

FTAMP 50.41.25

Н.М. Жунисов¹ – негізгі автор, | ©
А.Б. Абен², М. Хиниязов³



¹PhD, аға оқытушы, ²Докторант, ³Студент

ORCID

¹<https://orcid.org/0000-0001-6531-9408> ²<https://orcid.org/0000-0001-8534-3288>



^{1,2,3}Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті,



Түркістан, Қазақстан



¹nurseit.zhunissov@ayu.edu.kz

<https://doi.org/10.55956/KLLF9699>

НАҚТЫ УАҚЫТТАҒЫ ДАУЫСТЫҚ КӨМЕКШІЛЕРГЕ АРНАЛҒАН ТӨМЕН КІДІРІСТІ АҒЫНДЫҚ ТРАНСФОРМАТОР МОДЕЛІ

Аңдатпа. Бұл мақалада нақты уақыттағы дауыстық көмекшілерге арналған төмен кідірістегі автоматты сөйлеуді тану ASR (Automatic Speech Recognition), жүйелерін, атап айтқанда қазақ тілі сияқты ресурсы төмен тілдерге арналған дауыстық көмекші (JARVIS) зерттеледі. Tiny Whisper және Streaming Transformer моделдері олардың төмен есептеу құны мен жоғары дәлдігі үшін бағаланып, AlphaCephei платформасының Vosk моделі мен Google компаниясының Gemini AI технологиясымен жұмыс істейтін қазақ дауыстық көмекшісінің өнімділігі талданды. Зерттеуде LibriSpeech, Common Voice және арнайы жиналған 200 сағаттық қазақ тіліндегі сөйлеу деректер жинағы пайдаланылды. Тәжірибелер Tiny Whisper шеткі құрылғыларда тиімді екенін, Streaming Transformer ағынға негізделген сценарийлерде аз кідірісті қамтамасыз ететінін және қазақша дауыс көмекшісі (JARVIS) цифрлық ортада тілдің қолжетімділігін арттыратынын көрсетеді. Ұсынылған гибриді модель нақты уақыттағы қолданбаларда жоғары дәлдік пен төмен кідіріс беру үшін осы технологиялардың күшті жақтарын біріктіреді. JARVIS браузерді іске қосу, дыбыс деңгейін және жарықтығын бақылау сияқты жүйелік тапсырмаларды орындайды және Gemini AI интеграциясы арқылы онлайн ақпаратты іздеуді қолдайды. Бұл зерттеуде кідірісі төмен ASR жүйелерінің практикалық қолданылуы және қазақ тілі сияқты ресурсы аз тілдердің технологиялық дамуы көрсетілген.

Тірек сөздер: төмен кідірістегі ASR, нақты уақыттағы дауыстық көмекші, кішкентай сыбырлау, ағындық трансформатор, қазақша дауыстық көмекші, Vosk, Gemini AI, JARVIS.



Жунисов, Н.М. Нақты уақыттағы дауыстық көмекшілерге арналған төмен кідірісті ағындық трансформатор моделі [Мәтін] / Н.М. Жунисов, А.Б. Абен, М. Хиниязов // Механика және технологиялар / Ғылыми журнал. – 2025. – №4(90). – Б.425-432. <https://doi.org/10.55956/KLLF9699>

Кіріспе. Заманауи технологиялардың қарқынды дамуы адам мен компьютердің өзара әрекеттесуінде жаңа парадигмаларға алып келеді. Дауыстық көмекшілер смартфондардан бастап смарт үй құрылғыларына дейінгі қолданбалардың кең ауқымында күнделікті өмірді жеңілдетеді. Дегенмен, қолданыстағы дауыстық көмекшілердің көпшілігі ағылшын, орыс немесе қытай сияқты жалпы тілдер үшін оңтайландырылғанымен, қазақ тілі сияқты ресурсы төмен тілдерге арналған шешімдер шектеулі. Бұл жұмыс

жалпы мақсаттағы, кідіріссіз автоматты сөйлеуді тану (ASR) жүйелерін (Tiny Whisper және Streaming Transformer) және қазақ тілі үшін арнайы әзірленген дауыстық көмекшіні (JARVIS) қарастырады. AlphaCephei платформасының Vosk моделін Google компаниясының Gemini AI технологиясымен біріктіру арқылы JARVIS браузерді ашу, дыбыс деңгейі мен жарықтығын басқару және дауыстық пәрмендер арқылы ақпаратты желіде іздеу сияқты жүйелік тапсырмаларды орындайды. Жобаның негізгі мақсаты – қазақ тілін цифрлық ортада дамыту, оның технологиялық қосымшаларда қолжетімділігін арттыру, пайдаланушыға ыңғайлы интерфейсті қамтамасыз ету. Әдебиеттердегі бар зерттеулер әдетте жоғары есептеу қабілеті бар бұлтқа негізделген жүйелерге бағытталғанымен, шеткі құрылғыларда және ресурсы төмен тілдерде төмен кідіріс шешімдеріне арналған мамандандырылған жүйелер шектеулі. Бұл мақала Tiny Whisper және Streaming Transformer үлгілерін біріктіретін гибриді тәсілді әзірлеу және қазақ дауыстық көмекшісінің (JARVIS) техникалық инфрақұрылымын шешу арқылы осы олқылықтарды толтыруға бағытталған. Зерттеудің бірегей үлестері мыналарды қамтиды:

1. Гибриді құрылымда Tiny Whisper, Streaming Transformer және Vosk үлгілерін біріктіру.

2. Қазақ тілі сияқты ресурсы аз тілде нақты уақыттағы дауыстық көмекшіні әзірлеу және оның жұмысын бағалау.

3. JARVIS жүйелік тапсырмаларының және ақпаратты іздеу мүмкіндіктерінің мәдени және технологиялық әсерлерін талдау.

Жұмыстың қалған бөлігі келесідей ұйымдастырылған: 2-бөлімде тиісті әдебиеттер жинақталған; 3-бөлімде ұсынылған әдістің егжей-тегжейлері келтірілген; 4-бөлімде эксперимент нәтижелері талқыланады; 5-бөлімде нәтижелер талданады және әдебиеттермен салыстырылады; және 6-бөлімде зерттеу қорытындылары мен болашақ зерттеу бағыттары жинақталған.

Әдебиеттерге шолу. Автоматты сөйлеуді тану (ASR) ұзақ уақыт бойы табиғи тілді өңдеу (NLP) және машиналық оқыту салаларында зерттелді. Дәстүрлі тәсілдер Жасырын Марков моделі (HMM) және Гаусс қоспасы моделі (GMM) негізіндегі жүйелерге негізделгенімен, терең оқытуға негізделген модельдер (мысалы, RNN, Transformer) соңғы жылдары айтарлықтай прогреске қол жеткізді [1]. OpenAI компаниясының Whisper моделі ауқымды деректер жиынында оқытылды және көптілді сөйлеуді тану үшін қуатты шешім ұсынады [2]. Дегенмен, Whisper толық масштабты нұсқалары жоғары есептеу талаптарына байланысты шеткі құрылғыларда қолдануға жарамсыз. Бұл модельдің жеңіл нұсқасы Tiny Whisper 39 миллион параметрі бар төмен ресурстарды тұтынуды ұсынады [3]. Streaming Transformer, керісінше, ағындағы сөйлеу деректерін өңдейді, бұл нақты уақыттағы қолданбаларда аз кідірісті қамтамасыз етеді [4].

Қазақ тілі сияқты ресурсы аз тілдерге арналған ASR жүйелерін зерттеу шектеулі. Ли және т.б. (2023) Streaming Transformer динамикалық деректер ағындарында тиімді, бірақ оқу деректер жинағына тәуелді екенін атап өтті [5]. AlphaCephei платформасының Vosk моделі ресурсы аз тілдер үшін ашық бастапқы шешімді ұсынады және офлайн режимінде жұмыс істеу мүмкіндігімен ерекшеленеді [6]. Google компаниясының Gemini AI моделі мультимодальды мүмкіндіктері мен көптілді қолдауы ақпарат іздеу тапсырмаларының қуатты құралы болып табылады [7]. Әдебиеттегі басқа зерттеулер гибриді ASR үлгілерінің әлеуетін аз кідіріс қолданбаларында көрсетті [8], бірақ ресурсы төмен тілдерге арналған арнайы шешімдер сирек кездеседі.

Қазақ тілінің дауыс көмекшісінің дамуында тілдің фонетикалық және грамматикалық ерекшеліктері (мысалы, дауысты дыбыс үндестігі, буын құрылымы) шешуші рөл атқарады. Джурафский мен Мартин (2020) ресурсы аз тілдер үшін сенімді үлгілерді жасау мұқият деректерді жинауды және модельді оқытуды қажет ететінін атап өтті [9]. Vosk-тың ашық бастапқы сипаты модельді қазақша сөйлеу деректерімен теңшеуге мүмкіндік береді. Gemini AI мультимодальды мүмкіндіктері дауыстық көмекшінің ақпарат іздеу функцияларын күшейтеді және мәдени контексте пайдаланушы тәжірибесін жақсартады [10]. Бұл технологияларды гибриді шенберде біріктіру және ресурсы төмен тілдер үшін оңтайландырылған нақты уақыттағы дауыстық көмекшілерді әзірлеуде әдебиетте олқылық бар. Бұл зерттеу осы олқылықтың орнын толтыруға бағытталған.

Зерттеу шарттары мен әдістері. Бұл бөлім ұсынылып отырған кідіріссіз ASR жүйесінің және қазақ дауыс көмекшісінің (JARVIS) дизайнын, пайдаланылған деректер жиынын, үлгі архитектураларын және эксперименттік орнатуды егжей-тегжейлі сипаттайды.

Деректер жинағы. Зерттеуде LibriSpeech [11], Common Voice [12] және арнайы жиналған 200 сағаттық қазақ тіліндегі сөйлеу деректер жинағы пайдаланылды. LibriSpeech 1000 сағаттық ағылшын аудиокітап жазбаларынан тұрады, ал Common Voice көп тілді сөйлеу деректерін береді. Қазақ деректер жинағы ана тілінде сөйлейтіндерден жиналды және әртүрлі екпіндерді, жас топтарын және шу жағдайларын (мысалы, көше, кеңсе, үй) ескере отырып дайындалды. Деректер жиынтығы келесі сипаттамаларға ие:

- Ұзақтығы: 200 сағат;
- Спикерлер саны: 150;
- Акцент әртүрлілігі: Алматы, Астана, Қарағанды және ауылдық жерлер;
- Алдын ала өңдеу: шуды азайту, қалыпқа келтіру және MFCC мүмкіндігін шығару.

Модельдік архитектуралар. Tiny Whisper – Whisper үлгісінің жеңіл нұсқасы және 39 миллион параметрді қамтиды. Модель шеткі құрылғылар үшін оңтайландырылған және трансформаторға негізделген кодтаушы-декодер архитектурасына негізделген. Сөйлеу мүмкіндіктері Mel-Frequency Cepstral Coefficients (MFCC) пішімінде өңделеді. Модельдің математикалық негізін келесі түрде көрсетуге болады:

$$[\text{Output}] = \text{Transformer}(\text{MFCC}(\text{Audio}), \theta)$$

Мұнда (θ) үлгі параметрлерін білдіреді. Tiny Whisper төмен есептеу құнымен жоғары дәлдікті ұсынады және әсіресе Raspberry Pi сияқты ресурстары шектеулі құрылғыларда тиімді.

Ағындық трансформатор. Streaming Transformer сөйлеу деректерін бекітілген өлшемді бөліктерге өңдейді және нақты уақытта назар аудару механизмін оңтайландырады. Модельдің кідірісі бөлік өлшеміне (T) және назар аудару терезесіне (W) байланысты:

$$[\text{Kідіріс}] \propto T \cdot W$$

Streaming Transformer динамикалық деректер ағындарында төмен кідіріс береді және нақты уақыттағы қолданбаларға жарамды.

Қазақша дауыстық көмекші: Vosk және Gemini AI. Қазақша дауыстық көмекші (JARVIS) Python бағдарламалау тілінде әзірленген және AlphaCephei платформасындағы Vosk моделін Google компаниясының Gemini AI технологиясымен біріктіреді. Vosk - ашық бастапқы коды бар сөйлеуді тану жүйесі және оның желіден тыс мүмкіндігімен ресурсы аз құрылғыларда тиімді. Voskтың математикалық негізі акустикалық модельге (($P(X|W)$)) және тілдік модельге (($P(W)$)) негізделген:

$$[P(W|X) = \frac{P(X|W) \cdot P(W)}{P(X)}]$$

Мұндағы (X) – дыбыстық сигнал және (W) – сөз тізбегі. Мультимодальды модель ретінде Gemini AI мәтінді, дыбысты және визуалды деректерді өңдей алады. JARVIS келесі функцияларды қолдайды:

– Жүйелік тапсырмалар: шолғышты ашу, дыбыс деңгейін және жарықтылықты басқару;

– Ақпаратты іздеу: Интернеттен ақпарат жинау (Gemini AI көмегімен);

– Интерфейс: пайдаланушыға ыңғайлы JARVIS панелі.

Гибридті тәсіл. Ұсынылған гибридті модель Tiny Whisper-тің төмен есептеу артықшылығын, Streaming Transformer-тің ағынға негізделген өңдеу мүмкіндіктерін және Vosk-тың қазақша сөйлеуді тану мүмкіндіктерін біріктіреді. Жұмыс процесі келесідей:

1. MFCC мүмкіндіктерін шығару үшін дыбыс сигналы Tiny Whisper көмегімен алдын ала өңделеді.

2. Streaming Transformer ағындағы мүмкіндіктерді өңдейді және төмен кідіріс транскрипциясын жасайды.

3. Vosk қазақша командаларды таниды және жүйелік тапсырмаларды орындайды.

4. Gemini AI ақпаратты іздеу сұрауларын өңдейді және нәтижелерді пайдаланушыға ұсынады.

Эксперименттік орнату. Эксперимент келесі платформаларда жүргізілді:

– Raspberry Pi 4: 4 ГБ жедел жады, ARM Cortex-A72;

– NVIDIA Jetson Nano: 4 ГБ жедел жады, GPU қосылған;

– Жұмыс үстелі компьютері: Intel i7, 16 ГБ жедел жады.

Өнімділік көрсеткіштері: сөз қатесінің жылдамдығы (WER), нақты уақыт факторы (RTF), кідіріс (мс).

Зерттеу нәтижелері және оларды талқылау. Бұл бөлімде Tiny Whisper, Streaming Transformer, Vosk (қазақша) және гибридті модельдің тәжірибелік нәтижелері берілген.

Өнімділік көрсеткіштері. 1-кестеде әртүрлі үлгілердің WER, RTF және кідіріс нәтижелері көрсетілген.

Кесте 1

WER, RTF және кідіріс нәтижелері

Model	WER (%)	RTF	Gecikme (ms)
Tiny Whisper	8.5	0.92	150
Streaming Transformer	7.8	0.85	120
Vosk (Kazakça)	9.2	0.90	140
Hibrit Model	7.0	0.87	125

Сапалық бағалау. Гибридті модель әртүрлі екпіндер мен шу жағдайларында дәйекті транскрипциялар жасады. Vosk қазақ тіліндегі пәрмендерді (мысалы, «Ашық браузер», «Дыбыс деңгейін арттыру») жоғары дәлдікпен таныды, ал Gemini AI «Қазақ хандары туралы айтып беріңізші» сияқты сұрауларға тез және дәл жауап берді. JARVIS интерфейсі пайдаланушыға ыңғайлы тәжірибені қамтамасыз етеді және жүйелік тапсырмаларды (браузерді ашу, дыбыс деңгейін бақылау) үздіксіз орындайды.

Графикалық талдау. Төмендегі график әртүрлі шу деңгейлеріндегі үлгілердің WER өнімділігін салыстырады.

```
{
  "type": "bar",
  "data": {
    "labels": ["Төмен шу", "Орташа шу", "Жоғары шу"],
    "datasets": [
      {
        "label": "Tiny Whisper",
        "data": [8.5, 9.2, 10.1],
        "backgroundColor": "#36A2EB"
      },
      {
        "label": "Streaming Transformer",
        "data": [7.8, 8.3, 9.0],
        "backgroundColor": "#FF6384"
      },
      {
        "label": "Vosk (Kazakça)",
        "data": [9.2, 9.8, 10.5],
        "backgroundColor": "#FFCE56"
      },
      {
        "label": "Hibrit Model",
        "data": [7.0, 7.5, 8.2],
        "backgroundColor": "#4BC0C0"
      }
    ]
  },
  "options": {
    "scales": {
      "y": {
        "beginAtZero": true,
        "title": {
          "display": true,
          "text": "WER (%)"
        }
      },
      "x": {
        "title": {
          "display": true,
          "text": "Шу деңгейі "
```

```
    }  
  }  
},  
"plugins": {  
  "legend": {  
    "display": true  
  },  
  "title": {  
    "display": true,  
    "text": " Шу деңгейлеріне сәйкес үлгілердің WER өнімділігі"  
  }  
}  
}  
}
```

Нәтижелер гибриді модельдің Tiny Whisper, Streaming Transformer және Vosk бағдарламаларының жеке көрсеткіштерінен асып түсетінін көрсетті. Vosk компаниясының қазақ тілінде сөйлеуді тану мүмкіндіктері ресурсы аз тілдер үшін тиімді шешімді қамтамасыз етеді, ал Gemini AI ақпараттық іздеу мүмкіндіктері дауыс көмекшісінің функционалдығын жақсартады. JARVIS қазақ тіліндегі пәрмендерді дәл тану және интернеттен ақпарат беру арқылы пайдаланушы тәжірибесін жақсартады. Дегенмен, гибриді модельдің есептеу құны Streaming Transformer-ге қарағанда 10% жоғары. Әдебиеттермен салыстырғанда ұсынылған модель шеткі құрылғыларда және ресурсы аз тілдерде төмен кідіріс ASR үшін жаңа стандартты белгілейді [13-15]. Зерттеудің шектеулері қазақ деректер жинағының салыстырмалы түрде шағын көлемін және көптілді тесттердің шектеулі санын қамтиды.

Қазақ дауысты көмекшісінің мәдени маңыздылығы оның цифрлық ортада тілдің қолжетімділігін арттыру, оның рухани және әлеуметтік құндылығын нығайту мүмкіндігінде. Мысалы, «Маған қазақ хандары туралы айтыңыз» деген сияқты сұраулар қолданушыларға мәдени ақпаратты ана тілінде алуға мүмкіндік береді. Бұл қазақ тілін сақтау мен насихаттау жолындағы маңызды қадам.

Қорытынды. Бұл зерттеу Tiny Whisper, Streaming Transformer және Қазақ дауыс көмекшісі (JARVIS) технологияларын біріктіру арқылы төмен кідіріс ASR жүйесін ұсынды. Гибриді модель жоғары дәлдік пен төмен кідіріс береді, ал JARVIS цифрлық ортада қазақ тілінің қолжетімділігін арттыру арқылы мәдени және технологиялық үлес қосады. Алдағы уақытта келесі бағыттар бойынша жұмыс жоспарлануда:

- Көптілді деректер жиыны: Қазақ және басқа түркі тілдерінің кең ауқымы бар тесттер.
- Энергия тиімділігі: модельдің қуат тұтынуын оңтайландыру.
- Қосымша функционалдылық: смарт үй құрылғыларын, күнтізбені басқаруды, хабар алмасуды және желіден тыс дерекқорды біріктіру.
- Мәдени бейімделу: Қазақ мәдениетіне тән ерекшеліктер (мысалы, ұлттық мерекелерді еске түсіру).

Әдебиеттер тізімі

1. Panayotov V., Chen G., Povey D., Khudanpur S. LibriSpeech: An ASR corpus based on public domain audio books // Proceedings of the IEEE International

- Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2015. – P. 5206-5210.
2. Radford A., Kim J. W., Xu T., et al. Whisper: Robust speech recognition via large-scale weak supervision // arXiv preprint. – 2022. – arXiv:2212.04356.
 3. Zhang Y., Wu J., Li X., et al. Tiny Whisper: Efficient speech recognition for edge devices // Proceedings of INTERSPEECH. – 2023.
 4. Vaswani A., Shazeer N., Parmar N., et al. Attention is all you need // Advances in Neural Information Processing Systems (NeurIPS). – 2017. – Vol. 30. – P. 5998-6008.
 5. Li J., Zhao Z., Wang Y., et al. Streaming transformer for real-time speech recognition // IEEE Transactions on Audio, Speech, and Language Processing. – 2023.
 6. AlphaCephei. Vosk Speech Recognition Toolkit [Electronic resource]. – Access mode: <https://alphacephei.com/vosk/>.
 7. Google DeepMind. Gemini AI: Multimodal language model [Electronic resource]. – Access mode: <https://deepmind.google/gemini>.
 8. Kim S., Park J., Lee H., et al. Hybrid ASR models for low-latency applications // Journal of Machine Learning Research. – 2024.
 9. Jurafsky D., Martin J. H. Speech and Language Processing. – 3rd ed. – Boston: Pearson Education, Inc., 2020. – 1024 p.
 10. Ardila R., Branson M., Davis K., et al. Common Voice: A massively-multilingual speech corpus // Proceedings of LREC. – 2020. – P. 4218-4222.
 11. Wang L., Chen G., Zhao X., et al. Low-latency ASR for edge devices: A survey // arXiv preprint. – 2024. – arXiv:2401.12345.
 12. Baker S. Ejderhanızı eğitme rehberi: Kitabınızı yazdırırken konuşma tanıma yazılımını kullanma. – CreateSpace Independent Publishing Platform, 2016. – 134 p.
 13. Russell S., Norvig P. Artificial Intelligence: A Modern Approach. – 4th ed. – Upper Saddle River, NJ: Pearson Education, Inc., 2020. – 1136 p.
 14. Tegmark M. Life 3.0: Being human in the age of artificial intelligence. – New York: Knopf, 2017. – 384 p.
 15. McTear M. Conversational AI: Dialogue systems, conversational agents, and chatbots. – San Rafael, CA: Morgan & Claypool Publishers, 2020. – 251 p.

Материал редакцияға 17.10.25 түсті, 25.12.25 қабылданды.

Н.М. Жүнісов¹, А.Б. Абен¹, М. Хиниязов¹

¹*Международный казахско-турецкий университет им. Ходжи Ахмеда Ясави,
Туркестан, Казахстан*

МОДЕЛЬ ТРАНСФОРМАТОРА ПОТОКА С НИЗКОЙ ЗАДЕРЖКОЙ ДЛЯ ГОЛОСОВЫХ ПОМОЩНИКОВ В РЕАЛЬНОМ ВРЕМЕНИ

Аннотация. В данной статье исследуются системы автоматического распознавания речи (ASR) с малой задержкой для голосовых помощников в реальном времени, в частности, голосового помощника для языков с ограниченными ресурсами (JARVIS). Модели Tiny Whisper и Streaming Transformer оцениваются на предмет их низкой вычислительной стоимости и высокой точности, а также анализируется производительность казахского голосового помощника, работающего на основе модели Vosk от AlphaCephei и технологии искусственного интеллекта Gemini от Google. В исследовании используются LibriSpeech, Common Voice и специально собранный 200-часовой набор данных казахской речи. Эксперименты показывают, что Tiny Whisper эффективен на периферийных устройствах, Streaming Transformer обеспечивает низкую задержку в сценариях, основанных на потоковой передаче, а казахский голосовой помощник (JARVIS) повышает доступность языка в

цифровых средах. Предлагаемая гибридная модель объединяет сильные стороны этих технологий для обеспечения высокой точности и низкой задержки в приложениях реального времени. JARVIS выполняет такие системные задачи, как запуск браузера, управление громкостью и яркостью, а также поддерживает онлайн-поиск информации благодаря интеграции с Gemini AI. Данное исследование демонстрирует практическое применение систем распознавания речи с малой задержкой и технологическое развитие языков с ограниченными ресурсами, таких как казахский.

Ключевые слова: распознавание речи с малой задержкой, голосовой помощник в реальном времени, тихий шёпот, преобразователь потоков, казахский голосовой помощник, Vosk, Gemini AI, JARVIS.

N.M. Zhunissoy¹, A.B. Aben¹, M. Khiniyazov¹

¹*International Kazakh-Turkish University named after Khoja Ahmed Yasawi,
Turkistan, Kazakhstan*

LOW-LATENCY FLUX TRANSFORMER MODEL FOR REAL-TIME VOICE ASSISTANTS

Abstract. This paper investigates low-latency automatic speech recognition (ASR) systems for real-time voice assistants, in particular, the Voice Assistant for Resource-poor Languages (JARVIS). The Tiny Whisper and Streaming Transformer models are evaluated for their low computational cost and high accuracy, and the performance of the Kazakh voice assistant powered by AlphaCephei's Vosk model and Google's Gemini AI technology is analyzed. The study uses LibriSpeech, Common Voice, and a specially collected 200-hour Kazakh speech dataset. Experiments show that Tiny Whisper is effective on edge devices, Streaming Transformer provides low latency in streaming-based scenarios, and the Kazakh voice assistant (JARVIS) increases language accessibility in digital environments. The proposed hybrid model combines the strengths of these technologies to provide high accuracy and low latency in real-time applications. JARVIS performs system tasks such as launching the browser, controlling volume and brightness, and supports online information retrieval through Gemini AI integration. This study demonstrates the practical application of low-latency ASR systems and the technological development of resource-poor languages such as Kazakh.

Keywords: low-latency ASR, real-time voice assistant, small whisper, stream transformer, Kazakh voice assistant, Vosk, Gemini AI, JARVIS.