DOI: **10.37943/24AHNP6638**

Zhanat Makhataeva
PhD, Senior Data Scientist
zhanat.mahataeva@nu.edu.kz, orcid.org/0000-0001-9366-7047
Private Institution "Institute of Smart Systems and Artificial Intelligence" (ISSAI), Kazakhstan
Al-Farabi Kazakh National University, Kazakhstan
Nursultan Atymtay
Bachelor Student, Computer Science, School of Engineering and Digital Sciences (SEDS)
nursultan.atymtay@nu.edu.kz, orcid.org/0009-0003-4323-3023
Nazarbayev University, Kazakhstan
Rakhat Meiramov
MS, Data Scientist
rakhat.meiramov@nu.edu.kz, orcid.org/0000-0003-0800-288X
Private Institution "Institute of Smart Systems and Artificial Intelligence" (ISSAI), Kazakhstan
Al-Farabi Kazakh National University, Kazakhstan
Guldana Nauryzbaikyzy
PhD Candidate, Lecturer, Department of Foreign Languages
gnauryzbaikyzy@zhubanov.edu.kz, orcid.org/0000-0001-9272-8952
Private Institution "Institute of Smart Systems and Artificial Intelligence" (ISSAI), Kazakhstan
K. Zhubanov Aktobe Regional University, Kazakhstan
Kulzat Sadirova
Doctor of Philological Sciences, Professor, Department of Philology
kulzat.sadirova.70@mail.ru, orcid.org/0000-0001-6092-8191
Private Institution "Institute of Smart Systems and Artificial Intelligence" (ISSAI), Kazakhstan
K. Zhubanov Aktobe Regional University, Kazakhstan
Huseyin Atakan Varol
PhD, General Director of ISSAI and Professor of Robotics, Department of Robotics, School of Engineering and Digital Sciences (SEDS)
ahvarol@nu.edu.kz, orcid.org/0000-0002-4042-425X
Private Institution "Institute of Smart Systems and Artificial Intelligence" (ISSAI), Kazakhstan
Nazarbayev University, Kazakhstan
Al-Farabi Kazakh National University, Kazakhstan

MULTILINGUAL AUTOMATIC SPEECH RECOGNITION INTERFACE FOR TYPING: USABILITY STUDY AND PERFORMANCE EVALUATION FOR KAZAKH, RUSSIAN, AND ENGLISH

Abstract: We present a multilingual automatic speech recognition (ASR) system for Kazakh, Russian, and English designed for the trilingual community of Kazakhstan. Although prior research has shown that speech-based text entry can outperform conventional keyboard typing for human–computer interaction and interaction with large language models (LLMs), little is known about its performance and usability in low-resource multilingual contexts, particularly for Kazakh. To address this gap, we fine-tuned a Whisper-based model on additional Kazakh speech data, achieving a large reduction in Kazakh word error rate (WER) from 21.55% with the OpenAI baseline to 8.84%, while preserving competitive performance for Russian and English. We then conducted a user study with 38 participants from Nazarbayev University, who performed dictated reading and editing tasks in all three languages. We evaluated performance using WPM, CPM, WER, and CER, and assessed usability and cognitive effort using the System Usability Scale (SUS) and the Raw NASA Task Load Index (NASA-TLX). Participants reached high speech-based typing speeds without editing and moderate speeds with editing across all three languages. Importantly, there were no statistically significant differences between Kazakh, Russian, and English in error rates, cognitive load, or perceived usability. Users reported low cognitive load (NASA-TLX < 40) and consistently high usability (SUS > 80%), indicating that the interface is efficient, easy to use, and requires minimal mental effort. These results demonstrate that Kazakh-

adapted Whisper enables accurate, usable, and low-effort multilingual ASR, and highlight the potential of speech-driven text entry systems for trilingual contexts such as Kazakhstan.

Keywords: automatic speech recognition (ASR); cognitive load; usability; human-computer interaction (HCI); human-AI interaction; speech-based typing.

Introduction

Multilingual ASR Systems and LLMs

Speech technologies such as Automatic Speech Recognition (ASR) [1], speaker recognition [2], and Text-to-Speech (TTS) [3] systems are introducing novel communication methods in the fields of human-robot interaction, dialogue systems, and intelligent social agents. It is remarkable how a single Large Language Model (LLM) can be adapted to perform different tasks including writing, coding, utilizing search tools, chatbots, virtual assistants, and embodied agents [4]. LLMs as cutting-edge artificial intelligence (AI) systems are data hungry and have billions of parameters that need to be trained on massive text corpora [5]. Generally, LLMs have revolutionized the reality of AI and natural language processing (NLP) at their core, introducing a foundational shift in millions of people's everyday lives [6].

Speech-Based Text Entry Interfaces

Modern interactive input methods showcase a microphone button alongside text entry windows, indicating that voice-based entry mode is added to conventional typing-based text entry methods. As an example, Google's Gboard and Yandex Keyboard integrated a microphone icon into their keyboards. When the user taps or taps-and-holds the mic icon, the system's microphone is activated, and the spoken words appear as text in the communication window. Apple's iOS Dictation also uses the mic integrated into the on-screen keyboard. During the dictation, the on-screen keyboard keeps being open, allowing the user to switch between keyboard typing and speech-based typing. Custom web applications enable users to click a custom button to record speech-based input, view the resulting transcription as text in the communication box, and then edit it before proceeding. In commercial systems (i.e., Gboard and Yandex), dictation can run continuously across fields. Custom web applications enforce a one-phrase-at-a-time workflow for structured data collection. Usually, dictation-supported systems use common visuals, including mic icons, real-time instructions such as "Start speaking" and "Recording".

In addition, many speech-based and typing-based input methods are integrated with error correction frameworks [7]. Authors in [8] describe a mechanism for dynamic propagation of user feedback that progressively adapts the system to different speakers and lexical contexts. In [9], the authors integrated LLM with an audio encoder supporting speech-based communication with LLMs. In another work, a speech recognition system was integrated with LLM to deal with transcription errors, helping to increase the accuracy of the system [10]. Commercial UIs enhanced with voice-based input methods offer both touch-based typing and voice input for making corrections to the transcribed input text. For example, Google Docs' voice-based typing interface underlines uncertain words and offers a few alternatives for correcting the corresponding words. Specifically, users can right-click on an underlined word to see suggestions as a potential correction. Gboard and Yandex Keyboards enable users to apply voice commands for text corrections, such as "delete last word" or "clear," which removes recognized words. Some systems use, "Fix it" feature that performs auto-correction of the grammar as the post-dictation text processing

.

Kazakhstan Context and Fine-Tuned ASR Model

In this work, we present a multilingual ASR interface designed for the trilingual community of Kazakhstan. We evaluate the system's usability in an ASR-based typing task in three languages: Kazakh, Russian, and English. The system is deployed as a web application that integrates OpenAI's Whisper large-v3-turbo model fine-tuned by the Institute of Smart Systems and Artificial Intelligence (ISSAI). The resulting model, issai/whisper-turbo, demonstrates accurate speech recognition in Kazakh while maintaining high performance in Russian, English, and Turkish. Fine-tuning was performed using the Common Voice 12.0 dataset for Russian and English [11], the Kazakh Speech Corpus 2 (KSC2) [12], and the Turkish Speech Corpus (TSC) [13].

Quantitative evaluation shows that issai/whisper-turbo achieves a word error rate (WER) of 8.84% on Kazakh, a substantial improvement over the OpenAI Whisper baseline (21.55%). For English, the model achieves 5.82% WER vs. 5.15% baseline, and for Russian 6.15% WER vs. 5.89% baseline. These results highlight that our fine-tuning significantly enhances Kazakh recognition while preserving strong performance for high-resource languages, validating the model's effectiveness for real-world multilingual usage.

Aim of the study

We created a web application that simulates speech typing and editing processes. To evaluate how fast people could type via speech in three languages, Kazakh, Russian, and English, we designed a user study with 38 participants. During the user study, participants were asked to read aloud texts in three languages to create the ASR-based text transcriptions. Users could also make edits to the transcribed texts using the computer keyboard. During the experimental study, we explored the usability of the presented speech-based typing interface and evaluated the cognitive load of participants after using the system in each of the three languages. The aim of this study was to evaluate the performance and usability of a multilingual ASR interface for Kazakh, Russian, and English, with the hypothesis that speech-based text entry can be performed with comparable efficiency, usability, and cognitive load across the three languages

The rest of the paper is structured as follows: In the next section of the paper, we present the literature review on how ASR systems are integrated in various spheres of human-technology interaction and communication. We also provide an overview of the background research prior to the development of ASR systems for the Kazakh language. Then we present the methodology part of the paper with an overview of the user study design, the method used for the data collection, and analysis. This is followed by the part of the paper where we present the results of the user study and discussion. The paper is concluded in the final part of the paper.

Literature Review

Current trends in language models show that they are becoming increasingly multimodal and multilingual, meaning that interaction with LLMs via text and typing is extended by other modalities and communication patterns [6]. According to Fathullah et al. [9], interaction with LLMs purely via text may be limited due to the wide range of information structures that are difficult to capture in text but are naturally encoded in voice and visual inputs. For example, voice inputs to LLMs could provide information on speaker emotions, while images present contextual environmental information, making communication with LLMs faster and more efficient. Adhikary et al. [14] claim that speech-based interactions outperform typing when users are moving or multitasking. According to the authors, speech lets users focus more on what they want to say rather than how to type it, reducing human mental and physical workloads. Fig. 1 compares three input paradigms for communicating with LLMs. The traditional keyboard-only interface is a low-bandwidth channel which is slow and unnatural. Moreover, it comes with

substantial information loss since voice-based and visual cues are not transmitted. The hybrid approach proposed in this study uses voice to generate text and the keyboard for editing. From our experiments it approximately doubles input speed and partially improves naturalness, yet it remains constrained by a text-only bottleneck that discards prosodic and visual information. Future multimodal systems accept high-bandwidth, parallel streams (speech, text, video), reducing information loss and enabling LLMs to form more holistic, context-rich interpretations, thereby supporting faster and more natural human-AI interaction [15].

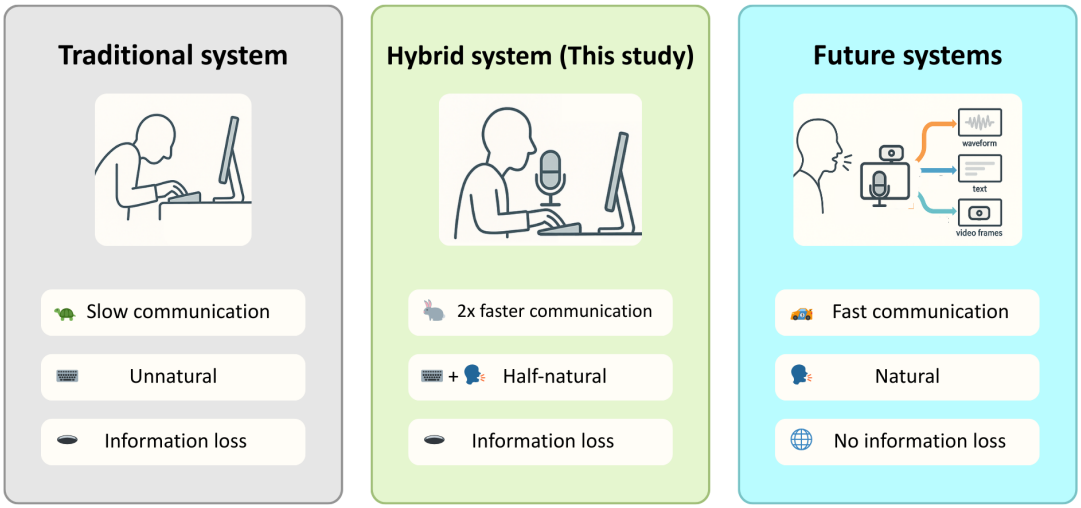

Figure 1. Comparison of Traditional, Hybrid (This study), and Future Human-AI Interaction Systems

There is a growing interest in multilingual ASR systems in bilingual and trilingual countries. Multilingual ASRs aim to preserve and enhance the practical use of native languages during human AI interaction and communication with LLMs. In [16],[17], the authors present multilingual ASR systems for Dutch-Frisian, Arabic-English, and Arabic-Malay languages. A speech emotion recognition system that can recognize emotional context for different languages is presented in [18]. Authors discuss how emotional cues can be understood differently depending on the language and culture. Overall, multilingual ASR systems enhanced with emotion recognition modality could be the next step for a multimodal communication framework with LLMs and AI agents.

There are also many works exploring speech-based technologies in education. For example, in [19], [20], the authors provide an overview of AI teaching assistants in online education. Kim et al. [21] in their work show that students view AI assistants as technically helpful, while limited emotionally. A meta-analysis [19] shows that learners gain more when chatbots offer quick, personalised feedback. In [22], authors discuss the effectiveness of AI chatbots in language practice. A review of 32 chatbot systems for English language learners' practice speaking and listening is presented in [23].

Early development of speech-based systems for the Kazakh language faced significant challenges due to a lack of linguistic and technological resources. In recent years, foundational datasets for Kazakh have been created to support the advancement of ASR systems, including large-scale speech corpora composed of transcribed audio from diverse speakers and sources such as media broadcasts and online content [12]. Additionally, publicly available resources have been developed for other NLP tasks, such as sentiment analysis, question answering, machine translation, and emotional TTS synthesis. These datasets have played a critical role in enabling research and development in Kazakh language technologies and continue to support progress in AI-driven language applications.

Methods and Materials

System Description

Commercial voice-based text entry applications involve steps such as speaking, viewing the transcribed text, and then editing the text. Many systems share common design principles, such as a button to start recording, displaying the transcribed text in real time, punctuation support, and an after-dictation editing flow. Correction workflows also overlap with manual edits and voice-based commands. In our application, transcribed text is highlighted in yellow where a mismatch occurred between the original and transcribed texts, helping users to quickly navigate through the text during editing (see Fig. 2). This way, users could manually correct any misrecognized words in the transcribed text.

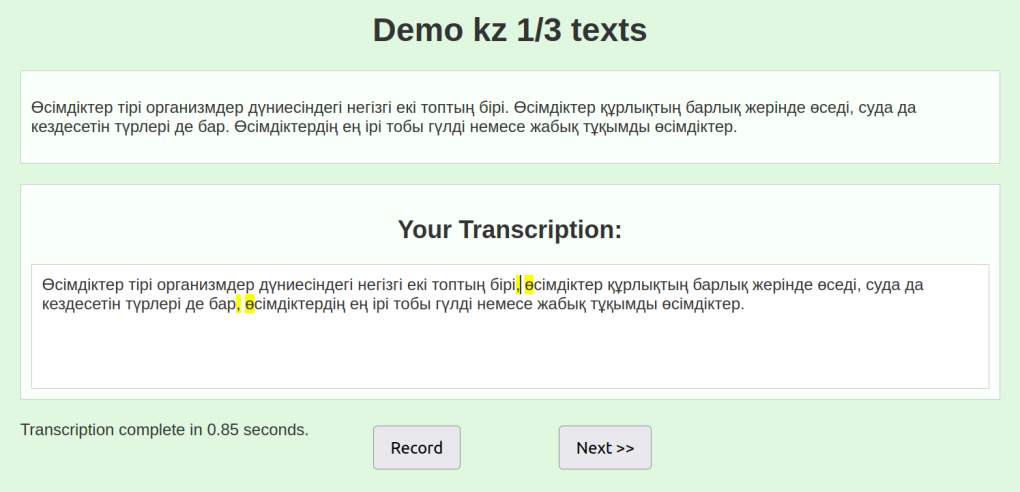

Figure 2. Editable Transcription Interface with Real-Time Error Highlighting

Our web application consists of five pages (registration, experiments, break, results, end). The interface presents the text to be read, provides a voice recording button using the MediaRecorder API, shows live ASR transcription, and highlights character-level mismatches in yellow. Users can correct errors before moving forward using a control button that advances the experiment. Front-end logic handles recording, rendering, and editing, while a Python backend executes ASR via the Flask framework.

The web application performs data logging of information presented in Table 1 for data collection and further statistical analysis.

Table 1. Variables and their descriptions stored for data analysis

Variable	**Description**
P	Presented text (string).
P_words	Number of words in P.
S	Returned text by ASR (string).

S_words	Number of words in S.
T	Transcribed text after user editing (string).
T_words	Number of words in T.
Time_talking (seconds)	Time elapsed from pressing the Record button until pressing Stop.
Time_asr (seconds)	Time from pressing Stop until the ASR result is received.
Time_edit (seconds)	Time from the first keyboard press during editing until the “Next” button is pressed.
Time_server (seconds)	Computed as Time_talking + Time_asr.
Time_total (seconds)	Computed as Time_talking + Time_asr + Time_edit.
WPM_asr	Calculated as S_words divided by (Time_asr/60).
WPM_server	Calculated as S_words divided by (Time_server/60).
WPM_user	Calculated as T_words divided by (Time_total/60).
CPM_server	Calculated as Number_of_characters divided by (Time_server/60).
CPM_user	Calculated as Number_of_characters divided by (Time_total/60).
CER_asr (%)	Used the provided formula (see Methods section) to compute the character error rate between S and P, then multiply by 100.
CER_user (%)	Compute the CER between T and P, then multiply by 100.
WER_asr (%)	Used the provided formula (see Methods section) to compute the word error rate between S and P, then multiply by 100.

WER_user (%)	Compute the WER between T and P, then multiply by 100.
Backspaces	Number of backspaces recorded during text editing in the current trial.

For audio-to-text transcription, we have used OpenAI's Whisper large-v3-turbo model, fine-tuned by ISSAI to achieve high recognition quality in Kazakh, Russian, English, and Turkish. The model (issai/whisper-turbo) is hosted on Hugging Face and is accessible by request rather than publicly downloadable. Fine-tuning was performed on 8× NVIDIA A100 GPUs over 7 epochs (learning rate 5×10^{-7}, batch size 16) using the following corpora: Common Voice 12.0 from Mozilla [11] (Kazakh: 3.8 h; Russian: 291 h; Turkish: 134 h; English: 3758 h), Kazakh Speech Corpus 2 (KSC2) [12] (1096 h), and Turkish Speech Corpus (TSC) [13] (218 h). Whisper large-v3-turbo was selected as an optimal trade-off between speed and multilingual recognition accuracy for real-time usability of the speech-based typing system. Researchers who require access to the fine-tuned weights or the fine-tuning recipe may contact us at issai@nu.edu.kz. In addition, to support reproducibility and facilitate future research, we publicly release the system implementation at: https://github.com/Nurikman/ASR_interface_ISSAI.

Experimental Procedure

We conducted a user study with 38 participants (17 female, 21 male; age range 20–37, M = 26.89, SD = 5.81) from the Nazarbayev University (NU) community in Astana, Kazakhstan. Participants were recruited among students, researchers, faculty, and staff. Ethics approval was obtained from NU's Institutional Research Ethics Committee, and all participants provided informed consent. Participants were randomly assigned to three groups (Group A: n=12; Group B: n=13; Group C: n=13), each performing speech-based typing tasks in different language orders: English-Kazakh-Russian (A), Kazakh-Russian-English (B), and Russian-English-Kazakh (C). At registration, demographic data (age, gender, education, occupation) were collected, and users were assigned a participant ID and task order. The speech-based typing interface (Fig. 2) displayed a short passage for reading aloud, an editable ASR-generated transcript, and controls to start/stop recording and move to the next passage. Each participant read 27 short texts (3 passages per language, split into 3 segments). Audio was recorded using the MediaRecorder API, chunked into 25-second segments, and transcribed via the Whisper-Turbo ASR model on Hugging Face. The final transcriptions and inference times were returned to the client.

Methods and Materials

We assessed participants' cognitive load and system usability during speech-based typing tasks in Kazakh, Russian, and English. After each language speech-based typing task, participants completed the paper-based Raw NASA Task Load Index (NASA-RTLX) to evaluate cognitive load across six dimensions: mental, physical, and temporal demand, performance, effort, and frustration [24]. Usability was measured using the System Usability Scale (SUS) [25], also administered after each language speech-based typing task. Participants took short breaks between language tasks and completed a demographic survey at the end, reflecting on language proficiency, speaking habits, and comfort with the system.

For the statistical analysis, we used means (M), standard deviations (SD), and Shapiro-Wilk tests for normality. One-Way ANOVA and post-hoc Tukey HSD tests were applied to evaluate differences in age, gender distribution, typing speed (i.e., words per minute - WPM, characters

per minute - CPM), accuracy (i.e., word error rate - WER, character error rate - CER), cognitive load (NASA-RTLX), and usability (SUS) across the three languages.

Task-Level Measures

For each trial, we recorded both text outputs and timing features of the interaction. The interface displayed a reference prompt P (the "presented text"), and the participant was instructed to read it aloud. The ASR system produced a raw transcription S ("system output"), after which the participant was allowed to edit this transcription to obtain a final corrected version T ("user-edited text") We denote by $P_{words}, S_{words}, T_{words}$ the number of word tokens in $P, S, and\ T$, respectively.

We also logged timing signals for each phase of the interaction. $Time_{talking}\ (s)$ is the duration from when the participant pressed the Record button until they pressed Stop (i.e., active speech production). $Time_{asr}\ (s)$ is the duration from Stop until the ASR hypothesis S was returned to the interface (model inference time). $Time_{edit}\ (s)$ is the duration from the first manual keystroke in the editable transcript until the participant confirmed the transcription by pressing "Next." We define $Time_{server}\ (s)$ as the sum of speech and inference time,

$$Time_{server} = Time_{talking} + Time_{asr}, \quad (1)$$

and $Time_{total}\ (s)$ as the full end-to-end interaction time including manual correction,

$$Time_{total} = Time_{talking} + Time_{asr} + Time_{edit}, \quad (2)$$

Using these quantities, we computed multiple throughput measures. $\boldsymbol{WPM_{asr}}$ is defined as

$$WPM_{asr} = \frac{S_{words}}{Time_{asr}\ /\ 60}, \quad (3)$$

capturing the instantaneous decoding rate of the ASR system alone. $\boldsymbol{WPM_{server}}$ reflects effective speech-to-text throughput including speaking and inference,

$$WPM_{server} = \frac{S_{words}}{Time_{server}\ /\ 60}, \quad (4)$$

and $\boldsymbol{WPM_{user}}$ reflects the end-to-end effective text entry rate experienced by the participant after corrections,

$$WPM_{user} = \frac{T_{words}}{Time_{total}\ /\ 60}, \quad (5)$$

In parallel, we computed $\boldsymbol{CPM_{server}}$ and $\boldsymbol{CPM_{user}}$ as the number of produced characters (in the ASR hypothesis for CPM_{server} and in the final corrected text for CPM_{user}) divided by $Time_{server}\ /\ 60$ and $Time_{total}\ /\ 60$, respectively, yielding character-level entry speed in characters per minute.

Transcription accuracy was quantified at both the system and user levels. $\boldsymbol{WER_{asr}}\ (\%)$ and $\boldsymbol{CER_{asr}}\ (\%)$ are the word error rate and character error rate, respectively, between S and P, multiplied by 100. $\boldsymbol{WER_{user}}\ (\%)$ and $\boldsymbol{CER_{user}}\ (\%)$ are the same metrics computed between T and P, multiplied by 100. WER and CER follow standard edit-distance definitions, i.e.,

$$WER = \frac{S + D + I}{N} \times 100\%, CER = \frac{S_c + D_c + I_c}{N_c} \times 100\%, \quad (6), (7)$$

where S, D, and I are word-level substitutions, deletions, and insertions with respect to the reference, and N is the total number of reference words; $S_c, D_c, I_c,$ and N_c are the analogous quantities at the character level. Finally, we recorded $Backspaces$, defined as the number of backspace keypresses during the edit phase of that trial. This serves as a proxy for manual correction effort.

Results

The study involved 38 participants with an average age of 26.89 ± 5.81 years, randomly assigned to three groups (A, B, C) to perform speech-based typing tasks in Kazakh, Russian, and English in varied orders. One-way ANOVA showed no significant differences in age ($F(2,35)=0.85$, $p=0.44$) or gender distribution ($F(2,35)=2.53$, $p=0.09$) across groups.

Typing performance, measured in WPM, differed significantly across languages. System WPM (typing without editing) showed significant variation ($F(2,111)=101.61$, $p<0.001$), as did user WPM (including editing time) ($F(2,111)=51.01$, $p<0.001$). According to Tukey HSD, Kazakh WPM values were significantly lower than both Russian and English. Mean system WPM was 90.87 ± 16.48 for Kazakh, 130.07 ± 16.41 for Russian, and 144.56 ± 18.01 for English. Corresponding user WPM values were 40.31 ± 16.06, 68.81 ± 17.4, and 76.98 ± 16.37, respectively (Fig. 3a–b).

Typing performance, measured in CPM, also showed significant differences across languages. System CPM was significantly different across groups ($F(2,111)=7.3$, $p<0.001$), particularly between Kazakh and Russian. User CPM differences were significant for Kazakh-Russian and Kazakh-English pairs ($F(2,111)=11.5$, $p<0.001$). Mean system CPM values were 667.96 ± 118.02 (Kazakh), 759.49 ± 99.92 (Russian), and 716.94 ± 94.14 (English), while user CPM values were 264.55 ± 98.25, 362.22 ± 96.93, and 335.06 ± 78.32, respectively (Fig. 3c–d).

Figure 4a-d presents speech-based typing accuracy metrics, including system and user WER and CER. System WER was highest for Kazakh (26.32 ± 8.34), followed by Russian (11.86 ± 2.38) and English (9.65 ± 2.52) (Fig. 4a). User WER followed a similar trend: Kazakh (3.40 ± 6.33), Russian (1.29 ± 6.25), and English (0.95 ± 2.11) (Fig. 4b). One-way ANOVA showed significant differences for system WER ($F(2,111)=114.63$, $p<0.001$) and user WER ($F(2,111)=3.88$, $p=0.023$). Tukey HSD confirmed significant differences between Kazakh-Russian and Kazakh-English for system WER, and between Kazakh-English for user WER.

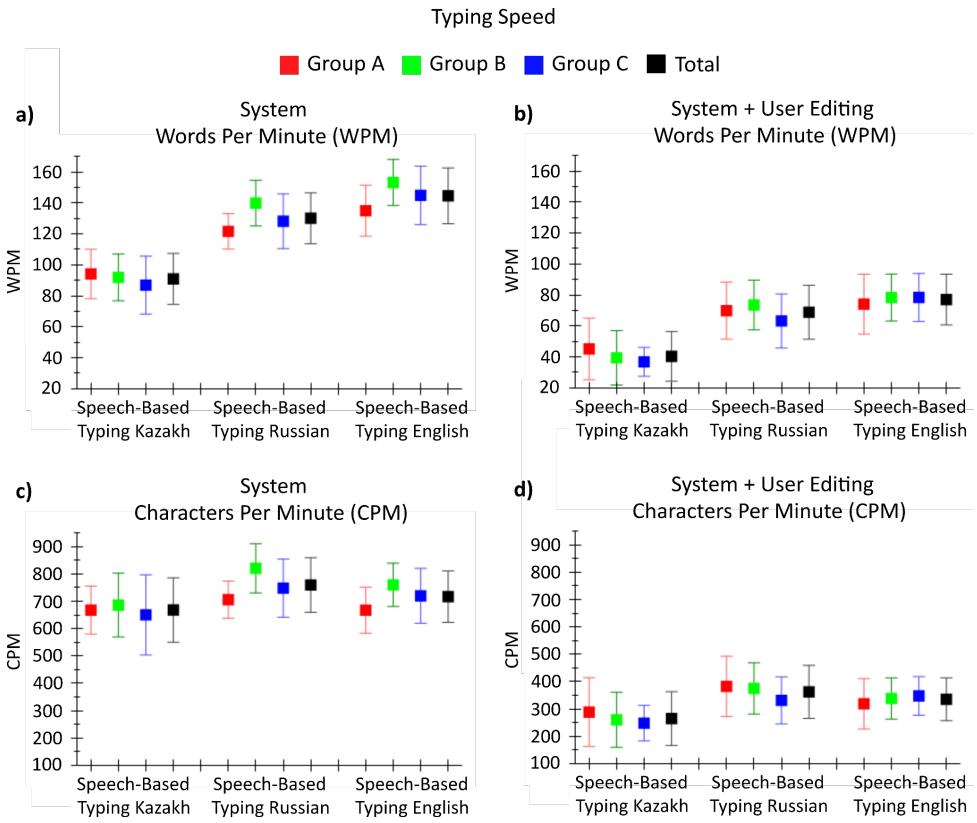

Figure 3. Typing speed: a) system WPM not edit case, b) user WPM edit case, c) system CPM not edit case, and d) user CPM edit case

System CER was highest for Kazakh (5.28 ± 2.7), compared to English (3.85 ± 1.29) and Russian (3.80 ± 1.39) (Fig. 4c). User CER values were lower overall, Kazakh (0.52 ± 0.98), Russian (0.40 ± 0.80), and English (0.33 ± 0.82) (Fig. 4d). One-way ANOVA revealed a significant difference in system CER across languages ($F(2,111)=7.33$, $p=0.001$), with Tukey HSD indicating significant differences for Kazakh-Russian and Kazakh-English pairs. However, user CER differences were not statistically significant ($F(2,111)=0.43$, $p=0.65$).

Figure 5a-f presents cognitive load ratings across six NASA-TLX dimensions (i.e., mental, physical, and temporal demand; effort; frustration; and perceived performance) for Kazakh, Russian, and English. One-way ANOVA revealed no statistically significant differences across languages for any dimension. While Kazakh showed slightly higher average scores in mental demand (29 ± 30.26), effort (34 ± 27.43), and frustration (17 ± 22.20), these differences were not statistically significant. Performance ratings were comparable across languages: 66 ± 28.13 (Kazakh), 73 ± 22.30 (Russian), and 69 ± 21.62 (English).

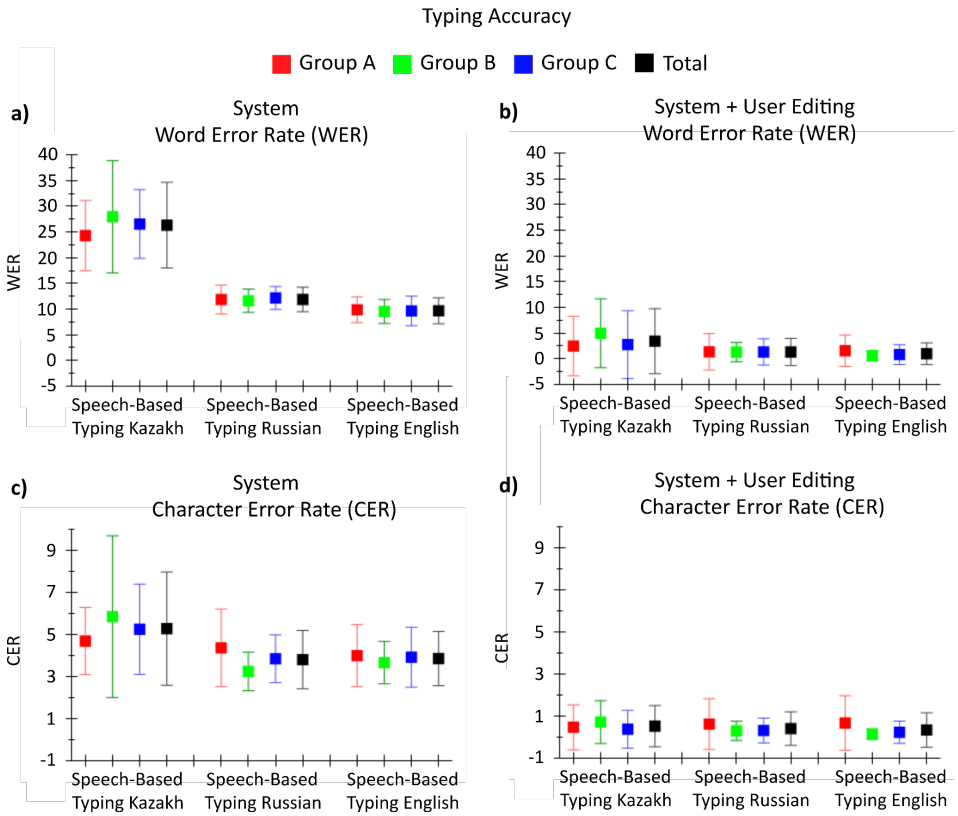

Figure 4. Typing accuracy: a) system WER not edit case, b) user WER edit case, c) system CER not edit case, and d) user CER edit case

As shown in Fig. 6, usability ratings for speech-based typing were high across all languages: 86.58 ± 14.29 (Kazakh), 86.91 ± 10.97 (Russian), and 88.95 ± 9.54 (English). One-way ANOVA revealed no significant differences between languages ($F(2,111)=0.45$, $p=0.64$), indicating similarly high usability (>80%) among the Kazakhstani population.

Demographic results are summarized in Fig. 7a-d. Most participants rated their typing experience in all three languages as "very comfortable" or "somewhat comfortable," with Kazakh receiving slightly higher "somewhat comfortable" ratings (34.21%) and a higher "neutral" response rate (10.53%) compared to Russian and English (5.26%) (Fig. 7a). In terms of language background (Fig. 7b), 57.89% reported Kazakh as their first language and 42.11% Russian; none reported English. Fluency was highest in Russian (73.68%), followed by English (13.16%) and Kazakh. Intermediate proficiency was lowest in Russian, being 5-6 times lower than in Kazakh or English.

Daily language use (Fig. 7c) showed limited use of English, with 65.79% using it 0–25% of the time. Russian and Kazakh were used more frequently: 55.26% and 60.53% reported using them 26–50% of the time, respectively. Only 2.63% reported using Kazakh 76–100% of the time; no participants reported this level of use for Russian or English. Educational levels (Fig. 7d) were predominantly at the Master's level (44.74%), followed by graduate students (21.05%), undergraduates (18.42%), and smaller proportions of Bachelor's and PhD holders (7.89% each).

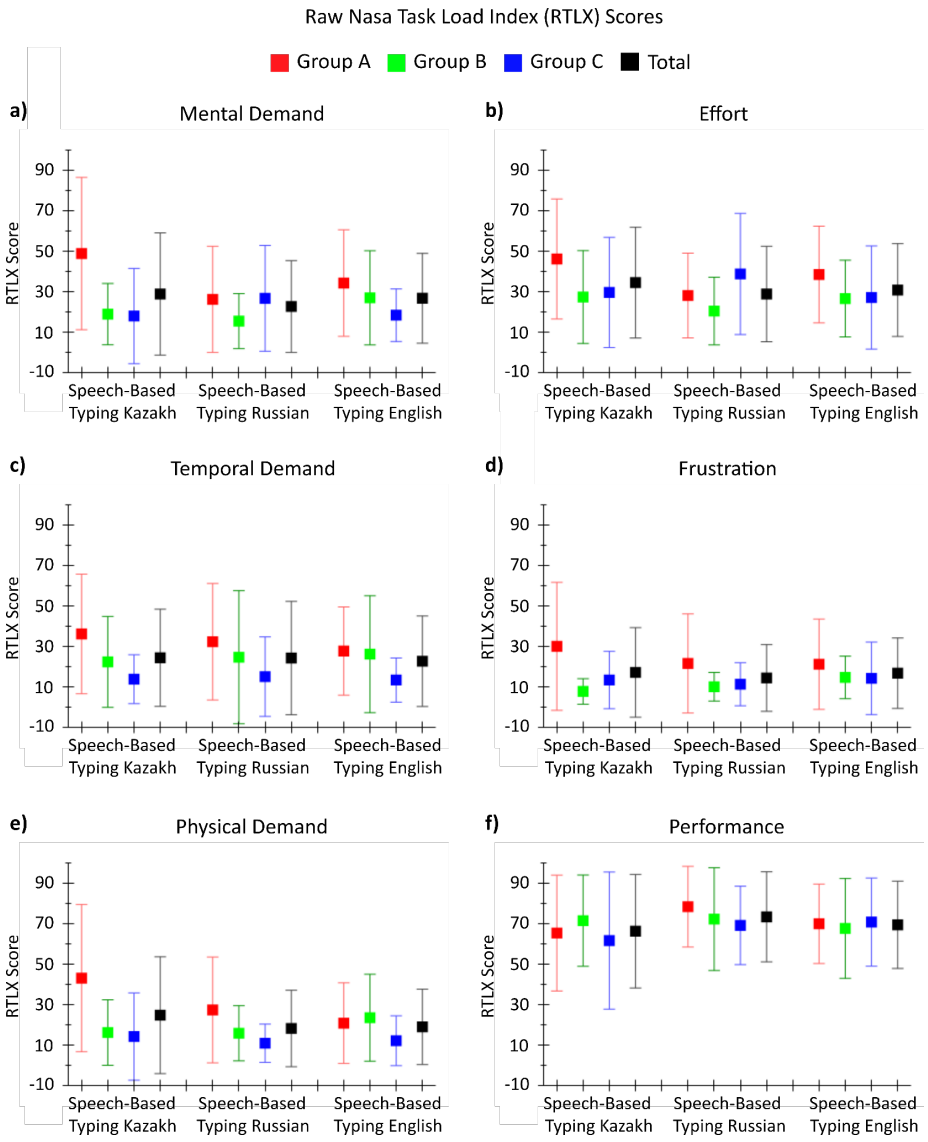

Figure 5. RTLX ratings: a) mental demand, b) effort, c) temporal demand, d) frustration, e) physical demand, and f) performance

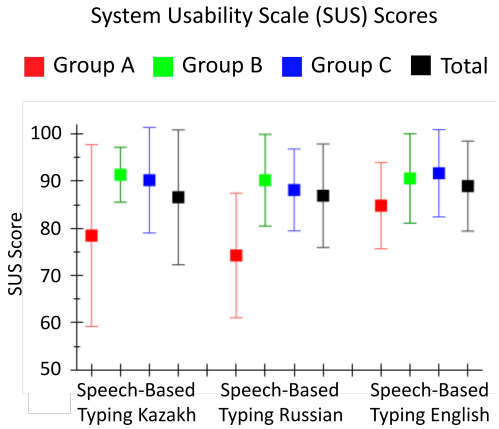

Figure 6. SUS ratings for Kazakh, Russian, and English language speech-based typing

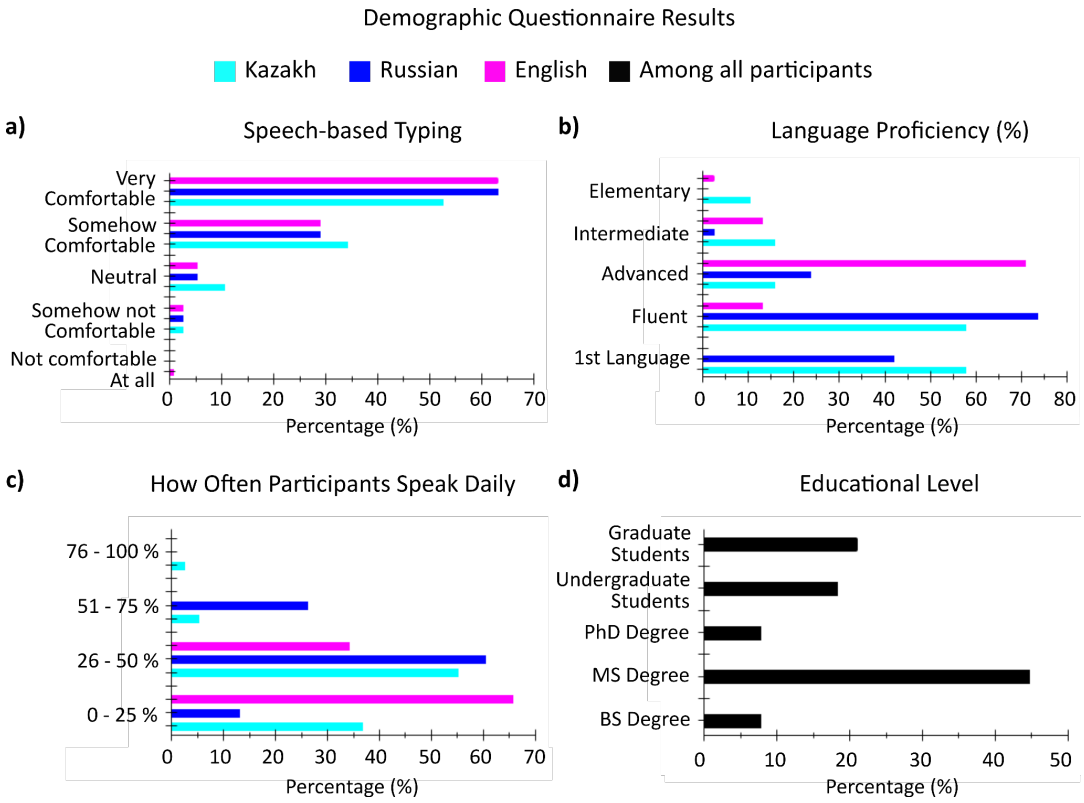

Figure 7. Demographic questionnaire results expressed in percentage: a) speech-based typing experience, b) language proficiency in three languages, c) how often participants type in each of the three languages during the day, and d) educational level

Discussion

In the era of AI, new interaction and communication modalities with AI and LLMs can further impact the usability of languages. In general, in our times of AI and technological development, speech has become an essential modality of embodiment, communication, and interaction between humans and AI-enhanced social robots, conversational agents, and voice assistants [9]. While one set of languages becomes more popular, accessible, and tech-supported, the others become less popular. This could disadvantage people who speak native languages with less support from the global tech community. Individuals who speak native languages may be hindered from integrating AI technologies and innovations into their daily lives due to the language barrier.

In our previous study [26], we found that keyboard typing in Kazakh was slower and less accurate than in Russian and English, with higher cognitive load. Kazakh typing speed averaged 23.04 ± 6.59 WPM, which was 1.27 times and 1.41 times lower than Russian (29.15 ± 7.58) and English (32.53 ± 8.31), respectively. Kazakh typing also had a higher character error rate (CER = 5.73 ± 5.00) compared to Russian (5.24 ± 5.27) and English (3.22 ± 3.59). Participants reported lower comfort and frequency of use when typing in Kazakh. In the current study, speech-based typing outperformed keyboard input across all languages. For Kazakh, participants achieved 90.87 ± 16.48 WPM (no edits) and 40.31 ± 16.06 WPM (with edits), which is 3.94× and 1.75× faster than keyboard typing. For Russian, the speed was 130.07 ± 16.41 WPM (no edits) and 68.81 ± 17.4 WPM (with edits), i.e., 4.48 times and 2.36 times faster. In English, users reached 144.56 ± 18.01 WPM and 76.98 ± 16.37 WPM, showing 4.44 times and 2.37 times improvements over keyboard input. Despite the speed advantage, Kazakh showed significantly higher word error rates (WER), being 2.22 vs. 2.64 times higher than Russian and 2.73 vs. 3.58 times higher than English (no edits/with edits). CER was also elevated for Kazakh, 1.39 vs. 1.30 times higher than Russian and 1.37 vs. 1.58 times higher than English. Nonetheless, speech-

based typing resulted in lower reported effort and higher perceived performance compared to keyboard input, particularly in Kazakh.

Our study has certain limitations concerning the diversity of the participant sample: 38 participants were recruited from NU. In the future these limitations can be addressed by involving more participants of various domains from different regions. Further works might be focused on expanding the Kazakh speech corpus with more diverse data, advancing research in emotional speech recognition, and deploying the system in real-world human-AI interaction scenarios to assess its practical applicability.

Overall, communication with LLMs in the Kazakh language via keyboard typing could become much less effective over time. ASR offers a faster alternative. Prior studies [26] have shown that voice input on smartphones allows significantly faster text entry, nearly 3 times faster than typing, when users are certain of what they want to say. However, users often prefer text in uncertain situations due to easier message editing and greater comfort. Similar findings were reported by Ruan et al. [27], where speech input was 2.93 and 2.87 times faster than typing in English and Mandarin, respectively.

Despite speed advantages, speech input poses challenges: editing is time-consuming, and users may feel uncomfortable speaking aloud in public or noisy settings. Therefore, designing audio-based interfaces requires attention to both technical performance and human factors. Deep learning systems like Deep Speech 2 [28] show promise in handling spontaneous speech, accents, and background noise, making speech-based interaction with LLMs more feasible across languages, including Kazakh.

ASR technology plays a key role in speaker identification and authentication, enabling speaker-independent and multi-speaker recognition systems. Target-speaker ASR, which identifies and responds to a specific user's voice, presents a promising direction for future research in the Kazakh language. Investigating its development and evaluating usability in terms of cognitive load and performance could be particularly valuable.

Conclusion

In this paper, we presented a multilingual ASR system and applied it in a speech-based typing user study in three languages: Kazakh, Russian, and English. The system was able to transcribe read speech into written text and was evaluated with 38 participants (17 female and 21 male), including students, researchers, faculty, and staff from NU in Astana, Kazakhstan. We investigated typing speed with and without editing the ASR-transcribed text, measured in WPM and CPM. In addition, we assessed participants' cognitive load and usability through NASA-RTLX and SUS.

The results indicate that users could type via speech in Kazakh at 90.87±16.48 WPM without editing and 40.31±16.06 WPM with editing, which is 3.94 and 1.75 times faster than keyboard typing speeds reported in our previous study. Russian and English also showed higher results, with speech-based typing being 4.48 and 2.36 times faster in Russian and 4.44 and 2.37 times faster in English compared to keyboard typing. Across all three languages, participants reported low cognitive load and high usability, with SUS scores above 80%. These findings suggest that the developed system can serve as a foundation for practical voice interfaces and educational applications in the Kazakh language and can be scaled to real-world multilingual services requiring fast and accessible text entry. As part of this work, we additionally employed ISSAI's fine-tuned Whisper model (issai/whisper-turbo) to support the system. This model significantly improves recognition accuracy in Kazakh (8.84% WER vs. 21.55% baseline) while maintaining comparable performance in English (5.82% vs. 5.15 baseline) and Russian (6.15% vs. 5.89 baseline). While not the central focus of this study, these improvements demonstrate the feasibility of enhancing ASR for low-resource languages without sacrificing performance in high-resource ones.

Acknowledgements

This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24993001).

References

[1] Yu, D., & Deng, L. (2016). Automatic speech recognition (Vol. 1). *Springer.*

[2] Bai, Z., & Zhang, X. L. (2021). Speaker recognition based on deep learning: An overview. *Neural Networks, 140,* 65-99. https://doi.org/10.1016/j.neunet.2021.03.004

[3] Ning, Y., He, S., Wu, Z., Xing, C., & Zhang, L. J. (2019). A review of deep learning-based speech synthesis. *Applied Sciences, 9* (19), 4050. https://doi.org/10.3390/app9194050

[4] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). A comprehensive overview of large language models. ACM Transactions on Intelligent Systems and Technology, 16(5), 1–72. https://doi.org/10.1145/3744746

[5] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877-1901.

[6] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models [Preprint]. arXiv. https://arxiv.org/abs/2108.07258

[7] Luz, S., Masoodian, M., Rogers, B., & Deering, C. (2008, December). Interface design strategies for computer-assisted speech transcription. In *Proceedings of the 20th australasian conference on computer-human interaction: designing for habitus and habitat* (pp. 203-210). https://doi.org/10.1145/1517744.1517812

[8] Vashistha, A., Sethi, P., & Anderson, R. (2017, May). Respeak: A voice-based, crowd-powered speech transcription system. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 1855-1866). http://dx.doi.org/10.1145/3025453.3025640

[9] Fathullah, Y., Wu, C., Lakomkin, E., Jia, J., Shangguan, Y., Li, K., ... & Seltzer, M. (2024, April). Prompting large language models with speech recognition abilities. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 13351-13355). IEEE. https://doi.org/10.1109/ICASSP48485.2024.10447605

[10] Yang, C. H. H., Gu, Y., Liu, Y. C., Ghosh, S., Bulyko, I., & Stolcke, A. (2023, December). Generative speech recognition error correction with large language models and task-activating prompting. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 1-8). IEEE. https://doi.org/10.1109/ASRU57964.2023.10389673

[11] Mozilla Foundation. (2022). Common Voice (Version 12.0) [Data set]. https://huggingface.co/datasets/mozilla-foundation/common_voice_12_0

[12] Mussakhojayeva, S., Khassanov, Y., & Varol, H. A. (2022). KSC2: An industrial-scale open-source Kazakh speech corpus. Proceedings of Interspeech 2022, 1367–1371. https://doi.org/10.21437/Interspeech.2022-421

[13] Mussakhojayeva, S., Dauletbek, K., Yeshpanov, R., & Varol, H. A. (2023). Multilingual Speech Recognition for Turkic Languages. Information, 14(2), 74. https://doi.org/10.3390/info14020074

[14] Adhikary, J., & Vertanen, K. (2021). Text entry in virtual environments using speech and a midair keyboard. IEEE Transactions on Visualization and Computer Graphics, 27(5), 2648–2658. https://doi.org/10.1109/TVCG.2021.3067776

[15] Schneider, J. (2020). Humans learn too: Better human-AI interaction using optimized human inputs [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2009.09266

[16] Yılmaz, E., van den Heuvel, H., & Van Leeuwen, D. (2016). Investigating bilingual deep neural networks for automatic recognition of code-switching frisian speech. *Procedia Computer Science, 81,* 159-166. https://doi.org/10.1016/j.procs.2016.04.044
[17] Abushariah, A. A., Ting, H. N., Mustafa, M. B. P., Khairuddin, A. S. M., Abushariah, M. A., & Tan, T. P. (2022). Bilingual automatic speech recognition: A review, taxonomy and open challenges. *IEEE Access, 11,* 5944-5954. https://doi.org/10.1109/ACCESS.2022.3218684
[18] Heracleous, P., & Yoneyama, A. (2019). A comprehensive study on bilingual and multilingual speech emotion recognition using a two-pass classification scheme. *PloS one, 14* (8), e0220386. https://doi.org/10.1371/journal.pone.0220386
[19] Wu, R., & Yu, Z. (2024). Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis. *British Journal of Educational Technology, 55* (1), 10-33. https://doi.org/10.1111/bjet.13334
[20] Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education, 20* (1), 56. https://doi.org/10.1186/s41239-023-00426-1
[21] Kim, J., Merrill, K., Xu, K., & Sellnow, D. D. (2020). My teacher is a machine: Understanding students' perceptions of AI teaching assistants in online education. *International Journal of Human–Computer Interaction, 36* (20), 1902-1911. https://doi.org/10.1080/10447318.2020.1801227
[22] Belda-Medina, J., & Calvo-Ferrer, J. R. (2022). Using chatbots as AI conversational partners in language learning. Applied Sciences, 12(17), 8427.
[23] Jeon, J., Lee, S., & Choi, S. (2023). A systematic review of research on speech-recognition chatbots for language learning: Implications for future directions in the era of large language models. *Interactive Learning Environments, 32* (8), https://doi.org/10.1080/10494820.2023.2204343
[24] National Aeronautics and Space Administration. (n.d.). NASA-TLX: Paper/pencil version. https://humansystems.arc.nasa.gov/groups/tlx/tlxpaperpencil.php
[25] Brooke, J. (1.996). SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, I. L. McClelland, & B. Weerdmeester (Eds.), Usability evaluation in industry (pp. 189–194). Taylor & Francis.
[26] Makhataeva, Z., & Varol, H. A. (2025). Evaluation of Typing Speed, User Experience, and Cognitive Load Across Kazakh, Russian, and English Languages among Kazakhstani Users. J*ournal of Educational Sciences, 83* (2), 2520-2634. https://doi.org/10.26577/JES202583214
[27] Ruan, S., Wobbrock, J. O., Liou, K., Ng, A., & Landay, J. A. (2018). Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1(4), 1-23. https://doi.org/10.1145/3161187
[28] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Zhu, Z. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. In M. F. Balcan & K. Q. Weinberger (Eds.), Proceedings of the 33rd International Conference on Machine Learning (Vol. 48, pp. 173–182). PMLR.