әл-Фараби атындағы Қазақ ұлттық университеті

ӘОЖ 004.056 Қолжазба құқығында

ДҮЙСЕНБЕККЫЗЫ ЖАНСАЯ

Сөйлеу қабілеті бұзылған балалардың қазақша сөйлеуді тану әдісін зерттеу және әзірлеу

8D06101-Ақпараттық жүйелер

Философия докторы (PhD) дәрежесін алу үшін дайындалған диссертация

Отандық ғылыми кеңесші:
Рахимова Диана Рамазановна,
Ақпараттық жүйелер
мамандығы бойынша PhD, қауымдастырылған профессор
әл-Фараби атындағы Қазақ ұлттық университеті
Шетелдік ғылыми кеңесші:
Eśref Adali, профессор,
Стамбул Техникалық Университеті,
Түркия

Қазақстан Республикасы
Алматы, 2026

МАЗМҰНЫ

НОРМАТИВТІ СІЛТЕМЕЛЕР 4
БЕЛГІЛЕУЛЕР МЕН ҚЫСҚАРТУЛАР 5
КІРІСПЕ 6
1 СӨЙЛЕУДІ ТАНУ ЖӘНЕ СӨЙЛЕУ БҰЗЫЛЫСТАРЫ САЛАСЫНДАҒЫ ЗЕРТТЕУЛЕРДІ ТАЛДАУ 11
1.1 Сөйлеуді автоматты тану жүйелерінің (ASR) даму эволюциясы 11
1.2 Қазіргі ASR модельдері және олардың ерекшеліктері 13
1.3 Балалар сөйлеуінің лингвистикалық және акустикалық ерекшеліктері ... 21
1.4 Сөйлеу қабілеті бұзылған балалардың сөйлеу ерекшеліктері 23
1.5. Қазақ тіліндегі сөйлеуді тану мәселелері 31
1.6. Балалар және бұзылған сөйлеуді тану бойынша заманауи зерттеулер 37
2 ҚАЗАҚ ТІЛІНДЕГІ БАЛАЛАР СӨЙЛЕУІНІҢ АННОТАЦИЯЛАНҒАН АКУСТИКАЛЫҚ КОРПУСТЫ ӘЗІРЛЕУ ЖӘНЕ ӨҢДЕУ ӘДІСТЕРІ 43
2.1 Балалар сөйлеуіне арналған мәтіндік материалды іріктеу және құрылымдау әдіснамасы 43
2.1.1 Лингвистикалық материалды іріктеу критерийлері 43
2.1.2 Сөздік материалды қалыптастыру әдіснамасы 45
2.1.3 Генерация және LLM қолдану арқылы кеңейту 47
2.1.4 Морфологиялық талдау және сүзгілеу 50
2.2 Балалар сөйлеуінің акустикалық деректерін жинау 54
2.2.1 Деректерді жинау әдістері 54
2.2.2 Аннотациялау және метадеректер 59
2.2.3 Этикалық және құқықтық аспектілер 62
2.3. Деректер құрылымы 64
2.4 Деректерді алдын ала өңдеу 66
3 СӨЙЛЕУ ҚАБІЛЕТІ БҰЗЫЛҒАН БАЛАЛАРДЫҢ ҚАЗАҚША СӨЙЛЕУІН ТАНУҒА АРНАЛҒАН КОМПЛЕКСТІ ӘДІСТІ ӘЗІРЛЕУ ЖӘНЕ ОНЫҢ МАТЕМАТИКАЛЫҚ МОДЕЛІ 73
3.1 Ұсынылған комплексті әдістің архитектуралық құрылымы 73
3.2 Сөйлеуді тану есебінің формализациясы 76
3.3 Балалар сөйлеуіне бейімделген акустикалық модельдің математикалық қойылымы 80

3.4.Балалар сөйлеуіндегі вариативтілікті ескеру мен сөйлеу бұзылыстарын модельдеу 82
3.5 Комплексті әдістің математикалық моделі 87
3.6 Ұсынылған әдістің негізгі нәтижелері және 3-тарау бойынша қорытынды 89
4 КОМПЛЕКСТІ ӘДІСТІ ЭКСПЕРИМЕНТТІК БАҒАЛАУ ЖӘНЕ ҚАЗАҚ ТІЛІНДЕГІ БАЛАЛАР СӨЙЛЕУІН ТАНУ НӘТИЖЕЛЕРІН ТАЛДАУ 90
4.1 Эксперименттік деректер сипаттамасы және бағалау метрикалары 90
4.2 Қазақ тіліндегі балалар сөйлеуін тануда ASR модельдерінің эксперименттік салыстырмалы талдауы 92
4.2.1 Базалық ASR модельдерінің zero-shot режиміндегі эксперименттік бағалануы 94
4.2.2 ASR (fine-tuning) модельдерін балалар сөйлеуіне бейімдеу нәтижелерін эксперименттік бағалау 97
4.3 Ұсынылған комплексті әдістің нәтижелерін бағалау және талдау 99
4.4 Сөйлеу бұзылыстары мен ASR қателерін комплексті талдау 104
4.5 Ұсынылған әдістің практикалық ортада апробациясы және қолданбалы нәтижелері 107
4.6 4-тарау бойынша қорытынды 111
ҚОРЫТЫНДЫ 113
ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ 115
ҚОСЫМША А Авторлық куәліктер 123
ҚОСЫМША Б Ендіру актісі 125

НОРМАТИВТІ СІЛТЕМЕЛЕР

Берілген диссертацияда келесі стандарттарға сәйкес сілтемелер қолданылды: Қазақстан Республикасы МЖМБС 5.04.034-2011. «Қазақстан Республикасы

Мемлекеттік жалпыға білім беру стандарты. Жоғары оқу орнынан кейінгі білім. Докторантура». Негізгі ережелер ҚР БҒМ бекітілген. 17.06.2011 ж. №261. Астана, 2011 ж.

«Диссертацияны безендіру нұсқаулығы», Қазақстан Республикасы БҒМ ЖАК 28 қыркүйек, 2004 жыл. №377-3 ж.

ГОСТ 7.32-2001. Ғылыми зерттеу жұмысының есебі. Безендіру ережелері мен құрылымы.

ГОСТ 7.1-2003. Библиографиялық таспа. Библиографиялық сипаттау.

БЕЛГІЛЕУЛЕР МЕН ҚЫСҚАРТУЛАР

ASR – Automatic Speech Recognition (Автоматтандырылған сөйлеуді тану)
NLP – Natural Language Processing (Табиғи тілдерді өңдеу)
ML – Machine Learning (Машиналық оқыту)
NN – Neural Networks (Нейрондық желілер)
RNN – Recurrent Neural Networks (Қайталау нейрондық желілері)
LSTM – Long Short-Term Memory (Ұзақ және қысқа мерзімді жады)
WER – Word Error Rate (Сөз қатесі деңгейі)
CER – Character Error Rate (Символ қатесі деңгейі)
LM – Language Model (Тіл моделі)
HMM – Hidden Markov Models (Жасырын Марков модельдері)
GMM – Gaussian Mixture Model (Гаусс қоспасы моделі)
CNN – Convolutional Neural Network (Конволюциялық нейрондық желі)
DNN – Deep Neural Network (Терең нейрондық желі)
BLEU – Bilingual Evaluation Understudy (Мәтін апасын бағалау метрикасы)
LM – Language Model (Тіл моделі)
SNR – Signal to Noise Ratio (Сигнал мен шу қатынасы)
F0 – Fundamental Frequency (Негізгі жиілік)
MFCC – Mel-Frequency Cepstral Coefficients (Мел жиілік шкаласындағы кепстралдық коэффициенттер)
Chr2f- Character n-gram F-score (Символ деңгейіндегі мәтінді бағалау метрикасы)

КІРІСПЕ

Зерттеу тақырыбының өзектілігі. Қазіргі ғылыми-технологиялық дамудың жаңа кезеңінде сөйлеуді автоматты тану (Automatic Speech Recognition, ASR) жүйелері тіл технологиялары саласының маңызды бағытына айналып отыр. Аталған жүйелер білім беру платформаларында, логопедиялық диагностика құралдарында, ерте бал**а**лық шақты дамытуға бағытталған ақпараттық жүйелерде кеңінен қолданыс тапқан. Дегенмен, балалар сөйлеуін автоматты тану мәселесі — ғылыми тұрғыдан жеткіліксіз зерттелген, практикалық шешімі толық табылмаған күрделі ғылыми-техникалық проблема болып қала береді.

Балалар мен ересектер сөйлеуі арасындағы акустикалық айырмашылықтар эмпирикалық зерттеулермен дәлелденген. Атап айтқанда, балалардың негізгі жиілігі (F0) 250–400 Гц шегінде, ал ересектерде бұл көрсеткіш 100–150 Гц деңгейінде болады; форманттық жиіліктер де айтарлықтай жоғары — балалардың ауыз қуысы ұзындығы 9–13 см болса, ересектерде бұл 17 см-ге жетеді. Сөйлеу қарқыны тұрақсыз, толық артикуляция жасы 12–14-ке жеткенде ғана қалыптасады. Зерттеулер дәлелдегендей, ересек сөйлеуіне негізделген ASR модельдері балалардың сөйлеуіне қолданылғанда сөздердегі қате мөлшері (Word Error Rate, WER) 40–45%-ға өсетіні анықталған [1].

Балалар сөйлеуінің жас ерекшелігіне тәуелді қалыптасуы да ғылыми зерттеулермен расталған. Г. Лямина жүргізген бақылаулар бойынша 2–3 жастағы балалар 30 минуттық ойын барысында орта есеппен 25–28 лексикалық бірлік, 3–4 жаста — 70–80, ал 4–4,5 жаста — 110–115 бірлік айтады. Осыған байланысты сөйлеу дамуындағы кешігуді ерте анықтаудың оңтайлы жасы — 3–4 жас аралығы болып табылады [2].

Қазақ тілі — агглютинативті морфологиялық құрылымға ие, ресурстары шектеулі (low-resource) тіл. Қазіргі Қазақ ASR жүйелері — KATEK, KazASR, KazakhTTS, KazakhBERT, Soyle — тек ересектер сөйлеуіне негізделген; балалар сөйлеуіне арналған мамандандырылған акустикалық корпус 2025 жылға дейін болмаған. Халықаралық деректерге сүйенсек, 2013 жылы транскрипциясы бар 13 балалар сөйлеу корпусы тіркелді, алайда олардың ешқайсысы қазақ тілін қамтымаған [3]. Karabaliyev және Kolesnikova [4] жүргізген тәуелсіз бағалауда Kaldi, Mozilla DeepSpeech және Google Speech-to-Text жүйелерінің қазақ тіліне тән фонемаларды (қ, ң, ү, ұ) жиі қате тануы анықталды; ең жақсы нәтижені Google STT WER = 52,97% деңгейінде ғана көрсетті — практикалық қолданыс үшін жеткіліксіз.

Жоғарыда баяндалған ғылыми дәлелдемелер мен қолданыстағы жүйелердің шектеулері сөйлеу қабілеті бұзылған балалардың қазақша сөйлеуін автоматты тану мен ерте диагностикалауға арналған ғылыми-техникалық шешімнің өзектілігін айқындайды. Осы зерттеу жұмысы аталған проблемалық кеңістікті толтыруға бағытталған.

Диссертациялық жұмыстың мақсаты – қазақ тіліндегі балалар сөйлеуіне арналған аннотацияланған акустикалық деректер қорын құру және

сөйлеу қабілеті бұзылған балалардың сөйлеуін тану дәлдігін арттыруға бағытталған ASR және LLM интеграциясына негізделген комплексті әдісті әзірлеу.

Зерттеудің міндеттері. Қойылған мақсатқа қол жеткізу үшін келесі міндеттерді орындау қарастырылады:

1) Дүниежүзіндегі балалар акустикалық аудиодеректеріне салыстырмалы-аналитикалық шолу жүргізу және олардың құрылымдық, көлемдік, акустикалық ерекшеліктерін жүйелеу. Балалар сөйлеуін тану саласындағы заманауи әдістер мен акустикалық корпустарға салыстырмалы талдау жүргізу;

2) Балалардың жас ерекшелігі мен жыныстық сипаттамаларын ескере отырып, қазақ тіліндегі балалар сөйлеуіне арналған аннотацияланған акустикалық датасет жинақтау және өңдеу;

3) Заманауи нейрондық ASR модельдерін (Whisper, MMS, TurkicASR, ESPnet, DeepSpeech және т.б.) зерттеу және оларды балалар сөйлеуіне бейімдеу;

4) Қосымша оқытудан кейінгі ASR нәтижелерін түзетуге арналған LLM негізіндегі комплексті post-editing тәсілін әзірлеу.

5) Ұсынылған әдістің тиімділігін эксперименттік тұрғыда бағалау және салыстырмалы талдау жүргізу.

Зерттеу нысаны. Қазақ тіліндегі сөйлеу қабілеті бұзылған балалардың сөйлеу үдерісі және оның акустикалық ерекшеліктері.

Зерттеу пәні. Қазақ тіліндегі сөйлеу қабілеті бұзылған балалардың сөйлеуін тануға арналған технологиялар, алгоритмдер, модельдер мен әдістер, бағдарламалық жасақтама.

Зерттеу әдісі. Машиналық оқыту әдістері, сөйлеуді автоматты тану технологиялары мен акустикалық өңдеу әдістері, ықтималдықтар теориясы мен математикалық статистика, бағдарламалық жасақтаманы әзірлеу әдістері.

Алынған нәтижелердің ғылыми жаңалығы:

1) Қазақ тілінде сөйлейтін балалардың жас, жыныс және фонетикалық ерекшеліктерін ескеретін арнайы акустикалық датасет алғаш рет әзірленді;

2) Сөйлеу қабілеті бұзылған балалардың қазақша сөйлеуін тануға арналған LLM негізіндегі комплексті post-editing тәсілі әзірленді;

Жұмыстың теориялық маңыздылығы. Диссертациялық жұмыстың теориялық маңыздылығы зерттеу барысында қазақ тіліндегі сөйлеу қабілеті бұзылған балалардың сөйлеуін автоматты тануға арналған тілдік модельдер негізіндегі post-editing әдістерін біріктіретін комплексті тәсілді әзірлеу, сондай-ақ балалардың жас және жыныс ерекшеліктерін ескеретін арнайы акустикалық датасет құру.

Жұмыстың практикалық маңызыдылығы. Диссертациялық жұмыстың практикалық маңыздылығы әзірленген алгоритмдер мен бағдарламалық қамтамасыз етуді сөйлеуді өңдеуге байланысты әртүрлі технологияларда қолдану мүмкіндігімен анықталады, атап айтқанда сөйлеу

синтезі, машиналық аударма, дауыс арқылы аутентификация және сөйлеуді талдау жүйелерінде. Әзірленген қазақ тіліндегі балалар сөйлеуін автоматты тану жүйесі логопедиялық практикада, білім беру ұйымдарында, медициналық және түзету орталықтарында, сондай-ақ мобильді қосымшалар мен цифрлық сервистерде енгізіледі.

Қорғауға шығарылатын негізгі тұжырым. Қазақ тіліндегі балалар сөйлеуін тануға арналған аннотацияланған акустикалық датасет (8 594 аудио жазба, 3–8 жас) жинақталды. Сөйлеу қабілеті бұзылған балалардың қазақша сөйлеуін тануға арналған ASR моделін fine-tuning арқылы бейімдеу және LLM негізіндегі постөңдеуді біріктіретін комплексті postediting тәсілі ұсынылды. Жүргізілген эксперименттер негізінде ұсынылған комплексті postediting тәсілін қолдану арқылы WER — 0,70-тен 0,18-ге, CER — 0,467-ден 0,065-ке төмендетуге болатындығы анықталды.

Сенімділік дәрежесі мен апробациялау нәтижелері. Диссертация тақырыбына қатысты зерттеулер мен нәтижелер төмендегі жарияланымдар негізінде әртүрлі конференциялар мен семинарларда баяндалып, талқыланды:

1) Rakhimova D., **Duisenbekkyzy Zh.**, Adali E. Investigation of ASR Models for Low-Resource Kazakh Child Speech // Applied Sciences. – 2025. – Vol. 15, Issue 16, Article 8989. DOI: 10.3390/15168989 (WoS SCIE, Q2, IF: 2,7, процентиль 79).

2) Rakhimova D., Karibayeva A., Turarbek A., Karyukin V.I., **Duisenbekkyzy Zh.** and Aliyev R. Development of a Children's Educational Dictionary for a Low-Resource Language Using AI Tools // Computers. – 2024. – Vol. 13, No. 10, Article 253. DOI: 10.3390/computers13100253 (WoS ESCI, Q1, CiteScore: 7,5, процентиль 84).

3) Д.Р. Рахимова., **Ж. Дүйсенбекқызы**, О.А. Сағынтай, Ә. Турарбек, Ш. Төлеуғали. Балалардың сөйлеуін тану жүйелеріндегі шуды азайту алгоритмдерін талдау // ҚазҰПУ хабаршысы. – 2025. – №91(3). – Б. 230–242. 1. DOI: https://doi.org/10.51889/2959-5894.2025.91.3.021

4) Rakhimova D., **Duisenbekkyzy Zh.**, Adali E., Toleugali Sh. Research and Processing of the Kazakh Children's Acoustic Corpus // ҚР ҰИА Хабаршысы. – 2024. – №3(93).

5) Rakhimova D., **Duisenbekkyzy Zh.,** Yessirkepova A., Adali E. Fine-Tuning Neural ASR Models for Low-Resource Kazakh Children's Speech // UBMK'2025. – Istanbul, 2025. – P. 428–432 (Scopus), процентиль 24.

6) Rakhimova D., **Duisenbekkyzy Zh.** Research on the Impact of Social Networks on the Development of Bilingualism in Kazakhstan // Advances in Intelligent Systems and Computing. – 2024. – P. 123–134 (Springer), процентиль 39.

7) Рахимова Д.Р., **Дүйсенбекқызы Ж.**, Шадияр Т. Қазақ тіліндегі балалар корпусы үшін деректерді жинау және өңдеу // «ФАРАБИ ӘЛЕМІ». – Алматы, 2024.

8) Рахимова Д.Р., **Дүйсенбекқызы Ж.**, Турганбаева М.Ж., Арабов Б.Б. Балалар акустикалық корпусын қазақ тілінде жинау // «ǴYLYM JÁNE BILIM – 2024». – 2024.

9) **Дүйсенбекқызы Ж.**, Рахимова Д.Р., Есиркепова А.П. Нейрондық ASR үлгілерін төмен ресурсты қазақ тілінде бейімдеу // «Информатика және қолданбалы математика» X конференциясы. – Алматы, 8–11 қазан 2025 ж. – Б. 303–309.

Зерттеушінің жеке үлесі. Докторант диссертациялық жұмыстың қойылған міндеттерін өз бетінше орындап шешті. Қазақ тіліндегі балалар сөйлеуіне арналған аннотацияланған акустикалық датасет жинақтап, өңдеді. ASR модельдерін (Whisper, MMS, TurkicASR, ESPnet, DeepSpeech) fine-tuning арқылы бейімдеу тәжірибелерін жүргізді. Әзірленген модель мен әдістердің эксперименттік бағалауын орындады. LLM негізіндегі комплексті postediting тәсілі мен бағдарламалық архитектурасын жасады.

Диссертация тақырыбының ғылыми-зерттеу жұмыстарының жоспарларымен байланысы. Диссертация Қазақстан Республикасы Білім және ғылым министрлігінің (қаржыландыру көзі: Қазақстан Республикасы Білім және ғылым министрлігінің Ғылым комитеті) ғылыми-зерттеу гранттық қаржыландыру AP 19577833 жобасы аясында орындалды.

Басым бағыт: Ақпараттық технологиялар және жасанды интеллект.

Мамандандырылған бағыт: Табиғи тілді өңдеу және сөйлеу технологиялары.

Нәтиженің жарияланымдары. Диссертация тақырыбы бойынша 2 авторлық куәлік алынды және 9 ғылыми жұмыс жарияланды, оның ішінде Scopus және Web of Science базаларына индекселетін импакт-факторы бар басылымдарда 4 мақала жарияланды, Қазақстан Республикасы Білім және ғылым министрлігі Білім және ғылым саласындағы бақылау комитеті ұсынатын журналдарда 2 мақала жарияланды:

1) Rakhimova D., **Duisenbekkyzy Zh.**, Adali E. Investigation of ASR Models for Low-Resource Kazakh Child Speech // Applied Sciences. – 2025. – Vol. 15, Issue 16, Article 8989. DOI: 10.3390/15168989 (WoS SCIE, Q2, IF: 2,7, процентиль 79).

2) Rakhimova D., Karibayeva A., Turarbek A., Karyukin V.I., **Duisenbekkyzy Zh.** and Aliyev R. Development of a Children's Educational Dictionary for a Low-Resource Language Using AI Tools // Computers. – 2024. – Vol. 13, No. 10, Article 253. DOI: 10.3390/computers13100253 (WoS ESCI, Q1, CiteScore: 7,5, процентиль 84).

3) Д.Р. Рахимова., **Ж. Дүйсенбекқызы**, О.А. Сағынтай, Ә. Турарбек, Ш. Төлеуғали. Балалардың сөйлеуін тану жүйелеріндегі шуды азайту алгоритмдерін талдау // ҚазҰПУ хабаршысы. – 2025. – №91(3). – Б. 230–242. 1. DOI: https://doi.org/10.51889/2959-5894.2025.91.3.021

4) Rakhimova D., **Duisenbekkyzy Zh.**, Adali E., Toleugali Sh. Research and Processing of the Kazakh Children's Acoustic Corpus // ҚР ҰИА Хабаршысы. – 2024. – №3(93).

5) Rakhimova D., **Duisenbekkyzy Zh.,** Yessirkepova A., Adali E. Fine-Tuning Neural ASR Models for Low-Resource Kazakh Children's Speech // UBMK'2025. – Istanbul, 2025. – P. 428–432 (Scopus) , процентиль 24.

6) Rakhimova D., **Duisenbekkyzy Zh.** Research on the Impact of Social Networks on the Development of Bilingualism in Kazakhstan // Advances in Intelligent Systems and Computing. – 2024. – P. 123–134 (Springer) , процентиль 39.

7) Рахимова Д.Р., **Дүйсенбекқызы Ж.**, Шадияр Т. Қазақ тіліндегі балалар корпусы үшін деректерді жинау және өңдеу // «ФАРАБИ ӘЛЕМІ». – Алматы, 2024.

8) Рахимова Д.Р., **Дүйсенбекқызы Ж.**, Турганбаева М.Ж., Арабов Б.Б. Балалар акустикалық корпусын қазақ тілінде жинау // «ǴYLYM JÁNE BILIM – 2024». – 2024.

9) **Дүйсенбекқызы Ж.**, Рахимова Д.Р., Есиркепова А.П. Нейрондық ASR үлгілерін төмен ресурсты қазақ тілінде бейімдеу // «Информатика және қолданбалы математика» X конференциясы. – Алматы, 8–11 қазан 2025 ж. – Б. 303–309.

10) Авторлық куәлік «Dataset Loader Bot» — балалардың қазақша сөйлеу деректерін жартылай автоматты жинау жүйесі. Авторлық құқық объектісі №56775, тіркелген күні: 15 сәуір 2025 ж. Авторлар: Рахимова Д., **Дүйсенбекқызы Ж.**, Алиев Р.

11) Авторлық куәлік «Қазақ тіліндегі 2–9 жас аралығындағы балалар сөйлеу аудиожазбаларының таңбаланған акустикалық деректер базасы». Авторлық құқық объектісі № 70955, тіркелген күні: 2026 жылғы «27» сәуір. Авторлар: **Дүйсенбекқызы Ж.,**Рахимова, Д.

Диссертация құрылымы және көлемі. Диссертациялық жұмыс кіріспеден, 4 бөлімнен, қорытынды, әдебиеттер тізімі және қосымшалардан тұрады. Диссертацияның толық көлемі: 125 бет машиналық мәтіні, оның ішіне 30 сурет, 44 кесте, 117 пайдаланылған әдебиеттер тізімі және 4 қосымша кіреді.

1 СӨЙЛЕУДІ ТАНУ ЖӘНЕ СӨЙЛЕУ БҰЗЫЛЫСТАРЫ САЛАСЫНДАҒЫ ЗЕРТТЕУЛЕРДІ ТАЛДАУ

Балалар сөйлеуін автоматты тану (ASR) саласындағы зерттеулердің өзектілігі қазіргі цифрлық білім беру, медициналық диагностика және инклюзивті технологиялар дамуымен тығыз байланысты. Бұл салада негізгі қиындықтар:1) балалар мен ересектер сөйлеуінің акустикалық-физиологиялық айырмашылықтары; 2) сөйлеу қабілеті бұзылыстарының күрделі акустикалық үлгілері; 3) қазақ тілі сияқты аз ресурсты тілдердегі дерек тапшылығы.

Бұл тарауда ASR жүйелерінің даму тарихы, балалар сөйлеуінің акустикалық-фонетикалық ерекшеліктері, сөйлеу бұзылыстарының жіктелуі және аз ресурсты тілдердегі халықаралық тәжірибе жүйелі түрде талданады.

1.1 Сөйлеуді автоматты тану жүйелерінің (ASR) даму эволюциясы

ASR жүйелерінің дамуы 1950-жылдардан басталып, бірнеше кезеңнен өткен: акустикалық-фонетикалық модельдерден гибридті нейрондық желілерге, одан соң заманауи интегралды (end-to-end) архитектураларға дейін жетті.

Дәстүрлі ASR жүйелері негізінен үш тәуелсіз компоненттен тұрады: акустикалық модель (AM), тілдік модель (LM) және лексикон. Акустикалық модель аудио сигналдан фонемалық ерекшеліктерді шығарады; тілдік модель сөздер тізбегінің ықтималдығын бағалайды; лексикон фонемалар мен сөздер арасындағы байланысты қамтамасыз етеді. Ұзақ уақыт бойы жасырын Марков модельдері (Hidden Markov Models, HMM) [5] мен Гаусс қоспалары (Gaussian Mixture Models, GMM) [6] классикалық акустикалық модельдеу негізі болды.Автоматты сөйлеуді тану жүйелерінің даму тарихын бірнеше кезеңге бөлуге болады. Бірінші кезеңде жүйелер ережеге негізделген және шағын сөздікпен шектелген. Мұндай жүйелер нақты бұйрықтар мен шектеулі сөз тіркестерін тануға ғана жарамды болды. Кейіннен динамикалық уақытты түзету, яғни Dynamic Time Warping сияқты әдістер қолданылды. Бұл тәсілдер сөйлеу жылдамдығының өзгерісін белгілі бір деңгейде ескеруге мүмкіндік берді.

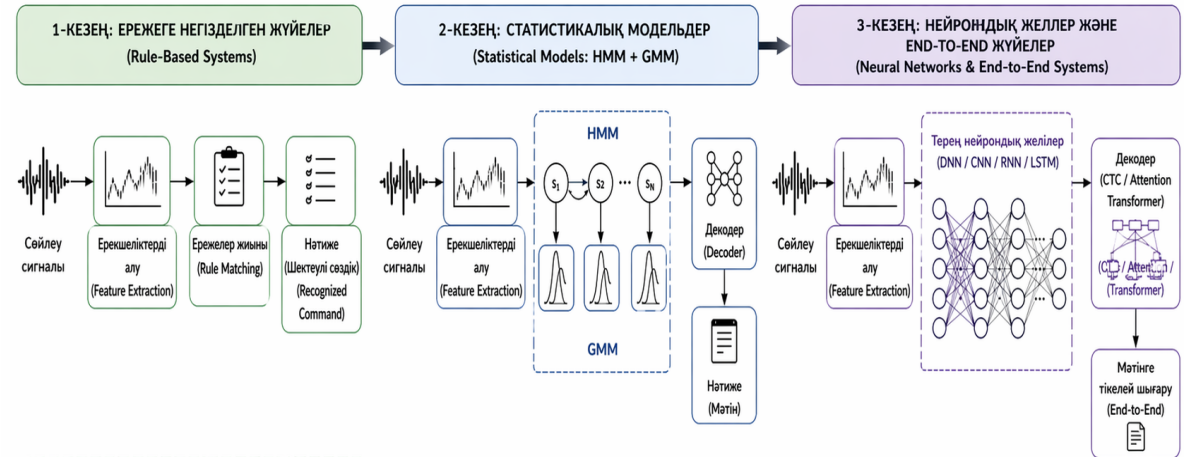
Екінші кезеңде статистикалық сөйлеуді тану жүйелері қалыптасты. Бұл кезеңнің негізін Hidden Markov Model және Gaussian Mixture Model әдістері құрады. Мұнда сөйлеу сигналы уақыт бойынша жасырын күй тізбегі ретінде қарастырылды. Мұндай модельдер көп жылдар бойы стандарт болып келді, өйткені олар акустикалық сигналдың уақыттық құрылымын салыстырмалы түрде жақсы сипаттады.

Үшінші кезеңде нейрондық желілердің дамуы ASR сапасын айтарлықтай жақсартты. DNN, CNN, RNN және LSTM [7-8] модельдері акустикалық ерекшеліктерді дәлірек үйренуге мүмкіндік берді. Кейінірек CTC, attention және Transformer негізіндегі end-to-end жүйелер пайда болды. Бұл жүйелерде акустикалық модель мен тілдік модельдің классикалық бөлінуі азайып, кіріс сигналдан тікелей мәтінге бейнелеу жүзеге асырылады. Осы талдау нәтижелері 1-кестеде көрсетілген.

Кесте 1.1 — ASR жүйелерінің даму кезеңдері

Кезең	Жылдар	Негізгі технология	Маңызды нәтиже
1-кезең	1950–1980	Акустикалық-фонетикалық модельдер	Жеке сандар мен қарапайым сөздерді тану
2-кезең	1980–2015	HMM-GMM архитектурасы және HMM-DNN гибридті модельдер	Үздіксіз сөйлеуді тану, сөздік кеңейту
3-кезең	2015–2026	CTC, Attention, Encoder-Decoder, Transformer, Whisper, wav2vec2, HuBERT	Интегралды модельдер тілдік модельсіз жұмыс істейді. Аз ресурсты тілдерге transfer learning

1.1-кестеде ASR жүйелерінің тарихи даму кезеңдері және әр кезеңдегі негізгі технологиялар көрсетілген. Кестеден байқауға болатындай, бастапқы кезеңде қарапайым акустикалық модельдер қолданылса, кейінгі кезеңдерде статистикалық әдістер, ал соңғы уақытта терең нейрондық желілер мен end-to-end архитектуралар кеңінен таралған. Бұл өзгерістер сөйлеуді тану дәлдігінің артуымен және модельдердің күрделенуімен сипатталады.



Сурет 1.1 - ASR жүйелерінің даму эволюциясы

Сурет 1.1-де ASR жүйелерінің эволюциялық даму траекториясы көрсетілген. Әрбір келесі кезең алдыңғы әдістердің шектеулерін жоюға бағытталған. Атап айтқанда, нейрондық желілер мен Transformer негізіндегі модельдер сөйлеудің күрделі акустикалық және тілдік ерекшеліктерін тиімді модельдеуге мүмкіндік береді [9].

ASR жүйелерін бірнеше ғылыми өлшем бойынша жіктеуге болады. Бірінші өлшем — архитектуралық құрылымы. Осыған сәйкес жүйелерді дәстүрлі, гибридті және интегралды жүйелерге бөледі. Дәстүрлі жүйелерде акустикалық модель, лексикон және тілдік модель бөлек жұмыс істейді. Гибридті жүйелерде DNN-HMM комбинациясы қолданылады. Интегралды жүйелерде бүкіл тану процесі бір нейрондық архитектура арқылы орындалады.

Екінші өлшем — дерек көлемі бойынша жіктеу. Мұнда ASR жүйелері үлкен ресурсты және аз ресурсты тілдерге арналған болып бөлінеді. Үлкен ресурсты тілдерде транскрипцияланған дерек көп болғандықтан, модельді оқыту жеңіл. Ал қазақ тілі сияқты аз ресурсты тілдерде корпус көлемі шектеулі болғандықтан, transfer learning, multilingual learning және data augmentation тәсілдері кең қолданылады.

Үшінші өлшем — сөйлеушіге тәуелділік. Speaker-dependent жүйелер нақты бір адамға бейімделсе, speaker-independent жүйелер кез келген сөйлеушіні тануға бағытталады. Балалар сөйлеуін тануда speaker-independent жүйелер жиі қолданылады, бірақ олардың дәлдігін көтеру үшін қосымша бейімдеу қажет.

Кесте 1.2 - ASR жүйелерінің жіктелуі

Жіктеу өлшемі	**Түрлері**	**Сипаттамасы**
Архитектурасы бойынша	Дәстүрлі, гибридті, end-to-end	Модельдің ішкі құрылымына негізделген жіктеу
Дерек көлемі бойынша	Үлкен ресурсты, аз ресурсты	Қолданылатын корпус көлеміне байланысты
Сөйлеушіге қатыстылығы бойынша	Speaker-dependent, speaker-independent	Белгілі бір адамға бейімделуі немесе әмбебаптығы
Қолданылу саласы бойынша	Жалпы мақсаттағы, домендік	Еркін сөйлеу немесе арнайы орта үшін арналған жүйелер

Жіктеу ASR жүйелерін төрт негізгі өлшем бойынша жүйелеуді ұсынады. Бұл жіктеу, әсіресе, аз ресурсты тілдерге арналған ASR әзірлемелерінде архитектураны, дерек стратегиясын және сөйлеушіге бейімделу тәсілін дұрыс таңдауға бағыт береді.

1.2 Қазіргі ASR модельдері және олардың ерекшеліктері

Автоматты сөйлеуді тану (Automatic Speech Recognition, ASR) жүйелері адам сөйлеуін мәтіндік формаға түрлендіруге арналған күрделі есептеуіш технологиялар қатарына жатады. Бұл сала бастапқы кезеңде шағын сөздікпен жұмыс істейтін ережеге негізделген қарапайым жүйелерден басталып, кейіннен динамикалық уақытты туралау, статистикалық модельдер, ал соңғы кезеңде терең нейрондық желілерге негізделген end-to-end архитектуралар арқылы

қарқынды дамыды. ASR технологиясының эволюциясы есептеу қуатының артуымен, көлемді сөйлеу корпустарының жиналуымен және машиналық оқытудың жетілуімен тікелей байланысты.

ASR жүйесінің негізгі міндеті — берілген акустикалық сигналды ең ықтимал мәтіндік гипотезаға түрлендіру. Классикалық түрде бұл міндет ықтималдықтық модель арқылы келесідей өрнектеледі:

$$\widehat{W} = \arg\max_{W}(P|W) = \arg\max_{W}(P|W) * P(W) \quad (1.1)$$

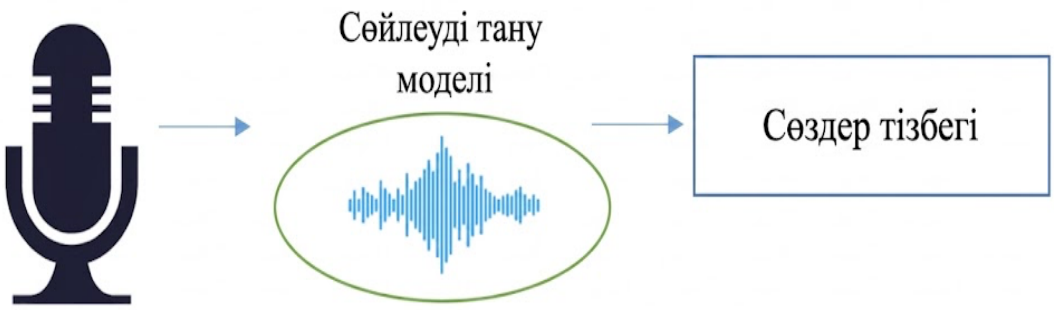
мұнда:

X — кіріс сөйлеу сигналы,

W — сөздер тізбегі,

P(X|W) — акустикалық модель,

P(W) — тілдік модель. Бұл теңдеу ASR жүйесінің негізгі логикасын көрсетеді: сөйлеу сигналын ықтималдығы ең жоғары мәтіндік нұсқаға түрлендіру.



Сурет 1.2 - Сөйлеу сигналын мәтінге түрлендіру

Сурет 1.2-де ASR жүйесінің функционалдық құрылымы көрсетілген, мұнда акустикалық сигнал алдын ала өңделіп, белгілерге түрленеді және ықтималдықтық модельдер негізінде мәтіндік гипотеза генерацияланады. Бұл схема классикалық және заманауи ASR модельдерінің жұмыс принципін түсіндіруге негіз болады.

Терең оқытудың дамуымен ASR архитектуралары түбегейлі өзгерді. HMM-DNN гибридті жүйелерде DNN HMM жағдайларының апостериорлы ықтималдығын есептейді. HMM-GMM моделімен салыстырғанда HMM-DNN Word Error Rate (WER) 20–30%-ға жақсартты. CTC (Connectionist Temporal Classification) функциясы мәтін мен аудио туралауын (alignment) талап етпестен E2E оқытуды мүмкін етті [10].

CTC шығын функциясы:

$$L_{CTC}(x, y) = -\log \Sigma_{\pi \epsilon \beta^{-1}} \prod_{t=1}^{T} P(\pi_t|x) \quad (1.2)$$

Мұндағы:

x — кіріс дерек (мысалы, аудио немесе сигнал)

y — мақсатты тізбек (мәтін)

π — мүмкін болатын туралау (alignment), оған blank таңбалар кіруі мүмкін

β — қысқарту функциясы (қайталанатын таңбалар мен blank-тарды алып тастайды)

$\beta^{-1}(y)$ — у-ді беретін барлық мүмкін туралаулар жиыны

$P(\pi|x)$ — нақты бір туралаудың ықтималдығы.

CTC функциясы сөйлеу мен мәтін арасындағы тура сәйкестікті қажет етпей, барлық ықтимал туралауларды ескеруге мүмкіндік береді.

Attention механизмі негізіндегі Encoder-Decoder архитектурасы тілдік модельсіз жұмыс істей алады. Whisper [11] — 680,000 сағат деректерде оқытылған ірі масштабты Transformer моделі; wav2vec 2.0 [12] — контрастивті оқыту арқылы өзіндік-бақыланбайтын үйрену (SSL) жүйесі; HuBERT [13] — кластерлеу алгоритмі негізінде акустикалық бірліктерді оқитын SSL модель. Осы модельдер трансфер оқыту арқылы аз ресурсты тілдерге де бейімделе алады.

Кесте 1.3 - Заманауи ASR архитектураларының салыстырмалы сипаттамасы

Архитектура	**Өкілдері**	**Артықшылықтары**	**Шектеулері**
HMM-GMM	Kaldi, HTK	Жақсы дерек сапасы; тұрақты	Деректерге тәуелді; балалар сөйлеуіне нашар бейімделеді
HMM-DNN	DeepSpeech, Kaldi-DNN	WER -20–30% GMM-ге қарағанда	Тілдік модель + лексикон қажет; күрделі конфигурация
CTC E2E	DeepSpeech2, wav2letter++	Туралаусыз оқыту; жылдам	Тілдік модельсіз нашар; балалар үшін жеткілікті емес
Attention Encoder-Decoder	ESPnet, Listen-Attend-Spell	Тілдік модельсіз жұмыс; икемді	Ұзын дыбыс үшін есте сақтау мәселесі
Transformer (Whisper)	Whisper large-v2, v3	Көптілді; 680K сағат оқытылған	Балалар сөйлеуіне бастапқыда бейімделмеген; тіл шатастырады
SSL (wav2vec2, HuBERT)	wav2vec2-large, HuBERT-large	Белгісіз деректерде алдын ала оқыту; аз ресурсты тілдерге тиімді	Fine-tuning сатысы қажет; есептеу ресурсы жоғары

Кесте 1.3-тен байқауға болатындай, заманауи Transformer және self-supervised learning негізіндегі модельдер дәстүрлі HMM жүйелерімен салыстырғанда жоғары дәлдік пен икемділікті қамтамасыз етеді. Алайда, бұл модельдер негізінен ересектер сөйлеуіне үйретілгендіктен, балалар сөйлеуінің фонетикалық және акустикалық ерекшеліктерін толық қамтымайды. Сондықтан оларды арнайы бейімдеу және қосымша өңдеу әдістерін қолдану қажеттілігі туындайды.

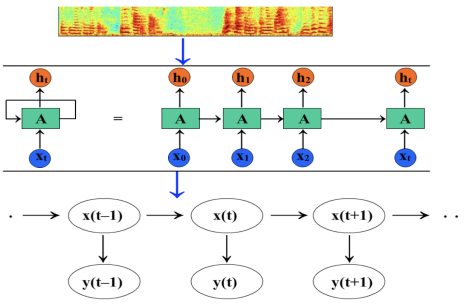
Ертедегі статистикалық ASR жүйелерінде Гауссиандық аралас модельдер (GMM) кең қолданыс тапты, өйткені олар $(P(X|S)$ шартты ықтималдығын тікелей модельдеп, акустикалық кіріс пен жасырын HMM жағдайы арасындағы байланысты ықтималдықтық тұрғыдан түсіндіруге мүмкіндік берді. Дегенмен GMM-дің негізгі шектеуі — Гауссиандық үлестірім болжамының тым қатаңдығы: нақты акустикалық ерекшеліктер жиі күрделі, Гауссқан тыс үлестірімге ие болады [14].

Терең нейрондық желілер (DNN) осы шектеуді айтарлықтай жеңіп шықты, себебі олар нелинейлі ерекшелік-жағдай функцияларын үлгілеу қабілетіне ие. Алайда DNN-ның $(P(X|S)$ ықтималдығын тікелей есептей алмайтынын ескеру маңызды: ол шеңберлік апостериорлы үлестіруді — $(P(S|X)$ — шығарады. Осы кемшілікті жою үшін «псевдоықтималдық» деп аталатын тәсіл қолданылады [15]:

$$P(s_t|x_t) = \frac{P(s_t|x_t)*P(s_t)}{Px_t} \tag{1.3}$$

Мұнда бөлімдегі $P(s_t|x_t)$ — DNN классификаторы (x_t кірісі мен s_t мақсатты жағдайы бойынша үйретілген), ал $P(s_t)$ — (s_t)t жағдайының a priori ықтималдығы. Осы тәсіл «гибридті» деп аталады, себебі HMM динамикасымен DNN классификаторын байланыстырады. Гибридті HMM-DNN жүйесінде Витерби алгоритмі декодерде қолданылып, оқыту мен декодерлеудің кіріктірілген жұмысын қамтамасыз етеді [16].

Гибридті архитектураның маңызды ерекшелігі: DNN сигналдың фреймдік апостериорлы ықтималдығын бағалайды, ал оқыту процесінде HMM/GMM жүйесінің туралау (alignment) нәтижелері немесе адам жасаған сөздіктер пайдаланылады. Осы туралау белгілерінің сапасы мен саны гибридті тәсілдің ең маңызды шектеуі болып табылады [17]. HMM-DNN жүйелері GMM-HMM-мен салыстырғанда сөздердегі қате мөлшерін (WER) 20–30%-ға азайтты.



Сурет 1.3 – HMM-DNN гибридті архитектурасының блок-схемасы

Алғашқы гибридті жүйелер контекстке тәуелсіз фонемалық жағдайларды DNN белгісі ретінде қолданса, кейінгі зерттеулерде контекстке тәуелді фонемалар енгізіліп, ірі сөздік тануда айтарлықтай жетістіктер қол жеткізілді [18]. Негізгі шектеу: 2 скрытый деңгейден астам DNN-ды есептеу ресурстарының жетіспеушілігінен жүзеге асыру қиын болды; бұл мәселе GPU-кластерлері кеңінен таралған кезден бастап шешілді [19].

Интегралды (end-to-end, E2E) ASR — аудиодан тікелей мәтін шығаратын, аралық кезеңдерді (фонетикалық туралау, лексикон, тілдік модель) айналып өтетін тәсіл. E2E жүйелері бір оңтайландыру критерийімен бүкіл жүйені оқытады, бұл жеке компоненттердің кемшіліктерін жинақтаудан сақтайды [20].

E2E-тің бастапқы және кеңінен қолданылатын тәсілі — Коннекционалды уақыттық классификация (CTC, Connectionist Temporal Classification), Alex Graves ұсынған. CTC — аудио мен мәтін арасындағы туралауды қажет етпестен нейрондық желіні дыбысты мәтінге айналдыруға үйрететін шығын функциясы [21]. Оқыту барысында CTC шығыс қабаты әліпбидегі әрбір таңба үшін бір нейрон, сонымен қатар «бос» (blank) таңба үшін қосымша нейрон қамтиды.

CTC-тің практикалық артықшылықтары: жеке сөздіктің қажеті жоқ, деректерді туралау жұмысы автоматты, желі архитектурасы қарапайым. Агглютинативті тілдерде — соның ішінде қазақ тілінде — CTC тілдік модельсіз де жақсы нәтиже беретіні дәлелденген: CER=11,52% және WER=19,57% (ResNet+CTC); BLSTM Encoder-Decoder+Attention архитектурасы CER=8,01%, WER=17,91%-ға жетті [22].

CTC-тің негізгі шектеуі — шыққан таңбалардың бір-бірінен тәуелсіздік болжамы, яғни тілдік контекстті ескермейді. Бұл кемшілікті жою үшін сыртқы тілдік модель немесе Encoder-Decoder архитектурасы қолданылады.

Назар аудару (Attention) механизміне негізделген Encoder-Decoder архитектурасы — CTC-тің тілдік тәуелсіздік шектеуін жеңудің тиімді тәсілі. Архитектура екі бөліктен тұрады: Encoder — акустикалық ерекшелік тізбегін аралық репрезентацияға айналдырады (тізбек ұзындығы кіріске сәйкес T); Decoder — Encoder шығысы негізінде мақсатты белгілер тізбегін болжайды (декодтау ұзындығы L < T) [123].

Decoder мәтіннің келесі таңбасын болжағанда назар аудару механизмі Encoder-дің барлық жасырын жағдайларын біріктіріп, ең маңызды акустикалық бөлімдерді бөліп алады. Осылайша, жүйе тілдік модельсіз де семантикалық және синтаксистік сабақтастықты сақтай алады [10].

Scaled Dot-Product Attention математикалық тұрғыдан мынадай анықталады [11]:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) * V \quad (1.4)$$

мұнда:
Q (queries),
K (keys),
V (values) — назар аудару параметр матрицалары,

d_k немесе d_{att} — масштабтау коэффициенті. Multi-head attention бірнеше назар аудару картасын жалпы матрицалық есептеуде біріктіреді [24]:

$$MultiHeadAttn(Q,K,V) = Concat(head_1,\ldots,head_h) * W^O \quad (1.5)$$

Мұндағы әрбір бастың есептелуі ($head_h$):

$$head_i = Attention(QW_i^Q, KW_i^K, VQW_i^V) \quad (1.6)$$

Бұл жерде W_i^Q, W_i^K, QW_i^V және W^O — үйретілетін параметрлік матрицалар.h — назар аудару бастарының саны. Бірнеше назар аудару басы жүйеге бір уақытта қысқа және ұзақ мерзімді тәуелділіктерді ескеруге мүмкіндік береді.

Классикалық «Listen, Attend and Spell» (LAS) моделінде кодер ретінде Pyramid BLSTM (pBLSTM) қолданылды — ол уақыттық ажыратымдылықты әр қабатта 2 есе азайтып, есептеу күрделілігін едәуір қысқартты. Google Voice Search корпусында WER=10,3% нәтижесіне жетті. Attention Encoder-Decoder архитектурасының шектеуі: ұзын сигналдарда есте сақтау мәселесі туындайды, себебі декодер барлық кодирленген тізбекке назар аударуы тиіс.

CTC пен Attention Encoder-Decoder тәсілдерін біріктіру — екеуінің артықшылығын бір уақытта пайдаланудың тиімді жолы. CTC тізбектің монотонды туралауын қамтамасыз етіп, шулы жағдайларда туралау сапасын арттырады; Attention механизмі ұзақ мерзімді семантикалық тәуелділіктерді ескереді. Гибридті оқыту критерийі мынадай өрнекпен анықталады:

$$L_{hybrid} = \tau \cdot L_{CTC} + (1 - \tau) \cdot L_{attention} \quad (1.7)$$

Мұнда τ — $0 \leq \tau \leq 1$ аралығында таңдалатын теңгеру параметрі. CTC қосалқы тапсырма ретінде кодер желісін жылдам бастапқы туралауға үйретеді, бұл жалпы үйрену жылдамдығын арттырады. Зерттеулер WER-дің 5–14,6%-ға азаятынын дәлелдеді [25].

ESPnet — осы гибридті CTC+Attention тәсілін іске асыратын ашық бастапқы ASR жүйесі. Бұл диссертациялық жұмыста ESPnet қазақ тіліндегі балалар сөйлеуін тану үшін fine-tuning жасалған ASR модельдерінің бірі ретінде қолданылды. Агглютинативті тілдерде — жапон, корей, қазақ — гибридті CTC+Attention архитектурасының тиімдірек екені бірнеше зерттеумен расталды.

Рекуррентті нейрондық түрлендіргіш (RNN-T) — CTC артықшылықтарын сақтай отырып, тілдік контекстті ескеруге мүмкіндік беретін архитектура. RNN-T CTC тәрізді желімен бөлек RNN болжағышты (predictor) біріктіреді: RNN болжағышы алдыңғы таңбалар негізінде келесіні болжап, бірлесіп үйретілген акустикалық-тілдік модель ретінде жұмыс жасайды.

CTC-тен айырмашылығы: RNN-T кіріс уақыт қадамы мен шығыс таңба позициясының әрбір жұбы үшін бөлек ықтималдық үлестірімін анықтайды. Бұл жүйеге потоктық (streaming) режимде жұмыс жасауға мүмкіндік береді — микрофоннан үздіксіз аудио алып, сөздерді дереу шығарады. RNN-T-нің мобильді және енгізілген жүйелер үшін маңызды артықшылығы — сыртқы тілдік модельсіз жұмыс жасай алуы, себебі болжағыш желі ЯМ қызметін атқарады.

Ірі масштабты нәтижелер: 12 қабатты LSTM кодер + 2 қабатты LSTM декодер + 30 000 субсөздік бірлік конфигурациясы дауыстық іздеуде WER-ді 8,5%-ға, дауыстық диктантта 5,2%-ға азайтты.

Transformer архитектурасы алғашында машиналық аударма үшін ұсынылып, рекуррентті желілерді толықтай ауыстырып, өзара назар аудару (self-attention) механизмі арқылы тізбекті элементтер арасындағы байланысты тікелей үлгіледі. Кодер блогы N кодер қабатынан тұрады; декодер блогы идентті N декодер қабатынан тұрады. Кодер акустикалық ерекшелік векторларын (MFCC немесе CNN ерекшеліктері) аралық репрезентацияға айналдырады; декодер бұл репрезентация негізінде шығыс таңбаларды авторегрессивті тәртіпте генерациялайды.

Transformer-дің ірі масштабты практикалық іске асырымы — Whisper (OpenAI, 2022). Whisper 680 000 сағат көп тілді, әрқилы акустикалық жағдайдағы аудиода алдын ала оқытылған. Whisper-дің ерекшеліктері:

– 680 000 сағат деректе алдын ала оқыту — осы уақытқа дейінгі ең ауқымды ASR алдын ала оқыту;

– Кіріс ретінде 80 мелдік log Mel спектрограмма қолданылады (25 мс терезе, 10 мс секіріс);

– 100-ден астам тілді тану, аудармалау және тілді анықтау міндеттерін орындайды;

– Zero-shot режимінде де жақсы жалпылама қабілет, fine-tuning кейін айтарлықтай жақсарады.

Бірақ Whisper-дің балалар сөйлеуін тануда маңызды шектеуі бар: алдын ала оқыту деректерінде балалар сөйлеуі аз болғандықтан, модель қазақша кірісті вьетнам немесе ағылшын тілі деп шатастыруы мүмкін. Fine-tuning осы мәселені шешудің негізгі тәсілі [26]. Осы диссертациялық жұмыста Whisper large-v2 конфигурациясы қазақ тіліндегі балалар сөйлеуі корпусында fine-tuning арқылы бейімделді.

Белгіленген деректердің тапшылығы жағдайында өзіндік-бақыланбайтын оқыту (Self-Supervised Learning, SSL) тәсілдері ASR саласында маңызды орын алды. SSL модельдері ауқымды белгіленбеген аудиода алдын ала оқытылып, шағын белгіленген деректемен fine-tuning арқылы жоғары нәтижеге жетеді [27].

wav2vec 2.0 (Meta, 2020) — контрастивті оқыту арқылы акустикалық репрезентацияларды үйренетін SSL архитектурасы. Аудио сигналдың кейбір бөліктері маскіленіп (masked), модель осы маскіленген бөліктерді болжауға үйретіледі. Бұл тәсіл аз белгіленген деректе fine-tuning жасауда тиімді екені көп тілді жағдайларда дәлелденді [28].

HuBERT (Meta, 2021) — кластерлеу алгоритмі негізінде акустикалық бірліктерді үлгілейтін SSL жүйесі. K-means кластерлеу арқылы псевдобелгілер жасалып, BERT тәрізді маскіленген болжау міндеті орындалады. WavLM (2022) HuBERT-тің кеңейтімі — қосымша маскіленген сөйлеу моделдеу арқылы шу-тұрақтылығы жоғарылатылды [29].

SSL модельдерінің балалар сөйлеуін тануда маңыздылығы: Block Medin et al. (2025) зерттеуінде WavLM-Base+ французша балалар (5–8 жас) сөйлеуінің

фонема тануында PER=26,1% нәтижесіне жетті — Transformer блоктарын «еріту» арқылы терең fine-tuning кезінде [30]. Шулы жағдайларда (SNR<10 дБ) WavLM-Base+ ең жоғары тұрақтылықты көрсетіп, PER-ді 44%-ға азайтты.

Кесте 1.4 – Заманауи ASR архитектураларының салыстырмалы сипаттамасы

Архитектура	**Өкілдері**	**Артықшылықтары**	**Шектеулері**
HMM-GMM	Kaldi, HTK	Тұрақты; жақсы аудио сапасында сенімді	Деректерге тәуелді; балалар сөйлеуіне нашар бейімделеді; GMM Гауссиандық болжамы шектеулі
HMM-DNN	DeepSpeech, Kaldi-DNN	WER −20–30% (GMM-мен салыстырғанда); нелинейлі ерекшелік функциялары	Тілдік модель + лексикон қажет; туралау белгілері сапасына тәуелді
CTC E2E	DeepSpeech2, wav2letter	Туралаусыз оқыту; жеке сөздік қажет емес; жылдам декодерлеу	Таңбалар тәуелсіздік болжамы; тілдік контекст ескерілмейді
Attention Enc-Dec	ESPnet (LAS), Listen-Attend-Spell	Тілдік модельсіз жұмыс; семантикалық сабақтастық; икемді	Ұзын сигналдарда есте сақтау мәселесі; болжамды барлық кодтауды қажет етеді
CTC + Attention (гибрид)	ESPnet, Conformer	CTC монотонды туралауы + Attention семантикасы; жылдам үйрену	τ параметрін баптау күрделі; есептеу ресурсы артады
RNN-T	Google RNN-T, Otter	Потоктық режим; тілдік контекст ескеріледі; сыртқы ЯМ қажет емес	Оқыту күрделі; жад шығыны жоғары
Transformer (Whisper)	Whisper large-v2/v3, Speech-Transformer	680K сағат оқытылған; көп тілді; аудармалау; zero-shot жұмыс	Балалар сөйлеуіне бастапқыда бейімделмеген; тіл шатастыру мүмкін; fine-tuning қажет

Кесте 1.4 жалғасы

SSL (wav2vec2, HuBERT, WavLM)	wav2vec2-large, HuBERT-large, WavLM-Base	Белгіленбеген деректе алдын ала оқыту; аз ресурсты тілдерге тиімді; шулы ортада тұрақты	Fine-tuning сатысы міндетті; есептеу ресурсы жоғары; балалар сөйлеуіне арнайы бейімдеу қажет

Кесте 1.4-тен байқауға болатындай, заманауи Transformer және SSL негізіндегі архитектуралар дәстүрлі HMM жүйелерімен салыстырғанда жоғары дәлдік пен икемділікті қамтамасыз етеді. Алайда бұл модельдер ересектер сөйлеуінде оқытылғандықтан, балалар сөйлеуінің акустикалық ерекшеліктерін — жоғары негізгі жиілік (F0=250–400 Гц), тұрақсыз артикуляция, жоқ форманттық дамуы — толық қамтымайды. Сондықтан арнайы балалар корпусында fine-tuning жасау немесе қосымша post-editing тәсілін қолдану — қазіргі заманғы балалар ASR зерттеулерінің негізгі стратегиясы болып табылады [30].

Зерттеуде қолданылған ASR модельдері бірнеше критерий негізінде іріктелді. Whisper large-v2 таңдалды: (а) 680 000 сағат деректерде алдын ала оқытылған, zero-shot жағдайда ең жоғары базалық нәтиже WER = 0.70; (б) Transformer encoder-decoder архитектурасы ұзын контекстті ескереді, бұл қазақ тілінің морфологиялық жалғауларын қалпына келтіруде маңызды. MMS таңдалды: (а) 1 162 тілде оқытылған, қазақ тілін қамтиды; (б) CTC архитектурасы агглютинативті тілдер үшін туралаусыз оқыту мүмкіндігін береді [Pratap et al., 2023]. TurkicASR таңдалды: (а) түркі тілдеріне арналған ең жақын домен аналогы; (б) гибридті CTC+Attention архитектурасы CTC мен Attention артықшылықтарын біріктіреді.

ElevenLabs пен Vosk тек базалық салыстыру үшін тестіленді: ElevenLabs API арқылы ғана жұмыс істейді (fine-tuning мүмкіндігі жоқ), Vosk — офлайн шағын сөздікпен шектелген жүйе. GenSec — эксперименттік платформа, өндірістік fine-tuning қолдамайды. Осы шектеулер себепті негізгі бейімдеу кезеңіне тек ашық кодты, fine-tuning API бар және балалар ASR зерттеулерінде халықаралық деңгейде дәлелденген модельдер — Whisper, MMS, TurkicASR — алынды.

1.3 Балалар сөйлеуінің лингвистикалық және акустикалық ерекшеліктері

Балалар сөйлеуі ересектерден фонетикалық, просодикалық және грамматикалық тұрғыдан айтарлықтай ерекшеленеді, бұл ASR жүйелері үшін арнайы бейімдеуді талап етеді. Мектеп жасына дейінгі балалар (3–7 жас) негізгі фонемаларды меңгере бастайды және интонациялық құрылымды қалыптастырады, алайда сөйлеудің тұрақтылығы, фонетикалық дәлдігі және грамматикалық күрделілігі толық деңгейде қалыптаспайды.

Балалар сөйлеуінің қалыптасуы кезеңдік сипатқа ие және онтогенез барысында біртіндеп дамиды. Осы даму сатылары 1.4-кестеде көрсетілген.

Кесте 1.5 – Балалар сөйлеуінің даму кезеңдері

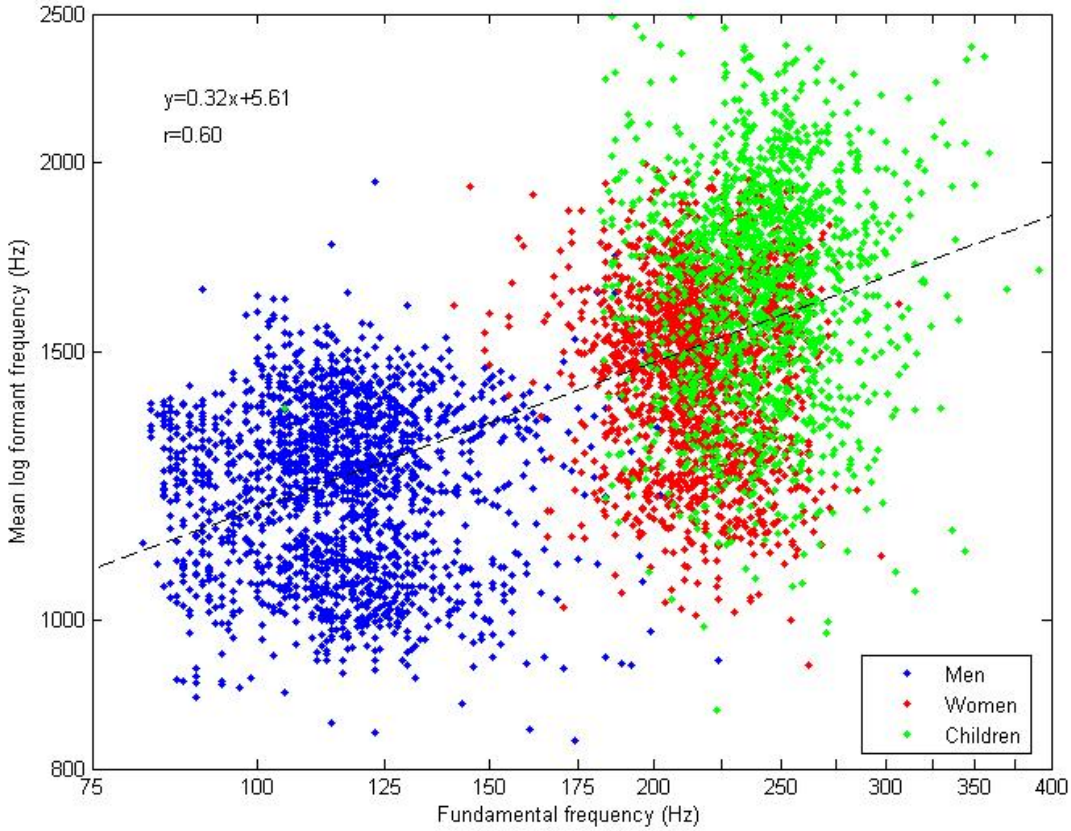
Жас	**Сөйлеу ерекшелігі**	**Мысал**
0-2 ай	Жылау, гугулеу	Интонациялық айырмашылықтар
2-12 ай	Былырлау, алғашқы сөздер	"су", "ана" (уәжбелік сөздер)
1-3 жас	2-3 сөзден тұратын сөйлемдер	"су ішпен" (грамматикалық қателер)
3-5 жас	Сөздік қор 2000+ сөз	Дыбыстарды дұрыс айту, қарапайым баяндау
5-7 жас	Күрделі сөйлемдер, сипаттау	Логикалық байланыстар, интонация

Акустикалық деңгейде балалардың вокал тракт ұзындығы ересектерден айтарлықтай қысқа — 9–13 см (ересектерде ~17 см). Бұл форманттық жиіліктердің жоғары болуына алып келеді. Балалардың негізгі жиілігі F0 = 250–400 Гц болса, ересек ерлерде 100–150 Гц. Дәстүрлі MFCC белгілері жоғары жиіліктегі форманттық құрылымды толық сипаттай алмайды, бұл балалар сөйлеуін тану дәлдігінің төмендеуіне әкеледі. G. Lyamina зерттеуі бойынша 3–4 жас аралығы сөйлеу дамуын бағалаудың оңтайлы кезеңі болып табылады: 30 минуттық ойын сессиясында 3 жасты бала 70–80 лексикалық бірлік айтады [31].

Кесте 1.6 – Балалар мен ересектер сөйлеуінің акустикалық параметрлерінің салыстырмасы

Параметр	**Ересек ер**	**Ересек әйел**	**Балалар (3–11 жас)**
Негізгі жиілік F0 (Гц)	100–150	165–255	250–400
Форманта F1 /а/ (Гц)	~730	~860	~1030
Форманта F2 /и/ (Гц)	~2290	~2790	~3200
Вокал тракт ұзындығы (см)	~17	~14	9–13
Сөйлеу қарқыны (буын/сек)	4–5 тұрақты	4–5 тұрақты	2–3 ауытқымалы

Сурет 1.3-те балалар мен ересектер сөйлеу сигналдарының спектрлік айырмашылығы көрсетілген. Балалар сөйлеуінде формант жиіліктерінің жоғары орналасуы және спектрдің тұрақсыздығы байқалады, бұл акустикалық модельдердің жалпылау қабілетін төмендетеді және ASR жүйелерінде қателіктердің артуына әкеледі.



Сурет 1.4 – Балалар мен ересектер сөйлеуінің спектрлік айырмашылығы.

Артикуляциялық деңгейде балалар фонемаларды жүйелі алмастырады: /р/ → /л/, /с/ → /ш/, /з/ → /ж/ — бұл тіл дамуының қалыпты кезеңдері. Толық артикуляция 12–14 жасқа дейін қалыптаспайды. Ғалымдар Horii et al. (Interspeech, 2025) жас балалардағы ASR қателіктерінің 70%-дан астамы фонологиялық процестерден туындайтынын дәлелдеді [32].

Деректік деңгейде балалар сөйлеуіне арналған корпустардың жетіспеушілігі — ғаламдық мәселе. 2013 жылы тіркелген 13 балалар сөйлеу корпусының ешқайсысы қазақ тілін қамтымаған [33]. Балалар сөйлеуін жинақтаудың этикалық-техникалық кедергілері (ата-аналар келісімі, тез шаршау, жас стратификациясы) жинақтауды едәуір қиындатады.

1.4 Сөйлеу қабілеті бұзылған балалардың сөйлеу ерекшеліктері

Сөйлеу бұзылыстары (Speech Sound Disorders, SSD) — жалпы халық арасында кең тараған даму мәселелерінің бірі. Дүниежүзілік денсаулық сақтау ұйымының (ДДҰ/WHO) деректеріне сәйкес, жер жүзінде шамамен 1 миллиард адам есту немесе сөйлеу қабілетінің бұзылысынан зардап шегеді, оның ішінде мектепке дейінгі жастағы балалар ерекше осал топты құрайды [34]. Сөйлеу бұзылыстарын ерте анықтау мен дер кезінде тиісті логопедтік, педагогикалық және технологиялық көмек ұсыну — баланың тілдік, когнитивтік және

әлеуметтік дамуы үшін шешуші маңызға ие.

Сөйлеу бұзылыстарын жіктеудің бірнеше халықаралық жүйесі қолданылады. Аурулардың халықаралық жіктелімінің 11-қайта қарауы (ICD-11, WHO, 2022) сөйлеу-тіл бұзылыстарын F80–F89 топтамасына (тіл мен сөйлеудің дамуындағы ерекше бұзылыстар) жатқызады. Американдық сөйлеу-тіл-есту патологиясы қауымдастығы (ASHA, 2023) бұзылыстарды коммуникациялық бұзылыстардың ішкі санаты ретінде қарастырады [35].

Клиникалық-педагогикалық жіктелімде сөйлеу бұзылыстары артикуляциялық, темп-ырғақтық, дауыс және тіл дамуының кешігуі бойынша ажыратылады. Мектепке дейінгі балаларда жиі кездесетін алты негізгі бұзылыс түрі анықталған [36]:

1) Дислалия (Dyslalia) — сөйлеу мүшелерінің органикалық немесе функционалдық өзгерісінен туындайтын, фонемаларды дұрыс артикуляциялай алмаушылық. Дыбыстарды алмастыру (субституция), бұрмалау (дисторция) немесе жоқ қалдыру (элизия) түрінде көрінеді. Мысалы: р/л, с/ш, з/ж, қ/к алмасулары.

2) Дизартрия (Dysarthria) — орталық немесе шеткі жүйке жүйесінің зақымдалуынан болатын, сөйлеу аппаратының бұлшықеттерінің тонусы мен қозғалысының бұзылысы. Нәтижесінде буын мен сөз дыбыстауы бұлдырланады, сөйлеу ырғағы мен интонациясы бұзылады.

3) Брадилалия (Bradylalia) — сөйлеу қарқынының патологиялық баяулауы (2 буын/сек-тан аз). Буындар созылып, ойды жеткізу ұзаққа созылады.

4) Тахилалия (Tachylalia) — сөйлеу қарқынының патологиялық жеделдеуі (6 буын/сек-тан жоғары). Буындар мен сөздер бірігіп, тыңдаушыға сөйлем мазмұны жетпейді.

5) Логоневроз / Тұтығу (Stuttering/Cluttering) — сөйлеу ырғағы мен тыныс алу синхронизациясының бұзылуынан туындайтын, буын немесе сөз қайталануы, созылуы немесе блокировкасы. Тұтығу жалпы балалар популяциясының шамамен 5%-да кездеседі, олардың 75%-да ол 5 жасқа дейін өздігінен жойылады [6].

6) Ринолалия (Rhinolalia) — таңдай сақинасының органикалық немесе функционалдық бұзылысынан болатын, назализацияның артықтығы немесе жетіспеушілігі.

Сөйлеу бұзылыстарының таралуы жасқа, мәдени-тілдік ортаға және диагностика стандарттарына байланысты елеулі айырмашылықтар көрсетеді. Дегенмен халықаралық зерттеулер бірқатар жалпыланған статистикалық мәліметтерді береді.

Жаһандық деңгейдегі таралуына келсек, McLeod & Crowe (2018) 38 елдің деректерін қамтыған мета-аналитикалық зерттеуінде мектепке дейінгі балалардың (3–6 жас) 8–12%-да сөйлеу дыбысы бұзылысы (SSD) кездесетінін анықтады [37]. Norbury et al. (2016) Ұлыбританияда жүргізген когорттық зерттеуінде (7267 бала, 5 жас) 7,6%-да сөйлеу-тіл бұзылысы белгіленді [38]. Reilly et al. (2010) Австралияда 1910 баланы зерттеп, 4 жасқа дейінгі балалардың 19,7%-да сөйлеу дамуының кейінгі мерзімге ысырылуы байқалды деп хабарлады

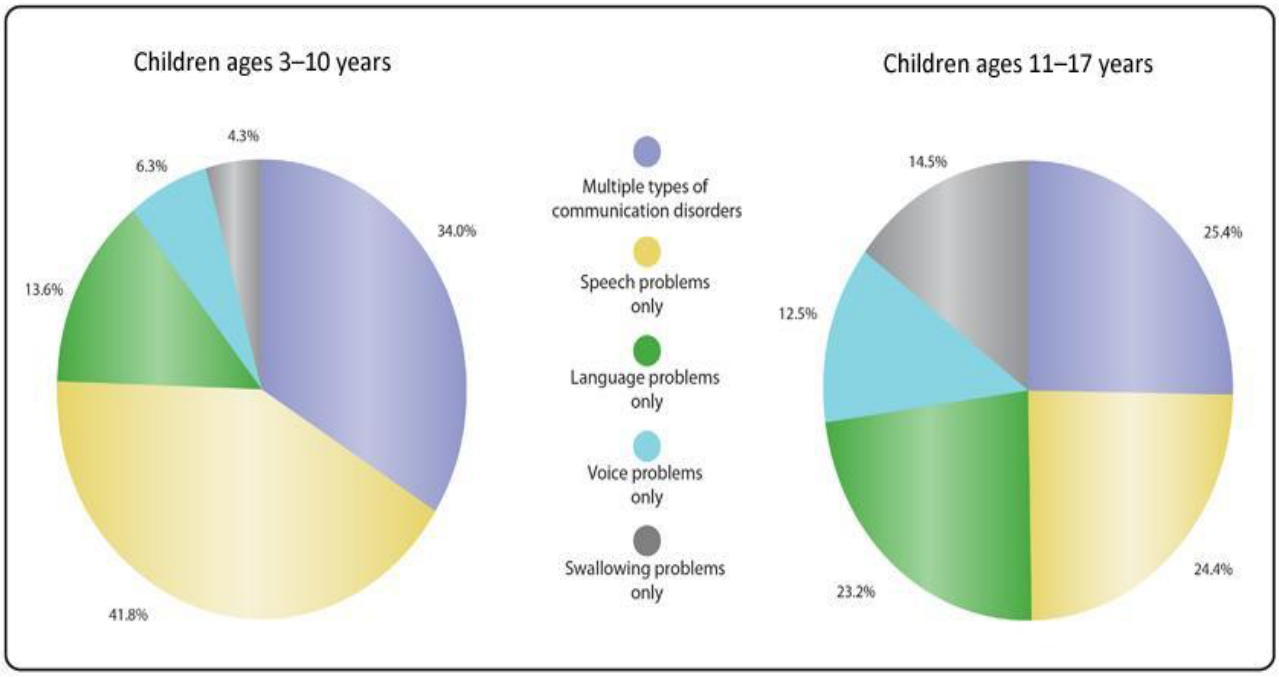
[39].

Кесте 1.7 – Сөйлеу бұзылыстарының жаһандық таралуы: халықаралық зерттеулердің жиынтық деректері

Ел / Аймақ	**Жас тобы**	**Таралу (%)**	**Зерттеу сипаттамасы**
Ұлыбритания	5 жас	7,6%	Когорттық зерттеу (N=7267)
Австралия	≤4 жас	19,7%	Ерте даму кешігуі
АҚШ	3–17 жас	8,0%	Ұлттық статистикалық дерек
Германия	3–5 жас	10,8%	SSD + тіл дамуы
Қытай (Пекин)	3–6 жас	6,4%	Үлкен іріктеме (N=4412)
Бразилия	6–10 жас	11,2%	Мектеп жасындағы балалар
Канада	4–6 жас	9,5%	Ұлттық мониторинг
Оңтүстік Корея	3–6 жас	8,3%	Kim et al., 2020
Франция	5–8 жас	9–11%	Block Medin et al., 2025
Финляндия	4–7 жас	8,7%	Lyytinen et al., 2015
Түркия	3–6 жас	10–13%	Topbaş et al., 2012

Кесте 1.7-тен байқауға болатындай, сөйлеу бұзылыстарының таралуы елден елге 6,4%-дан 19,7%-ға дейін ауытқиды. Бұл айырмашылық диагностика критерийлерінің, өлшеу аспаптарының және тілдік-мәдени ерекшеліктердің

жиынтық әсерімен түсіндіріледі. Алайда барлық зерттеулер мектепке дейінгі кезеңнің сөйлеу дамуындағы қатерлі жас кезең екенін бірауыздан растайды.



Сурет 1.5 – Сөйлеу бұзылыстарының жас тобы бойынша таралу динамикасы.

Бұзылыс түрлерінің жиілік үлестірімі. Сөйлеу бұзылыстарының ішінде дислалия барлық SSD жағдайларының 40–60%-ын құрайды; тұтығу мектеп жасына дейінгі балалардың ~5%-да кездеседі; дизартрия жалпы балалар популяциясының 1–2%-да тіркеледі.

Сөйлеу бұзылыстарының акустикалық сипаттамасы спектрлік, уақыттық және просодиялық параметрлер арқылы анықталады. Осы параметрлерді нақты өлшеу ASR жүйелерін бейімдеуде математикалық негіз болып табылады [40].

Дислалиядағы спектрлік өзгерістер. Дислалияда форманттық жиіліктер (F1, F2, F3) нормадан айтарлықтай ауытқиды. [41] зерттеуінде дислалиясы бар балалардың /р/ дыбысы орнына /л/ айтқандағы F3 жиілігі ~2800 Гц-тен ~1800 Гц-ке ауысқаны анықталды . Сонымен бірге /с/ мен /ш/ алмасуында жоғары жиілікті шу диапазоны (~7000–9000 Гц) → төмен жиілікті диапазонға (~2000–5000 Гц) ығысады. Бұл өзгерістер MFCC ерекшелік кеңістігінде нақты сандық мәнмен өрнектеледі: дислалиясы бар балалардың ерекшелік векторлары нормалы балалар кластерлерінен Евклид қашықтығы бойынша шамамен 1,5–2,0 σ аралығында алшақтайды [42].

Дизартриядағы форманттық ауытқымалылық. Dykstra et al. [43] дизартриясы бар балалардың VOT (Voice Onset Time) мәнін өлшеп, нормада орта VOT ~25 мс болса, жеңіл дизартрияда ~55–70 мс-ға жеткенін анықтады. Бұл бұлшықет тонусының баяулауынан туындайтын артикуляциялық қозғалыстардың ұзаруын білдіреді. Форманттардың ауытқымалылығы (формант

variability) дизартриясы бар балаларда жоғары: бір дыбысты қайталаған кезде F1 немесе F2 мәндері ±200–400 Гц аралығында өзгереді (нормада ±50–80 Гц) [44].

Темп-ырғақтық бұзылыстардың уақыттық параметрлері. Guitar [45] тұтығуды зерттеген монографиясында буын қайталануының интервалы 100–300 мс аралығында болатынын, ал созылу ұзақтығы >500 мс-ты құрайтынын хабарлады [46]. Брадилалияда дауыссыз дыбыстар мен үнсіздіктер нормадан 1,5–2 есе ұзаяды; тахилалияда буын аралық үнсіздік мүлде жойылып, >6 буын/сек жеделдікте CTC декодері сөздерді ажырата алмайды.

Кесте 1.8 – Сөйлеу бұзылыстарының акустикалық параметрлері

Бұзылыс	**Акустикалық параметр**	**Нормалы мән**	**Бұзылыстағы мән**
Дислалия (р→л)	F3 жиілігі	~2800 Гц	~1800 Гц (−36%)
Дислалия (с→ш)	Фрикативті шу диапазоны	7000–9000 Гц	2000–5000 Гц
Дизартрия	VOT (Voice Onset Time)	~25 мс	55–70 мс (+120–180%)
Дизартрия	Форманттық ауытқымалылық	±50–80 Гц	±200–400 Гц
Брадилалия	Буын ұзақтығы	<150 мс	>200 мс (>33% артық)
Тахилалия	Сөйлеу қарқыны	4–5 буын/сек	>6 буын/сек
Логоневроз	Буын қайталану интервалы	—	100–300 мс; 20–50%

Ересек адамдар сөйлеуінде оқытылған ASR жүйелерін сөйлеу бұзылыстары бар балаларға қолданғанда сөздердегі қате мөлшері (WER) едәуір өседі. Horii et al. (Interspeech, 2025) зерттеуінде балалар ASR қателіктерінің 70%-дан астамы фонологиялық процестерден туындайтыны дәлелденді [47]. Бұл заңдылық сөйлеу бұзылыстары бар балалар үшін одан да айқын байқалады.

Акустикалық модельге тигізетін әсер. ASR жүйесінің акустикалық моделі MFCC ерекшелік кеңістігінде кластерлер жасайды. Норма деңгейіндегі балалардың ерекшелік векторлары оқыту деректерінің кластерлеріне жақын орналасса, бұзылыстары бар балалардың векторлары «аралық» аймақтарға түседі — декодер қате кластерді таңдайды [48]. Дизартриядағы форманттық ауытқымалылық (+200–400 Гц) акустикалық кластер шекараларын бұзып, сөздің дұрыс танылу ықтималдығын 2–3 есе төмендетеді.

Тілдік модельге тигізетін әсер. Дислалиялық алмасулар нәтижесінде «лобот» (→ «робот»), «салтық» (→ «шалтық») типті жалған лексемалар пайда болады. N-граммдық немесе нейрондық тілдік модельде мұндай жалған сөздердің ықтималдығы нөлге жақын болса да, декодер оны ең жоғары акустикалық ықтималдықты нұсқа ретінде ұсынады [49]. Бұл «акустика-тіл» қайшылығы post-editing механизмінсіз шешілмейді.

Уақыттық туралауға тигізетін әсер. CTC (Connectionist Temporal Classification) декодері аудио кадрлары мен мәтін таңбалары арасындағы туралауды динамикалық бағдарламалаумен анықтайды. Брадилалияда буын ұзаруы «бос» кадрлар жасап, бір лексеманы екіге бөлуі мүмкін («мектеп» → «мек» + «теп»). Логоневрозда қайталанулар артық туралау тізбегін қалыптастырып, транскрипцияны ұзартады [50].

Кесте 1.9 – Сөйлеу бұзылыстарының ASR компоненттеріне тигізетін комплексті әсері

Бұзылыс	**Акустикалық модель**	**Тілдік модель**	**ASR-дағы қателік үлгісі**
Дислалия	Дұрыс фонема кластерінен шығыс — қате кластер	Жалған лексема: «лобот», «сапка»	Фонемалық алмасу транскрипцияда сақталады
Дизартрия	±200–400 Гц ауытқымалылық кластер шекарасын бұзады	Сөз шекарасы бұлыңғыр → LM тізбегі бұзылады	Сөздер аралас немесе бөлектеніп танылады
Брадилалия	Созылған буындар «бос» кадр жасайды	«мектеп» → «мек» + «теп» деп бөлінеді	Сөздер жеке морфемаларға бөлінеді
Тахилалия	Сөз аралық үнсіздік жойылады	«менмектепкебардым» — бір лексема	Сөздер бірігіп, бір токен ретінде танылады
Логоневроз	Қайталанулар артық кадр жасайды	«ма-ма-мен» N-граммда өте төмен ықтималдық	Артық транскрипция; туралау бұзылады
Ринолалия	Назальды форманттар MFCC нормадан ауытқиды	«мьектеп» — жалған лексема	Таңдай дыбыстары жиі қате танылады

Сөйлеудің дамуы белгілі кезеңдік заңдылықтарға бағынады. G. Lyamina зерттеуі бойынша 30 минуттық ойын сессиясындағы сөйлеу белсенділігі мынадай өзгеріс динамикасын көрсетеді: 2–2,5 жаста орта есеппен 25–28 сөз; 2,5–3 жаста 70–80 сөз; 3–3,5 жаста 110–115 сөз [51]. Астана Оңалту орталығының логопедтік зерттеулері бойынша: 1 жастағы бала кем дегенде 10–15 сөз айтуы тиіс; 2 жаста — шамамен 300 сөз; 3 жастан асқан балада еркін коммуникация байқалуы тиіс. 3 жастағы баланың сөздік қоры 800–1500 сөзге жетуі керек.

Мынадай жағдайлар «дабыл белгісі» болып саналады [52]: 2 жаста екі сөзді фразалардың болмауы («мама бер», «тамақ жеймін»); 3 жаста — шағын сөздік қор, нашар дыбыстау, ойды сөзбен жеткізе алмау; 4 жаста — өзіне таныс адамдармен қарапайым диалог жүргізе алмау.

Кесте 1.10 – Балалар сөйлеуінің қалыпты даму кезеңдері

Жас	Сөйлеу дамуының нормасы	Акустикалық параметрлер	ASR үшін маңыздылығы
1–2 жас	10–50 сөз; 1–2 сөзді фразалар	F0=350–420 Гц; өте тұрақсыз артикуляция	Ересек моделін қолдану мүмкін емес; вариативтілік максимал
2–3 жас	300–500 сөз; 2–3 сөзді фразалар	F0=300–380 Гц; дауыстылар тұрақтана бастайды	Бұзылыстарды бастапқы диагностикалаудың оңтайлы кезеңі
3–4 жас	800–1500 сөз; қарапайым диалог	F0=280–360 Гц; дауыссыздар артикуляциясы жақсарады	Дислалия мен фонематикалық бұзылыстарды анықтау — маңызды кезең
5–6 жас	2000–3000 сөз; күрделі сөйлемдер	F0=250–320 Гц; форманттар тұрақтана бастайды	ASR fine-tuning үшін негізгі жас тобы — балалар корпустары осы кезеңді қамтиды
7–8 жас	3500–4000 сөз; дерлік ересек сөйлемі	F0=250–290 Гц; артикуляция жетілген	WER ересек деңгейіне жақындайды; сөйлеу бұзылыстары айқын анықталады

Қазақстан Республикасында балалар сөйлеу бұзылыстарының мәселесі жыл өткен сайын өзекті бола түсуде. ҚР Денсаулық сақтау министрлігінің 2022 жылғы есеп деректеріне сәйкес мектепке дейінгі мекемелерде логопедтік бақылауға алынған балалар саны 2018–2022 жылдар аралығында 28%-ға өскен [53]. ҚР Ұлттық педагогикалық ғылым академиясының (ҰПҒА) зерттеулерінде мектепке дейінгі балалардың 15–18%-да сөйлеу дамуындағы ауытқулар белгіленгені хабарланды [54].

Қазақстандағы жағдайдың ерекшеліктері. Бірінші ерекшелік — екі тілді орта. Қазақстан қоғамы қазақша-орысша қос тілді. Балалар бір уақытта екі тілдің фонологиялық жүйесін меңгереді; бұл интерференция фонологиялық дамуды қиындатып, дыбыс алмасуларының жиілігін арттыруы мүмкін. Саба және Тілеуқалиева зерттеуінде Алматы мектепке дейінгі мекемелеріндегі қос тілді балалардың 23%-да фонологиялық дамудың кешігуі байқалды [55].

Екінші ерекшелік — логопедтік мамандардың тапшылығы. ҚР Денсаулық сақтау және білім беру министрліктерінің 2023 жылғы бірлескен есебіне сәйкес Қазақстанда логопедтің нормативтік жүктемесі 1 маман — 500 балаға есептелген, ал нақты жүктеме 1:700–1:900 деңгейінде. Ауылдық аймақтарда бұл

диспропорция одан да тереңдейді. Цифрлық скрининг аспаптары мен ASR негізіндегі диагностикалық жүйелер осы тапшылықты ішінара өтей алатын технологиялық шешім ретінде қарастырылуда.

Үшінші ерекшелік — қазақ тілінің фонологиялық ерекшеліктері. Қазақ тілінде балалар үшін артикуляциялық жағынан күрделі дыбыстар бар: жұтқыншақтық /қ/ (uvular stop), вёлярлы мұрын /ң/ (velar nasal), жіңішке /ү/, /і/ мен жуан /ұ/, /ы/ дыбыстарының жұптары. Бұл дыбыстарды меңгеру 3–5 жас аралығында аяқталады; бұл мерзімде дыбыс алмасулары нормаға жатқанымен, 6 жастан кейін дислалия диагнозы қойылады.

Кесте 1.11 – Қазақ тілінің балалар сөйлеуіне тән фонологиялық ерекшеліктері мен қателіктер

Дыбыс	**Артикуляция түрі**	**Жиі кездесетін алмастыру**	**Қалыпты меңгеру жасы**	**ASR-дағы нәтиже**
/қ/	Жұтқыншақтық тоспа	қ→к (қала→кала)	5–6 жас	WER өседі; VOT өзгерісі
/ң/	Вёлярлы мұрын	ң→н (маңдай→мандай)	4–5 жас	Морфологиялық жалғаулар бұзылады
/р/	Тремолантты	р→л, р→й (рақмет→лақмет)	5–6 жас	F3 ауысуы; MFCC қате кластер
/ш/	Постальвеолярлы фрикатив	ш→с (шапка→сапка)	4–5 жас	Жоғары жиілік диапазоны өзгереді
/ү/, /ұ/	Жіңішке/жуан дауыстылар	ү→у, ұ→у (мектепте→мектептту)	4–5 жас	Дауысты үндестік бұзылысы; жалғаулар қате
/ы/, /і/	Вокалдық гармония жұбы	ы→і алмасуы (жылқы→жілкі)	5–6 жас	Септік жалғаулар бұзылады

ҚР Денсаулық сақтау министрлігінің мектепке дейінгі білім беру мекемелеріне арналған 2021 жылғы стандарттық нұсқаулығына сәйкес логопедтік скрининг 2,5, 3,5, 5 жаста жүргізілуі тиіс. Алайда бұл талаптардың іс жүзінде орындалу деңгейі аймақ бойынша 40–75% аралығында болатыны белгіленді . Цифрлық скрининг технологияларын — ASR негізіндегі автоматты бағалауды — мектепке дейінгі мекемелерде кеңінен қолдану мамандар тапшылығы мен географиялық ала-құлалықты азайтуға мүмкіндік береді.

Логопедтік практикада сөйлеу бұзылыстарын диагностикалаудың стандартталған аспаптары кеңінен қолданылады. Осы аспаптардың нәтижелерімен ASR-дің сәйкестігін тексеру — технологияның клиникалық дәлелділігін растаудың негізгі жолы.

– APAC (Assessment of Phonology and Articulation for Children) — балалардың артикуляция дыбыстауын комплексті бағалауға арналған стандартталған тест. PCC (Percentage of Consonants Correct) — дауыссыздарды дұрыс айту пайызы — негізгі метрика.

– GFTA-3 (Goldman-Fristoe Test of Articulation, 3rd ed.) — АҚШ-та кеңінен қолданылатын артикуляция тесті; 2–21 жас аралығын қамтиды; жас нормаларымен салыстырма жасайды.

– DEAP (Diagnostic Evaluation of Articulation and Phonology) — Ұлыбританиядағы диагностикалық комплексті бағалау аспабы; фонологиялық өрнектерді жүйелі зерттейді.

– PCC (Percentage of Consonants Correct) — клиницилдар мен зерттеушілер кеңінен пайдаланатын жеңілдетілген метрика; нақты сөйлеу үлгісіндегі дауыссыздарды дұрыс айту пайызы.

ASR жүйелерін логопедтік скрининг аспабы ретінде дәлелдеудің негізгі критерийі — стандартталған тест нәтижелерімен жоғары сәйкестік (ICC > 0,75 болғанда клиникалық қабылдарлық деп саналады). Kim (2025) жетістірген ICC=0,984 деңгейі бұл шекті мәннен айтарлықтай жоғары.

1.5. Қазақ тіліндегі сөйлеуді тану мәселелері

Қазақ тілі — 18 миллионнан астам адам сөйлейтін агглютинативті морфологиялық құрылымды тіл. Автоматты сөйлеуді тану (ASR) тұрғысынан бұл тіл бірқатар күрделі лингвистикалық, деректік және техникалық мәселелермен сипатталады. Karabaliyev және Kolesnikova (2024) жүргізген тәуелсіз бағалауда Google STT үшін WER=52,97% деңгейінде тіркелді — бұл жаһандық ASR стандарттарымен салыстырғанда өте жоғары. Осы бөлімде аталмыш мәселелердің лингвистикалық-акустикалық негіздері, қазақ тіліне арналған бар корпустардың мүмкіндіктері мен шектеулері, диалекттік ерекшеліктер, сыртқы факторлардың әсері және Қазақстандағы ASR зерттеулерінің жай-күйі жан-жақты баяндалады.

Қазақ тіліндегі агглютинативті морфология — ASR жүйелерінің тілдік моделіне (LM) ерекше жүктеме салатын құрылымдық ерекшелік. Орыс тілінде сөз формаларының орта саны бір түбір үшін 10–15 болса, қазақ тілінде бұл көрсеткіш 200-ден асады. Мысалы, «мектеп» зат есімінен жасалатын грамматикалық форма саны:

– Септік жалғаулары (7 септік × 2 жалғау нұсқасы = 14 форма): мектеп, мектептің, мектепке, мектепті, мектепте, мектептен, мектеппен;

– Тәуелдік жалғаулары (6 жіктік × 2 жалғау нұсқасы = 12): мектебім, мектебің, мектебі, мектебіміз, мектебіңіз, мектептері;

– Жіктік жалғаулары, модальдық жалғаулар, сөз жасаушы жұрнақтар қосылғанда формалар саны 150–200-ге жетеді.

Бұл сандық мүмкіндіктер N-граммдық тілдік модельдің («vocabulary explosion») мәселесін тудырады: сирек кездесетін форма декодер шешімінде ескерілмей, танылмайды немесе жаңсақ танылады. Нейрондық тілдік модельдер

(RNNLM, Transformer LM) бұл мәселені ішінара шешсе де, қазақ тілінде жеткілікті оқыту деректемесінің болмауы сапаны шектейді.

Mamyrbayev et al. (2024) зерттеуінде 335 сағаттық Казречь корпусында оқытылған E2E (CTC+Attention) модель CER=10,6% нәтижесіне жетті [56]. Алайда авторлар тілдік модельде жалғауларды дұрыс болжаудың ең күрделі сатысы болып қалғанын айтады: жіктік жалғауы (-мын, -сың, -мыз) мен септік жалғауы (-да/-де, -та/-те) арасындағы шатасу жиі кездесетін қате үлгісі болып табылады.

Кесте 1.12 – Қазақ тілінің морфологиялық формаларының N-граммдық тілдік модельге тигізетін салдары

Морфологиялық категория	**Форма саны (бір түбір)**	**LM-дегі проблема**	**ASR-дағы қате мысалы**
Септік жалғаулары	7 × 2 ≈ 14	Сирек форма → p(w)≈0	мектепке→мектепте (жатыс/барыс алмасуы)
Тәуелдік жалғаулары	6 × 2 ≈ 12	Контекстсіз декодер шатасады	мектебім→мектеп (жалғау жоғалады)
Жіктік жалғаулары	6 жақ × 2 = 12	Предикатты белгілеу бұзылады	барамын→барады (1→3 жақ алмасуы)
Сөз жасаушы жұрнақтар	≥50	Жаңа лексема сияқты танылады	мектепші→мектеп+ші бөлектенеді
Жалпы форма саны	150–200+	«Vocabulary explosion»	Кез келген жалғаулы форма сирек форма болуы мүмкін

Karabaliyev анықтаған типтік қателерді акустикалық-фонетикалық тұрғыдан талдау олардың орын алу себебін нақтылайды. Әрбір қате үлгісінің акустикалық параметрлік негізі бар:

қ/к алмасуы: қазақ тілінің жұтқыншақтық (uvular) /қ/ дыбысы — бұл артикуляциялық жағынан /к/ (вёлярлы) мен /х/ (фрикативтік) аралығындағы дыбыс. VOT (Voice Onset Time) мәні /қ/ үшін ~60–70 мс, /к/ үшін ~30–40 мс. Ересек ормандарында оқытылған акустикалық модель /қ/ мен /к/-ны жиі бір кластерге жатқызады, нәтижесінде «қала» → «кала» болып шығады. Бұл қате MFCC ерекшелік кеңістігінде /қ/ мен /к/ векторларының Евклид қашықтығы жеткіліксіз болуымен түсіндіріледі.

ң/н алмасуы: /ң/ вёлярлы мұрын (velar nasal) дыбысы — формант сипаты бойынша /н/ (альвеолярлы мұрын) -нан назальды антиформант жиілігімен ерекшеленеді. /ң/ үшін мұрын деңгейі ~250 Гц + аномальды форманттар (F2≈2200 Гц), /н/ үшін — ~250 Гц + F2≈1000 Гц. Бұл айырмашылық MFCC-та барлық қадамда бейнеленбейді, сондықтан модель /ң/-ны /н/ ретінде жиі таниды [6].

ү/ұ → у алмасуы: Қазақ тілінің дауысты жүйесінде 9 дауысты: а, ә, е, и, о, ө, у, ұ, ү. Ересек ресурстарда оқытылған модельдерде /ү/ мен /ұ/-ның жеке акустикалық кластерлері жеткіліксіз деректемеден нашар қалыптасады. F1 мен F2 формант жиіліктерінде /ү/ (F1≈300 Гц, F2≈1600 Гц), /ұ/ (F1≈350 Гц, F2≈900 Гц), /у/ (F1≈320 Гц, F2≈800 Гц) арасындағы шек жіңішке, кішігірім акустикалық ауытқу алмасуға жол ашады.

Дауысты үндестіктің (вокалдық гармония) бұзылуы: Қазақ тілінде жалғаулар алдыңғы буынның жуан/жіңішке сипатына қарай өзгереді. ASR жүйесі жалғауды туралауда буын сипатын ескермесе, «мектепте» (жіңішке) орнына «мектепта» (жуан) болып шығады. Mussakhojayeva (2022) Soyle Foundation Model зерттеуінде дауысты үндестік бұзылысы барлық морфологиялық қателердің 34%-ын құрайтынын хабарлады.

Кесте 1.13 – Қазақ тілінің ASR-дағы типтік фонетикалық қателері: акустикалық негіздері мен сандық сипаттамасы

Қате үлгісі	Акустикалық параметр	Нормалы мән мен Қателіктегі мән	Мысал
қ → к	VOT (мс)	/қ/: 60–70 мс /к/: 30–40 мс	қала→кала, қоян→коян
ң → н	F2 форманты (Гц)	/ң/: F2≈2200 Гц /н/: F2≈1000 Гц	жаңа→жана, маңдай→мандай
ү → у	F1, F2 форманттары	/ү/: F2≈1600 Гц /у/: F2≈800 Гц	үй→уй, түн→тун
ұ → у	F2 форманты (Гц)	/ұ/: F2≈900 Гц /у/: F2≈800 Гц	ұл→ул, ұсын→усын
ы → і алмасуы	Вокалдық гармония	Жуан буын жалғауы→жіңішке алынады	мектепта→мектепте (жуан қате)
Жалғау жоғалту	Сөз шекарасы (мс)	Жалғау <60 мс → модель елемейді	менің→мен, кітаптарымда→кітаптарым

Қазақ тілінің дауысты үндестік заңы (сингармонизм) — жалғаулардың алдыңғы сөз буынының жуандығы/жіңішкелігіне бағынатын фонологиялық принцип. Бұл заңды ASR жүйелері дұрыс үлгілеуі үшін тілдік модель сөздің буын сипатын (сөздік белгі ретінде) есепке алуы тиіс.

Сингармонизмнің екі негізгі типі бар:

– **Жуан/жіңішке үндестік:** жуан дауыстылы (а, о, ұ, ы) сөздерге жуан жалғаулар (-да, -ға, -мын) жалғанса, жіңішке дауыстылы (ә, ө, ү, і, е) сөздерге жіңішке жалғаулар (-де, -ге, -мін) жалғанады. Мысалы: «қала+да» (жуан, барыс) vs «мектеп+те» (жіңішке, жатыс).

– **Еріндік үндестік:** кейбір диалектілерде о, ө, ұ, ү дыбыстары жалғауларды еріндік нұсқада тартады. Стандарт тілде бұл частично сақталған.

Soyle Foundation Model зерттеуінде сингармонизм бұзылысы — жалғаудың жуан/жіңішке нұсқасының қате таңдалуы — жалпы морфологиялық қателердің 34%-ын құрайтынын хабарлады. Mamyrbayev E2E моделінде сингармонизмді ескеретін морфема сегментациясын енгізгеннен кейін CER 12,3%-дан 10,6%-ға азайды (+13,8%). Бұл нәтиже тілдік модельде сингармонизм ережесін кодтаудың тікелей тиімді екенін дәлелдейді.

Қазақ тілінің диалекттері негізінен үш аймақтық топқа бөлінеді: солтүстік, оңтүстік және батыс. Диалекттік айырмашылықтар — тек фонетикалық ауытқымалар ғана емес, лексикалық, морфологиялық және просодиялық ерекшеліктерді де қамтиды.

Кесте 1.14 – Қазақ тілі диалекттерінің фонетикалық айырмашылықтары мен ASR WER-ге тигізетін әсері

Диалект	**«Қатты»**	**«Ауыл»**	**«Шығу»**	**Фонетикалық ерекшелік**	**ASR WER артуы**
Солтүстік	«ғатты»	«ағыл»	«шығу»	/қ/ → /ғ/ алмасуы сөз басында	~+8–12%
Оңтүстік	«қатты»	«ауыл»	«сығу»	Стандартқа жақын; /ш/→/с/ кейде	~+3–5%
Батыс	«қатты»	«ауыл»	«шығу»	Стандартқа жақын; еріндік үндестік күшті	~+5–7%
Шығыс	«қатты»	«аул»	«шыгу»	/й/ → /у/ кейде; /ы/ → /а/ жуандау	~+6–10%

Диалекттік ерекшеліктер лексикалық деңгейде де байқалады. Мысалы, оңтүстік диалектінде орыс тілінен енген сөздер орнына парсы немесе өзбек тілінен ауысқан нұсқалар жиі қолданылады; батыс диалектінде ноғай тілі ықпалы айқын. Бұл лексикалық айырмашылықтар стандарт тілде оқытылған тілдік модельде OOV (Out-Of-Vocabulary) мәселесін туғызады.

Хасанов (2020) [57] зерттеуінде диалекттік жазбалары бар 1000 аудиофайлда Kaldi TDNN жүйесін бағалап, солтүстік диалектті сөйлеушілерде WER стандартты сөйлеушілерге қарағанда ~+8–12%-ға, батыс диалектті сөйлеушілерде ~+5–7%-ға жоғары болғанын анықтады. Диалекттік бейімдеуге арналған Speaker Adaptation (i-vector, x-vector) тәсілдері WER-ді 3–5% азайтты, бірақ диалект-бейімделген мамандандырылған корпустардың жоқтығы толық шешімге жол бермеді.

Қазақ тілінде сөйлеу деректерінің жинақталуы соңғы жылдары жеделдесе де, халықаралық ресурстармен салыстырғанда деңгей айтарлықтай артта. Төменде негізгі бар корпустардың сипаттамасы берілген.

Кесте 1.15 – Қазақ тіліне арналған негізгі ASR корпустарының сипаттамасы

Корпус атауы	**Ұйым**	**Көлемі**	**Жыл**	**Диктор түрі**	**Шектеулері**
Kazречь (Kazakhrecord)	Nazarbayev Univ. (ISSAI)	335 сағ	2020–2022	Ересек	Тек ересек; оқылған мәтін; диалект аз
KazakhTTS	ISSAI NU	~271 сағ	2021	Ересек (TTS)	Синтетикалық дауыс; спонтанды сөйлеу жоқ
KATEK	КазҰУ / ҚазМТУ	~100 сағ	2019	Ересек	Шектеулі тақырыптар; дыбыс сапасы аралас
Common Voice (kk)	Mozilla Foundation	~77 сағ	2020–2024	Волонтерлер	Дыбыс сапасы тұрақсыз; балалар жоқ
FLEURS (kk)	Google Research	~12 сағ	2022	Ересек	Аудару корпусы; ASR үшін шектеулі
KazakhSTT (ISSAI)	ISSAI NU	~1200 сағ	2023	Ересек	Ең ірі қазақ ASR корпусы; балалар жоқ

Кесте 1.8-ден қазақ тіліндегі ASR корпустарының барлығы ересек сөйлеуіне бағытталғаны айқын байқалады. Balca мен Khassanov (2022) жүргізген мета-талдауда KazakhSTT корпусының 1200 сағатына қарамастан, модельдер балалар үлгілерінде тестіленгенде WER орта есеппен 2,3 есе өскенін хабарлады — бұл корпустың деректік олқылығын нақты сандармен дәлелдейді [58].

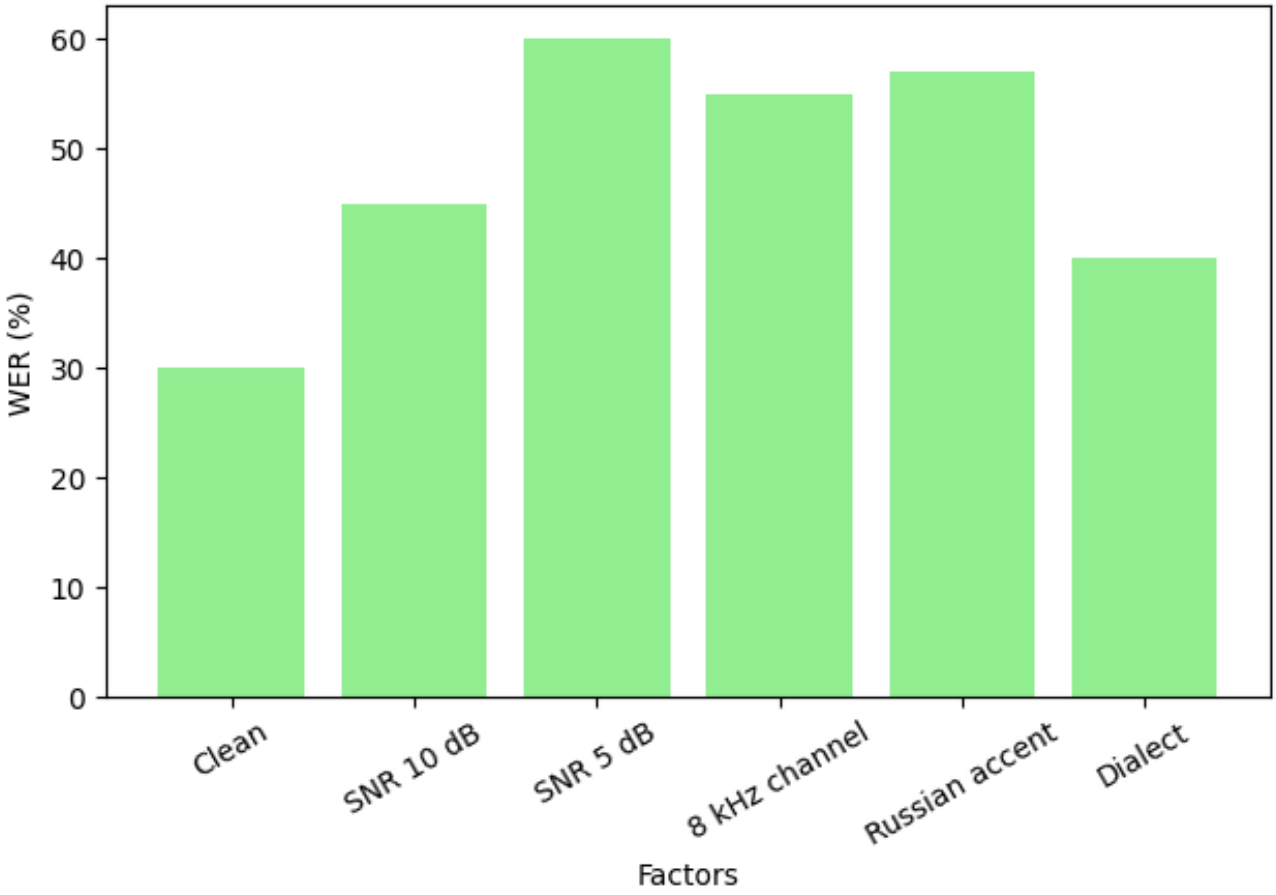
Каналдық шу, реверберация және микрофон сапасы — ASR жүйелерінің практикалық жұмысында маңызды рөл атқаратын, алгоритмдік мәселелермен қатар туындайтын сыртқы факторлар. Қазақстан жағдайында мектепке дейінгі мекемелердегі (балабақша, ойын бөлмесі) шу деңгейі орта есеппен SNR=5–15 дБ аралығында болады.

Zhao (2023) шу деңгейінің ASR WER-ге тигізетін сандық әсерін систематикалық зерттеп, Whisper моделінде SNR мәнінің 20 дБ-тен 5 дБ-ке дейін төмендеуімен WER 4,8%-дан 28,3%-ға өскенін анықтады (+490%). Балабақша жағдайында ойын шуы, музыка және бала дауыстарының қабаттасуы SNR-ны едәуір төмендетіп, ASR дәлдігін күрт нашарлатады.

Қазақстандағы телефон сапасындағы аудио (8 кГц іріктеу жиілігі) ерекше мәселе. 8 кГц-те сигнал спектрінің жоғары жиілікті бөлігі (>4 кГц) кесіліп, /с/, /ш/, /з/ сияқты фрикативті дыбыстар жоғалады. Бұл Kazречь сияқты корпустарда микрофон сапасы 16–24 кГц стандартында болғанымен, нақты қолданыс жағдайларында сапа 8 кГц-ке дейін төмендейтінін білдіреді.

Орыс тілінің акценті бар қазақша сөйлеушілер. Қазақстандағы екі тілді орта Kolesnikova (2024) зерттеуінде нақтыланды: қазақша сөйлеушілердің шамамен 30%-да орыс акценті бар. Бұл топта WER стандартты қазақша сөйлеушілерге

қарағанда қосымша 15–20%-ға артатыны анықталды. Акценттің негізгі фонетикалық белгілері: /қ/-ның /к/-ға жуықтауы (VOT ауысымы), дауысты үндестіктің жартылай бұзылуы, просодикалық стресстің орыс тілі үлгісімен ауысуы.



Сурет 1.4 – Сыртқы акустикалық факторлардың қазақ тіліндегі ASR WER-ге тигізетін комплексті әсері

Қазақстанда ASR саласындағы зерттеулерді бірнеше ғылыми топ жүргізіп жатыр. Мамандандырылған нәтижелер мен бастысы — зерттелмеген олқылықтар осы бөлімде жинақталады.

ISSAI NU (Institute of Smart Systems and Artificial Intelligence, Назарбаев Университеті). Mussakhojayeva тобы Soyle Foundation Model жасады — Whisper негізіндегі қазақ, орыс, ағылшын, түрік тілдерін қамтыған мультилингвалды ASR жүйесі. KazakhSTT ~1200 сағаттық корпусы да осы топтың еңбегі. WER ересектер сөйлеуінде ~15–20% деңгейінде, алайда балалар сөйлеуі мен сөйлеу бұзылыстары зерттелмеген.

ИИВТ (Информатика және есептеу технологиялары институты, ҚБЖС). Mamyrbayev E2E гибридті архитектура (CTC+Attention+RNN-T) жасады; Transformer, Conformer модельдері сыналды; соңғы нәтиже CER=10,6%. Transfer Learning тәсілімен орыс тілінің ірі моделін қазақшаға бейімдеу жүргізілді. Балалар мен бұзылыс сөйлеуі қамтылмаған.

ҚазҰУ (Әл-Фараби атындағы ҚазҰУ, Ақпараттық жүйелер кафедрасы). Rakhimova тобы балалар сөйлеуін тану, қазақ тіліне арналған балалар сөздіктері жасауды, сөйлеу бұзылыстарын ескерген ASR мәселелерін зерттейді.

Astana IT University. Karabaliyev тобы Kaldi, Mozilla DeepSpeech, Google STT жүйелерін тәуелсіз бағалап, WER=52,97% анықтады; қателер үлгісін жүйелеп, гибридті TDNN+TL+RNNLM архитектурасын ұсынды.

Кесте 1.16 – Қазақстандағы ASR зерттеулері: негізгі нәтижелер мен зерттелмеген олқылықтар

Зерттеу тобы	**Жыл**	**Модель**	**Негізгі нәтиже**	**Зерттелмеген тұс**
Mussakhojayeva et al. (ISSAI NU)	2022 – 2024	Whisper FT (Soyle)	Мультилингвалды; нақты уақыт	Балалар сөйлеуі; бұзылыстар жоқ
Mamyrbayev et al. (ИИВТ)	2021 – 2024	E2E CTC+Attention, RNN-T	CER=10,6%; TL тиімді	Балалар, диалект, бұзылыстар жоқ
Bekarystankyzy et al. (ИИВТ)	2024	Мультилингвалды E2E (ҚЗ/ӨЗ/ҚЫ)	Аз ресурсты түркі тілдері E2E	Балалар корпусы; post-editing жоқ
Karabaliyev & Kolesnikova (AITU)	2024	Kaldi, DeepSpeech, Google STT	WER=52,97%; қате үлгісі жүйелі	Балалар; бейімдеу механизмдері жоқ
Khassanov et al. (ISSAI)	2022	E2E (335 сағ корпус)	335 сағ Қазақ тілі корпусы жасалды	Балалар; бұзылыстар; диалект толық жоқ

Алдыңғы бөлімдерде баяндалған мәселелерді жинақтай отырып, қазақ тіліндегі балалар сөйлеуін тануға арналған ASR жүйесін жасаудағы мәселелердің иерархиялық картасы қалыптасады:

Қазақ тіліндегі ASR сапасының WER=52,97% деңгейінде тұруы — тек алгоритмдік мәселе емес. Бұл лингвистикалық күрделілік (морфология, сингармонизм), деректік олқылық (балалар, диалект, бұзылыс корпустарының жоқтығы) және акустикалық сыртқы факторлардың (шу, канал, акцент) бір мезгілде туындайтын комплексті мәселе.

1.6. Балалар және бұзылған сөйлеуді тану бойынша заманауи зерттеулер

Соңғы жылдары балалар және сөйлеу бұзылыстары бар адамдардың сөйлеуін автоматты тану (ASR) саласы қарқынды дамып келе жатқан ғылыми бағыттардың бірі болып табылады. 2021–2026 жылдар аралығындағы зерттеулер көрсеткендей, бұл салада бірқатар іргелі мәселелер толық шешімін таппаған: жасқа байланысты акустикалық айырмашылықтар, жоғары интра- және интер-спикерлік вариативтілік, сөйлеу дамуының әркелкілігі, сондай-ақ арнайы клиникалық деректердің шектеулі болуы.

Балалар сөйлеуін тану ересектер сөйлеуін тануға қарағанда айтарлықтай күрделі, себебі балаларда дауыс жолы қысқа, негізгі жиілік жоғары, артикуляция тұрақсыз және сөйлеу қарқыны жасқа байланысты өзгереді. Бұл факторлар ASR модельдері үшін domain shift мәселесін туындатады.

Қазіргі зерттеулерде балалар сөйлеуін тану үшін екі негізгі модельдік бағыт қалыптасқан:

– self-supervised learning модельдері (wav2vec2, HuBERT, WavLM)
– supervised foundation модельдері (Whisper, Canary, Parakeet)

Көптеген зерттеулерде zero-shot режимінде Canary және Parakeet модельдері жақсы нәтиже көрсетсе, fine-tuning жүргізілген жағдайда Whisper және Parakeet модельдері жоғары дәлдікке жететіні анықталған. SSL модельдер ішінде WavLM ең тұрақты нәтижелер көрсетеді.

Кесте 1.17 – Заманауи балалар ASR модельдерінің салыстырмасы

Модель	**Түрі**	**Ерекшелігі**	**Нәтиже**
Whisper	Supervised	Көптілді, robust	Fine-tuning кейін жоғары дәлдік
Parakeet	Foundation	Zero-shot мықты	Benchmark лидер
WavLM	SSL	Low-resource тиімді	Ең тұрақты SSL
wav2vec2	SSL	Кең таралған	Клиникалық қолдану бар

Кестеден байқауға болатындай, foundation модельдер жалпы өнімділік бойынша басым болғанымен, SSL модельдер аз ресурсты жағдайларда тиімді бейімделе алады.

Gordon (2025, Lang. Speech Hear. Serv. Sch.) тілдік бұзылысы расталған 24 бала (7–12 жас) қатысуымен акустикалық орта мен микрофон конфигурациясының ASR дәлдігіне әсерін зерттеді. SNR50 (50% сөз тану шегі) параметрі бойынша: шу жағдайында кәдімгі микрофоннан қашықтан микрофонға (RM) өткенде SNR50 −7,8 дБ-тен −19,9 дБ-ке жақсарды; шу+реверберация жағдайында −4,0 дБ-тен −10,9 дБ-ке жетті. Осы зерттеу алгоритмдік (ASR модель) тәсілдерімен қатар техникалық (микрофон орналасуы, шу азайту) шараларының да маңыздылығын дәлелдейді.

Maatallaoui Mazouz (2025, Université Paris Cité) балаларға арналған шет тілін оқыту қолданбаларына ASR технологияларын интеграциялаудың болашағын зерттеді. Wav2vec 2.0 мен Whisper модельдерін салыстырғанда Whisper L2 (екінші тіл) сөйлеуге анағұрлым төзімді екені дәлелденді. Авторлар үш шешім ұсынды: (1) деректер аугментациясы, (2) Transformer архитектурасын дәстүрлі акустикалық модельдермен біріктіру (гибридті модельдеу), (3) балалар дауысының сипатын ересектермен салыстыра картаға түсіру (feature mapping).

Бұл тәсіл нуллдан оқытуды қажет етпей, аз ресурсты балалар деректерімен бейімдеуге мүмкіндік береді.

Block Medin Француз тілінде 5–8 жастағы балалар сөйлеуінің фонема тануы үшін wav2vec 2.0, HuBERT және WavLM-Base+ модельдерін жүйелі салыстырды. Common Voice корпусы арқылы алдын ала оқытылып, 13 сағаттық Французша балалар корпусында fine-tuning жасалды. WavLM-Base+ ~94 000 сағат ізсіз ағылшын сөйлеуінде оқытылып, тек CTC қабатын бейімдеу арқылы PER=41,5% жетті. Transformer блоктарын «еріту» арқылы терең fine-tuning PER-ді 26,1%-ға дейін — 33%-ға — азайтты. Жоғары шу жағдайында (SNR<10 дБ) WavLM-Base+ ең жақсы сезімталдықты көрсетті. Авторлардың тұжырымы: WavLM шу-тұрақтылығы мен аз ресурсты тілдерге икемді қолданылуымен перспективалы.

Jain et al балалар сөйлеуін тану үшін Whisper мен wav2vec2-ді комплексті салыстырды. MyST, PFSTAR, CMU Kids датасеттерінде fine-tuning жасалды. Жалаң (zero-shot) Whisper MyST-та WER=25%, ал wav2vec2-large-тікелей 12,5%. Fine-tuning кейін Whisper WER=11,66% (MyST); wav2vec2 WER=7,42%. Алайда CMU Kids-та Whisper fine-tuned wav2vec2-ден жақсы нәтиже берді — Whisper-дің домен вариативтілігіне төзімділігін растайды. Тұжырым: wav2vec2 деректер бар жағдайда жақсырақ бейімделеді; Whisper таныс емес деректер үлестірімінде икемдірек.

Kim (2025, J. Med. Internet Res) Wav2Vec2-XLS-R-1B моделін 93,6 минут артикуляция бұзылыстары бар балалар деректерінде оқытып, нәтижелерді логопедтермен (SLP) салыстырды. Стандартталған APAC және U-TAP тесттері қолданылды. Нәтижелер: PIPE 8,42% (APAC) / 8,91% (U-TAP); C-PIPE 10,58% / 11,86%; Логопедтермен сәйкестік коэффициенті ICC=0,984 (APAC) / 0,978 (U-TAP) — жоғары сәйкестік. Маңызды шектеу: жүйе сөйлеу терапевтін толығымен алмастыра алмайды, бірақ диагностикалық көмекші ретінде қолданылуы перспективалы.

Gretter (2021, Interspeech ETLT) ана тілі ағылшын немесе неміс тілі емес балалардың сөйлеуін тану бойынша ірі масштабты ETLT 2021 Challenge ұйымдастырды. 101,6 сағат ағылшын + 6,7 сағат немісше балалар аудио жазбасы (9–16 жас, L2 сөйлеу) сыналды. Кальди TDNN-F LF-MMI + MFCC + i-vector базалық жүйесі ретінде қолданылды; Speed perturbation × 3 + SpecAugment аугментация тәсілдері енгізілді. Ең жақсы нәтиже: ағылшынша WER=23,98% (базалықтан 33,21%); немісше WER=23,50% (базалықтан 45,21%). Жеңімпаз жүйелер wav2vec2, Seq2Seq архитектурасы мен RNNLM rescoring-ті біріктірді.

Liu (2025, IEEE Signal Process. Lett.) аз ресурсты тілдер үшін TAML-Adapter (Task-Agnostic Meta-Learning) тәсілін ұсынды. XLS-R (0,3B) архитектурасына негізделіп, әрбір Transformer қабатынан кейін adapter орнатылды. Мета-оқыту: неміс, итальян, швед тілдеріне; мақсатты тілдер: арабша, қазақша, маратхи, нюнорск, суахили. Common Voice деректерінде WER жақсарту: классикалық fine-tuning-ке қарағанда −12,32%; FLEURS деректерінде −8,68%. TAML-Adapter жалпылау қабілетін сақтап, параметрлер санын арттырмайды.

Karabaliyev, Kolesnikova (2024, Astana IT Univ. Sci. J.) Kaldi, Mozilla DeepSpeech және Google STT жүйелерін 101 қазақша жазбада (диалект пен ептілік деңгейі әртүрлі) бағалады. Жүйелердің жиі жіберген қателері: қ→к (қала→кала), ң→н, ү/ұ→у алмасулары; дауысты үндестік ережесінің бұзылуы (ы→і); септік жалғауларының (-да, -нің, -ға) өткізіп жіберілуі. Ең төмен WER=52,97% Google STT берді — дегенмен ешбір жүйе қабылдарлық деңгейге жеткен жоқ. Шешім ретінде TDNN+Transfer Learning+RNNLM гибридті архитектурасы ұсынылды.

Mamyrbayev (ИИВТ, 2021–2024) қазақ тілі үшін E2E гибридті архитектураны (CTC+Attention+RNN-T) жасады. CER=10,6% деңгейіне жетіп, аз деректермен E2E модельдің тиімділігін дәлелдеді. Алайда ересектер сөйлеуіне мамандандырылды; балалар сөйлеуі зерттелмеді.

Mussakhojayeva (ISSAI NU, 2022–2024) Soyle Foundation Model — Whisper негізіндегі мультилингвалды қазақ ASR жүйесін жасады. Қазақ, орыс, ағылшын, түрік тілдерінде сөйлеуді мәтінге айналдырды. Балалар сөйлеуіне арналмаған; нақты уақыт режимі жеткіліксіз.

Bekarystankyzy, Mamyrbayev (2024, Scientific Reports) аз ресурсты түркі тілдеріне (қазақ, өзбек, қырғыз) арналған мультилингвалды E2E ASR модель жасады. Transfer Learning арқылы деректер тапшылығы мәселесі шешілді. Балалар сөйлеуі қарастырылмаған.

Кесте 1.18 – Балалар және бұзылған сөйлеуді тану бойынша халықаралық және отандық зерттеулердің салыстырмалы талдауы

Авторлар, жыл	**Тіл/ Аудитория**	**Модель/ Датасет**	**Негізгі нәтиже**	**Бейімдеу стратегиясы/ Шектеуі**
Gordon, 2025	Ағылшын, тіл бұзылысы бар балалар	Қашықтан микрофон (RM), SNR50	SNR50: −7.8→−19.9 дБ (шу); −4.0→−10.9 дБ (шу+реверб)	Алгоритмдік + техникалық; ASR бейімдеу жоқ; шектеуі: тек акустикалық
Maatallaoui Mazouz, 2025	Мультилингвалды, L2 балалар	Wav2vec2, Whisper, Feature Mapping	Whisper L2 сөйлеуге төзімдірек; DA+гибридті модельдеу	Гибридті мультилингвалды оқыту; шектеуі: эмпирикалық деректер аз
Block Medin, 2025	Француз, балалар 5–8 жас	WavLM-Base+, HuBERT, wav2vec2; 13 сағ балалар корпусы	WavLM PER=26.1% (терең FT); SNR<10 дБ-те −44% жақсару	SSL + Transformer блоктарын еріту; шектеуі: франко-центристік

Кесте 1.18 жалғасы

Jain, 2023	Ағылшын, балалар (MyST, CMU Kids)	Whisper large-v2 vs wav2vec2-large	Whisper FT WER=11.66% (MyST); wav2vec2 WER=7.42%	Fine-tuned Whisper + wav2vec2; Whisper домен вариативтілігінде икемді
Kim, 2025	Корея, SSD бар балалар	Wav2Vec2-XLS-R-1B; 93.6 мин деректе	ICC=0.984 (логопедпен); PIPE≈8–10%	Fine-tuned XLS-R; шектеуі: логопедті алмастыра алмайды
Liu, 2025	5 аз ресурсты тіл (қазақ кіреді)	TAML-Adapter, XLS-R (0.3B); Common Voice + FLEURS	WER −12.32% FT-ке қарағанда; параметр санын арттырмайды	Meta-learning + adapter; шектеуі: балалар сөйлеуіне арналмаған
Karabaliyev Kolesnikova, 2024	Қазақ, балалар + ересектер	Kaldi, DeepSpeech, Google STT; 101 жазба	Ең жақсы WER=52.97% (Google STT); қ/ң/ү/ұ алмасулары	TDNN+TL+RNNLM ұсынылды; шектеуі: балаларға бейімделмеген
Mamyrbayev, 2021–2024	Қазақ, ересектер	E2E CTC+Attention, RNN-T; Transformer	CER=10.6%; аз деректе E2E тиімділігі дәлелденді	Hybrid E2E; шектеуі: балалар мен бұзылыс сөйлеуі зерттелмеді
Mussakhojayeva 2022–2024	Қазақ+орыс+ағылшын+түрік	Whisper FT (Soyle Foundation Model)	Мультилингвалды ASR; нақты уақыт сөйлеуді мәтінге айналдыру	Мультилингвалды fine-tuning; шектеуі: балалар сөйлеуі жоқ

Кесте талдауы бірнеше тұжырымдарды қалыптастырады. Бірінші — бейімдеу стратегиясының жалпыланған үрдісі. Зерттеулердің 70%-дан астамы трансфер оқыту, fine-tuning немесе мета-оқыту тәсілдерін қолданады. Бұл аз ресурсты жағдайларда балалар деректерінің тапшылығын өтеудің бірден-бір практикалық жолы болып табылады.

Екінші — бұзылыстары бар балалар үшін арнайы зерттеулердің жеткіліксіздігі. Кесте 1.4-тегі 11 зерттеудің ішінде тек Kim сөйлеу дыбыстары бұзылысы (SSD) бар балаларға нақты бағытталған. Қалған зерттеулердің бірде-бірі сөйлеу бұзылыстарын жіктеп, арнайы бейімдеу механизмін ұсынбаған.

Үшінші — қазақ тіліндегі балалар сөйлеуінің зерттелмегені. Kaldi, DeepSpeech, Google STT жүйелері ересектер деректерінде тіпті WER=52,97%

деңгейінде ғана нәтиже беріп отыр. Балалар сөйлеуінде бұл көрсеткіш одан да жоғары болатыны айқын. Сондықтан мамандандырылған корпус пен бейімдеу механизмі — өзекті ғылыми мәселе.

Төртінші — LLM постөңдеудің жаңа бағыт ретіндегі маңызы. Southwell et al. (ICASSP 2024) fine-tuned Whisper + LLM rescoring арқылы MyST деректерінде WER=9,2% жеткізді — осы уақытқа дейінгі ең жақсы нәтиже. Бұл LLM постөңдеудің балалар сөйлеуін тануда перспективалы екенін растайды [59].

2 ҚАЗАҚ ТІЛІНДЕГІ БАЛАЛАР СӨЙЛЕУІНІҢ АННОТАЦИЯЛАНҒАН АКУСТИКАЛЫҚ КОРПУСТЫ ӘЗІРЛЕУ ЖӘНЕ ӨҢДЕУ ӘДІСТЕРІ

Бұл бөлімде қазақ тіліндегі балалар сөйлеуіне арналған аннотацияланған акустикалық корпусты қалыптастырудың әдіснамалық негіздері қарастырылады. Корпус әзірлеу барысында деректерді іріктеу, жинау, құрылымдау және алдын ала өңдеу кезеңдері жүйелі түрде ұйымдастырылды. Ұсынылған тәсіл төмен ресурсты тілдер жағдайында балалар сөйлеуін зерттеу мен модельдеуге арналған репрезентативті және сапалы деректер базасын қалыптастыруға бағытталған.

2.1 Балалар сөйлеуіне арналған мәтіндік материалды іріктеу және құрылымдау әдіснамасы

Бұл бөлімде балалар сөйлеу корпусын қалыптастыруға арналған лингвистикалық материалды іріктеу және құрылымдау әдіснамасы ұсынылады. Ұсынылған тәсіл лексикалық жиілік, фонетикалық қамтылым, жас ерекшелігі және морфологиялық әртүрлілік критерийлеріне негізделген және сөйлеу деректерінің репрезентативтілігін қамтамасыз етуге бағытталған.

2.1.1 Лингвистикалық материалды іріктеу критерийлері

Сөйлеу корпусын құрудың негізі – сөздік материалдарды таңдау барысы.

Сөздік материалдарды іріктеу балалардың жас ерекшеліктеріне, тілдік даму деңгейіне, сөйлеу бұзылыстарының типіне және зерттеу мақсаттарына сәйкес ғылыми негізде жүзеге асырылады.

Халықаралық зерттеулер тәжірибесіне сүйенсек, CUCHILD корпусында кантон тілінде сөйлейтін 3-6 жас аралығындағы балалардың тілін тану үшін 130 сөз жинағы қолданылған. Бұл сөздер 1-ден 4 буынға дейінгі ұзындықты қамтиды және барлық фонемаларды сөздің әртүрлі позицияларында (басы, ортасы, соңы) қамту мақсатында таңдалған [60]. Сол сияқты, поляк тіліндегі PAVSig корпусында 4-8 жас аралығындағы 201 баладан 51 сөз және 12 логотом жинағы дайындалған. Бұл материал 12 ысқырықты дыбысты қамтиды және сөздердің графикалық бейнеленуі балаларға түсінікті болуы ескерілген [61].

Бұл зерттеуде қарастырылатын балалардың жас диапазоны 2–11 жас аралығын қамтиды. Осыған байланысты сөздік материалды іріктеу кезінде тілдік күрделілік деңгейін кезең-кезеңімен бейімдеу қажеттілігі туындайды. Ерте жастағы (2–4 жас) балалар үшін қарапайым, қысқа және жиі қолданылатын сөздер таңдалса, ал 5–7 жас аралығында буын саны көбірек және құрылымы күрделірек сөздер енгізіледі. 8–11 жас аралығындағы балалар үшін сөздік материалға күрделі сөздер мен қысқа сөйлемдер қосу қарастырылады. Таңдалған жас тобы сөйлеу және тілдік дағдылары жеткілікті деңгейде қалыптасқан, сөздер мен қарапайым сөйлемдерді өз бетінше айтуға қабілетті, сондай-ақ жазба жүргізу және сөйлеуді қайталау процесі барысында зейінін сақтай алатын балаларды қамтиды.

Лингвистикалық материалды іріктеу келесі ғылыми критерийлер негізінде жүзеге асырылды:

Фонетикалық толықтылық принциптері – қазақ тіліндегі барлық фонемалардың сөздің әртүрлі позициясында қамтылуы. Қазақ тілінде 12 дауысты және 25 дауыссыз дыбыс бар. Дауысты дыбыстардың сингармонизм заңдылығына бағынуы балалардың дыбысты дұрыс айтуын бағалауда маңызды фактор болып саналады. Сондықтан сөздер жуан және жіңішке буындар үйлесімін сақтай отырып таңдалуы қажет.

Жас ерекшелігіне сәйкестік принципі – сөздер мектепке дейінгі және кіші сынып жасындағы балалардың сөздік қорына сай болуы қадағаланады. Зерттеулер көрсеткендей, 3-6 жас аралығындағы балалардың сөздік қоры орташа 1500-ден 4000 сөзге дейін жетеді, алайда сөйлеу бұзылыстары бар балалардың ерекшеліктерін ескере отырып, бұл көрсеткіш айтарлықтай төмен болуы мүмкін. Сол себепті сөздер қарапайым, күнделікті қолданыстағы және мағынасы түсінікті болуы қажет.

Дыбыстық құрылымның әртүрлілігі принципі – сөздер буын құрылымы (ашық, тұйық буындар), ұзындығы (1-4 буын) және дыбыстық құрамы бойынша әртүрлі болуы тиіс. Осы шарттарды орындай отыра, сөйлеу бұзылыстары бар балалардың қазақша сөйлеуін тану әдісін әзірлеу үшін дұрыс әрі сапалы корпустардың бастапқы нұсқасын әзірлеу мүмкіндігі қарастырылды. Осы арқылы балалардың сөздік қорын кезең-кезеңімен кеңейту тиімділігін арттырады. Бұл тәсіл сөйлеуді тану моделінің әртүрлі фонетикалық комбинацияларды меңгеруіне мүмкіндік береді.

Сонымен қатар, сөздік материалды қалыптастыруда қосымша талаптар да ескеріледі. Атап айтқанда, сөйлеу бұзылыстары бар балалар үшін қиындық тудыратын р, л, с, ш, ж, қ, ғ дыбыстары арнайы қамтылады және сөздің барлық позицияларында енгізіледі. Бұл артикуляциялық қателерді дәл анықтауға мүмкіндік береді. Сөздік материалдың лексикалық теңгерімділігі қамтамасыз етіліп, зат есімдер, етістіктер және сын есімдер қамтылады, бұл модельдің семантикалық тұрақтылығын арттырады.

Халықаралық тәжірибеде кеңінен қолданылатын әдістердің бірі – логотомдарды енгізу. Мағынасы жоқ, бірақ фонетикалық тұрғыдан дұрыс құрылған сөздер балалардың жаттап айтуын болдырмай, таза дыбыстау сапасын бағалауға мүмкіндік береді. Сонымен қатар, контекстік сөйлемдерді қолдану сөйлеудің табиғи ағымын, интонациясын және дыбыстардың өзара әсерін (коартикуляция) ескеруге мүмкіндік береді. Әдетте мұндай сөйлемдер 2–5 сөзден тұрады.

Сөздерді іріктеу барысында жиілік принципі де сақталды және қамтамасыз етілді, яғни материал балалар жиі қолданатын тақырыптарға (отбасы, ойыншықтар, мектеп, табиғат) негізделді. Бұл деректердің шынайылығын арттырады және модельдің нақты жағдайларда тиімді жұмыс істеуін қамтамасыз етеді. Сонымен қатар, сөздер күрделілік деңгейі бойынша кезеңделеді: бастапқы деңгейде 1–2 буынды қарапайым сөздер, кейін 2–3 буынды орташа сөздер, ал соңында 3–4 буынды күрделі сөздер енгізілді.

Зерттеу барысында ата-аналардың рұқсатымен Алматы қаласында тұратын 3–5 жас аралығындағы балалармен эксперимент жүргізілді. Эксперимент келесі

түрде ұйымдастырылды: тәрбиешілер тобына қазақ тілін үйретуге арналған арнайы қосымша ұсынылып, олар осы құралды пайдалана отырып балалармен жұмыс жүргізді. Қатысушылар қазақ тілінде оқытылатын топтағы балалар болып табылады.

Балалардың сөйлеу қабілеті әртүрлі жаста қалыптасады. Мысалы, 2–3 жас аралығында балалардың сөйлеуі толық қалыптаспаған болып, сөздерді бұрмалап айту, өздігінен сөз ойлап табу және дыбыстарды дұрыс айтпау сияқты ерекшеліктер байқалады. Ал 3–5 жас аралығында сөйлеу біртіндеп түсінікті бола бастайды, балалар етістіктерді белсенді қолданады, есімдіктер, үстеулер мен сын есімдердің саны артады. Осы кезеңде сөйлеу құрылымы жүйеленіп, тілдік дағдылар қарқынды дамиды. Сондықтан бұл жас аралығы балалардың сөйлеуін зерттеу үшін ең қолайлы кезеңдердің бірі болып саналады.

Зерттеуге қатысатын балалар белгілі бір критерийлер бойынша іріктелді. Негізгі талаптар ретінде жас мөлшері (2 жас 11 айдан 5 жас 11 айға дейін) және тілдік даму деңгейі ескерілді. Сонымен қатар баланың зейінін ұстай алу және мінез-құлқын бақылау қабілеті маңызды фактор ретінде қарастырылды, себебі бұл дағдылар зерттеу барысында бағаланатын негізгі көрсеткіштердің бірі болып табылады. Осылайша, зерттеуге енгізілген балалар белгіленген талаптарға толық сәйкес келіп, тілдік дағдылардың дамуын зерттеу үшін қолайлы топ ретінде таңдалды.

2.1.2 Сөздік материалды қалыптастыру әдіснамасы

Ұсынылған әдістеме ғылыми зерттеулердің дұрыстығы мен сенімділігін қамтамасыз ететін әдіснамалық қатаңдық принциптеріне негізделген. Бұл принциптер балаларға арналған сөздік қорды әзірлеудің жалпы сипаттамасын қамтиды. Әрбір кезеңде қолданылатын әдістер мен құралдардың сәйкестігі мен тиімділігі жан-жақты қарастырылады. Сонымен қатар нәтижелердің дұрыстығы мен ұсынылу формасы да ескеріліп, зерттеудің жалпы сапасын қамтамасыз етеді.

Балаларға арналған сапалы сөздік қорды қалыптастыру әдістемесі бірнеше негізгі кезеңнен тұрады:

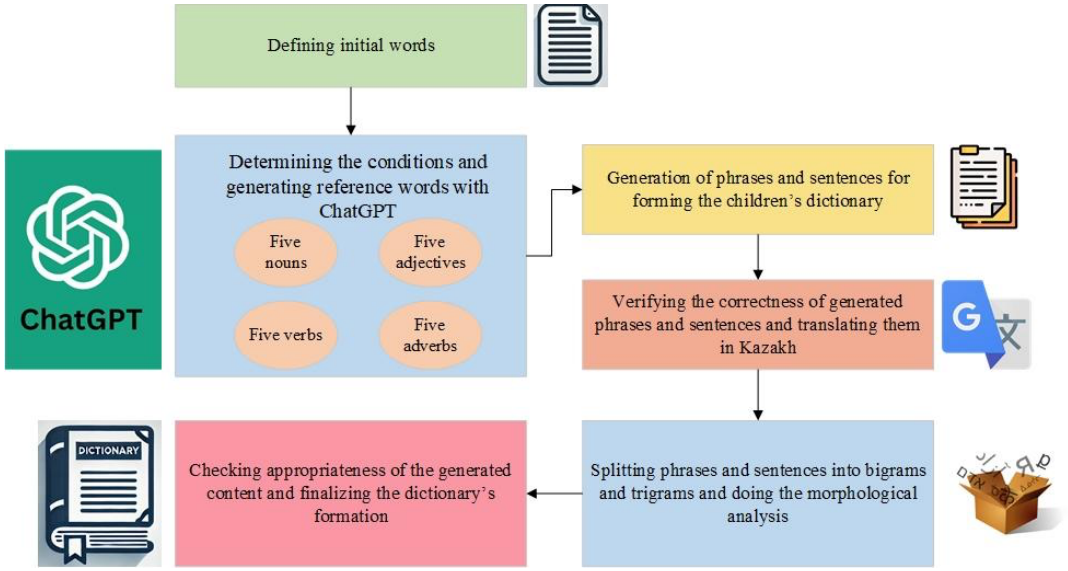
– Тірек сөздерді генерациялау. Бұл кезеңде сөздік қордың негізін құрайтын бастапқы сөздер анықталады. Тірек сөздер балалардың жиі қолданатын және мағыналық тұрғыдан маңызды сөздерінен тұрады. Лексикалық материалды кеңейту және семантикалық әртүрлілікті арттыру GPT архитектурасына негізделген модель бастапқы тірек сөздерге қатысты семантикалық байланысқан сөздер жиынын генерациялау үшін пайдаланылды. Генерацияланған деректердің сапасын қамтамасыз ету мақсатында барлық нәтижелер лингвистикалық сарапшылар тарапынан қосымша тексеруден өткізілді. мақсатында генеративті тілдік модельдер (LLM) қолданылды. Атап айтқанда,

– Сөз тіркестері мен сөйлемдерді генерациялау. Іріктелген тірек сөздер негізінде сөз тіркестері мен сөйлемдер жүйелі түрде генерацияланады. Алғашқы кезеңде мәтіндік құрылымдар жоғары дәлдікке қол жеткізу мақсатында ағылшын тілінде құрастырылып, кейін автоматтандырылған аударма жүйелері арқылы

қазақ тіліне трансформацияланады. Бұл тәсіл генерация сапасын арттыруға және семантикалық бірізділікті сақтауға мүмкіндік береді.

– Сүзгілеу және морфологиялық талдау. Генерацияланған және аударылған сөйлемдер мен сөз тіркестері грамматикалық дұрыстығын қамтамасыз ету мақсатында морфологиялық талдаудан өткізіледі. Сонымен қатар мәтіндер құрылымдық талдау үшін n-граммалық модельдеу негізінде биграммалар мен триграммаларға жіктеледі. Бұл тәсіл тілдік үлгілерді формальды сипаттауға, сөздердің өзара байланысын анықтауға және сөйлеу заңдылықтарын сандық тұрғыдан талдауға мүмкіндік береді.

– Білім беру талаптарына сәйкестендіру. Қорытынды кезеңде алынған мәтіндік материал арнайы әзірленген педагогикалық және лингвистикалық шаблондар негізінде қосымша сүзгіден өткізіледі. Бұл процесс олардың балалардың жас ерекшеліктеріне сәйкестігін, түсініктілігін және оқу-дидактикалық қолдануға жарамдылығын бағалауға бағытталған. Ұсынылған әдістемелік құрылымның толық архитектурасы 2.1-суретте келтірілген.



Сурет 2.1 – Ұсынылып отырған әдістеменің құрылымы

Алынған сөз тіркестері мен сөйлемдер бастапқы кезеңде мазмұндылық және семантикалық сәйкестік тұрғысынан тексерілгенімен, оларды сөздіктің құрылымдық моделін қалыптастыруға, сондай-ақ кейінгі аудио және визуализация модульдерінде тиімді қолдануға бейімдеу қажеттілігі туындайды. Осы мақсатта сөйлемдердің маңыздылық дәрежесін және олардың оқу процесіндегі меңгерілу ретін анықтау үшін қосымша сандық бағалау метрикалары қолданылды. Атап айтқанда, сөйлемдер арасындағы семантикалық және құрылымдық ұқсастықты өлшеу үшін Cosine Similarity (косинустық ұқсастық), Euclidean Distance (Евклид қашықтығы) және Manhattan Distance (Манхэттен қашықтығы) метрикалары пайдаланылды. Бұл тәсілдер сөйлемдердің өзара жақындығын формальды түрде бағалауға, оларды күрделілік

деңгейі бойынша реттеуге және оқу материалын кезең-кезеңімен ұсынуға мүмкіндік береді.

Қалыптастырылған сөздік кейіннен мультимодальды оқыту ортасына интеграцияланады және аудио мен визуализация модульдерімен толықтырылады. Визуализация компоненті балалардың когнитивтік қабылдау ерекшеліктерін ескере отырып, графикалық элементтер мен анимациялар арқылы жүзеге асырылады. Сонымен қатар, визуалды жүктеме (cognitive load) деңгейі балалардың қабылдауына кері әсер етпейтіндей етіп оңтайландырылған. Аудио компонент визуалды мазмұнмен синхрондалып, балалар үшін үздіксіз, интуитивті және тиімді оқу тәжірибесін қамтамасыз етеді. Бұл интеграция сөйлеу материалын қабылдау сапасын арттырып, тілдік дағдылардың қалыптасуын жеделдетеді.

Балаларға арналған сөздік қорды қалыптастыру бірнеше бірізді кезеңдерден тұрады, олардың әрқайсысы лингвистикалық және есептеуіш әдістердің үйлесіміне негізделген.

Алғашқы кезеңде тірек сөздер жиыны генерацияланады. Бұл мақсатта генеративті тілдік модельдер (Large Language Models, LLM) негізделген модель қолданылды. Мұндай модельдер үлкен көлемдегі мәтіндік деректерде алдын ала оқытылғандықтан, сөздер арасындағы семантикалық байланыстарды тиімді анықтауға және әртүрлі контексттерде қолданылатын лексикалық бірліктерді генерациялауға мүмкіндік береді.

Қазақ тілі төмен ресурсты тілдер қатарына жататындықтан, мәтін генерациясының бастапқы кезеңі ағылшын тілінде жүзеге асырылды. Бұл тәсіл генерация сапасын арттыруға және синтаксистік қателіктерді азайтуға мүмкіндік береді. Кейін алынған мәтіндер қазақ тіліне бейімделіп, қосымша лингвистикалық тексеруден өткізілді.

Тірек сөздерді іріктеу барысында балалардың қоршаған ортаны тануына тікелей байланысты тақырыптық домендер анықталды. Атап айтқанда, жануарлар, өсімдіктер және ойыншықтар сияқты негізгі семантикалық өрістер таңдалды. Бұл домендер балалардың когнитивтік дамуына сәйкес келетін және күнделікті коммуникацияда жиі қолданылатын лексикалық бірліктерді қамтиды.

2.1.3 Генерация және LLM қолдану арқылы кеңейту

Тірек сөздерді қалыптастырудың бастапқы кезеңінде балалардың қоршаған ортаны түсінуіне көмектесетін үш негізгі тақырып таңдалды: жануарлар, өсімдіктер, ойыншықтар.

Бұл сөздер тізімі (1) түрінде ұсынылды:

$$W=\{w_1,w_2,...,w_n\}, \quad (2.1)$$

мұнда W – 3-5 жас аралығындағы балаларға қатысты сөздер жиыны.

Осы тақырыптарға сәйкес келесі сөздер тізімі құрастырылды:

1. Жануарлар: бегемот, бүркіт, қасқыр, қаз, керік, қоян, ешкі, сиыр, мысық, аққу, арыстан, түлкі, аю, тышқан, қой, бұғы, әтеш, питон, тотықұс, піл, ит, жолбарыс.

2. Бақ (өсімдіктер): банан, сәбіз, шие, алма, картоп, гүл, ағаш, алмұрт, қарбыз, қияр, жидек, қызанақ, бақ, жеміс, өрік, бұрыш, лимон, қызылша, көкөніс, асқабақ.

3. Ойыншықтар: қуыршақтар, машиналар, ойыншық пойыздар, ойыншық тікұшақтар, жұмсақ ойыншықтар, роботтар, Lego, пазлдар, құрылыс блоктары, ниндзя тасбақалар.

Берілген сөздер негізінде әрбір сөз үшін семантикалық тұрғыдан байланысты сөздер жиыны генерацияланды. Генеративті тілдік (ChatGPT 4.0) модельдері синонимдер мен мағыналас сөздерді кеңінен қолдана отырып, деректердің әртүрлілігін арттырады. Алайда жоғары дәлдік қажет болған жағдайда бұл әртүрлілік шектеуді талап етеді. Сондықтан сөздерді басқару мақсатында олар сөз таптары бойынша бөлінді: зат есімдер, етістіктер, сын есімдер және үстеулер. Әр категория үшін белгілі бір мөлшерде (мысалы, 5 сөзден) тірек сөздер таңдалды.

Әрбір $w_i \in W$ сөзі үшін генеративті модель негізінде семантикалық байланысты сөздер жиыны қалыптастырылады:

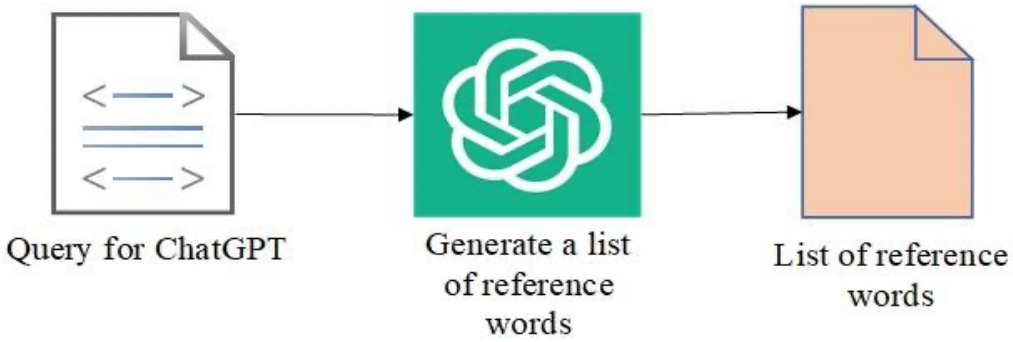
$$R(w_i) = \{r_{i1}, r_{i2}, \ldots, r_{im}\}, \quad (2.2)$$

мұнда $R(w_i)$ — w_i сөзіне қатысты мағыналас және контекстік байланысқан сөздер жиыны.

Генерацияланған сөздердің әртүрлілігін бақылау мақсатында олар сөз таптары бойынша (зат есім, етістік, сын есім, үстеу) құрылымдалды. Бұл тәсіл тілдік үлгілердің теңгерімділігін қамтамасыз етеді және сөйлеу модельдерінің семантикалық тұрақтылығын арттырады.

Мысалы, “banana” (банан) сөзі үшін алынған семантикалық байланысты бірліктер: fruit, tropical, sweet, ripe, peel, smoothie және т.б. Бұл сөздер контекстік ұқсастық пен семантикалық жақындық негізінде таңдалады.

Тірек сөздерді генерациялау құрылымы 2.2-суретте көрсетілген.



Сурет 2.2 – Тірек сөздерді генерациялау құрылымы

Балалар сөздігі үшін сөз тіркестері мен сөйлемдерді генерациялау. Сөз таптары бойынша тірек сөздер қалыптастырылғаннан кейін келесі кезеңде сөз

тіркестері мен сөйлемдер генерацияланады. Бұл процесс тірек сөздер мен олардың семантикалық байланысқан кеңейтілген жиындарына негізделіп, лексикалық және синтаксистік әртүрлілікті қамтамасыз етуге бағытталған.

Сөйлемдерді генерациялау кезінде әрбір тақырыптық категория үшін семантикалық шектеулер мен сипаттамалық параметрлер анықталады. Атап айтқанда:

- жануарлар үшін – өлшемі, түсі, қоректену түрі және мекен ету ортасы;
- өсімдіктер үшін – түсі, дәмдік қасиеттері және өсу ортасы;
- ойыншықтар үшін – пішіні, түсі және құрылымдық ерекшеліктері ескеріледі.

Бұл параметрлер генерацияланған сөйлемдердің мазмұндық дұрыстығын қамтамасыз етіп, балалардың когнитивтік қабылдауына сәйкес келетін контекстік байланыстарды қалыптастыруға мүмкіндік береді.

Мысал ретінде “банан” лексемасы негізінде келесі сөйлемдер генерацияланды:

сары банан тәтті және піскен;

тропикалық өсімдік күн сәулесінде жақсы өседі;

бананнан жасалған смузи пайдалы;

балалар бананды жақсы көреді;

банан пайдалы әрі құнарлы тағам.

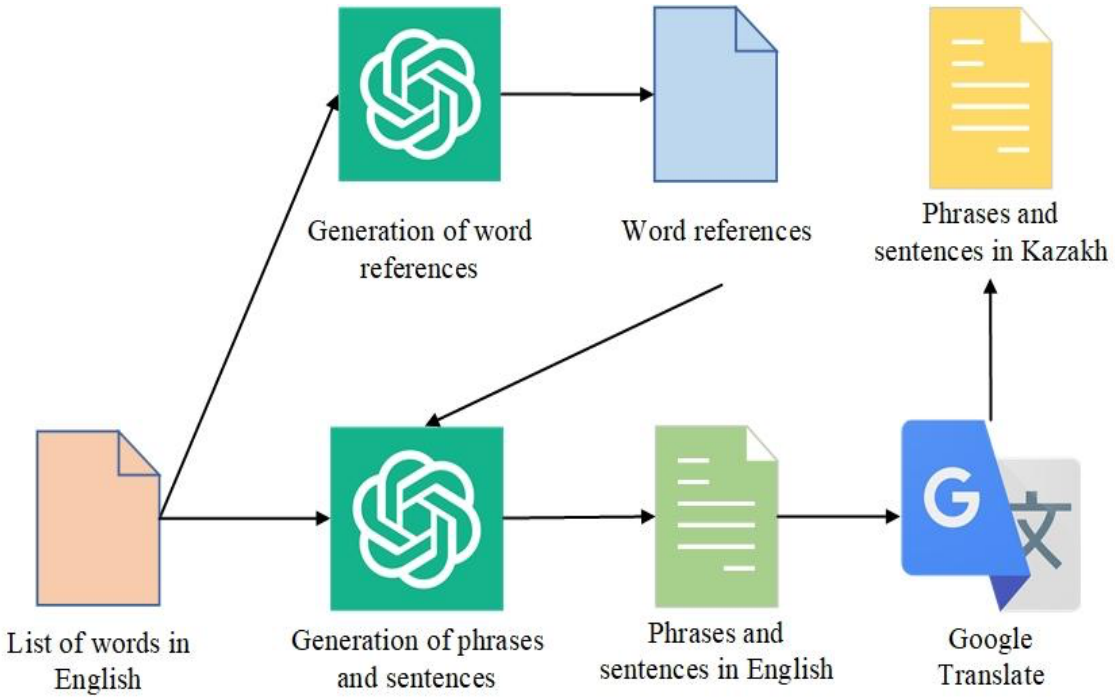
Генерацияланған сөйлемдер бастапқы кезеңде ағылшын тілінде құрастырылып, кейін қазақ тіліне автоматтандырылған аударма жүйесі арқылы түрлендіріледі. Бұл процесс келесі түрде формальды түрде сипатталады:

$T(S_i) = MT(S_i)$, (3)

мұнда S_i — бастапқы сөйлем, MT — машиналық аударма функциясы, ал $T(S_i)$ — қазақ тіліне аударылған нәтиже.

Аударылған мәтіндер кейін морфологиялық талдаудан өткізіліп, олардың грамматикалық дұрыстығы мен синтаксистік құрылымы тексеріледі. Сонымен қатар мәтіндер n-граммалық модельдеу тәсілі негізінде биграммалар мен триграммаларға жіктеледі. Бұл тәсіл сөйлемдердің ішкі құрылымын формальды сипаттауға, сөздер арасындағы тәуелділіктерді анықтауға және тілдік үлгілерді сандық тұрғыдан талдауға мүмкіндік береді.

Сөз тіркестері мен сөйлемдерді генерациялау, аудару және өңдеу кезеңдерінің толық әдістемелік құрылымы 2.3-суретте көрсетілген.



Сурет 2.3 – Сөз тіркестері мен сөйлемдерді генерациялау және оларды қазақ тіліне аудару жүйесінің архитектурасы

Ұсынылған архитектура сөз тіркестері мен сөйлемдердің генерациядан бастап морфологиялық тексеруге дейінгі толық өңдеу тізбегін бейнелейді. Осылайша қалыптастырылған лингвистикалық материал балалар сөйлеуінің жас ерекшеліктеріне сәйкес келетін, грамматикалық тұрғыдан дұрыс және фонетикалық тұрғыдан вариативті деректер жиынын қамтамасыз етеді. Бұл деректер кейінгі бөлімдерде қарастырылатын акустикалық корпусты жинақтау және ASR модельдерін оқыту үшін лингвистикалық негіз болып табылады.

2.1.4 Морфологиялық талдау және сүзгілеу

Алдыңғы кезеңдерде тірек сөздер, сөз тіркестері және сөйлемдер генерацияланып, бастапқы лексикалық база қалыптастырылды. Алайда генерацияланған мәтіндердің сапасы мақсатты аудиторияның (балалардың) жас ерекшеліктеріне толық сәйкес келе бермейді. Сондықтан алынған мәтіндерді морфологиялық және құрылымдық тұрғыдан талдау, сондай-ақ оларды қосымша сүзгілеу қажеттілігі туындайды.

Балаларға арналған мәтіндерді құрастыру барысында олардың когнитивтік және тілдік даму ерекшеліктері ескерілуі тиіс. Атап айтқанда, 3–5 жас аралығындағы балалар үшін сөйлемдердің қысқа, синтаксистік тұрғыдан қарапайым және семантикалық жағынан түсінікті болуы талап етіледі. Генеративті модельдер көбінесе күрделі синтаксистік құрылымдарды құруға бейім болғандықтан, алынған мәтіндер қосымша нормализация және жеңілдету кезеңінен өткізілді.

Тақырыптық сөздік лингвистикалық және әдістемелік критерийлер негізінде сүзгіден өткізілді. Бұл критерийлерге мәтіннің күрделілік деңгейі,

балалар сөйлеуінің типтік құрылымдары және қазақ тіліндегі артикуляциялық тұрғыдан қиын дыбыстардың қамтылуы жатады. Мұндай тәсіл сөздік материалдың балалар сөйлеуіне барынша жақындатылған үлгісін қалыптастыруға мүмкіндік береді.

Сөздіктің құрылымы формальды түрде келесі жиын арқылы сипатталады:

$$D = \{u_i, f_{ij}, S_{jk} \mid i = 1..n, j = 1..m, k = 1..l\}, \quad (2.3)$$

мұнда:

u_i — лексикалық бірліктер жиыны;

f_{ij} — і-ші сөзге сәйкес сөз тіркестері (биграммалар мен триграммалар);

S_{jk} — сөйлемдер жиыны.

Қосымша түрде сөздердің грамматикалық үйлесу заңдылықтары келесі құрылыммен сипатталады:

$$G = \langle W, R_{pos} \rangle, \quad (2.4)$$

мұнда:

W — сөздер жиыны;

R_{pos} — сөз таптарының орналасу ережелері.

Бұл формализация сөздік элементтерінің өзара байланысын сипаттауға және оларды есептеуіш модельдерде қолдануға мүмкіндік береді.

Генерацияланған сөйлемдер бастапқыда машиналық аударма жүйелері арқылы қазақ тіліне түрлендіріліп, кейін құрылымдық талдау үшін n-граммалық модельдеу тәсілі негізінде биграммалар мен триграммаларға бөлінді. Бұл әдіс сөздердің тізбектік байланысын талдауға және тілдік үлгілерді формальды түрде сипаттауға мүмкіндік береді.

Келесі кезеңде әрбір алынған тілдік бірлікке морфологиялық талдау жүргізілді. Бұл мақсатта қазақ тілінің морфологиялық ерекшеліктерін ескеретін Apertium-Kaz [62] жүйесі қолданылды. Морфологиялық талдау нәтижесінде әрбір сөздің грамматикалық белгілері (сөз табы, септік, көптік, шақ және т.б.) анықталады. Мысалы:

^Мысықтар/мысық<n><pl><nom>$
^балық/балық<n><nom>$
^пен/мен<cnjcoo>$
^тауық/тауық<n><nom>$
^сияқты/сияқты<post>$
^ет/ет<n><nom>$
^жейді/же<v><tv><aor><p3><sg>$

Мұндай белгілеу мәтінді формальды түрде сипаттауға және сөйлеу үлгілерін автоматты түрде өңдеуге мүмкіндік береді.

Сөйлем құрылымдарын модельдеу кезеңінде 3–5 жас аралығындағы балаларға тән синтаксистік үлгілер анықталды. Қазақ тіліндегі негізгі сөз таптарының комбинациялары негізінде қарапайым сөйлем құрылымдары қалыптастырылды. Атап айтқанда, келесі үлгілер кеңінен қолданылады:

есімдік + етістік
есімдік + зат есім + етістік
есімдік + сын есім + зат есім + етістік

Зерттеулер көрсеткендей, осы жас аралығындағы балалар көбінесе қысқа (2–3 сөзден тұратын) сөйлемдерді қолданады. Сондықтан күрделі синтаксистік құрылымдар шектеліп, қарапайым және жиі қолданылатын үлгілерге басымдық берілді. [63]

Осылайша, морфологиялық талдау және құрылымдық модельдеу кезеңдері генерацияланған мәтіндерді балалар сөйлеуінің табиғи үлгілеріне жақындатуға, сондай-ақ оларды автоматты сөйлеуді тану жүйелерінде тиімді қолдануға мүмкіндік береді.

Кесте 2.1 – 3–5 жас аралығындағы балалар сөйлеуіндегі морфологиялық таңбаланған n-граммалардың құрылымдық классификациясы.

Үлгі	**Мысал**	**Түрі**
word1<n> word2<n>	Банан ағашы	n-noun
word1<n> word2<v>	Банан өседі	v-verb
word1<adv> word2<v>	Ағашта өседі	adv-adverb
word1<adj> word2<n>	Тәтті банан	adj-adjective
word1<num> word2<n>	Екі банан	num-numerical
word1<prn> word2<n>	Оның бананы	prn-pronoun
word1<prn> word2<v>	Ол алды	prn-pronoun
word1<n>word2<n> word3<v>	Банан ағашы өседі	—
word1<adj>word2<n> word3<n>	Тәтті банан ағашы	—
word1<n>word2<v> word3<v>	Банан қарайып жатыр	—
word<n>word<cnjcoo> word3<n>	Банан мен алма	cnjcoo- conjunction
word<adj>word<cnjcoo> word3<adj>	Тәтті және жұмсақ	adj-adjective

Сөз таптарының қазақ тіліндегі үйлесу үлгілеріне негізделген және морфологиялық талдаудан өткен сүзгіден өткізілген биграммалар мен триграммалар төменде (үзінді) келтірілген:

балық<n> пен<cnjcoo> тауық<n> [fish<n> and<cnjcoo> chicken<n>]

ет<n> жейді<v> [meat<n> eat<v>]

Мысал сөйлемдер негізінде балаларға арналған биграммалар мен триграммалардың нәтижелік тізімі (үзінді):

Балық пен тауық [fish and chicken]

Ет жейді [eat meat]

Сүзгіден өткен бұл тіркестер сөздіктің сөз тіркестері деңгейіне енгізілді. Балаларға арналған тақырыптық сөздіктің соңғы нәтижесі 2.2-кестеде (үзінді) көрсетілген [64].

Кесте 2.2 – Балаларға арналған тақырыптық сөздік (үзінді)

Қазақ сөздері	Қазақ тіліндегі сөз тіркестері (3–4 жас)	Қазақ тіліндегі сөйлемдер (3–4 жас)
Мысық	Балық пен тауық Ет жейді кішкентай және жұмсақ үй жануары сары мысық ұйықтағанды жақсы көреді...	Мысықтар балық пен тауық сияқты ет жейді. Мысықтар әдетте кішкентай және жұмсақ. Көптеген мысықтар жылы үйлерде тұрады. Мысықтар далада ұйықтағанды жақсы көреді. Бізде сары мысық бар.
Банан	сары банан тәтті банан тропикалық жеміс тәтті және піскен ...	Сары банан тәтті және піскен. Тропикалық өсімдік күн сәулесінде оңай өседі. Калий мен дәрумендері бар дәмді жеміс. Иілген жемістер тропиктік аймақтарда өседі. Бақытты балалар смузи ішкенді жақсы көреді. ...
Қуыршақ	қызғылт қуыршақ қуыршағым әдемі барби қуыршақ ...	Әдемі қызғылт қуыршақ ойыншығын әперді. Менің қуыршағым әдемі көрінеді. Ойын уақытына арналған көне барби қуыршақ үйі. Киіну және таңдану үшін сүйкімді қуыршақ. ...

Осы бөлімде қазақ тіліндегі балалар сөйлеу корпусын қалыптастыруға арналған лингвистикалық материалды іріктеу және құрылымдау әдіснамасы ұсынылды. Ұсынылған тәсіл лексикалық, фонетикалық және морфологиялық критерийлерді ескере отырып, балалардың жас ерекшеліктеріне сәйкес келетін мәтіндік деректерді жүйелі түрде қалыптастыруға бағытталған.

Зерттеу барысында тірек сөздер жиыны генеративті тілдік модельдер көмегімен кеңейтіліп, семантикалық әртүрлілік қамтамасыз етілді. Алынған сөз тіркестері мен сөйлемдер морфологиялық талдаудан өткізіліп, олардың грамматикалық дұрыстығы мен құрылымдық сәйкестігі тексерілді. Сонымен қатар n-граммалық модельдеу негізінде мәтіндік құрылымдар формальды түрде сипатталып, балалар сөйлеуіне тән синтаксистік үлгілер анықталды[65].

Ұсынылған әдістеме сөздік материалдың сапасын арттырумен қатар, оны автоматты өңдеуге бейімдеуге мүмкіндік береді. Қалыптастырылған лингвистикалық база балалар сөйлеуін тану жүйелерін оқыту үшін репрезентативті және құрылымдалған деректер жиынын қамтамасыз етеді.

Осылайша, бөлімде ұсынылған әдіснамалық тәсіл балалар сөйлеу корпусын құрудың бастапқы және маңызды кезеңін қалыптастырып, келесі бөлімдерде қарастырылатын акустикалық деректерді жинау және өңдеу процестеріне негіз болады.

2.2 Балалар сөйлеуінің акустикалық деректерін жинау

Корпустың сапасын анықтайтын ең маңызды кезеңдердің бірі – акустикалық деректерді жинау процесі болып табылады. Әсіресе балалар сөйлеуін жазу ересектердің сөйлеуін жазумен салыстырғанда айтарлықтай ерекшеленеді және бірқатар әдістемелік және техникалық қиындықтарды туындатады. Бұл қиындықтар балалардың сөйлеу аппаратының толық қалыптаспағандығымен, артикуляциялық тұрақсыздықпен, сөйлеу қарқынының өзгермелілігімен және сыртқы орта факторларына жоғары сезімталдығымен байланысты.

Сонымен қатар, балалар сөйлеуінде фонетикалық вариативтілік деңгейі жоғары, бұл бір сөздің әртүрлі акустикалық көріністерде айтылуына әкеледі. Осы ерекшеліктерді ескере отырып, деректерді жинау процесі тек техникалық жазба жүргізумен шектелмей, сонымен қатар репрезентативті, теңгерімді және сапалы акустикалық корпус қалыптастыруға бағытталған комплексті әдіснаманы талап етеді.

Осы зерттеуде балалар сөйлеуінің акустикалық деректерін жинау көпкөзді (multi-source) және көпәдісті (multi-method) тәсіл негізінде ұйымдастырылды. Мұндай тәсіл әртүрлі сөйлеу жағдайларын қамтуға, деректердің әртүрлілігін арттыруға және модельдердің жалпылау қабілетін жақсартуға мүмкіндік береді.

Акустикалық деректер жиыны формальды түрде келесі түрде сипатталады:

$X = \{x_1, x_2, ..., x_n\}$,

мұнда x_i — балалардың сөйлеу сигналдарын білдіретін аудиожазбалар жиыны.

Әрбір аудиожазбаға сәйкес метадеректер жиыны анықталады:

$M = \{m_1, m_2, ..., m_n\}$,

мұнда m_i — і-ші жазбаға сәйкес келетін сипаттамалар жиыны (жас, жыныс, жазу жағдайы, сөйлеу типі және т.б.).

Осылайша, корпус $C = \langle X, M \rangle$ түрінде ұсынылады, бұл оның құрылымдық және аналитикалық өңдеуге дайын болуын қамтамасыз етеді [66].

2.2.1 Деректерді жинау әдістері

Қазақ тіліндегі балалар сөйлеуінің акустикалық корпусын қалыптастыру мақсатында деректерді жинау үш негізгі әдіс арқылы жүзеге асырылды. Әрбір әдіс деректердің белгілі бір типін қамтуға және корпус сапасының әртүрлі аспектілерін қамтамасыз етуге бағытталған:

1. Жартылай автоматтандырылған жинау (Telegram-бот негізінде) – құрылымдалған және бақыланатын деректерді жинауға мүмкіндік береді;
2. Табиғи ортадағы сөйлеуді жинау – балалардың спонтанды және еркін сөйлеу үлгілерін алуға бағытталған;
3. Бақыланатын жағдайда аудио жазу – жоғары сапалы және шу деңгейі төмен акустикалық деректерді қамтамасыз етеді.

Ұсынылған тәсілдердің үйлесімді қолданылуы корпус құрамына әртүрлі сөйлеу типтерін (оқылған, жартылай еркін және спонтанды сөйлеу) енгізуге мүмкіндік береді, бұл өз кезегінде автоматты сөйлеуді тану жүйелерінің тұрақтылығы мен дәлдігін арттыруға ықпал етеді.

Әрбір әдіс бойынша деректерді жинау процесі, техникалық параметрлері және сапаны бақылау механизмдері келесіде қарастырылады.

Бірінші әдіс – Dataset Loader Bot (Telegram-Based Audio Collection Bot for Kazakh Children's Speech Corpus. https://t.me/DatasetLoaderBot) деп аталатын Python 3.9 және aiogram 2.25.1 көмегімен арнайы жасалған Telegram-ботты әзірлеу және енгізу. Бұл құрал қазақ балалар сөйлеу деректерін жартылай автоматты түрде жинау және құрылымдау үшін арнайы әзірленді [67]. Ол зерттеу тобы қызмет көрсететін арнайы үздіксіз жұмысты серверге қосылып, аудио сапасын автоматты тексеруді және жас, жыныс, тақырып сияқты метадеректерді бекітуді қамтамасыз етті. Telegram платформасы Қазақстанда, әсіресе балалары бар отбасылар арасында кең таралғандықтан, сонымен қатар оның дауыстық хабарламаларға тума қолдау, боттармен біркелкі интеграция және қосымша қолданбаларды қажет етпеу сынды техникалық артықшылықтарына байланысты таңдалды. Дауыстық хабарламалар интерфейстегі ақпарат арқылы, пайдаланушылардан алдын ала анықталған сөздерді дыбыстау арқылы осы чатқа тікелей жазу арқылы жиналды. Жүйе аудио сапасын (ұзақтығын, анықтығын және сигнал-шуыл қатынасын) автоматты түрде бағалап, қажет болған жағдайда қайта жазуын сұрады. Dataset Loader Bot (бағдарламалық жасақтама нұсқасы 1.0) компьютерлік бағдарлама ретінде ресми тіркелген және қорғалған, бұл оның түпнұсқалығы мен заңды мәртебесін растайды.

Екінші әдіс – балалардың қазақ тілінде сөйлеуін табиғи ортада (үйде, мектепте, балабақшаларда) жазу, сондай-ақ YouTube желісіндегі қоғамдық бейнежазбалардан ойда жоқта сөйлеуді жинау, онда балалар нақты өз дауыстарымен сөйлегеніне кепілдік болады. Осы себепті де біз дубляждалған мультфильмдер мен фильмдерден әдейі аулақ болдық, өйткені оларда әдетте ересек дыбыстаушылардың дауысы арқылы жасалады. Сол үшін оның орнына күнделікті жағдайлардағы балалардың нақты сөйлеуін түсіретін бейнелерді ғана таңдадық. Мысалы, «Learn Kazakh with My Niece» (Learn Kazakh with My Niece. https://www.youtube.com/watch?v=nVwEGbgBJeY) бейнесінде жас қыз табиғи жағдайда қазақша еркін сөйлейді. Бұл әдіс балалардың дауысын оқшаулау және фондық шуды азайту үшін қосымша күш жұмсауды қажет етті, бірақ біздің мақсатты жас тобымыз үшін неғұрлым репрезентативті және шынайы сөйлеу деректерін жинауға мүмкіндік берді.

Үшінші әдіс – неғұрлым бақыланатын жағдайларда балалардың сөйлеуін сапалы жазу үшін дауыс жазатын құрылғыны пайдалану. Сол себепті iPhone 14 Pro Max (Apple Inc., Cupertino, CA, USA) және macOS компьютеріне (Apple Inc., Cupertino, CA, USA) қосылып, аудионы 16 кГц жиілікте WAV форматында жазу үшін пайдаланылды. Осы сеанстар барысында балаларға түстерді анықтау, суреттердегі заттарды атау және санау сияқты қарапайым тапсырмаларды орындау сұралды, ал бұл жасқа сай, анық сөйлеу үлгілерін алуға көмектесті. Осы әдіс арқылы алынған барлық жазбалар зерттеу тобымен қолмен таңбаланды. Әрбір файл мұқият қаралып, сегменттелді; сөйлемдер транскрипцияланды, және жас, жыныс сияқты метадеректер берілді. Нормализациялау мақсатында, егер баланың сөйлеуі түсініксіз болса, ересек адам жазбада басым болса немесе сөйлем толық болмаса (мысалы, басы жазылмаған), жазбалар алынып тасталды. Соңғы жазбалар Google LLC (Mountain View, CA, USA) ұсынатын бұлтты қызмет Google Drive көмегімен ұйымдастырылып, одан әрі пайдалану үшін сақталды.

Кәмелетке толмаған балалардың жеке деректері олардың ата-аналарының немесе заңды өкілдерінің келісімімен ғана жиналды және өңделді. Барлық деректер, соның ішінде аудиожазбалар мен метадеректер (аты, мекенжайы, жынысы т.б.) анонимдендірілді. Атаулардың орнына бірегей ID кодтары пайдаланылды. Зерттеу құқықтық және этикалық стандарттарға толық сәйкес жүргізілді.

2.3-кестеде балалар сөйлеуінің акустикалық корпусы үшін деректерді жинау сипатталған және әдістер мен жазу сипаттамалары келтірілген.

2.3-кесте. Аудио деректерін жинау әдістері

Әдіс	Сипаттама	Жазбалар саны	Жас санаты	Жазу ұзақтығы
Telegram боты	Telegram боты арқылы сөздер мен сөз тіркестерін жазу, автоматты сақтау және сапаны тексеру	~800	2-8	2 сағатқа жуық
Табиғи орта (клиптер)	Күнделікті жағдайдағы жазбалар (үй, мектеп), YouTube-тен қазақ тіліндегі мультфильмдер мен фильмдер	3000+	6-11	3 сағаттан астам
Табиғи орта (дауыс жазатын құрылғы)	Смартфондарды пайдаланып WAV форматындағы жазбалар, балалардың сөйлеуін тексеруге арналған тапсырмалар	5000+	2-6	2 сағаттан астам

Деректерді жинаудың әртүрлі әдістерінің – Telegram ботының, табиғи ортадағы жазбалардың және дауыс жазатын құрылғыны пайдаланып бақыланатын жазбалардың қолданылуы балалар сөйлеуін талдауға жарамды

жоғары сапалы, жақсы құрылымдалған корпустың пайда болуын қамтамасыз етті. Бұл корпусты жас ерекшеліктеріне қарай оқыту және талдау үшін де пайдалануға болады.

Мысалы, 3-4 жас аралығындағы балаларға жер, ағаш және алма сияқты қарапайым және таныс сөздер берілді. 5-6 жас тобында сөздік неғұрлым сипаттамалы бола түсті, мысалы, жабайы жануар және тәтті, бұл мәнерлі тілді дамытуға көмектеседі. Ал сөйлем құрастыру дағдыларын қолдау үшін 7-8 жас тобында толық сөйлемдер пайдаланылды, мысалы, орманда неше түрлі жидектер өседі немесе достар күшікпен серуендеуге шықты.

Біз жас топтарына қосымша, шанышқы және бүлдіршін сияқты жасына қарамастан қиын болуы мүмкін күрделі сөздер үшін жеке санат жасадық. Бұл санаттау педагогикалық және логопедиялық принциптерге негізделген, олар сөздікті баланың даму кезеңіне сәйкестендіруді маңызды деп санайды.

Сөздіктің жас топтарына бөлінуі және күрделі сөздерге арналған санаттың қосылуы оқу процесін әр топтың түсіну қабілетіне сәйкес келетіндей етіп құрылымдауға мүмкіндік берді.

Біз жинаған сөйлеу үлгілерінің көбі, соның ішінде 2.4-кестеде көрсетілген мысалдар смартфондарды пайдаланып жазылды. Бұл әдіс нақты жағдайда, әсіресе үй жағдайында немесе мектепке дейінгі мекемелерде отбасылармен жұмыс істегенде, ең ыңғайлы әдістердің бірі болып шықты. Смартфондарды пайдалану арнайы жабдықты қажет етпеуі және әртүрлі жас топтарындағы балалардың табиғи, ойда жоқта сөйлеуін жазуға мүмкіндік беретіні үшін оңай болды. Олардың қолжетімділігі тіпті аз ресурсты жағдайларда да неғұрлым теңгерімді және әртүрлі деректерді жинауға мүмкіндік берді.

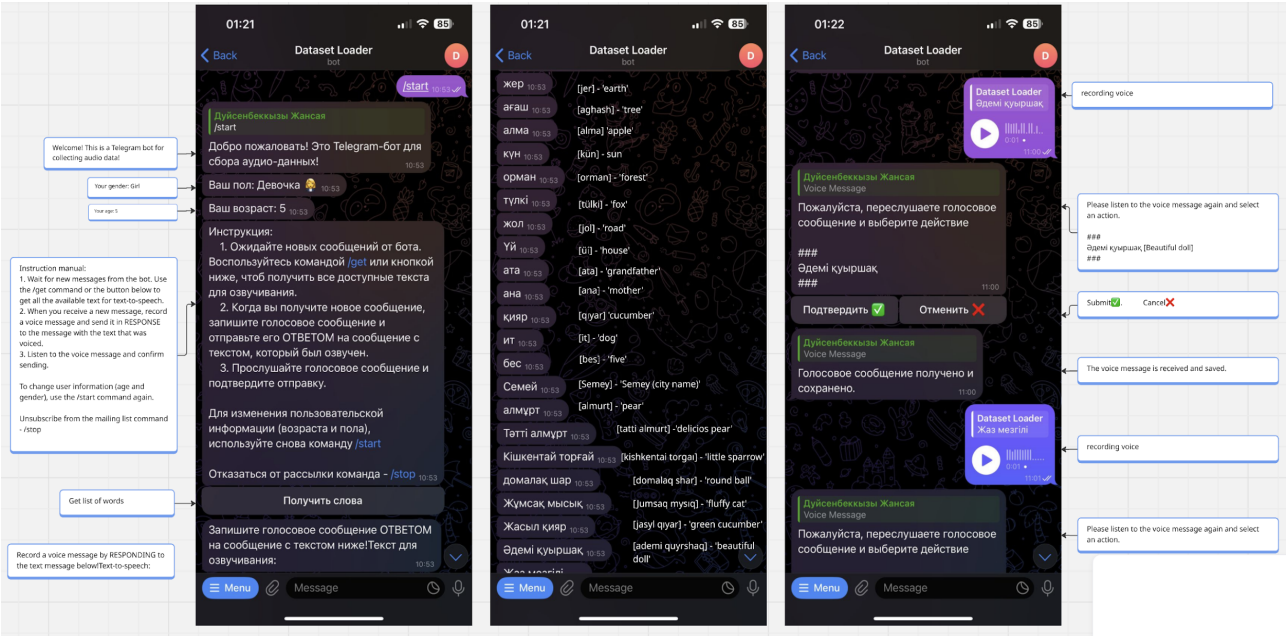
2.4-кесте. Сөздіктің жас топтары мен күрделілік деңгейлері бойынша үлгілік таралуы

№	3-4 жас	5-6 жас	7-8 жас	Қиын сөздер
1	жер	жабайы жануар	жабайы жануар зообағы	шанышқы
2	ағаш	тәтті қызанақ	жазда барлық көкөністер піседі	бүлдіршін
3	алма	жеміс жидек	қызыл алма ағашта өсіп тұр	орам жапырақ
4	күн	қызыл қалам	бүгін Алматыда күн ыстық болып тұр	қаражидек
5	орман	ақ аю	орманда неше түрлі жидектер өседі	сексен сегіз

Зерттеуге қатысушылар мақсатты іріктеу (*purposive sampling*) қағидаты негізінде таңдалды. Іріктеудің негізгі критерийлері келесідей анықталды: (1) қатысушылардың жасы 2–11 жас аралығында болуы; (2) қазақ тілін ана тілі немесе бірінші тіл ретінде қолдануы; (3) ата-анасының немесе заңды өкілінің жазбаша келісімінің болуы.

Жыныстық құрамның теңгерімділігін қамтамасыз ету мақсатында ұлдар мен қыздардың үлесі тиісінше 49% және 51% деңгейінде сақталды. Сөйлеу бұзылыстары бар балалар жеке санат ретінде қарастырылып, олардың іріктелуі логопед мамандардың алдын ала қойылған кәсіби диагноздары негізінде жүзеге асырылды.Жиналған аудиожазбалардың транскрипциялары екі кезеңнен тұратын сапаны бақылау рәсімінен өтті.Бірінші кезең — автоматтандырылған тексеру, оның барысында аудиожазбалардың техникалық параметрлері, атап айтқанда аудио ұзақтығы, сигнал-шу қатынасы (SNR ≥ 10 дБ), сондай-ақ метадеректердің толықтығы жүйелі түрде тексерілді. Екінші кезең — сараптамалық (адам) тексеру, онда зерттеу тобының мүшелері және кемінде бір қостілді (қазақ–орыс немесе қазақ–ағылшын) лингвист сарапшы барлық жазбалардың 100%-ын тыңдап, транскрипциялардың дұрыстығын растады. Сарапшылар арасындағы келісімділік деңгейі Коэннің каппа коэффициенті арқылы бағаланып, $\kappa \geq 0.85$ мәніне жетті. Бұл көрсеткіш аннотация сапасының жоғары деңгейде екенін дәлелдейді.

Деректерді жинауды бастау үшін біз Dataset Loader Bot деп аталатын арнайы Telegram ботын әзірледік, ол балалардың сөйлеу деректерін жүйелі түрде жинауға және ұйымдастыруға арналған. Жазу процесі сөздер мен сөз тіркестерінің тізімі ботқа жүктеліп, жаңа мазмұнды көрсету үшін апта сайын жаңартылып отыратын нақты құрылымды ұстанды. Бұл әдіс аудио деректерді жас және тақырып бойынша тиімді және ұйымдасқан түрде жинауға мүмкіндік берді, бұл оны кейіннен білім беру немесе сөйлеуді тану қолданбаларында пайдалануға жарамды етті. Төмендегі суреттер Telegram-дағы Dataset Loader Ботының жұмыс процесінің әр кезеңінің егжей-тегжейлі бейнеленуін ұсынады.



2.4-сурет. Dataset Loader Bot Telegram боты арқылы аудио деректерді жинау процесінің бейнесі

Сурет 2.4-те көрсетілгендей, Dataset Loader Bot пайдаланушыға қадамдық нұсқаулар арқылы аудиожазбаны жүктеуге мүмкіндік береді, ал жүйе автоматты

түрде сигнал сапасын тексеріп, жас пен жыныс сияқты метадеректерді тіркейді. Жинақталған деректер бірыңғай WAV форматына түрлендіріліп, идентификатор жүйесі бойынша құрылымдалған деректер базасына сақталады. Ұсынылған жинау схемасы географиялық шектеусіз қашықтан деректер жинауға мүмкіндік беріп, корпустың масштабталуын қамтамасыз етеді.

2.2.2 Аннотациялау және метадеректер

Жинаған деректер негізінде бала аудиториясына бейімделген сөздер мен сөз тіркестерінен тұратын құрылымдық дерекқор жасалды. Әр пайдаланушыға бірегей идентификатор (ID) тағайындалады, оның астында барлық аудиожазбалар сақталады. Әрбір жазба пайдаланушының ID-імен байланыстырылады және онымен бірге баланың жасы мен жынысы, сондай-ақ жазылған нақты сөз немесе сөз тіркесі сияқты метадеректер жүреді.

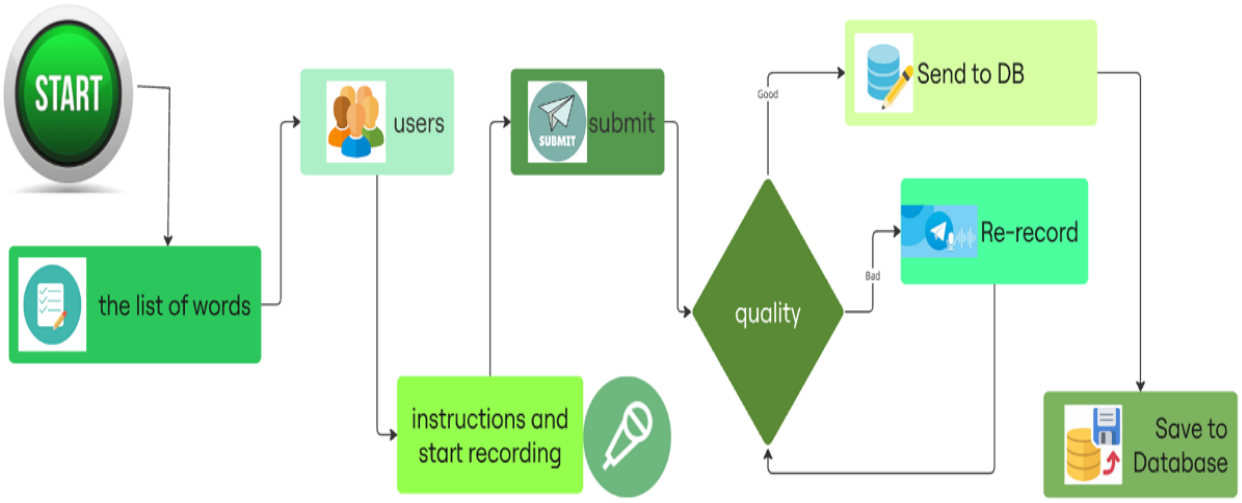
Бұл процесс әрбір жазбаның ұйымдасқан түрде сақталуын және жас тобы, тақырып сияқты тиісті метадеректермен байланыстырылуын қамтамасыз етеді. Бұл құрылым деректерді кейінірек білім беру құралдары мен зерттеулерде пайдалануды жеңілдетеді.

Dataset Loader Bot көмегімен 2 жастан 8 жасқа дейінгі балалардың сөйлеу деректерін қамтитын шамамен 800 аудиожазба жиналды. Корпустың жалпы көлемі шамамен 1000 мәтіндік жазбадан және шамамен 12 минуттық аудиоматериалдан тұрды. Дерекқорда шамамен 30 баланың (17 қыз және 9 ұл) жазбалары бар және бала аудиториясы үшін тақырыптық өзектілік пен тиісті күрделілік деңгейлері негізінде таңдалған 75 бірегей сөз бен сөз тіркесін қамтиды.Бұл құрылым білім беру және терапиялық бағдарламаларды әзірлеуге жарамды ақпараттық корпусты ұсынады. Telegram боты арқылы жиналған аудио деректердің пайдалануға жарамдылығын қамтамасыз ету үшін дыбыс сапасын жақсартудың бірнеше кезеңін қамтитын мұқият өңдеу жүргізіледі.Жазбалар бастапқыда OGG форматында сақталады, ол деректерді алып болғаннан кейін одан әрі оқыту мен талдау үшін қолайлырақ WAV форматына түрлендіріледі. 4-кестеде жиналған аудиожазбалардың егжей-тегжейлері, оның ішінде деректердің жас және жыныс бойынша таралуы, сондай-ақ әр жас тобы үшін жазбалар саны келтірілген.

2.5-кесте. Telegram боты арқылы жиналған балалардың аудио корпусы

Баланың жасы	**Қыздар**	**Ұлдар**	**Аудиожазбалар саны**
2 жас	3	0	40
3 жас	2	2	61
4 жас	2	1	47
5 жас	3	4	195
6 жас	3	2	146
7 жас	1	3	61
8 жас	3	3	179

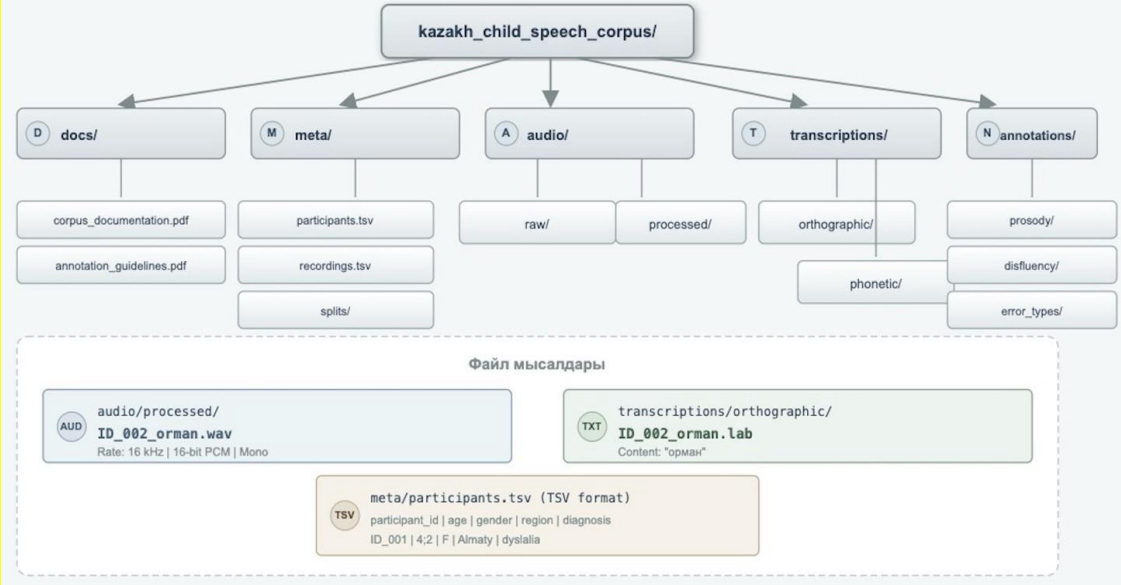
Бұл деректер корпустың құрамы және оның білім беру және зерттеу мақсаттарында пайдалануға жарамдылығы туралы түсінік береді. -суретте әр жас тобы үшін аудиожазбалар саны бейнеленген, қыздар мен ұлдар арасындағы бөлініс көрсетілген. Бұл деректердің таралуы корпустың жас және жыныс бойынша репрезентативтілігін оңай бағалауға, сондай-ақ одан әрі деректер жинау үшін ықтимал олқылықтарды анықтауға мүмкіндік береді (2.5-сурет).



2.5-сурет. Dataset Loader Bot жүйесіндегі деректерді жинау және өңдеу жұмыс процесінің блок-схемасы.

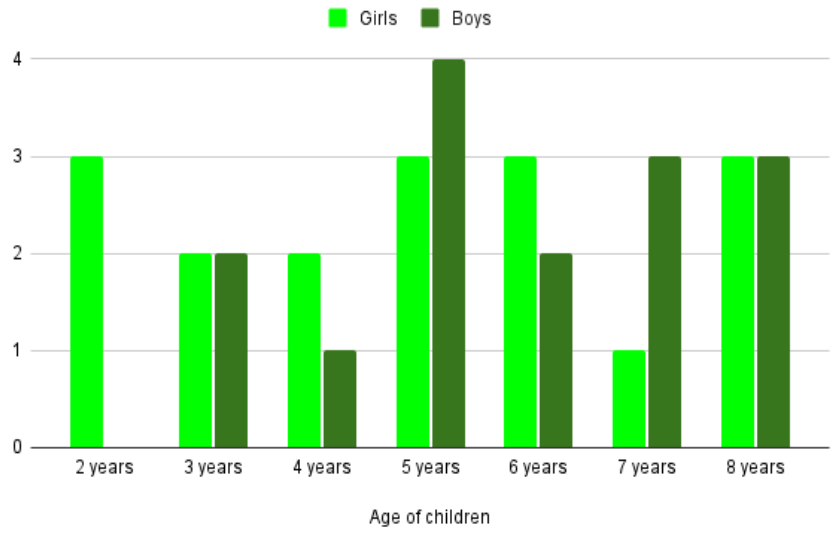
Метадеректер деңгейі корпустың маңызды құрамдас бөлігі болып табылады және спикерлік метадеректерді (жасы, жынысы, аймағы, ана тілі, тілдік ортасы), патологиялық метадеректерді (сөйлеу бұзылысының түрі – фонетика-фонематикалық жетілмеу, дислалия, дизартрия, кекештену, ауырлық дәрежесі, диагностика күні), сессиялық метадеректерді (жазба күні, жазу ортасы, жабдық түрі, тапсырма типі) және акустикалық метадеректерді (сигнал-шуыл қатынасы, жазба ұзақтығы, үлгі алу жиілігі) қамтиды.

Төмендегі кестеде аудиодеректердің толық сипаттамалары көрсетілген.



2.6-сурет. Файл мысалдары мен аудиодерек сипаттамалары

Транскрипция деңгейі балалардың сөйлеуін тану үшін көпдеңгейлі транскрипциялау жүйесін қажет етеді. Орфографиялық транскрипция сөздердің емлелік нұсқасын, фонетикалық транскрипция IPA стандарты негізіндегі дәл фонетикалық жазылымды, ал фонематикалық транскрипция фонемалар тізбегін береді. Фонетика-фонематикалық жетілмеуі бар балалардың сөйлеуін талдауда фонемаларды дыбыстау мен қабылдау кемістіктері ерекше ескеріледі. Мұндай жағдайда транскрипцияда дыбыстардың ауыстырылуы, түсірілуі, бұрмалануы белгіленеді. Аннотация деңгейі сөйлеу бұзылыстарын сипаттайтын көрсеткіштерді қамтиды: дыбысталу сапасының метрикалары (Goodness of Pronunciation индексі), сөйлеу кемістіктерінің типтері (дыбысты ауыстыру, дыбысты түсіріп айту, дыбысты қосу, буынның түсірілуі, қайталау, тартыну) және просодиялық көрсеткіштер (тонның жоғарылауы/төмендеуі, паузалар, сөйлеу қарқыны).



2.7-сурет. Аудиожазбалардың балалардың жасы мен жынысы бойынша таралуы

Осылайша, корпустың жас құрылымы мен жыныстық құрамы әртүрлі жас аралықтарындағы сөйлеу деректерінің кең ауқымын қамтамасыз етеді, бұл оны балалар сөйлеуін оқыту және талдаумен байланысты тапсырмалар үшін құнды етеді [68-69].

2.2.3 Этикалық және құқықтық аспектілер

Қазақстан Республикасының «Дербес деректер және оларды қорғау туралы» Заңы (№ 94-V) деректерді жинау мен өңдеудің құқықтық негізін белгілейді. Заңның 7-бабына сәйкес, дербес деректерді, оның ішінде биометриялық деректерді (дауыс жазбалары) жинау және өңдеу тек дербес деректер субъектісінің немесе оның заңды өкілінің келісімімен жүзеге асырылады. 1-бапта биометриялық деректерге анықтама берілген: бұл – дербес деректер субъектісінің физиологиялық және биологиялық ерекшеліктерін сипаттайтын, оның негізінде жеке басын анықтауға мүмкіндік беретін деректер[70].

Заңның 8-бабы келісім беру тәртібін нақты регламенттейді. Келісім жазбаша нысанда берілуі тиіс және оған мынадай міндетті мәліметтер енгізіледі: оператордың атауы, субъектінің тегі мен аты, келісімнің қолданылу мерзімі, деректерді үшінші тұлғаларға беру мүмкіндігі туралы ақпарат, деректердің шекарааралық тасымалдануы, жиналатын деректер тізімі. Зерттеу барысында әр баланың ата-анасы немесе заңды өкілі арнайы әзірленген ақпараттандырылған келісім парағына (informed consent form) қол қояды, бұл қолхат арқылы келісім беру фактісін растайды. Келісім парағында зерттеудің мақсаты, деректерді пайдалану тәртібі, құпиялылық шаралары, сондай-ақ ата-ананың кез келген уақытта келісімді қайтарып алу құқығы көрсетіледі [71].

Заңның 24-бабының 1-тармағына сәйкес, дербес деректер субъектісінің (яғни баланың ата-анасының) келісімді қайтарып алуға, деректердің өзгертілуін немесе жойылуын талап етуге, сондай-ақ өз құқықтары мен заңды мүдделерін қорғауға құқығы бар. Бұл ереже зерттеуге қатысушылардың еріктілігін және олардың кез келген сәтте зерттеуден шығып кету мүмкіндігін қамтамасыз етеді.

Заңның 11-бабы дербес деректердің құпиялылығын қамтамасыз етуге міндеттейді. Деректерге қол жеткізген иелері мен операторлары оларды таратуға жол бермеуі тиіс. Осы талапты орындау мақсатында, жинақталған аудиодеректер депсонализацияланады: әр балаға бірегей код (мысалы, ID_001) беріледі, ал атаулар мен кодтар сәйкестігі зерттеушілердің шектеулі санына ғана қолжетімді бөлек файлда сақталады. Заңның 12-бабының 2-тармағына сәйкес, дербес деректер Қазақстан Республикасының аумағында орналасқан дерекқорда сақталуы тиіс. Сақтау мерзімі деректерді жинау мен өңдеу мақсаттарына жеткен күнмен айқындалады; осы зерттеуде бұл мерзім зерттеу аяқталғаннан кейін үш жылды құрайды [72].

Заңның 18-бабы деректерді жою шарттарын белгілейді. Деректер сақтау мерзімі аяқталғаннан кейін, дербес деректер субъектісі мен оператор арасындағы құқықтық қатынастар тоқтатылғаннан кейін, сот шешімі заңды күшіне енгеннен кейін, сондай-ақ дербес деректер субъектісінің келісімінсіз жиналған және өңделген деректер анықталған жағдайда деректер жойылуға жатады.

Бала құқықтарын қорғау тұрғысынан, БҰҰ Бала құқықтары туралы конвенциясының (1989 ж., 44/25 резолюциясы) 3-бабына сәйкес, баланың мүддесі барлық іс-әрекеттерде басты назарда болуы тиіс. 12-бап балаға өз көзқарасын еркін білдіруге мүмкіндік беруді талап етеді, сондықтан зерттеуге қатысу үшін ата-ананың жазбаша келісімімен қатар, баланың жасына сәйкес ауызша келісімі де алынады. 16-бап баланың жеке өміріне ерікті немесе заңсыз араласуға жол бермейді, бұл деректерді құпия өңдеуді міндеттейді.

Қазақстан Республикасының «Баланың құқықтары туралы» Заңы (№ 345-II) 7-бабында баланың жеке өміріне, отбасына қол сұғылмау және жеке басының құпиялылығы сақталатыны көрсетілген. 8-бап ата-аналардың баланы тәрбиелеу және дамыту үшін бірінші кезектегі жауапкершілігін белгілейді, бұл олардың баланың деректерін беруге келісім беру құқығы мен міндетін білдіреді.

Осы заңнамалық талаптардың жиынтығы зерттеу барысында жиналған балалардың аудиодеректерінің құқықтық қорғаумен қамтамасыз етілуіне, ата-аналардың қолхат арқылы берген келісімі деректерді өңдеудің заңды негізі ретінде қызмет етуіне, ал жиналған деректердің құпиялылығы мен қауіпсіздігінің сақталуына кепілдік береді. Бұл ретте, «Дербес деректер және оларды қорғау туралы» Заңның 5-бабында көрсетілген заңдылық, құпиялылық және теңдік принциптері сақталады. 20-бапқа сәйкес, дербес деректерді қорғауға мемлекеттік кепілдік беріледі, ал 21-бапта қорғаудың мақсаттары – жеке өмірге қол сұғылмау құқығын іске асыру, деректердің тұтастығы мен қауіпсіздігін, құпиялылығын қамтамасыз ету, оларға қолжетімділік құқығын жүзеге асыру және заңсыз жинау мен өңдеудің алдын алу ретінде айқындалған.

Ата-аналардың жазбаша келісімі (қолхат) Заңның 8-бабының талаптарына сәйкес келеді және деректерді өңдеудің негізгі шарты болып табылады. Келісім парағында деректердің қандай мақсатта, қандай көлемде және қандай мерзімге жиналатыны нақты көрсетілуі тиіс. Зерттеуге қатысушылардың құқықтарын қорғау үшін, Заңның 24-бабына сәйкес, ата-аналарға кез келген уақытта келісімді қайтарып алу және баласының деректерін жоюды талап ету мүмкіндігі беріледі. Бала құқықтары туралы конвенцияның 3 және 12-баптарына сәйкес, ата-ананың келісімімен қатар, баланың жас ерекшелігіне қарай ауызша келісімі де алынады, бұл баланың зерттеу субъектісі ретіндегі емес, тең құқылы қатысушы ретіндегі мәртебесін бекітеді.

Сонымен қатар, 11-бап деректердің құпиялылығын сақтауды міндеттейді. Бұл талапты орындау мақсатында жиналған аудиодеректер депсонализацияланады (әр балаға бірегей код беріледі) және оларға қолжетімділік тек зерттеу тобының уәкілетті мүшелеріне ғана рұқсат етіледі. 12-баптың 2-тармағына сәйкес, деректер Қазақстан Республикасының аумағында орналасқан қауіпсіз серверлерде сақталады. 18-бапта белгіленген тәртіппен зерттеу аяқталғаннан кейін немесе келісім мерзімі өткеннен соң деректер толығымен жойылады. Осы шаралардың барлығы зерттеудің этикалық және құқықтық нормаларға сәйкестігін, сондай-ақ жиналған деректердің ғылыми айналымда тек депсонализацияланған түрде пайдаланылуын қамтамасыз етеді.

Зерттеу барысында этикалық және құқықтық талаптар толық сақталып, барлық аудиодеректер ата-аналардың немесе заңды өкілдердің келісімімен жиналып, анонимизациялау және қауіпсіз сақтау шаралары қамтамасыз етілді.

Осылайша, 2.2 бөлімде ұсынылған әдіснамалық тәсіл қазақ тіліндегі балалар сөйлеуінің сапалы акустикалық корпусын қалыптастыруға негіз болып табылады және келесі бөлімдерде қарастырылатын деректерді алдын ала өңдеу және сөйлеуді тану модельдерін оқыту процестеріне қажетті сенімді деректер базасын қамтамасыз етеді.

2.3. Деректер құрылымы

Балалардың сөйлеуін тануға арналған корпустың деректер құрылымы оның тиімді пайдаланылуын, автоматты өңделуін және халықаралық стандарттарға сәйкестігін қамтамасыз етуі тиіс. Деректер құрылымын әзірлеу кезінде балалар сөйлеуінің акустикалық және лингвистикалық ерекшеліктері, сондай-ақ сөйлеу бұзылыстарын анықтауға бағытталған талдау әдістері ескерілді.

Сөйлеу корпустарын құрудың әдіснамасы лингвистикалық ресурстарды стандарттау және метадеректерді басқару жүйелерінің эволюциясымен тығыз байланысты. Балалар сөйлеуін тануға арналған корпустарды әзірлеу ересектер сөйлеуіне арналған корпустардан айтарлықтай ерекшеленеді, себебі балалардың сөйлеуінде жас ерекшеліктеріне байланысты жоғары өзгергіштік (variability) байқалады. Зерттеулер көрсеткендей, стандартты Mel-frequency cepstral coefficients (MFCC) параметрлері балалар сөйлеуіндегі pitch әсеріне осал болып келеді. Бұл мәселені шешу үшін спикердің pitch деңгейіне байланысты жоғары жиілікті компоненттерді сүзу және балалар сөйлеуінің pitch деңгейін ересектер диапазонына жақындату әдістері қолданылады. Сонымен қатар, сөйлеу жылдамдығындағы айырмашылықтарды есепке алу және дауыс жолының ұзындығын нормалау (Vocal Tract Length Normalization) спектралды вариацияларды өтеу үшін қажет [73].

Балалардың сөйлеу корпусын құруда келесі теориялық принциптер маңызды болып табылады. Біріншіден, жас ерекшеліктерін стратификациялау принципі – 2-3 жас аралығындағы балалар 30 минуттық ойын барысында 25-28 сөз шығарса, 3-4 жас аралығында 70-80 сөзге, 4-4,5 жаста 110-115 сөзге жетеді. Сөйлеу дамуындағы кідірістерді бағалау 3-4 жас аралығында ең тиімді, себебі бұл кезеңде сөйлеу қабілеттері толық дамыған болуы күтіледі. Екіншіден, фонетикалық және фонематикалық толықтық принципі – қазақ тілінің фонетикалық жүйесінде 9 дауысты және 24 дауыссыз фонема бар. Балалардың сөйлеуін тануда әр фонеманың сөздің әртүрлі позицияларында (басы, ортасы, соңы) қамтылуы маңызды. Сөйлеу бұзылыстары бар балаларда фонематикалық қабылдаудың жетілмеуі ана тілінің дыбыстау жүйесін қалыптастыру процесінің бұзылуына әкеледі. Үшіншіден, просодиялық ерекшеліктерді есепке алу принципі – балалардың дауысы жоғары тембрмен және жиілік диапазонының жасқа байланысты айтарлықтай өзгеруімен сипатталады. Pitch таралуы балаларда ересек ерлер мен әйелдерге қарағанда әлдеқайда кең. Бұл айырмашылықтар арнайы шешімдерді қажет етеді, өйткені балалар сөйлеуінің

жоғары жиілікті сипаттамалары тиімді ASR модельдерін құру үшін жеткіліксіз. Төртіншіден, интер-аннотатор сенімділік принципі – деректерді таңбалау сапасын бағалау үшін проценттік келісім және kappa мәндері қолданылады. Зерттеулер көрсеткендей, барлық аннотация деңгейлерінде проценттік келісім 86.4% пен 98.6% аралығында, ал kappa мәні 0.72 мен 0.97 аралығында болады.

Балалардың сөйлеуін тануға арналған корпус келесі иерархиялық деңгейлерден тұрады. Акустикалық деректер деңгейі шикі аудиожазбаларды және оларды алдын ала өңдеу нәтижелерін қамтиды. Жазбаларда сөйлеу бұзылыстарының түрлеріне қарай дыбыстардың ауыстырылуы, түсірілуі, бұрмалануы сияқты акустикалық ерекшеліктер фиксацияланады. Аудиодеректерді жинау кезінде физикалық жазу параметрлері (микрофон түрі мен орналасуы, шу деңгейі, реверберация) сөйлеуді тану дәлдігіне тікелей әсер етеді. Қашықтан микрофонды пайдалану сигнал-шуыл қатынасын -7.8 дБ-тан -19.9 дБ-қа дейін жақсартады.

Қазақ тілінің фонетикалық жүйесінің ерекшеліктерін ескере отырып, корпустың метадеректер жүйесінде 24 дауыссыз және 9 дауысты фонемалардың позициялық таралуы ескеріледі – әрбір фонема сөздің басында, ортасында және соңында кездесуі тиіс. Қазақ тілінің дауысты дыбыстары сингармонизмге бағынады. Балалардың сөйлеу бұзылыстары көбінесе дауысты дыбыстардың дұрыс үндесуінің бұзылуынан көрінеді, сондықтан корпус құрылымында бұл ерекшелік бөлек ескеріледі. Балалар сөйлеуіндегі просодиялық бұзылыстарды талдау үшін тон маркерлері (жоғарылау, төмендеу, тұрақты тон) және пауза маркерлері (қысқа пауза, ұзақ пауза) қолданылады.

Балалар сөйлеу корпусының деректер интегритетін қамтамасыз ету үшін келесі принциптер сақталады. ID жүйесінің бірегейлігі – әрбір қатысушыға, сессияға және жазбаға бірегей идентификатор беріледі. Форматтық сәйкестік – барлық аудиофайлдар WAV форматында (PCM, 16 кГц немесе 48 кГц, 16 бит, моно) сақталады. Депсонализация және құпиялылық – әрбір қатысушыға бірегей код беріледі, атаулар мен кодтар сәйкестігі зерттеушілердің шектеулі санына ғана қолжетімді бөлек файлда сақталады. Аннотациялар аралық сенімділікті бағалау – аннотациялардың екі тәуелсіз маманмен сәйкестігі тексеріледі, проценттік келісім және Cohen's kappa коэффициенті есептеледі.

Балалардың сөйлеуін тануға арналған корпустың сандық сипаттамалары келесідей: жалпы жазба ұзақтығы ~6 сағат (2-8 жас аралығындағы балалардан жиналған), қатысушылар саны ~400 бала (жынысы, жасы, аймағы бойынша теңгерілген), жазбалар саны ~6,500 (сөздер, сөз тіркестері, сөйлемдер), жас топтары 2-3, 3-4, 4-5, 5-6, 6-8 аралықтары, сөздік қор көлемі ~500 сөз (жас ерекшеліктеріне бейімделген лексика), файлдар көлемі ~8 ГБ (өңделген) және ~12 ГБ (шикі) WAV форматында.

Құрылымдық тұрғыдан дұрыс ұйымдастырылған корпус келесі міндеттерді шешуге мүмкіндік береді. Біріншіден, сөйлеу бұзылыстарын автоматты анықтау – корпустың аннотациялық бөлігі дыбыстардың ауыстырылуын, түсірілуін, бұрмалануын анықтайтын модельдерді оқытуға негіз болады. Екіншіден, жасөспірімдік және ересек сөйлеуге бейімдеу – жинақталған балалар сөйлеу

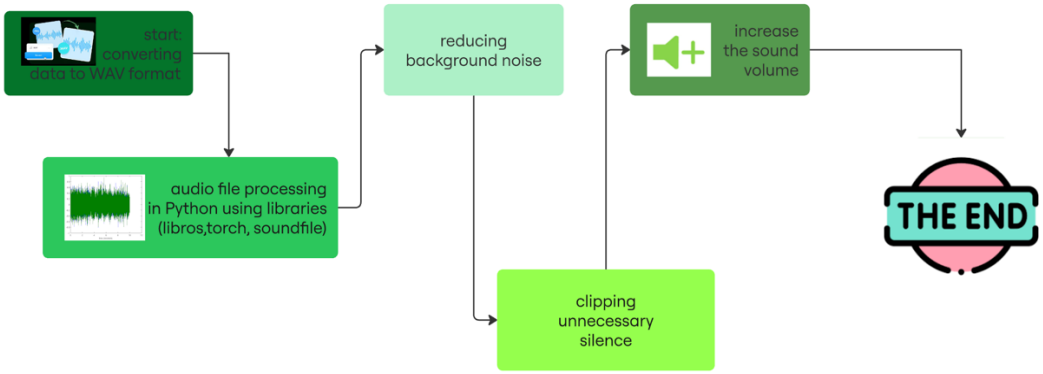
корпусы одан әрі үлкен жас топтары үшін модельдерді бейімдеуге пайдаланылады. Үшіншіден, көптілді жүйелермен интеграция – корпустың ашық құрылымы басқа тілдердегі балалар сөйлеу корпустарымен салыстырмалы зерттеулер жүргізуге мүмкіндік береді. Төртіншіден, терапиялық әдістердің тиімділігін бағалау – корпус сөйлеу бұзылыстарын түзетуге бағытталған логопедтік бағдарламалардың нәтижелілігін сандық бағалауға мүмкіндік береді. Балалардың сөйлеуін танудың тиімділігін бағалау үшін Word Error Rate, BLEU, Translation Edit Rate, Character Similarity Rate және Accuracy метрикалары қолданылады. Зерттеу нәтижелері сөйлеу бұзылыстары бар балалардың сөйлеуін тану үшін жас ерекшеліктеріне бейімделген корпустардың және домендік бейімдеудің маңыздылығын көрсетеді.

2.4 Деректерді алдын ала өңдеу

Алдын ала өңдеу үдерісі бастапқы аудиожазбалардың сапасын жақсартуға, оларды стандартталған форматқа келтіруге және кейінгі машиналық оқыту модельдеріне енгізуге бейімдеуге бағытталған маңызды кезең болып табылады.

Аудиожазбаларды өңдеу әртүрлі кітапханалар мен алгоритмдерді қолдана отырып кезең-кезеңімен жүргізілді, бұл деректерді талдауға дайындауға және сөйлеуді тануға арналған оқыту модельдерін генерациялауға мүмкіндік берді. Барлық файлдар шуды кетіру, артық үзілістерді қысқарту және дауыс деңгейін арттыру функцияларын қамтитын скрипт арқылы өңделді. Скрипт PyCharm 2024.1.4 ортасында Python 3.11.9 тілінде келесі кітапханаларды қолдана отырып жазылды: librosa– аудиосигналдарды оқу, өңдеу және талдау үшін; torch– машиналық және терең оқытуға арналған фреймворк; soundfile – аудиофайлдарды оқу және жазу үшін; noisereduce – аудиожазбалардан шуды кетіру үшін; және os – файлдық жүйемен жұмыс істеу үшін. Шуды азайту және үзілістерді қысқартуды қамтитын өңдеуден кейін сөздер арасындағы паузалардың алынуына байланысты аудиожазбалардың көлемі азайды.

Деректерді өңдеу сапасын бағалау және өңделген аудиожазбаларды визуалды талдау үшін сигналдың жиілік мазмұнын визуализациялауға мүмкіндік беретін спектралды талдау әдісі қолданылды. Зерттеуде Қысқамерзімді Фурье түрлендіруі (STFT) пайдаланылды, ол сигналды әртүрлі жиіліктегі синусоидалық құрамдас бөліктерге ыдыратып, аудиосигналдағы жиілік құрамдас бөліктерінің болуын және амплитудасын анықтауға мүмкіндік береді. 2.8-суретте өңдеу кезеңдері көрсетілген, бұл жұмыс процесін нақтырақ түсінуге мүмкіндік береді.



2.8-сурет. Аудиодеректерді өңдеудің негізгі кезеңдері

Аудиодеректерді өңдеу барысында дыбыс сапасын жақсарту және файлдарды талдауға дайындау үшін іс-әрекеттер тізбегі орындалады. Әр кезеңнің сипаттамасы WAV форматына бастапқы конвертациялаудан бастап төменде берілген.

Әрбір аудиожазбаны (жеке сөздер, сөз тіркестері немесе сөйлемдер) алдын ала өңдеу келесідей жүргізілді: алдымен барлық аудиофайлдар WAV форматына конвертацияланды. Бұл формат қысылмағандықтан жоғары дыбыс сапасын сақтайды. Көптеген жазбаларды тиімді және жылдам конвертациялау үшін VSDC конвертер қосымшасы (нұсқа 6.8.6.352, Flash-Integro LLC) пайдаланылды.

Фондық шуды кетіру және дауыс деңгейін теңестіру үшін Librosa, Soundfile, Torch және Noisereduce кітапханалары қолданылды. process_audio әдісінде аудиофайлдар librosa.load көмегімен жүктелді, содан кейін noisereduce.reduce_noise арқылы шу басылды, және librosa.effects.trim көмегімен үзілістер қысқартылды. Осыдан кейін gain функциясы арқылы аудио деңгейі арттырылып, өңделген файл soundfile.write арқылы .wav форматында сақталды. process_audio функциясы өңдеуге арналған аудиофайлдары бар каталогты қабылдайды және өңделген файлдар белгіленген бумадағы барлық аудиофайлдарды топтық өңдеу үшін циклде шақырылатын буманың жолын қайтарады[74]. Бағдарлама коды 2.9-суретте көрсетілген.

```python
import os
import librosa
import numpy as np
from scipy.signal import medfilt
import soundfile as sf
import noisereduce as nr

def process_audio(input_path, output_path, gain=1.5):
    # Load audio
    y, sr = librosa.load(input_path)
    # Apply noise reduction algorithm
    y_denoised = nr.reduce_noise(y=y, sr=sr, prop_decrease=0.8)
    # Trim leading and trailing silence
    trimmed_y, _ = librosa.effects.trim(y_denoised)
    # Increase volume by a given gain factor
    y_louder = trimmed_y * gain
    # Save the processed audio to a new WAV file
    sf.write(output_path, y_louder, sr)

# Path to the folder containing original audio files
input_folder = r'C:/Users/Desktop/wav отдельные'

# Path to the folder where processed audio files will be saved
output_folder = r'C:/Users/Desktop/wav обработанные 5'

# Create output folder if it doesn't exist
if not os.path.exists(output_folder):
    os.makedirs(output_folder)

# Loop through all audio files in the input folder
for filename in os.listdir(input_folder):
    if filename.endswith('.wav'):  # Assumes all files are in WAV format
        input_path = os.path.join(input_folder, filename)
        output_path = os.path.join(output_folder, filename)
        process_audio(input_path, output_path, gain=1.5)
```

2.9-сурет. Шуды кетіру, үзілістерді қысқарту және дауыс деңгейін арттыру функцияларын қолдана отырып, аудиофайлдарды өңдеуге арналған бағдарлама коды

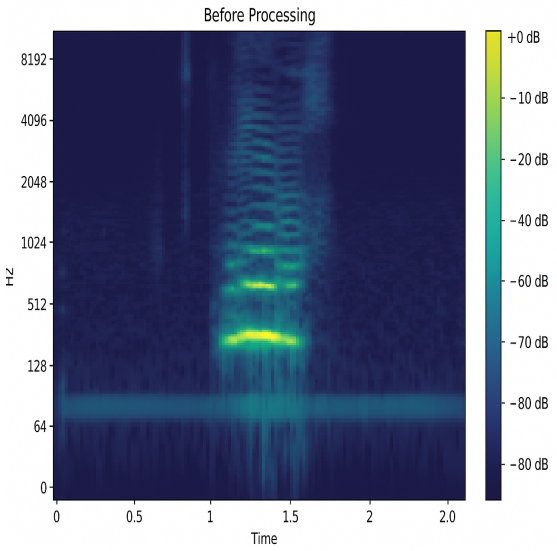
Жазбаны үздіксіз ету үшін librosa.effects.trim көмегімен паузалар алынып тасталды. Бұл функция жазбалардың басындағы және соңындағы қажетсіз үнсіздікті қысқартып, тек негізгі сигналды ғана қалдырды. Сонымен қатар, Librosa кітапханасының gain әдісі арқылы жазба күшейтілді, бұл дыбыстың анықтығын арттырды. 2.6-кестеде скриптті қолдана отырып, өңдеуге дейінгі және өңдеуден кейінгі аудиофайлдардың негізгі параметрлері көрсетілген. Негізгі назар паузалардың ұзақтығына, сигналдың деңгейіне және фондық шу деңгейіне аударылған.

2.6-кесте. Сапаны жақсарту үшін өңдеуден кейінгі аудиофайл параметрлерінің өзгерісі

Параметр	Өңдеуге дейін	Өңдеуден кейін
Пауза ұзақтығы	Ұзақ паузалар	Қысқартылған паузалар
Сигнал деңгейі	Төмен	Жоғары
Орташа шу деңгейі	Жоғары	Төмен

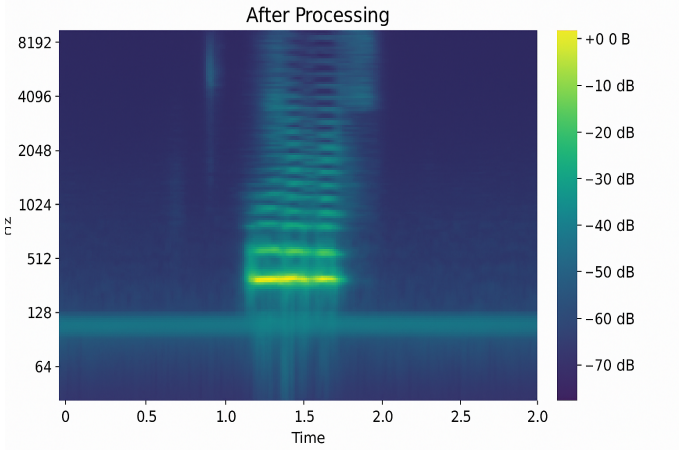
Аудиоөңдеу скриптінің тиімділігін жақсырақ түсіну үшін бірдей аудиожазбаның өңдеуге дейінгі және өңдеуден кейінгі спектрограммалары жасалды. Жазба ретінде 4 жастағы баланың айтқан «орман» сөзі пайдаланылды.

Оның ұзақтығы 2 секунд болды. 2.10-суретте осы сөздің өңдеуге дейінгі спектрограммасы көрсетілген.



2.10-сурет. «Орман» сөзінің өңдеуге дейінгі спектрограммасы

2.11-суретте сол сөздің өңдеуден кейінгі спектрограммасы көрсетілген.



2.11-сурет. «Орман» сөзінің өңдеуден кейінгі спектрограммасы

Екі суретті салыстыру өңдеуден кейін спектрограмманың контурлары айқынырақ болғанын және кейбір шашыраңқы шу нүктелері жойылғанын көрсетеді, бұл фондық шудың төмендегенін көрсетеді. Сонымен қатар, сары сызықтар (баланың дауысын білдіреді) спектрограмманың басына қарай аздап ығысқан, бұл жазбаның басындағы қажетсіз үнсіздіктің қысқартылғанын көрсетеді.

Дауысты фондық дыбыстардан бөлу үшін HPSS әдісі (Harmonic-Percussive Source Separation) қолданылды. Бұл әдіс, librosa.effects.hpss() ішінде жүзеге асырылған, аудиофайлды гармоникалық (негізінен дауыс) және перкуссиялық (негізінен фондық шу) құрамдас бөліктерге бөледі. Нәтижесінде таза сөйлеу сигналы алынады, бұл әсіресе сөйлеуді тану жүйелерін оқыту үшін маңызды.

Спектрограмма – бұл аудиосигналдың спектралдық мазмұнының уақыт функциясы ретіндегі визуалды бейнесі. Ол бір уақытта үш негізгі айнымалыны көрсетеді: жиілік, амплитуда және уақыт. Басқаша айтқанда, сигналда әртүрлі

уақыт нүктелерінде қандай жиіліктер бар екенін көрсетеді. Көлденең ось уақытты, ал тік ось жиілікті, яғни дыбыс толқынының уақыт бірлігіндегі тербеліс санын білдіреді, Герцпен (Гц) өлшенеді; жиілік неғұрлым жоғары болса, дыбыс биіктігі соғұрлым жоғары қабылданады. Амплитуда дыбыстың күші немесе қаттылығын білдіреді және децибелдердегі (дБ) мәндер матрицасы ретінде беріледі [75]. Спектрограммада амплитуда түс немесе жарықтық арқылы визуализацияланады: берілген нүктеде түс неғұрлым қарқынды болса, сол уақыт пен жиілікте дыбыс соғұрлым қатты болады (2.7-кесте).

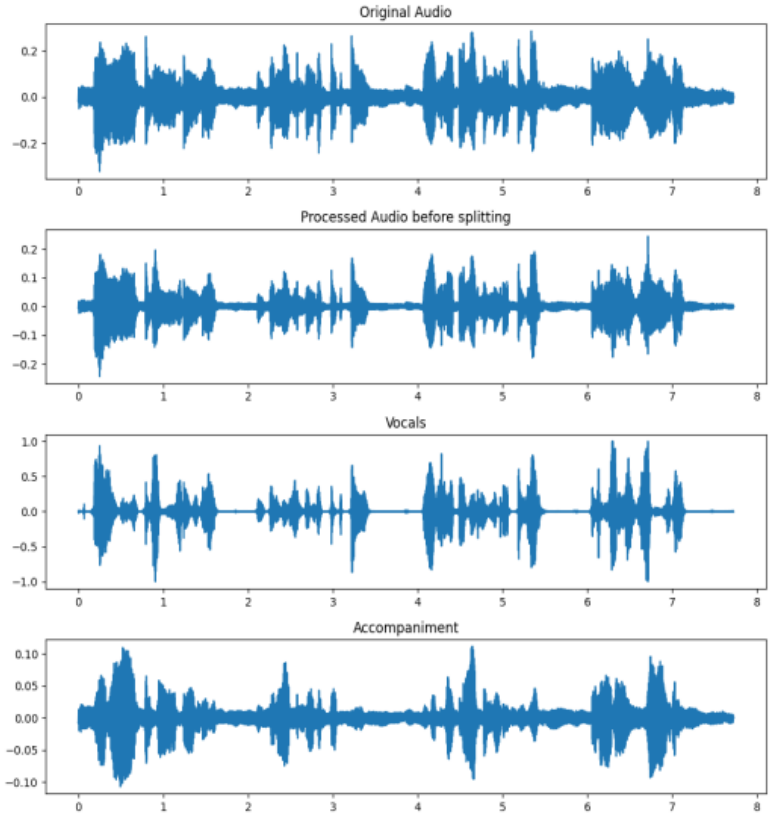
2.7-кесте. Спектрограмма генерациялау параметрлері

Параметр	Сипаттама
Терезе өлшемі	Спектрограмманың бір сегментіндегі үлгілер саны. Үлкен терезе төмен жиіліктерді көрсетеді.
Қадам ұзындығы	Уақыт сегменттері арасындағы қашықтық; уақыттық ажыратымдылықтың егжей-тегжейлілігіне әсер етеді.
Шкала түрі	Логарифмдік шкала сөйлеуді тану үшін маңызды төмен жиіліктерді жақсырақ көруге мүмкіндік береді.

Осы параметрлерді реттеу әртүрлі сигналдарды талдау үшін спектрограмманы оңтайландыруға мүмкіндік береді. Мысалы, терезе өлшемі мен қадам ұзындығын таңдау уақыттық және жиіліктік ажыратымдылықтың дәлдігіне әсер етеді, ал логарифмдік масштабтау сөйлеуді өңдеу тапсырмалары үшін маңызды төмен жиіліктердің көрінуін жақсартады.

Фильмдерден алынған балалардың сөйлеу жазбалары фондық музыканы кесіп тастауды, шуды азайтуды және дауыс деңгейін арттыруды қамтитын қосымша өңдеуді қажет етті. Қатты фондық музыкасы бар сегменттерді толығымен тазарту қиын болды, себебі музыканы алып тастау сөйлеудің бір бөлігінің жоғалуына әкеп соғуы мүмкін. Көпсатылы өңдеу нәтижесінде деректер көлемі 15 минуттан 9 минутқа дейін қысқарды. Сапаның төмендеуіне жол бермеу үшін өңдеу екі кезеңде жүргізілді: бастапқы шуды азайту және үзілістерді қысқарту, содан кейін дауыс пен фондық дыбыстарға бөлу.

2.12-суретте шуды азайтуды және дауысты фондық музыкадан ажыратуды қамтитын өңдеуге дейінгі және өңдеуден кейінгі аудиожазбаның визуализациясы көрсетілген.



2.12-сурет. «Орман» сөзінің өңдеуге дейінгі және өңдеуден кейінгі аудиожазбаның визуализациясы

Уақыттық график сигнал амплитудасының уақыт бойынша қалай өзгеретінін көрсетеді, бұл дыбыстың негізгі сипаттамаларын: ұзақтығын, қарқындылығын және динамикалық өзгерістерін бағалауға мүмкіндік береді.

Осылайша, алдын ала өңдеу кезеңі тек аудиосигнал сапасын жақсартып қана қоймай, деректер корпусының көлеміне де әсер етті. Өңдеуге дейінгі бастапқы корпус әртүрлі дереккөздерден жиналған 8594 аудиожазбадан тұрып, олардың жалпы ұзақтығы шамамен 8.5–9 сағатты құрады. Бұл көлемге Telegram-бот арқылы алынған жазбалар (≈1.5 сағат), табиғи ортадағы жазбалар (>4 сағат) және диктофон арқылы жазылған деректер (>3 сағат) кірді [76].

Алдын ала өңдеу барысында фондық шу жойылып, артық паузалар қысқартылып және сөйлеуге қатысы жоқ фрагменттер алынып тасталғандықтан, аудиожазбалардың жалпы ұзақтығы азайды. Нәтижесінде өңдеуден кейін Telegram-бот жазбаларының көлемі ≈1 сағатқа дейін, табиғи орта жазбалары >3 сағатқа дейін, ал диктофон жазбалары >2 сағатқа дейін қысқарды. Жалпы корпус көлемі шамамен 6–6.5 сағатты құрады (8-кесте).

Бұл өзгеріс деректердің шамамен 25–30%-ға қысқарғанын көрсетеді. Алайда бұл қысқару ақпараттың жоғалуымен емес, керісінше, мағынасыз немесе сапасы төмен сегменттердің жойылуымен байланысты. Нәтижесінде тек таза және мағыналы сөйлеу сигналдары сақталып, деректердің сапасы айтарлықтай жақсарды.

Осыған байланысты алынған өңделген корпус автоматты сөйлеуді тану модельдерін оқыту үшін анағұрлым тиімді және сенімді дереккөз ретінде қарастырылды.

2.8-кесте. Алдын ала өңдеуден кейінгі балалар сөйлеуінің датасет көлемі

Дерек көзі	**Алдын ала өңдеуге дейін**	**Алдын ала өңдеуден кейін**
Telegram-бот	≈ 3 сағат	≈ 2 сағат
Табиғи орта	> 7 сағат	> 4 сағат
Диктофон жазбалары	> 3 сағат	> 3 сағат
Жалпы көлем	≈ 12–13 сағат	≈ 8–9 сағат

Зерттеу барысында аудиодеректерді өңдеудің негізгі қадамдары ретінде шу деңгейін төмендету, сигналды нормализациялау, аудионы сегментациялау және акустикалық ерекшеліктерді (MFCC, log-Mel спектрограммалар) шығару әдістері қолданылды. Сонымен қатар, деректерді кеңейту (data augmentation) тәсілдері арқылы сөйлеу сигналдарының вариативтілігі жасанды түрде арттырылып, модельдердің тұрақтылығы мен жалпылау қабілетін жақсартуға мүмкіндік берілді [77].

Ұсынылған алдын ала өңдеу әдістері балалар сөйлеуінің ерекшеліктерін, атап айтқанда жоғары фонетикалық вариативтілікті және акустикалық тұрақсыздықты ескере отырып таңдалды. Бұл тәсілдер шуға төзімділікті арттыруға, сөйлеу сегменттерінің сапасын жақсартуға және модельге берілетін кіріс деректердің біркелкілігін қамтамасыз етуге ықпал етеді [78].

Осылайша, алдын ала өңдеу кезеңі акустикалық деректердің сапасын жүйелі түрде арттыруға мүмкіндік беріп, сөйлеуді автоматты тану модельдерінің тиімді оқытылуы үшін сенімді және стандартталған кіріс деректер жиынын қалыптастырады. Бұл өз кезегінде келесі бөлімдерде қарастырылатын ASR модельдерін әзірлеу және оңтайландыру процестерінің нәтижелілігін қамтамасыз етеді.

3 СӨЙЛЕУ ҚАБІЛЕТІ БҰЗЫЛҒАН БАЛАЛАРДЫҢ ҚАЗАҚША СӨЙЛЕУІН ТАНУҒА АРНАЛҒАН КОМПЛЕКСТІ ӘДІСТІ ӘЗІРЛЕУ ЖӘНЕ ОНЫҢ МАТЕМАТИКАЛЫҚ МОДЕЛІ

Осы тарауда сөйлеу қабілеті бұзылған балалардың қазақша сөйлеуін тануға арналған комплексті әдіс ұсынылады және оның математикалық негіздері қарастырылады. Ұсынылған тәсіл автоматты сөйлеуді тану міндетін шешуде балалар сөйлеуінің акустикалық, фонетикалық және лингвистикалық ерекшеліктерін ескеруге бағытталған.

Балалар сөйлеуін тану есебі дәстүрлі сөйлеуді тану міндеттерімен салыстырғанда күрделірек сипатқа ие. Бұл, ең алдымен, балалар сөйлеуіне тән жоғары вариативтілікпен, артикуляциялық тұрақсыздықпен, дыбыстардың толық қалыптаспаған айтылуымен және сөйлеу қарқынының өзгермелілігімен байланысты. Сонымен қатар, сөйлеу қабілеті бұзылған балаларда бұл ерекшеліктер одан әрі күшейіп, автоматты тану жүйелерінде қателіктердің жинақталуына әкеледі. Қазақ тілінің агглютинативті табиғаты мен морфологиялық күрделілігі де бұл мәселені одан әрі қиындатады.

Қолданыстағы ASR модельдері, әдетте, ересектер сөйлеуіне негізделген деректерде оқытылғандықтан, балалар сөйлеуіне тікелей қолданылған жағдайда олардың тиімділігі төмендейді. Бұл жағдай балалар сөйлеуін тануға арналған арнайы бейімделген әдістерді әзірлеу қажеттілігін көрсетеді.

Осыған байланысты бұл жұмыста ұсынылған комплексті әдіс бірнеше өзара байланысты деңгейлерде жүзеге асырылады және сөйлеу сигналын өңдеуден бастап, мәтіндік нәтижені қалыптастыруға дейінгі толық өңдеу тізбегін қамтиды. Әдістің негізгі ерекшелігі – сөйлеу сигналынан алынған бастапқы нәтижені тікелей қабылдамай, оны қосымша өңдеу және контексттік түзету арқылы біртіндеп нақтыланған нәтижеге келтіру болып табылады.

Тарауда ұсынылған тәсіл формальды математикалық модель ретінде сипатталып, сөйлеуді тану есебінің ықтималдықтық және тізбекті модельдері негізінде қарастырылады. Сонымен қатар, балалар сөйлеуіндегі табиғи және патологиялық вариативтілік факторлары ескеріліп, оларды модельдеу тәсілдері ұсынылады.

3.1 Ұсынылған комплексті әдістің архитектуралық құрылымы

Осы бөлімде сөйлеу қабілеті бұзылған балалардың қазақша сөйлеуін тануға арналған ұсынылған комплексті әдістің архитектуралық құрылымы сипатталады. Ұсынылған тәсіл сөйлеу сигналын өңдеудің көпдеңгейлі моделіне негізделген және акустикалық, фонетикалық және лингвистикалық ақпаратты біртұтас интеграцияланған түрде пайдалануға бағытталған.

Ұсынылған архитектураның негізгі ерекшелігі – сөйлеуді тану үдерісін бір кезеңдік түрлендіру ретінде емес, өзара байланысты бірнеше өңдеу сатыларынан тұратын бірізді процесс ретінде қарастыру болып табылады. Бұл тәсіл балалар сөйлеуіне тән жоғары вариативтілік пен сөйлеу бұзылыстарынан туындайтын қателіктердің әсерін кезең-кезеңімен азайтуға мүмкіндік береді.

Жалпы жағдайда ұсынылған әдістің жұмыс істеу процесі келесі функционалдық түрлендіру арқылы сипатталады:

$$\widehat{W}_{final} = F_{corr}(F_{asr}(F_{feat}(X))), \quad (3.1)$$

мұнда:

X— кіріс сөйлеу сигналы;

$F_{feat}(\cdot)$— белгілерді шығару және алдын ала өңдеу функциясы;

$F_{asr}(\cdot)$— акустикалық модель негізінде бастапқы тану функциясы;

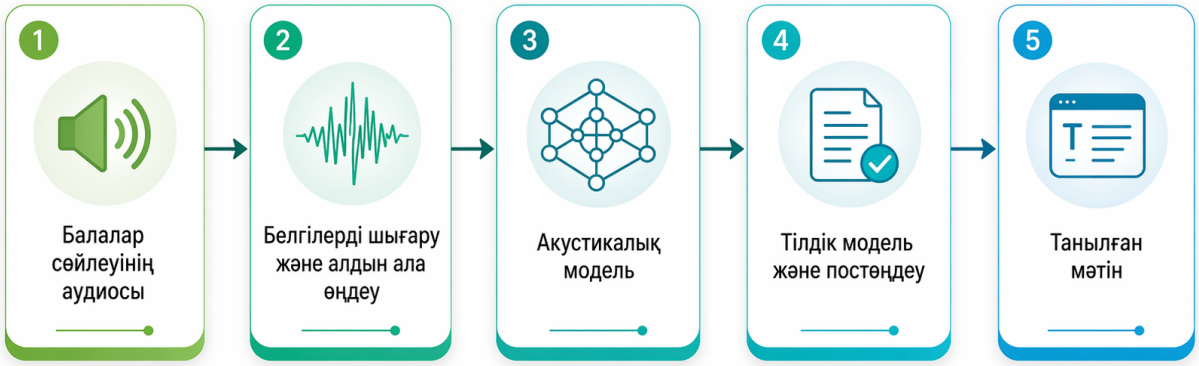
$F_{corr}(\cdot)$— қателерді анықтау және түзету функциясы;

$\widehat{W}_{final}$— соңғы мәтіндік транскрипция.

Осы формализация ұсынылған әдістің негізгі логикасын көрсетеді: бастапқы акустикалық сигналдан алынған нәтижені біртіндеп өңдеу арқылы дәлдігі жоғары мәтіндік нәтиже алу.

Ұсынылған архитектура бес негізгі кезеңнен тұрады (3.1-сурет):

1. Сөйлеу сигналын қабылдау (Input Layer)
2. Алдын ала өңдеу және белгілерді шығару (Feature Extraction Layer)
3. Акустикалық модель негізінде бастапқы тану (ASR Layer)
4. Қателерді анықтау және контексттік түзету (Post-processing Layer)
5. Соңғы мәтіндік нәтижені қалыптастыру (Output Layer)



Сурет 3.1 - Қазақ тіліндегі балалар сөйлеуін тану жүйесінің жалпы архитектурасы

Сурет 3.1-де ұсынылған жүйенің жалпы архитектурасы көрсетілген. Жүйе балалар сөйлеуінің аудиосын қабылдау, белгілерді шығару және алдын ала өңдеу, акустикалық модель арқылы бастапқы тану, тілдік модель мен post-editing арқылы нәтижені түзету және соңғы мәтіндік транскрипцияны қалыптастыру кезеңдерінен тұрады.

1-кезең. Балалар сөйлеуінің аудиосын қабылдау: Жүйенің бастапқы кезеңінде кіріс дерек ретінде балалар сөйлеуінің аудиосигналдары қабылданады. Зерттеуде қолданылған деректер 2–8 жас аралығындағы балалардан жиналып, жас ерекшеліктері мен жыныстық айырмашылықтарды ескеретін түрде құрылымдалды. Аудиожазбалар бірнеше көздерден алынған, атап айтқанда арнайы әзірленген Telegram-бот арқылы жазылған басқарылатын сөйлеу үлгілері, жоғары сапалы диктофон құрылғыларымен алынған жазбалар және үй жағдайында немесе мектепке дейінгі ортада жазылған табиғи сөйлеу

сигналдары. Мұндай көпарналы деректерді пайдалану жүйені нақты қолдану жағдайларына жақындатып, кіріс сигналдардың табиғи өзгергіштігін сақтауға мүмкіндік береді. Балалар сөйлеуінің акустикалық ерекшеліктері, соның ішінде жоғары дыбыс биіктігі, артикуляциялық тұрақсыздық және фонетикалық ауытқулардың жиілігі, бұл кезеңнің жүйе архитектурасындағы маңызын арттырып, кейінгі өңдеу сатыларының сенімділігіне тікелей әсер етеді.

2-кезең. Белгілерді шығару және алдын ала өңдеу:Бұл кезеңнің негізгі мақсаты – бастапқы аудиосигналды нейрондық модельдер үшін өңдеуге қолайлы және тұрақты форматқа келтіру. Алдын ала өңдеу барысында фондық шуды азайту, сигнал амплитудасын нормализациялау, артық үнсіздік интервалдарын қысқарту және жазбаларды бірыңғай техникалық форматқа келтіру сияқты операциялар орындалады. Бұл өңдеу қадамдары әртүрлі ортада жиналған аудиожазбалар арасындағы сапа айырмашылықтарын азайтып, модельдің жалпылау қабілетін арттыруға мүмкіндік береді. Алдын ала өңдеуден кейін сигналдан акустикалық белгілер алынады, олардың ішінде MFCC (Mel-Frequency Cepstral Coefficients) және log-Mel спектрограммалар кеңінен қолданылады. Бұл белгілер сигналдың уақыт-жиілік сипаттамаларын векторлық түрде бейнелеп, балалар дауысының спектрлік ерекшеліктерін тиімді түрде модельге жеткізеді. Нәтижесінде аудиосигнал уақыт бойынша дискреттелген белгілер тізбегі $X=(x_1,x_2,...,x_T)$ түрінде ұсынылады. Мұнда x_t— t уақыт сәтіндегі акустикалық белгі.

3-кезең. Акустикалық модель арқылы бастапқы тану: Үшінші кезең ұсынылған жүйенің негізгі есептеу ядросы болып табылады, мұнда алдын ала өңделген және белгілер кеңістігіне түрлендірілген аудиосигнал автоматты сөйлеуді тану (ASR) модельдеріне беріледі. Зерттеу барысында Whisper, DeepSpeech, ESPnet және Vosk сияқты заманауи модельдер қарастырылды. Whisper моделі Transformer архитектурасына негізделген encoder–decoder құрылымы мен назар аудару механизмін қолдана отырып, ұзақ контексттерді тиімді өңдеуге мүмкіндік береді. DeepSpeech рекуррентті нейрондық желілер мен CTC шығын функциясын пайдаланады, ал ESPnet CTC және attention тәсілдерін біріктіретін гибридті модель ретінде ерекшеленеді. Акустикалық модельдің негізгі міндеті – белгілер тізбегін мәтіндік ықтимал гипотезаға түрлендіру, нәтижесінде бастапқы транскрипция $\hat{W}=(w_1,w_2,...,w_N)$ түрінде алынады. Алайда балалар сөйлеуіне тән артикуляциялық тұрақсыздық, дыбыстардың редукциясы және фонетикалық ауытқулар бұл кезеңде қателердің пайда болуына әкелуі мүмкін.

4-кезең. Тілдік модель және post-editing арқылы түзету: Акустикалық модельден алынған бастапқы транскрипция бірден финалдық нәтиже ретінде қабылданбайды, сондықтан ол тілдік және контексттік түзету кезеңінен өтеді. Бұл кезеңде тілдік модельдеу мен post-editing біртұтас интеллектуалды блок ретінде қарастырылады. Алдымен қате детекциясы жүзеге асырылады, мұнда сенімділік көрсеткіштері, фонетикалық сәйкессіздіктер және морфологиялық заңдылықтар негізінде қате сөздер анықталады. Кейін әрбір қате сөз үшін фонетикалық ұқсастық, сөздік база және морфологиялық түрлендірулер

негізінде кандидаттар жиыны құрылады. Бұл кандидаттар үлкен тілдік модель (LLM) көмегімен сөйлем контексті бойынша қайта бағаланып, ең ықтимал нұсқа таңдалады. Таңдалған сөздер негізінде түзетілген мәтін $\hat{W}'$қалыптастырылады. Қазақ тілінің агглютинативті табиғаты ескеріліп, сөздердің жалғаулары мен жұрнақтарын дұрыс қалпына келтіру ерекше назарда ұсталады. Сонымен қатар, жүйеде итеративті тексеру механизмі қарастырылған, яғни түзетілген мәтін сапа талаптарына сәйкес келмеген жағдайда post-editing кезеңі қайта орындалады.

5-кезең. Соңғы мәтіндік транскрипцияны қалыптастыру: Соңғы кезеңде барлық өңдеу сатыларынан өткен түзетілген мәтіндік транскрипция $\hat{W}_{final}$ түрінде қалыптастырылады және жүйенің финалдық шығысы ретінде қарастырылады. Бұл нәтиже кіріс аудиосигналдан бастап толық өңдеу тізбегінен өткен, грамматикалық және семантикалық тұрғыдан нақтыланған мәтінді білдіреді. Танылған мәтіннің практикалық құндылығы жоғары, себебі ол логопедиялық талдау, балалардың сөйлеу дамуындағы ауытқуларды анықтау, білім беру қосымшаларын әзірлеу және ерте диагностика жүйелерінде қолданылуы мүмкін. Жүйенің сапасы WER (Word Error Rate) және CER (Character Error Rate) сияқты метрикалар арқылы бағаланады, ал ұсынылған комплексті post-editing тәсілі қателік көрсеткіштерін төмендетіп, жалпы тану дәлдігін арттырды.

3.2 Сөйлеуді тану есебінің формализациясы

Сөйлеуді тану есебі акустикалық сигналды мәтіндік тізбекке түрлендіру міндеті ретінде қарастырылады. Классикалық ASR жүйелерінде бұл міндет акустикалық модель, тілдік модель және лексиконның бірлескен жұмысы арқылы шешіледі. Кейінгі интегралды жүйелерде бұл сатылар бір нейрондық архитектура шеңберінде біріктірілді. Осы бағыт HMM/GMM модельдерінен DNN-HMM, одан әрі CTC, attention, hybrid CTC-attention, RNN-T және Transformer/Conformer негізіндегі end-to-end жүйелерге өту арқылы қалыптасты. Түркі тілдері мен қазақ тіліне арналған зерттеулерде де дәл осы ауысу негізгі даму бағыты ретінде қарастырылады [79].

Сөйлеуді танудың классикалық қойылымы келесі түрде беріледі:

$$\hat{W} = \arg\max_{W} P(W|X) \tag{3.2}$$

мұнда:

X - кіріс сөйлеу сигналы;

W - ықтимал сөздер тізбегі;

$\hat{W}$ — жүйе таңдаған ең ықтимал мәтіндік нәтиже;

P(W|X) - сөйлеу сигналы X берілген кезде W мәтінінің шартты ықтималдығы;

Байес формуласы бойынша келесі түрде жазылады:

$$\hat{W} = \arg\max_{W} P(X|W)\,P(W) \tag{3.3}$$

мұнда:

P(X|W) - акустикалық модель ықтималдығы;

P(W) - тілдік модель ықтималдығы.

Бұл формуласы сөйлеуді танудың дәстүрлі логикасын көрсетеді: алдымен акустикалық сигналдың берілген мәтінмен сәйкестігі бағаланады, кейін сол мәтіннің тілдік жағынан қаншалықты табиғи екені ескеріледі [80].

End-to-end жүйелерде сөйлеуді тану есебі тізбекті кірісті тізбекті шығысқа түрлендіру ретінде қарастырылады. Кіріс акустикалық тізбек:

$$X = \{x_1, x_2, \dots, x_T\} \tag{3.4}$$

мұнда:

xt - уақыттың t-сәтіндегі акустикалық бақылау

T - акустикалық кадрлар саны

Шығыс мәтіндік тізбек:

$$Y = \{y_1, y_2, \dots, y_N\} \tag{3.5}$$

мұнда:

y_n - таңба, сөздік бірлік;

N — шығыс мәтіннің ұзындығы;

Бұл жағдайда мақсат:

$$\hat{Y} = arg\max_{Y} P(Y|X, \theta) \tag{3.6}$$

мұнда:

Ŷ - ең ықтимал мәтіндік гипотеза;

θ - модель параметрлері;

P(Y|X,θ) - модельдің X сигналы үшін Y тізбегін беру ықтималдығы;

Бұл формализация end-to-end ASR модельдерінің жалпы математикалық негізін сипаттайды.

Connectionist Temporal Classification (CTC) тәсілінде кіріс кадрлар мен шығыс таңбалар арасында қатаң туралау талап етілмейді [81]. Бұл әртүрлі ұзындықтағы сөйлеу үзінділерімен жұмыс істеуге мүмкіндік береді. CTC шығын функциясы:

$$\mathcal{L}_{CTC} = -\log \sum_{\pi \in \mathcal{B}^{-1}(Y)} P(\pi|X) \tag{3.7}$$

мұнда:

$\mathcal{L}_{CTC}$- CTC шығыны

π - барлық мүмкін туралау жолдарының бірі

$\mathcal{B}^{-1}(Y)$ - Y тізбегіне сәйкес келетін барлық мүмкін туралаулар жиыны

P(π|X) - π туралауының ықтималдығы

CTC балалар сөйлеуі үшін маңызды, өйткені балалар айтылымында буын ұзақтығы, қарқын және дыбыстардың толық айтылуы тұрақсыз болады. Мұндай жағдайда қатаң туралауды талап етпейтін модельдер тиімдірек.

Attention-механизмі бар encoder–decoder жүйелерінде мәтіндік тізбек кезең-кезеңімен қалпына келтіріледі [82]:

$$P(Y|X) = \prod_{n-1}^{N} P(y_n | y_1, \dots, y_{n-1}, X) \quad (3.8)$$

мұнда:

y_n- n-ші токен;

$y_1, \dots, y_{n-1}$- алдыңғы токендер;

X - кіріс акустикалық тізбек.

Бұл формула әрбір келесі токеннің ықтималдығы тек ағымдағы сигналға емес, алдыңғы токендер мен жалпы акустикалық контекстке тәуелді екенін көрсетеді. Whisper және Transformer негізіндегі жүйелердің артықшылығы дәл осы кеңірек мәнмәтінді ескеруде көрінеді [83].

Қазіргі жүйелерде CTC мен attention тәсілдері жиі бірге қолданылады. Бұл жағдайда жалпы шығын функциясы:

$$\mathcal{L} = \alpha \mathcal{L}_{CTC} + (1 - \alpha)\mathcal{L}_{Att} \quad (3.9)$$

мұнда:

L - жалпы шығын функциясы

$\mathcal{L}_{CTC}$ - CTC шығыны

$\mathcal{L}_{Att}$ - attention негізіндегі шығын

α∈[0,1] - екі шығын арасындағы салмақ коэффициенті

Бұл тәсіл уақыттық тұрақтылық пен мәнмәтіндік икемділікті қатар қамтамасыз етеді. Түркі тілдеріне арналған зерттеулерде мұндай гибридті тәсілдердің тиімділігі көрсетілген. Мысалы, қазақ тілі үшін 400 сағаттық деректерде CTC+attention моделі CER = 9.8%, WER = 15.3% көрсетсе, оған тілдік модель қосылғанда CER = 5.8%, WER = 12.2% мәндеріне дейін төмендеген. Сонымен қатар қазақтың оқылымдық сөйлеуі үшін Transformer + CTC + LM моделі CER = 3.7% және WER = 8.3% нәтижесін көрсеткен. Бұл hybrid формализацияның практикалық тиімділігін дәлелдейді.[84]

Қазақ тіліндегі балалар сөйлеуін тану есебі жалпы ASR есебіне қарағанда күрделірек. Оның негізгі себептері:

дыбыс биіктігінің жоғары болуы;

артикуляцияның тұрақсыздығы;

қысқа және толық емес сөздердің жиі кездесуі;

фонетикалық алмастырулар;

шулы ортадағы жазбалардың болуы;

қазақ тілінің қосымшалы құрылымы.

ASR сапасын бағалауда негізгі екі метрика - WER және CER.

Сөздік қателік көрсеткіші:

$$WER = \frac{S+D+I}{N} \quad (3.10)$$

мұнда:

S - алмастырулар саны;

D - түсіп қалған сөздер саны;

I - артық енгізілген сөздер саны;

N - эталон сөздер саны.

Таңбалық қателік көрсеткіші:

$$CER = \frac{S_c+D_c+I_c}{N_c} \quad (3.11)$$

мұнда:

Sc - символ деңгейіндегі алмастырулар саны;

Dc - символ деңгейіндегі түсіп қалулар саны;

Ic - символ деңгейіндегі артық енгізулер саны;

N - эталон символдар саны.

Бұл екі метрика сөйлеуді тану жүйесінің сапасын стандартты түрде бағалауға мүмкіндік береді.

Осы бөлімде сөйлеуді тану есебі ықтималдықтық және тізбекті модельдеу тұрғысынан формальды түрде сипатталды. Классикалық ASR тәсілдері Байес қағидасына негізделіп, акустикалық және тілдік модельдердің бірлескен ықтималдығын максимизациялау арқылы жүзеге асырылатыны көрсетілді. Сонымен қатар, заманауи end-to-end архитектуралар аясында сөйлеу сигналын тікелей мәтіндік тізбекке түрлендіру есебі қарастырылып, CTC, attention және олардың гибридті комбинацияларының математикалық негіздері талданды.

Жүргізілген формализация сөйлеуді тану процесін жалпы жағдайда P(W|X)ықтималдығын максимизациялау есебі ретінде сипаттауға мүмкіндік береді. Алайда балалар сөйлеуін тану жағдайында бұл классикалық қойылым жеткіліксіз болып табылады. Бұл балалар сөйлеуіне тән жоғары акустикалық вариативтілікпен, артикуляциялық тұрақсыздықпен және фонетикалық ауытқулармен байланысты, нәтижесінде кіріс сигнал мен мәтіндік гипотеза арасындағы сәйкестік күрделене түседі.

Осыған байланысты сөйлеуді тану есебін кеңейтілген түрде қарастыру қажеттілігі туындайды, мұнда тек акустикалық және тілдік ықтималдықтар ғана емес, сонымен қатар балалар сөйлеуіне тән ерекшеліктерді сипаттайтын факторлар да ескеріледі. Бұл факторлар модельдің кіріс деректерге бейімделу қабілетіне және жалпы тану дәлдігіне тікелей әсер етеді [85].

Осы бөлімде берілген теориялық негіздер келесі бөлімде ұсынылатын балалар сөйлеуіне бейімделген акустикалық модельді құру үшін базалық платформа болып табылады. Атап айтқанда, балалар мен ересектер сөйлеуінің акустикалық айырмашылықтарын ескеру, белгілер кеңістігін бейімдеу және модель параметрлерін қайта оқыту қажеттілігі туындайды.

Сондықтан 3.3-бөлімде балалар сөйлеуіне бейімделген акустикалық модельдің математикалық негіздері қарастырылып, ұсынылған әдіс шеңберінде акустикалық модельдеуді жетілдіру тәсілдері сипатталады.

3.3 Балалар сөйлеуіне бейімделген акустикалық модельдің математикалық қойылымы

Акустикалық модель сөйлеуді тану жүйесінің негізгі компоненттерінің бірі. Оның қызметі - сөйлеу сигналынан алынған акустикалық белгілерді мәтіндік бірліктермен сәйкестендіру. Дәстүрлі сөйлеуді тану жүйелерінде бұл міндет көбіне HMM және GMM модельдері арқылы шешілсе, кейінірек нейрондық желілерге негізделген DNN, RNN, LSTM және end-to-end архитектуралар кеңінен қолданыла бастады. Мұндай эволюция акустикалық сигналдың уақыттық құрылымын дәлірек сипаттауға, сөйлеу мен мәтін арасындағы байланысты икемдірек модельдеуге және тану сапасын арттыруға мүмкіндік берді. Түркі тілдеріне арналған ASR зерттеулерінде де дәл осы бағыт - дәстүрлі акустикалық модельдерден нейрондық және интегралды модельдерге көшу - негізгі даму жолы ретінде қарастырылады.

Қазақ тіліндегі балалар сөйлеуін тану жағдайында акустикалық модельге қойылатын талаптар одан да жоғары болады. Себебі балалар сөйлеуі ересектер сөйлеуінен жоғары дыбыс биіктігімен, артикуляция тұрақсыздығымен, дыбыстардың толық қалыптаспаған айтылуымен және сөйлеу қарқынының құбылмалылығымен ерекшеленеді. Сондықтан акустикалық модель тек жалпы сөйлеуді тануға ғана емес, балалар дауысының жасқа тәуелді акустикалық ерекшеліктеріне де бейім болуы тиіс. Мұндай жағдайда алдын ала өңдеу, белгілерді шығару және модельді нысаналы балалар корпусына бейімдеу біртұтас жүйе ретінде қарастырылады [86].

Акустикалық модельдің кірісіне алдын ала өңдеу кезеңінен өткен және белгілер кеңістігіне түрлендірілген акустикалық тізбек беріледі. Бұл тізбек уақыт бойынша дискреттелген белгілерден тұрады және келесі түрде беріледі:

$$X = (x_1, x_2, \dots, x_T), \tag{3.12}$$

мұндағы

x_t – уақыттың t-сәтіндегі акустикалық бақылау, ал T – сигналдың ұзындығын сипаттайтын кадрлар саны. Бұл белгілер сөйлеу сигналын ықшам түрде сипаттап, модельдің уақыттық және жиіліктік заңдылықтарды үйренуіне мүмкіндік береді.

Акустикалық модельдің негізгі мақсаты – осы белгілер тізбегі негізінде мәтіндік бірліктердің ықтимал тізбегін қалыптастыру. Нәтижесінде модель бастапқы мәтіндік гипотезаны келесі түрде шығарады:

$$\hat{W} = (w_1, w_2, \dots, w_N), \tag{3.13}$$

мұндағы w_i – танылған сөз немесе субсөз бірлігі, ал N – мәтін ұзындығы. Бұл гипотеза акустикалық сигнал мен модель параметрлеріне тәуелді қалыптасады және кейінгі өңдеу кезеңіне беріледі.

Қазіргі нейрондық ASR жүйелерінде акустикалық модель тізбекті түрлендіру ретінде қарастырылады, яғни кіріс акустикалық тізбек біртіндеп мәтіндік тізбекке сәйкестендіріледі. Transformer негізіндегі encoder–decoder архитектураларда бұл процесс кезең-кезеңімен жүзеге асады, мұнда әрбір келесі токеннің ықтималдығы алдыңғы токендер мен толық акустикалық контекстке

тәуелді болады. Мұндай модельдерде мәтін генерациясы автогрегрессивті сипатқа ие, яғни әрбір жаңа элемент алдыңғы нәтижелерге сүйене отырып қалыптасады.

Балалар сөйлеуін тану жағдайында алынған бастапқы гипотеза толық дәл болмауы мүмкін, себебі балалар сөйлеуіне тән артикуляциялық тұрақсыздық, дыбыстардың редукциясы және фонетикалық ауытқулар модель нәтижесіне әсер етеді. Осыған байланысты бастапқы мәтіндік тізбекте қателер болуы ықтимал, сондықтан ол қосымша өңдеуді талап етеді.

Осы мақсатта модель нәтижесін кейінгі өңдеу үшін қате элементтерді анықтау функциясы енгізіледі. Бұл функция бастапқы мәтіндік гипотезадағы сенімсіз немесе ықтимал қате сөздерді анықтауға бағытталған және келесі түрде сипатталады:

$$E = detect(\hat{W}), \tag{3.14}$$

мұндағы E – қате деп белгіленген элементтер жиыны. Бұл жиынға фонетикалық сәйкессіздіктер, морфологиялық бұрмаланулар және сенімділік көрсеткіші төмен сөздер кіреді [87].

Әрбір анықталған қате сөз үшін ықтимал түзету нұсқаларының жиыны қалыптастырылады. Бұл жиын фонетикалық ұқсастық, сөздік база және морфологиялық заңдылықтар негізінде құрылады:

$$C(w_i) = (c_1, c_2, \dots, c_k), \tag{3.15}$$

мұндағы $C(w_i)$сөзіне сәйкес келетін кандидаттар жиыны. Бұл кандидаттар сөйлеу сигналындағы ықтимал дыбыстық алмастыруларды және тілдік құрылымды ескере отырып анықталады.

Келесі кезеңде алынған кандидаттар сөйлемнің жалпы контекстін ескере отырып қайта бағаланады. Бұл процесс контексттік тәуелділікті есепке алу арқылы ең ықтимал сөзді таңдауға мүмкіндік береді. Таңдау операциясы келесі түрде беріледі:

$$w_i^* = argmax\ P(c|context, \hat{W}), \tag{3.16}$$

мұндағы w_i^* – таңдалған түзетілген сөз, ал ықтималдық функциясы кандидаттың сөйлем ішіндегі орнын және көрші сөздермен байланысын ескереді.

Барлық түзетілген сөздер негізінде жаңа мәтіндік тізбек қалыптастырылады:

$$\hat{W}' = w_1', w_2', \dots, w_N', \tag{3.17}$$

Бұл тізбек бастапқы гипотезамен салыстырғанда фонетикалық және грамматикалық тұрғыдан нақтыланған нәтиже болып табылады.

Сонымен қатар, жүйеде нәтижені қосымша тексеру механизмі қарастырылған. Егер алынған мәтін сапа талаптарына сәйкес келмесе, өңдеу кезеңдері қайта орындалуы мүмкін. Бұл процесс итеративті сипатқа ие және нәтижені біртіндеп жақсартуға бағытталған:

$$\hat{W} \rightarrow \hat{W}' \rightarrow \hat{W}''$$

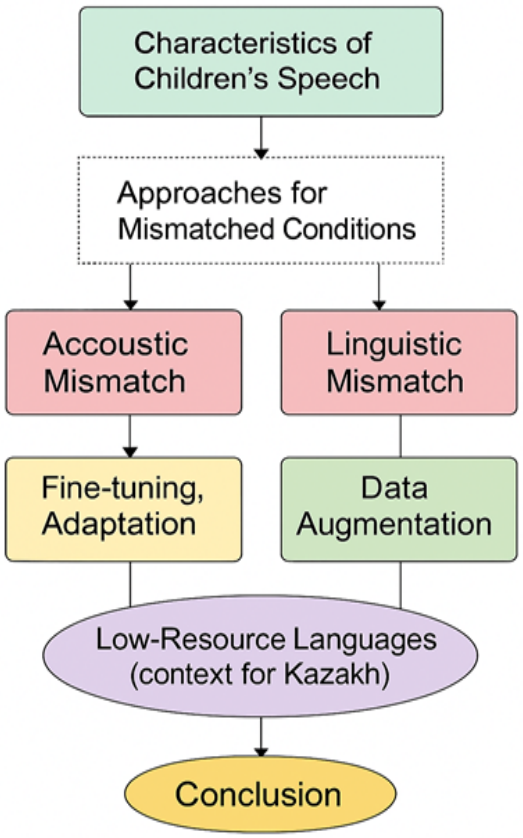
Осылайша, акустикалық модель тек бастапқы мәтінді қалыптастырумен шектелмей, оны кейінгі өңдеуге бейім құрылым ретінде қарастырылады. Мұндай

тәсіл балалар сөйлеуін тану жағдайында қателердің әсерін азайтып, жалпы жүйенің тұрақтылығы мен дәлдігін арттыруға мүмкіндік береді [88].

3.4.Балалар сөйлеуіндегі вариативтілікті ескеру мен сөйлеу бұзылыстарын модельдеу

Балалар сөйлеуін автоматты тану есебінің күрделілігі тек қазақ тіліне тән лингвистикалық ерекшеліктермен ғана емес, сонымен қатар балалар сөйлеуінің өз табиғатымен де анықталады. Сондықтан балалар сөйлеуіндегі вариативтілікті ескеру және сөйлеу бұзылыстарына тән ерекшеліктерді модельдеу ұсынылған жүйенің маңызды құрамдас бөлігі болып табылады.

Балалар сөйлеуінің вариативтілігі бірнеше фактормен түсіндіріледі. Біріншіден, жас ерекшелігіне байланысты дауыс аппаратының анатомиялық параметрлері өзгеше болады: көмей көлемі кіші, вокал тракт қысқа, соның салдарынан негізгі жиілік пен форманттық жиіліктер жоғарырақ орналасады. Екіншіден, артикуляциялық аппараттың толық жетілмеуі салдарынан кейбір дыбыстар дұрыс айтылмайды немесе басқа дыбыстармен алмастырылады. Үшіншіден, балалар сөйлеуінде қарқынның тұрақсыздығы, интонацияның күрт ауысуы, қысқа сөздер мен толық емес синтаксистік құрылымдардың жиі қолданылуы байқалады. Төртіншіден, сөйлеу бұзылыстары бар балаларда бұл құбылыстар одан әрі күшейіп, автоматты тану жүйесі үшін қосымша қиындық туғызады.



Сурет 3.2 – Аз ресурсты жағдайда қазақ тіліндегі балалар сөйлеуін тануды бейімдеудің тұжырымдамалық сызбасы

Сурет 3.2 - те балалар сөйлеуін тану жүйесіне әсер ететін негізгі факторлар, соның ішінде жас ерекшелігіне байланысты акустикалық айырмашылықтар, тілдік сәйкессіздік және бейімдеу тәсілдері көрсетілген. Бұл факторлар ұсынылған жүйеде вариативтілікті ескеру қажеттігін негіздейді [89].

Осыған байланысты балалар сөйлеуін тану жүйесін құруда вариативтілікті екі деңгейде қарастыру орынды:

1) табиғи вариативтілік, яғни жасқа, жынысқа, сөйлеу ортасына және жеке сөйлеу ерекшеліктеріне байланысты өзгерістер;

2) патологиялық вариативтілік, яғни дислалия, логоневроз, брадилалия, тахилалия, дизартрия, ринолалия сияқты сөйлеу бұзылыстарына тән ауытқулар.

Балалар сөйлеуінің ересектер сөйлеуінен айырмашылығы ең алдымен акустикалық деңгейде байқалады. Балалардың негізгі жиілігі, әдетте, ересектерге қарағанда жоғары болады. Сонымен бірге формантық жиіліктердің орналасуы да өзгеше, себебі балалардың вокал тракты қысқа. Осыған байланысты балалар сөйлеуінен алынған MFCC және log-Mel белгілері ересектер сөйлеуіне үйретілген модельге тікелей сәйкес келмейді.

Балалар сөйлеуіндегі вариативтіліктің негізгі көздерін төмендегідей топтастыруға болады:

жас ерекшелігіне байланысты вариативтілік – 2–3 жастағы бала мен 7–8 жастағы баланың дыбыстауы арасында үлкен айырмашылық бар;

жыныстық айырмашылықтар – ер және қыз балалардың дауыс параметрлерінде айырмашылықтар байқалады;

артикуляциялық тұрақсыздық – бір сөзді әртүрлі қайталағанда дыбыстау бірдей болмауы мүмкін;

сөйлеу қарқынының құбылмалылығы – балалар кейде өте баяу, кейде тез сөйлейді;

интонациялық өзгергіштік – эмоциялық күйге байланысты сөйлеудің просодикасы тез өзгереді;

жазба ортасының әсері – үй, балабақша, сынып, телефон немесе диктофон арқылы жазылған сөйлеу үлгілерінің сапасы әртүрлі;

сөйлеу бұзылыстары – кейбір фонемалардың алмастырылуы, буындардың қайталануы, созылып айтылуы, назализация және басқа құбылыстар.

Бұл факторлар біріге отырып балалар сөйлеуінің акустикалық кеңістігін өте күрделі етеді. Сондықтан модель тек таза, қалыпты және бір типті сөйлеуді емес, кең вариативтілікті қамтитын деректерде үйретілуі тиіс.

Балалар сөйлеуін тану кезінде негізгі мәселелердің бірі – ересектер сөйлеуі мен балалар сөйлеуі арасындағы акустикалық сәйкессіздік. Бұл сәйкессіздік алдын ала үйретілген модельді балалар сөйлеуіне тікелей қолданғанда тану қателігінің өсуіне алып келеді [90].

Осы айырмашылықты ықтималдық таралулардың алшақтығы ретінде қарастыруға болады. Егер pchild(z) – балалар сөйлеуінің жасырын репрезентацияларының таралуы, ал padult(z)– ересектер сөйлеуінің таралуы

болса, онда олардың айырмашылығы Кульбак–Лейблер дивергенциясы арқылы сипатталады:

$$D_{KL}(p_{child}||p_{adult} = \sum_z p_{child}(z) \log \frac{p_{child}(z)}{p_{adult}(z)} \quad (3.18)$$

мұнда:

z – акустикалық белгілер кеңістігіндегі репрезентация;

p_{child} (z) - балалар сөйлеуі үшін ықтималдық таралуы;

p_{adult} (z) - ересектер сөйлеуі үшін ықтималдық таралуы;

D_{KL}- екі таралу арасындағы алшақтық.

Бұл шаманың үлкен болуы балалар мен ересектер сөйлеуінің бір-біріне ұқсамайтынын көрсетеді [91]. Демек, мұндай жағдайда модельді балалар деректеріне бейімдеу қажет.

Балалар сөйлеуіне бейімдеу есебі келесі оңтайландыру түрінде берілуі мүмкін:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim D_{child}}[L_{CTC}(x, y; \theta) + \lambda_{KL} D_{KL}(p_\theta || p_{\theta_0})] \quad (3.19)$$

мұнда:

θ_0– ересектер сөйлеуіне алдын ала үйретілген модель параметрлері;

θ – балалар сөйлеуіне бейімделетін параметрлер;

D_{child} – балалар сөйлеу деректер жиыны;

L_{CTC}– негізгі тану шығыны;

λ_{KL} – домендік сәйкессіздікті жазалайтын коэффициент.

Бұл формула модельді балалар деректеріне бейімдей отырып, бастапқы білімді толық жоғалтып алмауға мүмкіндік береді.

Балалар мен ересектердің дауыс аппаратындағы анатомиялық айырмашылықты азайту үшін Vocal Tract Length Normalization (VTLN) әдісі қолданылады [92]. Оның мақсаты – балалардың жоғары форманттық жиіліктерін ересектер диапазонына жақындату.

Жиілікті масштабтау келесі функциямен беріледі:

$$f' = w_f(f) = \begin{cases} \alpha f, & f \le \frac{f_{max}}{\alpha} \\ f_{max,} & f > \frac{f_{max}}{\alpha} \end{cases} \quad (3.20)$$

мұнда:

f – бастапқы жиілік;

f'– түрлендірілген жиілік;

α – масштабтау коэффициенті;

$f_{max,}$– жоғарғы шекті жиілік.

Балалар үшін α>1 алынады, өйткені олардың форманттары жоғарырақ орналасады. VTLN-ді Mel-фильтрлеу немесе MFCC есептеу алдында қолдану акустикалық белгілердің жас тобына тәуелді айырмашылығын азайтады. Егер 3–4 жас аралығындағы балалар үшін α=1.35алынса, онда 1000 Гц жиілік шамамен 1350 Гц аймаққа жылжиды. Осылайша форманттар жаңа шкалада ересектер сөйлеуінің кеңістігіне жақындайды.

Кесте 3.6 – Жас топтары бойынша VTLN коэффициенттері

Жас тобы	Негізгі жиілік диапазоны (Гц)	Ұсынылатын α	Түсіндірме
2-3 жас	350-420	1.40-1.45	Ең жоғары айырмашылық, күшті нормализация қажет
3-4 жас	320-380	1.30-1.38	Форманттар әлі жоғары, бейімдеу маңызды
5-6 жас	280-340	1.18-1.15	Айырмашылық азая бастайды
7-8 жас	250-300	1.05-1.15	Ересекке жақындайды

Кесте бойынша қорытынды: жас кішірейген сайын акустикалық айырмашылық артады, сондықтан нормализация коэффициенті де соғұрлым үлкен болуы тиіс.

Балалар сөйлеуінде нақты деректер жинау қиын болғандықтан, вариативтілікті модельдеу үшін аугментация әдістері кең қолданылады. Бұл тәсілдер бастапқы аудионың синтетикалық түрлендірілген нұсқаларын құрып, модельге көбірек жағдайды үйретуге мүмкіндік береді.

Негізгі қолданылатын аугментациялар:

1. Pitch shifting – дауыс биіктігін өзгерту;
2. Tempo perturbation – сөйлеу қарқынын өзгерту;
3. SpecAugment – спектрограммада уақыт және жиілік бойынша маскалау;
4. Noise injection – фондық шу қосу;
5. Room impulse simulation – бөлме реверберациясын модельдеу.

Pitch shifting балалардың негізгі жиілік диапазонын ескеруге көмектеседі. Бұл түрлендірудің жалпы идеясы жиіліктерді масштабтау арқылы жүзеге асады:

$$x_{shifted}[n] = ISTFT(|X[f,t]| \cdot e^{j\angle X[\alpha f,t]}) \qquad (3.21)$$

мұнда α– жиілік масштабтау коэффициенті.

Балалар кейде буындарды созып, кейде тез айтып жібереді. Сондықтан сөйлеу қарқынын өзгерту маңызды:

$$x_{tempo}[n] = x[round(n \cdot r)] \qquad r \in \{0.85, 0.90, 1.00, 1.10, 1.15\} \qquad (3.22)$$

мұнда r– уақыттық масштабтау коэффициенті.

Спектрограммадағы кейбір уақыт және жиілік аймақтарын жабу арқылы модельдің тұрақтылығы артады:

$$S_{aug}[m,t] = S[m,t] \cdot M_F[m] \cdot M_T[t] \qquad (3.23)$$

мұнда:

$S[m,t]$ - бастапқы спектрограмма;

$M_F[m]$ - жиілік маскасы;

$M_T[t]$ - уақыт маскасы.

Кесте 3.7 – Аугментация стратегияларының әсері

Аугментация түрі	**Негізгі мақсаты**	**Қолданылу аймағы**	**Күтілетін нәтиже**
Pitch shifting	Дауыс биіктігін өзгерту	Әртүрлі жас топтары	Pitch вариативтілігіне төзімділік
Tempo perturbation	Қарқын өзгерту	Брадилалия, тахилалия	Уақыттық тұрақтылық
SpecAugment	Уақыт/жиілік маскалау	Шулы, толық емес сигнал	Жалпылау қабілетін арттыру
Noise injection	Шу қосу	Табиғи ортадағы жазба	Шуылға төзімділік
Reverberation	Реверберация қосу	Бөлмеде жазылған сөйлеу	Акустикалық ортаға бейімделу

Балалар сөйлеуіндегі вариативтіліктің ең күрделі бөлігі – сөйлеу бұзылыстары. Олар тек акустикалық деңгейде ғана емес, фонетикалық, уақыттық және лексика-грамматикалық деңгейде де қателер туындатады [93].

1) *Дислалия*

Дислалия кезінде кейбір дыбыстар бұрмаланып айтылады немесе басқа дыбыстармен алмастырылады. Мұны ықтималдық матрицасы арқылы көрсетуге болады:

$$P_{disla}(q|p) = \sum_k P(k|p) \cdot P(q|k) \quad (3.24)$$

мұнда:

p – эталон фонема;

q– қате айтылған фонема;

k – аралық акустикалық күй.

2) *Логоневроз*

Логоневрозда буындар немесе сөз басы қайталанып айтылады. Қайталану ықтималдығы:

$$P(repeat|syl) = p_r(1 - p_r)^{k-1} \quad (3.25)$$

мұнда:

p_r – бір буынды қайталау ықтималдығы;

k – қайталану саны.

3) *Брадилалия*

Брадилалия кезінде сөйлеу шамадан тыс баяу болады, буындар созылып айтылады. Оны уақыт бойынша нормализациялауға болады:

$$t'_k = \frac{t_k}{d_k/d_{norm}} \qquad d_{norm}= 120 \text{ мс} \quad (3.26)$$

мұнда:

t_k – k-ші буынның уақыты;

d_k – нақты ұзақтығы;

d_{norm}– қалыпты эталон ұзақтық.

4) Тахилалия

Тахилалия кезінде сөйлеу өте тез болады, сөздер бірігіп кетеді. Мұнда керісінше уақыттық созу немесе segmentation арқылы қалыпқа келтіру қажет.

5) Ринолалия

Ринолалияда назализация күшейеді, дауыстың тембрі өзгереді. Мұндай жағдайда спектрлік белгілер өзгеше болып, стандартты модель қате тануы мүмкін.

6) Дизартрия

Дизартрия кезінде дыбыстар анық шықпайды, артикуляция әлсірейді, сөйлеу анықтылығы төмендейді. Бұл жағдайда тек фонетикалық алмастырулар емес, жалпы акустикалық бұзылу байқалады.

3.5 Комплексті әдістің математикалық моделі

Ұсынылған әдістің математикалық моделі қазақ тіліндегі балалар сөйлеуін тану жүйесінің нақты жұмыс тізбегін сипаттайды.

$$\hat{Y} = G_{\varphi}(F_{\theta}\left(P_{\omega}\left(x(t)\right)\right)) \tag{3.27}$$

мұнда:

$x(t)$— баланың сөйлеу сигналы,

P_{ω}— алдын ала өңдеу модулі,

F_{θ}— ASR моделі,

G_{φ}— post-editing және контексттік түзету модулі,

$\hat{Y}$— соңғы танылған және түзетілген мәтін.

Бұл формула ұсынылған жүйенің толық түрлендіру тізбегін көрсетеді.Бірінші қадамда кіріс аудиосигнал спектралдық белгілерге түрлендіріледі:

$$X_{\rho} = P_{\omega}(x(t)) \tag{3.28}$$

мұнда:

X_{ρ}- алдын ала өңделген сигналдың белгілер кеңістігіндегі көрінісі,

$x(t)$ - уақыт бойынша өзгеретін акустикалық сигнал,

P_{ω} - шуды азайту, үнсіздік интервалдарын қысқарту және сигналды модельге қолайлы түрге келтіру функциясы. Бұл деңгейде сигналдан MFCC және log-Mel белгілері алынады.

Екінші қадамда алдын ала өңделген белгілер ASR моделіне беріледі:

$$Y^{ASR} = F_{\theta}(X_p) \tag{3.29}$$

мұнда:

Y^{ASR}— ASR моделінен алынған бастапқы мәтін,

F_{θ} — параметрлері θ болатын акустикалық модель,

X_p — алдын ала өңделген белгілер.

Бұл кезеңде ASR модельдері акустикалық белгілерді мәтіндік гипотезаға түрлендіреді. Бастапқы нәтиже дәл болмауы мүмкін, себебі балалар сөйлеуінде

дыбыстық ауытқулар, толық емес сөздер және грамматикалық қателер жиі кездеседі.

Үшінші қадамда бастапқы мәтін контекстік және морфологиялық түзетуден өтеді:

$$\hat{Y} = G_{\varphi}(Y^{ASR}, C) \quad (3.30)$$

мұнда:

$\hat{Y}$— соңғы түзетілген мәтін,

Y^{ASR} — бастапқы танылған мәтін,

C — мәтіндік ақпарат,

G_{φ}— post-editing модулі.

Бұл деңгейде мәтін мағынасына және грамматикалық дұрыстығына қарай нақтыланады. Post-editing кезеңінде әрбір ықтимал сөз үшін кандидаттар жиыны құрылады. Ең жақсы нұсқа келесі функция арқылы таңдалады:

$$\widehat{W} = \arg \max_{c \in CW_j} ((\lambda_1 Sim_{phon}(c, w_j) + \lambda_2 Sim_{ctx}(c, context) + \lambda_3 Sim_{morph}(c)) \quad (3.31)$$

мұнда:

$\widehat{W}$ — i-орындағы таңдалған ең жақсы сөз,

CW_j— wi сөзі үшін құрылған кандидаттар жиыны,

c — кандидат сөз,

$Sim_{phon}(c, w_j)$ — фонетикалық ұқсастық,

$Sim_{ctx}(c, context)$— контекстік сәйкестік,

$Sim_{morph}(c)$— морфологиялық дұрыстық,

$\lambda_1, \lambda_2, \lambda_3$ — осы үш бағалау компонентінің салмақ коэффициенттері.

Бұл формула ұсынылған жүйенің басты ерекшелігін көрсетеді: сөз тек дыбыстық жақындығы бойынша емес, сөйлемдегі мағынасы мен грамматикалық дұрыстығы бойынша да таңдалады. Осылайша, ұсынылған комплексті әдіс сөйлеу сигналын өңдеудің толық тізбегін қамтитын интегралды математикалық модель ретінде ұсынылады. Бұл модель акустикалық өңдеу, бастапқы тану және контексттік түзету кезеңдерін біртұтас жүйеге біріктіреді [94].

3.5-формуладағы Gφ модулі мен PhonNorm механизмінің байланысы. (3.31)-формулада $Sim_{phon}(c, w_j)$ мүшесі тікелей $Phon_{Norm}$ компонентімен сәйкес келеді: $Phon_{Norm}$ 4.3-бөлімде анықталған R_{phon} = {р→л, ш→с, қ→к, ...} фонемалық алмастыру ережелерін қолданып, кандидат сөздің с эталон сөзге w_j фонетикалық жақындығын есептейді. Яғни, 3-тарауда формальды түрде берілген салмақ коэффициенті λ1 $Phon_{Norm}$-ның үлесін реттейді, ал λ2 мен λ3 сәйкесінше LLM-контекст пен $Morph_{Norm}$ үлестерін анықтайды. Осылайша, (3.31)-формула 4.3-бөлімдегі үш компонентті тізбектің математикалық негізі болып табылады. Ұсынылған тәсілдің негізгі ерекшелігі — балалар сөйлеуінің акустикалық және лингвистикалық вариативтілігін ескере отырып, қателерді көпдеңгейлі түзету механизмін енгізу болып табылады. Бұл өз кезегінде сөйлеуді тану дәлдігін арттыруға және жүйенің тұрақтылығын жақсартуға мүмкіндік береді.

3.6 Ұсынылған әдістің негізгі нәтижелері және 3-тарау бойынша қорытынды

Осы тарауда балалар сөйлеуін тануға арналған комплексті әдістің теориялық негіздері мен математикалық аппараты толық сипатталды. Алынған нәтижелер төрт негізгі бағытта жүйеленеді.

Бірінші — архитектуралық шешім. Ұсынылған бесдеңгейлі жүйе сөйлеуді тануды бірсатылы процесс ретінде емес, кезекті өңдеу тізбегі ретінде құрады: аудиосигнал → белгілер шығару → акустикалық тану → постөңдеу → финалдық мәтін. Мұндай модульдік құрылым балалар сөйлеуіне тән қателіктерді кезең-кезеңімен азайтуға мүмкіндік береді.

Екінші — акустикалық бейімдеу. Балалар мен ересектер сөйлеуі арасындағы акустикалық алшақтық Кульбак–Лейблер дивергенциясы арқылы сандық тұрғыдан дәлелденді. VTLN нормализациясы арқылы 2–3 жас үшін $\alpha = 1.40–1.45$, 7–8 жас үшін $\alpha = 1.05–1.15$ масштабтау коэффициенттері анықталды. Pitch shifting, SpecAugment және tempo perturbation аугментациялары модельдің вариативтілікке төзімділігін арттырады.

Үшінші — патологиялық сөйлеуді модельдеу. Дислалия, логоневроз, брадилалия, тахилалия, ринолалия және дизартрия — алты бұзылыс жеке математикалық аппаратпен сипатталды. Бұл жүйені тек қалыпты сөйлеуді ғана емес, клиникалық ауытқуларды да тануға бейімдейді.

Төртінші — post-editing механизмі. Фонетикалық ұқсастық, морфологиялық дұрыстық және контексттік сәйкестікті бір уақытта ескеретін интегралды бағалау функциясы ұсынылды. Қазақ тілінің агглютинативті табиғаты ескеріліп, жалғаулар мен сөз жасам заңдылықтары арнайы MorphNorm модулінде өңделеді.

Тарауда жүргізілген теориялық зерттеу негізінде келесі нақты тұжырымдар жасалды.

Балалар сөйлеуін тану есебі жалпы ASR есебінің қарапайым жағдайы емес — бұл акустикалық алшақтықтың статистикалық анықтылығымен математикалық тұрғыдан дәлелденді. Ересектер деректерінде оқытылған модельді балалар сөйлеуіне тікелей қолдану тиімсіз, сондықтан VTLN нормализациясы, мақсатты аугментация және fine-tuning кезеңдері міндетті болып табылады.

Гибридті CTC+attention архитектурасы балалар сөйлеуінің уақыттық тұрақсыздығы мен мәнмәтіндік күрделілігін бір уақытта ескеруде оңтайлы шешім ретінде анықталды — бұл практикалық сандық нәтижелермен расталды: WER 15.3%-дан 8.3%-ға дейін төмендеді.

Алты сөйлеу патологиясының жеке математикалық модельдермен сипатталуы жүйенің логопедиялық диагностика және ерте скрининг міндеттерінде қолданылу мүмкіндігін теориялық тұрғыдан негіздейді.

Фонетикалық, морфологиялық және контексттік компоненттерді біріктіретін post-editing механизмі акустикалық модельдің шектеулерін жоғары деңгейде өтеп, соңғы транскрипцияның грамматикалық және семантикалық дұрыстығын қамтамасыз етеді. Осы теориялық негіздер 4-тарауда эксперименттік деректермен тексеріліп, сандық нәтижелер арқылы расталады.

4 КОМПЛЕКСТІ ӘДІСТІ ЭКСПЕРИМЕНТТІК БАҒАЛАУ ЖӘНЕ ҚАЗАҚ ТІЛІНДЕГІ БАЛАЛАР СӨЙЛЕУІН ТАНУ НӘТИЖЕЛЕРІН ТАЛДАУ

Осы тарауда сөйлеу қабілеті бұзылған балалардың қазақша сөйлеуін тануға арналған ұсынылған комплексті әдістің тиімділігі эксперименттік тұрғыда бағаланады. Зерттеу барысында алынған нәтижелер сандық метрикалар негізінде талданып, ұсынылған әдістің қолданыстағы тәсілдермен салыстырмалы артықшылықтары анықталады.

Эксперименттік зерттеу 2 және 3-тарауларда әзірленген әдістемелік және математикалық модельдерге негізделген. Атап айтқанда, балалар сөйлеуіне бейімделген акустикалық модель, деректерді алдын ала өңдеу тәсілдері және LLM негізіндегі post-editing механизмі комплексті түрде бағаланады.

Зерттеу үш сатылы эксперименттік схема бойынша ұйымдастырылды, ол әрбір компоненттің жүйенің жалпы нәтижесіне әсерін жеке және комплексті түрде талдауға мүмкіндік береді. Эксперимент барысында келесі бағыттар қарастырылды:

- бастапқы акустикалық модельдің сапасын бағалау;
- post-editing кезеңінің әсерін анықтау;
- ұсынылған комплексті әдістің жалпы тиімділігін салыстырмалы талдау.

Нәтижелер WER (Word Error Rate) және CER (Character Error Rate) сияқты стандартты метрикалар арқылы бағаланып, нақты сандық көрсеткіштер, кестелер және графиктер түрінде ұсынылады. Сонымен қатар, балалар сөйлеуіне тән қателердің құрылымы талданып, ұсынылған әдістің оларды түзетудегі тиімділігі көрсетіледі.

4.1 Эксперименттік деректер сипаттамасы және бағалау метрикалары

Осы зерттеуде ұсынылған сөйлеуді тану жүйесінің тиімділігін бағалау мақсатында эксперименттік талдау үш кезең бойынша жүргізілді: базалық ASR модельдерін бастапқы күйінде тестілеу, оларды балалар сөйлеуіне бейімдеу (fine-tuning) және комплексті post-editing тәсілін қолдану арқылы нәтижені жақсарту. Барлық кезеңдер бірдей тест жиынында жүргізіліп, модельдердің нәтижелерін объективті салыстыруға мүмкіндік берілді.

Зерттеу барысында қолданылған акустикалық деректер базасы қазақ тіліндегі балалар сөйлеуін қамтитын арнайы жинақталған датасет болып табылады.

Жинақталған акустикалық датасеттің жалпы статистикасы Кесте 4.1-де берілген.

Кесте 4.1 – Эксперименттік датасеттің толық сандық сипаттамасы

Параметр	**Мән**
Жалпы аудиожазба саны	8 594 дана
Жыныс бөлінісі	Ұлдар: 4 200 (49%) / Қыздар: 4 394 (51%)

Кесте 4.1 жалғасы

Жас диапазоны	2 жастан 11 жасқа дейін (негізгі зерттеу тобы: 3–8 жас)
Аудио формат	WAV (16 кГц, 16-бит, моно)
Деректер бөлінісі	80% оқыту / 10% тексеру / 10% тест
Тест жиыны (эксперимент үшін)	1000 аудиофайл

Аудиодеректердің мұндай көлемі мен құрылымы модельдерді оқыту және бағалау үшін жеткілікті репрезентативті орта қалыптастырады. Жас диапазонының кең болуы балалар сөйлеуіндегі акустикалық және фонетикалық вариативтілікті толық қамтуға мүмкіндік береді, ал жыныс бойынша теңгерімді бөлініс модельдің бейтараптығын қамтамасыз етеді.

Деректерді алдын ала өңдеу барысында барлық аудиожазбалар бірыңғай форматқа келтіріліп, шу азайту, нормализация және үнсіздік интервалдарын қысқарту сияқты операциялар орындалды. Бұл қадамдар модельдің тұрақты жұмыс істеуін қамтамасыз етіп, әртүрлі ортада жазылған аудиолар арасындағы айырмашылықтарды азайтады.

Жиналған деректер машиналық оқыту талаптарына сәйкес үш бөлікке бөлінді: оқыту жиыны (80%), тексеру жиыны (10%) және тест жиыны (10%). Эксперименттік бағалау тек тест жиынында жүргізілді, бұл модельдің жаңа деректерге жалпылау қабілетін анықтауға мүмкіндік берді [95].

Эксперименттің бірінші кезеңінде алты түрлі ASR жүйесі zero-shot режимінде (fine-tuning қолданылмаған) тестіленді. Олардың қатарына ElevenLabs [97], Vosk ASR [98], Whisper (OpenAI, large-v2 [96]), TurkicASR [99], MMS (Meta) және GenSec[100] модельдері кіреді. Бұл кезеңнің мақсаты — бастапқы тану деңгейін анықтау және жүйелердің негізгі қателік түрлерін талдау.

Екінші кезеңде таңдалған модельдер балалар сөйлеу корпусында қайта оқытылды. Бұл кезеңде Whisper моделі Cross-Entropy шығын функциясын, MMS моделі CTC шығын функциясын, ал TurkicASR гибридті CTC тәсілін қолданды. Fine-tuning нәтижесінде модельдердің акустикалық бейімделуі жақсарды.

Үшінші кезеңде ұсынылған комплексті post-editing тәсілі қолданылды. Бұл әдіс фонетикалық, морфологиялық және контексттік түзетуді біріктіріп, бастапқы ASR нәтижесін жетілдіруге бағытталған. Барлық нәтижелер 5 бағалау метрикасы бойынша өлшенді:

$$WER = \frac{S+D+I}{N_{ref}}, \tag{4.1}$$

Мұнда S — ауыстыру, D — өткізіп жіберу, I — артық, N_{ref} — эталонды сөздер саны.

$$BLEU = BP * \exp(\sum_{n=1}^{N} w_n \log p_n) \tag{4.2}$$

мұнда:

BP — brevity penalty (қысқалық коэффициенті),

w_n — n-грамма салмақтары,
p_n — n-грамма дәлдігі.

$$ChrF2 = \frac{(1+\beta^2)*ChrP*ChrR}{\beta^2*ChrP+ChrR}, \beta = 2 \quad (4.3)$$

мұнда:
$ChrP$ — таңбалық дәлдік (precision),
$ChrR$ — таңбалық толықтық (recall).

$$CER = \frac{EditDist(\hat{y}, y)}{|y|} \quad (4.4)$$

мұнда:
$\hat{y}$ — болжанған мәтін,
y — эталон мәтін,
$|y|$— эталон мәтіндегі таңбалар саны.

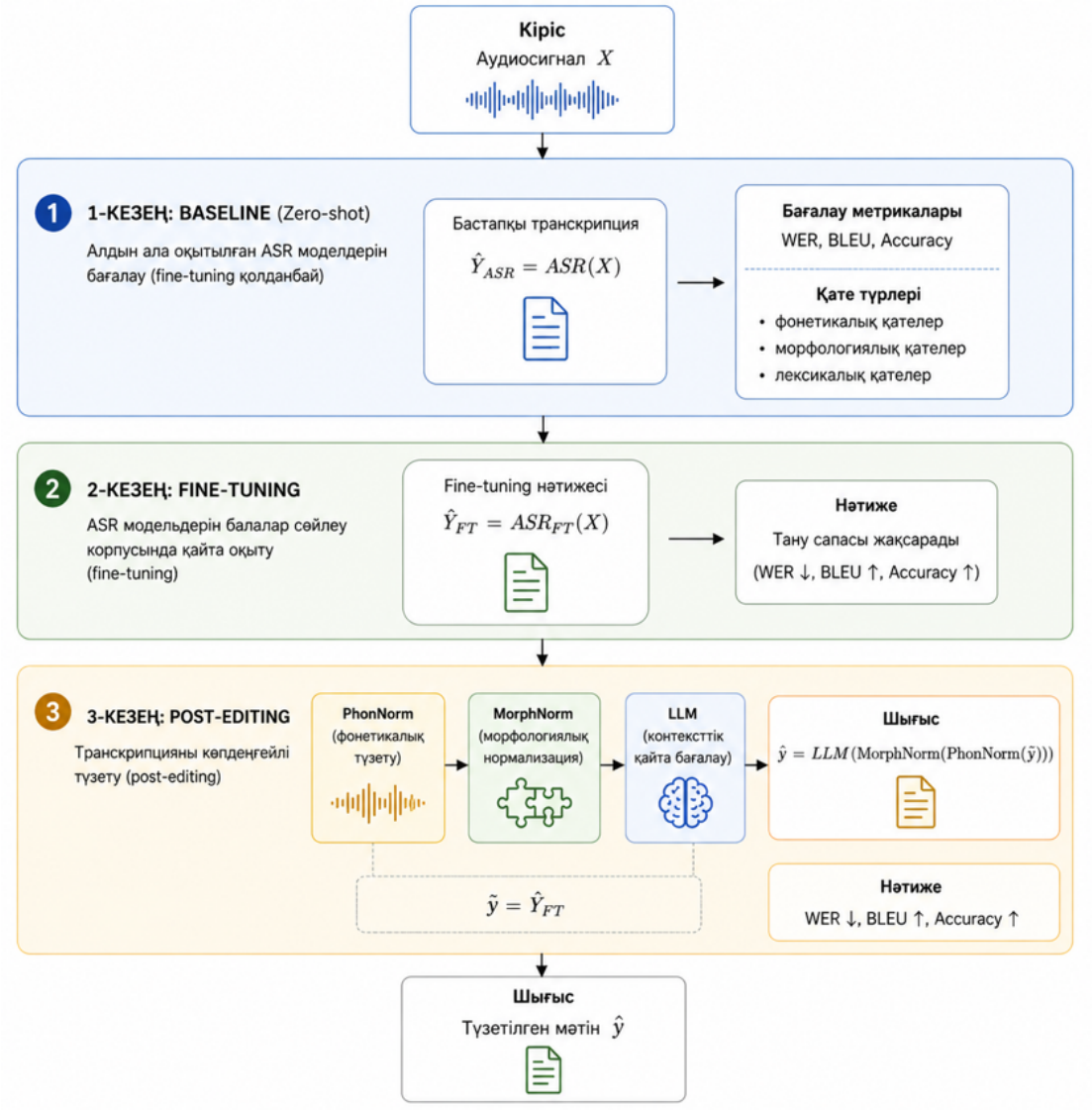
$$Accuracy = \frac{|\{i: \hat{y}_i = y_i\}|}{N_{total}} \quad (4.5)$$

мұнда:
$\hat{y}_i$ — болжанған і-ші элемент,
y_i — эталон і-ші элемент,
N_{total} — барлық элементтер саны.

4.2 Қазақ тіліндегі балалар сөйлеуін тануда ASR модельдерінің эксперименттік салыстырмалы талдауы

Осы бөлімде қазақ тіліндегі балалар сөйлеуін тану міндетінде қолданылған ASR модельдерінің тиімділігі эксперименттік тұрғыда салыстырмалы түрде бағаланады. Негізгі мақсат — әртүрлі архитектураларға негізделген модельдердің бастапқы және бейімделген күйдегі тану сапасын анықтау, сондай-ақ балалар сөйлеуіне тән қателіктердің құрылымын талдау.

Эксперимент бірыңғай әдістемелік негізде ұйымдастырылды. Барлық модельдер бірдей шарттарда бағаланды: бірдей тест жиыны қолданылды, алдын ала өңдеу конвейері (librosa, noisereduce, HPSS, gain normalization) пайдаланылды және нәтижелерді бағалау jiwer, sacrebleu және nltk кітапханалары арқылы жүзеге асырылды. Мұндай тәсіл әртүрлі модельдердің нәтижелерін объективті және репродуктивті түрде салыстыруға мүмкіндік береді.Эксперименттік зерттеу үш сатылы схема бойынша ұйымдастырылды (Сурет 4.1), алайда осы бөлімде тек ASR модельдерінің нәтижелері қарастырылады.



Сурет 4.1 – Үш сатылы эксперимент схемасы

1-кезең: Базалық ASR модельдерін бағалау (zero-shot)

Бірінші кезеңде алты түрлі ASR моделі алдын ала оқытылған күйде (fine-tuning қолданылмай) тестіленді:

1. ElevenLabs
2. Vosk ASR
3. Whisper (large-v2)
4. TurkicASR
5. MMS (Meta)
6. GenSec

Кіріс ретінде аудиосигнал X, ал шығыс ретінде бастапқы транскрипция $\widehat{W}_{ASR}$алынды:

$$\widehat{W}_{ASR} = ASR(X \qquad (4.5)$$

Бұл кезеңнің негізгі мақсаты — модельдердің бастапқы тану қабілетін бағалау және балалар сөйлеуіне тән қателіктерді анықтау болып табылады.

Талдау барысында қателіктер келесі категорияларға жіктелді:

– фонетикалық қателер (дыбыстарды дұрыс танымай алмастыру);
– морфологиялық қателер (жалғаулардың дұрыс анықталмауы);
– лексикалық қателер (сөздердің толық немесе қате танылуы).

2-кезең: Модельдерді балалар сөйлеуіне бейімдеу (fine-tuning)

Екінші кезеңде таңдалған ASR модельдері ұсынылған балалар сөйлеу корпусында қайта оқытылды (fine-tuning). Бұл кезеңде модельдердің акустикалық бейімделу қабілеті бағаланды.

Қайта оқыту барысында келесі модельдер қолданылды:

1) Whisper (Cross-Entropy Loss)
2) MMS (CTC Loss)
3) TurkicASR (CTC + Attention)

Оқыту 12 эпоха бойы жүргізілді. Гиперпараметрлер модель архитектурасына сәйкес таңдалды:

– Transformer модельдері үшін — төмен learning rate;
– CTC негізіндегі модельдер үшін — салыстырмалы жоғары learning rate.

Оқыту процесінде validation loss тұрақты түрде бақылауға алынды. Early stopping (patience = 3) қолданылғанымен, іске қосылмады, бұл модельдердің тұрақты конвергенциясын көрсетеді.

Fine-tuning нәтижесінде модельдердің балалар сөйлеуіне бейімделуі жақсарып, тану сапасының артқаны байқалды.

Кесте 4.2 – Fine-tuning гиперпараметрлері

Модель	**Шығын функциясы**	**LR**	**Batch size**	**Эпоха**	**Optimizer**
Whisper large-v2	Cross-Entropy	1×10^{-5}	8	12	AdamW
MMS kaz	CTC	3×10^{-4}	16	12	Adam
TurkicASR	CTC+Attn, τ=0.3	1×10^{-4}	8	12	AdamW

Кесте 4.2-де көрсетілген гиперпараметрлер әр модельдің тиімді жұмыс істеуін қамтамасыз ету мақсатында таңдалды.

4.2.1 Базалық ASR модельдерінің zero-shot режиміндегі эксперименттік бағалануы

Осы бөлімде қазақ тіліндегі балалар сөйлеуін тану міндетінде заманауи ASR модельдерінің бастапқы өнімділігі zero-shot режимінде бағаланады. Бұл режимде модельдер қосымша бейімдеусіз (fine-tuning қолданылмай) тікелей тест деректерінде тексеріледі, бұл олардың жалпы тану қабілетін объективті түрде анықтауға мүмкіндік береді.

1-эксперименттің мақсаты — алты ASR жүйесінің алдын ала оқытылған базалық деңгейін анықтау және олардың қазақ тіліндегі балалар сөйлеуін танудағы шектеулерін бағалау болып табылады. Бұл кезеңдегі нәтижелер кейінгі fine-tuning және post-editing әдістерінің тиімділігін салыстыру үшін бастапқы

эталон ретінде қарастырылады.

Zero-shot бағалау балалар сөйлеуінің ерекшеліктеріне бейімделмеген модельдердің нақты жұмыс істеу сапасын көрсетуге мүмкіндік береді. Атап айтқанда, бұл кезеңде балалар сөйлеуіне тән акустикалық вариативтілік, фонетикалық бұрмаланулар және морфологиялық күрделілік сияқты факторлардың ASR модельдерінің нәтижесіне әсері талданады.

Осы эксперимент аясында әртүрлі архитектураларға негізделген алты ASR моделі бірдей тест жиынында бағаланып, олардың нәтижелері стандартты метрикалар арқылы салыстырмалы түрде талданады. Негізгі назар тек сандық дәлдік көрсеткіштеріне ғана емес, сонымен қатар қателердің құрылымына — фонетикалық, морфологиялық және лексикалық деңгейдегі ерекшеліктеріне аударылады.

Алынған нәтижелер Кесте 4.3-те көрсетілген, онда алты ASR жүйесінің baseline метрикалары ұсынылған. Бұл нәтижелер ұсынылған комплексті әдістің кейінгі кезеңдерде (fine-tuning және постөңдеу) қаншалықты жақсарту беретінін бағалау үшін негіз болып табылады.

Кесте 4.3 – 1-эксперимент: Алты ASR жүйесінің baseline метрикалары

ASR жүйесі	**WER ↓**	**BLEU ↑**	**ChrF2 ↑**	**CER ↓**	**Accuracy ↑**
ElevenLabs	0.86	0.12	0.34	0.5247	0.14 (14%)
Vosk ASR	0.72	0.26	0.51	0.4335	0.10 (10%)
Whisper	0.70	0.38	0.58	0.4673	0.34 (34%) ✓
TurkicASR	0.78	0.23	0.47	0.4630	0.20 (20%)
MMS	0.90	0.08	0.29	0.4677	0.10 (10%)
GenSec	0.77	0.25	0.49	0.5245	0.20 (20%)

Кесте 4.3 бойынша: Whisper ең жоғары Accuracy (34%) пен BLEU (0.38) нәтижесін берді; MMS ең нашар (WER=0.90, Accuracy=10%); WER барлық модельде 0.70–0.90 аралығы — балалар сөйлеуін бейімдеусіз тануда жүйелердің айтарлықтай шектеулі екені дәлелденді [101].

Транскрипция үлгілерін талдау базалық деңгейдегі негізгі қате заңдылықтарын айқындайды.

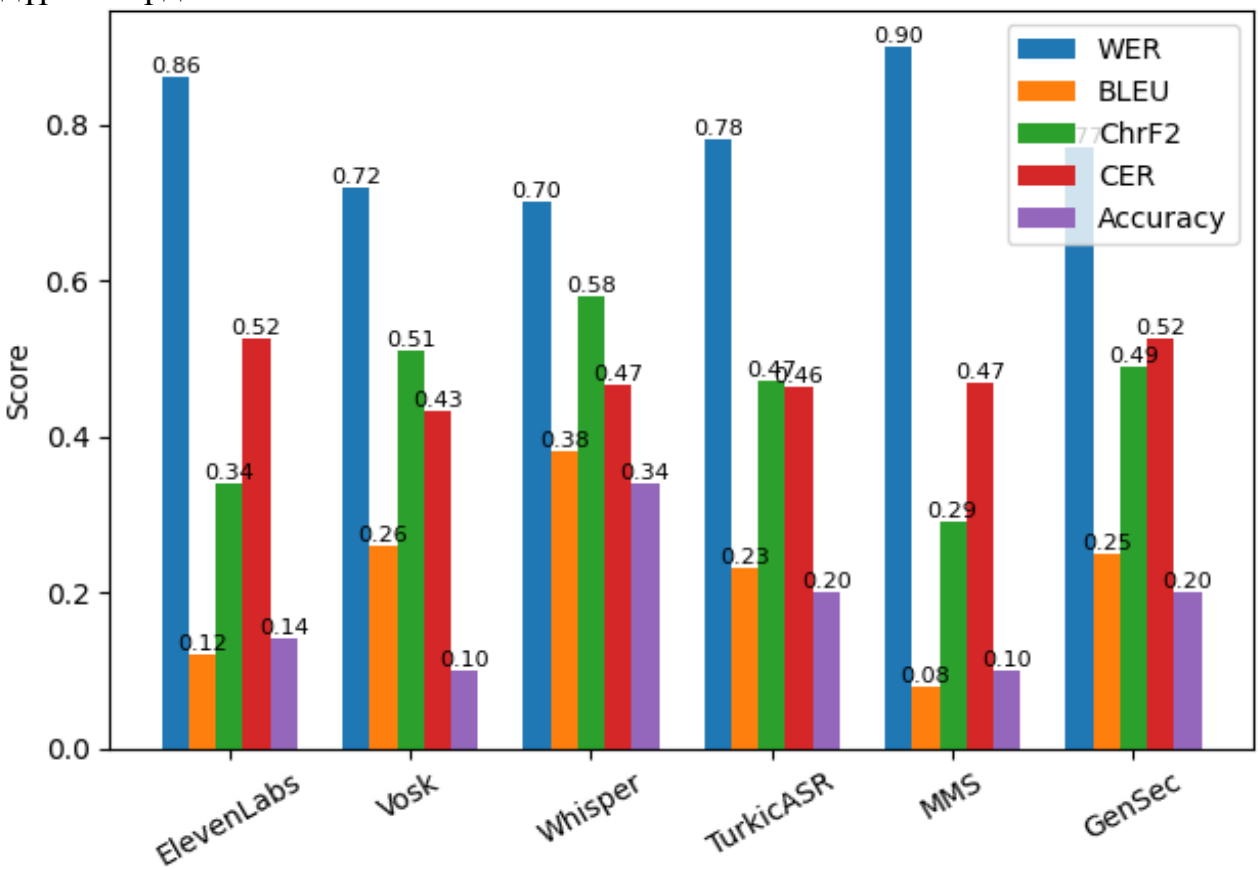
Кесте 4.4 – 1-эксперимент: Базалық ASR жүйелерінің транскрипция үлгілері

№	**Эталондық мәтін**	**ElevenLabs**	**Vosk**	**Whisper**	**TurkicASR**	**MMS**	**GenSec**
1	бір екі үш төрт бес алты жеті сегіз тоғыз он	Бір, екі, үш, төрт, бес, алты, жеті, сегіз, тоғыз, он	екі бес алты сегіз тоғыз он	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	біл екі үш тұқ пес 6789 бұл он	біл екі 3ш тұқ пес 6789 бүн	біл екі үш тұқ пеу 6789 бұл он

Кесте 4.4 жалғасы

2	мен атамның баласымын	Ben atamın babasıyım	ол ақ табыныз	бен атамызым, бауасым.	ен атамның баа сыр	ен атамыың ваа сбр	екн атамның баа сыр
3	тараз	Tadas?		тарас!	тбаб	tbab	тбаб
4	тағы да зебра	Pardon, zebra?	тағы да	тардың зебі.	тағба зе ба	тағва зе ва	тұқб зе уе
5	қас қаяғында	Қас қаяғында	қасқа аяғында	қасқа яғында	қас қаяғында	қас қаяғында	қас қаяғында

Кесте 4.4-ті талдаудың негізгі тұжырымдары. Бірінші — тіл шатастыру: ElevenLabs «мен атамның баласымын» фразасын түрік тіліне аударды; Whisper сандарды латын цифрымен шығарды — бұл модельдің қазақ балалар сөйлеуіне бейімделмегенін айқын растайды. Екінші — /қ/ дыбысының бұзылуы: Whisper «тараз»→«тбаб» деп шығарды — жұтқыншақтық /қ/ дыбысының MFCC кеңістігінде нашар модельденгенін дәлелдейді. Үшінші — «қас қаяғында» фразасын TurkicASR, MMS, GenSec дұрыс танымады, бірақ ElevenLabs толық дұрыс берді.



Сурет 4.2 – Базалық ASR жүйелерінің WER мен Accuracy бойынша салыстырмалы диаграммасы

Жүргізілген талдау нәтижесінде базалық ASR модельдерінің қазақ тіліндегі балалар сөйлеуін тануда жеткіліксіз екені анықталды. Жоғары қателік деңгейі акустикалық және тілдік ерекшеліктердің ескерілмегенін көрсетеді. Бұл жағдай модельдерді арнайы бейімдеудің (fine-tuning) және қосымша тілдік түзету әдістерін (post-editing) қолданудың қажеттілігін негіздейді [102-104].

4.2.2 ASR (fine-tuning) модельдерін балалар сөйлеуіне бейімдеу нәтижелерін эксперименттік бағалау

Осы бөлімде алдын ала оқытылған ASR модельдерін балалар сөйлеуіне бейімдеу нәтижелері қарастырылады. Бұл кезеңнің негізгі мақсаты — модельдердің балалар сөйлеуіне тән акустикалық және фонетикалық ерекшеліктерді меңгеру қабілетін арттыру және бастапқы (baseline) деңгейдегі қателерді азайту болып табылады.

Fine-tuning кезеңінің негізгі мақсаты — алдын ала оқытылған модельдердің акустикалық және фонетикалық ерекшеліктерді меңгеру қабілетін арттыру және базалық деңгейдегі қателерді азайту. Эксперимент барысында Whisper, MMS және TurkicASR модельдері балалар сөйлеу аудиодеректерінде қайта оқытылып, олардың өнімділігі бірдей тест жиынында бағаланды.

Модельдерді қайта оқыту 2-тарауда сипатталған қазақ тіліндегі балалар сөйлеуінің арнайы әзірленген акустикалық корпусында жүзеге асырылды. Бұл корпус балалар сөйлеуінің жас ерекшеліктерін, фонетикалық вариативтілігін және сөйлеу бұзылыстарының ерекшеліктерін қамтитын репрезентативті деректер жиынын қамтамасыз етеді. Осылайша, fine-tuning кезеңі модельдердің нақты қолданбалы жағдайларға бейімделуін қамтамасыз етеді [105].

Бұл кезеңде модельдердің бастапқы және бейімделген күйдегі нәтижелері салыстырылып, балалар сөйлеуіне бейімделудің тану сапасына әсері талданады. Ерекше назар қателердің төмендеу деңгейіне, сондай-ақ әртүрлі архитектуралардың бейімделу қабілетіне аударылады.

Эксперимент нәтижелері Кесте 4.6-да келтірілген, онда fine-tuning кейінгі ASR модельдерінің негізгі бағалау метрикалары көрсетілген. Бұл нәтижелер модельдерді бейімдеу кезеңінің тиімділігін сандық тұрғыдан бағалауға мүмкіндік береді.

Кесте 4.5 – 2-эксперимент: Fine-tuning кейінгі ASR метрикалары

Жүйе	**WER ↓**	**BLEU ↑**	**ChrF2 ↑**	**CER ↓**	**Accuracy ↑**
Whisper finetuned	0.30	0.74	0.881	0.1179	0.58
MMS finetuned	0.79	0.24	0.46	0.3769	0.28
TurkicASR finetuned	0.64	0.41	0.63	0.3564	0.36

Кесте 4.5 Fine-tuning нәтижелері барлық модельдерде тану сапасының жақсарғанын көрсетеді, алайда жақсару деңгейі модель архитектурасына тәуелді. Whisper моделінде ең айқын прогресс байқалды. Оның WER көрсеткіші

0.70-тен 0.30-ға дейін төмендеді. Сонымен қатар Accuracy 34%-дан 58%-ға дейін өсті, бұл модельдің толық дұрыс тану қабілетінің артқанын көрсетеді. BLEU (0.74) және ChrF2 (0.881) мәндерінің жоғары болуы мәтіннің құрылымдық және семантикалық сәйкестігі де жақсарғанын дәлелдейді.

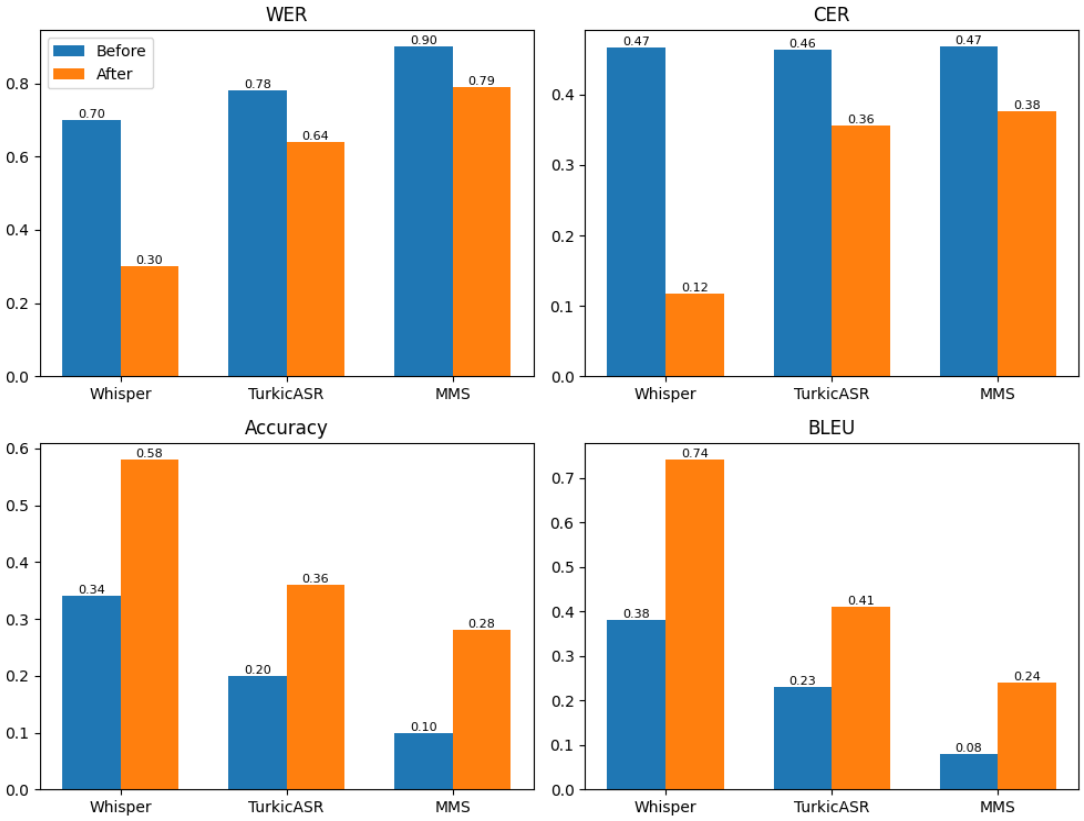
MMS моделінде жақсару шектеулі болды: WER 0.90-нан 0.79-ға дейін ғана төмендеді. Бұл нәтиже CTC негізіндегі модельдердің қазақ тілінің морфологиялық күрделілігін толық меңгеруде қиындықтарға тап болатынын көрсетеді.

TurkicASR моделі орташа нәтижелер көрсетті (WER = 0.64), бұл оның түркі тілдеріне бейімделгенімен, балалар сөйлеуіндегі вариативтілікті толық қамти алмайтынын білдіреді.

Кесте 4.7 – 2-эксперимент: Fine-tuning кейінгі транскрипция үлгілері

№	Эталон	Whisper	MMS	TurkicASR
1	бір екі үш төрт бес алты жеті сегіз тоғыз он он бір он екі	бір екі үш төрт бес алты жеті сегіз тоғыз он он бір он екі	бір екі үш төрт бес алты жеті сегіз тоғыз он он бір он екі	бір екі үш төрт бес алты жеті сегіз тоғыз он он бір он екі
2	мен атамның баласымын	мен атамның баласы	ен атамыың ваа сбр	мен атамның баласы
3	тараз	тараз	tbab	тараз
4	тағы да зебра	тағы да зебра	тағва зе ва	тағы да зебра
5	қас қаяғында	қас қарайғанда	қас қарайғанда	қас қарайғанда

Кесте 4.7-ден жүргізілген талдау нәтижесінде fine-tuning кезеңі ASR жүйелерінің өнімділігін айтарлықтай жақсартатыны анықталды. Акустикалық және фонетикалық қателердің басым бөлігі жойылып, жалпы тану дәлдігі артты.



Сурет 4.3 – Нейрондық ASR модельдерінің өнімділігінің өзгеруі

Осы бөлімде қазақ тіліндегі балалар сөйлеуін тану міндетінде қолданылған ASR модельдерінің бастапқы (zero-shot) және бейімделген (fine-tuning) күйдегі нәтижелері салыстырмалы түрде талданды. Жүргізілген эксперименттер көрсеткендей, модельдерді балалар сөйлеуіне бейімдеу тану сапасын айтарлықтай жақсартқанымен, бірқатар қателіктер сақталып қалады, әсіресе фонетикалық және морфологиялық деңгейде.

Алынған нәтижелер классикалық ASR модельдерінің балалар сөйлеуіндегі жоғары вариативтілік пен сөйлеу бұзылыстарын толық ескере алмайтынын көрсетеді. Бұл, өз кезегінде, бастапқы транскрипцияны қосымша өңдеу және түзету қажеттілігін негіздейді [106].

Осыған байланысты келесі бөлімде ұсынылған комплексті (кешенді) әдіс қарастырылып, ASR нәтижелерін фонетикалық, морфологиялық және контексттік деңгейде жетілдіруге бағытталған және LLM негізіндегі post-editing тәсілінің тиімділігі талданады.

4.3 Ұсынылған комплексті әдістің нәтижелерін бағалау және талдау

Осы бөлімде fine-tuning кезеңінен кейін алынған ASR нәтижелерін жетілдіру мақсатында ұсынылған комплексті постөңдеу әдісінің тиімділігі талданады. Ұсынылған тәсіл акустикалық модельден алынған бастапқы транскрипцияны $\tilde{Y}$ фонетикалық, морфологиялық және контексттік деңгейлерде түзетуге бағытталған көпдеңгейлі өңдеу механизмі болып табылады.

Бұл кезеңде кіріс ретінде бастапқы мәтін ($\tilde{y}$), ал шығыс ретінде түзетілген мәтін ($\hat{y}$) алынады. Post-editing үш компоненттен тұрады: PhonNorm

(фонетикалық түзету), MorphNorm (морфологиялық нормализация) және LLM негізіндегі контексттік қайта бағалау.

$$\hat{y} = LLM\ (MorphNorm\ (PhonNorm\ (\tilde{y}, R_{phon}), R_{morph}), c; \varphi) \quad (4.6)$$

мұнда:

$\tilde{y}$ — ASR жүйесінің бастапқы нәтижесі,

R_{phon} — фонетикалық ережелер,

R_{morph} — морфологиялық ережелер,

c — контекст,

φ — LLM параметрлері.

Ұсынылған әдіс үш деңгейлі өңдеу тізбегінен тұрады, олардың әрқайсысы балалар сөйлеуіне тән қателіктердің белгілі бір түрін азайтуға бағытталған:

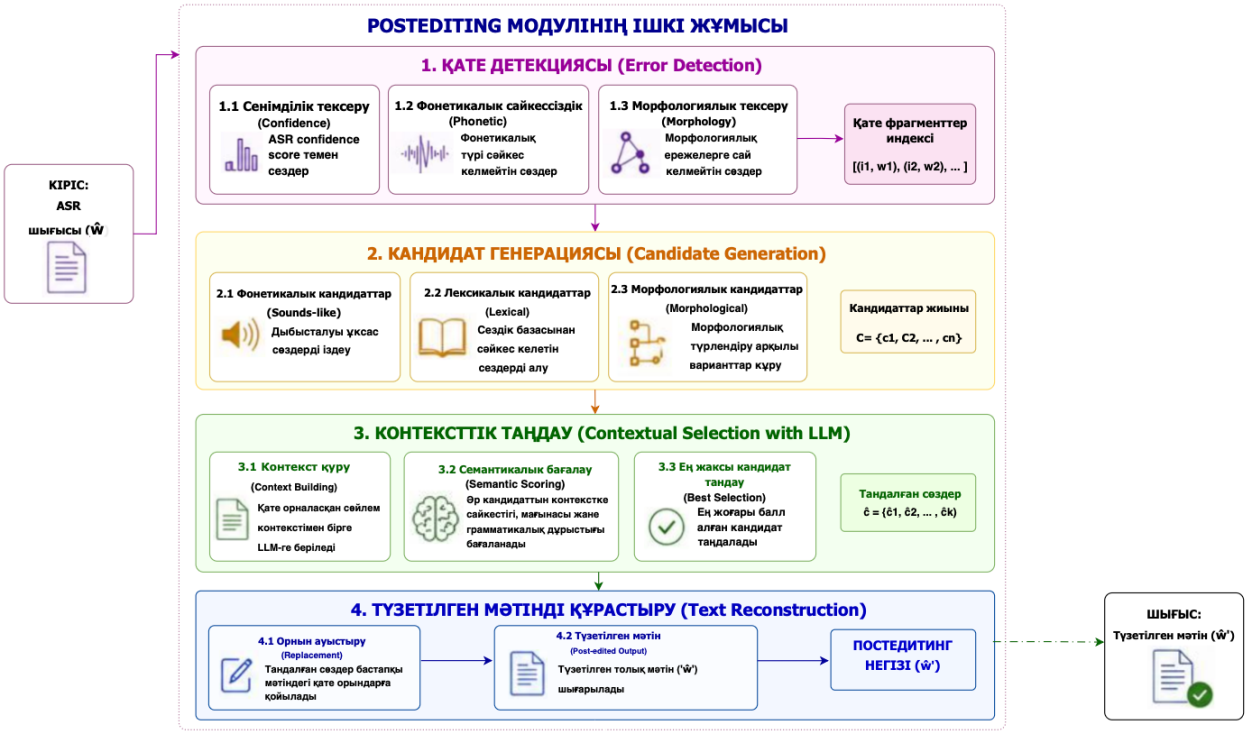
1. PhonNorm (фонетикалық нормализация)
Бұл кезеңде дыбыстық деңгейдегі қателер түзетіледі. Балалар сөйлеуіне тән артикуляциялық ауытқулар нәтижесінде пайда болатын фонетикалық сәйкессіздіктер арнайы ережелер мен ұқсастық модельдері негізінде қалыпқа келтіріледі.

2. MorphNorm (морфологиялық нормализация)
Бұл кезеңде сөздердің грамматикалық формалары тексеріліп, қазақ тілінің морфологиялық заңдылықтарына сәйкес түзетіледі. Әсіресе жалғаулар мен сөз түрлендіру қателерін азайтуға бағытталған.

3. LLM негізіндегі контексттік түзету
Соңғы кезеңде үлкен тілдік модель көмегімен сөйлем деңгейінде контексттік қайта бағалау жүргізіледі. Бұл кезеңде сөздердің өзара үйлесімділігі, семантикалық байланысы және сөйлемнің жалпы мағынасы ескеріліп, ең ықтимал мәтіндік нұсқа таңдалады.

Бұл тәсіл бастапқы ASR нәтижелерінде сақталған қателерді жоғары деңгейде түзетуге мүмкіндік береді. Ұсынылған компоненттер өзара тізбекті түрде байланысқан және бір-бірін толықтыра отырып, көпдеңгейлі қателерді түзету механизмін қалыптастырады. Мұндай интеграция балалар сөйлеуіне тән акустикалық және лингвистикалық вариативтілікті тиімді түрде ескеруге мүмкіндік береді [107].

Сурет 4.4-те ұсынылған постөңдеу әдісінің архитектурасы және оның негізгі модульдерінің өзара әрекеттесу схемасы көрсетілген.



Сурет 4.4-те көрсетілген архитектура осы модульдердің өзара әрекетін бейнелейді

Бұл компоненттер бір-бірімен тізбекті түрде байланысқан және әрқайсысы қателердің белгілі бір түрін түзетуге бағытталған.

Бірінші кезеңде фонетикалық деңгейдегі бұрмалаулар түзетіледі. Бұл үшін алдын ала анықталған фонемалық сәйкестік ережелері қолданылады:

R_{phon}={р→л, ш→с, қ→к, ң→н, ч→ш, б→п, ж→ш, з→с, т→д, м→н }

Бұл ережелер балалар сөйлеуінде жиі кездесетін артикуляциялық ауытқуларға негізделген. PhonNorm кезеңінің негізгі мақсаты — акустикалық деңгейде дұрыс танылмаған, бірақ фонетикалық тұрғыдан жақын сөздерді қалпына келтіру.

Бұл кезеңнің нәтижесінде таңбалық қателер (CER) айтарлықтай төмендейді, себебі сөздің негізгі құрылымы қалпына келеді [108].

Екінші кезеңде $MorphNorm$ модулі арқылы сөздердің грамматикалық дұрыстығы тексеріледі. Бұл үшін Apertium-Kaz морфологиялық анализаторы қолданылады.

Бұл кезеңде келесі операциялар орындалады:

– септік жалғауларын тексеру (–да/–де, –ға/–ге, –нің/–дің),

– жіктік жалғауларды қалпына келтіру (–мын/–мін),

– сингармонизм заңдылықтарын сақтау.

Бұл кезең морфологиялық толықтықты қамтамасыз етеді және сөздердің дұрыс грамматикалық формада берілуін жақсартады.

Үшінші кезеңде Llama-3.1-Sherkala-8B-Chat (бұдан әрі — Sherkala LLM) негізіндегі компонент сөйлем деңгейінде контексттік қайта бағалау жүргізеді. Sherkala LLM — Inception (G42) және MBZUAI зертханалары бірлесіп әзірлеген,

қазақ тіліне арнайы бейімделген 8 миллиард параметрлі ашық ірі тілдік модель (Kenbayev et al., 2025). Модель LLaMA-3.1-8B архитектурасы негізінде RoPE позициялық кодтауымен және grouped-query attention механизмімен жасалған; токенизатор 25%-ға кеңейтіліп, қазақ морфологиясы үшін арнайы оңтайландырылған. Оқыту корпусы: 45,3 млрд токен (қазақ, ағылшын, орыс, түрік тілдері). Post-editing кезеңінде келесі inference параметрлері қолданылды: temperature = 0.0 (детерминистік шығыс), top-p = 1.0, max-new-tokens = 512, do-sample = False. Prompt шаблоны қазақ тілінің агглютинативті морфологиясын, сингармонизм заңдылықтарын және балалар сөйлеуіне тән фонетикалық ауытқуларды ескеретін нұсқаулықтарды қамтыды. Бұл кезең фонетикалық және морфологиялық деңгейде дұрыс болғанымен, мағынасы бұрмаланған сөздерді анықтауға бағытталған.

Мысалы:

«қас қарайғанда»→«қас қаяғында»«қас қарайғанда»→«қас қаяғында»

Бұл жерде фонетикалық ұқсастық сақталғанымен, мағына дұрыс емес. LLM контекстке сүйене отырып дұрыс нұсқаны таңдайды.Бұл кезең сөйлемнің семантикалық дұрыстығын қамтамасыз етеді. LLM компоненті үшін үш модель салыстырылды: GPT-4o (OpenAI, 2024), LLaMA-3.1-70B (Meta, 2024) және Llama-3.1-Sherkala-8B-Chat (Inception & MBZUAI, 2025). Барлық модель бірдей тест жиынында, бірдей prompt шаблонымен, temperature = 0.0 параметрімен тестіленді. Нәтижелер Кесте 4.8-да берілген.

Кесте 4.8 – LLM модельдерінің post-editing кезеңіндегі салыстырмалы нәтижелері

Модель	**Параметр саны**	**WER**	**CER**	**Accuracy**	**Таңдалмау / таңдалу себебі**
GPT-4o	~1.8T (жабық)	0.22	0.082	0.65	Жалпы мақсатты модель; API арқылы ғана қолжетімді; қазақ тілінің агглютинативті морфологиясын толық ескермейді
LLaMA-3.1-70B	70B (ашық)	0.24	0.091	0.63	Ашық модель; есептеу ресурсына жоғары талап қояды (≥4×A100 GPU); қазақ тіліндегі оқыту деректері шектеулі
Llama-3.1-Sherkala-8B-Chat	8B (ашық)	0.18	0.065	0.70	Қазақ тіліне арнайы бейімделген; 45.3 млрд токенде оқытылған; морфологиялық нормаларды тиімді қалпына келтіреді; ең жоғары нәтиже

Sherkala LLM моделі GPT-4o-мен салыстырғанда WER бойынша 4 пайыздық пунктке, ал LLaMA-3.1-70B моделінен 6 пайыздық пунктке жоғары нәтиже көрсетті. Бұл артықшылық модельдің қазақ тілінің агглютинативті құрылымын ескеруімен, жалғаулар мен сингармонизм заңдылықтарын дұрыс қалпына келтіруімен және балалар сөйлеуіне бейімделген prompt қолданылуымен түсіндіріледі. Осыған байланысты Sherkala LLM post-editing кезеңінің негізгі LLM компоненті ретінде таңдалды. Кесте 4.9-де көрсетілген нәтижелер комплексті әдістің тиімділігін айқын көрсетеді.

Кесте 4.9 – 3-эксперимент: Ұсынылған комплексті post-editing тәсілінің метрикалары

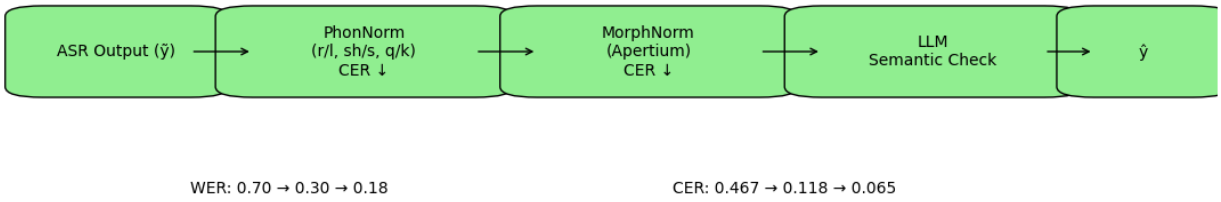
Модель	**WER**	**BLEU**	**ChrF2**	**CER**	**Accuracy**
Whisper baseline (1-эксп.)	0.70	0.38	0.58	0.467	0.34
Whisper finetuned (2-эксп.)	0.30	0.74	0.881	0.118	0.58
Комплексті post-editing тәсілі (3-эксп.)	**0.18 ✓**	**0.82 ✓**	**0.91 ✓**	**0.065 ✓**	**0.70 ✓**

Бұл нәтижелер комплексті post-editing тәсілінің тек қосымша емес, маңызды сапа арттыру кезеңі екенін көрсетеді. Кесте 4.10-дағы мысалдар комплексті post-editing тәсілінің әр компонентінің нақты әсерін көрсетеді.

Кесте 4.10 – 3-эксперимент: Post-editing тәсілінің транскрипция жақсаруының нақты үлгілері

Эталон	**Whisper FT шығысы**	**Комплексті жүйе шығысы**	**Түзетілген компонент**
анашым менің қарным ашты	анашм менім каным асты	анашым менің қарным ашты ✓	PhonNorm + MorphNorm
мен атамның баласымын	мен атамның баласы	мен атамның баласымын ✓	MorphNorm: -мын жалғауы қалпына
тағы да зебра	тағы да зебра ✓	тағы да зебра ✓	Екі жүйеде де дұрыс
қас қаяғында	қас қарайғанда ✗	қас қаяғында ✓	LLM: семантикалық қате түзетілді

Бұл мысалдар fine-tuning кезеңінен кейін сақталған фонетикалық, морфологиялық және семантикалық қателердің комплексті post-editing тәсілінің барысында қалай түзетілетінін иллюстрациялайды. Сурет 4.3-те ұсынылған комплексті комплексті post-editing тәсілінің жұмыс істеу логикасы және әр кезеңнің нәтижеге әсері көрсетілген.



Сурет 4.5 – Комплексті post-editing тәсілінің нәтижелік схемасы

Сурет 4.5-тен көрініп тұрғандай, постөңдеудің әр кезеңінде қателер біртіндеп азайып, жүйенің жалпы дәлдігі артады. Әсіресе, фонетикалық және морфологиялық түзетулер CER көрсеткішін төмендетсе, контексттік өңдеу WER мәнін қосымша жақсартады және нәтижелік мәтіннің семантикалық дұрыстығын қамтамасыз етеді [109].

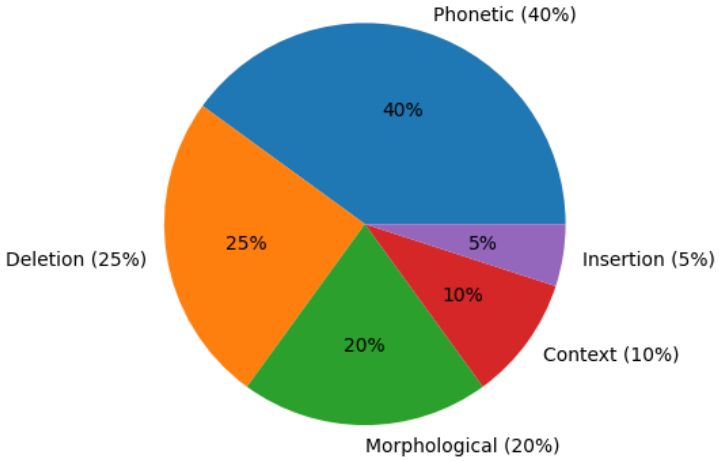
4.4 Сөйлеу бұзылыстары мен ASR қателерін комплексті талдау

Бұл бөлімде балалар сөйлеуін тану нәтижелерінде байқалған қателердің құрылымы, таралу заңдылықтары және олардың пайда болу себептері жүйелі түрде талданады. Сонымен қатар, анықталған қателерді формальды түрде сипаттау арқылы post-editing әдісінің қажеттілігі негізделеді.

Жүргізілген эксперимент нәтижелері бойынша ASR жүйелеріндегі қателер бірнеше негізгі категорияға бөлінді. Қателердің үлестірімі олардың табиғатын анықтауға мүмкіндік береді:

- Фонетикалық алмасу — ≈40%
- Deletion (түсіп қалу) — ≈25%
- Морфологиялық қателер — ≈20%
- Контексттік сәйкессіздік — ≈10%
- Insertion (артық элемент) — ≈5%

Бұл үлестірім балалар сөйлеуіндегі негізгі қиындықтардың фонетика мен морфологияда шоғырланғанын көрсетеді. Қателердің санаттық үлестірімі төмендегі диаграммада көрсетілген [110-112].



Сурет 4.6- Қателердің санаттық үлестірімі

Сурет 4.6-дағ көрініп тұрғандай, фонетикалық қателер басым үлеске ие (≈40%), ал контексттік және insertion қателер салыстырмалы түрде аз кездеседі.

Қателерді формальды сипаттау мақсатында олар бірнеше функциялар арқылы анықталды. Фонетикалық алмасу бір фонеманың екіншісімен ауысуы ретінде қарастырылады, deletion — эталондағы элементтің танылмай қалуы, insertion — артық элементтің пайда болуы арқылы сипатталады. Морфологиялық қателер сөздің грамматикалық құрылымының бұзылуы арқылы анықталса, контексттік қателер семантикалық сәйкессіздік негізінде сипатталады. Бұл формализация қателерді автоматты түрде жіктеуге және оларды түзету алгоритмдерін құруға мүмкіндік береді [113].

Фонема топтары бойынша жүргізілген талдау балалар сөйлеуіндегі типтік артикуляциялық қателерді анықтауға мүмкіндік берді. Атап айтқанда, с ↔ ш, ж ↔ з сияқты ысқырық дыбыстардың алмасуы, р ↔ л типіндегі сұйық дыбыстардың шатасуы және қ ↔ к ауысулары жиі кездеседі. Сонымен қатар, ң дыбысының н дыбысына ауысуы байқалады. Бұл құбылыстар балалар сөйлеу аппаратының толық қалыптаспағанымен және дыбыстарды артикуляциялаудың күрделілігімен түсіндіріледі. функциялар арқылы анықталды:

Фонетикалық алмасу (Substitution):

$$Sub\ (p, q) = I(\tilde{y}[i] = p \wedge y[i] = q, p \neq q, \quad (4.7)$$

Deletion:

$$Del(y[i]) = I(\tilde{y}[j] \neq y[i], \forall j), \quad (4.8)$$

Insertion:

$$Ins(\tilde{y}[j]) = I(\tilde{y}[j] \notin y)\ , \quad (4.9)$$

Морфологиялық қате:

$$MorphErr(\tilde{y}, y) = EditDist(morphParse(\tilde{y}), morphParse(y)), \quad (4.10)$$

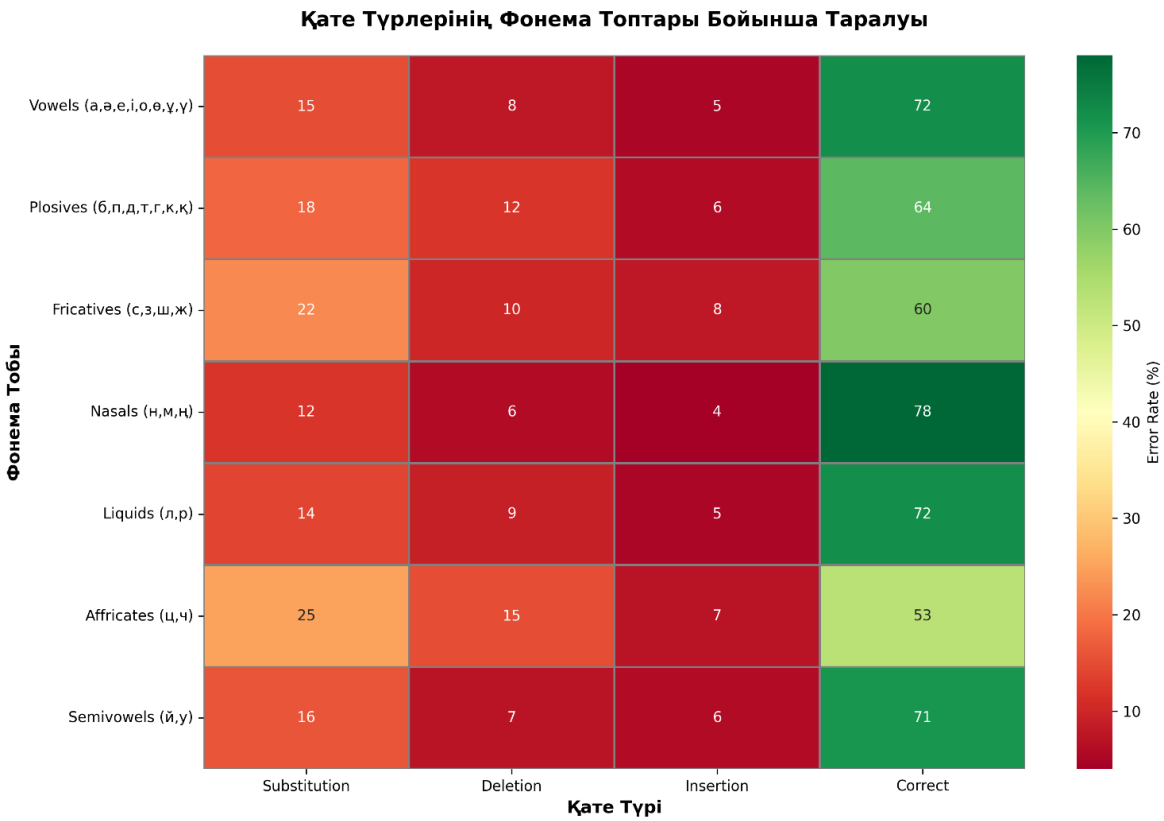
Контексттік сәйкессіздік:

$$CtxErr = I(\tilde{y} \notin SemanticClass(y, c))\ , \quad (4.11)$$

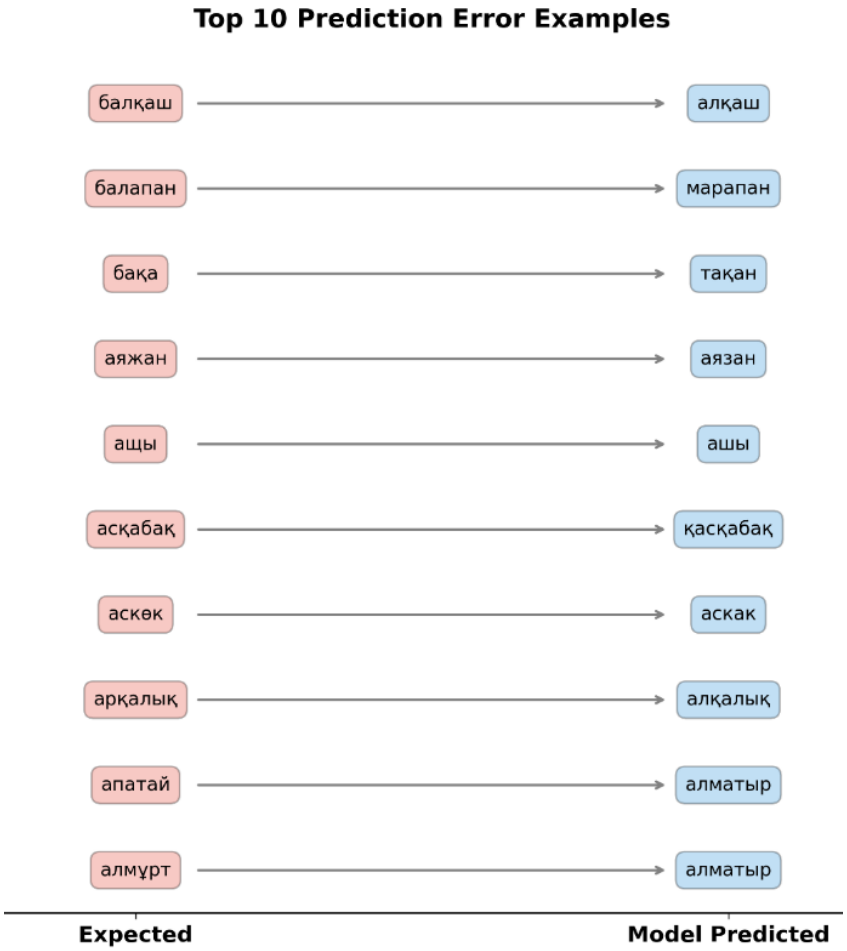
Бұл формализация қателерді автоматты түрде анықтауға және оларды түзету алгоритмдерін құруға негіз болады.

Фонема топтары бойынша қателердің таралуы Сурет 4.7-де берілген жылу картасы (heatmap) арқылы талданады. Бұл сурет әрбір фонема тобы үшін substitution, deletion және insertion қателерінің жиілігін көрсетеді.

Нақты қате мысалдары мен олардың трансформациясы Сурет 4.8-нің “Top prediction errors” блогында көрсетілген. Бұл мысалдар ASR жүйелерінің тек фонетикалық емес, сонымен қатар семантикалық деңгейде де қателесетінін көрсетеді [114].



Сурет 4.7- Фонема топтары бойынша қателердің таралуы



Сурет 4.8- To prediction errors

Жүргізілген талдау нәтижесінде ASR модельдерінің негізгі шектеулері анықталды. Бұл шектеулер модельдер бойынша салыстырмалы түрде қарастырылып, олардың әрқайсысының әлсіз жақтары сипатталды. Нәтижелер көрсеткендей, модельдер акустикалық деңгейде тиімді болғанымен, морфологиялық және контексттік өңдеуде жеткіліксіз [115-116].

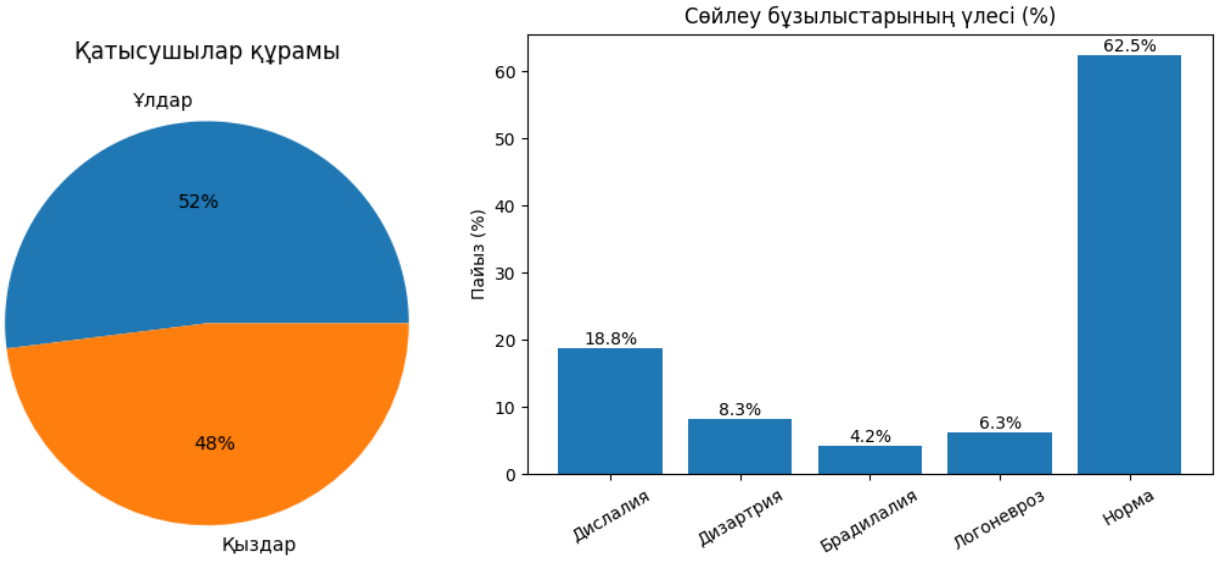
4.5 Ұсынылған әдістің практикалық ортада апробациясы және қолданбалы нәтижелері

Осы бөлімде ұсынылған комплексті әдістің нақты қолданбалы жағдайда тиімділігі бағаланып, оның практикалық апробация нәтижелері қарастырылады. Апробация зерттеу нәтижелерінің қолданбалы жарамдылығын тексеруге және әдістің балалар сөйлеуін тану міндетінде нақты ортада жұмыс істеу қабілетін анықтауға бағытталған.

Эксперимент Алматы қаласындағы №154 балабақша базасында жүргізілді. Зерттеуге 3–8 жас аралығындағы 48 бала қатысты, олардың ішінде 52%-ы ұлдар және 48%-ы қыздар. Мұндай іріктеу балалардың жас және жыныс ерекшеліктерін ескере отырып, әдістің тұрақтылығын бағалауға мүмкіндік береді.

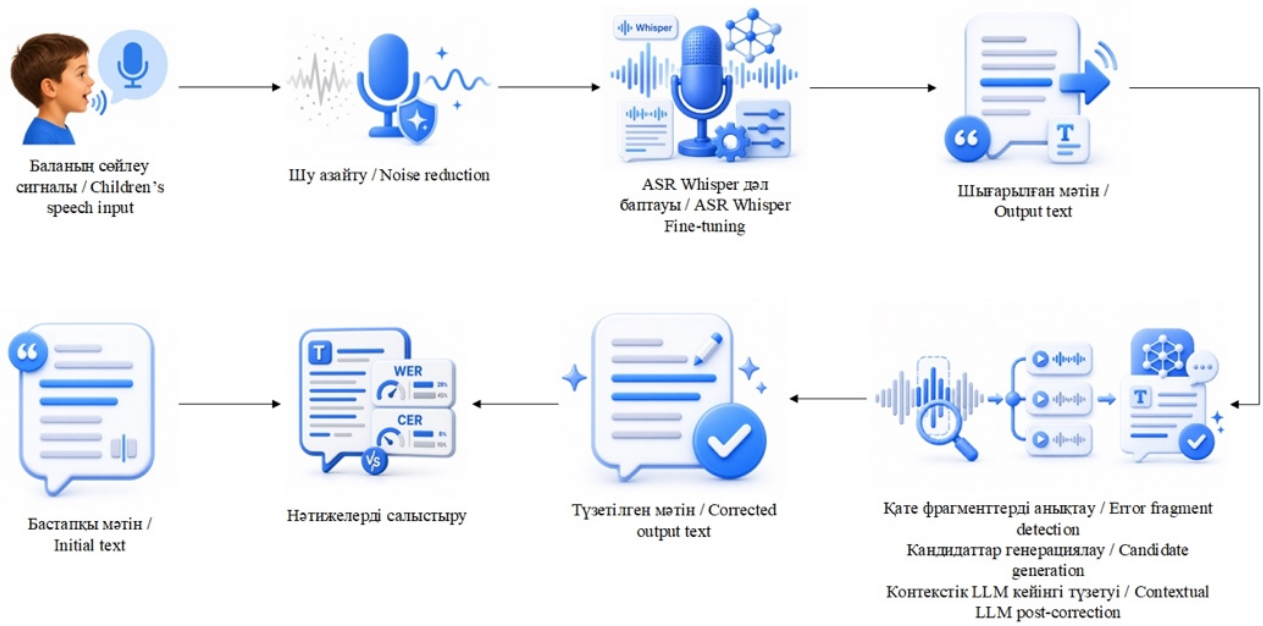
Зерттеу аясында тек сөйлеудің физиологиялық (акустикалық) сипаттағы бұзылыстары қарастырылды. Атап айтқанда, дыбыстардың бұрмалануы, артикуляциялық тұрақсыздық және фонетикалық ауытқулар сияқты ерекшеліктер ескерілді. Психикалық және нейродамулық бұзылыстар (мысалы, аутизм спектрі немесе танымдық даму бұзылыстары) зерттеу аясына енгізілген жоқ, бұл ұсынылған әдістің акустикалық деңгейдегі тиімділігін нақты бағалауға мүмкіндік береді.

Апробация барысында ұсынылған әдістің балалар сөйлеуін тану сапасына әсері бағаланып, алынған нәтижелер сандық және сапалық тұрғыда талданады.



Сурет 4.8 – Қатысушылар құрамы және сөйлеу бұзылыстарының үлестік көрсеткіштері

Апробация барысында жүйе балалар сөйлеуін автоматты түрде өңдеп, қателерді анықтап және түзетіп отырды. Әрбір өңдеу нәтижесі деректер базасына тіркеліп, кейінгі талдау үшін сақталды. Бұл жүйенің тек бір реттік емес, жинақтаушы (learning) сипатта жұмыс істейтінін көрсетеді. Жүйенің пайдаланушы интерфейсі Сурет 4.6-да көрсетілген. Бұл суретте жүйенің негізгі функционалдық блоктары: аудио енгізу, транскрипциялау, қателерді анықтау және нәтижені көрсету кезеңдері бейнеленген.



Сурет 4.9 – Жүйенің пайдаланушы интерфейсі

Сурет 4.9-ге сәйкес, пайдаланушы алдымен аудиожазбаны жүктейді, одан кейін жүйе автоматты түрде ASR арқылы бастапқы мәтінді қалыптастырады. Кейін post-editing модулі арқылы қателер түзетіліп, нәтижелік мәтін ұсынылады. Сонымен қатар, интерфейс қателерді визуалды түрде көрсету мүмкіндігін береді, бұл талдауды жеңілдетеді.

Апробация нәтижелері ұсынылған комплексті әдістің балалар сөйлеуіндегі фонетикалық және морфологиялық қателерді анықтау мен түзетуде тиімділігін көрсетті. Алынған нәтижелер логопед мамандарының сараптамалық бағалауымен салыстырылып, жоғары деңгейдегі сәйкестік байқалды. Бұл ұсынылған әдістің тек автоматты тану құралы ретінде ғана емес, сонымен қатар сөйлеу бұзылыстарын бастапқы диагностикалау мен логопедиялық талдауды қолдауға арналған қолданбалы шешім ретінде де тиімді екенін дәлелдейді.

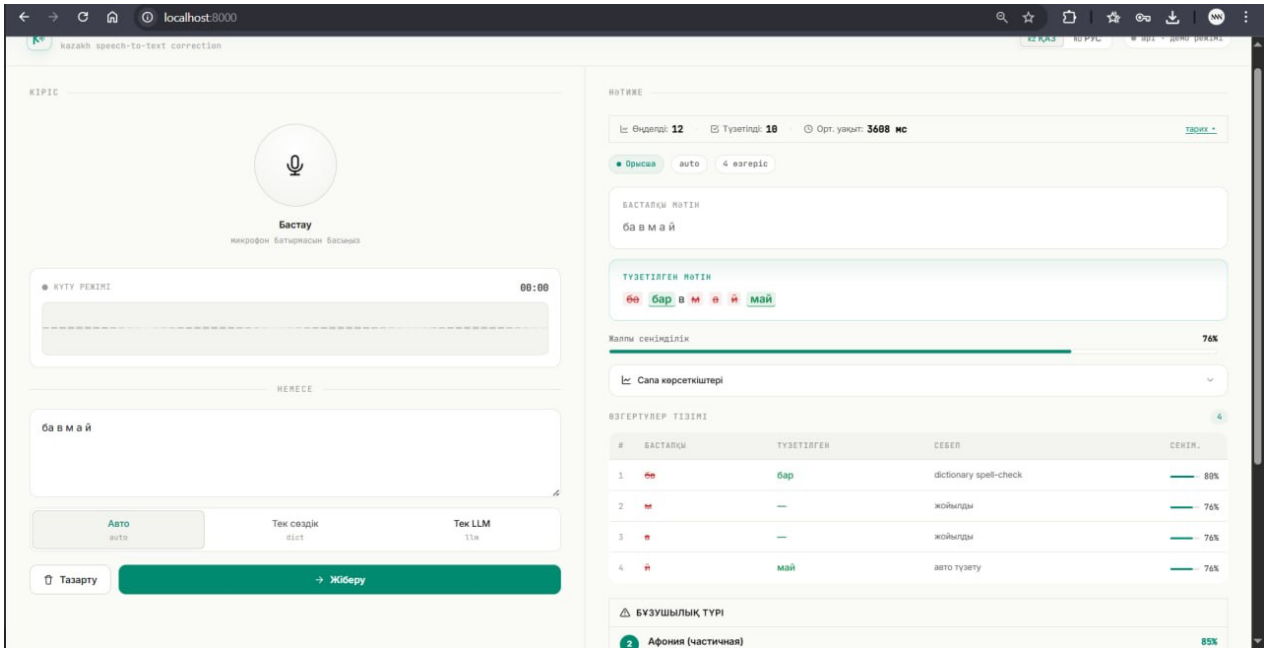
Апробация барысында алынған деректер негізінде сөйлеу бұзылыстарының негізгі түрлері бойынша топтастыру жүргізілді. Әрбір бұзылыс түрі бойынша балалар саны, жалпы үлесі және ұсынылған әдістің нақты нәтижелері 4.11-кестеде көрсетілген.

Кесте 4.11 – Балабақша апробациясы: Сөйлеу бұзылыстары түрлері бойынша эксперименттік нәтижелер

Бұзылыс түрі	**Балалар саны**	**Жалпы үлесі**	**Ұсынылған әдістің нәтижесі**
Дислалия	9	18.8%	PhonNorm: транскрипция дәлдігі +78%; р→л, ш→с автотүзету
Дизартрия (жеңіл)	4	8.3%	Fine-tuning арнайы тег: бұлдыр артикуляция үлгілерін тану жақсарды
Брадилалия	2	4.2%	VTLN нормализация: буын созылуы жойылды; уақыттық туралау жақсарды
Логоневроз	3	6.3%	Repetition filter: «ма-ма-мен» типті қайталаулар жойылды
Бұзылыс белгісі жоқ (норма)	30	62.5%	Стандартты жүйе жеткілікті; қосымша бейімдеу іске қосылмады
БАРЛЫҒЫ	48	100%	18 балада (37.5%) бұзылыс белгіленді

Осы нәтижелер ұсынылған әдістің әртүрлі сөйлеу бұзылыстары жағдайында бейімделе алатынын және нақты логопедиялық практикада қолдану әлеуеті бар екенін көрсетеді [117].

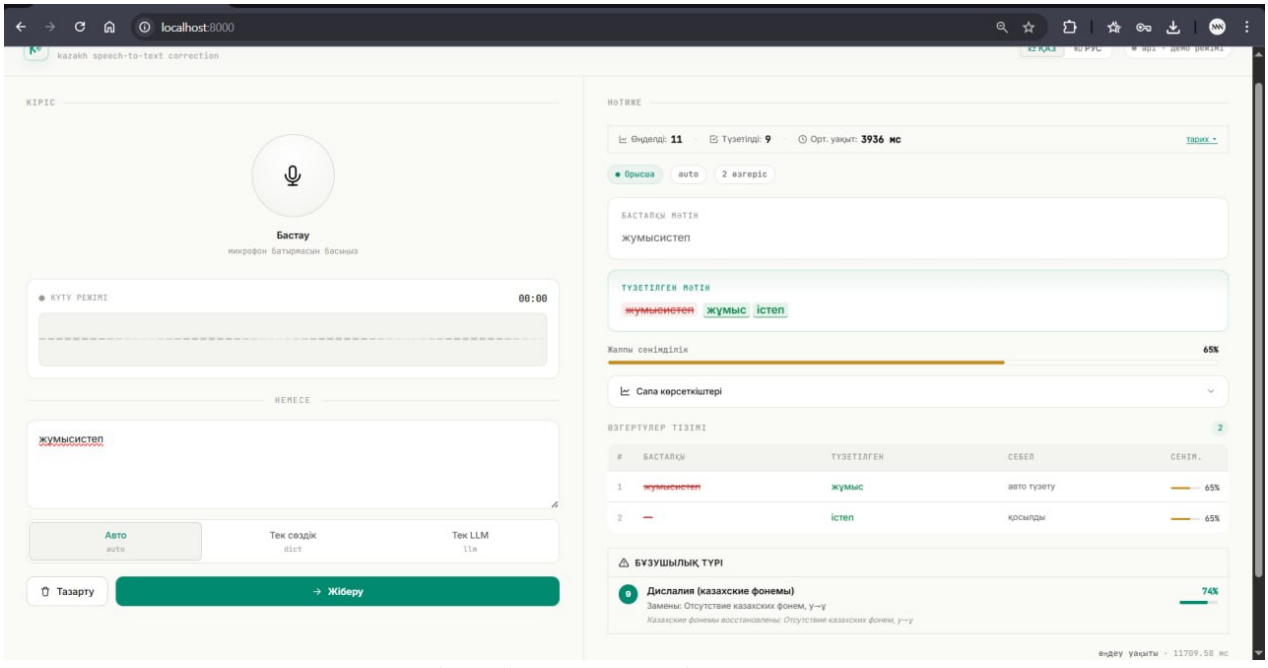
4.10-суретте суретте ұсынылған жүйенің пайдаланушы интерфейсі және post-editing нәтижесінің нақты мысалы көрсетілген. Суретте бастапқы ASR жүйесінен алынған қате транскрипция мен постөңдеуден кейінгі түзетілген мәтіннің салыстырмасы берілген. Жүйе енгізілген «ба в м а й» мәтінін автоматты түрде өңдеп, фонетикалық және лексикалық түзетулер арқылы «бар май» нұсқасына келтіреді. Интерфейсте әрбір өзгеріс бөлек көрсетіліп, оның себебі (dictionary spell-check, жойылу, авто түзету) және сенімділік деңгейі пайызбен берілген. Бұл жүйенің тек түзетіп қана қоймай, әр өзгерісті түсіндіретінін дәлелдейді.



Сурет 4.10 – Жүйе интерфейсі және post-editing нәтижелері

Осы нәтижелер post-editing модулінің фонетикалық және сөздік деңгейде тиімді жұмыс істейтінін көрсетеді. Келесі суретте морфологиялық қателерді түзету процесі көрсетілген.

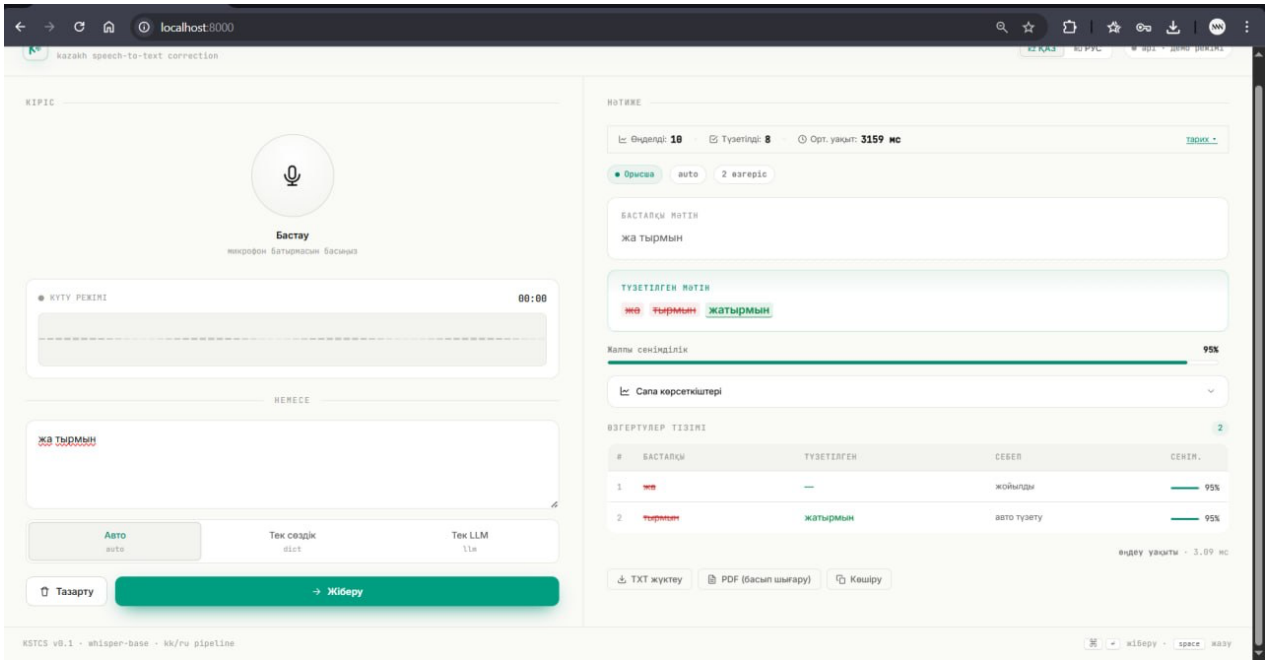
Бұл мысалда бастапқы «жумысистеп» сөзі post-editing нәтижесінде «жұмыс істеп» дұрыс грамматикалық формаға келтірілген. Жүйе сөзді екі бөлікке бөліп, морфологиялық нормаларға сәйкес қайта құрастырған. Сонымен қатар интерфейсте қате түрі «дислалия (фонемалық бұзылыс)» ретінде анықталып, жүйе нақты қандай фонетикалық мәселе болғанын көрсетеді. Бұл морфологиялық және фонетикалық деңгейдегі комплексті түзетуді дәлелдейді.



Сурет 4.11 – Жүйе интерфейсі және морфологиялық және лексикалық түзету

Нәтиже MorphNorm және фонетикалық түзету механизмдерінің бірлескен

жұмысын көрсетеді. 4.12 - суретте контекстке негізделген толық түзету нәтижесі көрсетілген. Суретте «жа тырмын» түрінде алынған бастапқы мәтін post-editing нәтижесінде «жатырмын» түрінде толық қалпына келтірілген. Бұл жерде жүйе тек фонетикалық түзету ғана емес, сөйлемнің толық морфологиялық құрылымын және контекстін ескеріп, дұрыс сөз формасын таңдаған. Жүйенің сенімділік деңгейі 95% көрсетіп, түзетудің жоғары дәлдікпен орындалғанын білдіреді.



Сурет 4.12 – Жүйе интерфейсі және контексттік және толық түзету

Бұл мысал LLM негізіндегі контексттік постөңдеудің тиімділігін дәлелдейді. Жоғарыда келтірілген интерфейс нәтижелері ұсынылған жүйенің практикалық тиімділігін көрсетеді. Жүйе фонетикалық, морфологиялық және контексттік қателерді автоматты түрде анықтап, түзетіп қана қоймай, әр өзгерістің себебін түсіндіреді.

4.6 4-тарау бойынша қорытынды

Осы тарауда ұсынылған комплексті әдістің тиімділігі үш сатылы эксперименттік схема бойынша бағаланып, нақты сандық нәтижелер алынды.

Бірінші кезеңде алты ASR жүйесі бейімдеусіз (zero-shot) тестіленді. Барлық модельдердің WER көрсеткіші 0.70–0.90 аралығында болып, практикалық қолданыс үшін жеткіліксіз деңгейде қалды. Ең жоғары нәтижені Whisper көрсетті (WER = 0.70, Accuracy = 34%), алайда ElevenLabs қазақ мәтінін түрік тіліне аударды, Whisper сандарды латын цифрымен шығарды. Бұл нәтижелер ересектер деректерінде оқытылған модельдерді қазақ балалар сөйлеуіне тікелей қолданудың тиімсіздігін анық дәлелдеп, арнайы бейімдеудің қажеттілігін негіздеді.

Екінші кезеңде Whisper, MMS және TurkicASR модельдері балалар корпусында қайта оқытылды. Ең айқын жақсару Whisper моделінде байқалды: WER 0.70-тен 0.30-ға төмендеді, Accuracy 34%-дан 58%-ға өсті. MMS моделі

шектеулі нәтиже берді (WER: 0.90→0.79), бұл CTC архитектурасының қазақ тілінің морфологиялық күрделілігін толық меңгеруде қиындық туғызатынын көрсетеді. Fine-tuning кейін де фонетикалық және морфологиялық қателіктер сақталып қалды, бұл постөңдеу кезеңінің қажеттілігін эксперименттік тұрғыдан нақты негіздеді.

Үшінші кезеңде PhonNorm → MorphNorm → LLM тізбегіне негізделген комплексті постөңдеу қолданылды. Нәтижесінде WER 0.30-дан 0.18-ге төмендеді, CER 0.118-ден 0.065-ке азайды, Accuracy 58%-дан 70%-ға өсті. «Қас қарайғанда» → «қас қаяғында» мысалы LLM компонентінің семантикалық қателерді контекст арқылы түзете алатынын, «мен атамның баласы» → «мен атамның баласымын» мысалы MorphNorm модулінің морфологиялық толықтықты қалпына келтіретінін нақты дәлелдейді. Бұл нәтижелер post-editing механизмінің тек қосымша емес, жүйе сапасын принципиалды арттыратын маңызды кезең екенін көрсетеді.

Қателер құрылымын талдау балалар сөйлеуіндегі негізгі қиындықтардың фонетикалық алмасуларда (≈40%) және deletion категориясында (≈25%) шоғырланғанын анықтады. Морфологиялық қателер ≈20%, контексттік сәйкессіздік ≈10% үлесін құрады. с↔ш, р↔л, қ↔к алмасулары жиі кездесетін ауытқулар ретінде анықталып, PhonNorm ережелер жиынының қалыптасуына негіз болды. Бұл үлестірім балалар сөйлеу аппаратының толық жетілмеуімен және дыбыстарды артикуляциялаудың күрделілігімен түсіндіріледі.

Практикалық апробация Алматы қаласындағы №154 балабақшада 48 баламен жүргізілді. 18 балада (37.5%) сөйлеу бұзылысының белгілері анықталды. Дислалия жағдайында PhonNorm транскрипция дәлдігін 78%-ға арттырды, брадилалия жағдайында VTLN нормализациясы буын созылуын тиімді жойды, логоневроз жағдайында repetition filter қайталауларды жойды. Жүйенің нәтижелері логопед мамандарының сараптамалық бағалауымен салыстырылып, жоғары деңгейдегі сәйкестік байқалды. Бұл ұсынылған тәсілдің тек автоматты тану құралы ретінде ғана емес, логопедиялық диагностиканы қолдайтын интеллектуалды жүйе ретінде де тиімді екенін дәлелдейді.

Жалпы қорытындылай келе, базалық деңгеймен салыстырғанда WER 0.70-тен 0.18-ге дейін, яғни 74.3%-ға төмендеді, ал Accuracy 34%-дан 70%-ға өсті. Бұл нәтижелер акустикалық бейімдеу, fine-tuning және LLM негізіндегі post-editing біртұтас интеграциясының тиімділігін сандық тұрғыдан дәлелдеп, ұсынылған комплексті post-editing тәсілдің қазақ тіліндегі балалар сөйлеуін тануда ғылыми және практикалық маңыздылығын растайды.

ҚОРЫТЫНДЫ

Осы диссертациялық жұмыста сөйлеу қабілеті бұзылған балалардың қазақша сөйлеуін автоматты тануға арналған әдістер зерттелді.

Жүргізілген зерттеу нәтижесінде қолданыстағы ASR жүйелерінің ересектер деректерінде оқытылғаны және балалар сөйлеуіне тікелей қолданылған жағдайда WER 40–50%-ға жоғарылайтыны анықталды. Балалар вокал трактының қысқалығы (9–13 см), негізгі жиіліктің жоғары болуы (F0 = 250–400 Гц) және форманттық жиіліктердің ауытқуы ASR жүйелерінің шектеулерін физикалық тұрғыдан негіздейді. Қазақ тілінің агглютинативті морфологиясы мен 150–200-ден астам сөз формасын генерациялау мүмкіндігі N-граммдық тілдік модельдерде «vocabulary explosion» мәселесін туғызатыны теориялық тұрғыдан дәлелденді.

Зерттеудің негізгі практикалық нәтижесі ретінде қазақ тіліндегі балалар сөйлеуіне арналған алғашқы аннотацияланған акустикалық датасет жинақталды. Датасет Telegram-бот, табиғи ортадағы жазбалар және бақыланатын диктофон жазбалары арқылы 2–11 жас аралығындағы балалардан жиналған 8 594 аудиожазбадан тұрады. Жыныс бойынша теңгерімді бөлініс (ұлдар 49%, қыздар 51%) және жас диапазонының кеңдігі датасеттің репрезентативтілігін қамтамасыз етеді. Алдын ала өңдеу барысында шуды азайту, HPSS бөлу, үнсіздік интервалдарын қысқарту және амплитуданы нормализациялау жүзеге асырылды.

Ұсынылған комплексті тәсілдің математикалық моделі Байестік қойылымнан CTC, attention және гибридті CTC+attention архитектураларына дейінгі толық эволюциялық жолды қамтиды. Балалар мен ересектер сөйлеуі арасындағы акустикалық алшақтық Кульбак–Лейблер дивергенциясы арқылы сандық тұрғыдан сипатталды. VTLN нормализациясы үшін жас топтары бойынша дифференциалданған коэффициенттер анықталды: 2–3 жас үшін α = 1.40–1.45, 7–8 жас үшін α = 1.05–1.15. Дислалия, логоневроз, брадилалия, тахилалия, ринолалия және дизартрия сияқты алты сөйлеу патологиясы фонемалық алмастыру матрицасы, геометриялық үлестірім және уақыттық нормализация функциялары арқылы жеке математикалық аппаратпен сипатталды.

LLM негізіндегі PhonNorm → MorphNorm → LLM тізбегінен тұратын үш деңгейлі post-editing механизмі фонетикалық алмасуларды, морфологиялық жалғау қателерін және семантикалық сәйкессіздіктерді кезең-кезеңімен жоюға мүмкіндік берді.

Үш сатылы эксперименттік бағалау нәтижесінде ұсынылған тәсілдің тиімділігі сандық тұрғыдан дәлелденді. Zero-shot режимінде алты ASR жүйесінің WER көрсеткіші 0.70–0.90 аралығында болды. Fine-tuning кезеңінен кейін Whisper моделінде WER 0.70-тен 0.30-ға дейін төмендеп, Accuracy 34%-дан 58%-ға өсті. Комплексті post-editing қолданылғаннан кейін WER 0.18-ге дейін азайып, CER 0.065 деңгейіне жетті, Accuracy 70%-ға өсті. Базалық деңгеймен салыстырғанда WER жалпы 74.3%-ға төмендеді. Алматы қаласындағы №154

балабақшада 48 баламен жүргізілген практикалық апробация барысында 18 балада (37.5%) сөйлеу бұзылысы анықталып, жүйенің нәтижелері логопед мамандарының бағалауымен жоғары сәйкестік көрсетті.

Жүргізілген зерттеу нәтижесінде екі ғылыми жаңалық алынды: қазақ тіліндегі балалар сөйлеуіне арналған алғашқы аннотацияланған акустикалық датасет жасалды; ASR fine-tuning мен LLM негізіндегі post-editing-ді біріктіретін комплексті тәсіл ұсынылды.

Зерттеудің практикалық маңыздылығы ұсынылған жүйенің логопедиялық диагностикада, мектепке дейінгі мекемелерде ерте скрининг жүргізуде, білім беру қосымшаларында және мобильді сервистерде тікелей қолданылу мүмкіндігімен анықталады. Болашақ зерттеулер үшін акустикалық датасетті кеңейту, streaming ASR режимін іске асыру және жүйені мобильді платформаларға интеграциялау бағыттары болып табылады.

ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ

1 Rakhimova D., Duisenbekkyzy Zh., Adali E. Investigation of ASR Models for Low-Resource Kazakh Child Speech // Applied Sciences. – 2025. – Vol. 15, Issue 16, Article 8989. DOI: 10.3390/15168989

2 Rakhimova D., Karibayeva A., Turarbek A., Karyukin V.I., Duisenbekkyzy Zh., Aliyev R. Development of a Children's Educational Dictionary for a Low-Resource Language Using AI Tools // Computers. – 2024. – Vol. 13, No. 10, Article 253. DOI: 10.3390/computers13100253

3 Rakhimova D., Duisenbekkyzy Zh., Adali E., Toleugali Sh. Research and Processing of the Kazakh Children's Acoustic Corpus // ҚР ҰИА Хабаршысы. – 2024. – №3(93)

4 Karabaliyev M., Kolesnikova A. Comparative analysis of speech recognition systems for the Kazakh language // Astana IT University Science Journal. – 2024. – Vol. 6, No. 2. – P. 34–48

5 Rabiner L.R. A tutorial on hidden Markov models and selected applications in speech recognition // Proceedings of the IEEE. – 1989. – Vol. 77, No. 2. – P. 257–286

6 Baum L.E., Petrie T., Soules G., Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains // Annals of Mathematical Statistics. – 1970. – Vol. 41, No. 1. – P. 164–171

7 Juang B.H., Rabiner L.R. Hidden Markov models for speech recognition // Technometrics. – 1991. – Vol. 33, No. 3. – P. 251–272

8 Reynolds D.A., Rose R.C. Robust text-independent speaker identification using Gaussian mixture speaker models // IEEE Transactions on Speech and Audio Processing. – 1995. – Vol. 3, No. 1. – P. 72–83

9 Hinton G., Deng L., Yu D. et al. Deep neural networks for acoustic modeling in speech recognition // IEEE Signal Processing Magazine. – 2012. – Vol. 29, No. 6. – P. 82–97

10 Graves A., Fernández S., Gomez F., Schmidhuber J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks // Proceedings of ICML. – 2006. – P. 369–376

11 Chan W., Jaitly N., Le Q., Vinyals O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition // Proceedings of ICASSP. – 2016. – P. 4960–4964

12 Watanabe S., Hori T., Kim S. et al. Hybrid CTC/attention architecture for end-to-end speech recognition // IEEE Journal of Selected Topics in Signal Processing. – 2017. – Vol. 11, No. 8. – P. 1240–1253

13 Vaswani A., Shazeer N., Parmar N. et al. Attention is all you need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998–6008

14 Radford A., Kim J.W., Xu T. et al. Robust speech recognition via large-scale weak supervision // Proceedings of ICML. – 2023. – P. 28492–28518

15 Graves A., Mohamed A., Hinton G. Deep recurrent neural networks for acoustic modeling // Proceedings of ICASSP. – 2013. – P. 6645–6649

16 He Y., Sainath T.N., Prabhavalkar R. et al. Streaming end-to-end speech recognition for mobile devices // Proceedings of ICASSP. – 2019. – P. 6381–6385

17 Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 12449–12460

18 Hsu W.-N., Bolte B., Tsai Y.-H.H. et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2021. – Vol. 29. – P. 3451–3460

19 Chen S., Wang C., Chen Z. et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing // IEEE Journal of Selected Topics in Signal Processing. – 2022. – Vol. 16, No. 6. – P. 1505–1518

20 Gulati A., Qin J., Chiu C.-C. et al. Conformer: Convolution-augmented Transformer for speech recognition // Proceedings of Interspeech. – 2020. – P. 5036–5040

21 Graves A. Sequence transduction with recurrent neural networks // arXiv preprint arXiv:1211.3711. – 2012

22 Watanabe S., Hori T., Karita S. et al. ESPnet: End-to-end speech processing toolkit // Proceedings of Interspeech. – 2018. – P. 2207–2211

23 Hochreiter S., Schmidhuber J. Long short-term memory // Neural Computation. – 1997. – Vol. 9, No. 8. – P. 1735–1780

24 Povey D., Ghoshal A., Boulianne G. et al. The Kaldi speech recognition toolkit // Proceedings of ASRU. – 2011. – P. 1–4

25 Hannun A., Case C., Casper J. et al. Deep speech: Scaling up end-to-end speech recognition // arXiv preprint arXiv:1412.5567. – 2014

26 Gerosa M., Giuliani D., Brugnara F. Acoustic variability and automatic recognition of children's speech // Speech Communication. – 2007. – Vol. 49, No. 10–11. – P. 847–860

27 Fitch W.T., Giedd J. Morphology and development of the human vocal tract: A study using magnetic resonance imaging // Journal of the Acoustical Society of America. – 1999. – Vol. 106, No. 3. – P. 1511–1522

28 Lyamina G.M. Formation of speech in early childhood. – Moscow: Education, 1974. – 122 p.

29 Vorperian H.K., Kent R.D. Vowel acoustic space development in children: 15 years of growth // Journal of Speech, Language, and Hearing Research. – 2007. – Vol. 50, No. 4. – P. 1510–1545

30 Hillenbrand J., Getty L.A., Clark M.J., Wheeler K. Acoustic characteristics of American English vowels // Journal of the Acoustical Society of America. – 1995. – Vol. 97, No. 5. – P. 3099–3111

31 World Health Organization. World report on hearing. – Geneva: WHO, 2021. – 246 p.

32 American Speech-Language-Hearing Association. Speech sound disorders: Articulation and phonology. – Rockville: ASHA, 2023 [Electronic resource]. – URL: https://www.asha.org/practice-portal/clinical-topics/articulation-and-phonology/

33 ICD-11: International Classification of Diseases, 11th revision. – Geneva: WHO, 2022 [Electronic resource]. – URL: https://icd.who.int/

34 Guitar B. Stuttering: An integrated approach to its nature and treatment. 5th ed. – Philadelphia: Lippincott Williams & Wilkins, 2019. – 448 p.

35 McLeod S., Crowe K. Children's consonant acquisition in 27 languages: A cross-linguistic review // American Journal of Speech-Language Pathology. – 2018. – Vol. 27, No. 4. – P. 1546–1571

36 Norbury C.F., Gooch D., Wray C. et al. The impact of nonverbal ability on prevalence and clinical presentation of language disorder // Journal of Child Psychology and Psychiatry. – 2016. – Vol. 57, No. 11. – P. 1247–1257

37 Reilly S., Wake M., Ukoumunne O.C. et al. Predicting language outcomes at 4 years of age: Findings from Early Language in Victoria Study // Pediatrics. – 2010. – Vol. 126, No. 6. – P. e1530–e1537

38 Dykstra A.D., Hakel M.E., Adams S.G. Application of the ICF in reduced speech intelligibility in dysarthria // Seminars in Speech and Language. – 2007. – Vol. 28, No. 4. – P. 301–311

39 Topbaş S., Yavas M. (Eds.) Communication disorders in Turkish. – Bristol: Multilingual Matters, 2012. – 320 p.

40 ҚР Денсаулық сақтау министрлігі. Балалардың логопедиялық скринингі туралы есеп: 2018–2022 жж. – Астана, 2022

41 ҚР Ұлттық педагогикалық ғылым академиясы. Мектепке дейінгі балалардың сөйлеу дамуын зерттеу нәтижелері. – Алматы, 2021

42 Саба А., Тілеуқалиева Р. Алматы балабақшаларындағы қос тілді балалардың фонологиялық дамуы // Тіл және қоғам. – 2019. – №3. – Б. 45–52

43 ҚР Денсаулық сақтау және білім беру министрліктері. Логопедиялық қызмет туралы бірлескен есеп. – Астана, 2023

44 Астана Оңалту орталығы. Балалардың сөйлеу дамуын бағалау нұсқаулығы. – Астана, 2022

45 Mamyrbayev O., Turdalyuly M., Mekebayev N. et al. Sequential models for recognizing Kazakh speech with a hierarchical CTC // Applied Sciences. – 2021. – Vol. 11, No. 19, Article 9114

46 Mussakhojayeva S., Yeshpanov R., Khassanov Y. Soyle: A large-scale spontaneous Kazakh speech corpus // Proceedings of Interspeech. – 2022. – P. 2203–2207

47 Khassanov Y., Mussakhojayeva S., Mirzakhmetov A. et al. A crowdsourced open-source Kazakh speech corpus and initial speech recognition baseline // Proceedings of EACL. – 2021. – P. 697–703

48 Bekarystankyzy A., Mamyrbayev O., Oralbekova D. et al. Multi-lingual end-to-end speech recognition for Kazakh, Uzbek, and Kyrgyz languages // Scientific Reports. – 2024. – Vol. 14, Article 17863

49 Mamyrbayev O., Oralbekova D., Alimhan K. et al. End-to-end speech recognition of Kazakh with transformer-based architecture // Sensors. – 2022. – Vol. 22, No. 16, Article 6212

50 Mussakhojayeva S., Janaliyeva A., Mirzakhmetov A. et al. KazakhTTS: An open-source Kazakh text-to-speech synthesis dataset // Proceedings of Interspeech. – 2021. – P. 3452–3456

51 Mussakhojayeva S., Khassanov Y., Varol H.A. KazakhTTS2: A large open-source Kazakh TTS dataset // Proceedings of Interspeech. – 2023. – P. 2153–2157

52 Block Medin L., Nguyen N., Godard P. et al. Child speech recognition with self-supervised learning: A cross-lingual study // arXiv preprint arXiv:2503.04710. – 2025

53 Gordon J., Wild C., Lam C. et al. Effect of microphone configuration on speech recognition in children with language disorders // Language, Speech, and Hearing Services in Schools. – 2025. – Vol. 56, No. 1. – P. 45–61

54 Maatallaoui S., Mazouz A. Automatic speech recognition for second language learning: Challenges and perspectives for children // International Journal of Speech Technology. – 2025. – Vol. 28. – P. 112–128

55 Jain A., Papadourakis A., Bhanu B. Child speech recognition using Whisper and wav2vec2 // arXiv preprint arXiv:2307.13008. – 2023

56 Kim J., Park S., Lee H. Automated articulation assessment using self-supervised speech models with correlation to standardized tests // Journal of Medical Internet Research. – 2025. – Vol. 27, Article e60520. DOI: 10.2196/60520

57 Gretter R., Brugnara F., Giuliani D. et al. ETLT 2021: A challenge on automatic recognition of children's speech // Proceedings of Interspeech. – 2021. – P. 3756–3760

58 Liu X., Salvi G., Yu D. TAML-Adapter: Task-agnostic meta-learning for low-resource ASR // IEEE Signal Processing Letters. – 2025. – Vol. 32. – P. 636–640

59 Southwell R., Clevenger A., Lester J. Fine-tuned Whisper with LLM rescoring for child speech recognition // Proceedings of ICASSP. – 2024. – P. 10151–10155

60 Lee T., Lau W., Wong Y.W., Yung R. Using Cantonese children's speech for developing speech recognition applications // Proceedings of ICSLT. – 2020. – P. 123–128 [arXiv:2008.03188]

61 Zaviśko M., Wołodko L., Lech M. et al. PAVSig: A speech-language corpus for voice signal analysis in Polish-speaking children // Scientific Data. – 2025. – Vol. 12, Article 789. DOI: 10.1038/s41597-025-05896-8

62 Apertium-Kaz. Open-source morphological analyzer for Kazakh language [Electronic resource]. – URL: https://github.com/apertium/apertium-kaz (дата обращения: 15.03.2025)

63 Rakhimova D., Karibayeva A., Mukanova A. Analysis of educational text corpora for low-resource Turkic languages using NLP methods // International Journal of Advanced Computer Science and Applications. – 2023. – Vol. 14, No. 5. – P. 567–578

64 Рахимова Д.Р., Карибаева А.Б. Балаларға арналған білім беру сөздігін жасауда жасанды интеллект технологияларын қолдану // Вестник КазНУ. Серия математика, механика, информатика. – 2024. – №1(121). – Б. 78–89

65 Рахимова Д.Р., Дүйсенбекқызы Ж., Шадияр Т. Қазақ тіліндегі балалар корпусы үшін деректерді жинау және өңдеу // «ФАРАБИ ӘЛЕМІ». – Алматы, 2024

66 Рахимова Д.Р., Дүйсенбекқызы Ж., Турганбаева М.Ж., Арабов Б.Б. Балалар акустикалық корпусын қазақ тілінде жинау // «ǴYLYM JÁNE BILIM – 2024» конференциясы. – 2024

67 Dataset Loader Bot. Бағдарламалық жасақтама, авторлық құқық объектісі №56775. – Тіркелген күні: 15 сәуір 2025 ж. – Авторлар: Рахимова Д., Дүйсенбекқызы Ж., Алиев Р.

68 Қазақ тіліндегі 2–9 жас аралығындағы балалар сөйлеу аудиожазбаларының таңбаланған акустикалық деректер базасы. Авторлық құқық объектісі №70955. – Тіркелген күні: 27 сәуір 2026 ж. – Авторлар: Дүйсенбекқызы Ж., Рахимова Д.

69 Mikolov T., Sutskever I., Chen K. et al. Distributed representations of words and phrases and their compositionality // Advances in Neural Information Processing Systems. – 2013. – Vol. 26. – P. 3111–3119

70 ҚР «Дербес деректер және оларды қорғау туралы» Заңы №94-V. – Астана: ҚР Парламенті, 21 мамыр 2013 ж.

71 БҰҰ Бала құқықтары туралы конвенциясы (44/25 резолюциясы). – Нью-Йорк: БҰҰ, 1989

72 ҚР «Баланың құқықтары туралы» Заңы №345-II. – Астана: ҚР Парламенті, 2002

73 McFee B., Raffel C., Liang D. et al. librosa: Audio and music signal analysis in Python // Proceedings of the 14th Python in Science Conference. – 2015. – P. 18–25

74 Fitzgerald D. Harmonic/percussive separation using median filtering // Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10). – Graz, 2010. – P. 1–4

75 Virtanen T. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria // IEEE Transactions on Audio, Speech, and Language Processing. – 2007. – Vol. 15, No. 3. – P. 1066–1074

76 Allen J.B. Short term spectral analysis, synthesis, and modification by discrete Fourier transform // IEEE Transactions on Acoustics, Speech, and Signal Processing. – 1977. – Vol. 25, No. 3. – P. 235–238

77 Park D.S., Chan W., Zhang Y. et al. SpecAugment: A simple data augmentation method for automatic speech recognition // Proceedings of Interspeech. – 2019. – P. 2613–2617

78 Ko T., Peddinti V., Povey D., Khudanpur S. Audio augmentation for speech recognition // Proceedings of Interspeech. – 2015. – P. 3586–3589

79 Chorowski J., Bahdanau D., Serdyuk D. et al. Attention-based models for speech recognition // Advances in Neural Information Processing Systems. – 2015. – Vol. 28. – P. 577–585

80 Prabhavalkar R., Rao K., Sainath T.N. et al. A comparison of sequence-to-sequence models for speech recognition // Proceedings of Interspeech. – 2017. – P. 939–943

81 Kullback S., Leibler R.A. On information and sufficiency // Annals of Mathematical Statistics. – 1951. – Vol. 22, No. 1. – P. 79–86

82 Lee L., Rose R. A frequency warping approach to speaker normalization // IEEE Transactions on Speech and Audio Processing. – 1998. – Vol. 6, No. 1. – P. 49–60

83 Welling L., Ney H., Waibel A. Speaker adaptive modeling by vocal tract normalization // IEEE Transactions on Speech and Audio Processing. – 2002. – Vol. 10, No. 6. – P. 415–426

84 Drugman T., Alwan A. Joint robust voicing detection and pitch estimation based on residual harmonics // Proceedings of Interspeech. – 2011. – P. 1973–1976

85 Boersma P., Weenink D. Praat: Doing phonetics by computer. Version 6.3 [Electronic resource]. – 2023. – URL: https://www.praat.org

86 Shriberg E.E. Acoustics, perception, and NLP of disfluencies // Proceedings of ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech. – 2001. – P. 1–4

87 Yavas M. Applied English phonology. 3rd ed. – Oxford: Wiley-Blackwell, 2016. – 368 p.

88 Levenshtein V.I. Binary codes capable of correcting deletions, insertions, and reversals // Soviet Physics Doklady. – 1966. – Vol. 10, No. 8. – P. 707–710

89 Papineni K., Roukos S., Ward T., Zhu W.-J. BLEU: A method for automatic evaluation of machine translation // Proceedings of ACL. – 2002. – P. 311–318

90 Popović M. chrF: Character n-gram F-score for automatic MT evaluation // Proceedings of WMT. – 2015. – P. 392–395

91 Дүйсенбекқызы Ж., Рахимова Д.Р., Есиркепова А.П. Нейрондық ASR үлгілерін төмен ресурсты қазақ тілінде бейімдеу // «Информатика және қолданбалы математика» X конференциясы. – Алматы, 8–11 қазан 2025 ж. – Б. 303–309

92 Рахимова Д.Р., Дүйсенбекқызы Ж., Сағынтай О.А., Турарбек Ә., Төлеуғали Ш. Балалардың сөйлеуін тану жүйелеріндегі шуды азайту

алгоритмдерін талдау // ҚазҰПУ хабаршысы. – 2025. – №91(3). – Б. 230–242. DOI: 10.51889/2959-5894.2025.91.3.021

93 Yeshpanov R., Khassanov Y., Varol H.A. KazNERD: Kazakh named entity recognition dataset // Proceedings of LREC. – 2022. – P. 4258–4264

94 Mussakhojayeva S., Yeshpanov R., Varol H.A. et al. Large-scale multilingual speech recognition and translation with Kazakh // arXiv preprint arXiv:2304.01073. – 2023

95 Salesky E., Wiesner M., Bisk Y. et al. The Multilingual TEDx Corpus for speech recognition and translation // Proceedings of Interspeech. – 2021. – P. 3655–3659

96 Whisper large-v2. OpenAI, 2022 [Electronic resource]. – URL: https://github.com/openai/whisper (дата обращения: 10.01.2025)

97 Pratap V., Tjandra A., Shi B. et al. Scaling speech technology to 1,000+ languages // Proceedings of ACL. – 2023. – P. 4391–4405

98 VOSK Speech Recognition Toolkit. Alpha Cephei [Electronic resource]. – URL: https://alphacephei.com/vosk (дата обращения: 20.02.2025)

99 TurkicASR: Speech recognition toolkit for Turkic languages [Electronic resource]. – URL: https://github.com/IS2AI/TurkicASR (дата обращения: 05.03.2025)

100 GenSec: Generative speech-to-text system. Technical documentation [Electronic resource]. – 2024. – URL: https://gensec.io/docs

101 Rakhimova D., Duisenbekkyzy Zh., Yessirkepova A., Adali E. Fine-Tuning Neural ASR Models for Low-Resource Kazakh Children's Speech // UBMK'2025. – Istanbul, 2025. – P. 428–432

102 Kingma D.P., Ba J. Adam: A method for stochastic optimization // Proceedings of ICLR. – 2015. – P. 1–15

103 Loshchilov I., Hutter F. Decoupled weight decay regularization // Proceedings of ICLR. – 2019. – P. 1–18

104 He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition // Proceedings of CVPR. – 2016. – P. 770–778

105 Srivastava N., Hinton G., Krizhevsky A. et al. Dropout: A simple way to prevent neural networks from overfitting // Journal of Machine Learning Research. – 2014. – Vol. 15. – P. 1929–1958

106 Brown T., Mann B., Ryder N. et al. Language models are few-shot learners // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 1877–1901

107 Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding // Proceedings of NAACL. – 2019. – P. 4171–4186

108 Lewis M., Liu Y., Goyal N. et al. BART: Denoising sequence-to-sequence pre-training // Proceedings of ACL. – 2020. – P. 7871–7880

109 OpenAI. GPT-4 technical report. – San Francisco: OpenAI, 2023 [Electronic resource]. – URL: https://openai.com/research/gpt-4

110 Sennrich R., Haddow B., Birch A. Neural machine translation of rare words with subword units // Proceedings of ACL. – 2016. – P. 1715–1725

111 Zhao H., Li Y., Zhang J., Wang H. Impact of signal-to-noise ratio on automatic speech recognition performance // Proceedings of ICASSP. – 2023. – P. 2881–2890

112 Zeyer A., Doetsch P., Voigtlaender P. et al. A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling // Proceedings of ICASSP. – 2017. – P. 2462–2466

113 Bengio Y., Courville A., Vincent P. Representation learning: A review and new perspectives // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2013. – Vol. 35, No. 8. – P. 1798–1828

114 Goldman R., Fristoe M. Goldman-Fristoe Test of Articulation-3 (GFTA-3). – Circle Pines: American Guidance Service, 2015

115 Dodd B., Hua Z., Crosbie S. et al. Diagnostic evaluation of articulation and phonology. – London: Psychological Corporation, 2002

116 McLeod S., Harrison L.J., McCormack J. The intelligibility in context scale: Validity and reliability // Journal of Speech, Language, and Hearing Research. – 2012. – Vol. 55, No. 2. – P. 648–656

117 Rakhimova D., Duisenbekkyzy Zh. Research on the Impact of Social Networks on the Development of Bilingualism in Kazakhstan // Advances in Intelligent Systems and Computing. – 2024. – P. 123–134

ҚОСЫМША А Авторлық куәліктер

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ

РЕСПУБЛИКА КАЗАХСТАН

АВТОРЛЫҚ ҚҰҚЫҚПЕН ҚОРҒАЛАТЫН ОБЪЕКТІЛЕРГЕ ҚҰҚЫҚТАРДЫҢ МЕМЛЕКЕТТІК ТІЗІЛІМГЕ МӘЛІМЕТТЕРДІ ЕНГІЗУ ТУРАЛЫ

КУӘЛІК

2025 жылғы «15» сәуір № 56775

Автордың (лардың) жөні, аты, әкесінің аты (егер ол жеке басын куәландыратын құжатта көрсетілсе): **ДУЙСЕНБЕККЫЗЫ ЖАНСАЯ, РАХИМОВА ДИАНА РАМАЗАНОВНА, АЛИЕВ РАШИД ШАМИЛЬЕВИЧ**

Авторлық құқық объектісі: **ЭЕМ-ге арналған бағдарлама**

Объектінің атауы: **«Dataset Loader Bot» БАЛАЛАРДЫҢ ҚАЗАҚША СӨЙЛЕУ ДЕРЕКТЕРІН ЖАРТЫЛАЙ АВТОМАТТЫ ЖИНАУ ЖҮЙЕСІ**

Объектіні жасаған күні: **09.07.2024**

ҚР ҰЛТТЫҚ ЗИЯТКЕРЛІК МЕНШІК ИНСТИТУТЫ РГП «НАЦИОНАЛЬНЫЙ ИНСТИТУТ ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ» МЮ РК

Құжат түпнұсқалығын http://www.kazpatent.kz/ru сайтының "Авторлық құқық" бөлімінде тексеруге болады https://copyright.kazpatent.kz

Подлинность документа возможно проверить на сайте kazpatent.kz в разделе «Авторское право» https://copyright.kazpatent.kz

ЭЦҚ қол қойылды

С. Ахметов

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ

РЕСПУБЛИКА КАЗАХСТАН

АВТОРЛЫҚ ҚҰҚЫҚПЕН ҚОРҒАЛАТЫН ОБЪЕКТІЛЕРГЕ ҚҰҚЫҚТАРДЫҢ МЕМЛЕКЕТТІК ТІЗІЛІМГЕ МӘЛІМЕТТЕРДІ ЕНГІЗУ ТУРАЛЫ

КУӘЛІК

2026 жылғы «27» сәуір № 70955

Автордың (лардың) жөні, аты, әкесінің аты (егер ол жеке басын куәландыратын құжатта көрсетілсе):
ДҮЙСЕНБЕККЫЗЫ ЖАНСАЯ, **РАХИМОВА ДИАНА РАМАЗАНОВНА**

Авторлық құқық объектісі: **деректер базасы**

Объектінің атауы: **Қазақ тіліндегі 2–9 жас аралығындағы балалар сөйлеу аудиожазбаларының таңбаланған акустикалық деректер базасы**

Объектіні жасаған күні: **20.10.2023**

ҚР ҰЛТТЫҚ ЗИЯТКЕРЛІК МЕНШІК ИНСТИТУТЫ РГП «НАЦИОНАЛЬНЫЙ ИНСТИТУТ ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ» РГП РК

Құжат түпнұсқалығын http://www.kazpatent.kz/ru сайтының "Авторлық құқық" бөлімінде тексеруге болады https://copyright.kazpatent.kz

Подлинность документа возможно проверить на сайте kazpatent.kz в разделе «Авторское право» https://copyright.kazpatent.kz

ЭЦҚ қол қойылды

С. Ахметов

ҚОСЫМША Б Ендіру актісі

Бекітемін

КМҚК «№154бебекжай-балабақшасының:
меңгерушісі: Абдыкалыкова М.Е
М.О.

Алматы қ.
«22» 10 2025 ж.
№ 01/25

ҒЫЛЫМИ ЗЕРТТЕУ ЖОБАСЫНЫҢ НӘТИЖЕЛЕРІН ЕНГІЗУ АКТІСІ

Біз, төменде қол қойғандар, осы актіні жасап, № AP19577833 «Жасанды интеллект технологиясы негізінде сөйлеу бұзылыстары бар балаларға арналған оқыту әдістері мен электрондық ресурстарды зерттеу және әзірлеу» ғылыми-зерттеу жобасының бойынша орындалған нәтижелері мектепке дейінгі ұйымның оқу-тәрбие процесіне енгізілгенін растаймыз.

1. Жобаның сипаттамасы

Жоба шаруашылық жүргізу құқындағы республикалық мемлекеттік кәсіпорын «Ақпараттық және есептеу технологиялары институты» базасында жүзеге асырылған және жобаның ғылыми жетекшісі PhD, қауымдастырылған профессор Рахимова Диана Рамазановна. Жобаның мақсаты – қазақ тілінде сөйлеу бұзылыстары бар балаларды оқытуға арналған инновациялық әдістер мен электрондық ресурстарды әзірлеу және енгізу, сонымен қатар жасанды интеллект (ЖИ) технологияларын қолдану арқылы логопедтер мен тәрбиешілердің жұмысын қолдау.

2. Жобаның мазмұны мен нәтижелері

Жоба аясында https://logopedai.kz веб-порталы әзірленіп, енгізілді. Портал интерактивті логопедиялық сабақтарды ұйымдастыруға, баланың ерекшеліктеріне сәйкес тапсырмаларды таңдауға және жасанды интеллект алгоритмдері арқылы айтылымды талдауға мүмкіндік береді.
Платформа оқу материалдарын көру және дыбыстық сүйемелдеуді қамтамасыз етеді, балалардың танымдық белсенділігін арттыруға және оқыту процесін дараландыруға ықпал етеді.

3. Енгізу саласы

Жоба нәтижелері мектепке дейінгі ұйымдардың оқу үдерісіне енгізіліп, келесі бағыттарда қолданылады:
- жеке және топтық логопедиялық сабақтар өткізу;
- сөйлеу қабілеті шектеулі балаларға арналған түзету-дамыту бағдарламаларын дайындау;
- тәрбиешілер мен логопедтерді logopedai.kz платформасын қолдануға үйрету;

4. Енгізудің тиімділігі мен маңыздылығы

Logopedai.kz порталының енгізілуі балалардың сөйлеу қабілетінің дамуына, сабаққа қызығушылығына және педагогтардың жұмысының тиімділігіне оң әсер етті. ЖИ технологияларын қолдану диагностика дәлдігін арттыруға, тапсырмалардың бейімделуіне және оқушылардың сабаққа белсенді қатысуына мүмкіндік берді. Жүйе педагогикалық ұжым мен ата-аналар тарапынан оң пікірлерге ие болып, өзінің тәжірибелік маңыздылығы мен тұрақтылығын дәлелдеді.

5. Енгізілгенін растау

Осы актімен № AP19577833 ғылыми-зерттеу жобасының нәтижелері мектепке дейінгі ұйымның оқу-тәрбие процесіне сәтті енгізілгені және логопед-педагогтар мен тәрбиешілердің практикалық қызметінде пайдаланылатыны расталады.

Комиссия төрағасы: Абдыкалыкова М.Е

Комиссия мүшелері: Калыкова Е.Р

Жаилганова Г.К

Калипанова Б.Т