ӘОЖ 004

ДИКТОФОН ЖАЗБАЛАРЫНДАҒЫ СӨЙЛЕУДІ ТАНУ ҮШІН НЕЙРОНДЫҚ ЖЕЛІЛЕРДІ ЗЕРТТЕУ ЖӘНЕ ҚОЛДАНУ

Толеу Ж.С.
магистрант Қ.Құлажанов ат. ҚазТБУ, Астана қ., Қазақстан

Ғылыми жетекшісі: PhD, қауым.проф. Ұзаққызы Н.
Қ.Құлажанов ат.ҚазТБУ Астана қ., Қазақстан

Мақалада диктофоннан дыбыстық сигналдарды танудың заманауи әдістері, атап айтқанда терең оқыту моделі (Deep Learning) қарастырылады. Сөйлеуді өңдеу кезеңдері, шуды басу және қазақ тіліндегі дыбыстарды тану мәселелері талданады. Бұл жұмыста диктофон құрылғыларынан алынған сөйлеу сигналдарын мәтінге автоматты түрде түрлендіру мәселесі қарастырылады. Нейрондық желілерге негізделген сөйлеуді тану жүйелерінің құрылымы, жұмыс істеу кезеңдері және практикалық қолдану жолдары сипатталған.

Кілт сөздер: сөйлеуді тану, диктофон, нейрондық желі, ASR, MFCC, CNN, RNN, Transformer.

Кіріспе

Сөйлеуді тану технологиялары қазіргі таңда ақпараттық жүйелердің ажырамас бөлігіне айналды. Диктофондарда жазылған аудио деректерді интеллектуалды өңдеу білім беру, медицина, журналистика салаларында кеңінен қолданылады.

Бұл бағыттағы негізгі құрал - жасанды нейрондық желілер.

Диктофон жазбаларын қолмен транскрипциялау - көп уақыт пен еңбекті қажет ететін процесс. Жасанды интеллекттің, атап айтқанда нейрондық желілердің дамуы бұл процесті автоматтандыруға мүмкіндік берді. Сөйлеуді автоматты тану (Automatic Speech Recognition - ASR) жүйелері қазіргі таңда дыбыс толқындарын талдап, оны жоғары дәлдікпен мәтінге түрлендіре алады.

Сөйлеу сигналдарын өңдеу саласы ауқымды зерттеулердің арқасында айтарлықтай өзгерістерге ұшырады. Сөйлеу сапасын жақсартуға (Speech Enhancement, se) және сөйлеуді автоматты түрде тануға (Automatic Speech Recognition, ASR) қызығушылық артып келеді, бұл ретте SE ASR тиімділігін арттырудың маңызды алдын ала кезеңі болып табылады. Бұл мақалада негізгі мәселелер қарастырылады, атап айтқанда сөйлеу сапасын сақтау және ASR жүйелерінде түсінікті болу қажеттілігі. Жақында терең оқыту әдістері осы мәселелерді шешудің күшті құралдарына айналды. Бұл жүйелі шолу шуды

басуға, акустикалық модельдеуге және фокустық диаграмманы қалыптастыруға баса назар аудара отырып, сөйлеуді жақсарту және тану әдістерін қарастырады. Терең нейрондық желілер (GNS), конволюциялық нейрондық желілер (SNS), кестелік форматтағы қайталанатын нейрондық желілер (RNS), ұзақ мерзімді жады бар желілер (Long Short-Term Memory, LSTM) және гибридті нейрондық желілер сияқты әртүрлі терең оқыту архитектуралары олардың жақсартудағы және танудағы рөлі тұрғысынан қарастырылады сөйлеу [1].

Шолу олардың қолданылуын, әрбір зерттеуге қатысатын функцияларды, пайдаланылатын дерекқорларды, өнімділікті және шектеулерді егжей-тегжейлі сипаттайды. Барлық ақпарат құрылымдық түрде ұсынылған .Уақыт сигналы бұл тәсіл әр әдістің күшті және әлсіз жақтары туралы құнды түсінік алуға мүмкіндік береді, бұл осы саланың одан әрі дамуына көмектеседі. Атап айтқанда, бұл LSTM-RNN модельдері өңдеуде өте жақсы жұмыс істейтінін көрсетеді, ал гибридті модельдер тапсырмаларды орындау нәтижелерін оңтайландыруда жоғары тиімділікті көрсетеді. Мақалада сөйлеуді жақсарту және тану мәселелерін шешу үшін тек терең нейрондық желілерді пайдаланатын 187 ғылыми жұмыстың жан-жақты статистикалық талдауы берілген. Мақалада осы саладағы соңғы жетістіктер де көрсетілген. Шолу 2012-2024 жылдар аралығындағы басылымдарды зерттейді, зерттеу тенденциялары мен заңдылықтарына жарық түсіреді. Ұсынылған шешімдер осы дамып келе жатқан саладағы зерттеушілердің біліміндегі олқылықтарды жоюға арналған [2].

Зерттеу материалдары мен әдістері

Сөйлеуді тану жүйесінің архитектурасы

Қазіргі заманауи ASR жүйелері бірнеше деңгейлі нейрондық желілерден тұрады. Негізгі модельдерге Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN) және соңғы жылдардағы революциялық Transformers архитектуралары жатады.

RNN және LSTM: Уақытқа тәуелді дыбыс тізбектерін өңдеу үшін қолданылады. Олар алдыңғы дыбыстардың контекстін "есте сақтауға" қабілетті.

Transformers (Whisper, Wav2Vec): Мәтін мен дыбыстың арасындағы ұзақ байланыстарды "Attention" механизмі арқылы анықтайды. Бұл модельдер диктофондағы әртүрлі акценттер мен сөйлеу жылдамдығын тануда өте тиімді.

Қазіргі таңда ақпараттық технологиялардың қарқынды дамуына байланысты адам мен компьютер арасындағы өзара әрекеттесу жаңа деңгейге көтерілді. Соның ішінде сөйлеуді тану жүйелері (Automatic Speech Recognition - ASR) ерекше орын алады. Бұл жүйелер адамның ауызша айтқан сөзін компьютерлік жүйе арқылы мәтінге немесе басқару командаларына айналдыруға мүмкіндік береді. Сөйлеуді тану технологиялары мобильді қосымшаларда, виртуалды көмекшілерде, банктік қызметтерде, білім беру

жүйесінде және мүмкіндігі шектеулі адамдарға арналған технологияларда кеңінен қолданылады.

Сөйлеуді тану жүйесінің архитектурасы бірнеше өзара байланысты модульдерден тұрады. Әрбір модуль белгілі бір функцияны атқарады және жүйенің жалпы дәлдігі мен сенімділігіне әсер етеді. Негізгі архитектура келесі кезеңдерден тұрады: дыбысты алу, алдын ала өңдеу, белгілерді шығару, акустикалық модельдеу, тілдік модельдеу және декодтау.

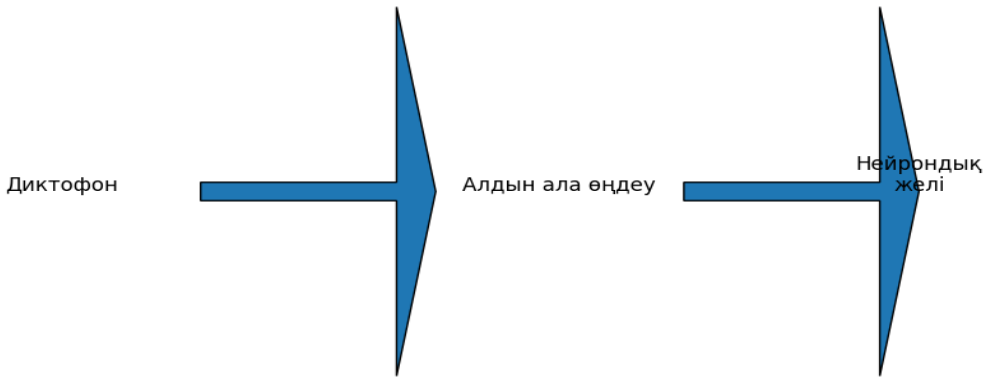
Сөйлеу сигналдарының ерекшеліктері

Сөйлеу сигналы - уақыт бойынша өзгеретін күрделі сигнал. Ол фонетикалық, акустикалық және лингвистикалық ақпаратты қамтиды. Диктофон арқылы жазылған сигналдарда шу, жаңғырық және басқа да кедергілер болуы мүмкін, сондықтан алдын ала өңдеу кезеңі өте маңызды.Сөйлеу сигналы - уақыт бойынша өзгеретін күрделі сигнал. Ол фонетикалық, акустикалық және лингвистикалық ақпаратты қамтиды. Диктофон арқылы жазылған сигналдарда шу, жаңғырық және басқа да кедергілер болуы мүмкін, сондықтан алдын ала өңдеу кезеңі өте маңызды [2,3].

Сөйлеу тану жүйесінің құрылымы

Сөйлеу тану жүйесі бірнеше негізгі кезеңнен тұрады: аудио сигналды қабылдау, алдын ала өңдеу, ерекшеліктерді шығару және нейрондық желі көмегімен тану. Бұл кезеңдердің әрқайсысы тану дәлдігіне тікелей әсер етеді.Сөйлеу тану жүйесі бірнеше негізгі кезеңнен тұрады: аудио сигналды қабылдау, алдын ала өңдеу, ерекшеліктерді шығару және нейрондық желі көмегімен тану. Бұл кезеңдердің әрқайсысы тану дәлдігіне тікелей әсер етеді.

Сөйлеу тану жүйесінің схемасы келесі суретте көрсетілген.



1-сурет. Сөйлеу тану жүйесінің схемасы

Дыбыстық сигналды өңдеу процесі

Диктофоннан алынған шикі дыбыс файлы тікелей желіге берілмейді. Ол бірнеше математикалық түрлендіруден өтеді:

Алдын ала өңдеу (Digital Signal Processing): Дыбыс сигналындағы артық шуды (background noise) жою және сигналды нормализациялау.

Белгілерді алу (Feature Extraction): Дыбыс толқынын математикалық векторларға айналдыру. Мұнда көбіне MFCC (Mel-frequency cepstral coefficients) қолданылады.

Дыбыстық сигналды өңдеу - қазіргі заманғы цифрлық технологиялардың маңызды салаларының бірі болып табылады. Ол дыбысты жазу, сақтау, өңдеу және талдау процестерін қамтиды. Дыбыстық сигналды өңдеу телекоммуникация, сөйлеуді тану жүйелері, музыкалық индустрия, медицина және мультимедиялық жүйелерде кеңінен қолданылады. Бұл саланың негізгі мақсаты - дыбыстық ақпаратты тиімді, сапалы және сенімді түрде өңдеу.

Төмендегі кестеде қазіргі таңда диктофон жазбаларын өңдеуде қолданылатын танымал нейрондық желі модельдерінің сипаттамасы берілген.

1кесте. ASR модельдерінің салыстырмалы талдауы

Модель атауы	Архитектурасы	Артықшылығы	Кемшілігі
DeepSpeech (Mozilla)	RNN / LSTM	Ашық бастапқы код, оқытуға ыңғайлы	Шулы ортада дәлдігі төмен
Wav2Vec 2.0 (Meta)	Self-supervised CNN	Аз деректермен жақсы жұмыс істейді	Есептеу ресурстарын көп қажет етеді
Whisper (OpenAI)	Transformer	Көптілділік (100+ тіл), жоғары дәлдік	Нақты уақытта (real-time) өңдеуі баяу
Kaldi	HMM / DNN	Параметрлерді терең баптау мүмкіндігі	Қолдану мен орнатудың күрделілігі

Диктофон жазбаларындағы қиындықтар

Дыбыстық сигналды өңдеудің алғашқы кезеңі - аналогтық сигналды сандық сигналға түрлендіру. Бұл процесс аналог-сандық түрлендіргіш (ADC) арқылы жүзеге асады. Түрлендіру кезінде үш негізгі қадам орындалады: дискретизация, кванттау және кодтау. Дискретизация кезінде сигнал белгілі бір уақыт аралығында өлшенеді, кванттау кезінде алынған мәндер белгілі бір деңгейлерге жуықтатылады, ал кодтау кезінде олар сандық кодқа айналады. Бұл кезең дыбыстың сапасына тікелей әсер етеді.

Диктофонды қолдану кезінде нейрондық желілер келесі кедергілерге тап болады:

Диаризация (Speaker Diarization): Жазбада екі немесе одан көп адам сөйлескенде, желінің "кім, не айтты?" дегенді ажырата алмауы. Қазіргі кезде бұл мәселе X-vectors және Clustering әдістерімен шешілуде.

Шалғайлық және акустика: Диктофон сөйлеушіден алыс орналса, дыбыс жаңғырығы (reverb) пайда болады. Мұны жою үшін De-reverberation нейрондық желілері қолданылады.

Диктофон жазбаларындағы негізгі қиындықтардың бірі - дыбыс сапасының төмендігі. Бұл көбінесе сыртқы шудың көп болуымен байланысты. Көшедегі көлік дыбыстары, адамдардың сөйлесуі, жел немесе тұрмыстық құрылғылардың дыбысы негізгі сөйлеу сигналын басып кетуі мүмкін. Сонымен қатар, микрофонның сапасы мен оның дыбыс көзіне қатысты орналасуы да жазбаның анықтығына тікелей әсер етеді.

Диктофондардың техникалық мүмкіндіктері де белгілі бір қиындықтар туғызады. Арзан немесе ескі құрылғыларда жиілік диапазонының тар болуы, төмен биттік тереңдік және дискретизация жиілігінің аздығы дыбыстың табиғилығын төмендетеді. Мұндай шектеулер сөйлеудің кейбір бөліктерінің жоғалуына немесе бұрмалануына әкелуі мүмкін. Сонымен қатар, батареяның тез отыруы немесе жадының шектеулі болуы жазбаның толық сақталмауына себеп болады.

Зерттеу нәтижелері.

Нейрондық желілер - адам миының жұмыс істеу принципіне негізделген есептеу модельдері. Олар кіріс деректерін қабылдап, өңдеп, белгілі бір нәтиже шығара алады. Нейрондық желілер үлкен көлемдегі деректермен жұмыс істеп, күрделі заңдылықтарды автоматты түрде үйрену қабілетіне ие. Осы қасиеттері оларды дыбыстық сигналдарды өңдеу саласында тиімді құралға айналдырды.

Диктофон жазбалары көбінесе шулы ортада, әртүрлі акустикалық жағдайларда жазылады. Мұндай жазбаларды дәстүрлі әдістермен өңдеу қиынға соғады. Нейрондық желілер шуды азайту, пайдалы сөйлеу сигналын бөліп алу және дыбыс сапасын жақсарту үшін қолданылады. Мысалы, терең нейрондық желілер фондық шуды автоматты түрде анықтап, оны басуға мүмкіндік береді. Бұл диктофон жазбаларының сапасын айтарлықтай арттырады.

Автоматты сөйлеуді тану жүйелерінде нейрондық желілер акустикалық және тілдік модельдер ретінде кеңінен қолданылады. Convolutional Neural Network (CNN) дыбыстық белгілерді тануға тиімді болса, Recurrent Neural Network (RNN) және LSTM модельдері уақыт бойынша өзгеретін сөйлеу сигналдарын өңдеуге қолайлы. Ал соңғы жылдары Transformer және end-to-end модельдер ASR жүйелерінің дәлдігін айтарлықтай арттырды. Бұл модельдер диктофон арқылы жазылған сөйлеуді тікелей мәтінге айналдыруға мүмкіндік береді.

Терең нейрондық желілер сөйлеу сигналындағы күрделі заңдылықтарды анықтауға мүмкіндік береді. CNN спектрлік белгілермен жұмыс істеуде тиімді болса, RNN және LSTM уақыттық тәуелділіктерді жақсы модельдейді. Transformer архитектурасы соңғы жылдары ең жоғары нәтижелер көрсетуде.Терең нейрондық желілер сөйлеу сигналындағы күрделі

заңдылықтарды анықтауға мүмкіндік береді. CNN спектрлік белгілермен жұмыс істеуде тиімді болса, RNN және LSTM уақыттық тәуелділіктерді жақсы модельдейді. Transformer архитектурасы соңғы жылдары ең жоғары нәтижелер көрсетуде [2,4].

Қолдану салалары

Диктофонды сөйлеу тану жүйелері автоматты хаттама жасау, дәрістерді мәтінге айналдыру, дауыстық көмекшілер және архивтік аудиоларды цифрландыру салаларында қолданылады.

Диктофон жазбаларындағы сөйлеуді тану үшін нейрондық желілерді қолдану көптеген салаларда өзекті және тиімді шешім болып табылады. Бұл технологиялар уақытты үнемдеп, ақпаратты өңдеу сапасын арттырады. Болашақта нейрондық желілерге негізделген ASR жүйелерінің дамуы диктофон жазбаларын өңдеуді одан әрі жеңілдетіп, қазақ тілінің IT саласында кеңінен қолданылуына мүмкіндік береді.

Қазақ тілінде қолдану ерекшеліктері

Қазіргі цифрлық технологиялардың дамуына байланысты қазақ тілін ақпараттық жүйелерде, соның ішінде дыбыстық жазбалар мен сөйлеуді тану жүйелерінде қолдану өзекті мәселеге айналуда. Алайда қазақ тілінің құрылымдық ерекшеліктері оны автоматты өңдеу мен қолдану барысында бірқатар қиындықтар туғызады.

Қазақ тілін диктофон және ASR жүйелерінде тиімді қолдану үшін арнайы тілдік корпустар құру, диалектілерді ескеретін модельдер жасау және нейрондық желілерді қазақ тіліндегі деректермен оқыту қажет. Сонымен қатар, заманауи end-to-end ASR модельдерін қолдану арқылы диктофон жазбаларынан мәтін алу дәлдігін арттыруға болады. Бұл бағыттағы жұмыстар қазақ тілінің IT саласында кеңінен қолданылуына мүмкіндік береді. Қазақ тілінде сөйлеуді автоматты танудың қолдану ерекшеліктері келесі кестеде берілген [5].

2-кесте. Қазақ тілінде сөйлеуді автоматты танудың қолдану ерекшеліктері

Ерекшелік түрі	**Сипаттамасы**	**Тануға әсері**	**Шешу тәсілдері**
Фонетикалық ерекшелік	Дауысты және дауыссыз дыбыстардың үндестігі	Дыбыстардың өзгеріп айтылуы қателік тудырады	Акустикалық модельді бейімдеу
Агглютинативті құрылым	Сөздерге көптеген қосымшалардың жалғануы	Сөз формалары саны өте көп	Морфологиялық талдау қолдану
Еркін сөз тәртібі	Сөйлемдегі сөздердің орны өзгеруі мүмкін	Тілдік модель күрделенеді	Контекстік нейрондық модельдер
Диалект	Аймақтық айтылу	Танудың дәлдігі	Әртүрлі деректер

айырмашылығы	ерекшеліктері	төмендейді	жинау
Дыбыстық деректердің аздығы	Қазақ тіліндегі корпус көлемі шектеулі	Модельді оқыту қиындайды	Data augmentation әдістері
Интонациялық ерекшелік	Сөйлеу ырғағы мен екпін өзгермелі	Мағынаны анықтау қиындайды	Просодикалық параметрлерді енгізу
Шу әсері	Фондық дыбыстардың болуы	Сигнал сапасы төмендейді	Шу фильтрациясы

Қазақ тілі - агглютинативті тіл, яғни сөз түбіріне жалғаулардың жалғануы арқылы сөз формалары өте көп болады. Бұл тілдік модельдің (Language Model) жұмысын қиындатады. Зерттеулер көрсеткендей, қазақ тілі үшін End-to-End модельдерін қолданып, оны үлкен дыбыстық корпуспен (мысалы, ISSAI Kazakh Speech Corpus) оқыту ең жақсы нәтиже береді.

Қазақ тілінің фонетикалық және морфологиялық ерекшеліктері нейрондық желілерді бейімдеуді талап етеді. Агглютинативті құрылым, сөз формаларының көптүрлілігі және ерекше дыбыстар қазақ тіліне арналған ASR жүйелерін оқытуда қиындықтар тудырады. Дегенмен, жеткілікті көлемдегі аудиодеректер мен мәтіндік корпустарды пайдалану арқылы нейрондық желілер қазақ тіліндегі сөйлеуді тану сапасын жақсарта алады [6].

Қазақ тілінде сөйлеуді автоматты тану ерекшеліктері сурет 2-де көрсетілген.

2-сурет. Қазақ тілінде сөйлеуді автоматты тану ерекшеліктері

Қорытынды

Бұл жұмыста диктофон жазбаларындағы сөйлеуді тану үшін нейрондық желілерді зерттеу және қолдану мәселелері қарастырылды. Диктофон арқылы алынған аудиожазбалар әртүрлі акустикалық ортада, шулы жағдайда және ауызекі сөйлеу формасында жазылатындықтан, оларды өңдеу мен тану күрделі есептердің бірі болып табылады. Осыған байланысты дәстүрлі әдістердің мүмкіндіктері шектеулі екені анықталды.

Зерттеу барысында нейрондық желілердің, әсіресе терең оқытуға негізделген модельдердің, диктофон жазбаларындағы сөйлеуді тану дәлдігін айтарлықтай арттыратыны көрсетілді. Нейрондық желілер шуды азайту, сөйлеу сигналын ажырату және күрделі тілдік заңдылықтарды үйрену қабілетінің арқасында ASR жүйелерінің тиімділігін жоғарылатады. CNN, RNN, LSTM және Transformer сияқты модельдер сөйлеу сигналдарының уақыттық және жиіліктік ерекшеліктерін дәл өңдеуге мүмкіндік береді.

Сонымен қатар, қазақ тіліндегі диктофон жазбаларын тану барысында тілдің фонетикалық және морфологиялық ерекшеліктерін ескеру қажеттілігі анықталды. Агглютинативті құрылым, сөз формаларының көптүрлілігі және ерекше дыбыстар нейрондық желілерді арнайы бейімдеуді талап етеді. Бұл бағытта сапалы аудиодеректер қорын құру және қазақ тіліне арналған тілдік модельдерді дамыту маңызды болып табылады.

Нейрондық желілер диктофон жазбаларындағы сөйлеуді тану мәселесін шешудің тиімді құралы болып саналады. Оларды білім беру, медицина, журналистика, құқық қорғау және IT салаларында кеңінен қолдану уақытты үнемдеп, ақпаратты өңдеу сапасын арттыруға мүмкіндік береді. Болашақта нейрондық желілерге негізделген ASR жүйелерін дамыту қазақ тілінің цифрлық кеңістікте толыққанды қолданылуына және ақпараттық технологиялар саласындағы маңызын арттыруға жол ашады.

Нейрондық желілер диктофон жазбаларын өңдеуде адам еңбегін 80-90%-ға дейін жеңілдетуге қабілетті. Трансформерлік модельдердің пайда болуымен дыбысты тану сапасы жаңа деңгейге көтерілді. Болашақта бұл технологиялар тек мәтінге айналдырып қана қоймай, сөйлеушінің эмоциясын және контекстік мағынасын тереңірек талдайтын болады.

Пайдаланған әдебиеттер тізімі

1. Глубокие нейронные сети для улучшения качества и распознавания речи: систематический обзор // Инженерный журнал Айн Шамс. - 2025.

2. Миао Й., Говайед М., Метце Ф. WFST негізіндегі декодтау арқылы терең RNN модельдерін қолдану: end-to-end сөйлеуді тану жүйесі. - 2015.

3. Хатон Ж.-П. Нейронные сети для автоматического распознавания речи: обзор. - Springer, 1999. - Б. 223-240.

4. Баймаханова А.С. және т.б. Deep Learning алгоритмдерін қолдану технологиясы, 2023.

5. Мамырбаев Ө.Ж., Құрметқан Т. Қазақ сөйлеуін автоматты тану модельдерін зерттеу // Қазақстан ҰҒА хабаршысы. - 2024.

6. Suleimenov Y., Omirzak N. Neural Network-Based Automatic Speech Recognition for Kazakh Language // International Journal of Computing. - 2023.