Z., & Prud'hommeaux, E. (2021, April). *Dependency parsing evaluation for low-resource spontaneous speech*. In Proceedings of the Second Workshop on Domain Adaptation for NLP (pp. 156-165).
[59] Lind, S. J. (2025). *Can AI-powered avatars replace human trainers? An empirical test of synthetic humanlike spokesperson applications*. Journal of Workplace Learning, 37(1), 19-40.
[60] Wang, C., & Zou, B. (2025). D-ID Studio: Empowering Language Teaching With AI Avatars. TESOL Journal, 16(2), e70034.
[61] Logeswari, P., Jebaraj, N. R. S., & BanuPriya, G. (2024). *Comparative analysis of AI tools for video production*. Journal of Information Technology Review, DLINE Journals.
[62] Cavanagh, T. M., & Kiersch, C. (2023). *Using commonly-available technologies to create online multimedia lessons through the application of the Cognitive Theory of Multimedia Learning*. Educational technology research and development, 71(3), 1033-1053.
[63] Swenson, A. (2023). *Teaching digital identity: opportunities, challenges, and ethical considerations for avatar creation in educational settings*. Brazilian Creative Industries Journal, 3(2), 41-58.