ӘОЖ 004.912; 004.93
ҒТАХР 28.23.37

https://doi.org/10.55452/1998-6688-2025-22-4-23-30

[1]*Оралбекова Д.,
PhD, аға ғылыми қызметкер, ORCID ID: 0000-0003-4975-6493,
*e-mail: dinaoral@mail.ru
[1]Мамырбаев О.,
PhD, профессор, бас ғылыми қызметкер, ORCID ID: 0000-0001-8318-3794,
e-mail: morkenj@mail.ru
[1]Еримбетова А.,
PhD, т.ғ.к., қауымдастырылған профессор, ORCID ID: 0000-0002-2013-1513,
e-mail: aigerian8888@gmail.com
[1]Бекарыстанқызы А.,
PhD, аға ғылыми қызметкер, ORCID ID: 0000-0003-3984-2718,
e-mail: akbayan.b@gmail.com
[1]Тұрдалыұлы М.,
PhD, аға ғылыми қызметкер, ORCID ID: 0000-0002-1470-3706,
e-mail: m.turdalyuly@gmail.com

[1]Ақпараттық және есептеуіш технологиялар институты, Алматы қ., Қазақстан

АВТОГРЕССИЯЛЫҚ ЕМЕС ДЕКОДТАУДЫҢ ҚАЗАҚ ТІЛІНДЕГІ СӨЙЛЕУДІ ТАНУДА ҚОЛДАНУ

Аңдатпа

Сөйлеуді тану саласында интегралдық модельдер дәстүрлі және гибридті тәсілдерді біртіндеп ығыстырып келеді. Олардың негізгі принципі – автогрессиялық декодтау, мұнда шығатын тізбек солдан оңға қарай қалыптасады. Алайда осы уақытқа дейін бұл әдіс дауысты мәтінге ауыстыруда ең жақсы нәтиже беретіні дәлелденген жоқ. Сонымен қатар, интегралдық модельдер тек алдыңғы контекстке сүйенеді, бұл анық емес немесе бұрмаланған дауыстарды өңдеуді қиындатады. Осыған байланысты Insertion әдісі ұсынылды, ол автогрессиялық декодтауды пайдаланбайды және шығатын деректерді кез келген ретпен генерациялайды. Осы жұмыста Insertion әдісі мен коннекциялық уақытша классификация (CTC) негізінде оқытылған қазақ тіліндегі сөйлеуді тану моделі қарастырылады. Жүргізілген эксперименттер бұл әдістің тану дәлдігін арттыруға мүмкіндік беретінін көрсетті. Автогрессиялық модельдерден айырмашылығы, Insertion әдісі тізбектерді өңдеуде үлкен икемділік береді, себебі ол шығатын деректерді қалыптастыруда қатаң тәртіпті қажет етпейді. Бұл декодтау кезіндегі кідірістерді азайтады және модельді анық айтылмаған сөздерге төзімді етеді. Сонымен қатар, Insertion әдісін CTC-мен үйлестіру аудио деректер мен мәтін транскрипциясының сәйкестігін жақсартуға мүмкіндік береді. Бұл агглютинативті тілдер, мысалы, қазақ тілі үшін өте маңызды. Эксперименттер нәтижесінде ұсынылған модельдің тану дәлдігі 10,2%-ға дейін жетіп, қазіргі таңда бәсекеге қабілетті көрсеткіштерге ие болды.

Тірек сөздер: коннекциялық уақытша классификация, интеграциялық модельдер, қазақ тіліндегі дыбысты тану, Insertion, Transformer.

Кіріспе

Сөйлеуді синтездеу және тану технологиялары роботтандырылған және автоматтандырылған жүйелердің ажырамас бөлігіне айналды. Олар call-орталықтарда операторлардың жүктемесін азайтуға, қоңырауларды мамандар арасында жылдам бөлуге, сондай-ақ банк саласында интеллектуалды дауыс көмекшілері мен басқа да сервистер түрінде кеңінен қолданылуда. Интегралдық (end-to-end, E2E) модельдер [1], оның ішінде коннекциялық уақытша классификация (CTC) [2], кодер-декодерлік архитектуралар мен назар аудару механизмдері [3],

сондай-ақ ағымдық дауысты өңдеуге арналған онлайн-модельдер [4, 5], дыбысты автоматты тану (ASR) жүйелерін құруды айтарлықтай жеңілдетті. E2E-модельдер дәстүрлі көпдеңгейлі жүйелерді бірегей терең нейрожелімен ауыстырып, оқыту және декодтау уақытын қысқартуға мүмкіндік береді. Сонымен қатар, олар тану процесін және оның кейінгі өңдеуін, мысалы табиғи тілдерді түсіну тапсырмалары үшін, оңтайландыруға жол ашады.

Соңғы жылдары E2E-модельдердің тиімділігін арттыру бойынша белсенді жұмыстар жүргізілуде, әсіресе CTC [6, 7] және назар аудару механизмдерін [8, 9] жетілдіруде үлкен жетістіктерге қол жеткізілді. Зерттеулер көрсеткендей, әртүрлі архитектураларды біріктіру арқылы интегралдық модельдерді қолдану дәстүрлі ASR жүйелерінен ғана емес, сондай-ақ жеке жүзеге асырылған E2E тәсілдерінен де жоғары нәтижелерге қол жеткізуге мүмкіндік береді. Осы саладағы айтарлықтай прогресс Transformer моделінің пайда болуы және оның CTC-мен біріктірілуі [10, 11] арқасында мүмкін болды, бұл дыбысты тану сапасын жаңа деңгейге көтерді.

Көптеген E2E-модельдер «тізбектен тізбекке» (seq2seq) тәсілімен жұмыс істейді, онда автогрессиялық декодтау қолданылады, мұнда шығатын тізбек солдан оңға қарай қалыптасады. Алайда бұл әдіс шектеулерге ие, ол тек алдыңғы контекстi ескеріп, болашақ сөздер туралы ақпаратты елемейді. Бұл анық емес айтылған дыбысты өңдеуде қателіктерге әкелуі мүмкін. Мұндай жағдайда тек алдыңғы емес, сондай-ақ кейінгі сөздерді ескеру маңызды. Бұндай көзқарас модельдің фонетикалық бұрмалауларға төзімділігін арттырады.

Машиналық аударма тапсырмаларында белсенді түрде автогрессиялық емес трансформерлер [12] қолданылады, және осы тәсілді сөйлеуді тануға бейімдеу бойынша әрекеттер жасалды [13]. Бұл әдіс негізінде декодтау процесінде жасыратын токендерді болжау жатыр. Модель жасырынып қалған элементтерді назар аудару механизмі арқылы қайта құруға оқытылады, бұл оған автогрессиялық емес тәсілмен тізбектерді генерациялауға мүмкіндік береді. Бұл әдіс автогрессиялық модельдермен салыстырғанда жоғары тиімділік көрсетті, бірақ оны қолдану кезінде қиындық туындайды. Бұл – шығатын тізбектің ұзындығын алдын ала анықтау қажеттілігі.

Осы мәселені шешу үшін Insertion әдісі ұсынылды, ол шығатын деректерді кез-келген ретпен генерациялауға мүмкіндік береді, бұл тізбектің ұзындығын болжау үшін қосымша элементтерді қажет етпейді [14]. Бұл тәсіл декодтау қадамдарының санын азайтып, өңдеу уақытын қысқартуға мүмкіндік береді.

Біздің зерттеуіміз қазақ тілін ақпараттық жүйелерде автоматты тану технологияларын дамытуға және енгізуге бағытталған. Алдыңғы жұмыстарда [15,16] қазақ тілін машиналық аударма мен морфологиялық талдау мәселелері қарастырылған, сондай-ақ дауысты идентификациялау және верификациялау жүйелері әзірленген [17]. Бұған дейін біз агглютинативті тілдер, мысалы қазақ және әзірбайжан тілдерінде тану үшін назар аудару механизмі бар CTC-модельдерін [18], сондай-ақ қазақ тілінде ағымдық тану үшін RNN-T моделін [19] іске асырдық. Алайда, қолданыстағы шешімдер қазақ тілін мәтінге түрлендірудің дәлдігін арттыру үшін жетілдіруді қажет етеді.

Бұл жұмыстың негізгі мақсаты – Insertion әдісіне негізделген қазақ тілін интегралдық тану моделін әзірлеу және тестілеу, сондай-ақ оның тиімділігін қолданыстағы тәсілдермен салыстыру. Зерттеу барысында келесі міндеттер қойылды:

- Қазақ тілін автоматты тану үшін Insertion моделін әзірлеп, оқыту.
- Декодтау дәлдігін арттыру үшін Insertion әдісін CTC-мен біріктіру.
- Қазақ тілін тану бойынша эксперименттер жүргізіп, нәтижелерді дәстүрлі автогрессиялық модельдермен салыстыру.
- Декодтау қадамдарының санын азайту және модельдің жұмысын жеделдету үшін Insertion әдісінің әсерін бағалау.

Ұсынылған тәсіл қазақ тілін тану дәлдігін арттыруға және сөздер мен таңбаларды танудағы қателіктерді азайтуға бағытталған. Бұл агглютинативті морфологиясы бар тілдерді өңдеуге перспективті болатынын көрсетеді.

Материалдар мен әдістер

Ең ықтимал сөздер тізбегі W барлық мүмкін болатын V тізбектері арасында P(W|A) ықтималдылығын максимизациялау арқылы анықталады (1). Бұл процесті келесі өрнекпен көрсетуге болады [20]:

$$W^* = \underset{W \in V^*}{\operatorname{argmax}} P(W|A) \quad (1)$$

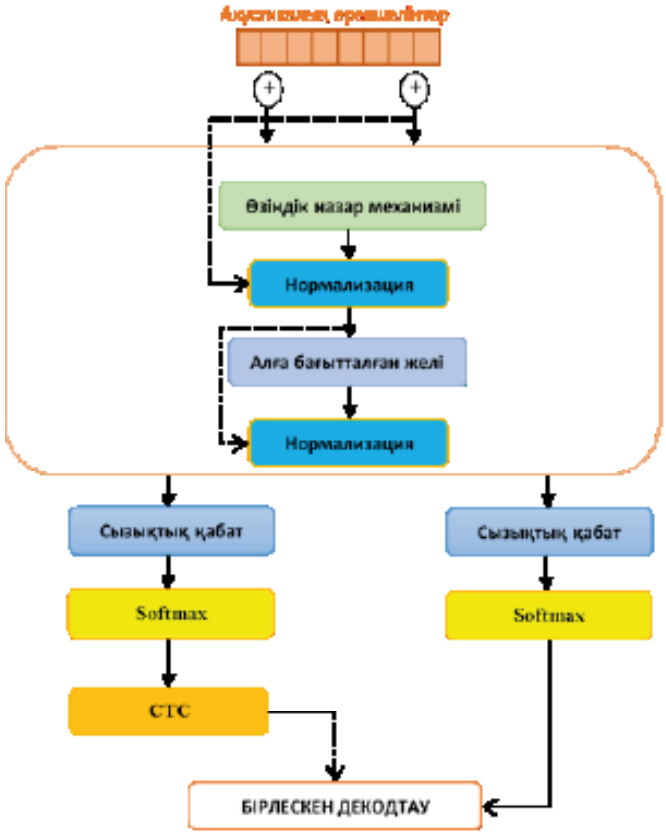
Бұл өрнекте акустикалық сипаттамалар A түріндегі ерекшелік векторларының тізбегі ретінде көрсетіледі, ал W – шығыс сөздер тізбегін білдіреді. Осылайша, интегралдық автоматты сөйлеуді тану жүйесінің негізгі міндеті – барлық ықтимал V* тізбектері үшін P(W | A) апостериорлы үлестірімін дәл бағалай алатын модель құру.

Insertion моделін іске асыру үшін тек сөздер тізбегін емес, олардың қосылу тәртібін де ескеру қажет. Order параметрі сөздердің декодтау процесінде нәтижелік тізбекке қандай тәртіппен енгізілетінін анықтайды. Мысалы, сөздерді бірінен соң бірін тізбектей қосудың орнына, модель алдымен негізгі сөздерді немесе ең ықтимал токендерді болжай алады, сосын оларды қалған элементтермен толықтырады. Бұл декодтауды икемді етеді, себебі модель әртүрлі жағдайларға бейімделіп, болашақ сөздер контексін пайдалана алады.

Order параметрін ескере отырып, (1) теңдеуі келесі түрде түрленеді:

$$P(W \mid A) = \sum_{Order} p(W, Order \mid A) = \sum_{Order} p(W_{Order} \mid A) p(Order|A) \quad (2)$$

Бұл тәсіл декодтау кезінде шығыс тізбегін құрудың бірнеше нұсқасын ескеруге мүмкіндік береді. Insertion әдісіне негізделген декодтауды іске асыру үшін Transformer нейрожелісі пайдаланылады, мұнда позициялық тәртіп механизмі қолданылады (1-сурет).



Сурет 1 – Insertion-CTC гибридті моделінің құрылымы

Әзірленген модель интегралдық әдіспен оқытылады, бұл E2E архитектураларының барлық артықшылықтарын сақтауға мүмкіндік береді. Бұл әдістің негізгі артықшылықтарының бірі – шығатын тізбектің ұзындығын алдын ала анықтаудың қажеттілігі жоқ, сонымен қатар бұл декодтау процесін икемді етеді.

(2) теңдеуде көрсетілгендей, сөздердің позицияларын болжау алдын ала жасалған сөздерге назар аудару механизмі арқылы енгізу әдісімен жүзеге асырылады. Позицияларды жаңарту бұрын есептелген ұсыныстарға әсер етпегендіктен, енгізу тәртібі туралы ақпарат кейінгі декодтау кезеңдерінде қайта қолданылуы мүмкін.

Шығыс тізбегін құру процесінде әр сөздің шартты ықтималдығы сөздердің және олардың позицияларының біріктірілген үлестірімін ескере отырып есептеледі. Бұл енгізу тәртібін оңтайландырады және декодтаудың дәлдігін арттырады.

Нәтижелер мен талқылау

Insertion әдісіне негізделген модельді оқыту үшін Ақпараттық және есептеуіш технологиялар институты жинаған көлемі 405 сағаттан астам тілдік корпус пайдаланылды. Бұл корпус әртүрлі жастағы және жыныстағы адамдардың қазақ тіліндегі жазбаларынан тұрады. Деректер құрамында телефон арқылы сөйлесу жазбалары мен транскрипциялары, сондай-ақ жаңалықтар көздерінен алынған аудиоматериалдар мен көркем аудиокітаптар бар.

Барлық дыбыс файлдары оқыту (85%) және тестілеу (15%) жиынтығына бөлінді, бұл модельді тиімді оқытуға және тексеруге мүмкіндік береді. Алынған корпус түрлі деректер жиынтығының тану сапасына әсерін зерттеуге, сондай-ақ модель параметрлерін талдап, олардың қазақ тілін декодтау дәлдігіне әсерін бағалауға мүмкіндік береді.

Аудиофайлдар .wav форматында ұсынылған, ал оларды өңдеуде жоғары дәлдікпен сандық деректерді алуға мүмкіндік беретін сызықтық импульс-кодтық модуляция (PCM) әдісі қолданылған. Жазбалар 44,1 кГц дискретизациясы және 16 биттік разрядтылыққа ие, бұл дауыс деректерін өңдеу үшін сапа стандарттарына сәйкес келеді.

Негізгі параметрлерді баптау

Модельді оқыту барысында кодер мен декодер үшін 10 блок таңдалды, бұндай таңдау тізбектерді өңдеуді жақсартып, есептеу күрделілігі мен болжау дәлдігі арасындағы балансты сақтауға мүмкіндік берді.

Декодтау сапасын жақсарту және жақындауды жылдамдату үшін Insertion моделіне коннекциялық уақытша классификация (CTC) интеграцияланды. Модельде әр қабатта 512 ұяшығы бар алты қабатты екі бағытты LSTM (BLSTM) қолданылды. Осынын арқасында контексті өңдеу жақсартылып, ұзақ сөйлемдерді тану дәлдігі артты. Артық оқытуды болдырмау үшін dropout = 0,1 коэффициентпен шектелді.

Модель параметрлерін оңтайландыру үшін градиентті төмендету тұрақтылығын қамтамасыз ететін және салмақтардың жылдам бейімделуін қамтамасыз ететін Adam алгоритмі пайдаланылды. Декодтау кезінде CTC салмағы 0,4 деңгейінде орнатылды. Осындай шектеу модельге аудио сигналын мәтіндік транскрипциямен жақсы үйлестіруге мүмкіндік берді.

Қосымша beam search іздеу ені 15 пайдаланылды, себебі осындай шектеу күрделі тіркестер мен нақты қазақ сөз түрлерін тану сапасын жақсартуға ықпал етті. Іздеу енін ұлғайту модельге соңғы оптималды болжауды таңдамас бұрын көбірек гипотезаларды ескеруге мүмкіндік берді, бірақ бұл есептеу шығындарын аздап арттырды.

Оқыту 100 кезең бойы жүргізілді, ол модельдің тереңірек жақындауына және шығыс деректерінің жақсы генерациясына мүмкіндік берді. Сонымен қатар, алдын ала оқытылған эмбеддингтер қолданылды. Олар оқытудың бастапқы кезеңін жылдамдатып, жалпы дәлдікті арттыруға ықпал етті.

Декодтау кезеңінде сөйлеу тізбектерін үйлестіру, сирек сөздерді тануды жақсарту және модельдің соңғы шығысынан тиімді нәтиже алу үшін тілдік модель қолданылды.

Эксперимент жүргізу және нәтижелерді алу

Қазіргі уақытта қазақ тіліндегі сөйлеуді тану бойынша интегралдық модельдерге арналған зерттеулер әлі шектеулі. Сонымен қатар, қазақ сөйлеуін автоматты тануға арналған дәстүрлі және гибридті тәсілдерді қарастырған бірнеше ғылыми жұмыстар бар.

Біздің алдыңғы жұмысымызда [18] трансферлік оқытуға негізделген тәсіл ұсынылған, ол қазақ және әзірбайжан тілдерінде интегралдық сөйлеуді тану үшін қолданылды. Зерттеу аясында агглютинативті тілдерді тиімді өңдеуге мүмкіндік беретін CTC-назар аудару модельдері әзірленді. Оқыту екі тілдік корпус бойынша бір уақытта жүргізілді, оның ішінде 400 сағаттық қазақ тіліндегі аудиолар болды.

Басқа зерттеуде [19] қазақ тіліндегі дыбысты тану үшін бейімделген RNN-T моделінің нұсқасы іске асырылды. Бұл жұмыста сөйлеу болжамдарын жақсартуға оң әсерін тигізген тілдік модель (ТМ) қолданылды. Модель 300 сағаттық корпус бойынша оқытылды, бұл соңғы транскрипциялардың дәлдігін айтарлықтай арттырды.

Аталған зерттеулерде алынған дыбысты тану нәтижелерін салыстырмалы талдау және осы жұмыста ұсынылған Insertion моделінің тиімділігі 1-кестеде ұсынылған.

Кесте 1 – Дыбысты тану жүйелері үшін CER көрсеткіштері

Модель	CER (%)	Деректер көлемі, сағ
End-to-end трансфермен, ТМ-сіз	14.23	400
LSTM және BLSTM негізіндегі RNN-T, ТМ-мен	10.6	300
Ұсынылған модель Insertion-CTC	10.2	405

Insertion моделін CTC және тілдік модельмен үйлестіріп қолдану тану дәлдігін 4%-ға арттыруға мүмкіндік берді. Алайда, осы компоненттердің жоқ болуы Insertion моделінің тиімділігіне теріс әсер етіп, болжау сапасын төмендетті. Сонымен қатар, Insertion әдісінің ерекшеліктерінің арқасында декодтау қадамдарының саны азайтылды. Бұл басқа модельдермен салыстырғанда оқыту мен сөйлеуді өңдеу процесін жылдамдатуға ықпал етті.

Қорытынды

Бұл жұмыс қазақ тіліндегі сөйлеуді тану үшін CTC және Insertion бірлескен интегралдық модельдерінің әсерін зерттеуге арналады. Эксперименттер қазақ тілінің 405 сағаттық аралас сөйлеу корпусымен жүргізілді және нәтижелер көрсеткендей, жүйе RNN негізіндегі тілдік модельдер қолданылғанда жоғары нәтижелерге қол жеткізе алды. Бұл модельдер негізіндегі декодтау есептеу шығындарын арттырмайды, сондықтан декодтау жылдамдығы баяуламайды. Сонымен қатар, Insertion моделі іске асырылғанда декодтау кезінде итерациялар саны азаяды. Осылайша, CER көрсеткішінің ең жақсы нәтижесі 10.2%-ды құрады, бұл бүгінгі таңда бәсекеге қабілетті нәтиже болып табылады. Ұсынылған әдіс икемді және айнымалылардың шартты тәуелсіздігін талап етпейді. Сонымен қатар, бұл модельді түркі тілдері сияқты туыс тілдерде сөйлеуді тану үшін қолдануға болады деген қорытынды жасауға болады.

Келесі жұмыстарда біз агглютинативті тілдерді тану үшін басқа Transformer негізіндегі модельдерді зерттеуді жоспарлап отырмыз.

Қаржыландыру туралы ақпарат. Бұл зерттеу Қазақстан Республикасының Ғылым және жоғары білім министрлігі Ғылым комитеті тарапынан қаржыландырылды (Грант № BR24992875).

ӘДЕБИЕТТЕР

1 Mamyrbayev, O., Oralbekova, D. Modern trends in the development of speech recognition systems. News of the National Academy of Sciences of the Republic of Kazakhstan, 4 (332), 42–51 (2020).

2 Graves, A., Fernandez, S., Gomez, F., and Schmidhuber, J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. ICML, Pittsburgh, USA, pp. 369–376 (2006).

3 Majhi, M.K., and Saha, S.K. An automatic speech recognition system in Odia language using attention mechanism and data augmentation. International Journal of Speech Technology, 27, 717–728 (2024). https://doi.org/10.1007/s10772-024-10132-6.

4 Cheng, L., Zhu, H., Hu, Z., and Luo, B. A Sequence-to-Sequence Model for Online Signal Detection and Format Recognition. IEEE Signal Processing Letters, 31, 994–998 (2024). https://doi.org/10.1109/LSP.2024.3384015.

5 Mundotiya, R.K., Mehta, A., Baruah, R., and Singh, A.K. Integration of morphological features and contextual weightage using monotonic chunk attention for part of speech tagging. Journal of King Saud University – Computer and Information Sciences, 34, 7324–7334 (2021).

6 Deng, K., Cao, S., Zhang, Y., Ma, L., Cheng, G., Xu, J., and Zhang, P. Improving CTC-based speech recognition via knowledge transferring from pre-trained language models (2022). https://doi.org/10.48550/arXiv.2203.03582.

7 Dingliwal, S., Sunkara, M., Ronanki, S., Farris, J., Kirchhoff, K., and Bodapati, S. Personalization of CTC Speech Recognition Models. 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, pp. 302–309 (2023). https://doi.org/10.1109/SLT54892.2023.10022705.

8 Zeyer, A., Kazuki, I., Ralf, S., and Hermann, N. Improved training of end-to-end attention models for speech recognition. arXiv preprint arXiv:1805.03294 (2018).

9 Wang, Z., Tao, Z., Shao, Y., and Ding, B. LSTM-convolutional-BLSTM encoder-decoder network for minimum mean-square error approach to speech enhancement. Applied Acoustics, 172, 107647 (2021).

10 Huang, Z., Wang, P., Wang, J., Miao, H., Xu, J., and Zhang, P. Improving Transformer Based End-to-End Code-Switching Speech Recognition Using Language Identification. Applied Sciences, 11 (19), 9106 (2021). https://doi.org/10.3390/app11199106.

11 Miao, H., Cheng, G., Gao, C., Zhang, P., and Yan, Y. Transformer-Based Online CTC/Attention End-To-End Speech Recognition Architecture. ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6084–6088 (2020). https://doi.org/10.1109/ICASSP40776.2020.9053165.

12 Li, F., Chen, J., and Zhang, X. A Survey of Non-Autoregressive Neural Machine Translation. Electronics, 12 (13), 2980 (2023). https://doi.org/10.3390/electronics12132980

13 Chen, N., Watanabe, S., Villalba, J., and Dehak, N. Listen and fill in the missing letters: Non-autoregressive transformer for speech recognition. arXiv preprint arXiv:1911.04908 (2020).

14 Fujita, Y., Watanabe, S., Omachi, M., and Chan, X. Insertion-Based Modeling for End-to-End Automatic Speech Recognition. INTERSPEECH 2020 (2020). https://doi.org/10.48550/arXiv.2005.13211.

15 Rakhimova, D., Sagat, K., Zhakypbaeva, K., and Zhunussova, A. Development and Study of a Post-editing Model for Russian-Kazakh and English-Kazakh Translation Based on Machine Learning. In: Wojtkiewicz K., Treur J., Pimenidis E., Maleszka M. (eds) Advances in Computational Collective Intelligence. ICCCI 2021. Communications in Computer and Information Science, vol. 1463. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88113-9_42.

16 Karyukin, V., Rakhimova, D., Karibayeva, A., Turganbayeva, A., and Turarbek, A. The neural machine translation models for the low-resource Kazakh–English language pair. PeerJ Computer Science, 9, e1224 (2023). https://doi.org/10.7717/peerj-cs.1224

17 Kydyrbekova, A., and Oralbekova, D. Speaker identification using distribution-preserving x-vector generation. News of the National Academy of Sciences of the Republic of Kazakhstan, Physico-mathematical series, 4 (352), 152–162 (2024). https://doi.org/10.32014/2024.2518-1726.314

18 Mamyrbayev, O., Alimhan, K., Oralbekova, D., Bekarystankyzy, A., and Zhumazhanov, B. Identifying the influence of transfer learning method in developing an end-to-end automatic speech recognition system with a low data level. Eastern-European Journal of Enterprise Technologies, 1 (9(115)), 84–92 (2022). https://doi.org/10.15587/1729-4061.2022.252801.

19 Mamyrbayev, O., Oralbekova, D., Kydyrbekova, A., Turdalykyzy, T., and Bekarystankyzy, A. End-to-End Model Based on RNN-T for Kazakh Speech Recognition. 2021 3rd International Conference on Computer Communication and the Internet (ICCCI), pp. 163–167 (2021). https://doi.org/10.1109/ICCCI51764.2021.9486811.

20 Soleymanpour, M., Johnson, M.T., Soleymanpour, R., and Berry, J. Synthesizing Dysarthric Speech Using Multi-Speaker TTS for Dysarthric Speech Recognition. ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, pp. 7382–7386 (2022). https://doi.org/10.1109/ICASSP43922.2022.9746585.

[1]*Оралбекова Д.,
PhD, старший научный сотрудник, ORCID ID: 0000-0003-4975-6493,
e-mail: dinaoral@mail.ru
[1]Мамырбаев О.,
PhD, профессор, главный научный сотрудник, ORCID ID: 0000-0001-8318-3794,
e-mail: morkenj@mail.ru
[1]Еримбетова А.,
PhD, к.т.н., ассоц. профессор, ORCID ID: 0000-0002-2013-1513,
e-mail: aigerian8888@gmail.com
[1]Бекарыстанқызы А.,
PhD, старший научный сотрудник, ORCID ID: 0000-0003-3984-2718,
e-mail: akbayan.b@gmail.com
[1]Тұрдалыұлы М.,
PhD, старший научный сотрудник, ORCID ID: 0000-0002-1470-3706,
e-mail: m.turdalyuly@gmail.com

[1]Институт информационных и вычислительных технологий КН МНВО РК,
г. Алматы, Казахстан

ПРИМЕНЕНИЕ НЕАВТОРЕГРЕССИОННОГО ДЕКОДИРОВАНИЯ ДЛЯ РАСПОЗНАВАНИЯ КАЗАХСКОЙ РЕЧИ

Аннотация

В области распознавания речи интегральные модели постепенно вытесняют традиционные и гибридные подходы. Основным их принципом является автогрессионное декодирование, при котором выходная последовательность формируется слева направо. Однако до сих пор не было доказано, что этот метод дает наилучшие результаты при преобразовании речи в текст. Кроме того, интегральные модели полагаются только на предыдущий контекст, что затрудняет обработку нечетких или искаженных звуков. В связи с этим был предложен метод Insertion, который не использует автогрессионное декодирование и генерирует выходные данные в произвольном порядке. В данной работе рассматривается модель распознавания казахской речи, обученная на основе метода Insertion и коннекционной временной классификации (CTC). Проведенные эксперименты показали, что этот метод позволяет повысить точность распознавания. В отличие от автогрессионных моделей, метод Insertion предоставляет большую гибкость в обработке последовательностей, поскольку не требует строгого порядка формирования выходных данных. Это снижает задержки при декодировании и делает модель более устойчивой к нечетко произнесенным словам. Кроме того, сочетание метода Insertion с CTC позволяет улучшить соответствие аудиоданных и текстовой транскрипции. Это особенно важно для агглютинативных языков, таких как казахский. По результатам экспериментов точность распознавания предложенной модели достигла 10,2%, что делает ее конкурентоспособной на сегодняшний день.

Ключевые слова: коннекционная временная классификация, интегральные модели, распознавание казахской речи, Insertion, Transformer.

[1]*Oralbekova D.,
PhD, Senior Researcher, ORCID ID: 0000-0003-4975-6493,
e-mail: dinaoral@mail.ru
[1]Mamyrbayev O.,
PhD, Professor, Chief Researcher, ORCID ID: 0000-0001-8318-3794,
e-mail: morkenj@mail.ru
[1]Yerimbetova A.,
PhD, Cand.Tech.Sc., Associate Professor, ORCID ID: 0000-0002-2013-1513,
e-mail: aigerian8888@gmail.com
[1]Bekarystankyzy A.,
PhD, Senior Researcher, ORCID ID: 0000-0003-3984-2718,
e-mail: akbayan.b@gmail.com
[1]Turdalyuly M.,
PhD, Senior Researcher, ORCID ID: 0000-0002-1470-3706,
e-mail: m.turdalyuly@gmail.com

[1] Institute of Information and Computational technology, Almaty, Kazakhstan

APPLICATION OF NON-AUTOREGRESSIVE DECODING FOR KAZAKH SPEECH RECOGNITION

Abstract

In the field of speech recognition, end-to-end models are gradually replacing traditional and hybrid approaches. Their main principle is autoregressive decoding, where the output sequence is formed from left to right. However, it has not yet been proven that this method provides the best results in converting speech to text. Moreover, end-to-end models rely solely on the previous context, which complicates the processing of unclear or distorted sounds. In this regard, the insertion method was proposed, which does not use autoregressive decoding and generates output data in an arbitrary order. This paper examines a Kazakh speech recognition model trained using the insertion method and Connectionist Temporal Classification (CTC). The experiments conducted showed that this method improves recognition accuracy. Unlike autoregressive models, the Insertion method provides greater flexibility in processing sequences, as it does not require a strict order for generating output data. This reduces decoding delays and makes the model more robust to poorly pronounced words. Furthermore, combining the Insertion method with CTC improves the alignment of audio data and text transcription. This is especially important for agglutinative languages such as Kazakh. According to the experimental results, the recognition accuracy of the proposed model reached 10.2%, making it competitive today.

Keywords: connectionist temporal classification, end-to-end models, Kazakh speech recognition, insertion, transformer.

Мақаланың редакцияға түскен күні: 24.02.2025