ҒТАХР 28.23.15

10.51889/2959-5894.2024.87.3.015

Қ.Е. Тұрғанбай[1*], С.Н. Исабаева[2], Е.Ж. Тенизбаев[3], Т.А. Жукова[3], Л.В. Игнашова[3]

[1]Л.Б. Гончаров атындағы қазақ автомобиль-жол институты, Алматы қ., Қазақстан
[2]Нұр-Мүбарак Египет ислам мәдениеті университеті, Алматы қ., Қазақстан
[3]Орталық Азия Инновациялық университеті, Шымкент қ., Қазақстан
[*]*e-mail: kuralai_12@mail.ru*

НЕЙРОНДЫҚ ЖЕЛІЛЕРДІ ТЕРЕҢ ОҚЫТУ АРҚЫЛЫ СӨЙЛЕУШІНІ ТАНУ ЕРЕКШЕЛІКТЕРІ

Аңдатпа

Мақалада жасанды интеллект арқылы терең нейрондық желіні оқыту арқылы сөйлеушіні тану жүйесі сипатталған. Акустикалық сөйлеушінің деректерін жинау сөйлеуші сигналын өңдеуді, акустикалық және тілдік модельдерді оқытуға арналған үлгіні сәйкестендіру алгоритмдерін, нейрондық желілерді, матрицаны көрсету, векторлық кванттау жолдарын қамтиды. Сөйлеушіні тану жүйелерінде қолданылатын терең нейрондық желілер (DNN − Deep Neural Network) сөйлеушіні тану жүйесінің әдістеріне негізделген. Сөйлеушіні тану үшін дәстүрлі әдістерден жаңа терең оқыту архитектурасына көшу және терең оқыту үлгілерін пайдалана отырып, Марков модельдері, акустикалық бірліктерді модельдеу, статистикалық модельдерде жаңа тәсілдерді салыстыру туралы айтылған. Қазіргі автоматты тану жүйелері негізінен Гаусс қоспасының моделіне (GMM − Gaussian Mixture Model) немесе терең нейрондық желіге (DNN − Deep Neural Network) негізделген және осы жұмыс негізінде акустикалық модельдеуді зерттеу үшін ықтималдық сызықтық дискриминанттық талдау (PLDA − Probability Linear Discriminant Analysis) жүргізілген. Эксперимент нәтижелері терең оқытуды қолдана отырып, осы нейрондық желіні модельдеу сөйлеушіні автоматты түрде танудың әртүрлі тапсырмаларында тиімді екенін көрсеткен. Қарастырылып отырған дерекқордағы нәтижелерден сөйлеушінің тануға қосқан үлесі мен байланысты шектеулерді қамтитын мүмкіндіктері сипатталған.

Түйін сөздер: сөйлеуші, тану, конволюция, сәйкестендіру, нейрондық желі, биометрия, аутентификация.

Қ.Е. Тұрғанбай[1], С.Н. Исабаева[2], Е.Ж. Тенизбаев[3], Т.А. Жукова[3], Л.В. Игнашова[3]
[1]Казахский автомобильно-дорожный институт им. Л. Б. Гончарова, г. Алматы, Казахстан
[2]Египетский университет исламской культуры «Нур - Мубарак», г. Алматы, Казахстан
[3]Центрально-Азиатский Инновационный университет, г.Шымкент, Казахстан

ОСОБЕННОСТИ РАСПОЗНАВАНИЯ ГОВОРЯЩЕГО ПОСРЕДСТВОМ ГЛУБОКОГО ОБУЧЕНИЯ НЕЙРОННЫХ СЕТЕЙ

Аннотация

В статье описывается система распознавания говорящего посредством обучения глубокой нейронной сети с помощью искусственного интеллекта. Сбор данных акустического говорящего включает обработку сигнала говорящего, алгоритмы сопоставления моделей для обучения акустическим и языковым моделям, нейронные сети, отображение матриц, пути векторного квантования. Глубокие нейронные сети (DNN − Deep Neural Network), используемые в системах распознавания говорящего, основаны на методах системы распознавания говорящего. Речь идет о переходе от традиционных методов к новой архитектуре глубокого обучения для распознавания говорящего и сравнении новых подходов в Марковских моделях, моделировании акустических единиц, статистических моделях с использованием моделей глубокого обучения. Современные системы автоматического распознавания в основном основаны на модели смеси Гаусса (GMM − Gaussian Mixture Model) или глубокой нейронной сети (DNN − Deep Neural Network), и на основе этой работы

был проведен вероятностный линейный дискриминантный анализ (PLDA − Probability Linear Discriminant Analysis) для изучения акустического моделирования.

Результаты эксперимента показали, что моделирование этой нейронной сети с использованием глубокого обучения эффективно в различных задачах автоматического распознавания говорящего. Из результатов в рассматриваемой базе данных описаны возможности говорящего, которые включают вклад в распознавание и связанные с этим ограничения.

Ключевые слова: говорящий, распознавание, свертка, идентификация, нейронная сеть, биометрия, аутентификация.

K. Turganbay[1], *S.* Issabayeva[2], *Y.* Tenizbayev[3], T. Zhukova[3], L. Ignashova[3]
[1]Kazakh automobile road institute named after L. B. Goncharov, Almaty, Kazakhstan
[2]Egyptian University of Islamic Culture «Nur-Mubarak», Almaty, Kazakhstan
[3]Central Asian Innovation University, Shymkent, Kazakhstan

FEATURES OF SPEAKER RECOGNITION THROUGH DEEP LEARNING OF NEURAL NETWORKS

Abstract

The article describes a speaker recognition system through deep neural network training using artificial intelligence. Acoustic speaker data collection includes speaker signal processing, model matching algorithms for teaching acoustic and language models, neural networks, matrix mapping, vector quantization paths. Deep Neural Networks (DNN − Deep Neural Network), used in speaker recognition systems, are based on the methods of the speaker recognition system. We are talking about the transition from traditional methods to a new deep learning architecture for speaker recognition and comparing new approaches in Markov models, modeling acoustic units, statistical models using deep learning models. Modern automatic recognition systems are mainly based on the Gauss Mixture Model (GMM − Gaussian Mixture Model) or deep neural network (DNN − Deep Neural Network), and based on this work, a probabilistic Linear discriminant analysis (PLDA − Probability Linear Discriminant Analysis) was conducted to study acoustic modeling. The results of the experiment showed that modeling this neural network using deep learning is effective in various tasks of automatic speaker recognition. From the results in the database under consideration, the speaker's capabilities are described, which include a contribution to recognition and related limitations.

Keywords: speaker, recognition, convolution, identification, neural network, biometrics, authentication.

Негізгі ережелер

Сөйлеушіні тану жүйесін құрудың бірыңғай тәсілі қалыптасты, оған төрт негізгі компонент кіреді: сигналды алдын-ала өңдеу, акустикалық модель, тілдік модель және гипотезалар - ізделініп, әр компонентке қатысты тұрақты зерттеулер жүргізілді. Сөйлеушіні тануды машиналық оқытуды қолдана отырып, сөйлеу белгілері мен сөйлеушіні анықтаудың моделі мен алгоритмі жасалынып зерттеу жұмыстары жүргізілді. Сөйлеу сигналдарының белгілерін анықтау процесінде дамыған нейрондық желінің моделі қолданылды; құрастырылған акустикалық корпус сөйлеушіні тану саласында зерттеулер жүргізуге мүмкіндік береді.

Сондықтан сөйлеушінің қазақ тіліндегі оқытылған акустикалық және тілдік модельдерін пайдаланылу қазіргі таңда өзекті мәселе болып отыр.

Кіріспе

Сөйлеу технологиялары операторлардың кейбір функцияларын компьютерге ауыстыруға мүмкіндік береді. Дауысты тану арқылы бүгінде сіз кітаптарды, смс-хабарламаларды оқи аласыз, құжаттар мен бүкіл веб-сайттарды дауыстай аласыз, тілдерді үйренуге көмектесетін интеллектуалды оқыту жүйелерін жасай аласыз. Осындай автоматты сөйлеу сигналын қабылдау жүйелерін жасаушылардың сөзсіз жетістіктеріне қарамастан, дауысты тану жүйелерін құруда әзірлеушілер қолданатын эмпирикалық тәсілді атап өткен жөн. Мәселе мынада, адамның сөйлеуді қабылдау механизмінде әлі де түсініксіз көп нәрсе бар, сондықтан сөйлеуді автоматты түрде тану саласындағы көптеген зерттеушілердің назары осы механизмдерге аударылады. Қысқа мәлімдемелерге арналған заманауи верификация

жүйелерін әзірлеу зерттеудің белсенді бағыты болып табылады, өйткені жүйенің әлеуетті пайдаланушылары қысқа мәлімдемелерді қалайды.

Дауысты тану әдісі дауыстың ерекше сипаттамаларының үйлесімі арқылы адамның жеке басын анықтайды. Алгоритмдерде сөйлеушінің жеке басы туралы шешім қабылданатын негізгі белгілер талданады: дауыс көзі, дауыс жолдарының резонанстық жиіліктері және олардың әлсіреуі, сондай-ақ артикуляцияны бақылау динамикасы. Биометриялық дауыстық аутентификация әдісі қолданудың қарапайымдылығымен сипатталады. Бұл әдіс қымбат жабдықты қажет етпейді, микрофон мен дыбыстық карта жеткілікті. Бірақ дауысты аутентификациялаудың биометриялық әдісін қолданған кезде бірқатар мәселелер туындайды. Ең маңызды мәселелердің бірі – дауысты анықтау сапасы. Қазіргі уақытта адамды дауыспен тану қателігінің ықтималдығы өте жоғары. Дауыстық сигналдан биометриялық параметрлерді дәлірек анықтау үшін жаңа алгоритмдерді әзірлеу қажет. Екінші маңызды мәселе – белгілі құрылғылардың шулы жағдайда тұрақсыз жұмыс істеуі. Маңызды мәселе – бір адамның дауысының әртүрлі көріністерімен дауысты анықтау: дауыс денсаулық жағдайына, жасына, көңіл-күйіне және т. б. байланысты.

Сөйлеушіні тану – үлгіні тану мәселесі. Дауыстық басып шығаруды өңдеу және сақтау үшін қолданылатын әртүрлі технологияларға жиілікті бағалау, жасырын Марков модельдері, Гаусс қоспасы модельдері, үлгіні сәйкестендіру алгоритмдері, нейрондық желілер, матрицалық бейнелеу, векторлық кванттау және шешім ағаштары жатады. Дауыстық іздері бар мәлімдемелерді салыстыру үшін олардың қарапайымдылығы мен өнімділігіне байланысты косинус ұқсастығы сияқты қарапайым әдістер дәстүрлі түрде қолданылады. Кейбір жүйелер сонымен қатар когорттық және әлемдік модельдер сияқты анти-спикер әдістерін қолданады. Спектрлік сипаттамалар негізінен сөйлеушінің сипаттамаларын көрсету үшін қолданылады. Сызықтық болжамды кодтау LPC (Linear Predictive Coding) – сөйлеушіні тану және сөйлеуді тексеру үшін қолданылатын сөйлеуді кодтау әдісі [1].

Стационарлық дауыстық биометриялық аутентификация жүйелерін құрудың негізгі мақсаты пайдаланушының жеке ерекшеліктерін ескеретін және анықталған сөйлеу сегменттерін синхрондайтын, тиімді кодтық фразалардың құрылысын біріктіретін дыбыстық сигналды мерзімінен бұрын өңдеу болып табылады деген қорытынды жасауға болады. бағдарламаны оқыту кезеңі, яғни сөйлеудің осы салаларында фазалық сәйкессіздік болмауы керек.

Пайдаланушы аутентификациядан өткен кезде бағдарлама статистикалық сипаттамалардың барлық түрлерін қорытындылап, таңдалған сөйлеу бөлімдерін нақты жіктеуі керек. Биометриялық сәйкестендіру жағдайында дыбыстарды жіктеуге қабілетті автоматты фрагментатор алдын-ала жасалған сөздіктің және жасалған дыбыстардың дерекқорына қол жеткізудің арқасында әр пайдаланушы үшін бөлек жасалуы керек.

Зерттеу әдіснамасы

Зерттеу жұмысында пайдаланушының негізгі тонусының кезеңін бақылау алгоритмі жасалды. Әр адамның дыбыстық файлды жазу кезінде есептелетін өзіндік дыбыстық кезең параметрлері бар. Қадамдық кезеңнің ұзақтығын математикалық күту көптеген адамдар үшін қолайлы болуы мүмкін болса да, ол жеке сипаттама болып саналады. Негізгі тональды кезеңнің минималды мәні негізгі әйел жынысына және 16 жасқа дейінгі адамдарға тән. Ер дауысында бұл мән айтарлықтай ерекшеленеді. Кейбір еркектерде дауыс негізгі сипатқа ие және олардың кезеңінің орташа мәні орташа адамнан асып түседі.

Сәйкестендіру және тұлғаны тексеру технологияларын қамтитын дауыс биометриясын сөйлеушіні тану технологияларымен шатастыруға болмайды. Сөйлеуді тану технологиясын қолдана отырып, ол сөйлеушіні дауысынан тани алады. Сондықтан қауіпсіздік саласында сөздерді тану технологияларын қолдану шектеулі. Керісінше, адамды дауыспен анықтау және тексеру технологиялары адамның сол күйінде көрінетінін растау қажет болған кезде қолданылады:

- деректерді енгізу;
- математикалық алгоритмдер;
- есептеу қуаты.

Кіріс дерекқорда сақталған биометриялық үлгі немесе дауыстық басып шығару. Биометриялық үлгінің сапасы көбінесе енгізу құрылғысының түріне (мысалы, кәсіби микрофон немесе ұялы телефон) және қоршаған ортаға (шулы көше немесе тыныш бөлме) байланысты. Дауыстық басып шығару сапасын автоматты түрде анықтайтын, содан кейін жақсы үлгіні алу үшін оны шудан тазартатын технологиялар бар.

Биометриялық жүйелердегі алгоритмдер алынған дауыстық басып шығаруды мәліметтер базасындағы үлгімен салыстыру үшін қолданылады. Алгоритм неғұрлым жетілген болса, салыстыру нәтижесі соғұрлым дәл болады. 1-суретте байланыс орталықтарының бірінде дауысты тану жүйесін қолдану мысалы көрсетілген.

Сурет 1. Тұлғаны тану жүйесін қолдану

Бүгінгі күні бірнеше биометриялық технологияларды біріктіретін жүйелер әзірленді, мысалы, дауыс пен саусақ ізі арқылы тұлғаны тексеру технологиясы. Екі биометриялық технологияның үйлесімі бір технологияның екіншісінің кемшіліктерін өтеуге мүмкіндік береді және керісінше, сонымен қатар операторға қауіпсіздік деңгейін бақылауға мүмкіндік береді.

Бұрын дауыс биометриясы саусақ іздері, бет пішіні және көз қарашығы арқылы анықтау және тексеру сияқты биометриялық әдістерге мүмкіндік берді. Дегенмен, жаңа алгоритмдер мен компьютердің деректерді өңдеудегі жоғары өнімділігі дауысты тану дәлдігін айтарлықтай жақсартуға мүмкіндік берді, бұл оны дәстүрлі емес, ыңғайлы дауыс биометриясын анықтау және тексеру әдістерінің күшті бәсекелесі етеді.

Қазіргі уақытта қол жетімді биометриялық технологиялардың ішінде дауыстық биометрия ең үнемді және пайдаланушыға ыңғайлы, сондықтан дауыстық биометриялық шешімдер жақын арада барлық жерде болады. Магниттік карталарды жоғалту немесе ұрлату, PIN кодтарын ұмыту кезінде қолданылатын арнайы сканерлеу құрылғылары көп ақша талап етеді. Керісінше, дауыстық биометрия кез-келген уақытта кез-келген жерде сәйкестендіруге мүмкіндік береді. Сізге тек ұялы немесе қалалық телефонды немесе микрофонды пайдалану керек.

Ерінмен сөйлеуді тану адамдар үшін өте қиын міндет. Қолда бар зерттеулерге сәйкес, нашар еститіндер 30 бір буынды сөздердің шектеулі жиынтығынан тек 17-12% және 30 көп буынды сөздердің 21-11% ғана. Сондықтан автоматты танудың шулы ортада, жеке сәйкестендіруде, сондай-ақ медициналық мақсаттарда пайдалану үшін үлкен практикалық әлеуеті бар.

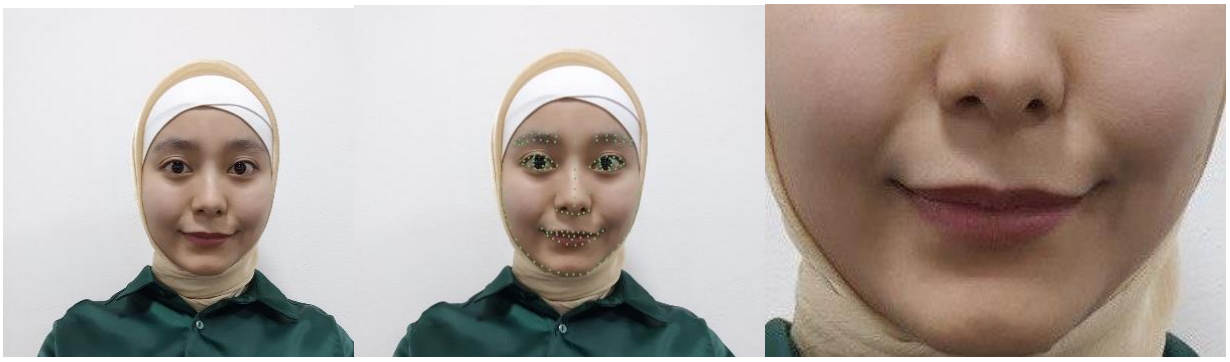
Зерттеу нәтижелері

Зерттеу барысында 34 сөйлеуші ағылшын тілінде сөйлейтін 1000 сөйлемнен тұратын бейне жазбадан тұратын деректер жинағы пайдаланылды.

Бұл ретте кейбір бейнежазбалар бүлінген, сүзілгеннен кейін әрқайсысы 3 секундқа созылатын 32904 бейнежазба қалады. Әрбір бейне 75 кадрға бөлінген. Деректер жаттығу (26304 жазба) және кейінге қалдырылған (6600 бейне жазба) үлгілеріне бөлінеді.

LipNet нейрондық желісі бейне тізбегін өңдеу арқылы бүкіл сөйлем деңгейінде ерін арқылы сөйлеуді сәтті таныған әлемде бірінші жүйе болды.

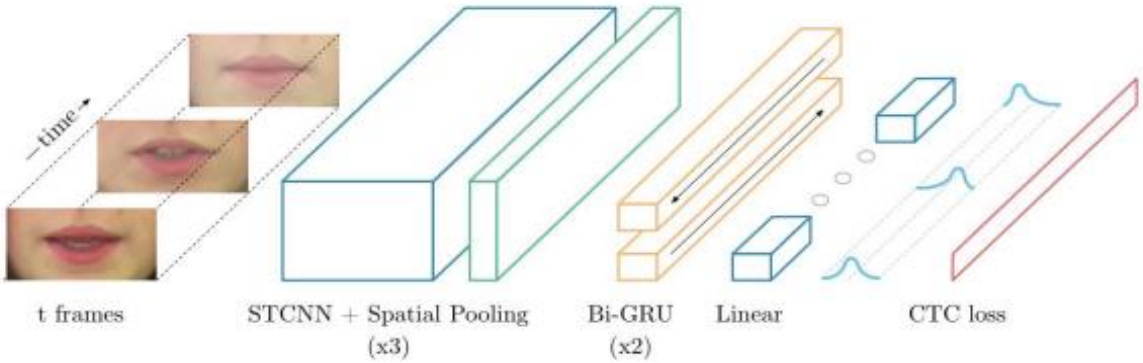
Беттің қажетті бөлігін бөлектеу үшін dlib кітапханасынан негізгі нүкте детекторы қолданылады [2] және алынған нүктелер беттің белгілі бір аймақтарына сәйкес келеді. Осы зерттеу жұмысында 2 суретте еріннің артикуляциялық қозғалысы ғана қолданылатындықтан, бұл тек ерінге сәйкес келетін бет аймағын, атап айтқанда 48-67 нүктелерін пайдалануға мүмкіндік береді.



а) б) в)

Сурет 2. Ерін аймағын таңдау:
(а) - бастапқы сурет, (б) - негізгі нүктелерді алу, (в) – түрлендіру нәтижесі

Кадрлар ағынын мәтінге аудару үшін LipNet – LSTM типті қайталанатын нейрондық желі (long short-term memory) архитектурасының нейрондық желісі қолданылады.



Сурет 3. Vipnet нейрондық желісінің архитектурасы

Біз зерттеуден алынған мәліметтер жиынтығын қолданамыз [3]. Vipnet нейрондық желісі кеңістіктік-уақыттық конволюциялық қабаттарды (STCNN – спатиотемпоральды конволюциялық нейрондық желі), басқарылатын қайталанатын блоктарды (GRU), сондай-ақ толық байланысқан қабатты пайдаланады.

Нейрондық желіні оқыту қадам өлшемі бар Adam оңтайландыру әдісі арқылы жүзеге асырылды, жоғалту функциясы ретінде CTC (Connectionist Temporal Classification) loss және нейрондық желінің мәтінге шығуын декодтау үшін - CTC beam decoder қолданылады.

Болжау сапасын бағалау үшін (кесте) формула бойынша есептелген WER (Word Error Rate) көрсеткіші 1-ші формулада пайдаланылды:

$$WER = \frac{S + D + I}{N} \quad (1)$$

мұндағы S-ауыстыру саны, D-жою саны, I-кірістіру саны, N – сөйлемдегі сөздер саны. I-вектор негізіндегі құрылым мәтіннен тәуелсіз динамикті танудың соңғы деңгейін анықтады. I – вектор Гаусс қоспасының үлгісінен (GMM − Gaussian Mixture Model) немесе терең нейрондық желіден (DNN – *Deep Neural Network*) [4] алынады, ал бэкенд үшін ықтималдық сызықтық дискриминанттық талдау (PLDA− *Probability Linear Discriminant Analysis*) кеңінен қолданылады. Айтылымның қысқа I – векторларының вариациясын модельдеудің бірқатар әдістері болды. Пайдаланушылардың үлкен саны ажыратуға үйретілген нейрондық желіні пайдаланады, бекітілген өлшемдегі кірістірілген дауыстарды жасайды және PLDA бағалау үшін енгізілген дауыстар пайдаланылады. DNN1 -екі сатылы әдіс: алдын-ала дайындық және дәл баптау. Бақыланатын DN моделінің жақсы инициализациясын табу үшін алдымен автоэнкодерді қолдана отырып, қысқа және ұзын векторлардың бірлескен көрінісін үйретеміз. Бұл әдісті DNN1 деп белгілейік. DNN2 -бір сатылы әдіс: жартылай бақыланатын оқыту. Екі сатылы әдіс, алдымен автоэнкодерді қолдана отырып, бірлескен көріністі үйрету керек, содан кейін бақыланатын дисплейді үйрету үшін дәл баптау керек. Қалпына келтіру қатесін азайту үшін автоматты кодер арқылы басқарылатын дисплейді бірлесіп үйрете алатын [5] бірыңғай жартылай басқару ортасын енгізу қажет. Бұл әдіс DNN2 ретінде белгіленеді. Біз алдыңғы бөлімде айтылғандай бірдей автоэнкодер платформасын қабылдаймыз, онда кодер мен декодер бар, бірақ мұнда кодер үшін WS айту арқылы кіріс қысқа I – вектор болып табылады.

Алынған модельді болжау үшін wer орташа бағасы – 22,7 %, бұл адамдардың сөйлеуді болжауын бағалаудан әлдеқайда аз – 47,7 % [6]. Осылайша, берілген жүйе адамдарға қарағанда бейне ағыны арқылы сөйлеуді тануға мүмкіндік береді.

Кесте 1. Тану нәтижелерінің мысалы

Түпнұсқа мәтін	*Болжамды мәтін*	*WER*
place red by m seven please	*place red by m seven please*	*0,0*
place green at t seven now	*place green at d seven now*	*0,1667*
place green in t six again	*place green at d six again*	*0,3333*
set blue by q seven now	*set blue by u seven now*	*0,1667*
place red in z six soon	*place red i c six soon*	*0,3333*
place white by n two soon	*place red by h two soon*	*0,1333*

Іске асырылған жүйені, мысалы, трахеостомия (тыныс алу жолдарының бітелуі кезінде түтікті трахеяға енгізу) немесе қатерлі ісікпен күресу үшін көмейді алып тастау сияқты алдыңғы ауруларға, жарақаттарға және медициналық манипуляцияларға байланысты сөйлеу дыбыстарын шығару қабілетін жоғалтқан адамдар қолдана алады. Сондай-ақ, мұндай жүйені адамның дауысын жазу мүмкін емес жағдайларда сөйлеуді тану үшін қолдануға болады, мысалы, шу деңгейінің жоғарылауы, адамның орнатылған құрылғыдан қашықтығы [7].

Спектрлік сәйкессіздікке байланысты шу болған кезде сөзді тану өнімділігі күрт төмендейді. Спектрлік қысу - мүмкіндік доменіндегі оқу және сынақ деректері арасындағы сәйкессіздікті азайтуға арналған тиімді мүмкіндіктерді шығару әдісі. Кәдімгі MFCC мүмкіндіктерін алуда динамикалық диапазонды азайту үшін Mel сүзгі банкінің энергияларына логарифм функциясы қолданылады. Түбірлік цестральды талдау логарифмдік функцияны тұрақты түбір функциясымен алмастырады және RCC коэффициенттерін береді. RCC коэффициенттері шуға жақсырақ төзімділікті көрсетті. RCC әдісінде қысылған сөйлеу спектрі (2) формулада көрсетілгендей есептеледі:

$$P_C(m) = P(m)^{\gamma} \quad , \quad 0 \le \gamma \le 1 \tag{2}$$

мұндағы қысу коэффициенті жиілік жолағына тәуелді және біркелкі емес спектрлік қысу деп аталады.

Шу тұрақты болса да, бұл параметрді басында калибрлеу керек. Дәстүрлі әдістерде тану нәтижелерінің орнына SNR сәйкес әрбір жолақ үшін қысу коэффициенті реттеледі, бұл сенімдірек болып көрінеді. SNR негізіндегі тәсілдерде тану кезеңінен өтемақы кезеңіне кері байланыс жоқ және оларға SNR бағалаушысы қажет екені анық.

Осылайша, оның өнімділігі SNR бағалауының дәлдігіне байланысты. Дегенмен, сөйлеуді тану классификациялық мәселе болып табылады және өтемақы әдістерінің параметрлеріндегі кез келген түзету тану көрсеткіштерінің жақсаруына әкелетіні орынды болып көрінеді [8].

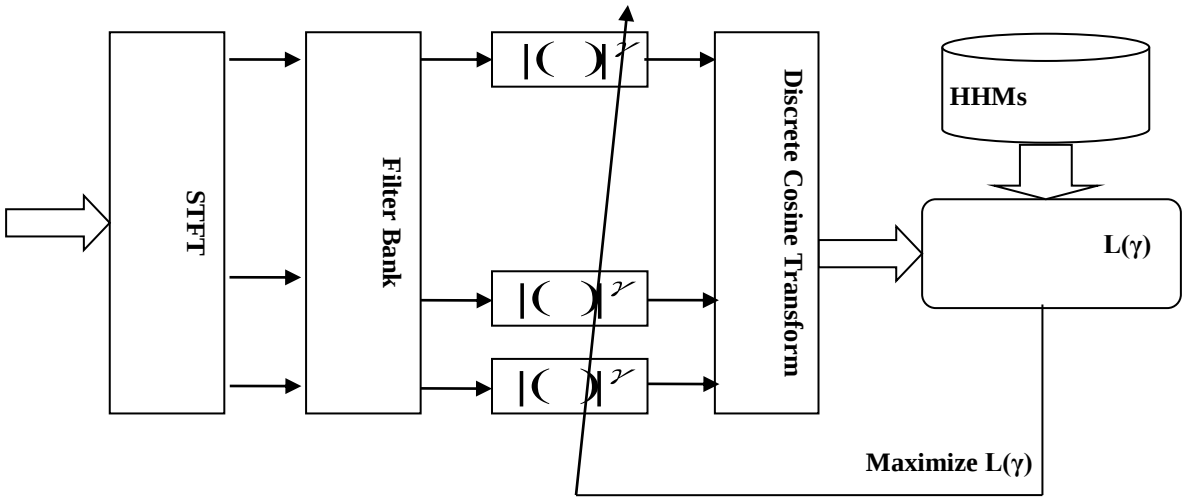
Бұл зерттеу алдыңғы қатардағы сөйлеушіні тану жүйелерінде біркелкі емес спектрлік қысуды қолданудың жаңа құрылымын ұсынады.

Сығымдау коэффициентін реттеу процесіне сөйлеуді тану жүйесін қосу арқылы тану жылдамдығы одан әрі жақсаратынын көрсетеміз.

Бұл сұлбаны жүзеге асыру үшін транскрипциясы берілген оператор пайдаланылады және қысу коэффициенті мен дұрыс модельдің ықтималдығы арасындағы байланыс тұжырымдалады, ұсынылған бұл әдіс екі фазадан тұрады: бейімдеу және декодтау.

Бейімдеу кезеңінде қысу коэффициенті дұрыс транскрипцияның акустикалық ықтималдығын барынша арттыру негізінде реттеледі, ал декодтау кезеңінде бұл оңтайландырылған қысу коэффициенті барлық кіріс сөзге қолданылады.

Сөйлеушіні тану құрылғысындағы MFCC [9] – сөйлеушіні анықтау мәселесін шешудің дәстүрлі тәсілі - мүмкіндіктерден шу әсерлерін жоюдың ұсынылған әдісі 4- суретте көрсетілген.



Сурет 4. Ұсынылған құрылымның схемасы

Бұл формулалар мүмкіндікті алу алгоритміне және акустикалық бірлік үлгісіне байланысты. Бұл жұмыста MFCC алгоритмі және сәйкесінше әр күйдегі Гаусс қоспасы бар жасырын Марков моделі мүмкіндіктерді алу және акустикалық бірлік модельдеу үшін пайдаланылады [9]. Статистикалық модельге негізделген сөйлеуді тану жүйелері күшейтілген сөйлеу сигналынан алынған байқалатын мүмкіндік векторларын тудыруы мүмкін акустикалық бірліктердің тізбегін табады. Бұл бақыланатын мүмкіндіктер кіріс сөйлеу сигналының да, спектрлік қысу векторының да функциясы болып табылады. Сөйлеушінің танушысы Байестің оңтайлы жіктеу (3) формуласына негізделген ең ықтимал гипотезаны алды:

$$\hat{w} = \arg\max_{w} P(Z(\boldsymbol{\gamma})|w)P(w) \tag{3}$$

мұнда байқалатын мүмкіндік векторлары γ -спектрлік қысу векторының функциясы болып табылады. $P(Z(\gamma)\,|\,w)$ және $P(w)$ осы формулада және сәйкес акустикалық және тілдік көрсеткіштер.

Акустикалық немесе экологиялық бейімделу әдістері сияқты, γ түзету үшін бізге белгілі транскрипциясы бар кейбір бейімделу деректері қажет. Айтылымның дұрыс w_C транскрипциясы белгілі деп есептейміз. Осылайша, мәнді елемеу мүмкін, себебі Р(w_C). мәніне қарамастан тұрақты. Сонда (4) теңдеуді мыналарды ескере отырып γ көбейтуге болады:

$$\hat{\gamma} = \arg\max_{\gamma}(P(Z(\gamma)|w_C)) \tag{4}$$

HMM негізіндегі сөйлеуді тануда акустикалық ықтималдық күйлердің жалғыз ықтимал тізбегі арқылы бағаланады. Егер S_C комбинациялық HMM ішіндегі барлық күй ретін көрсетсе және жалғыз ең ықтимал күй тізбегін көрсетсе, онда ең үлкен γ ықтималдық бағалауы келесідей жазылады:

$$\hat{\gamma} = \arg\max_{\gamma, s\in S_C}\left\{\sum_i \log(P(z_i(\gamma)|s_i)) + \sum_i \log(P(s_i|s_{i-1}, w_C))\right\} \tag{5}$$

(5) тармағына келетін болсақ, дұрыс транскрипцияның акустикалық ықтималдығын алу үшін бірізділік пен күй параметрлеріне қатысты максималды көбейту керек. Бұл бірлескен оңтайландыру итеративті түрде жасалуы керек. Шулы сөйлеу спектрлік алу сүзгісіне $Z(\gamma)$ беріледі және белгілі берілген мүмкіндік векторы γ шығарылады. Оңтайлы күй тізбегі $s = \{s_1,...,s_t\}$ дұрыс транскрипцияны w_C ескере отырып (6) көмегімен есептеледі.

$\hat{S}$ күйлер тізбегін Витерби алгоритмі арқылы оңай есептеуге болады. Күйлердің белгілі $\hat{S}$ тізбегін ескере отырып, біз мыналарды $\hat{\gamma}$ тапқымыз келеді:

$$\hat{\gamma} = \arg\max_{\gamma}\left\{\sum_i \log(P(z_i(\gamma)|\hat{s}_i))\right\} \tag{6}$$

Күйлердің берілген тізбегі үшін оптималды есептеудің жабық түрі бар шешім жоқ, сондықтан біз сызықты емес оңтайландыруды қолданамыз.

Алдымен пайдаланушы априорлы белгілі транскрипциясы бар айтылымды айтады, содан кейін айтылым бастапқы параметрлермен бекітілген спектрлік қысу блогы арқылы беріледі. Осыдан кейін Витерби алгоритмі арқылы күйлердің ең ықтимал тізбегі құрылады. Содан кейін күйлердің реттілігін ескере отырып, оптималды спектрлік қысу векторы жасалады. Тану арқылы берілген оңтайландырылған сүзгіні пайдаланып валидация жиынында орындалады, егер қажетті қателік деңгейі орындалса, алгоритм тоқтатылады, әйтпесе күйлердің жаңа тізбегі бағаланады.

Біз зерттеуден алынған мәліметтер жиынтығын қолданамыз [10].

Дискуссия

Нейрондық желілер негізінде сөйлеушіні автоматты түрде моделі мен алгоритмі жасалды. Бұл мақалада сөйлеушіні автоматты түрде тануда акустикалық және тілдік модельдерді оқытуға арналған үлгіні сәйкестендіру алгоритмдеріне салыстырмалы талдау жасалды. Онда осы модель төрт нейрондық желіден тұрады: біріншісі мәтінді фонемаларға (g2p) түрлендіреді, екіншісі - біз клондағымыз келетін сөйлеушіні таңбалар векторына (сандарға)

түрлендіру. Үшіншісі - алғашқы екеуінің шығуы негізінде Mel спектрограммасын синтездейді. Соңында, төртінші спектрограммалардан дыбыс алады. Бұл модель үшін сөйлеушіні нейрондық желілеріндегі жасанды нейрондық желілерді қолдана отырып, жіктеу алгоритмдері дауысты тану және белгілерді анықтау мәселелерінде жақсы нәтиже көрсетті. Эксперимент екі нейрондық модельді қамтиды: сөйлеушіні тану мәселесі/міндеті үшін қарастырылған алгоритмдерді қолдандық, салыстырмалы талдау жүргізілді. Салыстырмалы талдау мен тәжірибелердегі ең жақсы көрсеткіштер тірек векторлары мен көп қабатты персптрондарды қолдану арқылы алынды. Нәтижесінде 5904 парадиофайл алынды. Сөйлеушінің әр дауысы жазбада тіркелген сөйлеуші атымен белгіленеді. Алынған жиынтықтың мөлшері-1480x5904. Көрсетілгенді визуализациялау үшін 5904 белгісі бар векторлық кеңістіктің және екі-үш өлшемді кеңістіктің мөлшерін азайту үшін негізгі компоненттер әдісі қолданылады.

Берілген өлшем 1479 белгісіне дейін азайған кезде дисперсия 100% сақталады. Алайда, ұсынылған жіктеу модельдерімен және стандарттаушылармен жүргізілген тәжірибе көрсеткендей, бұл мөлшердің азаюы жіктеудің дәлдігіне сындарлы әсер етеді. Енді Robust scaler – 0,93 көп қабатты перцептрон әдісімен масштабтауда ең үлкен дәлдік көрсетіле бастады, ал Standard scaler және MaxAb Scaler әдістерімен масштабтауда олар өз нәтижелерін 0,90-дан 0,9-ға дейін жақсарғанымен, тірек векторлары әдісі екінші қатарға ауыстырылды.

Егер жіктеудегі негізгі компонент әдісін қолдана отырып, сөйлеу белгілерінің өлшемін 1479-ға дейін төмендетсе жіктеу дәлдігі өзгереді. SVC және MLP Classifier алгоритмдерін салыстырмалы талдаудың мақсаты спикерді тану тапсырмасын орындау үшін жіктеу алгоритмдерінің әсер ету дәрежесін анықтау болды.

Сөйлеушінің жинағында жүргізілген эксперименттер осы міндеттердің болашағы туралы мәселе қоюға болатындығын көрсетті.

Қорытынды

Бұл зерттеу жұмысында сөйлеушіні автоматты түрде тану тапсырмалары бойынша терең нейрондық желілерге негізделген акустикалық моделдеу қарастырылған. Эксперимент кезінде сөйлеушіні автоматты түрде тану тапсырмаларына сай корпус әзірленді және фонетикалық репрезентативті мәтінді, ауызекі сөйлеудегі айтылымды, яғни сөйлеушіні тану әдістері қарастырылды, нәтижесінде сөйлеушінің сөз транскрипциясының оңтайлы нұсқаларын тани алатын таңдау әдісі ұсынылды.

Зерттеу барысында, сөйлеушінің алдын-ала өңделген тапсырмалары қарастырылды, сөйлеу сигналынан шуды жою әдістері мен алгоритмдері, сөйлеушіні тану үшін спектрлік уақытты қысу жолдарының мүмкіндіктері анықталды.

Зерттеу нәтижесінде автоматты түрде сөйлеушіні тану жүйесінде құрастырылған корпуста дауысты орысша-қазақша аудармасында компьютерлік стенографиялық жазуда, компьютердің дауысын басқаруда, роботтандырылған және автоматтандырылған жүйелерде адамдарға пайдалануға тиімді болатыны анықталды.

Пайдаланылған дереккөздер тізімі

[1] Dehak, N., Кенни, P.J., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Фронтальный факторный анализ для проверки докладчика. IEEE Trans. Аудио Речь Ланг. Процесс. – № 19 (4), с. 788–798.

[2] Li, L., Wang, D., Zhang, C., Zheng, T.Z., 2016. Улучшение распознавания коротких высказываний за счет моделирования классов речевых единиц. IEEE Trans. Аудио Речь Ланг. Процесс. – № 24 (6), с. 1129–1139.

[3] Mekebayev N., Tuyebaev Ch., Sabrayev K., Yerkebay A. Researchof acoustic and linguistic modeling based on repetitive neural networks for speech recognition of children // Bulletin of physics & mathematical sciences. No1(77), 2022, https://doi.org/10.51889/2022-1.1728-7901.16, No1(77), 2022, 119-126

[4] О.Ж. Мамырбаев, М. Отман, А.Т. Ахмедиярова, А.С.Кыдырбекова, Н.О.Мекебаев «Голосовая верификация с использованием і-векторов и нейронных сетей с ограниченными данными обучения» Бюллетень Национальной академии наук РК Выпуск: 3, 2019, с.36-43

[5] Behbahani, Yasser Mohseni, Babaali, Bagher, Turdalyuly Mussa Persian sentences to phoneme sequences conversion based on recurrent neural networks // Open Computer Science. – 2016. - Issue-6. - P. 219–225.

[6] Bagher BabaAli, Waldemar Wojcik, Oken Mamyrbayev, Mussa Turdalyuly, Nurbapa Mekebayev. Speech Recognizer-Based Non-Uniform Spectral Compression for Robust MFCC Feature Extraction // Przeglad Elektrotechniczny. – 2018. - № 6 (94). – P. 90-93.

[7] Мамырбаев О.Ж., Мекебаев Н.О., Тұрдалыұлы М. Генетикалық алгоритм көмегімен сөйлеуді автоматты тануғы гендерлік сәйкестендіру // Алматы энергетика және байланыс университетінің хабаршысы. – 2018. – спец. вып. – Б. 120-129.

[8] Мамырбаев О.Ж., Тұрдалыұлы М., Мекебаев Н.О. Система распознавания слитной казахской речи на основе глубоких нейронных сетей // Вестник Алматинского университета энергетики и связи. – 2018. – спец. вып. – С. 130-135.

[9] A. Toleu, G. Tolegen, A. Makazhanov, (2021). «Character-aware neural morphological disambiguation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics 2», Association for Computational Linguistics, Vancouver, Canada. 666–671.

[10] Orken Mamyrbayev, Mussa Turdalyuly, Nurbapa Mekebayev, Kuralay Mukhsina, Alimukhan Keylan, Bagher BabaAli, Gulnaz Nabieva, Aigerim Duisenbayeva and Bekturgan Akhmetov. (2018) «Continuous Speech Recognition of Kazakh Language AMCSE». International Conference on Applied Mathematics, Computational Science and Systems Engineering. - Rome, Italy. v24 - 2019

References

[1] Dehak, N., Kenny, P.J., Dean, R., Dumouchel, P., Ouellet, P., (2011) Frontal factor analysis to verify the speaker. IEEE Trans. Lang's Audio Speech. Process. No. 19 (4), 788-798.

[2] Li, L., Wang, D., Zhang, C., Zheng, T.Z., (2016) Improving the recognition of short utterances by modeling classes of speech units. IEEE Trans. Lang's Audio Speech. Process. No. 24 (6), 1129-1139.

[3] Mekebayev N., Tuyebaev Ch., Sabrayev K., Yerkebay A. (2022) Researchof acoustic and linguistic modeling based on repetitive neural networks for speech recognition of children // Bulletin of physics & mathematical sciences. No1(77), 2022, https://doi.org/10.51889/2022-1.1728-7901.16, No1(77), 119-126

[4] O.Zh. Mamyrbaev, M.Otman, A.T.Ahmedijarova, A.S.Kydyrbekova, N.O.Mekebaev (2019) «Golosovaja verifikacija s ispol'zovaniem i-vektorov i nejronnyh setej s ogranichennymi dannymi obuchenija» [Voice verification using i-vectors and neural networks with limited training data]. Bjulleten' Nacional'noj akademii nauk RK Vypusk: 3, 36-43. (In Russian)

[5] Behbahani, Yasser Mohseni, Babaali, Bagher, Turdalyuly Mussa (2016) Persian sentences to phoneme sequences conversion based on recurrent neural networks // Open Computer Science. Issue-6. 219–225.

[6] Bagher BabaAli, Waldemar Wojcik, Oken Mamyrbayev, Mussa Turdalyuly, Nurbapa Mekebayev. (2018) Speech Recognizer-Based Non-Uniform Spectral Compression for Robust MFCC Feature Extraction // Przeglad Elektrotechniczny. № 6 (94), 90-93.

[7] O.Zh. Mamyrbaev, M., N.O.Mekebaev., M. Turdalyly. (2018) «Geneticalyk algorithm komegimen soyleudi automatty tanudagi genderlik saykestendiru» [Gender identification in automatic speech recognition using a genetic algorithm]. Almaty energetika zhane baylanys universitetinin khabarshysy] special issue, 120-129. (In Kazakh)

[8] O.Zh. Mamyrbaev, M. Turdalyly, N.O. Mekebaev. (2018) «The recognition system of the unified Kazakh speech based on deep neural networks» [Kazakh speech recognition system based on deep neural networks]. Bulletin of the Almaty University of Energy and Communications], special issue,130-135. (In Russian)

[9] A. Toleu, G. Tolegen, A. Makazhanov, (2021). «Character-aware neural morphological disambiguation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics 2», Association for Computational Linguistics, Vancouver, Canada. 666–671.

[10] Orken Mamyrbayev, Mussa Turdalyuly, Nurbapa Mekebayev, Kuralay Mukhsina, Alimukhan Keylan, Bagher BabaAli, Gulnaz Nabieva, Aigerim Duisenbayeva and Bekturgan Akhmetov. (2018) «Continuous Speech Recognition of Kazakh Language AMCSE». International Conference on Applied Mathematics, Computational Science and Systems Engineering. Rome, Italy. v24, 2019