әл-Фараби атындағы Қазақ ұлттық университеті

ҚР БҒМ ҒК Ақпараттық және есептеуіш технологиялар институты

ӘОЖ 004.934

Қолжазба құқығында

МЕКЕБАЕВ НУРБАПА ОТАНОВИЧ

Сөйлеулерді тану есептерінде машиналық оқытуды қолданып белгілерді анықтау және өңдеу алгоритмдерін зерттеу және құру

6D060200-Информатика

Философия докторы (PhD)
Дәрежесін алу үшін дайындалған диссертация

Ғылыми кеңесшілер:
Қалимолдаев М.Н.
ҚР ҰҒА академигі, ф.-м.ғ.д., профессор
Andrzej Smolarz
т.ғ.д., профессор

Қазақстан Республикасы
Алматы, 2020

МАЗМҰНЫ

НОРМАТИВТІК СІЛТЕМЕЛЕР

Бұл диссертация келесі стандарттарға сәйкес сілтемелер қолданылды:

ҚР МЖМБС 5.04.034 – 2011 «Қазақстан Республикасының Мемлекеттік жалпыға міндетті білім беру стандарты. Жоғары оқу орнынан кейінгі білім. Докторантура». Негізгі ережелер ҚР білім және ғылым министрімен бекітілген. «17» маусым 2011 ж. №261. Астана 2011.

«Диссертацияларды және авторефераттарды рәсімдеу бойынша нұсқаулық», ҚР БҒМ, Жоғары аттестаттау комитеті, Алматы, 2004.

МЕСТ 7.1-2003. Библиографиялық жазба.

БЕЛГІЛЕУЛЕР МЕН ҚЫСҚАРТУЛАР

ПДБА – Пайдаланушыларды дауыс бойынша анықтау
СҚАТ – Сөйлеуді қабылдаудың автоматтық теориясы
SALT – Speech Application Language Tags
MFCC – Мел кепстралды коэффициенті
ЖММ – Жасырын Марковтық моделі
DTW – Dynamic Time Warping
GMM – Caussian Mixture Model
HMM – Hidden Markov Model
АСМ – Ақпараттық сәйкестік толықтығы
ЗСӘ – Залалсыздандыру сүзгі әдісі
ҚСБ – Қарапайым сөйлеу бірліктері
АКМ – Автокорреляциялық матрица
СТҚ – Спектрлік тығыздық қуаттылығы
АСТ - аналогты-сандық түрлендіргіш
MLP – Multilayer Perceptron
DNN – Deep Neural Networks
CNN – Convolutional Neural Networks
RNN – Recurrent Neural Networks
LSTM – Long Short Term Memory
GRU - Gated Recurrent Unit
ASR – Automatic Speech Recognition
CTC – Connectionist Temporal Classification
KNN – K-nearest-neighbors
SDK – Software Development Kit
TTS – Text-to-Speech
КП – Көпқабатты перспетрон
ГШ – Гаустық шуыл
NN – нейрондық желілер
АО – Ақпараттық орталық
АР – Авторегрессия
VQ – Vector Quantization
NB – Naive Bayes
ҚР – Қазақтан Республикасы
БҒМ – Білім және ғылым министрлігі
ҒК – Ғылым комитеті
Гендерлік – ер және әйел дауысын анықтау

КІРІСПЕ

Зерттеу тақырыбының өзектілігі. Қазіргі уақытта информатиканың өзекті мәселелерінің бірі - сөйлеуді тану проблемасы. Компьютерлік жүйелерді пайдалану тиімділігі тікелей осыған байланысты, өйткені сөйлеу - бұл адамның қарым-қатынас жасаудың ең кең таралған және табиғи құбылыс болып саналады және ол ақпаратты енгізу және мобильді жүйелерді басқару процесін едәуір жылдамдатады. Ақпараттық технологиялар қарқынды дамып келеді және ақпарат алмасуда кеңінен қолданылуда. Осыған байланысты сөйлеуді танудың дамуы маңызды рөл атқарады.

Күнделікті өмірде тіл - адамзат жаратылысының табиғи көрінісі. Ғылым мен техниканың дамуы барысында ғалымдар мен инженерлер көптеген жылдар бойы пайдаланушы мен машина арасындағы ауызша байланыс мәселесін зерттеп келе жатқаны бәріне мәлім.

Көптеген компаниялар мен жеке әзірлеушілер сөйлеуді тану технологияларын жасауда біршама жетістіктерге жеткенін мойындау керек, бірақ олардың Қазақстанда әлі де кең қолданыста еместігін мойындау керек. Бұл сөйлеушінің сөйлеу мәнерінің ерекшеліктеріне және дыбыстық кедергілердің болуына байланысты.

Осыған байланысты, мәселені шешудің ең өзекті міндеті - сөйлеуді автоматты тануда сөйлеушіні анықтау.

Қолданыстағы сөйлеуді тану жүйелерін оңтайландыру адамның компьютермен өзара қарым-қатынасының тиімділігін едәуір жеңілдетуге және арттыруға мүмкіндік береді. Сондай-ақ сөйлеуді тану жүйелерін қолдану құқық қорғау қызметінің жұмысында аса үлкен маңызға ие.

Қарастырылған мәселе бойынша зерттеулердің өзектілігі қазіргі жүйелердегі бірқатар факторлармен күрделенеді - тілдердің әр түрлі құрылымы, шуылмен бірге өңделетін сөйлеу сигналдар нәтижелерінің төмендігімен және нәтиженің дикторға тәуелділігімен, жүйелердің жұмыс жылдамдығының жоғары болмау мәселелерімен түсіндіріледі.

Сөйлеуді танудың қазіргі жүйелері, негізінен жасырын Марков модельдеріне (ЖММ) негізделіп құрылған, олар сөйлеудегі бір фонеманың екінші бір фонемаға үйлесу ықтималдығын анықтайды. Осылайша Гаусс құрамасы (ГҚ) [1] арқылы белгілердің ықтимал бөліктерге бөлуін модельдеу көмегімен, бақыланып отырған сигналдың вариативтігін қамтамасыз етеді. Бұл әдісті 1989 жылы Лоуренс Робимер ұсынды және ол ұзақ уақыт сөйлеу сигналын модельдеуге негіз болып келді. Deep Belief Networks [2] өзінің қарқынды дамуы арқасында ЖММ-не балама болуда және тану үдерісінде жоғарғы дәлдікті қамтамасыз етіп отыр.

Л.Райнердің еңбектері жарияланған уақыттан бастап, сөйлеуді автоматты тану жүйелеріндегі сөйлеу сигналын сипаттау үшін мел-кепстралдық коэффициенттер пайдаланылады (MFCC Mel Frequency Cepstral Coefficients), олардың даму негізін Пол Мермельстайн қалаған болатын [3].

Сонымен қатар, соңғы уақыттарда қазіргі қолданыстағы MFCC диктордың сөйлеу ағынының [4] вариациялығына төзімді белгілер балама болуда, ол мейлінше сенімді жүйелерді құруға септігін тигізеді.

Қазіргі уақытта дауыстық биометрия жүйесін әзірлейтін көптеген шетелдік (Agnitio, Nuance, Voice Security Systems) және Ресейлік (Speech technologies, Speech Technologies Centre) компаниялар бар [5]. Жасалған жүйелердің көпшілігінде диктордыанықтау қателігінің ықтималдығы 1-3% құрайды, бірақ бұл қосымшалардың бірқатар кемшіліктері бар.

Сөйлеуді тануды анықтауда әлемдік тәжірибеде біршама жетістіктерге қол жеткізген Карнеги-Меллон университеті (АҚШ), Иллинойс университеті (АҚШ), Орегон университеті (АҚШ), шығыс Финляндия университетінде Томи Кинунен бастаған зерттеушілер тобы және т.б. шетелдік университеттер біршама жетістіктерге жеткен. Украинаның Донецк қаласындағы «Информатика және жасанды инттелект мемелекеттік университінің» ғалымы В.Ю.Шелепова, Ресей ғылым академиясының Санкт-Петербург мемлекеттік университетінің ғалымы А.А.Карпов және де Аграновский А.В., Леднов Д.А., Балакирев Н.Е., Малков М.А., ал отандық ғалымдардан Л.Н.Гумилев атындағы Еуразия ұлттық университетінің ғалымы А.Шәріпбай, Әл-Фараби Қазақ ұлттық университетінің ғалымы У.А. Тукеев, Назарбаев университетінің ғалымы Ж.А.Есенбаев, ҚР БҒМ ҒК Ақпараттық және есептеуіш технологиялар институтының ғалымдары Әмірғалиев Е.Н., Мұсабаев Р.Р., Мамырбаев Ө.Ж. айналысқан.

XXI ғасыр – жасанды интеллект ғасыры. Елімізде жасанды интеллект жүйесі кешенді түрде дамуда. Осы үдерісте аса өзектілікті әліде шешуін күтетін мәселелер аз емес. Солардың бірі – сөйлеуді автоматты тануда сөйлеушіні анықтау. Сөйлеуді тануды анықтаудың өзі тіл ерекшелігіне байланысты, сондықтан өзге типті тілдерге арналған қазіргі қол жеткізілген тәжірибелерді түркі тектес тілдер тобына жататын тілдер үшін қолдану едәуір күрделілікті туғызады. Осыған байланысты шетелдік тәжірибелерді принциптік негізге ала отырып, қазақ тілінің тілдік ерекшеліктері сөйлеуді танудың проблемаларын шешуге едуір маңызды. Сондықтан да сөйлеуді танудағы сөйлеушіні анықтаудың жаңа модельдері мен алгоритмдерін құруға арналған бұл жұмыс өзектілікті болып табылады.

Диссертациялық жұмыстың мақсаты: Сөйлеулерді тануда машиналық оқытуды қолдана отырып, сөйлеу белгілерін және сөйлеушіні анықтайтын модель мен алгоритмін құру және зерттеу.

Зерттеудің міндеттері. Белгіленген мақсатқа қол жеткізу үшін төмендегі міндеттерді шешу қажет:

- Сөйлеуді тану саласындағы сөйлеу белгілерін және сөйлеушіні анықтауға арналған әдістер мен заманауи жүйелерге сараптау жүргізу;
- Сөйлеуді тану үдерісінде сөйлеушінің дыбыстық сөйлеу белгілерін және сөйлеушінің мәліметтерін анықтауға арналған акустикалық корпусын құру;
- Нейрондық желілер негізіндегі гендерлік ерекшелігі мен сөйлеушіні анықтаудың моделі мен алгоритмін құру;

Зерттеу нысаны. Сөйлеуді тану және анықтау жүйесі.

Зерттеу пәні: Сөйлеудің акустикалық деректері және сөйлеуді автоматты түрде тануда сөйлеушіні анықтау әдістері мен алгоритмдері.

Зерттеу әдісі: Ақпараттар теориясы, сигналдар теориясы, нейрондық желілер теориясы және технологиялары, сөйлеулерді тану әдістері мен технологиялары, бейне тану теориясы мен технологиялары, бағдарламалық қамтама жобалау және құру технологиялары.

Жұмыстың ғылыми жаңалығы:

- Сөйлеуді тану үдерісінде сөйлеушінің дыбыстық сөйлеу белгілерін және сөйлеушінің мәліметтерін анықтауға арналған акустикалық корпусы құрылды;
- Машиналық оқыту саласындағы классификациялық алгоритмдер көмегімен және осы алгоритмдердің дәлдігін арттыра отырып сөйлеуші анықталды;
- Нейрондық желілер негізіндегі гендерлік ерекшелігі мен сөйлеушіні анықтаудың моделі мен алгоритмі құрылды;
- Зерттеу барысында алынған модель мен алгоритм көмегімен сөйлеу белгілерін және сөйлеушіні анықтауға арналған бағдарламалық қосымша құрылды.

Жұмыстың теориялық және практикалық маңызы. Зерттеу жұмысының теориялық маңызы сөйлеулерді тануда гендерлік ерекшеліктерін анықтауға арналған нейрондық желі модельдері мен алгоритмдерін жетілдіру болып табылады. Сонымен қатар қазіргі кезде пайдаланылып келе жатқан әдістердің жетілдірілген және ерекшелігі бар сөйлеу белгілерін анықтайтын жаңа әдістер әзірлеумен және эксперименттік зерттеумен сипатталады.

Диссертациялық зерттеудің практикалық маңызы сөйлеу сигналдарының белгілерін анықтау кезінде әзірленген, жетілдірілген нейрондық желі моделдерін қолдану; құрылған акустикалық корпус сөйлеуді тану саласында зерттеу жұмыстарын жүргізуге мүмкіндік береді.

Қорғауға шығарылған негізгі тұжырым. Сөйлеулерді тану және сөйлеу белгілерін анықтауға арналған 36 сағаттан астам сөйлеулерден тұратын қазақ тілінің акустикалық корпусы жасалған. Деректер базасында тіркелген сөйлеушілердің дыбыс ерекшеліктерін автоматты түрде сүзгілеуден өткізіп, сөйлеулер белгілерін анықтау.

Сөйлеу белгілерін анықтауда нейрондық желілердің жалпы құрылымы мен жетілдірілген архитектуралары қарастырылып, гендерлік ерекшеліктерін анықтауда қолданатын алгоритм мен модель құру, нейрондық желі MLP мен CNN архитектурасын салыстырып қайсысы жақсы нәтиже беретіні анықталды.

Сенімділік дәрежесі мен апробациялау нәтижелері. Диссертациялық жұмысымның тақырыбына байланысты зерттеулер мен нәтижелері төменде көрсетілген жарияланымдар негізінде көрсетілген әртүрлі конференция мен семинарларда баяндалды және талқыланды.

1) «Төртінші өнеркәсіптік революция жағдайындағы дамудың жаңа мүмкіндіктері» атты ҚР Президенті Н. Назарбаевтың Жолдауын іске асыру шеңберінде «Көліктегі инновациялық технологиялар: білім, ғылым, тәжірибе»

атты XLII Халықаралық ғылыми-практикалық конференциясында (Алматы, 18 сәуір, 2018).

2) Профессор М.Б.Айдархановтың және Валдамер Войцуктің 70 жылдығына, Е.Н.Амирғалиевтің 60 жылдығына арналған «Информатика және қолданбалы математика» атты IV Халықаралық ғылыми-практикалық конференциясында (Алматы, 25-29 қыркүйек 2019).

3) 3rd International Conference Applied Mathematics, Computational Science and Systems Engineering (Рим, Италия, 2018).

4) «11th Asian Conference on Intelligent Information and Database Systems» халықаралық ғылыми конференциясы (Йоджиякарта, Индонезия, 2019).

Диссертациялық тақырыбымның ғылыми бағдарламалармен байланысы.

Диссертациялық жұмыс Қазақстан Республикасының Білім және Ғылым министірлігі Ғылым комитетінің Ақпараттық және есептеуіш технологиялар институтында бекітілген PhD докторлық диссертациялар жоспарына және ЖТН –AP05131207 «Терең нейрондық желілерді пайдаланатын мультитілдік автоматты сөйлеуді тану технологиясын құру» жоба негізінде сәйкес орындалды.

Нәтижелердің жариялынымдары. Диссертация тақырыбы бойынша алынған нәтижелері жеті баспалық жұмыста жарияланды. Оның ішінде халықаралық реферативтік базаларына енетін басылымдарында жарияланған мақалалар:

1) Speech Recognizer-Based Non-Uniform Spectral Compression for Robust MFCC Feature Extraction. Przegląd Elektrotechniczny, ISSN 0033-2097, №6, 2018. (Scopus және Web of Science (Clarivate Analytics));

2) Automatic Recognition of Kazakh Speech Using Deep Neural Networks. 11th Asian Conference on Intelligent Information and Database Systems, ACIIDS 2019 (Scopus).

3) Voice verification using i-vectors and neural networks with limited training data. Bulletin of the national academy of sciences of the republic of Kazakhstan. May-jun 2019 (Web of Science).

4) Mamyrbayev O, Toleu A, Tolegen G, Mekebayev N. Neural Architectures for Gender Detection and Speaker Identification // Journal Cogent Engineering. ISSN: 2331-1916. – 2020. Volume 7, - Issue 1. (Scopus)

Жұмыс көлемі мен құрылымы. Диссертациялық жұмыс кіріспе, 3 тарау, қорытынды және пайдаланылған әдебиеттерден тұрады. Диссертацияның толық көлемі: 108 бет жазба мәтіні, соның ішінде 33 сурет, 10 кесте. 110 пайдаланылған әдебиеттер тізімі атаудан және 2 қосымшадан тұрады.

1 СӨЙЛЕУ СИГНАЛДАРЫН ӨҢДЕУ ӘДІСНАМАСЫ

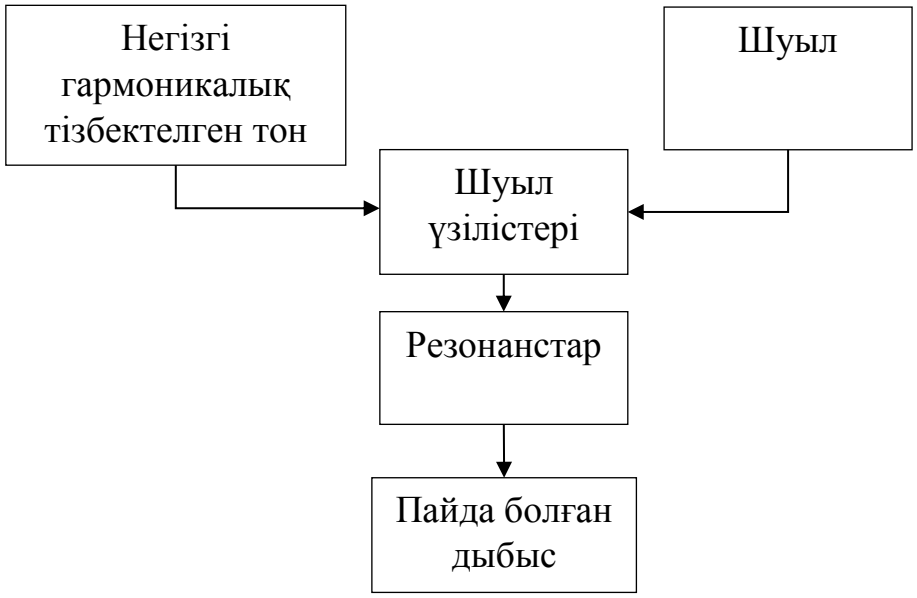
1.1 Сөйлеу сигналын алдын ала өңдеудің жолдары мен анықтау белгілерін анықтаудың ерекшеліктері

XXI ғасырдың басы сөйлеу технологияларының дамуы және өндіріске енгізілуімен ерекшеленеді. Сөйлеуді тану жүйесі ерекше маңызды орын алады. Бұл салада ірі зертеулерімен келесідей авторлардың Б. М. Лобанов, Т. К. Винцюк, А. В. Фролов, Л. Р. Рабинер, Р. В. Шафер, У. А. Ли, Д. Х. Клетт, Xuedong D. Huang, Hsiao-Wuen Hon, Alex Acero еңбектері танымал болды. Бұл кезеңде сөйлеу сигналдарын өңдеу саласында көптеген іргелі және қолданбалы шешімдер жасалынғандықтан аса маңызды болғаны Ресей, Еуропа, Америка ғалымдарының еңбектерінен көрінеді. Сөйлеу сигналдарын өңдеу саласындағы зерттеу жұмыстары қазіргі кезде де кеңінен және белсенді жүргізілуде.

Адамның жаратылысынан бастап сөйлеу – қарым-қатынас жасаудың қызметін атқаруда. Осы тұрғыдан сөйлеу мүмкіндіктерін түрліше сипаттауға болады. Олардың саны бойынша сипатталуы К.Шеннон әзірлеген ақпараттық теорияға негізделеді. Бұл теорияға сәйкес сөйлеуді оның ақпараттық құндылығымен сипаттауға болады. Сөйлеуді сипаттаудың екінші бір тәсілі оны сигнал түрінде ұсыну яғни акустикалық тербеліс түрінде танылады.

Сөйлеу қарым-қатынасының басталуында диктор миында абстракты қалыпта белгілі бір хабарлама пайда болады. Сөйлеудің пайда болу үдерісінде бұл хабарлама акустикалық сөйлеу тербелісіне айналды. Сөйлеу сигналдары көмегімен берілетін хабарлама дискреттік болып табылады яғни символдардың тізбегі түрінде ұсынылуы мүмкін. Сөйлеу сигналдары құрылымдалған дыбыстық символдар - фонемалар болып табылады [6].

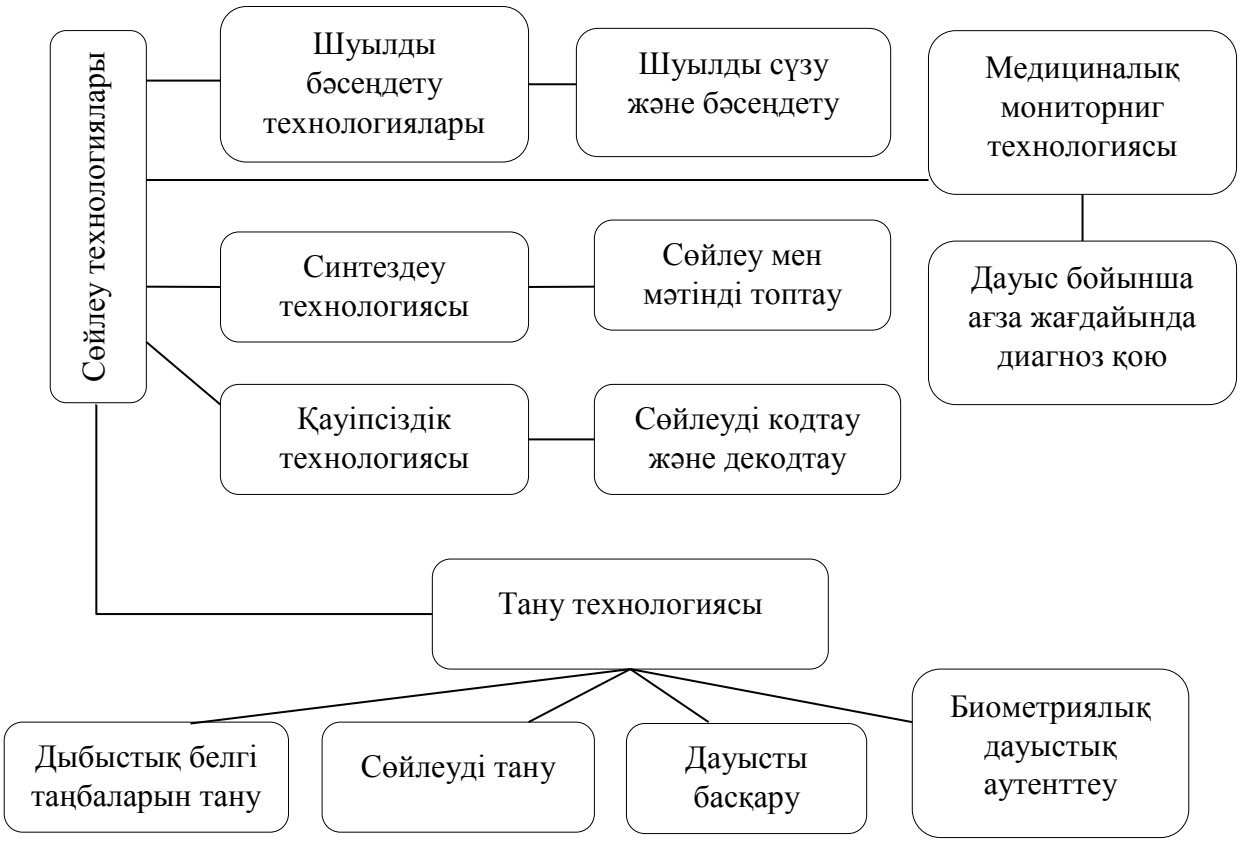
Физикалық тұрғыдан алғанда сөйлеу дыбыстар тізбесінен тұрады, олардың топтарының арасында үзілістер болады [7,8,9]. 1.1-суретте адам сөйлеуінің пайда болу сызбасы берілген.



Сурет 1.1 – Адам сөйлеуінің пайда болу сызбасы

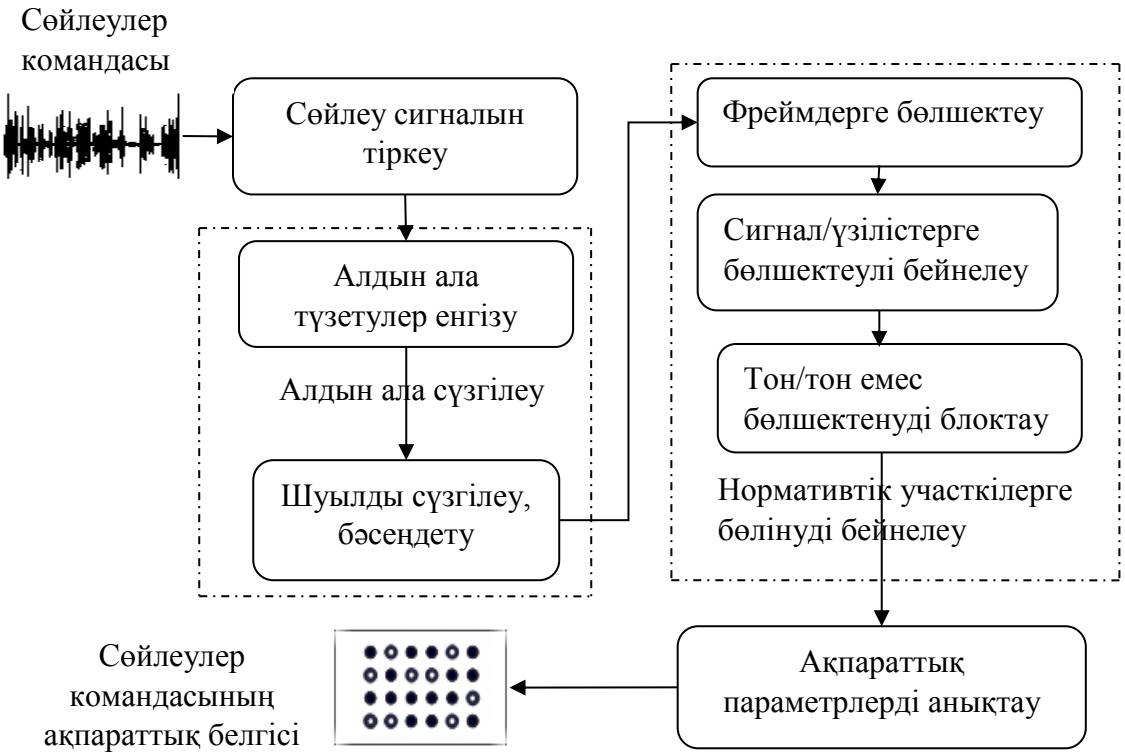
Сөйлеу сигналдарын өңдеу дегеніміз ол шуылды сүзу мен бәсеңдету, ақпараттық ағындарын күшейту, бөлшектеу, ақпаратты шығарып алу, сөйлеу сигналдарын кодтау, қысу және қалпына келтіру әрекеттері жасалатын ғылым саласы. Ол барлық бағытта кең таралған (1.2 сурет). Сөйлеу бұйрықтарын өңдеу бағыты дауысты басқару жүйелерінде төмендегідей міндеттерді қамтиды [10]:

- тіркеу;
- алдын ала түзетпелер жасау;
- шуылды сүзу мен бәсеңдету;
- фреймдерге бөлшектенуін бейнелеу;
- сигнал/үзілістер бөлшектерінің бейнеленуі;
- тон/тон емес бөлшектерінің бейнеленуі;
- ақпараттық параметрлерді анықтау.



Сурет 1.2 - Сөйлеу технологияларын түрлі бағыттарда қолдану

Жоғарыда көрсетілген міндеттерді шешетін сөйлеу командаларын өңдеудің алгоритмі 1.3-суретте көрсетілген.



Сурет 1.3 - Сөйлеу командаларының өңдеу алгоритмі

Тіркеу. Тіркеу дегеніміз сөйлеу командасын нақты уақыт режимінде дыбыстық тұрғыдан стандартты құралдарды қолданып сандық түрге келтіру. Стандартты құралдар: микрофон, алдын ала және негізгі күшейткіш, аналогты-сандық түрлендіргіш (АСТ) т.б.с.с.

Дыбыстық толқынның қысымы микрофон арқылы қабылданады және ол электрлік аналогтық сигналға айналады. Әрі қарай сөйлеу командасының ақпараттық бейнесі дискреттеу мен кванттауды жүзеге асыратын АСТ көмегімен аналогтық сигналдан, сандық сигналға түрлендіріледі [11].

Сөйлеу сигналдарын тіркеу төмендегідей қосымша мүмкіндіктерді көрсете береді:

- күшейтуді автоматты реттеу және әлсіз, күшті сөйлеу сигналдарының сапалы жазылуын қамтамасыз ететін жақын және алыс пайдаланушының деңгейін теңестіру;
- тіркеудің жұмыс параметрлерін жекелей және топтық қайта баптау және жазу процесін тоқтатпай түрлендіру;
- жазу арналарының санын немесе тіркелетін ақпараттың типтерін арттыру.

Алдын ала түзету. Алдын ала түзету адамның сөйлеу аппараттарында сөйлеу дыбыстарын айту кезінде пайда болатын табиғи бұрмалануды (6дБ әр октова бойынша) жоюға арналған [12].

Сөйлеу сигналын төмендегідей түрде тасымалдау функциясымен бар түзетуші сүзгі арқылы өткізіледі:

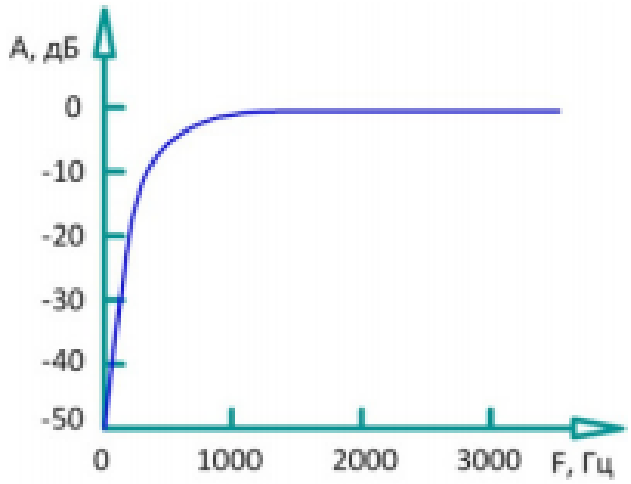
$$W(z) = \sum_{k=0}^{m} c_k z^{-k} \qquad (1.1)$$

Мұндағы c_k – тұрақты коэффициенті, m – бүтін сан (m>0). Көбінесе m=1, ал тасымалдау функциясы мына түрде болады:

$$W(z) = 1 - c_1 z^{-1} \qquad (1.2)$$

c_1 – коээфициенті – 0,4 тен -1,0 дейінгі аралықта алынады, k-1 ЭЕМ-де тіркелген нүктесі бар сүзгі оңай жасалғандықтан k-1 мәніне өте жуық. Көбінесе $c = -(1 - 1/16) \approx -0.95z^{-1}$ [13].

Алдын ала түзету спектрлік талдау алдында сигнал спектрін теңестіреді (1.4 сурет).



Сурет 1.4 - Сөйлеу сигналының спектрін теңестіру

Алдын ала түзету міндетті емес, сондықтан дауысты басқаратын көптеген жүйелерде түзету қарастырылмаған, мұнда талдау жасау сатысында адамның сөйлеу аппаратына тән дыбыс спектрінің бұрмалануы ескеріледі.

Сүзгіден өткізу. Шуылды сүзгіден өткізу және бәсеңдету – бұл акустикалық, сондай-ақ технологиялық себептерден туындаған шуылдың түсініктілігін арттыруға және мөлшерін азайтуға мүмкіндік беретін сөйлеу командаларын өңдеу кезеңі. Шуыл – өзінің уақытша және спектрлік құрылымдардың күрделілігімен сипатталатын әртүрлі физикалық сипаттағы кездейсоқ тербелістер [14]. Сөйлеу сигналдарына қатысты шуыл – ол сигналдың ақпараттық параметрлерін өзгертетін әртүрлі қарқындылық пен жиіліктегі периодтық дыбыстар жиыны [15].

Шуыл пайдалы сөйлеу сигналымен өзара әсер етуі бойынша аддитивтік және мультипликативтік болып бөлінеді. Аддитивтік шуыл пайдалы сигналмен араласып беріледі, бұл аса маңызды болмаса да ауытқушылық тудырады. Мультипликативтік шуыл пайдалы сигналмен бірігіп қайта көбейеді. Сөйтіп көптеген ауытқушылық туындатады, ол тіптен сөйлеу командасының параметрлін де өзгерте алады.

Сигналдың және шуылдың комбинациясы жалпы түрде келесідей жазылды:

$$S(r) = \left(k_c(r) + k_m(r)\right) \cdot e(r) + n(r) \qquad (1.3)$$

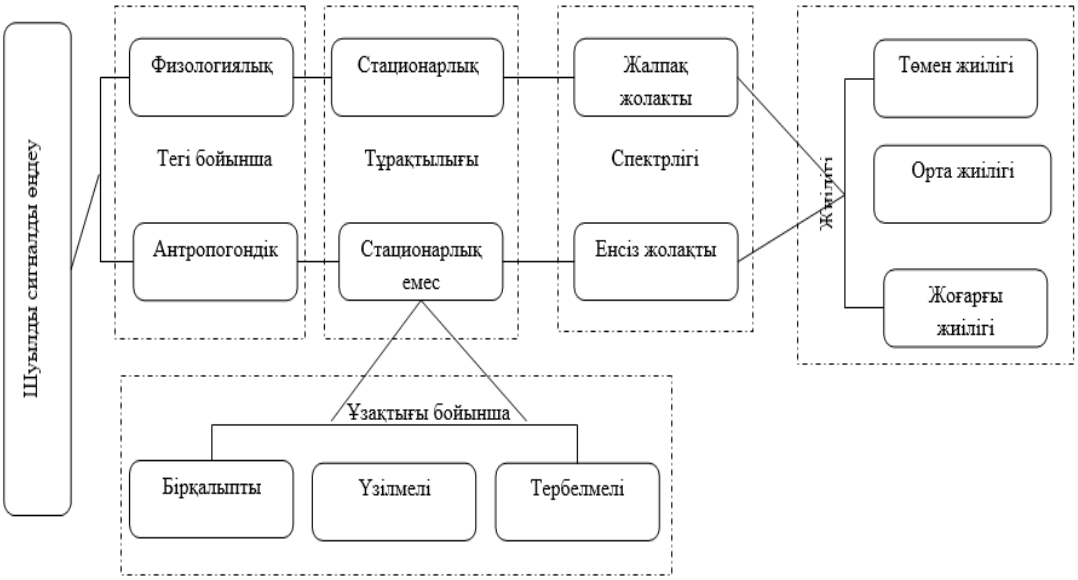
мұндағы $e(r)$ – пайдалы сөйлеу сигналы, $k_c(r)$ – пайдалы сөйлеу сигналын сипаттайтын коэффициент, $k_m(r)$ – мультипликативтік шуылды сипаттайтын коэффициент, $n(r)$ - аддитивтік шуыл.

I_c және шуыл $I_Ш$ сигналдың қарқындылық қатынасы [16-17]. Бұл қатынас сигнал/кедергі қатынасы деп аталады, ол шуылды, сүзгіден өткізу мен бәсеңдетуде маңызды рөл атқарады. Сигнал/кедергі қатынасы децибел деп аталатын логорифмдік өлшемсіз бірліктермен беріледі (dB,дБ).

$$M = 101gI_cI_Ш \qquad (1.4)$$

мұндағы $I_c, I_Ш$ – сигнал мен шуылдың қарқындылығы.

Сөйлеу сигналдарын сүзгіден өткізу және бәсеңдету кезіндегі жетістіктерді талдау негізінде және жүргізілген зерттеулері нәтижесінде, сөйлеу сигналдарының анықтылығына, түсінікілігіне ықпал ететін шуылдарды төмендегіше жіктеу ұсынылады (1.5 сурет):



Сурет 1.5 - Сөйлеу сигналындағы шуылдардың жиілігі

Сөйлеу командаларындағы шуылдың пайда болуына байланысты физикалық және антропогендік деп бөлуге болады. Шуылдың бірінші түріне олардың пайдалы сөйлеу сигналдарымен жүйесіз бірігіп кеткендегі әртүрлі қарқындылық пен жиіліктегі дыбыстар кешені жатады.

Шуылдың бұл түрінің пайда болуы сөйлеудегі ауытқушылықтармен тікелей байланысты (сөйлеу аппаратының артикуляциялық бөліміндегі жекелей немесе ағзалар кешеніндегі ауытқушылықтар). Сөйлеудегі ауытқушылықтарды зерттейтін ғылым «логопедия» деп аталады, ол түзету оқыту құралдары арқылы сөйлеудегі ауытқуларды меңгеруді және алдын алуды көздейді. Сөйлеудегі

ауытқулардан болатын шуылдарға формасы мен құрылымы айтылымындағы ауытқушылықпен тікелей байланысты көптеген дыбыстарды жатқызады:

- сөйлеу сигналдарының жылдамдығы мен ырғағы бұзылған (брадилалия, тахилалия, тежеліс, тұтығу);
- дауыстың бұзылуы (афония, дисфония, ринофения);
- сөйлеу сигналдарының шашыраңқылығы (афазия).

Айқын антрогендік шуылға физиологиялық шуылдан басқа шуылдың барлық түрі жатады. Антропогендік атауының өзі оның тікелей адамның әрекеттерінен пайда болатындығымен тікелей байланысты пайда болған. Кейбір әдебиеттерде оларды өндірістік немесе бейнелі шуыл деп атайды [18]. Анропогендік шуылдың шығу көздеріне автомобильдерден, теміржол пойыздарынан, ұшақтардан, өндірістік кәсіпорындардан, құрылыс және жөндеу жұмыстарынан, тұрмыстық және кеңселік техникадан т.б. туындайтын дыбыстар жатады.

Параметрлерінің тұрақтылығы бойынша барлық шуылдар тұрақты және тұрақты емес деп бөлінеді. Тұрақты шуыл – орта параметрлердің тұрақтылығынан сипатталады. Олар: қарқындылық (қуаттылық), қарқындылықтың спектрлер бойынша бөлінуі (спектрлік тығыздық), автокорреляциялық қызметтер. Тұрақты шуылдың классикалық үлгісі - бұл ақтаңдақ шуыл, оның спектрлік компоненттері барлық жиілік диапазонында біркелкі таратылады [19].

Тұрақты емес шуыл - қысқа уақытқа созылатын шу (орташа уақыттан қысқа) [20]. Олар ұзақтығы бойынша импульстік, үзілмелі және тербелмелі болып бөлінеді. Импульстік шуыл – ұзақтығы 1секунтдан аз бір немесе бірнеше дыбыстық сигналдардан құралады, оның деңгейі 7дБ-ден кемдеу. Үзілмелі шуылдың немесе үзік-үзік шуылдың деңгейі сатылап өзгеріп тұрады (шамамен 5дБ, одан көбірек) мұнда аралық ұзақтық деңгей тұрақты күйде қалып отырады ол 1 секунд және одан көбірек болады. Тербелмелі шуыл деңгейі уақыты бойынша үздіксіз өзгеріп тұрады [20].

Сөйлеу командаларындағы тұрақты емес шуылға көше шуылы, өтіп жатқан көлік шуылы, өндірістік жағдайлардағы жекелей тырсылдар, радиотехникадағы импульстік сирек қырылдар т.б. жатады.

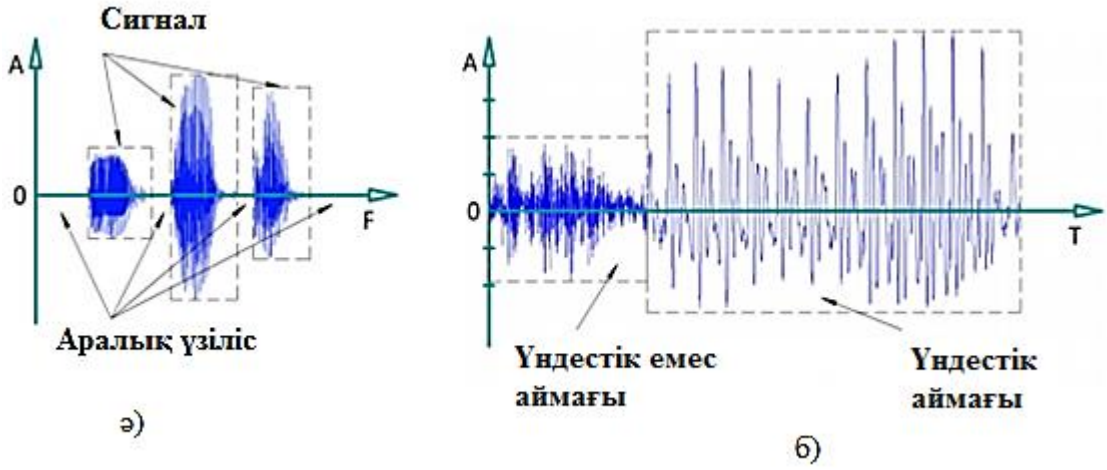
Спектрдің сипаттамасына қарай, шуыл кең жолақты және тар жолақты деп бөлінеді. Кең жолақты шуылдың үздіксіз спектрінің ені 1 октовадан артық болады. Тар жолақты шуыл (тон бойынша) дегеніміз дыбыс белгілері бір жиілікте естілетін шуыл. Жиілік сипаттамаларына байланысты шуыл төменгі жиілікте (<400 Гц) бөлінеді, орташа жиілікті (400-1000 Гц), жоғары жиілікті (> 1000 Гц) болып бөлінеді[21].

Сегменттеу. Сөйлеу командаларын өңдеудегі сегменттеу-сөйлеу ағынын сызықтық бөлікке бөлу сегменттер деп аталады. [22].

Сөйлеу сигналдары параметрлері мен сипаттамалары, әдетте уақыт ішінде тез өзгереді, күрделі формадағы тұрақты емес сигналдар болып табылады. Алайда, сөйлеуді өңдеудің көптеген әдістерінің негізінде уақыт өте келе сөйлеу сигналының қасиеттері баяу өзгереді деген болжамға негізделген. Бұл болжам

қысқа мерзімді талдау әдістерін қолдануға алып келеді, онда сөйлеу сигналының сегменттері әртүрлі қасиеттері бар жеке дыбыстардың қысқа бөліктері сияқты шығарылып, сол жерде өңделеді. Бірдей ұзындықтағы ақпараттық белгілердің жиынын алу үшін, сөйлеу сигналдарын фреймдер деп аталатын өзара тең үздіктерге/кесінділерге сегменттеу (бөлшектеп бейнелеу) бөлу керек (1.6[а] сурет) [23]. Мұндағы фреймдердің ағынын бекіту шектерінде сигнал туралы ақпараттың жоғалмауынан болдырмау үшін қолданылады. Ағынды бөгеу азайған сайын, нәтижесінде қарастырылып отырған аймақ үшін тән белгілер жиынының мөлшері азая түседі. Есептеуіш ресурстарды үнемдеу мақсатында кейде бөгеу алынды, өйткені ол мәліметтерді өңдеуді едуір баяулатады [24].

а)



Сурет 1.6 - Сегменттеу. а)-фреймдер, ә)- сигнал аралық үзіліс, б)-үндестік, үндестік емес аймақтар

Сигнал/үзіліс сегментациясы сөз тіркестерінің бастауы мен аяқталу сәттерін анықтау міндеті болып табылады [25]. Бұл міндет шуылдың бар болу жағдайында сөйлеу командаларын өңдеу саласында маңыздылығымен ерекшеленеді (1.6[ə] сурет). Атап айтқанда, дауысты басқару жағдайында

командаңың басталуы мен аяқталуы сәттерін дәлме-дәл анықтау маңызды болып табылады.

Сөйлеу сигналдарындағы аймақтардың үндестік/үндестік емес болып сегменттелуі өңдеу үдерісінде маңызды міндетке жатады (1.6^{б} сурет).

Тональды аймағы деп уақыт аралықтары барысында сөйлеу дыбыстарын топтау дауыс көзінің қатысумен жүзеге асуын айтады. Тональды емес аймақтарға сөйлеу дыбыстарының пайда болуы дауыс көзінің қатысуынсыз болатын уақыт аралықтарын жатқызады.

Сөйлеу командаларын талдауда тональдық аралықтарында құнды рөл атқарады. Оларды талдау арқылы акустикалық сипаттамалардың және сөйлеу сигналдарының мағыналық мәндерінің ақпараттық көріністерін жеткілікті түрде алуға болады. Кейбір жағдайларда сөйлеу командаларын өңдеуде басты мақсатқа айналады. Мұндай жағдайлар сөйлеудің маңызды параметрін, дикторды тану және анықтау міндетіндегі сөйлеушінің негізгі тонының жиілігін анықтайды [26].

Ақпараттық белгілерді анықтау. Ақпараттық параметрлерді анықтау деп сөйлеу сигналдарының ақпараттық белгілері мен сипаттамаларын анықтау міндетін айтады [27]. Адамның ақпараттық сөйлеу параметрлерін сипаттайтын негізгі түсініктің қалыптасуы сөйлеу аппаратының пішінімен, мөлшерлерімен, өзгеру динамикасымен байланысты және ол адамның эмоциялық көңіл-күй жағдайымен де түсіндіріледі. Сөйлеу сигналдарының ақпараттық параметрлерін бөлу және өзіндік зерттеу саласындағы жетістіктерді талдау негізінде барлық ақпараттық параметрлерді сөйлеу үлгілерін ажыратуға мүмкіндік беретін объективті белгілердің үш тобына бөлуге болады:

- спектрлік – уақыттық;
- кепстральдық;
- амплитудалық – жиілік.

Бірінші топ шартты түрде спектрлік және уақыттық белгілер болып бөлінеді. Спектрлік белгілерге жататындар:

- талданатын сөйлеу сигналдарының спектрінің орташа мәні;
- сигналдың спектр жолақтарында болуының салыстырмалы уақыты;
- жолақтардағы сөйлеу спектрінің медиандық мәні;
- жолақтардағы сөйлеу спектрінің салыстырмалы қуаты;
- сөйлеу спектрінің орама вариациясы.

Уақыттық белгілерге жататындар:

- сөйлеудің минимум құрылымдық белгілерінің сегментінің ұзақтығы (фонемалар, аллофондар, фифондар, трифондар);
- сегмент биіктігі;
- сегмент формасының коэффициенті.

Спектрлік – уақыттық белгілер сөйлеу сигналын оның физика-математикалық болмысында үш түрдегі компоненттердің бар болуына байланысты сипаттайды. Олар:

- дыбыстық толқынның кезең-кезеңдік (үндестік) аймақтары;

- дыбыстық толқынның кезеңдік емес аймақтары (шуылдық);
- сөйлеу үзілістері жоқ аймақтар.

Спектральды-уақыттық белгілер әр түрлі адамдардағы дауыстық импульстер спектрінің және олардың сөйлеу жолдарының сүзу функцияларының ерекшеліктерін бейнелеуге мүмкіндік береді, сөйлеушінің сөйлеу артикуляциялық органдарының қайта құру динамикасымен байланысты сөйлеу ағынының ерекшеліктерін сипаттайды және сөйлеушінің артикуляциялық органдарының өзара байланысын немесе қозғалыс синхрондығын көрсететін сөйлеу ағынының интегралды сипаттамалары болып табылады.

Кепстралдық белгілер:

- мел кепстралдық коэффициенттер (MFCC);
- тіркеу жиілігінің қуаттылығының коэффициенттері;
- сызықтық жорамал спектрінің коэффициенттері;
- сызықтық жорамал кепстрінің коэффициенттері.

Дауыстық басқарудың қазіргі жүйелерінің көпшілігі адам сөйлеуінің жолының жиілік сиппаттамаларына қол жеткізуге күшін шоғырландырады. Бұл бірінші модель коэффициенттері дыбыстардың бөлінулерін артықша қамтамасыз етуімен байланысты. Қозғалыс сигналын желілік жол сигналынан ажырату үшін кепстралдық талдауға жүгінеді [28].

Амплитудалық – жиілік белгілер:

- қарқындылық, амплитуда;
- энергия;
- негізгі тон жиілігі;
- форманттық жиіліктер.

Акустикалық тұрғыдан сөйлеу сигналы ауада таратылатын құрылымы жағынан күрделі дыбыстық тербелістер болып табылады. Олар өздерінің жиілігі «бір секундтағы тербеліс сандар), қарқындылығы (тербелістер амплитудасы) және ұзақтылығына бойынша сипатталады. Амплитудалық-жиілік белгілер адам үшін минимум қабылдау уақытындағы сөйлеу сигналы бойынша қажетті және жеткілікті ақпарат тасымалдайды.

Сөйлеу сигналдарын өңдеуді жалпылай тұжырымдағанда – сигналдарды сипаттау үдерісі болып табылады, одан әріде олар ақпараттық мазмұнды анықтау және пайдалану мүмкіндігі болуы үшін талап етілетін формаға айналдырады.

1.2 Сөйлеу сигналдарының акустикалық сипаттары

Акустикалық ақпарат деп - әдетте тасымалдаушылары акустикалық сигналдар болып табылатын ақпарат ретінде түсіндіріледі. Егер ақпарат көзі адамның сөзі болса, акустикалық ақпарат сөйлеу деп аталады.

Акустикалық сигналдардың бастапқы көздері механикалық тербеліс жүйелері болып табылады, мысалы, адамның сөйлеу мүшелері, ал екіншісі - әртүрлі түрдегі түрлендіргіштер, мысалы дауыс зорайтқыштар.

Акустикалық сигналдар бойлық механикалық толқындар болып табылады. Олар тербелмелі денемен шығады және акустикалық тербелістер (толқындар)

түрінде қатты денелерде, сұйықтықтарда және газдарда таралады, яғни әртүрлі ауытқулар әсерінен орта бөлшектерінің тербелмелі қозғалыстарында кездеседі.

20-20 000 Гц шегінде акустикалық тербелістердің жиілігі дыбыстық деп аталады (оларды адамның құлақтары қабылдай алады). Олар 20 Гц төмен болса инфрадыбыстық, ал 20000 Гц жоғары болса ультрадыбыстық деп аталады.

Акустикалық тербелістердің түріне байланысты қарапайым (тональды) және күрделі сигналдарды ажыратады. Тональды сигнал дегеніміз - синусоидальды заңына сәйкес тербеліс нәтижесінде пайда болатын сигнал. Күрделі сигнал гармоникалық құрамдастардың тұтас спектрін қамтиды. Сөйлеу сигналы күрделі акустикалық сигнал болып табылады.

Ақыл-ой тұрғысынан сөйлеуді үш сипаттамалық топқа бөлуге болады:

- тілдің фонетикалық сипаттамасы - оның дыбыстық құрамы тұрғысынан сөйлеуді сипаттайтын деректер;
- физикалық сипаттамалар - дыбысты акустикалық сигнал ретінде сипаттайтын шамалар мен тәуелділіктер;
- тілдің семантикалық немесе мағыналық жағы оның көмегімен берілетін ұғымдардың мағынасын сипаттайды.

Сөйлеудің жиілік спектрі гармоникалық құрауыштардың көп санын қамтиды, олардың амплитудасы жиіліктің өсуімен азаяды. Спектрлік құрамы бойынша сөйлеу дыбыстары бір-бірінен формант санымен және олардың жиілік спектрінде орналасуымен ерекшеленеді. Сөйлеу дыбыстарының форматтары шамамен 150-ден 8600 Гц-ға дейінгі жиіліктің кең аймағында орналасқан. Соңғы аралықта тек 12 000 Гц-қа дейінгі аумақта жатуы мүмкін Т2 формантты дыбыс жолағының құрамдастарынан асады. Алайда, сөйлеу дыбыстарының формантының басым бөлігі 300-ден 3400 Гц-ға дейінгі аралықта болады, бұл берілген сөйлеудің дәлме дәлдігін қамтамасыз ету үшін, осы жиілік жолағын жеткілікті деп есептеуге мүмкіндік береді.

Дыбыстың физикалық қасиеттері белгілі бір физикалық ортада, мысалы, ауада, таралатын әр түрлі дыбыстық қысымдағы толқындардың аса үлкен шектік қалыптары түрінде сипатталына алады. Берілген жұмыста тек бойлық дыбыстық толқындар [29] зерттелінеді, яғни ортаның белгілі бір молекулалары шартты түрде өздерінің орташа шептік қалыптарының толқынның таралу бағытымен түйісетін бағытта қозғалады. Толқындардың таралуы бір-бірінен сол толқынның жартысындай қашықтықта орналасқан молекулалар қарама-қарсы бағыттарда тербеліс түзеді, бұның өзі қысылу мен бәсеңдеудің бір-бірін алмастыратын аймақтарының пайда болуына алып келеді. Сөйтіп бір сәттік және статистикалық тұрақты қысымдар арасындағы айырма ретінде анықталынатын дыбыс қысымы деген – шектік қалып пен уақыт функциялары болып табылады.

Бұдан әріде дыбыстық толқындар тек ауада ғана таралады және таралу ортасы келесідегідей қасиетке ие:

1. Гомогендік яғни құрылым біртектілігі.
2. Изотроптық яғни орта қасиеттерінің бағыттан тәуелсіздігі.
3. Тұрақты яғни орта қасиеттерінің уақыттан тәуелділігі.

Дыбыс таралуы орта салмақ пен созылғыштық, иілгіштік қасиеттеріне ие. Идел газдың созылғыштығы, иілгіштігі, жұмсақтығы көлемнің созылып кеңейюімен және көлемнің қысылуымен анықталады.

Идеалды газдың көлемінің қысылуы немесе теріс кеңеюі мына төмендегідей анықталады:

$$-\Delta = -\frac{\delta V}{V_1} \tag{1.5}$$

мұнда V -көлем, V_1- көлемнің өзгерісі.

Идеалды газдың созылмалығы, жұмсақтығы мына төмендегідей көлемдік модульмен анықталады

$$k = \frac{\sigma p}{-\Delta}, \tag{1.6}$$

мұнда σp - қысымдық өзгеруі.

Дыбыстық таралуы адиабаттық үдеріс болып табылады, өйткені ұзын бойлық толқындардың кеңеюі мен қысылуы жылудың таралуынан жылдам жүреді. Жылу сиымдылығын C_p және C_v арқылы белгілейміз, ол жылу сыйымдылығының сәйкес келетін тұрақты қысымда және тұрақты көлемде болды. Сонда көлемдік модульді $\mu = \frac{C_p}{C_V}$:$k \approx \mu p$ адиабаттық экспонент көмегімен жақындату мүмкін болады.

Дыбыстың жылдамдығын Лаплас $C = \sqrt{\frac{k}{\rho}}$, адиабаттық үдеріс жағдайларында өлшеген, мұндағы ρ - ауа тығыздығы. Дыбыстың ауадағы жылдамдығы, негізінен атмосфералық жағдайларға (негізінде, температураға және аз дәрежеде ылғалдылыққа) байланысты. Ауа идеалды газ болып табылады деген болжаммен, ауа қысымы дыбыс жылдамдығына маңызды рөл атқармайды, өйткені қысым мен тығыздық жылдамдыққа бірдей әсер етеді, нәтижесінде бұл екі әсер бір-бірін жоққа шығарады.

Дыбыс қарқындылығына анықтама беру үшін жылдамдық потенциалы ұғымын енгіземіз. Консервативті және жай қосылған вектор өрісінде ағын жылдамдығын скаляр функцияның градиенті ретінде көрсетуге болады, оны жылдамдық потенциалы деп атайды.

Анықтама. Дыбыс қарқындығы немесе акустикалық қарқындылық дегеніміз – ρ дыбыстық қысымның жылдамдық потенциалына көбейтіндісі ϕ: $I_{\text{дауыс}} = p\phi$.

Тұжырым. Дыбыс қарқындылығы қайнар көзге дейінгі қашықтықтың квадратына кері пропорционал [29].

Дәлелдеме. Толқын теңдеуінің шешімнен [30] шығыс және кіріс дыбыс толқындарының ең үлгілі шекті-қалпы (суперпозиция) ретінде берілуі мүмкін:

$$p = \frac{A}{r} e^{j\omega t - jkr} + \frac{B}{r} e^{j\omega t + jkr} \tag{1.7}$$

мұнда A,B – қуат көздердің күші. $p = p_0 \frac{\partial \varphi}{\partial t}$, дыбыстық қысым мен жылдам потенциалының [30] арасындағы өзара қатынасты пайдалана отырып, дыбыс қарқындылығын былайша көрсетуге болады:

$$I_{\text{дауыс}} = \frac{p^2}{cp_0} \quad (1.8)$$

Бұл екі теңдеуден тиістіні таба аламыз.

1.3 Сөйлеулерді тану және белгілерін анықтауға арналған модельдер

Сөйлеулерді тану және белгілерін анықтау әдістерінің көп ретте бірегей ерекшелігімен өзгеленіп тұратынына қарамастан, тұтастай алғанда, сол әдістерінің әрбіріне тән төмендегідей сатыларын айырықша көрсетуге болады:

1. Кіріс сөйлеу сигналынан белгілерді бөліп алу.

2. Алдыңғы қадамда алынған белгілер векторларының негізінде сөйлеуді тану моделін (үлгісін) құру.

Жүйеде тіркелген дикторды анықтау үдерісі барлық қаралатын әдістердегі кірістік сөйлеу сигналы бойынша белгілі бір критерийлер негізінде сақталған модельдердің ең сәйкесін іздеп табуда қолданылады.

Кейбір әдістер сөйлеуді тануда келесідей моделдерді қолдануды ұсынады: жасырын Марков моделі және Гаусс қоспасы.

Белгілі бір параметрлер бойынша жіктеуге негізделген сөйлеуді танып анықтаудың мынандай әдістері бар: динамикалық программалау әдісі, векторлық кванттау.

1.3.1 Динамикалық бағдарламалау әдісі

Dynamic Time Warping (DTW) – белгілі бір уақыт аралығында екі өлшеу тізбегі арасындағы жақындықты табуға мүмкіндік беретін динамикалық бағдарламалау әдісі. Жалпы алғанда, бұл тізбектер әр түрлі ұзындықта болуы мүмкін, осыған байланысты өлшеу әртүрлі жылдамдықта жүргізілуі мүмкін.

Бұл берілген әдісте сақталынатын модель ретінде $Q = \{q_1, \dots, q_n\}$ оқыту таңдамасындағы кіріс сөйлеу сигналының белгілер векторларының желілік тізбегі болады. Мұнда $C = \{c_1, \dots, c_n\}$ тестілік таңдаудағы кіріс сөйлеу сигналы белгілерінің векторларының желілік тізбегі. Сондай-ақ $M_{m \times n}$ екі желілік тізбектердің теңестірілу матрицасы ұғымы ендіріледі, мұнда (i,j) қалыптарында мынандай теңестірулердің мәні орын алады: a_i және q_j элементтері, соған сәйкес C және Q желілік тізбелер, сондай-ақ $W = \{w_1, \dots, w_T\}$ матрицасының аралас элементтерінің индекстерінің жиыны, ал ол салыстырылатын желілік тізбектерінің элементтері арасындағы сәйкестіктерді анықтайды. Мұнда жиын элементтері төмендегідей шарттарды қанағаттандыруы тиіс.

1. $w_1 = (1,1), w_T = (m,n)$

2. Егерде $w_{t-1} = (a,b)$, онда $w_t = (a,b)$, онда $a - a \leq 1, b - b \leq 1$

DTW алгоритмінің мақсаты - 1-ші және 2-ші шарттарды қанағаттандыратын W жиынын табу,онда Q тізбегіне қатысты C тізбегінің толық бұрмалануы минималды болады, яғни:

$$DTW(Q,C) = min\left\{\frac{1}{T}\sqrt{\sum_{t=1}^{T} M(w_t)}\right\} \quad (1.9)$$

Бұл теңдеудің мәні Q және С желілік тізбегінің өзара жақындығын анықтайды. Бұл мәндерді табу үшін динамикалық бағдарламалау әдісі қолданылады.

Дауысты анықтау үшін әрбір сақталған үлгі үшін $DTW(Q_i, C)$ мәні есептелінеді. Минимумға қол жеткізілетін i мәні кіріс сөйлеу сигналының үлгісіне жақын дауыс үлгісі бар пайдаланушының нөмірін анықтайды.

DTW алгоритмінің негізгі артықшылығы – жүзеге асырудың қарапайымдылығында. Алайда диктордың мәтінге тәуелсіздік жағдайында танып анықтау міндетін шешу үшін бұл берілген алгоритмді қолдану күрделі болып табылады.

1.3.2 Векторлық кванттау

Векторлық кванттау (Vector Quantization (VQ)) - $A = \{a_1, \dots, a_n\}$ бұл - кіріс векторларының желілік тізбегі үшін $W = \{w_1, w_2 \dots, w_n\}$ коды бар векторларын векторлық кванттау міндеті болып табылады, оның өзі Q-ға сәйкес келетін кодтық векторынан алынған әрбір вектордың орнын басқан жағдайында ауытқушылықты минималдау міндеті болып қойылады.

Бұл берілген әдісте сөйлеу сигналының белгілерінің векторларының кіріс желілік тізбегінен алынған кодтары бар векторлардың көптеген жиыны пайдаланушы моделі болып табылады. Бұл көп жиынды құру үшін белгілер векторларының бастапқы желілік тізбегі L кластерлерге бөлінеді, сөйтіп кодтары бар векторлар ретінде олардың орта нүктелері алынады.

Кіріс сөйлеу сигналы бойынша сөйлеушіні анықтау үдерісі мына ретте жүргізіледі. a_{i} әрбір тестілік векторлар үшін k ең жақын кодтары бар векторлар анықталады. Егер r_{ij} – табылған ең жақындардың ішінде S_j сөйлеушіге тиеселі векторлар саны болсын, сонда a_i векторы S_j дикторға тиесілі екенінің ықтималдығы мына формуламен анықталады:

$$\text{P}\left(S_j \middle| a_j\right) = \frac{r_{ij}}{r} \quad (1.10)$$

Векторлық кванттау әдісі жүзеге асыруда – жеңіл және сөйлеушінің мәтінге тәуелсіз танып анықтау міндетінде қолданыла береді, алайда ол әр кезде тану дәлдігін бере алмайды.

1.3.3 Гаусс қоспасы моделі

Caussian Mixture Model (GMM) – гаусс қоспасы моделі, өлшенген M компоненті b ықтималдық тығыздық жиыны, $p(x)$ теңдеуімен беріледі:

$$p(x|\lambda) = \sum_{i=1}^{M} p_i b_i(\bar{x}) \quad (1.11)$$

мұнда $\bar{x}$ - кездейсоқ көлемінің D-өлшемді векторы, $p_i, 1 \leq i \leq M$-бөліну тығыздығының функциясы.

Гаусс қоспаларының моделі толық түрде математикалық болжау векторларымен, ковариациялық матрицалар және модельдің әрбір компанентіне арналған құрамалар өлшемімен анықталады:

$$\lambda = \{p_i \mu_i \sum_i\}, i = 1 \ldots M \quad (1.12)$$

мұндағы μ_i- математикалық болжау векторы және $\sum_i -$ ковариациялық матрица.

Бұл берілген әдісті қолдануда әрбір тұтынушы λ Гаусс қоспалары моделі арқылы беріледі.

1.3.4 Жасырын Марков моделі

Hidden Markov Model (HMM) – статистикалық жасырын Марковтық моделі(ЖММ), ол бақылаушылар негізінде жасырын параметрлерді жіктеу міндетін шешу үшін қолданылады. ЖММ – ақырғы автомат болып табылады, онда қалып-күйлер арасындағы өзара өтулер кейбір ықтималдықпен жүзеге асады, сөйтіп басталатын үдеріс бастапқы қалып-күйді береді. Уақыттың таңдалынған сәттері арқылы жаңа қалып-күйлерге өту жүзеге асады. Мұнда белгіленген ықтималдығы бар әрбір жасырын қалып-күйге бақыланып отырған қалып-күй сәйкес келеді. Мейілінше кең жайылған *пайдаланушыларды дауыс бойынша анықтау (ПДБА)* технологиялары – 60-жылдарда Баум қалыптастырған жасырын Марков моделі теориясын қолдану болып табылады.

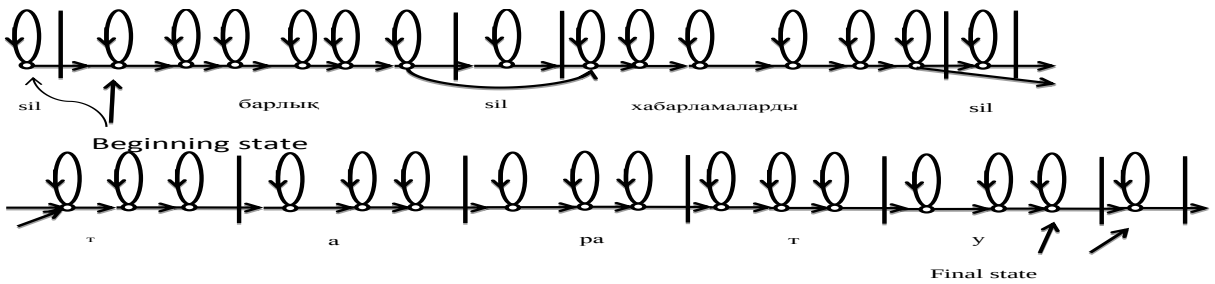
ЖММ қалып – күйге келтіру (оқыту) эталондар жиынының векторларына сәйкес келетін мүмкіндікті сигналдардың апостерорлық ықтималдығының максималдандыру бағыты бойынша ЖММ–нің параметрлерін таңдау арқылы көрінеді. Қалып – күйге келтірілген ЖММ –ді толық анықталынған сипаттамалары бар кейбір кездейсоқ сигналдар көзі ретінде қарастыруға болады. Белгісіз бейне (сөйлеу сигналы) «бақыланып отырғандардың» тізбегінің желісі түрінде беріледі. Содан кейін әр модель үшін осы үлгі бойынша жасалған реттіліктің құрылу ықтималдығы болады. Егер мейілінше үлкен ықтималдылық кейбір тіркелгендерден аз болса, онда анықталатын дауыс бір де бір қатысып отырған тұтынушыларға жатпайтыны туралы қорытынды жасалады.

ЖММ сөйлеу сигналының едуір өзгерістерін және стастикалық модельдеу аппаратындағы ауызекі сөйлеу тілі құрылымын модельдеуге мүмкіндік береді. ЖММ сөйлемдегі фонемаларды іс жүзінде көптеп қолдануды сипаттау үшін Марков тізбегін пайдаланады. Сигналды кластардың біріне жатқызу кезінде фонемаларды таңдау міндеті төмендегідей берілуі мүмкін

$$W_V(X): L_V(X) = maxL_r(X), r = \overline{1, R}, \quad (1.13)$$

мұнда $L_r(X)$ – ЖММ r-ге арналған мүмкіндікті функция. 1.7 суретте сөз бөліктерінің әртүрлі контекстік - тәуелді модельдерінен құрастырылған аяққы

қалыптардағы «барлық хабарламаларды тарату» сөйлемдері үшін жасырын Марков моделінің жұмыс сызбасы көрсетілген.



Сурет 1.7 - Жасырын Марков моделінің жұмыс схемасы

ЖММ–ның елеулі кемшілігі - фонемалардың анықталынатын кластары арасындағы ерекшелік сипаттамаларын бөліп көрсету мүмкіндігі жоқ.

1.3.5 Теориялық – ақпараттық тәсіл

Дауыс бойынша анықтау міндеттерін шешуде көптеген тәсілдер бар. Олардың ішінде *сөйлеуді қабылдаудың ақпараттық теориясы* (СҚАТ) шеңберлерінде профессор Совченко дайындаған теориялық-ақпараттық тәсіл көңіл аудартады [31], ол сөйлеу сигналдарын талдауды Кульбак Лейблердің ақпараттық өлшемдегі ақпараттық сәйкестік толықтығы минимумның критерийлері және минималдық топологиялық бірліктердің кластерлік моделі негізінде қарастырады [32]. Қазіргі уақытта бұл теория дауыс бойынша анықтау жүйелерімен жұмыс істеу кезінде перспективалы болып танылуы мүмкін.

СҚАТ өзінің негізінде ақпараттық сәйкестік минимумнің ұстанымын (АСМ) және залалсыздандыру сүзгісі әдісін (ЗСӘ) пайдаланады. Оның басқа таңдама жолдармен салыстырғанда тиімділігі және артықшылықтары (ПДБА-пайдаланушыны дауыс бойынша анықтау) жұмыс тәжірибесінен алынған бірқатар мысалдар берілген еңбекте көрсетілген [33]. Алайда қазіргі сәтте СҚАТ–ның артықшылықтарының және мүмкіндіктерінің бәрі бірдей тиісінше жарық көре және дами алған жоқ. Қазіргі уақытта жекелеген (оқшауланған) сөздер немесе тұтас сөз тіркестерінің күрделі сөйлеудің типтік бірліктері негізінде автоматты түрде анықтау мәселелерінде дәстүрлі әдістер мен таңдау жолдарға қарағанда, АСМ ұстанымының артықшылықтарын зерттеу өзектілікті болып отыр.

Сөйлеу функциясы адам ағзасының жоғары жүйке қызметінің өнімі және дерексіз немесе бейнелік ойлаудың тікелей шарты болып табылады. Адам санасындағы әрбір құбылыстың бейнесіне сәйкесінше сөйлеу «белгісі» қосалқы жүреді. Оның үстіне, әр түрлі адамдар қабылдауда бір бейненің өзі естілуде әр түрлі сөйлеу белгілеріне ие болады. Осының өзінде ауызекі сөйлеудің вариативтілігінің проблемасы көрініс табады. Берілген жұмыста [34] оның шешімі үшін тәсіл ұсынылған.

Ортақ бір атаулы белгілердегі бар айырмашылықтарға қарамастан, олардың барлығы да ортақ тән құбылыс деп танылады, әйтпесе сөйлеу өзінің ақпараттық

мәнін жоғалтқан болар еді. Сондықтан $x_j, j = \overline{1,J},\ j < \infty$, бір атаулы жүзеге асыру белгісі адам санасында біртиптес сигналдардың кластерінің типінің тиісті сөйлеу типіне топтастырылады деп тұжырымдауға болады. Сондай-ақ әрбір осындай кластердің берілген бейненің өзінің орталық этолондық белгісінің айналасында айқын көрсетілген шекаралары болады. Ол СҚАТ кілттік ұғым: $x_v, v \leq J$ сөйлеу белгісі, кейбір сөйлеу белгісінің x_* ақпараттық орталық-эталон түзілетін болады, ол үшін $\{x_j\}$ жүзеге асырудың көп ортақ бір атаулылық шегінде Кульбаку-Лейблер бойынша ақпараттық сәйкестік толықтығының минимумдық жиыны арқылы сипатталуы шарт, яғни:

$$x_* = x_v: \rho_v \triangleq \sum_{j=1}^{J} \rho_v(x_j) = \min_{i \leq J} \sum_{j=1}^{J} \rho_i(x_j), \qquad (1.14)$$

мұнда

$$\rho_i(x_j) \triangleq \iint ln\left[{dP_j(x)}/{dP_i(x)}\right] P_j(dx)$$

көрсетілген кластердегі j мен і сөйлеу сигналдарының арасындағы ақпараттық сәйкестік толықтығының көлемі (АСТК).

Сөйлеуді қабылдаудың ақпараттық теориясында берілгендерді сегменттер бойынша өңделген жағдайдағы тұтынушыларды автоматты анықтау міндеттерін шешу мейлінше қызығушылық туындатады. Мұндай жағдайда пайдаланушының үзіліссіз сөйлеуінде *қарапайым сөйлеу бірліктерін* (ҚСБ) анықтау жүреді. Тап осы қалыпта жоғарыда көрсетілген міндет бағдарламалық және техникалық әзірлемелердің көбінде жүзеге асырылады. Ұсынылып отырған W_x анықтау нысаны (біздің жағдайда – сөйлеуде біртектіліктің бөлініп алынған бөлшегі) $g(\omega_r)$ кластарынан R>1-де жатқызу талап етіледі, яғни фонемаларының типтерінің біреуіне. Бізге дәл таныс емес тұтынушының қайсы бірінен шығатын сөйлеу сигналынан X=$\{\overrightarrow{x_i}\}$ көп өлшемді қайта таңдау болып табылады деп есептейміз, ал $aP_0 = N(K_0)$-белгілі пайдаланушының статистикалық бейнесі рөліндегі (K_0) өзінің (nxn) автоковариациясының матрицасынан пайда болған бөліп орналастырудың қалыпты заңы. Сонда: X таңдалған дауыс P_0 пайдаланушыға жата ма деген мәселені шешу керек болады. Бұл бейнелерді тануда дихотомия яғни қабылданатын шешімдердің бинарлық («иә»-«жоқ» көптігі пайда болғанда жағдайға арналған бейнелерді тану міндеттерінің бірі болып есептеледі. Міндеттің бұлайша қойылуы «түзетуге келмейтіндік туралы» классикалық мәселенің терминдері еңбек ішінде егжей-тегжейлі зерттелген: $H_1: K_x = K_0$ деген n-өлшемдік Гаусстық бөліп орналастырудың АКМ (автокорреляциялық матрица) түріне қатысты болжам қабылданатын шешімдердің мәнділігі болып табылатын қайсы бір тіркелген деңгейдегі $H_1: K_x \neq K_0$ күрделі баламаға қарама-қарсылықта тексеріледі.

$$P(X \in W | H_0) = \alpha_0 = const \qquad (1.15)$$

мұндағы R^n таңдап алынған кеңістік шектеріндегі W сынилық сала кеңістік міндеттерін шешу критерийлеріне байланысты. Қайсыбір объектердегі [35] мүмкіндіктік қатынас критерийлеріне қарай $W: \lambda(X) = tr(K_X K_0^{-1}) + \ln|K_0^{-1} K_0| > \lambda_0$ жағдайына қол жеткізетін боламыз, мұнда λ_0 – бастапқы шекті деңгей, онда α_0 мәнінің деңгейі қайта есептелінеді. Алынған нәтиже өзінің негізінде шешуші r статистикасына арналған теңдеуді қайталайды, алайда ол тек Кульбаку-Лейбер бойынша шарттық сәйкестік толықтығының көлеміне өз бетінше айқын танылған болады.

$$I_n[P_X|P_r] = 0.5[tr(K_X K_0^{-1}) + \ln|K_0^{-1} K_0| - n] \quad (1.16)$$

мұнда P_X және P_r – Кульбак Лейблер бойынша тәртіпке келтірілген қос Гаусстық бөліп орналастыру.

Қайсыбір еңбекте $n \to \infty$ жағдайындағы ассимптотикада тап осылай көрсетілген, сонда АКМ (автокорреляциялық матрица) $К_1$ жолақтық құрылымы бар $P(X_r)$ сигналының Гаусстық бөліп тартылуы жағдайында ұтымды шешуші статистика үшін теңдеу төмендегідей түрге келтіріледі.

$$\rho_{x,r} \triangleq \frac{1}{F}\sum_{f=1}^{F}\left(\frac{G_x(f)}{G_r(f)} + ln\frac{G_r(f)}{G_x(f)}\right) - 1 \to \min|_{r=\overline{1,R}} \quad (1.17)$$

мұнда $G_x(f)$ - f дискреттік жиілік қызметіндегі X сигналының спектрлік тығыздық қуаттылығын (СТҚ) таңдамалы бағалау; $G_r(f)$ – эталондар сөздігіндегі (СТҚ) r сигналы; F – сигналдың жиілік диапозонының жоғарғы шекарасы немесе пайдаланылатын байланыс арнасы. Оның өзі сөйлеу сигналының АР (авторегрессия) моделі негізіндегі АСМ критерийінің белгілі таңбалануы болып табылады:

$$x(n) = \sum_{i=1}^{P} a(i)x(n-i) + \varepsilon(n) \quad (1.18)$$

мұнда $x(n)$ – сөйлеу сигналының саналуының басталуының n мәні; ал $a = \{a(i)\}$ – оның АР (авторегрессия) коэфициенттерінің векторы, Р-АР (авторегрессия) моделінің тәртібі, ал $\varepsilon(n)$ – математикалық күту мен σ^2 тіркелген дисперциясының нөлдік мәні бар Гаусстық шуыл (ГШ) типінің тудырушы үдерісі.

Дисперсия және оларды туындататын шуыл бойынша ҚСБ типінің АР (авторегрессия) моделін қалыптасудың қосымша жағдайларында оң жағындағы екінші қосылғыш нөлге тең ұқсастықта болады, сөйтіп АСМ (1.19) шешуші статистикасы үшін теңдеу төмендегідей мейлінше қарапайым түрде иеленеді:

$$\rho_{x,r} = \frac{1}{F}\sum_{f=1}^{F}\frac{\left|1+\sum_{m=1}^{P} a_r(m)\exp(-j\pi mf/F\right|^2}{\left|1+\sum_{m=1}^{P} a_x(m)\exp(-j\pi mf/F\right|^2} \quad (1.19)$$

Бұл кірістегі X сигналы мен жиілік саласындағы сөздіктен алынған r сигналы арасындағы АСТК (1.11) таңдау бағалауының негізіндегі ПДБА міндетіндегі есебіндегі мәселелеріндегі залалсыздандыру сүзгісі әдісінің (ЗСӘ) стандарттық таңбалануы болып табылады. Ол ең алдымен, Берг және т.б. әдістері сияқты авторегрессиялық талдаудың жедел есептеу рәсімдері негізіндегі бейімделінген вариантта берілген (АСМ) критериінің анықталу артықшылығының тиімді жүзеге асуы болып саналады.

Сөйтіп, Кульбак-Лейберлердің АСМ критерийлеріне және Гельмгольцтің акустикалық құбыр теориясынан алынған АР моделіне негізделіне отырып, бұл бөлімде қарастырылған СҚАТ зерттеушіге төмендегідей негізгі нәтижелерге қол жеткізеді: жиілік саласындағы АСМ ұстанымы (1.19); шуыл туындататын дисперсия бойынша (1.17), (1.18) кірістегі сигналдардың қалыптандырылуы бар залалсыздандыру сүзгі әдісі; және ең бастысы өзінің «геометриялық» - орталықтың АО (ақпараттық орталық) эталонының айналасына АСМ критерийі бойынша біріккен бірегей ортақ бір типтік жүзеге асуының көптігі арқылы ҚСБ кластерлік моделін қатаң түрде анықтау.

Теориялық-ақпараттық тәсіл шеңберінде бейнелерді анықтау теориясының әдіснамалық аппаратын қолдана отырып, дауысы бойынша сөйлеушіні анықтау міндетін төмендегідей түсіндіруге болады: талданылатын үдерістің берілгендерін сегменттік өңдеуді жасай отырып, біз $\{x_r\}, r = \overline{1,R}$ жүзеге асырылуының эталондық модельдерінің жиынына және $\rho(x_1 x_2)$ қатаң айқындалған ұқсастық өлшеміне қол жеткіземіз. Сөйтіп эталондық модель мен X үдерісінің берілгендерінің сегментін талдауды өзара салыстырғанда болжамдарды көпбаламалы тексеру есебі шығарылады:

$$W_r(X): \rho_{X,V} = \rho(X, x\_r)|_{V=r} \to min, r = \overline{1,R} \qquad (1.20)$$

Берілгендердің талданылатын біртекті (стационарлық) сегменті саны алдын ала белгісіз $g(w_r)$-сынды қолда бар R>1 кластарының біреуінің элементі ретінде анықталады. Мұндай әрбір класс өзінің тұрақты $\overrightarrow{W_r}, r = \overline{1,R}$ сипаттарының (белгілерінің) көптігімен толық анықталынған бейне (кластер) түрінде беріледі. Сөйтіп қолданылатын ұқсастық өлшем талданылатын бейне мен $\overrightarrow{W_X} \approx \overrightarrow{W_v}, v \leq R$ (1.20) эталондық көп модельдің ішінен біреуі арасындағы эквиваленттік қатынасты белгілеуге мүмкіндік береді.

Бұл міндетті шешуде ең ұтымды болып статистикалық (байестік) тәсіл саналады. Бұл жағдайда $\overrightarrow{W_r}, r = \overline{1,R}$ белгілерінің көп жиыны $\{\overrightarrow{R^n}, \overrightarrow{W_r}\}$ гипотетикалық негізгі жиынтығын бақылаулардың көпөлшемді таңдауды бөлу заңы болып табылады, мұнда n – таңдау өлшемі таңдалынатын сигналдан саналынып алынатын сан, ал $\overrightarrow{R^n}$ – көп өлшемді евклид кеңістігі. Солайынша белгіленген міндет (1.20) берілгендердің талданылатын сегменттерін бөлудің белгісіз заңы туралы болжамды тексеруге әкеледі [36-41]. Стационарлық сипаттарын сақтайтын берілгендердің сегменттерінің есептерінің бар саны бойынша минимумдарын біз минимумдық сөйлеу бірліктері (МСБ) ретінде

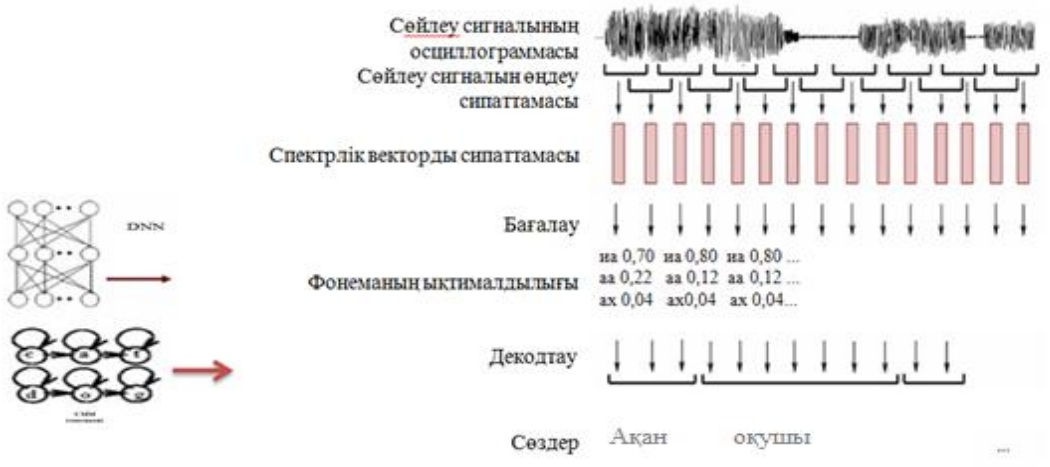
қабылдаймыз, оның өзі сигналдардың Гаусстық моделімен көрсетілген және кластар жиыны арқылы сипатталады. Мұндай тәсіл бірқатар артықшылықтарға ие, солардың ішінде: минимумдық сөйлеу бірліктерін анықтаудың дәлдігі мен сенімділігінің жеткілікті түрде жоғары екенін айтуға болады.

Жоғарыдағы қарастырылыған әдістер өзгерілмейтін парольдік сөз тіркестерін қолдана отырып, сөйлеушіні мәтіндік тәуелділік арқылы анықтауды жүзеге асыруға мүмкіндік береді. ҚСБ – ті қолдану әр түрлі сөйлеу парольдерін сөз тіркестеріне байланыссыз қолдануға яғни мәтінге тәуелсіз анықтауды қолдануға мүмкіндік береді.

1.4 Заманауи сөйлеулерді тану жүйелерінің архитектураларына шолу

Бұл бөлімде дәстүрлі сөйлеуді тану жүйесінің архитектурасына қысқаша шолу жасалады. 1.8 суретте сөйлеуді тану жүйесінің архитектурасы көрсетілген. Сөйлеуді тану жүйелерінің архитектурасын қарастырамыз. Олар төмендегідей болады:

- көп қабатты перцептрондар (Multilayer Perceptron; MLP) немесе терең нейрондық желілер (Deep Neural Networks; DNN);
- жиналмалы нейрондық желі (Convolutional Neural Networks), CNN);
- рекурренттік нейрондық желілер (recurrent Neural Networks, RNN);
- ұзақ және қысқа мерзімді жадысы бар нейрондық желілер (Long Short Term Memory; LSTM)];
- басқарылатын рекурренттік блок (gated Recurrent Unit; GRU);
- екі бағытты рекуррентті нейрондық желілер (Bidirectional Recurrent Neural Networks BRNNN)];
- қалдықты желілер (Residual Networks).



Сурет 1.8 - Сөйлеулерді тану жүйесінің архитектурасы

GMM/DNN архитектурасы

Сөйлеуді автоматты тану жүйесі сөйлеу сигналының белгілерін $X = (X_t \in \mathbb{R}^D | t = 1, \dots T)$ реттілікпен $W=(w_n \in v | n = 1, \dots N)$ сөйлеу тізбектеріне түрлендіруді орындайды, мұнда X_t- t кадріндегі D-өлшемдік сөйлеу белгісінің векторын, ал $w_n - v$ сөздіктегі n позициядағы сөзді білдіреді.

Кесте 1.1 – Әртүрлі тестілік жиындар үшін WER (%)

Моделдер	Шумсыз		Шуммен		Мәні
	Оқшауланған сөздер	Үздіксіз сөйлеу	Оқшауланған сөздер	Үздіксіз сөйлеу	
Негізгі бір бағытты контекстік тәуелді фонема	6.4	9.9	8.7	14.6	11.4
Негізгі екі бағытты контекстік тәуелді фонема	5.4	8.6	6.9	_	11.4
Интегралдық жүйе					
CTC графемасы	39.4	53.4			
RNN түрлендіргіш	6.6	12.8	8.5	22.0	9.9
Attention RNN түрлендіргіш	6.5	12.5	8.4	21.5	9.7
Attention 1-ші қабатты декодтау	6.6	11.7	8.7	20.6	9.0
Attention 2-ші қабатты декодтау	6.3	11.2	8.1	19.7	8.7

Сөйлеуді автоматты тану жүйесінің міндеттері Байес құрылымының шеңберінде қалыптасады. Сөйлеуді автоматты түрде тану мақсаты - s дыбыс сигналды W сөздер тізбегіне айналдыру. Бұл міндетті s кірістік сигнал бойынша сөздердің мейлінше [42] ықтималды тізбегінің ізденісі ретінде төмендегідей формулаға келтіруге болады:

$$\widehat{W} = \underset{W \in \delta}{\arg max}\, \mathrm{P}(W|S) \tag{1.21}$$

мұнда δ-гипотезалар жиыны.

Акустикалық модельдеу белгілер мен лингвистикалық бірліктердің, мысалы фонемалардың арасындағы статистикалық тәуелділіктерді құру үшін пайдаланылады. Тізбекті декодтау X белгілерінің тізбегін W сөздерінің тізбегіне айналдырады. Бұл қадамды Байес ережелерін пайдалана отырып төмендегідей сипаттауға болады.

$$\widehat{W} = \arg \max_{w \in \delta} \mathrm{P}_A(X|W)\mathrm{P}_L(W) \tag{1.22}$$

мұнда $\mathrm{P}_L(W)$ –тілдік модельдер арқылы алынатын априорлық ықтималдылық, ал $\mathrm{P}_A(X|W)$ – акустикалық модельдер негізіндегі ұқсас ақиқаттық функциясы.

Гибридтік HMM/DNN моделінде бақылаулардың ықтималдылығы нейрон желісі көмегімен есептеп шығарылады, сөйтіп S тізбегі қалпын P(W|X) факторизациялау үшін HMM-ді төмендегідей енгізеді.

$$\underset{w \in \delta}{argmax}\, P(W|X)$$

$$= \underset{w \in \delta}{arg\ max} \sum_S P(X|S,W)P(S|W)P(W) \tag{1.23}$$

$$\approx \underset{w \in \ \delta}{arg\ max} \sum_S P(X|S), P(S|W)P(W)$$

мұнда – $P(X|S)$ акустикалық модель, $P(S|W)$ – лексикон, $P(W)$ – тіл моделі, ал δ – сөздердің мүмкін болған тізбектерін көрсетеді.

1.5 Сөйлеуді танудың интегралдық әдісі

Сөйлеуді тану жағдайында интергралды тәсіл P(W|X) – ті «аумақты түрде түгелімен» есептеп шығаруға тырысады. Егер кірістік $X = (x_1, x_2, \dots, x_T)$ дыбыс белгілерінің тізбегі болатын болса, ал оған сәйкес келетін - $W = (x_1, x_2, \dots, x_T)$ сөздер тіркесі болады. Онда нейрон желісі $P(* |x_1), \dots, P(* |x_T)$, ықтималдығын есептеп шығарады. Мұнда ықтималдық көзі болып сөздер тізбегінің өзі емес, тек олардың жуық мәндері болады.

1.5.1 Шифрлеуші - дешифрлеуші механизмінің Attention-based моделі

Attention-based моделі ешқандай да Марковтық жорамалдар жасамайды. Модель тікелей p(C|X) апостериорлық ықтималдықты тауып отырды.

$$P(C|X) = \prod_{l=1}^{L} \underbrace{p(c_t|c_1, \dots, c_{l-1}, X}_{\Delta p_{at} C|X} \tag{1.24}$$

p_{att}(C|X) мұнда Attention-based мақсатты функциясын білдіреді. Берілген C- L ұзындығындағы символдар тізбегі. Ал $P(c_t|c_1, \dots, c_{l-1}, X)$ ықтималдылығы төмендегідей жолмен алынады:

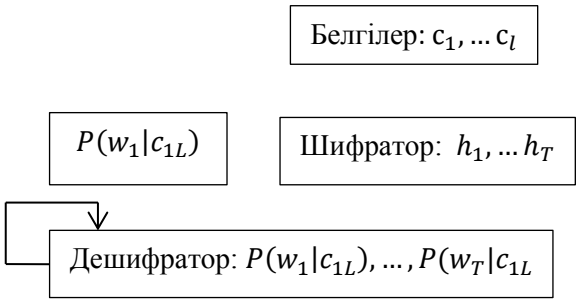
$$h_t = Encoder(X)$$

$$a_{lt} = \begin{cases} ContentAttention\ (q_{l-1}, h_t) \\ LocationArentin(\{a_{t-1}\}_{t-1}^{T}, q_{-1}, h_t) \end{cases} \tag{1.25}$$

$$r_l = \sum_{t=1}^{T} a_{lt} h_t$$

$$p(c_t|c_1, \dots, c_{l-1}, X) = Decor(r_1, q_{-1}, c_{t-1})$$

Жоғарыда қарастырылған теңдеу екі бөліктен тұрады: кодтау және декодтау. Берілген a_{lt} -теңдеуде h_t жасырын вектордың жұмсақ түзетпесін көрсетеді, сондай-ақ ол attention-нің ықпалы ретінде де белгілі. Бұл модельде r_1 жасырын векторлардың қосындылары жолымен есептеп шығарылатын символдар бойынша жасырын векторды білдіреді. Шифрлеуші кез келген нейрондық желі болуы мүмкін, мысалы: DNN, LSTM, BLSTM, CNN. 1.9- суретте модель схемасы бейнеленген.



Сурет 1.9 - Шифратор-дешифратор сөйлеуді тану жүйесі

Encoder желісі: X кірістік сөйлеу белгілерінің тізбегі h_t жасырын векторларға айналады. Кодтағыш желісі үшін артықшылық берілетін таңдау ол BLSTM яғни:

$$Enocoder(X) \triangleq BLSTM_t(X) \qquad (1.26)$$

Мұнда кодтаушының желісінің есептеп шығаратын күрделілігі, кірістік сигналдарды іштей таңдау есебінен төмендейді [43, 44].

Content-based attention mechanism: Content Attention(·) мына төмендегідей бейнеленеді:

$$e_{lt} = g^T \tanh\big(Lin(q_{l-1}) + LinB(h_t)\big) \qquad (1.27)$$
$$a_{lt} = Softmax(\{e_{lt}\}_{t=1}^{T})$$

мұнда g – оқытылатын параметрлер болып табылады, $\{e_{lt}\}_{t=1}^{T}$ – T өлшемді векторды білдіреді, ал солайынша tanh(·), Lin(·) жанамалық активациялық гиперболалық функцияны және сәйкесінше матрицаның оқытылатын параметрлері бар желілік (сызықтық) қабатты білдіреді.

Location-aware attention mechanism: Бұл контент негізіндегі attention-based функциясы орналасқан жеріне бағдарланған жұмысқа арналған. Берліген $a_{l-1} = \{a_{l-1}\}_{t=1}^{T} = [a_{l-1,1}, a_{l-1,2}, \dots, a_{l-1,T}]^T$ теңдеуді мынандай түрде жазуға болады $a_{l-1} = \{a_{l-1}\}_{t=1}^{T}$. Location Attention (·) бұны төмендегідей жаза аламыз:

$$\{f_t\}_{t=1}^{T} = K * a_{l-1}$$
$$e_{lt} = g^T \tanh(Lin(q_{l-1}) + Lin(f_t)) \qquad (1.28)$$
$$a_{lt} = softmax(\{e_{lt}\}_{t=1}^{T}))$$

мұнда $* \{e_{lt}\}_{t=1}^{T}$ белгілерінің T жиынын алу үшін K параметрмен бірге t кірістік белгінің өсінің бойындағы бір өлшемді 1-D жиынын білдіреді, сондай-ақ LinB(·) жылжыма векторларының параметрі бар метриканың оқытылған параметрлерімен бірге желілік (сызықтық) қабатты білдіреді.

Декодтау желісі: Декодтау желісі алдыңғы шығыстық C_{l-1} арқылы және q_{l-1} жасырын векторына тәуелдендірілген RNN-ді білдіреді. RNN-нің таңдаған таңдамасы болып табылатын LSTM төмендегідей берілген:

$$Decoder(\cdot) \triangleq softmax(LinB(LSTM_l(\cdot))$$

мұнда $\mathrm{LSTM}_1(\cdot)$ шығыстық ретінде q_l жасырын векторын жасайды

$$q_{1=LSTM=(T_t,q_6-1,C_l-1)}$$

r_l әріптер бойынша жасырын вектордың біріккен түрі (векторы), c_{l-1} – кірістік берілгендер ретінде танылатын алдыңғы қабаттық шығыстық берілгендерін білдіреді.

Оқыту функциясы: attention моделінің жаттығу функциясы төменде көрсетілген тізбектен есептеп шығарылады:

$$p_{att}(C|X) \approx \prod_{l=1}^{L} p(c_l|c_1^*, \dots, c_{l-1}^*, X) \triangleq p_{att}^*(C|X) \qquad (1.29)$$

мұнда c_1^*-алдыңғы символдардың басты ақиқаттығын білдіреді. Attention-ға негізделген тәсіл символдар комбинациясын білдіреді, ол әрбір L шығысында $c_1^*, \dots, c_{l-1}^*$ ақиқаттығының шартты тарихы бар көпкластық жіктеуге негізделеді, сөйтіп көрсетілген функцияны толық ескермейді.

1.5.2 Коннекциялық уақытша жіктеу негізіндегі модельдер (CTC)

Сөйлеуді танудағы нейрон желілері әдетте дыбыстық жазудың жекелеген фрагменттері көмегімен оқытылады. Бұл үшін әрбір кадрға сәйкес келетін жекелеген меткаларды бөліп көрсету талап етіледі, оның өзі дыбыстық жолмен транскрипцияларды түзету қажеттігін туындатады. Алайда түзетпе жасау тек нейрон желісін оқытудан кейін ғана сенімді, ал ол сегментация мен тану арасындағы циклдық тәуелділікке алып келеді. Сонымен қоса тек сөздердің транзакциясына ғана негізделген сөйлемді тану есептерінде бұл түзетпе пайда алып келмейді.

Коннекциялық уақытша классификация (Connectionist Temporal Classification; CTC) [45] - бұл кірістіктерді бастапқы және тізбекті шығыстықтарды түзетулерсіз сөздердің тізбегін тану үшін рекурренттік нейрон желілерінің оқытылуына мүмкіндік етеді.

Оқыту сатысы. CTC функциясы нейрон желісін оқыту үшін шығын функциясы ретінде пайдаланылатын функция болып табылатын тәсілді төмендегіше сипаттаймыз. Нейрон желісінің шығыс қабаты шығыстық тізбектің (әріптері, фонемалар, тыныс белгілері, ноталар) әрбір символы үшін бір-бір блокты және қосымша «рұқсаттама» («blank») символы үшін тағы бір блокты қамтиды, ол бос шығыстық символға сәйкес болады. w_m шығыс вектор softmax [46] функциясы көмегімен реттелінеді, ал ол *m* уақыты сәтінде *k* индексімен бірге

символдың (немесе «рұқсаттаманың») пайда болу ықтималдығы ретінде интерпретацияланады:

$$P(k, m\,|x) = \frac{\exp(w_m^k)}{\sum_{k=0}^{|w_m|} w_m^k} \quad (1.30)$$

мұнда kw_m^k – k-ның w_m элементі, ал а-w_m символының ұзындығы. Сонда: а-түзетпе үшінгі «рұқсаттама» индекстерінен және Т ұзындық символынан тұратын тізбек болды делік. Бұл жағдайда $P(\alpha\,|x)$ ықтималдығын уақыттың әрбір сәтіндегі символдардың пайда болу ықтималдығының өнімі ретінде түсінуге болады:

$$P(\alpha\,|x) = \prod_t^T P(\alpha_t, t|x) \quad (1.31)$$

Шығыстық w_m тізбегі үшін символдар арасына «рұқсаттамаларды» қойып шығу тәсілдері қанша болса мүмкін болатын түзетпелерде де сонша болады. Мұнда «−», «рұқсаттаманы» білдірсін делік. Мысалы, (а,-,б,в,-,-) және (-,-,а,-,б,в) түзетпелер (а,б,в) тізбегіне сәйкес келеді. Бірдей символдар тізбекпен пайда болған кезде, онда бұл қайталаулар жойылады: (а,б,б,в,в) және (а,-,б,-в,в) бұларға (а,б,в) сәйкес. В - оператор деп белгілесек, онда ол алдымен барлық қайталауларды, содан соң ««рұқсаттамаларды» жоятын болады. Осылайша w шығыстық тізбектің толық ықтималдығы барлық мүмкін болатын сәйкес түзетпелердің ықтималды жиынына тең болады:

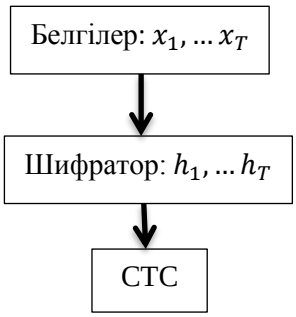
$$P(w\,|x) = \sum_{\alpha \in B^{-1}(w)} P(\alpha\,|x) \quad (1.32)$$

мұнда B^{-1}- оператор, ол В-ға кері барлық мүмкін болатын түзетпелер бойынша бұл жиын нейрон желісіне сегменттелмеген берілгендер арқылы жаттығуға мүмкіндік береді: Яғни меткалардың дәл орналасуын білмей тұрып, біз олардың пайда болу мүмкін орындардардың бәрін қосамыз. Бұл жиын динамикалық программа көмегімен есептеп шығарылады. W*- сөздердің мақсатты тізбегі болсын делік, сондай нейрон желісі CTC функциясын минималдауға оқытыла алатын болады.

$$CTC(x) = -logP(w^*\,|x) \quad (1.33)$$

Нейрон желісі градиентті пайдаланатын кез келген оңтайландырылған алгоритм көмегімен оқытыла алады. 1.10 суретте модельдің CTC сызбасы берілген, мұнда шифрлеуші DNN, LSTM, BLSTM, GNN немесе кез келген нейрон желісінің түрі бола алады. [46] еңбекте CTC тура және кері жүрісті алгоритм ұсынылған, өз кезегінде CMM үшін тура – кері жүрістік алгоритмге ұқсас динамикалық программалау алгоритмін қолданады [47]. Бұл алгоритмінің басты идеясы – барлық түзетпелер жиыны олардың өзінің шығыстық тізбектерінің сәйкес префикстерінің түзетпелері бойынша жиынға бөлу. Бұл

жиын рекурсивті тура және кері айнымалылар көмегімен тиімді есептелініп шығарыла алады.



Сурет 1.10 - CTC сөйлеуді тану жүйесі

Кодтау кезеңі. [46] еңбекте интегралды CTC модельдердің декодтаудың екі нұсқасы ұсынылған. Бірінші әдіс (шығыстық тізбекті ең жақсы түзету жолын іздестіруде) жорамалдауға негізделген, мұнда мейлінше ықтималды түзету ең ықтималды шығыстық тізбекке сәйкес:

$$\underset{w}{argmax}\, P(w \dashv | x) \approx B(\alpha^*) \tag{1.34}$$

мұнда $\alpha^* = \underset{\alpha}{argmax}\, P(\alpha \mid x)$.

Ең жақсы түзетуді есептеп шығару қарапайым есептеу болып саналады, өйткені α^*дегеніміз – әрбір уақыттық қадамындағы мейлінше белсенді шығыстар конкатенациясы. Алайда бұл сөздер тіркесінің мейлінше ықтималдығын табуға кепіл бола алмайды.

CTC Марковтық моделін пайдаланады. CTC сөйлеуді танудың интегралды жүйесінің артықшылықтарын толық пайдаланбайды, бірақ та оның символдардың шығысын көрсетуі әліде болса интегралды артықшылықтарға ие болуда.

2 МАШИНАЛЫҚ ОҚЫТУ АЛГОРИТМДЕРІ МЕН МОДЕЛЬДЕРІН СӨЙЛЕУДІ ТАНУ ЕСЕПЕТЕРІНДЕ ҚОЛДАНУ

2.1 Машиналық оқытудағы нейрондық желілер

Қазіргі таңда машиналық оқыту алгоритмдерін қолдану кең етек алуда. Машиналық оқыту алгоритмдері бұл деректер үлгілерін іздеу және тиісінше бағдарлама әрекеттерін өзгерту үшін қолданады.

Машиналық оқу (Machine Learning) әдісі – эмпирикалық мәліметтердегі заңдылықтарды анықтау үшін оқуға қабілетті алгоритмдерді құру әдістерін зерттейтін жасанды интелект саласы.

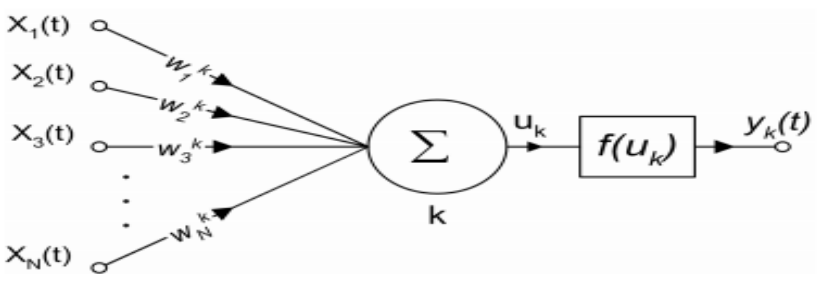
Қазіргі уақытта сөйлеуді тану өзекті проблема болып отыр. Осы проблеманы шешу мақсатында қолданылатын көптеген заманауи әдістер кең ауқымды есептеуіш ресурстарын талап етеді, ал көп жағдайда ол ресурстардың көлемі шектеулі.

Соңғы уақытта танудың кейбір кезеңдері үшін нейрондық желілерге негізделген әдістер барған сайын жиі қолданылуда. Осы бетбұрыстың себебі нейрондық желілерді қолданудың қарапайымдылығымен, сондай-ақ жұмыс жылдамдығының жеткілікті түрде жоғары болуымен түсіндіріледі.

Сөйлеу сигналын акустикалық-фонетикалық модельдеу үшін қолданылатын модельдердің басқа класы болып жасанды нейрон желілерінің модельдері саналады. Оның құрылымы мен жұмыс істеу принциптері нерв жүйелерінің биологиялық модельдеріне, әсіресе ми моделіне негізделеді. Нейрон желілері өз бетінше ұйымдастырылып, реттілігіне қарай алгоритмдердің көптүрлігі ретінде қаралады, өзара байланысқан көптеген бір типті және бір қатарда қызмет ететін элементтер немесе нейрондық және арнайы ұйымдасқан байланыстар көмегімен «сыртқы әлеммен» тұтасқан құбылыс болып танылады.

Уақыттың дискреттік сәттерінде кіріс байланыстары бойынша нейронға ақпарат беріледі, оның негізінде кейбір принциптерге сәйкес шығыс сигналы қалыптасады, ал ол өз кезегінде басқа нейрондардың кірістеріне беріледі. Сөйтіп нейрон желісінің негізгі элементі нейрон болып саналады.

1943 жылы ұсынған [47-49] Мак Каллока-Питсаның (2.1-сурет) нейрон моделі ең көп таралған модель болып есептеледі, соған сәйкес нейронда кіріс байланыстарының жиыны пайда болды, ал шығыс біреу ғана және ол параллельдене алады.



Сурет 2.1 – Нейрондық модель

Бұл модель қызметі келесідей: нейрон шығысына $x(t) = x_1(t), \dots, x_i(t) = \{x_1(t), \dots, x_N(t)\}$ кіріс векторлары беріледі, ал ол $w_k = \{w_{1k}, w_{2k}, \dots, w_{Nk}\}$ сенімді өлшемдік векторға көбейтіледі немесе басқаша айтқанда, $x_i(t)$ векторларының коэфициентері w_{Nk} сенімді коэфициенттерімен өлшенеді, ол төмендегі формулаға сәйкес туындайды:

$$u_k(t) = \sum_{i=0}^{N} w_{ik} x_i(t) \quad (2.1)$$

Нейрондық шығыс сигналы келесі түрде анықталады

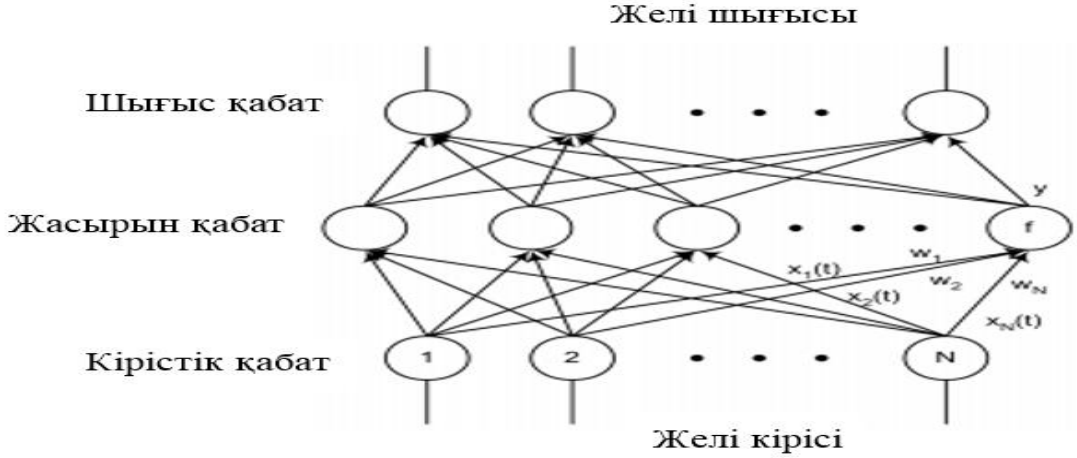
$$y_k = f(u_k(t)) \quad (2.2)$$

мұнда $f(u_k(t))$-нейрондық белсенділендіру функциясы. Белсенділендіру функциясы ретінде көп жағдайда желілі емес үздіксіз қызмет таңдалынады, мысалы сигмоидалдық функция төмендегідей:

$$f(x) = \frac{1}{1+e^{-\propto x}} \quad (2.3)$$

мұнда α-белсендіру функциясының түріне әсер ететін және пайдаланушы таңдайтын кейбір параметр.

1980 жылдардың соңынан бастап кейбір зерттеушілер тану жүйелерінде нейрон желілерінің моделдерін белсенді қолдана бастады. Бұл нейрон желілерінің көмегімен сөйлеуді тануға арналған жұмыстардың санына ықпал етті, ол бірнеше есе артты. 1989 жылы Lippmann 80 – жылдардың аяғына қарайғы сөйлеуді танудағы нейрон желілерінің модельдерінің жай – күйі туралы шолу жазды. Құрылымдық схемасы 2.2 – суретте көрсетілген көпқабатты перспетрон (КП) ең белгілі және мейлінше кең таралған нейрон желінің моделі болып табылады [50].

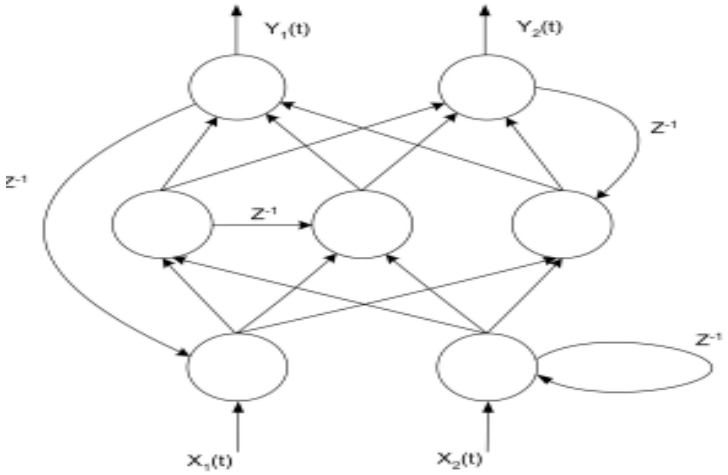


Сурет 2.2 – Көпқабатты перспетрон

Көпқабатты персептрон элементтері бірнеше қабаттарға бөлінген, ал қабаттың ішіндегі элементтерді желілі түрде реттелген және олар өзара ықпалсыз деп санауға болады. Желінің әрбір нейроны (кіріс қабатының нейрондарынан – перецепторларынан – басқа) кіріс сигналын алады, сөйтіп нейронның шығыс сигналы (соңғы қабаттардан басқа) келесі қабаттың нейрондарының кірісіне түседі. Осылайша КП сигналдың тек алға (кері байланыссыз) желінің кірісімен шығысына қарай (кері байланыссыз) таралуын қамтамасыз ететін байланыстары бар модель болып табылады. Аралық қабаттардың элементтері жасырын эелементтер деп, ал қабаттары жасырын қабат деп аталады. Нейрондардың өздері көп жағдайда МакКаллока-Питса моделімен сәйкес қызмет етеді, белсендіру функциясы ретінде (2.3) – формуладағы сигмондалдық функциясы таңдалынады.

2.1.1 Рекурренттік нейрондық желілер (RRN)

Контекстік ақпаратты пайдаланудың басқа бір тәсілі өздерінің желідегі топологиясына тәуелсіз түрде еркін нейрондардың арасына байланыстар енгізуден тұрады. Алайда желі КП болып қала беруі үшін бұл байланыстар уақытша бір қадам тежелуі керек болады. Бұл рекурренттік байланыс деп аталады. Мұндай желінің құрылымдық үлгісі 2.3-суретте көрсетілген.



Сурет 2.3 – Рекурренттік желі (RRN)

Сөйтіп нейрондардың белсенділігіне және нейрондардың алдыңғы уақыттағы қадамының белсенділігіне тәуелді екеніне көз жеткіземіз. Мұндай желі рекурренттік желі [51] немесе динамикалық [52] нейрондық желі деп аталады.

Бастапқыда RNN өзінің оқуы, талдауы және әзірленуі жағынан үлкен күрделігіне байланысты сөйлеуді тану жүйесі үшін сирек қолданылады. Дегенмен бірқатар зерттеулер нәтижесінде ВР алгоритмін жетілдірудің бірнеше нұсқасы ұсынылады. Атап айтқанда: рекурренттік ВР [53], тізбекті ВР [54], іс жүзіндегі уақыт аралығындағы рекурренттік оқу [55] уақытқа тәуелді

рекурренттік ВР алгоритмі [56,57] және сөйлеуді тану жүйелерінде рекурренттік құрылымдарды пайдалануды едәуір жеңілдеткен мейлінше кең тараған уақыттық ВР. Мұндай көпқабаттық персепрондық оқу алгоритмдерінің жетілдірулері фонемалар сияқты қысқа уақыттық акустикалық-фонетикалық бірліктерді тану сапасын арттырады, алайда сөз сияқты лингвистикалық бірліктерді ұсыну үшін қажетті тануды аз ғана жақсартады. Бұл нәтиженің теориялық негіздемесі [58] әдебиетте жасанды нейрондық желі (ЖНЖ)-ні негізгі құрылым етуге мүмкіндік бермейтін едуір маңызды кемшіліктерді анықтады. Олар:

- ЖНЖ-нің уақыттық вариативтілікке және сөйлеу сигналының тізбектілік табиғатына ұқсас көрсететін тетіктері жоқ;
- Қазіргі кезде ЖНЖ-нің динамикасы мен топологиясын анықтайтын параметрелердің көптеген қатары үшін сол параметрлерді есептеп шығаруға немесе таңдап алуға мүмкіндік беретін теориялық негіздер жоқ (олар әзірлеушінің құрастыруына қарай таңдалынады);
- Оқу рәсімін жеңілдететін алгоритм әзірленсе де, ЖНЖ ресурстарды өте көп қажетсінетін рәсім болып отыр.

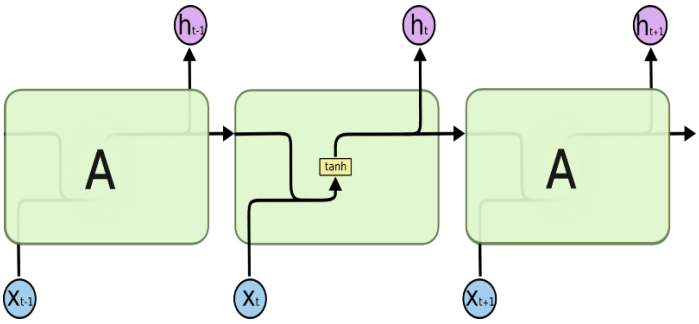
Нейрон желісі градиентті пайдаланатын кез-келген ұтымды алгоритм көмегімен оқытылады. Атап айтсақ LSTM. LSTM – көптеген мәселелерде стандартты нұсқасынан едәуір артықша рекурренттік нейрон желісінің әдеттен тыс жетілдірілуі. RNN-нің таңданарлық нәтижелерінің барлығы дерлік LSTM арқылы алынады.

2.1.2 LSTM желілері

Ұзақ мерзімді жад (Long short-term memory; LSTM) – ұзақ уақыт тәуелділікке оқу үшін қабілетті болатын рекурренттік нейрондық желілердің ерекше архитектурасы. Оларды 1997 жылы Зеппом Хохрайтер и Юргеном Шмидхубером (Jürgen Schmidhuber) ұсынды, ал содан кейін барлық басқа зерттеушілердің еңбектерінде жетілдіріліп, танымал болды. Олар көптеген әртүрлі міндеттерді жақсы шеше алады және қазіргі уақытта кең қолданыста.

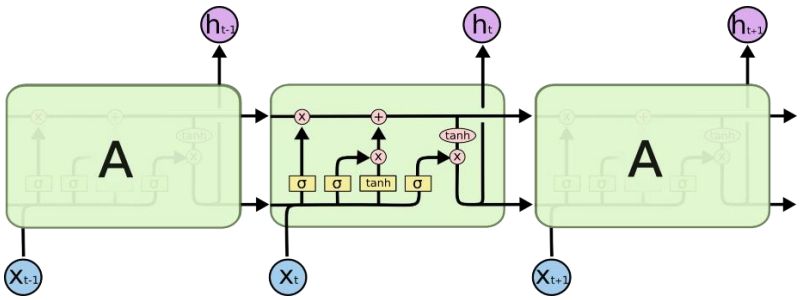
LSTM ұзақ уақытқа тәуелді болудан арылу үшін арнайы әзірленген Уақыттың ұзақ кезеңдеріне ақпаратты жадында сақтау – бұл олардың әдепкі амалы, бұл үшін ол күштеп оқуды қажетсінбейді.

Кез-келген RNN нейрон желісінің қайталанып келген модульдерінің тізбегінің формасына ие. Әдепкі RNN – де мұндай бір модульдің құрылымы өте қарапайым, мысалы ол tanh (гипорболалық тангенс) активациялық қызметі бар болатын бір қабат ретінде көрінеді.



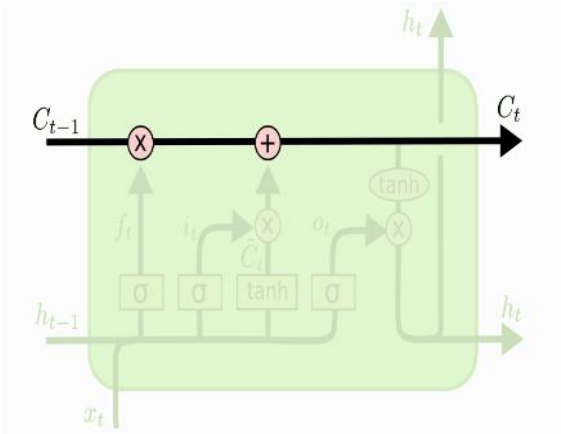
Стандартты RNN-де қайталанатын модуль бір ғана қабаттан тұрады

LSTM құрылымы, сондай-ақ тізбекті елестетеді, бірақ модульдер басқаша болады. Нейрон желісінің бір қалыпты орнына олар төртеуден тұрады және бұл қабаттар ерекше сипатта өзара әрекеттестікте болады.



LSTM-нің негізгі идеясы

LSTM – нің негізгі компоненті - яғни ол ұяның (call stste) жай-күйі, ол схеманың жоғарғы бөлігі бойынша өткен көлденең сызық. Ұяның жай – күйі ұласпалы жолақтарды еске салады. Ол бүкіл тізбек бойынша түзу өтеді, тек бірнеше қайта түзілімдерге ғана қатысады. Ақпарат өзгерістерге ұшырамай-ақ, оның бойымен жеңіл түрде ағып өте алады.



Сонымен бірге LSTM ұяның бойынан ақпаратты жоя алады; бұл үдеріс сүзгі (gates) деп аталатын құрылыммен реттелінеді. Олар сигмоидтық нейрон операцияларынан тұрады.

2.2 Сөйлеуді тану және белгілерін анықтауға арналған акустикалық корпустар немесе деректер қоры

Сөйлеу корпустары яғни мәліметтердің сөйлеу базалары – тіл ресурстарының маңызды бір түрі. Корпус құрамында тілдік, соның ішінде фонетикалық ресурстарды жасау, жинау, ұйымдастыру және басқару амалдарын атқаратын компьютерлік бағдарламалар болады.

Сөйлеуді автоматты тану саласындағы жетістіктер сөйлеу корпустарына қызығушылықты арттырды. Мұнда зерттеушілер тілдің дыбыстың бірліктерінің аса көп акустикалық нұсқаларымен ұшырасады. Ол контекстік вариантылық жүйесінен бастап сөйлеушінің психологиялық – физиологялық жай-күйіне дейінгі және сөйлеу материалын жазуда қолданылатын микрофонның техникалық сипаттары сияқты көздері иеленеді. Қазіргі тану жүйелері көптеген дикторлардан (100 адамнан кем емес) жазылған естілетін аса үлкен ауқымды сөйлеуден оқылады. Соңғы он жылда ережелер мен алгоритмдерді «қолмен» жасаудан корпустық модельдеуге көшу байқалады, бұл сөйлеуді автоматты синтездеу саласында да көрініс табады. Ол сөйлеудің просодикалық сипаттарын, оның көңіл-күй мазмұны мен реңдерін, сондай-ақ сөйлеушінің дауысының жеке ерекшеліктерін бейіндеуді модельдеуде аса маңызды болып табылады. Сөйлеу корпустары өз алдын да ғылыми қызығушылыққа ие, оған деген сұраныс әр түрлі тілдерде естілетін сөйлеуді талдау мен сипаттауға байланысты көптеген ғылыми мәселелерінде көрініс тауып отыр.

Сөйлеу корпустары дыбыстық және соған сәйкес мәтіндік файлдар түрінде көрінеді [59]. Дыбыстық файлдар сөйлеу элементтерін (дыбыстық, буындар, сөздер тіркестер) қамтиды, мәтіндік файлдарда тиісті сәйкес транкрипциялар орналасады. Әдетте сөйлеу корпустарының құрамы файлдарды редакторлар, транскрипциялауды автоматтандыру т.б.с.с. амалдарды қамтамасыз ететін көмекші бағдарламалық құралдар болады. Сөйлеу корпустарына сондай-ақ әр түрлі қызметтік ақпараттар (пайдалану бойынша нұсқаулар, бағдарламалардың бастапқы мәтіндері т.б.) бар файлдарда ендіріледі. Сөйтіп сөйлеуді тану үрдісінде сөйлеу корпусы ресурстарының қорының болуы аса маңызды болып табылатынын көреміз.

Сөйлеу корпусы – аудиофайлдар мәліметтерінің және мәтіндер транскрипцияларының базасы, ол сондай-ақ және мәтіндер корпустарының әр түрлілігі. Сөйлеу технологияларында сөйлеу корпустары, басқаларын айтпағанның өзінде, акустикалық модельдерді (кейінде сөйлеуді тану тетіктерінде қолданыс табуы мүмкін) жасау үшін пайдаланылады. Лингивистикада сөйлеу корпустары фонетиканы, диалектологияны, конверсациялық талдауларды т.б. салаларды зерттеуде қолданылады.

Сөйлеу корпустарының екі түрі болады:

1. Оқылған мәтіндер базасы, соның ішінде:
   - кітаптар мәтіндері;
   - жаңалықтардың сөйлеу арқылы тасымалдануы;
   - сөздер тізімі;
   - сандар тізбектілігі;

2. Қолма-қол айтылған сөйлеу базасы, соның ішінде:

- диалогтар яғни екі немесе одан көп адамдардың өзара әңгімелесуі;
- ауыз екі әңгімелер (мысалы, Buckeye Corpus);
- картографиялық түсіндірмелер – мұнда бір адам карта бойынша басқаларға маршрутты яғни жол бағыттарын түсіндіреді;
- тағайындау міндеттері, мұнда белгілер бір графикке негізделген кездесудің жалпы уақытын табу үшін екі адамның әрекеттері беріледі.

Сөйлеу копустарының ерекше бір түрі – тіл және шет тілдердің акценттерімен сөйлейтін адамдардың сөйлеуінен түзілген мәтіндер базалары.

Сөйлеу корпусы – сөйлеу үздіктерінің құрылымдалған жиыны, ол өзіне қолжетімділік үшін бағдарламалық құралдармен қамтамасыз етіледі. Сөйлеу үздіктері – сөйлеу сигналының сандық үздіктері болып табылатын базалық бірлік, ол белгілі бір типтердің ақпараттарымен бірге көрінеді. Қазіргі уақытта компьютерлік қосымшалар үшін және іргелі фонетикалық зерттеулер үшін үлкен, әр түрлі және ақпараттық «бай» (көпдеңгейлі) сөйлеу корпустарын және оларды әзірлеудің ыңғайлы әрі сенімді құралдарын жасау міндеттері барған сайын өзектілікке ие болуда. Сөйлеуді танудың мейлінше жоғары көрсеткішті қазіргі жүйелері көбінесе сөйлеу және тілдік құбылыстарды статикалық модельдеу әдістеріне негізделеді, сөйтіп көп диктордан (100 адамнан кем емес) жазылған үлкен ауқымдағы бағыттамасы берілген естілетін сөйлеу бойынша оқуды талап етеді.

Әртүрі өлшемегі акустикалық үзіктердің конкатенациясына негізделген мәтін бойынша сөйлеуді синтездеудің қазіргі тәсілдері де үлкен сөйлеу корпустарын пайдалануды көздейді [60]. Мамандар корпустың тәсілін (corpus-based approach) синтез технологиясының дамуы үшін, әсіресе, сөйлеудің просодикалық сипаттары мен сөйлеушінің жеке ерекшеліктерін модельдеуде негізгі анықтаушы санайды. Зерттеулер сондай-ақ бұл тәсілдің оқудың рәсімдерін форматтау, пайда болатын және бақыланатын қателерді түзету жүзеге асыратын интерактивтік оқыту үдерісін қолдану, стандарттандырылған негізінде (бір сөйлеу корпусының өз ішінде) әртүрлі қолданбалы жүйелер жұмысын бақылау мүмкіндігі мен объективті бағалау сияқты артықшылықтарын көрсетеді. Практика көрсеткендей, сөйлеу корпустары мен оқыту технологияларының жеткілікті жағдайында автоматты танудың прототиптік нұсқасын немесе сөйлеу синтезаторларын жасау айтарлықтай көп уақыт алмайды. Әдебиеттерде оның мерзімі екі айдан жарты жылға дейін көрсетілген. Ал, коммерциялық мақсатқа бағытталған әзірлеме үшін бұл маңызы жағдай.

Сөйлеу корпустары тек сөйлеуді тану технологиясының дамуы үшін ғана қызықты деу дұрыс болмаған болар еді. Көрнекті сөйлеу корпустарын қолдану, қазіргі сөйлеу технологияларының даму деңгейі және компьютерлік техниканың қуатының тұрақты түрде үдемелі өсуі ғалымдарға әр түрлі тілдік материалдар бойынша ірі ауқымды және статикалық сенімді фонетикалық зерттеулер жасауға мүмкіндік береді.

Мәліметтердің акустикалық базасы немесе акустикалық корпустар статистикалық әдістер мен машиналық оқытудың алгоритмдеріне негізделген

сөйлеуді тану мен синтездеу жүйелерін жасауда аса қажетті ресурсқа айналады. Мұндай базалардың негізгі қызметі – тілдегі дыбыстардың фонетикалық әр түрлілігін ескеретін белгілі бір тілдің акустикалық моделін жасау мүмкіндігі болып табылады [61].

Мәліметтердің берілген базасын жасаудың негізгі мақсаты қазақ тілінің акустикалық моделін жасауда қолдану. Бұл әзірленіп жатқан интеллектуалдық дауыс жүйесінде тану мен синтездеудің негізгі моделі болып табылады.

Бұл берілген жұмыста сқйлеушіні анықтаудың акустикалық корпусы жасалған. Ол әртүрлі стильдік жанрлардағы қазақ тілінің мәтіндік корпусынан арнайы алынып оқылатын сөйлемдерден тұрады. Дауыстауға арналған 29805 сөзден астам сөйлеулерден тұратын мәтіндік материал мәтіндік корпустан мұқият сұрыпталып алынып, топтастырылған. Барлық мәтіндік материал дикторынан оқуға тиіс жекелеген кешендерге бөлшектендірілген. Диктордың бір кешені 75 сөйлемнен тұрады.

Жиналған мәтіндік материалдарды дауыстау үшін дикторлар жастарының белгілеріне қарай тартылады.

Тілдің акустикалық корпусын жасау үшін акустикалық мәліметтерді жинау Алматы қаласының ҚР БҒМ ҒК Ақпараттық және есептеуіш технологиялар институтында жүргізілді. Ол үшін Vocalbooth.com фирмасының кәсіби мамандандырылған дыбыс жазатын, шудан оқшаудандырылған студиясы пайдаланылады (2.4-сурет). Аудио берілгендерді жазу кабинасы шудан оқшаулайтын екі қабаттан және сондай есіктен тұрады. Студияның ішкі қабаты пирамида бейнелі дыбыс жұтатын қызыл түсті акустикалық материалмен жабдықталған және кабина шу өткізбейтін ауа айналымы жүйесімен қамтамасыз етілген [98]. Студия жоғарғы сапада аудио берілгендер жазуға арналған.

Сурет 2.4 – Vocalbooth.com -фирманың шудан оқшаудандырылған кәсіби дыбыс жазу студиясы

Жазылған аудио материалдар .wav кеңейтілген түрде сақталынады. Әрбір сөйлем дербес файл ретінде сақталынып, ал оның атауы төмендегідей идентификаторлардан таңдалынып алынады:

<аймақ _коды> + <жынысы> + <туған_жылы> + <тектері _ТАӘ> + <білім _коды> + <мәтін _нөмірі>, <мәтіндегі _сөйлем _нөмірі>

Мысалы: диктор Оңтүстік Қазақстан облысының, Шардара қаласының тумасы, аты-жөні Ермекбай Қабылбек Шалқарұлы, ер жынысты, 1998 жылы туылған, жоғары білімді, академиялық дәрежесі бакалавр, 7-шы нөмірлі мәтінді және 85 сөйлемді оқыды, идентивтенуі 05M90MT3_T007_S085.

Барлық аудиоматериалдар бірыңғай сипатқа ие:

- файлдық кеңейтілу: .wav;
- сандық түрге айналдыру әдісі: PCM;
- дикреттік жиілігі: 8 кГц;
- разрядтығы: 16 бит;
- аудио арналардың саны: біреу (моно);

Дикторлар ретінде сөйлеу кезінде ешқандай мүкістігі жоқ адамдар таңдалып алынады. Ғылыми – зерттеушілік мақсаттар және бұдан әріде ол мәліметтерді пайдалану үшін алдын ала жасалған қалып бойынша дикторлармен сауалнамалар жүргізілді (2.5 сурет).

Жазу үшін әртүрлі жастағы (18 жастан 50 жасқа дейін) және әр жыныстық 212 диктордың сөйлеуі пайдаланылды. Бір диктордыдауыстау және жазу орташа есеппен 40-50 минут уақытты алды. Әрбір диктор үшін жеке-жеке файлдарға жазылған 100 сөйлемнен тұратын мәтін дайындалды. Әрбір сөйлем орташа есеппен 6-8 сөзден құралды. Сөйлемдер фонемалары мейлінше бай сөздерден таңдалынды. Мәтіндік мәліметтер қазақ тіліндегі жаңалықтар сайттарынан («Егемен Қазақстан», «Алматы ақшамы» т.б.) жиналды. Сондай-ақ электрондық түрдегі басқа да материалдар пайдаланылды. Барлығы 36 сағаттық аудио мәліметтер жазылды. Жазу кезінде транскрипциялар жасалды: мәтіндік файлдағы әрбір аудио файлдың сипаттамасы.

Диктор сауалнамасы

№	Тегі	Аты	Әкесінің аты
1	Ермекбай	Қабылбек	Шалқарұлы
2	Жынысы*:	Ер ✓	Әйел
3	Ұлты:	қазақ	
4	Туған жері:	Оңтүстік Қазақстан Облысы Шардара қаласы	
5	Туған жылы:	07.08.1998ж	
6	Тұрғылықты жері:	Алматы қ. Навои 37.	
7	Тел:	+7707-789-33-34	
8	e-mail:	yerlekbayk@mail.ru	
9	Білімі*:	Орта / Арнайы орта / Аяқталмаған жоғары / Жоғары ✓	
10	Академиялық дәрежесі*:	Бакалавр ✓ / Магистр / PhD	
11	Мамандығы:	математика	
12	Күні:	20.02.2018.	
13	Қолы:	Ермекбай.	

Ескерту қажет болған жағдайда астын сызыңыз.*

Сурет 2.5 – Дикторлармен сауалнама жүргізуге арналған үлгі қалыпты толтыру үлгісі

Жазу сессиясы.

Жазудың әрбір сессиясы алдында дикторға дайындалу үшін мәтіндік материал, сондай-ақ төмендегідей базалық нұсқаулар беріледі:

- жазу сессиясының алдында өзінің тегін атап айту;
- дауыстап қатесіз оқу (бірқалыпты күйде немесе диктордың қалауынша интонацияны түрлендіру);
- қысқартылған сөздерді дикторлар өз қалауынша оқиды: қысқарған қалыпта немесе қысқарғанды таратып оқиды (мысалы «ҚР ҒБМ» немесе «Қазақстан республикасының білім және ғылым министрлігі»;
- дикторлар тыныс белгілерін ескеруге, қажетті паузалар жасап отыруға тиіс.
- мәтіндегі тырнақшаларға көңіл бөлінбейді, олар еленбейді;
- адамдардың инициалдары егер оның толық түрін диктор білмейтін болса ескерілмейді, тек фамилиясы ғана оқылады.

Студияда жазу кезінде дыбыс операторынада қатысты, оның қолында диктордағыдай материал болады, сөйтіп барлық үдерісті бақылайды, керек болған жағдайда дикторларды түзетіп отырады. Бұдан басқа диктордың студия жағдайларына бейімделу үшін және өзінің оқуын бақылау үшін олар алдын ала тестілік жазудан өткізіледі.

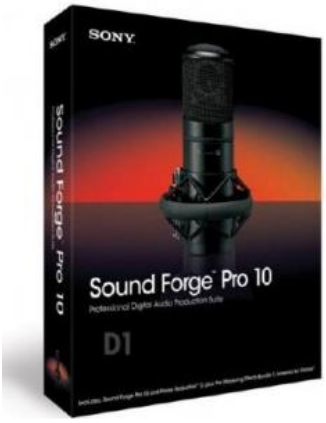
Жоғарыда аталған жұмыстардан кейін өте маңызды элементтердің бірі жасалынды. Ол – сөйлеуді тану жүйесіне арналған сөздік базасы. Барлық жазылған мәтіндер бір файлға жиналды, сосын қайталанатын сөздер жойылды. Содан соң олар алфавиттік ретпен сұрыпталып, фонемаларға көшірілді. Жасалынған сөздік фрагменттері 2.6-суретте көрсетілген. Оқылған сөздердің сөздігі қайталанбайтын 29000 аса сөзден құралды және олардың фонема түріндегі фонетикалық транскрипциясы берілді. Акустикалық сөйлеудің 36 сағатын және дискідегі 7 Гб жадын құрайды.

абай	a b a i
абайтану	a b a i t a n u
абылай	a b y l a i
абыз	a b y z
абзал	a b z a l
абақты	a b a k t y
абысын	a b y s y n
абажа	a b a zh a

Сурет 2.6 – Қазақ тілін тану жүйесіне арналған сөздік база фрагменті (үлгі)

Бұл жазу жеке оқуға арнайы арналған бағдарламадағы SONY5x кәсіби микрофон арқылы жүзеге асты. Волкалдық микрофон ретінде ол дауыстарға нәр, реңк және өткірлік сипаттарын беріп, тазалығы мен теңгермелілігін қамтамасыз

етеді. Сөйлеуді санға айландыру *Sony Sound Forge* Audio Studio сыртқы дыбыстық кәсіптік аудио-редактор көмегімен жасалынды.



SONY5x кәсіби микрофон

Sony Sound Forge Audio Studio – аудиоредактордың қуатты жаңа түрі, оның құрамына дауыспен жұмыс істеуге арналған әр түрлі утилеттер жиыны кіреді. Осы бағдарлама көмегімен сөздерді аудио файлдарды тиімді өңдей аламыз, сэмплдерді реадакциялай аламыз, дыбыстарды жаза аламыз, дыбыстық жазуларға көптеген эффектілерді енгіземіз, аудио мәліметтерді кодтай аламыз, аудио файлды бір форматтан екінші форматқа түрлендіре аламыз т.б.

Бағдарлама мүмкіндіктері:

- аудионы редакциялау бойынша барлық стандарттық мүмкіндіктер;
- 24 бит және 32/64 бит 192 кГц файлдарды қолдану;
- моно, стеро және көпарналы аудио файлдарды редакциялау;
- 20 – дан астам ендірілген DirectX аудио-плагин;
- көп арналы редакциялау мен өңдеу;
- әрекеттерді жою мен қайталаудың шектелмеген деңгейлері; және т.б.;

Бұл жұмыста біз қазақ тіліндегі сөйлеушіні анықтаудың акустикалық корпусын қалыптастырдық:

- Мәтіндік материалдарды жинау және дайындау;
- Дикторларды іздеу және іріктеу;
- Дикторларды жазу;
- Акустикалық деректерді өңдеу және құрылымдау;
- Акустикалық корпусты белгілеу және транскрипциялау.

Қазақ тілінің қалыптасқан акустикалық корпусы бүгінгі күннің алғашқы түрлерінің бірі болып табылады және осы зерттеу шеңберінде де, әртүрлі салалардан тыс зерттеушілер үшін де үлкен мүмкіндіктерге ие.

Жиналған акустикалық корпус негізінде қазақ тілінің дыбыстарын фонетикалық - акустикалық талдау және оларды басқа тілдердің дыбыстарымен салыстыру жүргізілді. Жұмыс нәтижелері қазақ тілін тану жүйесін құру үшін Қазақ тілі дыбыстарының физикалық сипаттамаларын тереңірек ашады, сондай-ақ қазақ дыбыстарын халықаралық фонетикалық алфавит нышандарында

көрсетуге мүмкіндік береді, бұл Қазақстанның ғылыми ортасындағы бірегей жетістік болып табылады.

2.3 Сөйлеу белгілерін анықтауға арналған классификациялық алгоритмдер

Ақпараттандыру дәуірінде көптеген жоғары технологиялық өнімдер біздің күнделікті өмірімізге енуде, сөйтіп өмірдегі әдепкі дағдыларды едәуір өзгертуде. Екінші жағынан, ақпараттық технологиялар адамға бағдарланған бағытта дамуын жалғастыруда. Нақты адамдарды танудың мейлінше қарапайым және ыңғайлы әдістерін біздерге биометриялық тану технологиясы ұсынып отыр, олар бірте бірте аутентификацияның кейбір қолданыста бар әдістерін алмастыруда. Сондықтан да адамдар оларды тиісінше басқара алуы үшін олар жан-жақты зерттелуі керек. Қоғамдық орындарда, құқық қорғау ұйымдарында [62] қолданылатын тұлғаны тану жүйелерімен Siri на iPhone, Bixby Voice на Galaxy [63] байланысының мобильдік құрылымдарындағы дауыс көмекшісі биометриялық танудың үлгілі нәтижелері болып табылады.

Адамды дауысы арқылы тану оның тұлғасын дауыстың қайталанбас сипаттарының жиыны бойынша тануға мүмкіндік беретін биометриялық динамикалық әдістеріне жатады. Диктордыы тану – сөйлеу толқыны формасына негізделген сөйлеушіні автоматты түрде тани алатын технология [63]. Ол сөйлеушіден тарайтын сөйлеу параметрлерінің физиологиялық және мінез-құлық сипаттарын бейнелейді. Диктордыы дәстүрлі тану жүйесі сияқты ол да екі сатыдан тұрады, атап айтқанда: оқыту мен тестілеу. Бұл сөйлеушінің танудың негізгі сатылары. Оқыту – үлгі ретінде жазылып қойған немесе сақталынған динамикадан фонетикалық сипаттарды іріктеп алу үдерісі, сондай-ақ оларды берілгендердің базасынан күмәнді дыбыс пен фонетикалық сипаттарды салыстыру үдерісі. Мел-кепстралдық коэффициент (MFCC) және желілік жорамалдау коэффициенті (LPCC) – сөйлеу сигналын талдауда жиі қолданылатын белгілердің екі танымал жиыны болып табылады. Танудың неғұрлым кең тараған модельдері – векторлық кванттау (VQ), уақыттың динамикалық өзгеруі (DTW) және жасанды нейрон жүйесі (ANN) [64].

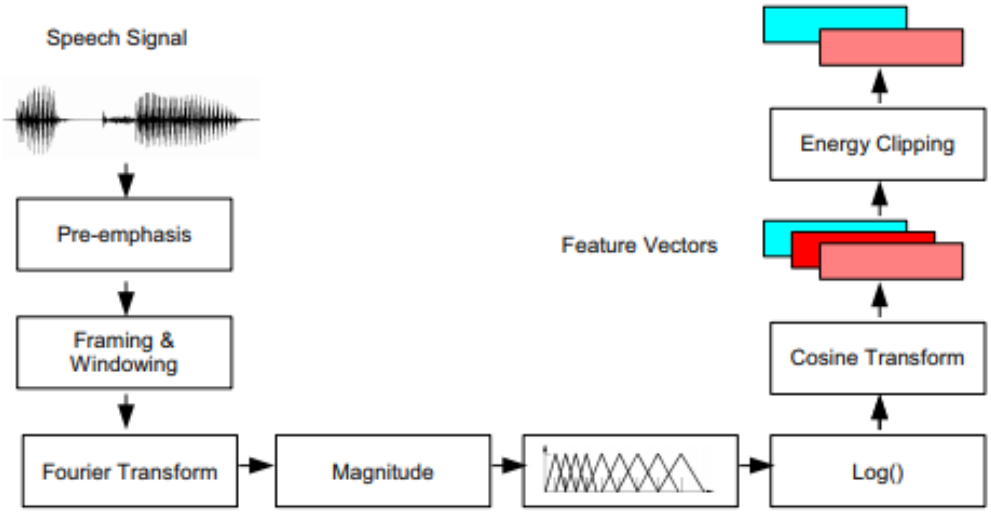
Қазақстанда қазақ тіліне арналған сөйлеу технологиялырын зерттеулер жүргізілуде. Қазақ тілі аглютинативті тілдер тобына жатады. Аглютинативтік тілдердің құрылымында әр түрлі форманттағы қосымшалар (жұрнақтар, жалғаулар) бірінен соң бірі кезегімен жалғанбалы болуы сөзді өзгертуде үстем тип болып табылады және де ол қосымшалардың әрқайсысы бір ғана мағынамен жүктелінген болып табылады. Түркі, Монғол, Корей тілдері аглютинативтік тілдерге жатады. Біздің елімізде қазақ тілінде тұлғаны тану жүйесі әлі де дами қойған жоқ, бұның өзі аталған бағыттағы зерттеу өзектілікке ие етеді.

Диссертациялық жұмыста жасанды нейронды желілерді қолдану арқылы классификациялық алгоритмдер көмегімен дауысты тану және белгілерін анықтау міндеттері қарастырылды.

Түрік, фин, қазақ тілдері үшін әлі күнге дейін тұлғаны тану мен сөйлеуді тану әрекеттері ағылшын жүйелерімен салыстырғандағы жоғары өнімділікке

жете алған жоқ. Бұның себебі тек тілді модельдеудегі қиындықтар ғана емес, сонымен қатар сөйлеу және мәтіндік оқытудың лайықты ресурстарының жоқтығы болып отыр [65,66] жұмыстарында жүйелер активті лексика мен тілдік модельдерді кластерлеу мен фокустау арқылы мүмкіндікті өлшемге дейін қысқартуға бағытталған.

Тану міндеттерінде сөйлеуді алдын ала өңдеу негізгі үдеріс болып табылады. Ғылыми зерттеу жұмысында біз MFCC-ті [67] динамиканың дауыстық функцияларын іріктеп алу үшін құрал ретінде таңдаймыз. Сөйлеу сигналын алдын ала өңдеу үдерісі 2.7 суретте көрсетілген.



Сурет 2.7 – MFCC нысандар векторларын алуға байланысты қадамдар

Сөйлеу сигналдары уақыттық аралығында өте тез және шұғыл өзгереді, ал егер уақыт аралығындағы сөйлеу сигналдарын жиілік шамасына айналдырсақ [68] онда тиісті спектрлік айқын көрінісін таба аламыз. Сөйлеу сигналдан кейін, жүйе сигналдарды фреймдерге бөледі, кадрдағы сөйлеу сигналдарының үздіксіздігін арттыру үшін терезе функциясын шақырады. Сандық сигналдарды спектрдің деректеріне түрлендіру үшін Фурье тез түрлендіруді пайдаланады және адамның есту спектралды деректерін имитациялауға арналған үшбұрышты жолақ фильтрді пайдаланады. Сөйтіп DCT блоктарында MFCC арқылы талдау қабілеті пайда болатын спектралды энергия бойынша берілгендерді санына қарай бағалау пайданылады. MFCC параметрлері 300-8000Гц, сондай-ақ 16 кепстарль талданатын жиілік диапазонында болады.

Диссертациялық жұмыс барысында эксперименттер жүргізу үшін «Интелектуалды жүйелердің компьютерлік инженериясы» зертханасы құрған сөйлеу деректер қоры қолданылды. Берілген мәліметтер жиыны әрбірінің 74-75-тен жазбасы жазылған 20 диктордың 1480 аудио жазбасынан тұрды. Әрбір аудиожазба қазақ тілінде орташа есеппен 6 секундтық ұзақтығы фраза болып табылады. Сөйлеушіні тану үшін біз төмендегідей мәліметтер жинадық: аты-жөні, жынысы, туған жері, туған жылы (2.1-кесте)

Кесте 2.1 – Дикторлар туралы мәліметтер

Label	Origin	Name	Middle name	Gender	birthplace	Year of birth
MZA	Masimkanova	Zhazira	Auezbekkyzy	female	Almaty	20.03.1982
IMT	Iskakova	Moldir	Tasbolatkyzy	female	Almaty	01.01.1994
DAZ	Duisenbaeva	Aigerim	Zhanbolatovna	female	Almaty	15.05.1995
ZEA	Zhetpisbaev	Erlan	Alibekovich	man	Almaty	23.05.1995
SSM	Samrat	Sanjar	Muhametkaliuly	man	Almaty	12.07.1996

Аудиоматриалдардың жазылу дәлдігін арттыру мақсатында шуылдан оқшауландырылған, Vocalbooth.com. фирмасының жоғары мамандықтағы дыбыс жазатын студиясы пайдаланылады.

Барлық аудиоматериалдар бірыңғай сипатқа ие:

- файлдың кеңейтілуі: .wav;
- сандық түрге түрлендіру әдісі: PCM;
- дискреттік жиілік: 44,1 кГц;
- разрядтық: 16 бит;
- аудиоканалдар саны: біреу (моно)

Бір дикторды сөйлеу мен жазу диктордың, құрал жабдықтарды, дубльдерді дайындауға кеткен уақытты қоса есептегенде, 40-50 минут уақыт алады, бұл әр дикторға шаққанда жалпы ұзындығы 7-8 минут болатын алынуға тиісті 74-75 файлға тең. Сөйлеушіні тану үшін біз төмендегідей классификациялық алгоритмдерді қарастырдық:

Extra-Trees алгортимі

Extra-Trees алгоритмі қайта құрылмаған шешімдер немесе регрессиялар ансамблін жасайды. Қадам негізіндегі ансамбльдің басқа әдістері алгоритмнің екі негізгі айырмашылығы ол кесу нүктелерін толығымен кездейсоқ түрде таңдап, тораптарды бұзады және ол қадамдарды өсіру үшін барлық оқыту іріктеуін пайдаланады. Extra-Trees алгоритмі 2.2-кестеде келтірілген.

Кесте 2.2 – Extra-Trees алгоритмі

Trees_node(M)

Кіріс сигналы: жергілікті оқыту жиынтығы M түйінге сәйкес келеді

Шығыс сигналы: $[a < a_c]$

– егер **Tree(S)** ақиқат болса, онда ешнәрсе қайтарылмайды

– Әйтпесе барлық тұрақты емес (S) кандидат атрибуттарының ішінен $\{a_1,...,a_K\}$ атрибуттарын таңдаймыз;

– K қадамдарды сызамыз $\{s_1,...,s_K\}$, мұнда s_i = **Random_split**(S, a_i), ∀i = 1,..., K;

–s_* қадамында Score(s_*, S) = $\max_{i=1,...,K}$ Score(s_i, S) деп келтіреміз.

Random_split(S, *a*)

Кірістер: S ішкі жиынтығы және a атрибуты

– a_{max}^{S} және a_{min}^{S} максималды және минималды мәнді S-ге белгілейді;
–Кездейсоқ кескін a_c нүктесін $[a_{min}^{S}, a_{max}^{S}]$ біркелкі сызамыз;
– $[a < a_c]$ қадамды қайтарамыз.

Tree(S)

Кіріс: S ішкі жиын
Шығыс: a логикалық мәні
– егер $|S| < n_{min}$ болса онда қайтару ақиқат;
– егер барлық атрибуттар S тұрақты болса, онда TRUE (ақиқат) мәнін қайтарамыз;
– егер шығыс S тұрақты болса, онда TRUE (ақиқат)мәнін қайтарамыз;
–болмаған жағдайда FALSE (жалған) мәнін қайтарамыз.

Ол екі параметрге ие: K әрбір торапқа кездейсоқ таңдалған атрибуттар саны, және n min торапты бөлу үшін ең аз үлгінің өлшемі. Ол ансамбль моделін жасау үшін бастапқы оқыту үлгісімен бірнеше рет қолданылады.

KNN алгоритмі

K-nearest-neighbors (KNN) алгоритмі сұраныс сценарийі мен деректер жиынтығындағы сценарийлер жиынтығы арасындағы қашықтықты өлшейді.

Тестілік үлгілеу объектілерінің әрқайсысын жіктеу үшін келесі әрекеттерді жүйелі түрде орындау қажет:

- оқыту үлгісінің әрбір объектісіне дейінгі қашықтықты есептеу;
- K оқыту үлгісінің объектілерін таңдап алу;
- Жіктелетін объектінің класы K-ең жақын көршілердің арасында жиі кездесетін класы.

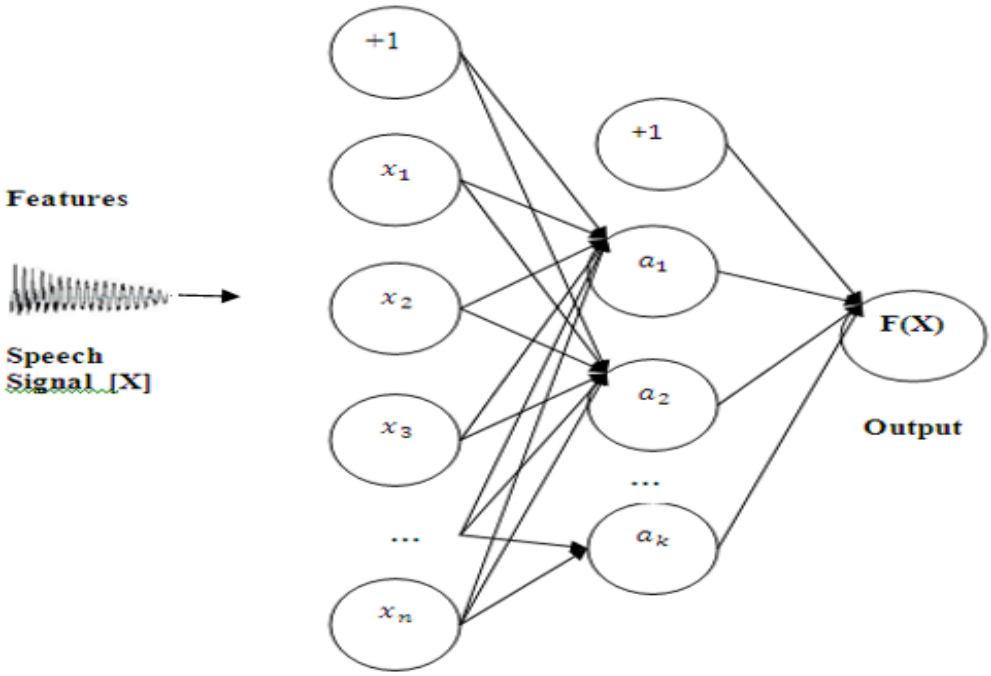
SVC алгоритмі

Сызықты бөлінетін екілік классификация мәселесін шешу үшін SVC қолдану үшін бізге қажеттілер:

- H құру үшін**,** мұндағы $H_{ij} = y_i y_j x_i \cdot x_j$
- α –ны табу;
- $\sum_{i=1}^{L} \alpha_i - \frac{1}{2}\alpha^T H \alpha$
- $\alpha_i \geq 0 \quad \forall_i$ және $\sum_{i=1}^{L} \alpha_i y_i = 0$ шектеулерді ескере отырып, барынша жоғарылату;
- QP шешімін пайдалану;
- $w = \sum_{i=1}^{L} \alpha_i y_i x_i$ есептеу;
- $\alpha_i > 0$ деген индекстерді тауып, *s* тірек векторларының санын анықтау;
- $b = \frac{1}{N_s} \sum_{s \in S} (\alpha_m y_m x_m \cdot x_s$ есептеу;
- x' әрбір жаңа нүктесі $y' = sgn(w \cdot x' + b$ есептеу жолымен жіктеледі.

MLP Classifier алгоритмі

Көп қабатты персептрон - бұл басқарылатын оқыту алгоритмі, ол $(X) = R_n$: $R_n \to R^0$ функциясын зерттейді. $X = x_1, x_2, \ldots, x_n$ белгілер жиынтығын есепке ала отырып ол кез келген жіктеу үшін сызықты емес функцияның аппроксиматорын зерттей алады (2.8-сурет).



Сурет 2.8 – MLP архитектурасы

Кіріс қабаты кіріс функцияларын білдіретін $x_1, x_2, \ldots, x_n$ тұрады. Шығыс қабаты соңғы жасырын қабаттан мәндерді алады және оларды шығару мәндеріне түрлендіреді.

Gaussian NB алгоритмі

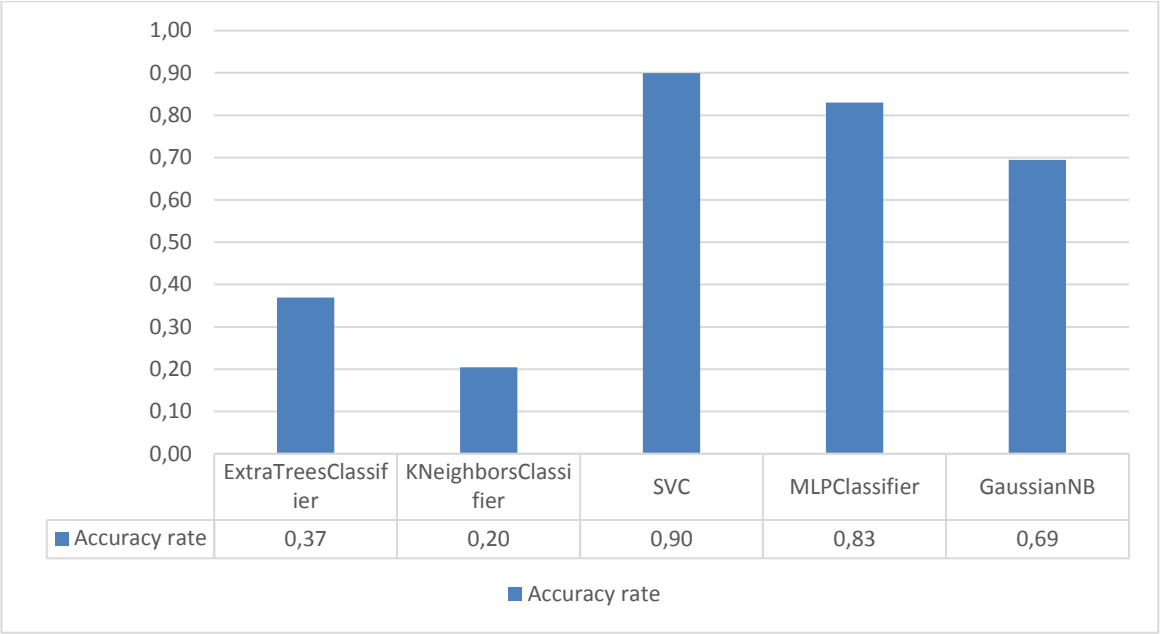
Naive Bayes-$x_1, x_2, \ldots, x_n$ қарапайым көбейтіндісіне пропорционал C_kкласына жататын $n+1$ мәліметтер нүктесінің ықтималдығын береді. n алдыңғы $p(C_k)$ класы және $p(C_a)\prod_{i=1}^{n} p(x_i|C_k)$ белгілерінің шартты ықтималдығы.

$$p(C_a)\prod_{i=1}^{n} p(x_i|C_a) > p(C_b)\prod_{i=1}^{n} p(x_i|C_b) \qquad (2.4)$$
$$p(C_a|x_1, \ldots, x_n) > p(C_b|x_1, \ldots, x_n)$$

Осылайша, $x_1, x_2, \ldots, x_n$ деректер нүктесі үшін кластың ең ықтимал тағайындалуы $k = 1, \ldots, K$ үшін $p(C_a)\prod_{i=1}^{n} p(x_i|C_k)$ есептеу және осы мәні ең үлкен болып табылатын C_k класына $x_1, x_2, \ldots, x_n$ беру жолымен табылуы мүмкін.

2.3.1. Классификациялық алгоритмдерді салыстырмалы талдау

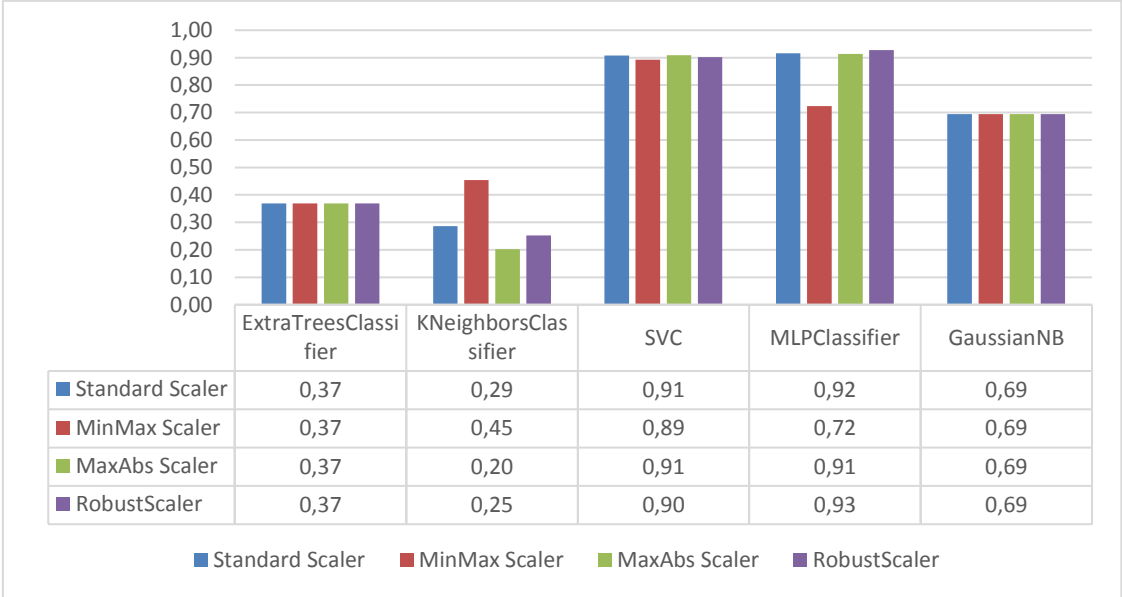
Жоғарыда қарастырылған алгоритмдерді сөйлеушіні тану есебі/міндеті үшін пайдаландық, салыстырмалы талдау жасадық [69]. Салыстырмалы талдау мен эксперименттердегі ең жақсы көрсеткіштер тірек векторлары мен көп қабатты персптрондарды қолдану арқылы алынды (2.9-сурет).



Сурет 2.9 – Берілгендер жиынындағы жіктеу дәлдігі

Дикторлардан көрініп тұрғандай, тірек векторлары мен көпқабатты персептрон әдісі тиісінше 0,90 және 0,83 өте жақсы нәтижелер көрсетті.

Нәтижелерді жақсарту үшін әртүрлі әдістер арқылы масштабтау жасадық, сонда нәтижелер едуір өзгерді (2.10-сурет).

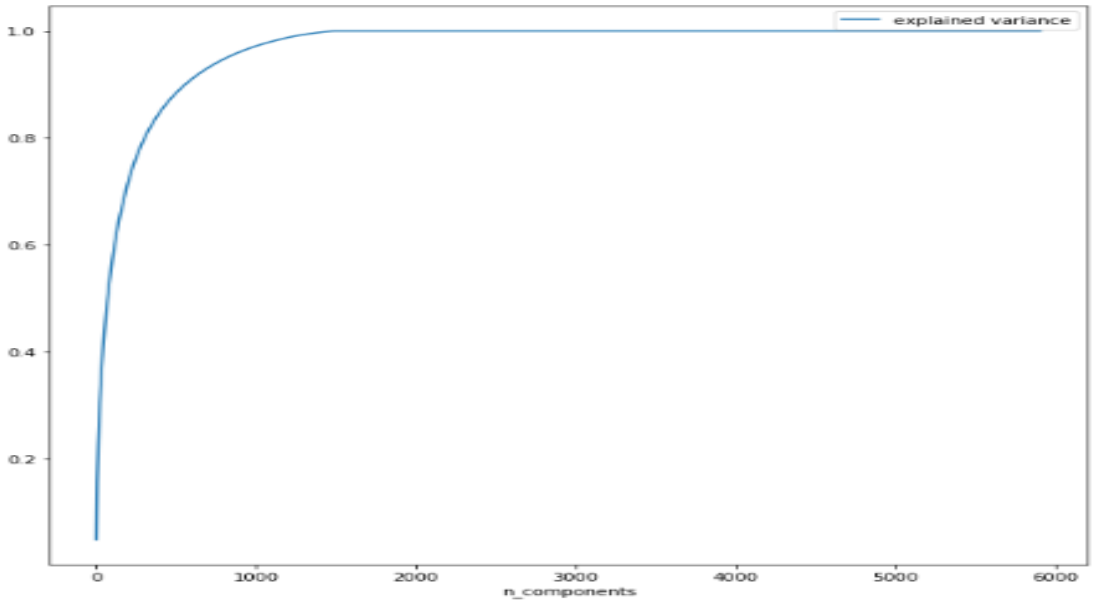


Сурет 2.10 – Әртүрлі әдістермен берілгендерді масштабтаудағы жіктеу дәлдігі

Нәтижесінде әрбір аудио файл үшін 5904 белгілер алынды. Әрбір аудиофайл дауысы жазылуда тіркелген диктордың аты-жөнімен таңбаланған. Берілгендердің алынған жиынының өлшемі – 1480x5904.

Берілгендерді визуалдау мақсатында 5904 белгісі бар векторлық кеңістіктің және екі-үш өлшемдік кеңістік өлшемдерін кішірейту үшін басты компоненттер

әдісі пайдаланылады. Басты компонент әдісімен кішірейтілген өлшемде дисперцияны сақтау 2.11- суретте көрсетілген.



Сурет 2.11 – Басты компонент әдісімен өлшемдіктің төмендеуі кезінде дисперсияны сақтау

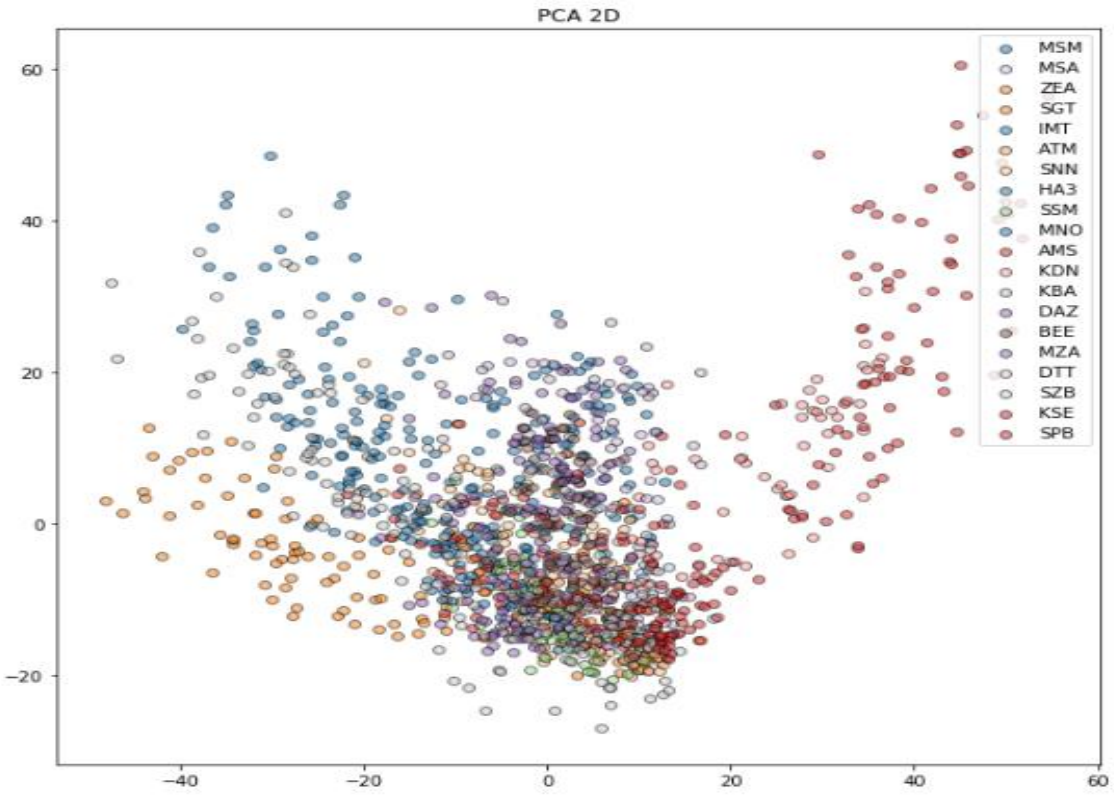
Берілген графикте көрініп тұрғандай, берілгендердің өлшемділігін 1479 белгілерге дейін кішірейту жағдайында дисперсия 100%-ға сақталады. Алайда классификациялық модельдерімен және берілгендердің стандарттаушыларымен жүргізілген тәжірибе мұндай өлшемдік төмендетілу жіктеу дәлдігіне сындарлы ықпал ететінін көрсетті.

Ендігі жерде неғұрлым үлкен дәлдікті Robust scaler әдісімен масштабтауда – 0,93 көпқабатты персептрон көрсете бастады, ал Standard scaler және MaxAb Scaler әдістері арқылы масштабтауда өзінің дәлдік бойынша нәтижелерін 0,90-нан 0,9-ге дейін жақсартқанымен, тірек векторлар әдісі екінші қатарға ығысты. Егер біз жіктеуде басты компонент әдісі көмегімен сөйлеу белгілерінің өлшемдігін 1479 дейін төмендетсек, онда жіктеу дәлдігі 2.3-кестеде көрсетілгендей өзгереді.

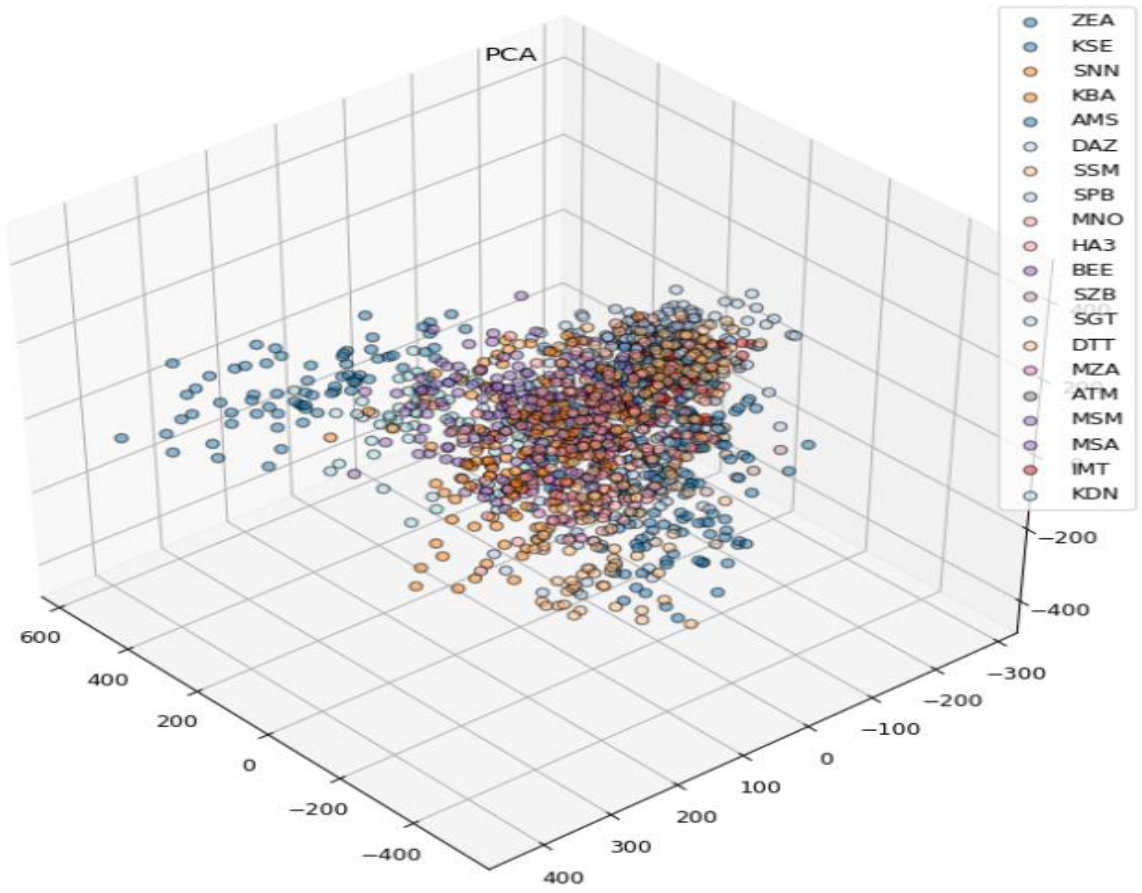
Кесте 2.3 – Берілгендер арқылы өлшемі азаятын деректер бойынша жіктеу дәлдігі.

Алгоритм атау	Standard Scaler	MinMax Scaler	MaxAbs Scaler	Robust Scaler
ExtraTreesClassifier	0.128125	0.128125	0.128125	0.128125
KNeighborsClassifier	0.043571	0.052143	0.050089	0.060982
SVC	0.051875	0.134732	0.097500	0.157679
MLPClassifier	0.002589	0.051875	0.082054	0.098393
GaussianNB	0.324286	0.324286	0.324286	0.324286

Сондай-ақ сөйлеушіні тану міндетін атқару үшін жіктеудің алгоритмдерінің ықпал ету дәрежесін анықтау SVC және MLP Classifier алгоритмдерін салыстырмалы талдаудың мақсаты болды [70]. Сөйлеу берілгендерінің оқыту жинағында жасалған эксперименттер бұл алгортимдердің келешегі туралы мәселе көтеруге болатынын көрсетті. Алынған нәтижелер 2.12-2.13 суреттерде көрсетілген.



Сурет 2.12 – Сөйлеу берілгендерінің екі өлшемдік көрінісі



Сурет 2.13 – Сөйлеу белгілерінің үш өлшемді көрінісі.

Әртүрлі әдістер арқылы берілгендерді масштабтаудағы жіктеу нәтижелері алдын ала жүргізілген эксперименттер барысында алынған нәтижелерден едәуір айырмашылықтары бар.

Берілген ғылыми жұмыста бірақатар классификациялық алгоритмдері және сөйлеуді алдын ала өңдеу мәселелері қарастырылды. Эксперименттердің нәтижелерін талдау нәтижесінде Robust scaler әдісі арқылы масштабтауда дәлдігі 0,93 көпқабатты персептрон ұсынылды, сөйтіп көпқабатты персептрон арқылы біз сөйлеу сигналын жіктеуге болатындығы анықталады. Одан соң алынған мәліметтерден сөйлеушіні тану процесін жүзеге асырдық .

3 СӨЙЛЕУЛЕРДІ ТАНУ ЕСЕПТЕРІНДЕ МАШИНАЛЫҚ ОҚЫТУДЫ ҚОЛДАНЫП БЕЛГІЛЕРДІ АНЫҚТАУ ЖӘНЕ ӨҢДЕУ МОДЕЛДЕРІ МЕН АЛГОРИТМДЕРІН ҚҰРУ

Машиналық оқыту - бұл бағдарлама нәтижеге қол жеткізу үшін амал-әрекеттерді мұқият орындауда оқытылатын яғни бейімделетін үдеріс. Ол өзінің оқуына сүйене отырып, кейбір міндеттерді дербес шешеді.

Машиналық оқыту әртүрлі алгоритмдерден тұратын жиын арқылы жүзеге асады, алайда адам сөйлеуін танудың міндеттерін шешуде жасанды нейрондық желілерді қолдану ең бір болашағы бар бағыттардың бірі болып саналады. Машиналық оқыту жасанды интеллектке негізделіп туындаған. Жасанды интеллектің алғашқы кезеңдерінде оны интелектуалдық өріс ретінде танып, зерттеушілердің машиналардың берілгендері бойынша оқытылуға қызығушылығы пайда болады. Жасанды нейрондық желілерді тереңдетіп оқыту әдісімен өздерінің берілгендерді жинақтау қабілеттілігі болғандықтан әртүрлі кірістік сигналдарды анықтау міндеттерін шешуде өздерін таныта алады.

Сондықтан да олар бұл мәселеге әр түрлі символдық тәсілдер арқылы және нейрон желілері деп аталған тәсілдер арқылы шешуге әрекет етуде. Нейрондық желілер-бұл әртүрлі спектрдегі көптеген міндеттерді шешуге мүмкіндік беретін бірегей құрал. Оның ішінде оларды тану есептерін шешу үшін қолданады. Бұл тарауда сөйлеу белгілерін анықтауда нейрондық желілердің жалпы құрылысы мен жекелеген ерекше архитектуралары қарастырылады. Гендерлік ерекшелігін анықтау және сөйлеушіні анықтауды қарастырамыз. Қолданылған алгоритмдерді және модельді салыстыра отырып, қайсысы жақсы нәтиже беретінін көрсетеміз.

3.1 Гендерлік ерекшелігін және сөйлеушіні анықтау алгоритмі мен моделі

Автоматты түрде гендерлік ерекшелігін анықтау дегеніміз сөйлеуді, талдауды, синтезді және гендерлік анықтауды кодтау үшін сөйлесу сигналдарын пайдалана отырып дайындалған жыныс өкілін жіктеу жүйесі. Әдетте гендерлік ерекшелігін тану жүйесін топтастыру жалпы тұтас екі деңгейде жүзеге асуы мүмкін, атап айтқанда интерфейстік жүйеде және ішкі жүйеде.

Соңғы жылдары табиғи тілді тануда едәуір ілгерілеушілік байқалады: әуелде ауызекі сөйлеу тілін тану содан соң оның жиынын тану. Мәтінді тану адамдар арасындағы қарым-қатынас пен дербес компьютер үшін өзектілікті мәселеге айналды. Мұнда сөйлеу ақпаратқа әмбебап қол жеткізудің кілті ретінде қарастырылады, өйткені сөйлеу режимі өзара әрекеттестігінің табиғи тәсілі болып табылады және ол аса сауаттылықты қажет ете бермейді.

Сөйлеу технологиясы біріншіден, сөйлеу және дыбыстық кодтау, екіншіден сөйлеудегі мәтіндер синтезі, үшіншіден сөйлеуді тану, төртіншіден сөйлеуді жетілдіру, бесіншіден ауызекі сөйлеуді тану болып кең түрде жіктелуі мүмкін.

Қазір бүкіл әлемде жыныстық белгілері бойынша танудың қауіпсіздік мәселесі сөйлеуді зерттеушілер тарапынан үлкен алаң туғызуда. Гендерлік тану ер адамның немесе әйел адамның дауысына қарай анықтауда қолданылады.

Қазіргі гендерлік тануды гендерлік анықтау және гендерлік тексеру деп жіктеуге болады.

Қазіргі мультимедиялық ақпараттық-іздеу жүйелерінде гендерлік жіктеу сөйлеуді тану, оның динамикасы адам бағдарламайтын компьютермен интелектуалдық өзара әрекеттестік, биомериялық әлеуметтік роботтар, аудио немесе видео контенттердің индексациясы т.б. сол сияқты бірнеше әлеуетті қосымшаларда қолданылады [70]. Гендерлік ерекшеліктерін автоматты тану сондай-ақ денсаулық сақтау жүйесінің кейбір мобильдік жағдайларында, мысалы көмей қуысы қатпарларындағы өскін сияқты патологиялар жағдайында пайдалы. Сонымен бірге цифрланудың жылдам дамуынан гендерлік анықтауда жаңа маңызды проблемалар туындады [71].

Сөйлеу ерекшеліктерін қалыпқа келтіру үшін пайдаланылатын гендер туралы ақпарат сөйлеуді тану кезінде сөз қателерін төмендетеді. Жалпы алғанда, сөйлеушіні, гендерлік ерекшелігін анықтау табиғи және жекеленген диалог жүйелерін арттыру үшін маңызды. Келесіде нейродық желі негізінде гендерлік ерекшелігін және сөйлеушіні анықтау алгоритмі мен моделі құрылды.

Гендерлік ерекшелігін және сөйлеушіні анықтау процесі.

$$h_i^1 = \sigma^{(1)}(\sum_j w_{ij}^{(1)} x_j^{(1)} + b_i^{(1)}) \quad (3.1)$$

$$h_i^2 = \sigma^{(2)}(\sum_j w_{ij}^{(2)} h_j^{(1)} + b_i^{(2)}) \quad (3.2)$$

$$\hat{y}_i = \sigma^{(3)}(\sum_j w_{ij}^{(3)} h_j^{(2)} + b_i^{(3)}) \quad (3.3)$$

$\sigma(z)$- сызықты емес активация функциясы;

h_i– жасырын қабат;

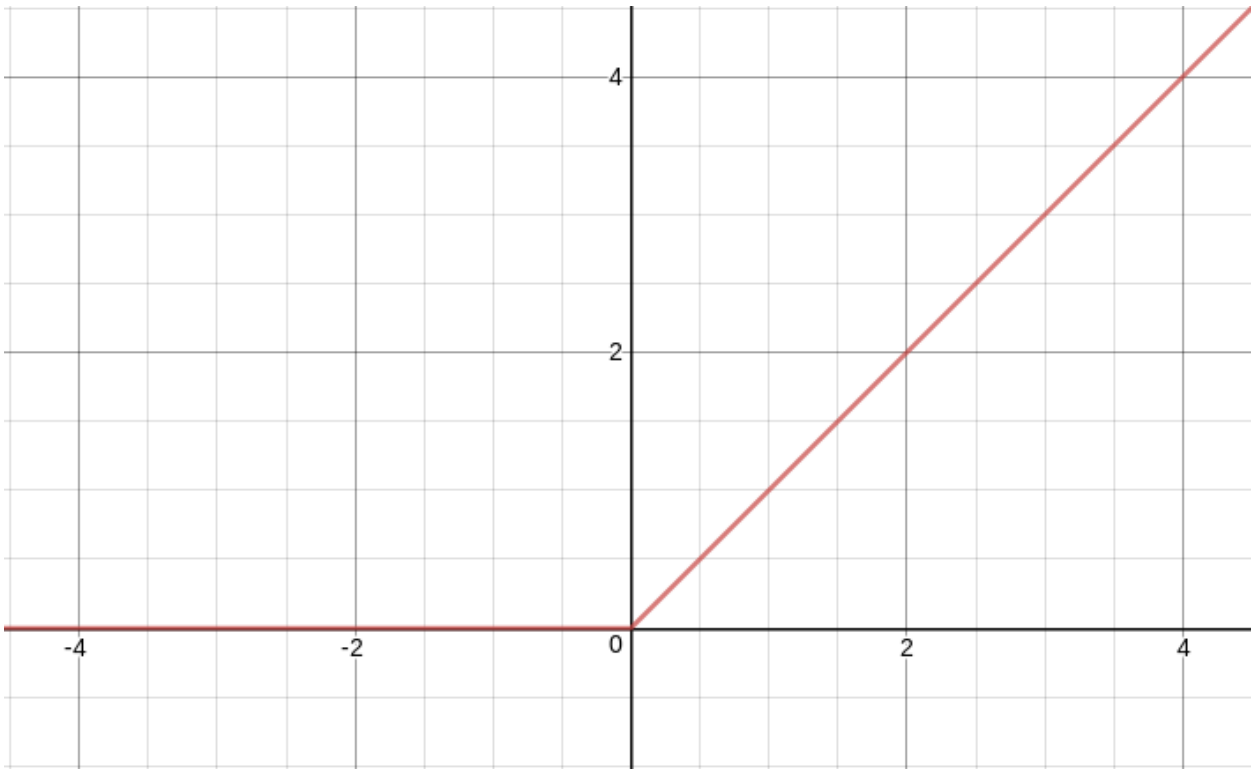
x_j – кіріс сигналы;

$\hat{y}$ - ($\hat{y}$ -циркумфлекс) модельдің шығысы;

$\theta = \left\{w_{ij}^{(1)}, w_{ij}^{(2)}, w_{ij}^{(3)}, b_i^{(1)}, b_i^{(2)}, b_i^{(3)}\right\}$ – модель параметрлері;

$\sigma^{(1)}(x)$ және $\sigma^{(2)}(x)$ активтендірудің тікелей функциялары:

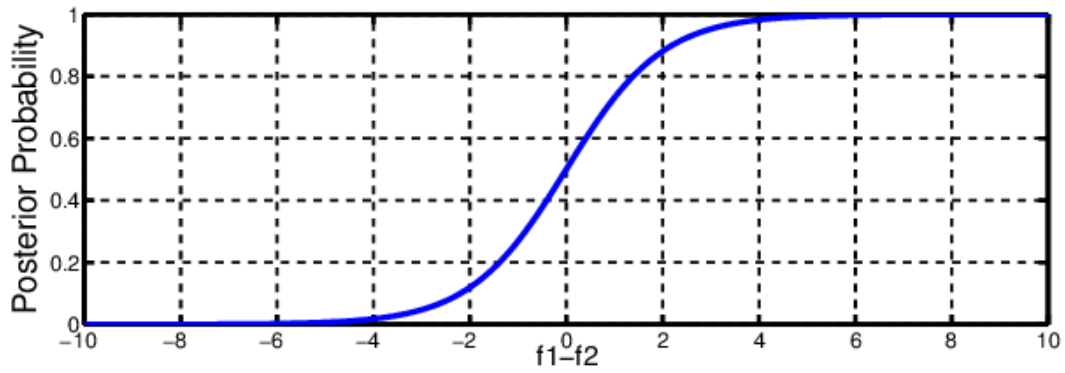
$$\sigma^{(1)}(x) = \sigma^{(2)} = x^+ = \max(0, x) \quad (3.4)$$



Сурет 3.1 – Relu сызықты емес функцияның графигі

$\sigma^{(3)}(z)$ - softmax функциясы, $i = 1, \dots, K$ үшін $z = (z_1, \dots, z_k) \in R$;

$$\sigma^{(3)}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \qquad (3.5)$$



Сурет 3.2 – softmax функциясының графигі

Softmax функциясының басты мақсаты моделдің шығысын ықтималдық жүйесіне қарай сәйкестендіру. Яғни шығысты 0 мен 1 сандар арасында алмастырады. (3.1)– ші теңдеумен (3.5) – ші теңдеу гендерлік ерекшелігін және сөйлеушіні анықтау процесі. Яғни сигналдың кірісінен бастап модель шығысына дейінгі есептеу. (3.1) теңдеудегі сигналдарды модельге оқытуға береміз. Ол өзінің бірінші жасырын қабатындағы параметрлермен әрекеттесіп есептеулер жасайды. h_i – бірінші қабаттың жасырын қабатының шығысы.

Модельдің оқыту процесі. (6) функцияға тренинг жасаудағы мақсатымыз шығын функциясының минимум жуықтау.

$$E_i = \frac{1}{2}\sum_{j=1}^{K}\left(\hat{y}_{ij} - y_{ij}\right)^2 \tag{3.6}$$

$$\frac{\partial E_i}{\partial w_{ij}^3} = \frac{\partial E_i}{\partial \hat{y}_{ij}} \cdot \frac{\partial \hat{y}_{ij}}{\partial \sigma^{(3)}(z)} \cdot \frac{\partial z^{(3)}}{\partial w_{ij}^3} \tag{3.7}$$

$$\frac{\partial E_i}{\partial w_{ij}^2} = \frac{\partial E_i}{\partial \hat{y}_{ij}} \cdot \frac{\partial \hat{y}_{ij}}{\partial \sigma^3(z)} \cdot \frac{\partial z^3}{\partial h_j^{(2)}} \cdot \frac{\partial h_j^{(2)}}{\sigma^{(2)}(z)} \cdot \frac{\partial z^{(2)}}{\partial w_{ij}^{(2)}} \tag{3.8}$$

$$\frac{\partial E_i}{\partial w_{ij}^1} = \frac{\partial E_i}{\partial \hat{y}_{ij}} \cdot \frac{\partial \hat{y}_{ij}}{\partial \sigma^3(z)} \cdot \frac{\partial z^{(3)}}{\partial h_j^{(2)}} \cdot \frac{\partial h_j^{(2)}}{\sigma^{(2)}(z)} \cdot \frac{\partial z^{(2)}}{\partial h_j^{(1)}} \cdot \frac{\partial h_j^{(1)}}{\partial \sigma^{(1)}(z)} \cdot \frac{\partial z^{(1)}}{\partial w_{ij}^{(1)}} \tag{3.9}$$

$$\Delta w_{ij}^{(3)} = -\eta \frac{\partial E_i}{\partial w_{ij}^{(3)}} \tag{3.10}$$

$$\Delta w_{ij}^{(2)} = -\eta \frac{\partial E_i}{\partial w_{ij}^{(2)}} \tag{3.11}$$

$$\Delta w_{ij}^{(1)} = -\eta \frac{\partial E_i}{\partial w_{ij}^{(1)}} \tag{3.12}$$

E_i - шығындар функциясы;

$\hat{y}$ - ($\hat{y}$ -циркумфлекс) модельдің шығысы;

y –нақты белгісі;

(3.6) – (3.12) функциялары моделдің тренинг жасау процесі. (3.6) теңдеу моделдің үйрену процесі осы теңдеулер арқылы жүзеге асырылып есептелінеді. (3.6) – шы теңдеу шығындар функциясы. Бұл жердегі басты идея біз моделдің әрбір оқыту қадамында модельдің шығысымен y –нақты белгісін салыстыру арқылы модельге тренинг жасаймыз. Шығын функциясының минумум нүктесін табу үшін, біз (3.7) –ші (3.9)-шы функцияларды қолданып, модельдің әрбір параметрлеріне байланысты жанама туындыны аламыз. Тренинг процесі моделдің шығысынан кірісіне қарай, кері бағытта жүргізіледі. (3.10) – шы және (3.12) – ші теңдеулер ескі параметрлерге оқытылған жаңа параметрлерді қосып беру процесі. η – оқыту жылдамдығы. η жоғары болған сайын шығын функцияның минимум нүктесіне қарай қадамы алшақтай түседі. η –ны кішірейткен сайын оқыту қадамы кішірейді. η нақты тренинг жасаған кезде 0,01 деп алынды.

Төменде гендерлік ерекшелігі мен сөйлеушіні анықтау алгроитмі көрсетілген
Input: $(X, Y) = (x_1, y_1), \ldots, (x_n, y_n)$ аудио сигналдар мен нақты таңбалар тізбегі
Output: $\hat{Y} = \hat{y}_1, \ldots, \hat{y}_n$ болжамды белгі.
Param: $\theta = \left\{w_{ij}^{(1)}, w_{ij}^{(2)}, w_{ij}^{(3)}, b_i^{(1)}, b_i^{(2)}, b_i^{(3)}\right\}$

1. **for** epoch ← 1 **to** total Epoch **do**
2. **for** I ← 1 to n do
3. $a_i \leftarrow mfcc(x_i)$
$\hat{y}_i \leftarrow a_i$ (1), (2) және (3) теңдеулерді қолданып a_i сигналын кіріс ретінде қарастырып модельдің шығысын есептеу; $\hat{y}$ - модельдің шығысы
4. **if** $\hat{y}_i \neq y_i$ then
$\Delta\theta \leftarrow y_i$-нақты аудио белгісімен $\hat{y}_i$ модель шығысы салысытырылады. Егер тең емес болса (7), (8), (9) функциялар теңдеулерін пайдаланып әрбір параметрлерге қатысты градиенттерді есептейді;
5. (10), (11) және (12) теңдеуі арқылы $\theta \leftarrow \Delta\theta$ параметрлерді жаңарту;
6. **end**
7. **end**
8. if *computeAccuracy* $(\hat{Y}, Y)$ > *precision then*
9. *# бір жолға аудио тізбегіне оқыту аяқталады.*
10. *# сол кезде модельдің жалпы дәлдігі есептеледі.*
11. *# сол модельдің дәлдігі күтілетін дәлдікпен салыстыру.*
12. *# егер модель дәлдігін күтілетін дәлдіктен үлкен болса онда тренинг процесі тоқтатылады.*
13. **end**
14. end.

Гендерлік ерекшелікті анықтау есебі қазақ тілінде айтылатын дыбыстық сигналдардағы автордың жынысын анықтауға бағытталады. Осыған ұқсас түрде қазақ тілінде айтылған аудио жазбаларды талдау жолымен сөйлеушіні анықтау дегеніміз автордың жеке тұлға ретіндегі ақпараттарын анықтау болып табылады.

Мұнда $X = x_1, x_2, \ldots x_n$ тізбегі бірқатар аудио сигналдарды білдіреді. Сонда $G = g_1, g_2, \ldots g_n$ дегеніміз – X дыбыстық сигналдарға сәйкес келетін гендерлік категорияларға арналған 0/1 бинарлық вектор. Бұл жерде біз 1-ді әйел дауысын белгілеу үшін, 0-ді ер дауысын белгілеу үшін пайдаланамыз. $S = s_1, s_2, \ldots s_n$ дегеніміз сөйлеушіні анықтауды білдіреді. Сөйлеушіні анықтау үшін өзгеше нөмір қолданамыз. Гендерлік ерекшеліктерін анықтау және сөйлеушіні анықтау элементтер төмендегідей түрде анықталады.

1) (X,G)=$(x_1, g_1),\ldots,(x_n, g_n)$ –гендерлік ерекшелігін анықтауға арналған.

2) (X,S)=$(x_1, s_1),\ldots,(x_n, s_n)$ –сөйлеушіні анықтауға арналған.

Осы екі міндеттер үшін X-ті тиісті сигналдарды енгізу және шығару ретінде пайдаланамыз, содан кейін MLP және CNN нейрондық желілерді гендерлік ерекшелікті анықтау және сөйлеушіні анықтау үшін модельдерді оқыту үшін пайдаланамыз.

Нейрондық желілерді функцияның - параметрлермен жіктеушісі ретінде қарастыруға болады, ал бірнеше қабатты нейрондық желіні келесі түрде анықталатын функциялар құрамы ретінде қарастыруға болады:

$$f_\theta(\cdot) = f_\theta^l(f_\theta^{(l-1)}(\ldots f_\theta^1)) \quad (3.13)$$

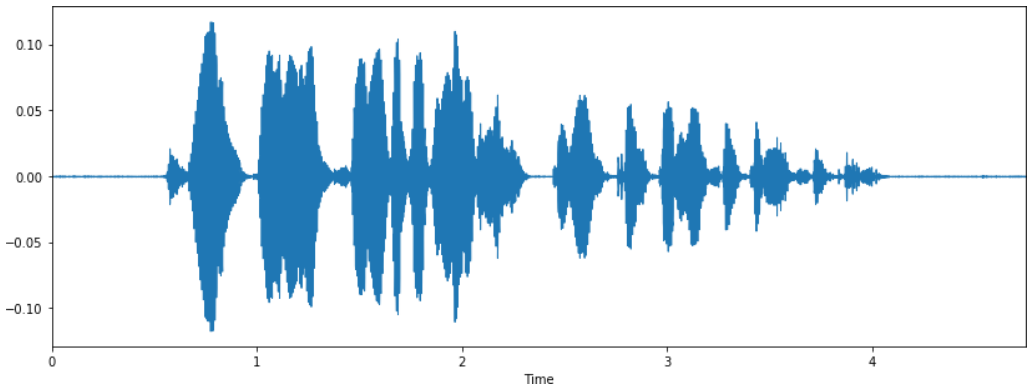
Мұнда θ-та нейрондық желінің параметрлерін білдіреді, ал l – қабаттар саны. Төменде гендерлік ерекшелігін және акустикалық жүйелерді анықтау міндеттеріне арналған екі нейрондық желілік архитектурасын сипаттаймыз:

1) feed-forward нейрондық желі, MLP-ға қатысты персептрон;
2) convolutional нейрондық желі, CNN-ді білдіреді

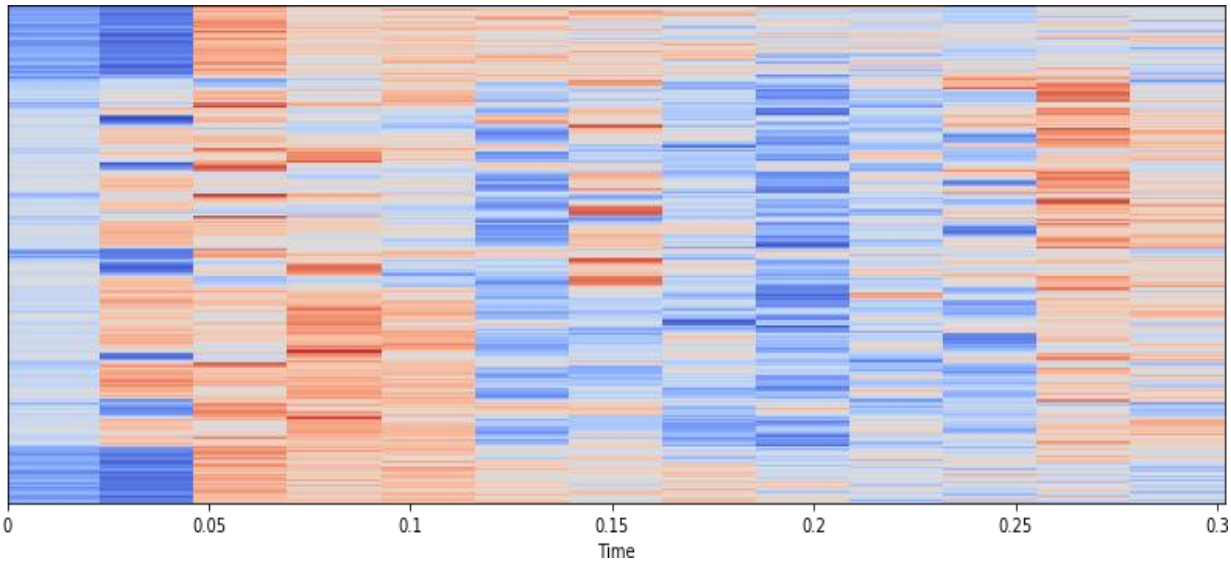
3.1.1 Сөйлеу сигналдарын алдын ала өңдеуде MFCC- ді қолдану

Сөйлеуді өңдеудің басқа көптеген есептері сияқты (сөйлеуді тану т.б.) функцияларды бөліп алу бірінші қадам болып табылады. Олар аудиосигналдардағы тілдік контентті анықтау үшін және қабаттасқан шуыл ақпаратты жою үшін қолданылуы мүмкін. Мел кепстралды коэфициенттері (MFCCx) [72] сөйлеуді өңдеу бойынша көптеген қосымша тіркемелерде кең қолданылатын ең соңғы заманауи сипаттама болып табылады. MFCC – ті сипаттаудан бұрын 3.1 суретте көрсетілген бастапқы дыбыстық сигналды көрсетейік. Бастапқы сигнал жүздеген және миллиондаған сандардан тұрады, оны лингивистикалық контент пен шуыл бар өте ұзын вектор ретінде қарастыруға болады. Мысалы, 3.3 суретте бастапқы дыбыстық тежелістердің нақты мәндері көрсетілген.

Бұл берілген ғылыми жұмыста MFCC – ті гендерлік ерекшеліктерін (дауыс зорайтқыштар) анықтау үшін қолданамыз. Іс жүзінде біз аудиосигналды талдау үшін Питонның LibROSA пакетін пайдаланамыз. Оның librosa.features.mfcc функциясы MFCC – ті бөліп алу үшін қолданылады. Бөліп алынған жеке ерекшеліктер төмендегі суретте берілген (3.4 - сурет.)



Сурет 3.3 – Бастапқы дыбыс сигналы

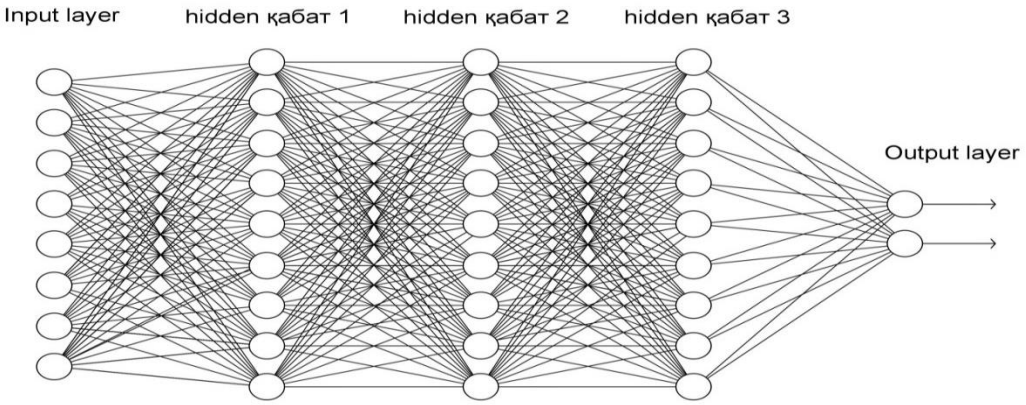


Сурет 3.4 – MFCC аудиосигнал функциясы

3.2 Генделік ерекшелігі мен сөйлеушінің дыбыс ерекшеліктерін тануға арналған MLP және CNN нейрондық желілері

Бұл модельді жақсырақ сипаттау үшін қарапайым нейрон желісінен бастаймыз. Бұрынан белгілі болғандай, бір қабатты персептрон [73,74] жасырын емес NN блоктарын білдірді, ол тек кіріс қабаты шығыс қабатынан тұрады.

Белгілерді желілік емес бөліп алудың болмауы, шығысты есептеп шығару кіріске сәйкес келетін өлшемдердің көбейтінді жиыннан тікелей алып тастау жолымен орындалады. Мұнда біз нейрондық желі архитектурасы MLP-ны [75,76,77] пайдаланамыз, ал оның өзі көптеген қабылдаулардан тұратын NNs болып табылады, сөйтіп MLP желілік емес функцияларды зерттей және бөліп ала алады. Жалпы айтқанда MLP кірістік қабаттан жасырын қабаттардың кейбір сандарынан және шығыстық қабаттардан тұрады. 3.5 суретте MLP – ның жалпы архитектурасы көрсетілген. Онда NN – нің кірістік қабаттан, бірнеше жасырын қабаттардан және шығыстық қабаттардан тұратыны көрінеді.



Сурет 3.5 – Гендерлік ерекшелігі мен сөйлеушіні анықтауға арналған нейрондық желі архитектурасы

$f(X, G) = (x_1, g_1), (x_2, g_2), \dots, (x_n, g_n)$ оқу үлгілері үшін модельге x_1-ді жеке дара немесе кіріс параметрлері ретінде топтастырған түрінде енгіземіз. Вектор мен матрицалар үшін ерекшеленіп белгіленген символдар пайдаланылады. Әдетте, векторлар баған векторлары ретінде қарастырылады. Біз төмендегідей мәндерді пайдаланамыз:

- W^l салмағы үшін $l-$ ші қабат;
- $l-$ ші қабатқа b^lығысуы;
- h^l көбейтінділер жиыны, оған l қабатқа арналған орын ауыстырылуы қосылады;
- o^lде $l-$ ші қабатқа шығарылады;
- m – қабаттар саны.

Содан кейін MLP-ның o шығысын есептеу келесідей болады:

1. W^{l}, b^{l}-ның барлық парамтрлерін қысқаша ықшамдап, $l = 0$ кіріс қабатты орнатамыз, бұл олардың x_{i} векторларындағы байланыстырылған кірістері үшін жасалады.
2. Жасырын қабаттың шығыс деректерін $l = 1$ ден $l =$ m -1-ге дейін есептеу:
   а) $h^{\mathrm{l}} = W^{\mathrm{l}} * o^{\mathrm{l}-1} + b^{\mathrm{l}}$;
   ә) $o^{\mathrm{l}} = \sigma(h^{\mathrm{l}})$ есептеу, мұнда σ белсенді функциясын білдіреді;
3. l=m шығыс қабаты үшін шығыс o^m мәнін есептеу.
   а) $h^{\mathrm{l}} = W^{\mathrm{l}} * o^{\mathrm{l}-1} + b^{\mathrm{l}}$ есептеу;
   ә) $o^m = \sigma(h^m)$ есептеу.

Біз W^l және b^l мәндерін жаңартып және жіктеу есептерінде кең қолданылатын нысан функциясының қиылысқан энтропиясының шығынын азайту жолдары арқылы MLPs мәнін оқытамыз.

$$E(X) = -\sum_i^n \sum_j^C (y_{ij} \log(o_{ij})) \tag{3.14}$$

мұнда o_{ij} – С ішіндегі Е әрбір j класы үшін болжамдаудың нәтижесі, ал y_{ij} – негізгі ақиқат. Бұл параметрлерді оқыту барлық берілген W^l және b^l мәндерін қатысына қарай $E(X)$ минимумына біріктіріледі. Бұл параметрлердің градиенттері кері тарату үдерісі барысында есептеп шығарылады, содан соң градиенттік төмендету көмегімен оларды η -оқыту жылдамдығына бейіндейді.

$$\Delta W^l = -\eta \frac{\partial E(X)}{\partial W^l} \tag{3.15}$$

$$\Delta b^l = -\eta \frac{\partial E(X)}{\partial b^l} \tag{3.16}$$

CNN архитектурасы

Жиналмалы нейрондық желілер (CNN) [78, 79, 80] екі өлшемді желілік топологиясы бар берілгендерді өңдеуге арналған нейрондық желінің мамандандырылған түрі болып табылады. CNN іс жүзінде практикада қолдануға өте ыңғайлы болды. Нысандарды бөліп алу үшін байланыстырылған қабаттарды толық қолданатын MLPs мәннен айыру үшін CNN екі маңызды идеяны пайдаланады. Олар модельді жақсартуға септігін тигізеді. Өзара әрекеттестігі аз және параметрлерді бірге пайдалану. Біріншісі - кіріс мәліметтеріне қарағанда, аз ядросы бар объектілерді шығару процесі. Мысалы, өңдеу негізінде аудио кірістік сигналдардың мыңдаған және миллиондаған сандары болуы мүмкін, мұндай аса ұзын векторды NN – ға енгізудің орнына CNN ықшам ақпараттарды өзіне сіңіріп алу жолымен шағын және маңызды нысандарды тауып көрсете алады. Параметрлерді бірлесіп пайдалану 2-D кірісі бойынша сырғымалы кіші ядроға арналған сол бастапқы параметрлерді пайдалану амалына жатады.

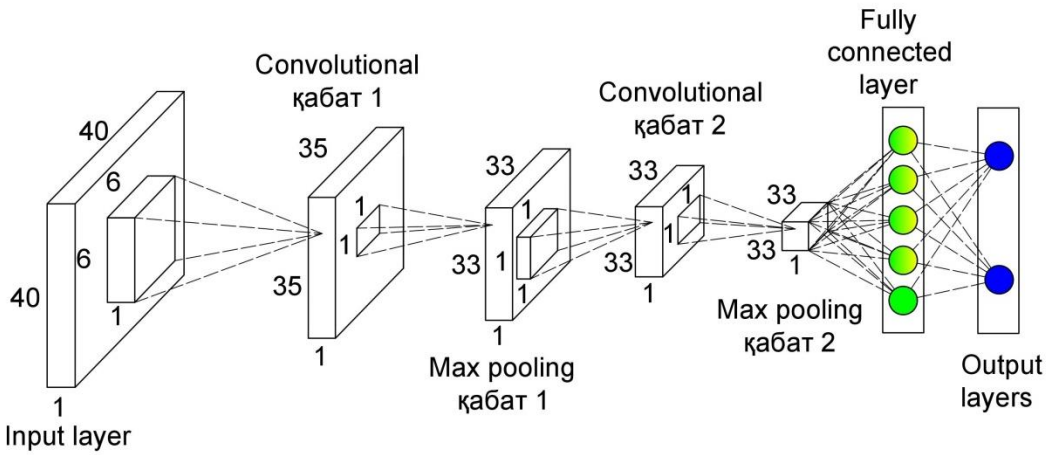
Типтік CNN үш кезеңнен тұрады:

1. Желілік активациялардың жиынын жасау үшін ықшамдалған қабаттарды пайдалану;

2. Әрбір сызықтық активтендіру бейсызық активтендіру функциясы арқылы өтеді;

3. Қабаттың шығыс деректерін одан әрі өзгерту үшін біріктіру функциясын пайдаланамыз.

CNN жалпы архитектурасы 3.6 суретте көрсетілген. Бұл архитектура кіріс қабатынан, екі орама қабаттарынан, максималды біріктірудің екі қабатынан және толығымен байланысқан қабаттан тұрады. Гендерліктің ерекшелігін анықтау және сөйлеушіні анықтау үшін, біз осындай архитектурамен CNN моделін жаттықтыратын боламыз.



Сурет 3.6 – Гендерлік ерекшелігі мен сөйлеушіні анықтауға арналған CNN архитектурасы

3.3 Гендерлік және сөйлеушінің дыбыс ерекшеліктерін тануға арналған нейрондық желілермен эксперимент жүргізу

Біз гендерлік ерекшелегін және сөйлеушіні анықтау есептері үшін MLP және CNN моделдерін бағалауға арналған эксперименттер топтамасын жүргіздік. Эксперименттер MLP және CNN моделдерін оқыту үдерісін бағалау үшін әзірленеді. Мұнда CNN моделді қалыптастырудағы әр түрлі конфигурациясы жағдайындағы бағалаудың соңғы нәтижелерімен бірге көрініс тапқан. F1 – бағалау мен дәлдік метрикалары модельді бағалау үшін хабарланады.

3.3.1 Гендерлікті және сөйлеушіні анықтаудағы деректер жиынын нейрондық желілерді оқыту

Гендерлік ерекшелігін және сөйлеушіні анықтау үшін 3.1-ші және 3.2-ші кестелерде берілгендер жиыны бойынша статистикалық мәліметтер берілген гендерлік ерекшелігін анықтауда барлығы 1125 оқу-тестілік жиын және 300 аудио жазбалар пайдаланылды. Ерлер мен әйелдер үшін аудио саны 600-ді және жаттығушылық жиынында 525-ті құрайды. Қалған 150 аудиолар – ерлер мен әйелдер үшін тестілік жиын болып табылады.

Кесте 3.1 – Гендерлік ерекшелігін анықтау үшін берілгендер жиындары

Деректер жинағы (Data sets)	#Ер (Male)	#Әйел (Female)	#Барлығы (Total)
Оқыту(Train)	600	525	1125
Тестілеу (Test)	150	150	300

3.1-кестеде сөйлеушіні анықтауға арналған оқыту және тест жинақтары келтірілген. Барлығы 19 сөйлеушінің бар екені және олардың әрбірі оқыту үшін 60 аудиоға ие екені көрінеді. Бұл шағын ғана оқыту жиынын, яғни тереңдетіп оқытудың жүктелген модельдерінің қайсыбір берілгендердің бұл жиынын MLP және CNN модельдерін оқытуға қолданамыз және бағалау жасаймыз.

Эксперимент жүргізгенде кіші және үлкен модельдері бар MLP және CNN нейрондық архитектуралары тестіленеді. Өнімділік пен өлшем арасындағы өзара байланысты бағалау үшін біз екі модельдің нұсқаларын жаттықтырамыз.

Кесте 3.2 – Сөйлеушіні анықтауға арналған берілгендер жиыны. Төменде тиісті сәйкес келетін ID жағдайындағы әрбір сөйлеушінің аудио жазбаларының саны көрсетілген.

Деректер жинағы Data sets	Сөйлеушінің ID-і (Speaker ID)	#Барлығы (Total)
	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19	
Оқыту (Train)	63 63 56 57 58 60 63 66 59 59 63 61 65 61 60 57 54 59 56	1140
Тестілеу (Test)	12 12 19 18 17 15 12 9 16 16 12 14 10 14 15 18 21 16 19	285

Үлкен және кіші модельдердің архитектуралары 3.3-ші және 3.4-ші кестелерде жалпыланып жинақталып көрсетілген.

Бастапқы гиперпараметрлер – ол тиісінше сәйкестендірілген үлкен және кіші модельдерге арналған 128,64,32 және 512,128 өлшемдерде орналастырғандарымыздың жасырын бірліктерінің саны. Біз Relu – ді барлық қабаттар үшін іске қосу функциясы ретінде пайдаланамыз, ал шығу деңгейі 0,15-ке тең.

Кесте 3.3 – Екі міндетті (есептер) үшінде пайдаланатын MLP гиперпараметрлер.

MLP	Кіші (Small)			Үлкен (Large)		
Қабаттар (Layers)	Layer-1 (қабат-1)	Layer-2 (қабат-2)	Layer-3 (қабат-3)	Layer-1 (қабат-1)	Layer-2 (қабат-2)	Layer-3 (қабат-3)
Жасырын бөліктер (Hidden units)	128	64	32	512	256	128
Шығару (Dropout)	0.15	0.15	0.15	0.15	0.15	0.15
Белсендіру функциясы (Activation function)	Relu	Relu	Relu	Relu	Relu	Relu

Кесте 3.4 – Екі міндетті (есептер) үшін де пайдаланатын CNN гиперпараметрлер.

CNN	Кіші (Small)					Үлкен (Large)				
Layers	Conv	maxP	Conv	maxP	Dense	Conv	maxP	Conv	maxP	Dense
Hidden units (Жасырын бөліктер)	128	-	64	-	32	512	-	256	-	128
Шығару (Dropout)	0.15	-	0.15	-	0.15	0.15	-	0.15	-	0.15
Белсендіру функциясы (Activation function)	Relu	-	Relu	-	Relu	Relu	-	Relu	-	Relu

Біз гендерлік ерекшелігі мен сөйлеушіні анықтау үшін MLP және CNN-нің үлкен және кіші модельдерін пайдаланамыз. Төменде біз бір-біріне қарамастан екі міндет үшін F1 дәлдігін талдау, кері қайтару және бағалау нәтижелерін баяндаймыз.

Басқадай көрсетілмесе, онда біз модельдің сипаттамаларын салыстыру жағадайындағы F1 бағалануына жүгінеміз.

3.3.2 Гендерлік және сөйлеушінің дыбыс ерекшеліктерін тану үшін эксперимент жүргізу

Ерлер мен әйелдер санаттары бойынша гендерлікті анықтаудың нәтижелері, сондай-ақ жалпы макро нәтижелер 3.5-кестеде берілген. Ең алдымен гендерлік ерекшелігін анықтау үшін CNN негізіндегі модельдердің нәтижелері MLP-ның негіздеріне қарағанда жақсы екенін растай аламыз. CNN-small моделі MLP-large моделінен озық екенін ал MLP-large алдыңғыларына қарағанда оқытатын параметрлерінің көп санын иеленетінін көреміз.

Модельдің өнімділігіне параметрлердің белгіленуі қалай әсер ететінін білу үшін біз архитектуралардың бір түрінің ғана нәтижелерін салыстырамыз, мұнда алайда модель параметрлерінің өлшемдері әр түрлі болады: үлкен MLP-мен салыстырғанда кіші MLP және үлкен CNN-мен салыстырғандағы кіші CNN. Модель нәтижелеріне негізделген MLP ішінде үлкен MLP-ның ерлер F1-бағалауы сол түрдегі кішілерден аздаған артықшылыққа ие (1% айналасында). Бұның әйелдер санатына қатысы жоқ. Модель өлшемі үлкейтілгенде F1 бағалауы MLP-large шамамен 6% айналасында төмендейтіні анық білінеді. Бұл құбылыстың ықтимал түсіндірмелерінің бірі әйелдерге арналған аудиофайлдардың саны ерлерге қарағанда аз, бұл 3.1-кестеде көрсетілген. MLP-large нәтижесінде MLP-small моделінен асып кетуі мүмкін емес. Жоғарыдағы 3.7-3.8-ші суреттерінің ішінен біз екі MLP-large моделін және MLP-small оқыту және тестілеу процесін салыстыра аламыз. Жоғарыдағы кестелерден көрініп

тұрғандай, бұл модельдер бір деректер жинағында 500-ге жуық рет оқытылды, екі үлгіні оқытудың дәлдігі 1-ге жақын, бұл модельдерді оқытудың келтірілген мысалдары үшін жақсы оқытылғанын көрсетеді, шамадан тыс қиыстырып келтірудің ешқандай елеулі белгілері мен жеткіліксіз қиыстырылғанын көрініп тұрған жоқ. Тестілеудің қисық сызықпен айшықталған үдерісінен оқыту барысында тестілеу нәтижелері тербелістері көрінеді. Біздің қолымызда әзірлемелер жиыны жоқ болғандықтан, тестілеудің ақырғы нәтижелері үшін барлық тестілеудің нәтижелерінің ішінен тестілеудің оқыту үдерісінен алынған ең жақсы нәтижелерін таңдаудың орнына біз соңғы модельдің нәтижелерін мәлім етеміз, бұл модельдерді жинақтаудың растығын көрсету үшін жасалады.

CNN-small және CNN-large-ні өзара салыстырғанда, ерлер үшін де, әйелдер үшін де едәуір жетілдірілгені көрінеді. Бұл avg макростарынан өте анық, 3.4-кестенің бағанынынан айқын білінеді. MLP-ға негізделінген модельдерді салыстыруда, CNN әр түрлі нәтижелер көрсетеді. Мұнда MLP модель өлшемін арттырғаннан жақсармайды, алайда CNN-нен жақсы нәтиже береді. Нейрондық архитектура тұрғысынан қарағанда, біз бұл нәтижелерді түсіндіре аламыз. Мысалы, MLPs аудиосигналдарды MFCC-тің мәндерін ұзын векторлар ретінде өңдейді. Алайда CNN формалы матрица болып (ұзындығы 13) табылатын MFCC-тің неғұрлым табиғи функцияларын өңдейді, содан соң мейлінше жоғары дерексіз функцияларды алу үшін орамның/ максимум біріктірілгеннің бірнеше қабаттарын пайдаланады. 6-суретте берілген салыстыруларға қарап, CNN негізіндегі модель гендерлік ерекшелігін анықтауға арналған конвергенция үшін қадамдарды азырақ иеленеді, ал MLP тұтастай алғанда 500 қадамды, CNN тек 50-қадамды иеленеді.

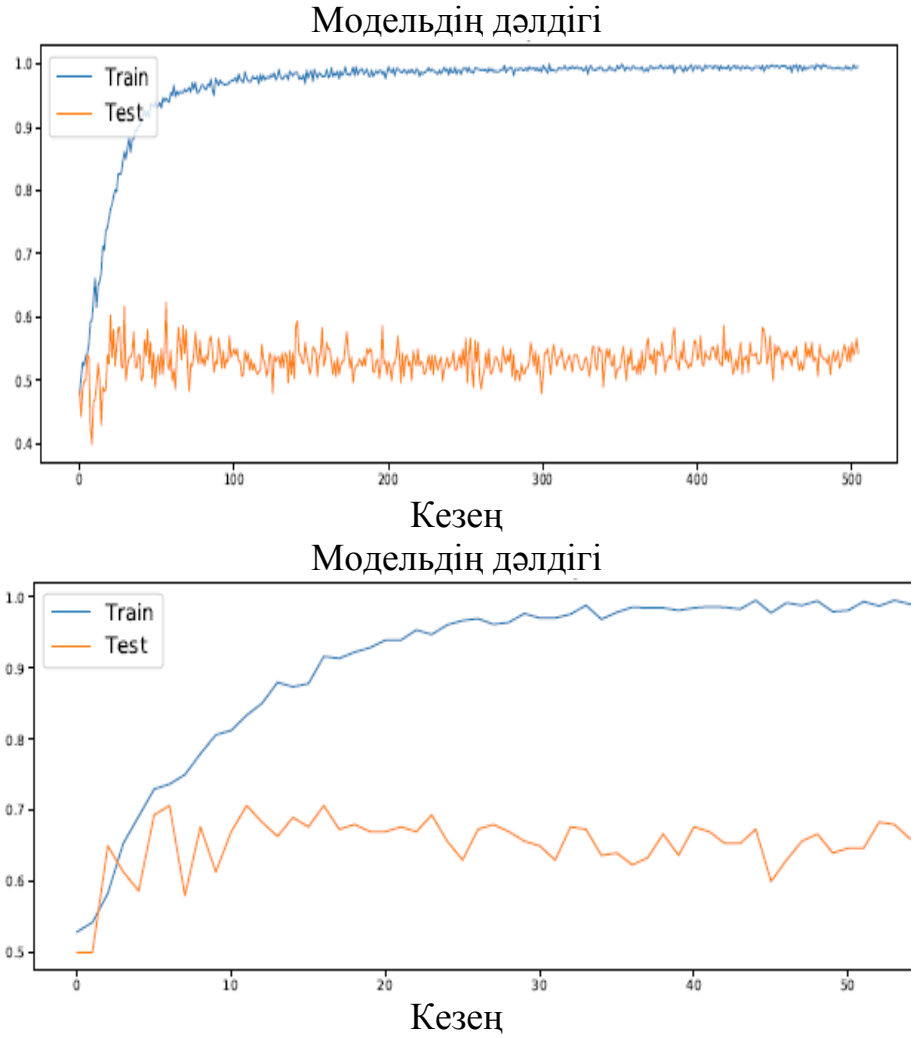
Сөйлеушіні анықтау әдісі

Бұл бөлімде сөйлеушіні анықтауды есептеу үшін нәтижелерді мәлім етеміз. Ең алдымен, макростың жалпы дәлдігі, шолу және F1-бағалауы 3.5-кестеде берілген. Екіншіден, 3.5-кестеде берілген әрбір баяндамашы бойынша дәлме-дәл нәтижелерді толық ұсынамыз. Сөйлеушіні анықтау үшін берілген құжаттағы нәтижелер гендерлік ерекшелігін анықтаудағы сияқты соншалықты жоғары емес, оның себебі – оқу аудиофайлдардың саны – әрбір диктор үшін бөлек көрсетілген.

Кесте 3.5 – Ерлер мен әйелдер категорияларына қатысты гендерлік ерекшелігін анықтаудағы дәлдік пен F1 бағалауының нәтижелері. Мұнда avg макростың, F1 бағалаудың жалпы дәлдігін береді.

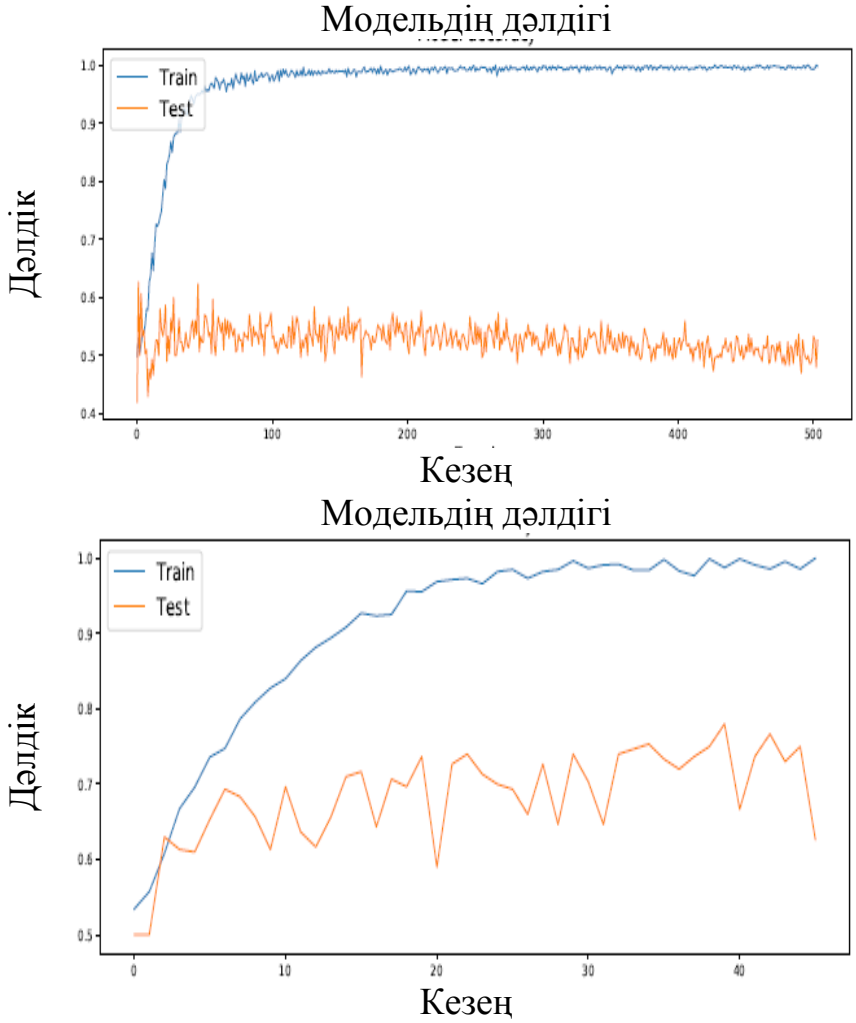
Моделдер Models	Әйел (Male)	Еркек (Female)	Marco avg.
	F1 нәтижесінің дәлдігі	F1 нәтижесінің дәлдігі	F1 нәтижесінің дәлдігі
1	2	3	4

MLP-кіші (small)	53.71	62.66	57.84	55.2	46	50.18	54.45	54.33	54.01
MLP- үлкен (large)	52.04	68	58.95	53.84	37.33	44.09	52.94	52.66	51.52
CNN-кіші (small)	63.29	**79.33**	70.41	**72.32**	54	61.83	67.80	66.66	66.12
CNN-кіші (үлкен)	**73.57**	68.86	**71.03**	70.62	**75.33**	**72.09**	**72.09**	**72**	**71.96**



Сурет 3.7 –MLP (жоғары) және CNN (төменгі) шағын модельдері бар гендерлік ерекшелігін анықтау үшін оқыту және тестілеу үдерістің дәлдігі.

Біз барлық сөйлеушіні анықтаудың бірыңғай моделін оқытатынымызды ескертеміз. Бұл модельдер берілген кірістік аудиосигнал үшін мәліметтер жиынындағы барлық сөйлеушілер үшін ықтималдықты есептеп шығарады. 3.6-кестеде көрсетілгендей дикторды анықтаудағы есептер үшін CNN-дер MLP-ден озып тұр, сөйтіп F1 бағалауында CNN-small –дың MLP –large ге қарағанда, шамамен 2% жуық жақсарған. CNN-small және CNN-large модельдерінің нәтижелері өзара салыстырыла алады. CNN үшін модель өлшемін арттырған жағдайда айтарлықтай жетілдірілуі байқалмайды. Керісінше, үлен MLP кіші MLP – ден 3% F1 бағалау артықшылығы бар. Бұл нәтижелер гендерлік ерекшелігін анықтаудан біраз өзгеше.



Сурет 3.8 – Үлкен MLP (жоғары) және үлкен CNN (төмен) модельдері бар гендерлік ерекшелігін анықтауға арналған оқыту мен тестілеу үдерістерінің дәлдігі

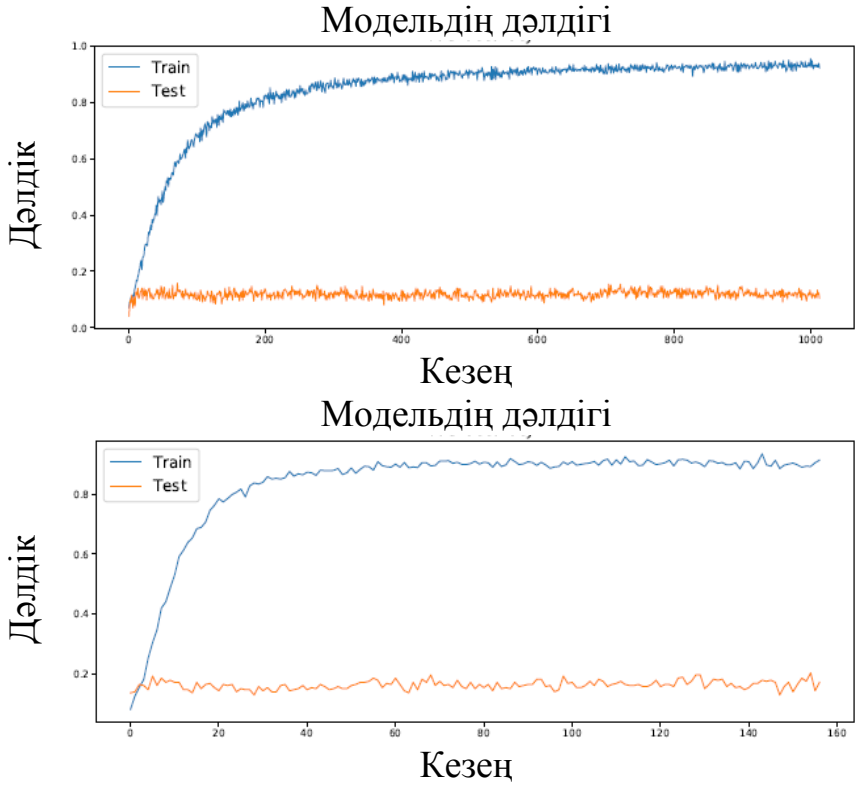
Осы көрсетілген модельдерді әр түрлі модельдің өлшемдермен салыстыру үлкен-CNN кіші CNN-нен біршама артықшылыққа ие болатынын көрсетеді. Бұл

мүмкіндіктің себебі бар: әрбір баяндамашыға арналған оқытатын мәліметтер жиынының шағын өлшемі CNN моделін жақсы оқыта алмайды, ал бірақ MLP үшін берілгендердің мұндай жиыны кішкене MLP моделіне қарағанда үлкен MLP үшін біршама жақсаруға қол жеткізу үшін жеткілікті болуы мүмкін.

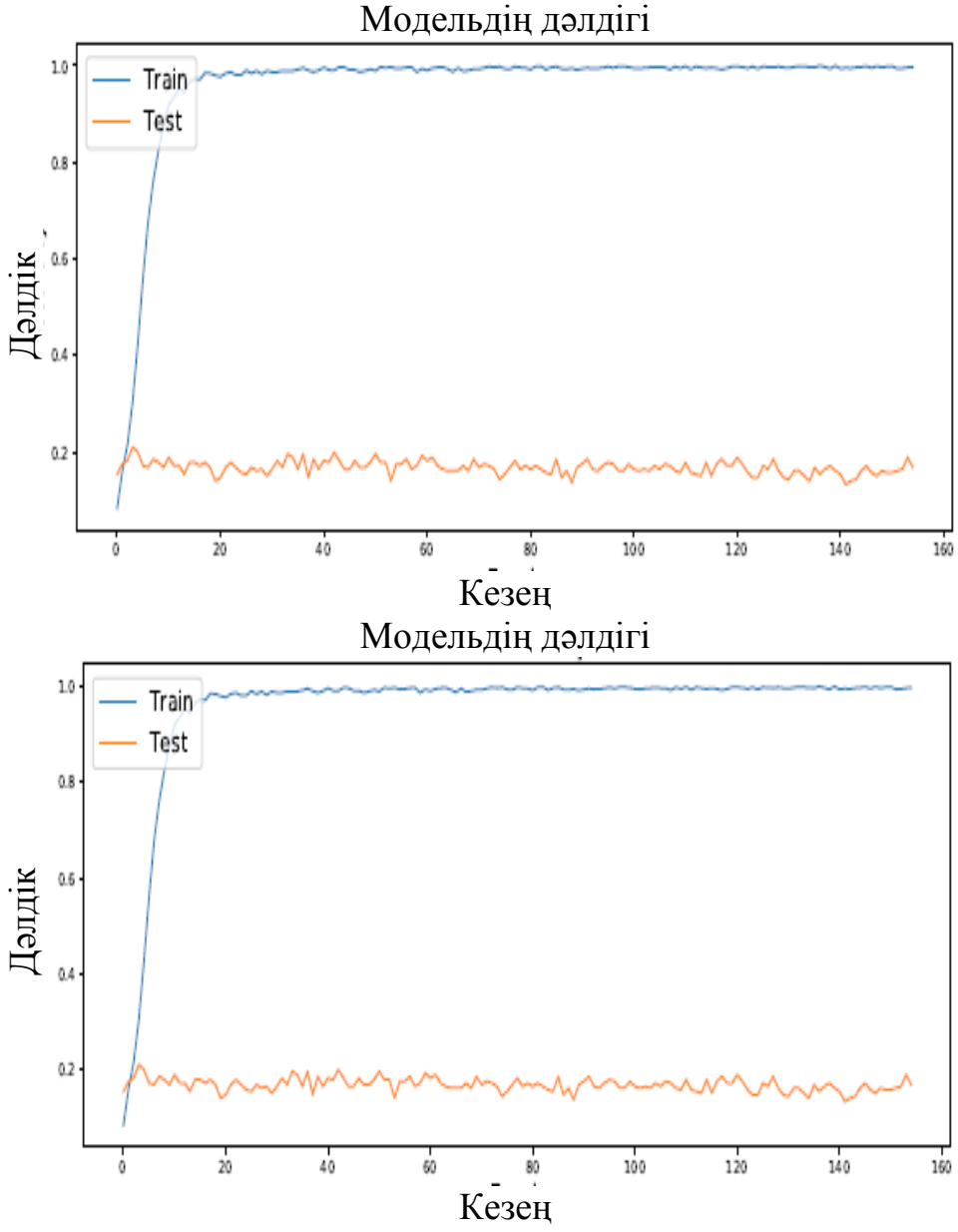
Кесте 3.6 – Сөйлеушіні анықтауға арналған жалпы макро дәлдікпен Ғ1 бағалаудың нәтижелері.

		MLP-Кіші (small)	MLP-Үлкен (large)	CNN-Кіші (small)	CNN-Үлкен (large)
Жалпы (Overall)	Дәлдік (Precision)	9.63	13.40	**16.41**	16.37
	Қайтару сигналы (Recall)	12.18	15.99	**19.08**	17.84
	F1- бағалау (F1-Score)	10.23	13.90	**16.05**	16.03

Әр түрлі өлшемдегі модельді қалыптауы бар MLP және CNN үшін оқыту дәлдігі 3.9-3.10 суреттерде көрсетілген. Мұнда MLP –small үшін 1000-нан көп қадам, ал MLP және CNN үшін бар болғаны қанша қадам керек етілетінін аңғарамыз.



Сурет 3.9 - MLP (жоғарғы) және CNN (төменгі) шағын модельдері бар сөйлеушіні анықтауға арналған оқыту мен тестілеу үрдістерінің дәлдіктері.



Сурет 3.10 –Үлкен MLP (жоғарғы) және үлкен CNN (төменгі) модельдері бар сөйлеушіні анықтауға арналған оқыту мен тестілеудің дәлдігі

Әрбір баяндамашы бойынша нәтижелер суретте немес кестеде берілген. Мұнда кейбір сөйлеушілердің нөлдік нәтижелері 3.7-кестеден байқалады, өйткені біз әрбір сөйлеуші үшін жеке-дара модель дайындаған жоқпыз. Оның орнына біз модель ішіне дыбыс жолдап, сөйтіп бұл кіріске арналған барлық сөйлеуші үшін (шығыстары-бірнеше) ықтималдықты есептеп шығарамыз. Сондай-ақ әрбір сөйлеушіге жеке-жеке модельді оқытуға болады (аудиоға кіру арнайы мақсаттағы сөйлеушіге арналған немесе арналмағанын анықтау үшін модельдің шығысы нөл мен бірлік аралығындағы бір ғана мәнді білдіреді) алайда бұл берілген еңбектің мақсатына жатпайды [82].

Кесте 3.7 – Сөйлеушіні өз идентификаторлары бойынша анықтау дәлдігі, нәтижелер және Ғ1 бағалауы.

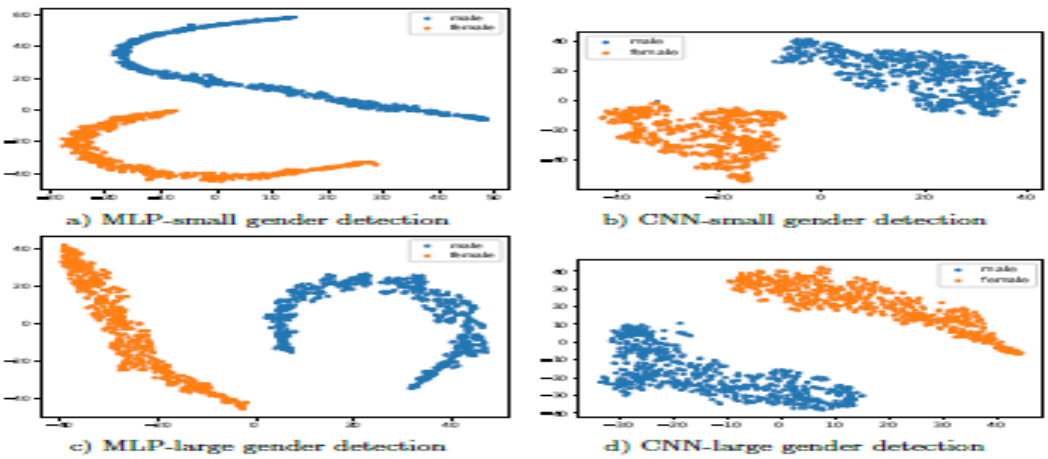
Speaker ID	Metrics	MLP-small	MLP-large	CNN-small	CNN-large
1	2	3	4	5	6
1	Precision	4.76	0.0	**21.42**	9.09
	Recall	8.33	0.0	**25**	8.33
	F1-score	6.06	0.0	23.7	8.69
2	Precision	9.0	**15.78**	12.90	13.33
	Recall	8.33	25	**33.33**	16.66
	F1-score	8.69	**19.35**	18.60	14.81
3	Precision	0.0	0.0	9.09	**16.66**
	Recall	0.0	0.0	5.26	**10.52**
	F1-score	0.0	0.0	6.66	**12.90**
4	Precision	8.3	**25**	21.42	20
	Recall	5.55	**16.66**	16.66	11.11
	F1-score	**6.66**	**20**	18.75	14.28
5	Precision	0.0	5.88	18.78	**25**
	Recall	0.0	5.88	11.76	**11.76**
	F1-score	0.0	5.88	14.28	**16**
6	Precision	0.0	0.0	16.66	**20**
	Recall	0.0	0.0	6.66	**26.66**
	F1-score	0.0	0.0	9.52	**22.85**
7	Precision	**13.33**	0.0	0.0	9.09
	Recall	**16.66**	0.0	0.0	8.33
	F1-score	**14.81**	0.0	0.0	8.69
8	Precision	26.31	**28.57**	20.83	14.70

	Recall	55.55	**66.66**	55.55	55.55
	F1-score	35.71	**40**	30.30	23.25
1	2	3	4	5	6
9	Precision	17.64	**33.33**	21.73	27.57
	Recall	18.75	25	31.25	**37.5**
	F1-score	18.18	28.	25.64	**32.43**
10	Precision	9.09	11.11	20	**25**
	Recall	6.25	6.25	6.25	**25**
	F1-score	7.4	8	9.52	**25**
11	Precision	14.28	31.25	**38.88**	27.27
	Recall	16.66	41.66	**58.33**	25
	F1-score	15.38	35.71	**46.66**	26.08
12	Precision	20.68	**46.15**	21.05	22.22
	Recall	4.28	**42.85**	28.57	28.57
	F1-score	27.90	**44.44**	24.24	25
13	Precision	4.54	0.0	**8.33**	6.66
	Recall	10	0.0	10	10
	F1-score	6.25	0.0	9.09	8
14	Precision	**23.07**	15.78	15.38	8.33
	Recall	**21.42**	21.42	**28.57**	7.14
	F1-score	**22.22**	18.18	20	7.69
15	Precision	0.0	**7.69**	0.0	0.0
	Recall	0.0	**6.69**	0.0	0.0
	F1-score	0.0	**7.14**	0.0	0.0
16	Precision	0.0	0.0	5.88	**6.66**
	Recall	0.0	0.0	5.55	**5.55**
	F1-score	0.0	0.0	5.71	**6.06**

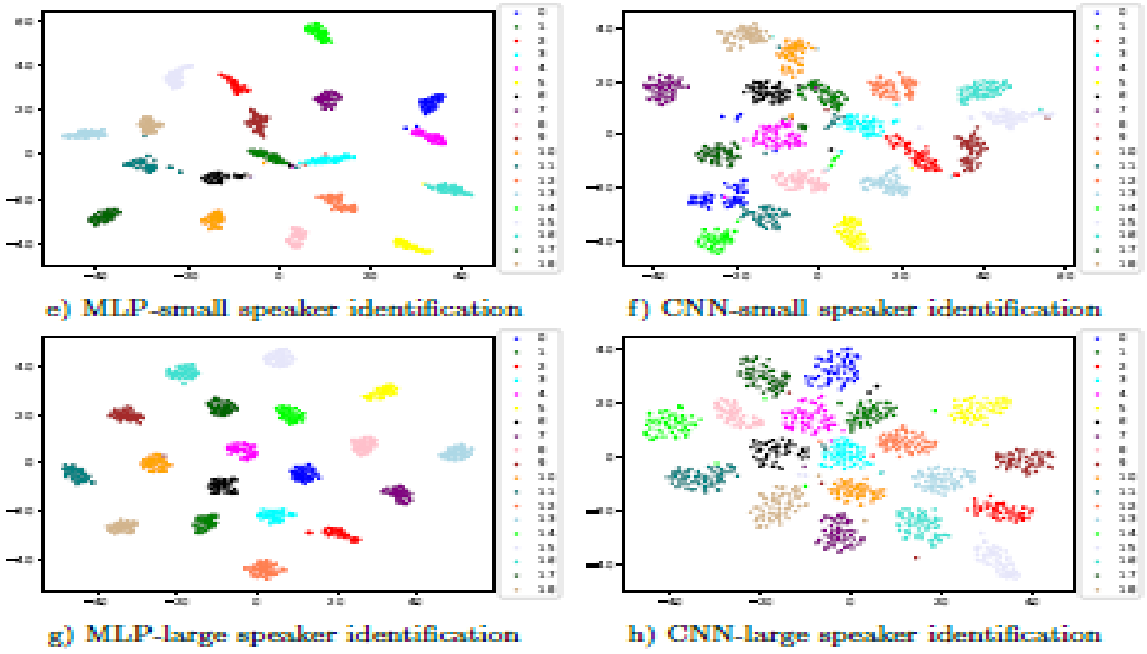
1	2	3	4	5	6
17	Precision	11.76	17.75	**33.33**	15.38
	Recall	9.52	**14.28**	9.52	9.52
	F1-score	10.52	**16.21**	14.81	11.76
18	Precision	9.0	0.0	20	**25**
	Recall	6.25	0.0	25	**31.25**
	F1-score	7.40	0.0	22.22	**27.77**
19	Precision	11.11	15.38	6.66	**18.18**
	Recall	5.26	**21.57**	5.26	10.52
	F1-score	7.14	**20.68**	5.88	13.33

Визуациялау.

Гендерлік ерекшелігін анықтау үшін және сөйлеушіні анықтау үшін модельдік анықтаудан кейін оқытатын аудиожазбаларды визуалдау 3.11-суретте көрсетілген. Біз кіріс дыбысты ұсыну ретінде әрбір модельдің шығыс қабатының алдындағы қабаттан нысан векторларын шығарып аламыз. 3.11-суреттен аңғарылатындай, аудиожазбалар екі топқа бөлінеді: ерлер және әйелдер тобы. MLP және CNN-ді визуалдау нәтижелері әр түрлі, CNN-ға қарағанда MLP нәтижесі аудио үлгілерінің іштей айқын байланыстылығын көрстетеді. Эксперименттік нәтижелерден көрініп тұрғандай, CNN моделі MLP моделінен гендерлік ерекшелігін анықтау үшін де, сондай-ақ дикторды ұқсатып анықтау үшін де озып тұр. MLP-нің шағын салада диктордың бірыңғай аудиожазбаларын орналастырып тұрғанын көре аламыз. Алайда CNN модельдері кең салада орналасқан, нәтижелері CNN моделі MLP-ға қарағанда жақсы жинақталуды көрсете алуы мүмкін. 3.11[a,e] суретте үлкен және кіші MLP-ны пайдаланғандағы көк түспен берілгендер ер адамның дауысын білдіреді. Ал 3.11[b,d] суретте үлкен және кіші CNN-ді пайдаланғандағы сары түспен берілгендер әйел даусын білідреді. 3.12[e,g] суретте үлкен және кіші MLP-ны пайдаланғандағы сөйлеушіні анықтауды білдіреді Ал 3.12[f,h] суретте үлкен және кіші CNN-ді пайдаланудағы сөйлеушіні анықтауды білдіреді.



Сурет 3.11 – MLP және CNN модельдерін оқытқаннан кейін оқитын аудиофайлдарды гендерлік ерекшелігін анықтау үшін визуализациялау



Сурет 3.12 – MLP және CNN модельдерін оқытқаннан кейін оқитын аудиофайлдарда сөйлеушіні анықтау үшін визуализациялау

Бұл бөлімде гендерлік ерекшелігін және сөйлеушіні анықтуда нейрондық желі архитектураларына салыстырмалы түрде талдау жасалды. Онда MLP – ға қарағанда CNN жақсы нәтиже көрсетті.

Эксперимент екі нейрондық архитектураны қамтиды: MLP және CNN, олар түрлі модельдік реттеулермен берілген. Екі бірдей есептеу үшін CNN моделі MLP-дан озық, мұнда гендерлік ерекшелігін анықтауда салыстырмалы ұтымдылық 10% - дан шамамен 20% - ға дейінгі нұсқаларда түрленіп отырады. Сөйлеушіні анықтаудағы салыстырмалы ұтымдылық 2% - дан 6% - ға дейін ауытқиды. Бұл нәтиже бірнеше аспектілермен белгіленген.

Біріншіден, сол берілген архитектураны салыстыру гендерлік ерекшелігін анықтайтын MLP үшін үлкен модельді орнату соншалықты көмектеспейтінін көрсетеді. Үлкен CNN кіші CNN-ге қарағанда жақсы және озық жұмыс істейді.

Екіншіден, CNN-сигналдардан нысандарды алып шығу үшін өзінің жинамаларының қабаттарын пайдаланады, бұның өзі модельді MFCC нысандарының бастапқы нұсқалары сақталған кірістің бар болуын иеленуге қабілетті етеді. MLP бірнеше өзара толық байланыстырылған қабаттардан тұрады. Олар матрицалы нысанды өңдей алмайды, және сонымен қатар MFCC-ның бастапқы нысандарын, нысандардың ұзын векторларына айналдыра алуға тиіс болады. Бұл MLP-ға қарағанда, екі есептеу (міндет) үшін де себептен CNN-нің жақсы жұмыс істейтініне дәлелді түсініктеме бола алады.

Үшіншіден, MLP және CNN жаттығу процесін салыстыруда бірінші қадамның көп мөлшерін алады, ал екіншісі едәуір аз алады. Визуалдау нәтижелері MLP бірдей диктордың аудиожазбаларын шағындау салаға орналастыратынын көрсетеді, ал CNN салыстырмалы түрде кең аймақта орналасқан, сөйтіп бұл нәтижелер CNN моделі MLP-ға қарағанда жақсы жинақтайтынын көрсетеді.

ҚОРЫТЫНДЫ

Диссертациялық жұмыста сөйлеулерді тану есептерінде машиналық оқытуды қолданып белгілерді анықтау алгоритмдері мен модельдері қарастырылды.

Бірінші тарауда сөйлеу сигналын алдын ала өңдеудің жолдары мен белгілерін анықтаудың ерекшеліктері, сөйлеу сигналдарының сипаттамасы, сөйлеуді тану және белгілерін анықтауға арналған әдіс-тәсілдер мен моделдерге талдау жасалды.

Екінші тарауда машиналық оқыту алгоритмдері мен модельдерін сөйлеуді тану есептерінде қолдану және машиналық оқытудағы нейрондық желілер, қарастырылды. Сөйлеушіні анықтауға арналған акустикалық корпус құрылды. Сөйлеушіні анықтауға арналаған классификациялық алгоритмдерге талдау жасалып SVC жоғары нәтиже көрсетті.

Үшінші тарауда сөйлеулерді тану есептерінде машиналық оқытуды қолданып белгілерді анықтау және өңдеу алгоритмдері мен модельдері құрылды. Сөйлеу сигналын алдын ала өңдеуде MFCC-ті қолданып гендерлік ерекшелігі анықталды. Гендерлік ерекшелігі мен сөйлеушінің дыбыс ерекшеліктерін тануға арналған MLP және CNN нейрондық желі архитектураларына салыстырмалы түрде талдау жасалып, CNN жақсы нәтиже көрсеткені анықталды.

Жұмыс нәтижесі бойынша барлық қойылған міндеттер орындалды. Зерттеу жұмысы барысында келесідей нәтижелерге қол жеткізілді:

- Сөйлеуді тану үдерісінде сөйлеушінің дыбыстық сөйлеу белгілерін және сөйлеушінің мәліметтерін анықтауға арналған акустикалық корпусы құрылды;
- Машиналық оқыту саласындағы классификациялық алгоритмдер көмегімен және осы алгоритмдердің дәлдігін арттыра отырып сөйлеуші анықталды;
- Нейрондық желілер негізіндегі гендерлік ерекшелігі мен сөйлеушіні анықтаудың моделі мен алгоритмі құрылды;
- Зерттеу барысында алынған модель мен алгоритм көмегімен сөйлеу белгілерін және сөйлеушіні анықтауға арналған бағдарламалық қосымша құрылды;
- Диссертациялық жұмыс барысында бағдарламалық қосымша негізінде екі авторлық куәлік алынды.

ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ

1 Rabiner L. A tutorial on hidden markov models and selected applications in speech recognition. –1989. – P. 257–286.

2 Larochelle H., Erhan D., Courville D. An Empirical Evaluation of Deep Architectures on Problems with Many Factors of Variation // International Conference on Machine Learning. –2007. – P. 473–480

3 Mermelstein P. Distance measures for speech recognition, psychological and instrumental // Pattern recognition and artificial intelligence. – 1976. – Vol. 116. –P. 374–388.

4 Grerzl F. Probabilistic and bottle-neck features for the BN features LVCSR of meetings // In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2007. – P. 4729–4732.

5 Agnitio — Voice Biometrics [Электронный ресурс] : официальный сайт компании «Agnitio». URL: http://www.agnitio.es. 29.10.2015.

6 Рабинер, Л. Р. Цифровая обработка речевых сигналов : пер. с англ. / Л. Р. Рабинер, Р. В. Шафер; под ред. М. В. Назарова и Ю. Н. Прохорова. – М. : Радио и связь, 1981. – 496 с.

7 Гробман, М. З. Выделение скрытых периодичностей и формантный анализ речи. Распознавание образов : теория и приложения / М. З. Гробман, В. И. Тумаркин. – М. : Наука, -1977. – 316 с.

8 Потапова, Р. К. О типологических особенностях слога. Распознавание образов: теория и приложения / Р. К. Потапова. – М. : Наука, 1977. –296 с.

9 Сорокин, В. Н. Элементы кодовой структуры речи. Распознавание образов: теория и приложения. – М. : Наука, 1977. – с. 42 – 60.

10 Рабинер, Л. Р. Цифровая обработка речевых сигналов / Л. Р. Рабинер, Р. В. Шафер. – М. : Радио и связь, 1981. – 496 с.

11 Сергиенко, А. Б. Цифровая обработка сигналов / А. Б. Сергиенко. – СПб. : Питер, 2002. – 608 с.

12 Кучерявый, А. А. Бортовые информационные системы : курс лекций / А. А. Кучерявый под ред. В. А. Мишина, Г. И. Клюева. – 2-е изд., перераб. и доп. – Ульяновск : УлГТУ, 2004. – 504 с.

13 Тэйлор, Р. Шум / Р. Тэйлор пер.с англ. Д. И. Арнольда. – М. : Мир, 1978. – 308 с.

14 Отт, Г. Методы подавления шумов и помех в электронных системах / Г. Отт ; пер. с англ. Б. Н. Бронина; под ред. М. В. Гальперина. – М. : Мир, 1979. – 318 с.

15 Алимурадов, А. К. Фильтрация речевых сигналов с использованием метода множе- ственной декомпозиции и оценки энергии эмпирических мод / А. К. Алимурадов, П. П. Чураков, А. Ю. Тычков // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2012. – № 4. – С. 50–61.

16 Михайлов, Е. В. Помехозащищенность информационно-измерительных систем // Е. В. Михайлов. – М. : Энергия, 1975. – 312 с.

17 Шахов, Э. К. Повышение помехоустойчивости цифровых средств

измерения // Э. К. Шахов. – Пенза : ППИ, 1983. – 48 с.

18 Методы автоматического распознавания речи : в 2 кн. : пер. с англ. / Д. Х. Клетт, Дж. А. Барнет, М. И. Бернстейн и др. под ред. У. Ли. – М. : Мир, 1983. – Кн. 2. – 392 с.

19 Методы автоматического распознавания речи : в 2 кн. : пер. с англ. // У. А. Ли, Э. П. Нейбург, Т. Б. Мартин и др. ; под ред. У. Ли. – М. : Мир. 1983. – Кн. 1. – 328 с.

20 Дигун, О. Г. Сигналы, помехи, шумы : учеб. пособие // О. Г. Дигун, В. И. Веприков. – Новочеркасск : НГТУ, 1994. – 94 с.

21 Болл, Р. М. Руководство по биометрии // Р. М. Болл, Дж. Х. Коннел, Н. К. Ратха ; пер с англ. Н. Е. Агапова. – М. : Техносфера, 2007. – 352 с.

22 Фролов, А. В. Синтез и распознавание речи. Современные решения // Г. В. Фролов. –М. : Связь, 2003. – 216 с.

23 Алимурадов, А. К. Определение частоты основного тона речевого сигнала с использо- ванием метода множественной декомпозиции на эмпирические моды / А. К. Алимура- дов, П. П. Чураков, А. Ю. Тычков // Модели, системы, сети в экономике, технике, при- роде и обществе. – 2012. –№ 1 (2). – С. 121–126.

24 Алимурадов, А. К. Определение частоты основного тона в задаче идентификации лич- ности по голосу / А. К. Алимурадов // Молодежь и наука: модернизация и инновацион- ное развитие страны : сб. тр. II Междунар. науч.-практ. конф. студентов и молодых ученых. – Пенза, 2012. – С. 399–403.

25 B. Schuller, A. Batliner, S. Steidl, D. Seppi., B. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. – 2011. – Vol. 53.– P.1062–1087.

26 Алимурадов, А. К. Выбор оптимального набора информативных параметров речевых сигналов для систем голосового управления / А. К. Алимурадов, П. П. Чураков, А. Ю. Тычков // Измерение. Мониторинг. Управление. Контроль. – 2013. – № 1 (3). – С. 16–20.

27 Алимурадов, А. К. Выбор оптимального набора информативных параметров речевых сигналов для систем голосового управления / А. К. Алимурадов, П. П. Чураков, А. Ю. Тычков // Измерение. Мониторинг. Управление. Контроль. – 2013. – № 1 (3). – С. 16–20.

28 Wölfel M., McDonough J. Distant speech recognition. John Wiley & Sons. – 2009. – P. 573– 580.

29 Howard D., Angus J. Acoustics and psychoacoustics. Taylor & Francis, – 2009. – P. 473– 480.

30 Тихонов А. Н., Самарский А. А. Уравнения математической физики. – М., 1999. –735с.

31 Baum L.E., Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. // Ann. Math. Stat. – 1966. – Vol. 37, – P. 1554 – 1563,

32 Савченко В. В. Информационная теория восприятия речи. // Известия вузов России. Радиоэлектроника. – 2007. – Вып.6. – С.3-9.

33 Гоноровский И.С. Радиотехнические цепи и сигналы. - М.: Радио и

связь, – 1986. – 512 с.

34 Савченко В. В., Акатьев Д. Ю., Карпов Н. В. Автоматическое распознавание элементарных речевых единиц методом обеляющего фильтра.

35 Савченко В.В. //АВТОМЕТРИЯ. – 1996. – №2. – С.77

36 Мамырбаев О.Ж. Кыдырбаева А.С., Ахмедиярова А.Т., М.Тұрдалыұлы. Н.О.Мекебаев. Систематический обзор и анализ особенностей идентификации по голосу // ҚБТУ хабаршысы. – 2019. – №2 (49). – Б. 120-133.

37 Савченко В.В. //Радиотехника и электроника. – 1999. – Т.44. – №1. – С.65

38 Савченко В. В., Савченко А. В. Принцип минимального информационного рассогласования в задаче распознавания дискретных объектов// Известия вузов России. Радиоэлектроника. – 2005. Вып.3. – С.10-19.

39 Савченко В. В. Информационная теория восприятия речи. // Известия вузов России. Радиоэлектроника. –2007. Вып.6. – С.3-9.

40 Савченко В. В., Акатьев Д. Ю., Карпов Н. В. Автоматическое распознавание элементарных речевых единиц методом обеляющего фильтра. // Известия вузов. Радиоэлектроника. – 2007. Вып.4. – С.11-19.

41 Кульбак С. Теория информации и статистика. – М.: Наука, – 1967. – 408 с.

42 Савченко В. В. Различение случайных сигналов в частотной области// Радиотехника и электроника. – 1997. – Т.42, – №4. – С. 426–431.

43 L.R. Rabiner, B.-H. Juang, Speech recognition: statistical methods, in: Encyclopedia of Linguistics, – 2006. – P. 1–18.

44 W. Chan, N. Jaitly, Q. Le, O. Vinyals, Listen, attend and spell: a neural network for large vocabulary conversational speech recognition, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2016. –P. 4960–4964.

45 J.K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, Y. Bengio, Attention-based models for speech recognition, in: Advances in Neural Information Processing Systems. – 2015, P. 577–585.

46 Y. Miao, M. Gowayyed, F. Metze, EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding, in: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). – 2015, –P. 167–174.

47 J. Chorowski, D. Bahdanau, K. Cho, Y. Bengio, End-to-end continuous speech recognition using attention-based recurrent NN: First results, arXiv: 1412.1602(2014).

48 D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, Y. Bengio, End-to-end attention-based large vocabulary speech recognition, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), – 2016, –P. 4945–4949.

49 McCulloch W. S., Pitts W. H. A logical calculus of ideas immanent in nervous activity // Bull. Math. Biophysics. – 1943. – Vol. 5. – P. 115-119.

50 Rosenblatt F. Principles of Neurodynamics // Spartan Books, New York. – 1959. – P. 210-229.

51 Lippmann R.P. Review of neural networks for speech recognition // Neural Computing. – 1989. –P. 1-38.

52 Rumelhart D. E., Hinton G. E., Williams R. J. Learning internal representations by error propagation // In: Rumelhart, D. E., G. E. Hinton, (eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol.1 Foundations., chapter 8. Bradford Books/MIT Press, Cambridge, MA. ISBN 0-262-18120-7. – 1986.

53 Pearlmutter B. A. Dynamic Recurrent Neural Networks // Technical Report CMU-CS-88-191, Carnegie-Mellon University, Computer Science Dept. Pittsburg, PA. – 1990.

54 Pineda F.J. Recurrent back-propagation and the dynamical approach to adaptive neural computation // Neural Computing. – 1989. No. 1. – P. 161–72.

55 Gori M., Bengio Y., R. De Mori BPS: a learning algorithm for capturing the dynamical nature of speech // Proceedings of the International Joint Conference on Neural Networks, Washington, DC, IEEE, New York. – 1989, –P. 643–644.

56 Williams R. J., Zipser D. A learning algorithm for continually running fully recurrent neural networks // Neural Comput. – 1989. No. 1. – P. 87–111.

57 Pearlmutter B. A. Learning state space trajectories in recurrent neural networks // Neural Comput. – 1989. No. 1. – P. 263–269.

58 Sato M. A real time learning algorithm for recurrent analog neural networks // Biol. Cybernet. 62. – 1990. – P. 237–241

59 Bengio Y., Simard P., Frasconi P. Learning long-term dependencies with gradient descent is difficult // IEEE Transaction on Neural Networks . – 1994. – P. 157–166.

60 Кривнова, О. Ф. Речевые корпуса на новом технологическом витке / О. Ф. Кривнова // Речевые технологии. – 2008. – № 2.– С.13 – 23.

61 Hunt A. , Black A. W. Unit selection in a concatenative speech synthesis system using a large speech database // ICASSP-96. – 1996.–vol. 1, – P. 373–376.

62 Gibbon, D.,Moore, R., Winski, R. (Editors) Handbook of Standards and Resources for Spoken Language Systems Mouton de Gruyter. –1997. – P. 113–165.

63 C. Zhan, W. Li and P. Ogunbona, “Face Recognition from Single Sample based on Human Face Perception,” International Conference Image and Vision Computing New Zealand. – 2009. – P. 56–61.

64 Beigi, Homayoon. Fundamentals of speaker recognition. Springer Science & Business Media. – 2011. – P. 942.

65 S. Yella, N. Gupta and M. Dougherty, "Comparison of pattern recognition techniques for the classification of impact acoustic emissions", Transportation Research Part C: Emerging Technologies. – 2007. – vol. 15. – no. 6. – P. 345–360.

66 Serizel, R., Giuliani, D. Vocal tract length normalization approaches to DNN-Based children’s and adults’ speech recognition. IEEE Workshop on Spoken Language Technology. – 2014. – P. 135–140.

67 Popović, B., Ostrogonac, S., Pakoci, E., Jakovljević, N., Delić, V. Deep neural network based continuous speech recognition for Serbian using the Kaldi toolkit. In: Ronzhin, A., Potapova, R., Fakotakis, N. – SPECOM 2015. – vol. 9319. –

P. 186–192.

68 Orken Mamyrbayev, Mussa Turdalyuly, Nurbapa Mekebayev, Keylan Alimhan, Aizat Kydyrbekova, Tolganay Turdalykyzy “Automatic Recognition of Kazakh Speech Using Deep Neural Networks” Lecture Notes in Computer Science. – 2019. – P. 465– 474.

69 Mamyrbayev O.Zh., Kunanbayeva, M.M., Sadybekov K.S., Kalyzhanova A.U., Mamyrbayeva A.Z. “One of the methods of segmentation of speech signal on syllables” bulletin of the national academy of sciences of the republic of kazakhstan. – 2015. – P. 286–290.

70 H. Harb and L. Chen, Voice-based gender identification in multimedia applications, Journal of Intelligent Information Systems. – 2005. – P. 184–190.

71 D. Ververidis and C. Kotropoulos. Automatic speech classification to five emotional states based on gender information. In Proc. XII European Signal Processing Conf. – Vienna, Austria. – 2004. –Vol. 1. – P. 341–344.

72 S. Greenberg, J. Hollenback, and D. Ellis, “Insights intospoken language gleaned from phonetic transcription of the Switchboard corpus”, in Proc. ICSLP. – 1996. –Vol. 1. – P. 32–35.

73 Sahidullah, M., Saha, G.: Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. Speech Communication54. – 2012. –Vol. 54. – P. 543–565.

74 Freund, Y., Schapire, R.E.: Large margin classi_cation using the perceptron algorithm. Mach. Learn. – 1999. –Vol. 3. – P. 277–296.

75 Auer, P., Burgsteiner, H., Maass, W.: A learning rule for very simple universal approximators consisting of a single layer of perceptrons. Neural Netw. – 2008. –Vol. 5. – P. 786–795.

76 Schmidhuber, J.: Deep learning in neural networks: An overview. – 2015. – P. 85–117

77 Toleu, A., Tolegen, G., Makazhanov, A.: Character-aware neural morphological disambiguation. In: Association for Computational Linguistics, Vancouver, Canada. – 2008. –Vol. 2. – P. 666–671.

78 Mamyrbayev, O., Turdalyuly, M., Mekebayev, N., Alimhan, K., Kydyrbekova, A.,Turdalykyzy, T.: Automatic recognition of kazakh speech using deep neural net-works. – 2019. – P. 465–474.

79 Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classication with deep convolutional neural networks. Curran Associates, Inc. – 2012. – P. 1097–1105.

80 Collobert, R., Weston, J.: A unied architecture for natural language processing: Deep neural networks with multitask learning. – 2008. – P. 160–167.

81 van den Oord, A., Dieleman, S., Schrauwen, B.: Deep content-based music recommendation. – 2013.– P. 2643–2651.

82 Ali, Md Sadek, Md Shariful Islam, and Md Alamgir Hossain. "Gender recognition system using speech signal." International Journal of Computer Science, Engineering and Information Technology (IJCSEIT). – 2018. –Vol. 2. – No. 1 – P. 1–9.

83 Alías, Francesc, Joan Claudi Socoró, and Xavier Sevillano. "A review of

physical and perceptual feature extraction techniques for speech, music and environmental sounds." Applied Sciences. – 2016. – No. 5. – P. 143–150

84 Subramanian, Hariharan, P. Rao, and S. D. Roy. "Audio signal classification." EE Dept, IIT Bombay. – 2004.– P. 1–5.

85 Bach, Jörg-Hendrik, Jörn Anemüller, and Birger Kollmeier. "Robust speech detection in real acoustic backgrounds with perceptually motivated features." Speech Communication. – 2011. –Vol. 53. – No. 5. – P. 690–706.

86 Campbell, J. P., "Speaker recognition: a tutorial" Proceedings of IEEE. – 1997. –Vol. 85. – No. 9 – P. 1437–1462.

87 Sadaoki Furui, "Digital Speech Processing, Synthesis and Recognition", 2nd edition. – 1989. – 320 p.

88 Jian Yang; Zhang, D.; Frangi, A.F.; Jing-yu Yang; "Two-dimensional PCA: a new approach to appearance-based face representation and recognition," Pattern Analysis and Machine Intelligence, , IEEE Transactions. – 2004. –Vol. 26. – No. 1. – P. 131–137.

89 Jolliffe. I. T. Principal Component Analysis", Second Edition Aapo Hyvarinen, Juha Karhunen and Erkki Oja. "Independent Component Analysis". – 2001. – 518 p.

90 Stouten, F., Duchateau, J., Martens, J.-P., Wambacq, P.: Coping with disfluencies in spontaneous speech recognition: acoustic detection and linguistic context manipulation. Speech Commun. – 2006.– P. 1590–1606.

91 Tsiaras, V., Panagiotakis, C., Stylianou, Y.: Video and audio based detection of filled hesitation pauses in classroom lectures. In: Proceedings of the 17th European Signal Processing Conference (EUSIPCO 2009), Glasgow, Scotland. – 2009.– P. 834–838.

92 Psutka, J., Ircing, P., Psutka, J.V., Hajič, J., Byrne, W.J., Mirovsky, J.: Automatic transcription of Czech, Russian, and Slovak spontaneous speech in the MALACH project. In:Proceedings of Eurospeech, Portugal, Lisboa. – 2005.– P. 1349–1352.

93 Young, S., et al.: The HTK Book (for HTK Version 3.4), Cambridge, UK. – 2009. – 375 p.

94 Karpov, A., Kipyatkova, I., Ronzhin, A.: Very large vocabulary ASR for spoken Russian with syntactic and morphemic analysis. In: Proceedings INTERSPEECH-2011, Florence, Italy. – 2011.– P. 3161–3164.

95 Serizel, R., Giuliani, D.: Vocal tract length normalization approaches to DNN-Based children's and adults' speech recognition. In: IEEE Workshop on Spoken Language Technology. – 2014.– P. 135–140.

96 Behbahani, Y.M., Babaali, B., Turdalyuly, M.: Persian sentences to phoneme sequences conversion based on recurrent neural networks. Open Comput. Sci. 6, 219–225 (2016) – 2016.– P. 219–225.

97 Yu, D., Deng, L.: Automatic Speech Recognition.. Springer, London. – 2014. – 315 p.

98 О.Ж. Мамырбаев, М. Тұрдалыұлы, Н.О. Мекебаев. Кіріккен қазақ сөйлеін тану жүйесі // ҚБТУ хабаршысы. – 2018. – № 3 (46). – Б. 129-133.

99 Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi “Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) Techniques”. – 2010. –Vol. 2 – P. 138–143.

100 Watcher, M. D., Matton, M., Demuynck, K., Wambacq, P., Cools, R., “Template Based Continuous Speech Recognition”, IEEE Transaction on Audio, Speech, & Language Processing. – 2007. –Vol. 4. – P. 1377–1390.

101 Gupta, R., and Sivakumar, G., “Speech Recognition for Hindi Language”, IIT BOMBAY. – 2006. 97 p.

102 Қалимолдаев М.Н., Мамырбаев О.Ж., Н.О. Мекебаев., Тұрдалыұлы М. Машиналық оқытуды қолдануда дауыстың гендерлік жіктелінуі // ҚазҰТЗУ хабаршысы – 2019. – № 6 (136). – Б.229–233.

103 Ingle V., Proakis J. Digital Signal Processing Using Matlab V4 – Boston: ITP. – 1997. – 420 p.

104 Rabiner, L. Juang, B. H., Yegnanarayana, B., “Fundamentals of Speech Recognition”, Pearson Publishers. – 2010. – 450 p.

105 Orken Mamyrbayev, Nurbapa Mekebayev, Mussa Turdalyuly, Nurzhamal Oshanova,Tolga Ihsan Medeni and Aigerim Yessentay. Voice Identification Using Classification Algorithms//We are IntechOpen, the world’s leading publisher of Open Access books Built by scientists, for scientists. London**.** – 2019. – 1 – 14 p.

106 Mamyrbayev O, Toleu A, Tolegen G, Mekebayev N. Neural Architectures for Gender Detection and Speaker Identification // Journal Cogent Engineering. ISSN: 2331-1916. – 2020. – Vol.7. – Issue 1. – P.1–13.

107 О.Ж. Мамырбаев, Н.О. Мекебаев, М. Тұрдалыұлы. Қазақ сөйлеуін тануда іргелі және қолданбалы зерттеуге арналған фонетикалық мәтін // «Төртінші өнеркәсіптік революция жағдайындағы дамудың жаңа мүмкіндіктері» атты ҚР Президенті Н. Назарбаевтың Жолдауын іске асыру шеңберінде «Көліктегі инновациялық технологиялар: білім, ғылым, тәжірибе» атты XLII Халықаралық ғылыми-практикалық конференцияның материалдары. – Алматы, 2018. – Т. 2. – Б. 81-87.

108 О.Ж. Мамырбаев, М. Тұрдалыұлы, Н.О. Мекебаев. Қазақ тілі сөйлеуінің акустикалық және тілдік модельдерін құру // Материалы XIV Международной Азиатской школы-семинара «Проблемы оптимизации сложных систем». – Алматы, 2018. – Т. 2. – Б. 344-347.

109 О.Ж. Мамырбаев, Н.О. Мекебаев, М. Турдалыулы. Алгоритмы и архитектуры систем распознавания речи // Материалы III Международной научной конференции «Информатика и прикладная математика» посвященная 80-летнему юбилею профессора Бияшева Р.Г. и 70-летию профессора Айдарханова М.Б. – Алматы, 2018. – Т. 2. – С. 108-121.

110 О.Ж. Мамырбаев, Н.О. Мекебаев, М. Турдалыулы. Алгоритмы и архитектуры систем распознавания речи // Международной научной конференции «Информатика и прикладная математика» посвященная 80-летнему юбилею профессора Бияшева Р.Г. и 70-летию профессора Айдарханова М.Б. – Алматы, 2018. – Т. 2. – С. 108-121.

ҚОСЫМША А

ЭЕМ-ге арналған бағдарлама «System of automatic creation vocabulary for ASR» авторлық куәлік 2019 жылғы 22 қаңтар № 1425.

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ

РЕСПУБЛИКА КАЗАХСТАН

АВТОРЛЫҚ ҚҰҚЫҚПЕН ҚОРҒАЛАТЫН ОБЪЕКТІЛЕРГЕ ҚҰҚЫҚТАРДЫҢ МЕМЛЕКЕТТІК ТІЗІЛІМГЕ МӘЛІМЕТТЕРДІ ЕНГІЗУ ТУРАЛЫ

КУӘЛІК

2019 жылғы « 22 » қаңтар № 1425

Автордың (лардың) жөні , аты, әкесінің аты (егер ол жеке басын куәландыратын құжатта көрсетілсе):

МАМЫРБАЕВ ОРКЕН ЖУМАЖАНОВИЧ; ТҰРДАЛЫҰЛЫ МЫСА ; МЕКЕБАЕВ НҰРБАПА ОТАНОВИЧ; ТҰРДАЛЫҚЫЗЫ ТОЛҒАНАЙ ; СЕЙТҚАЛИ БЕКЖАН НҰРЛАНҰЛЫ ДУЙСЕНБАЕВА АЙГЕРИМ ЖАНБОЛАТОВНА

Авторлық құқық объектісі: ЭЕМ-ге арналған бағдарлама

Объектінің атауы: System of automatic creation of vocabulary for ASR

Объектіні жасаған күні: 01.06.2018

СВИДЕТЕЛЬСТВО

О ВНЕСЕНИИ СВЕДЕНИЙ В ГОСУДАРСТВЕННЫЙ РЕЕСТР ПРАВ НА ОБЪЕКТЫ, ОХРАНЯЕМЫЕ АВТОРСКИМ ПРАВОМ

№ 1425 от « 22 » января 2019 года

Фамилия, имя, отчество, (если оно указано в документе, удостоверяющем личность) автора (ов):

МАМЫРБАЕВ ОРКЕН ЖУМАЖАНОВИЧ; ТҰРДАЛЫҰЛЫ МЫСА ; МЕКЕБАЕВ НҰРБАПА ОТАНОВИЧ; ТҰРДАЛЫҚЫЗЫ ТОЛҒАНАЙ ; СЕЙТҚАЛИ БЕКЖАН НҰРЛАНҰЛЫ ДУЙСЕНБАЕВА АЙГЕРИМ ЖАНБОЛАТОВНА

Вид объекта авторского права: программа для ЭВМ

Название объекта: System of automatic creation of vocabulary for ASR

Дата создания объекта: 01.06.2018

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

Подписано ЭЦП

Оспанов Е. К.

ЭЕМ-ге арналған бағдарлама «Мульти язычное распознавание речи MultiSpeech» авторлық куәлік 2020 жылғы 09 қаңтар, № 7844.

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ

РЕСПУБЛИКА КАЗАХСТАН

АВТОРЛЫҚ ҚҰҚЫҚПЕН ҚОРҒАЛАТЫН ОБЪЕКТІЛЕРГЕ ҚҰҚЫҚТАРДЫҢ МЕМЛЕКЕТТІК ТІЗІЛІМГЕ МӘЛІМЕТТЕРДІ ЕНГІЗУ ТУРАЛЫ

КУӘЛІК

2020 жылғы «29» қаңтар № 7844

Автордың (авторлардың) аты, аты, әкесінің аты (егер ол жеке басын куәландыратын құжатта көрсетілсе):
МАМЫРБАЕВ ОРКЕН ЖУМАЖАНОВИЧ, ТУРГАЛЫҰЛЫ МУСА , ТУРГАЛЫҚЫЗЫ ТОЛҒАНАЙ , КЫДЫРБЕКОВА АЙЗАТ СЕЙЗБЕКОВНА, МЕКЕБАЕВ НУРБАЛА ОТАНОВИЧ, СЕЙТҚАЛИ БЕКЖАН НҰРЛАНҰЛЫ, АХМЕТОВ БЕКТҰРҒАН СҰЛТАНҒАЛИЕВИЧ

Авторлық құқық объектісі: ЭЕМ-ге арналған бағдарлама

Объектінің атауы: Мультиязычное распознавание речи MultiSpeech

Объектіні жасаған күні: 09.01.2020

Подписано ЭЦП

Батаева К.О.

ҚОСЫМША Ә

Dimentionality reduction

```
from sklearn.manifold import TSNE

def flatten(l): return flatten(l[0]) + (flatten(l[1:]) if len(l) > 1 else []) if type(l) is list else [l]

def reduce_dim_into_two_TSNE(freq_list):
    #size = len(freq_list)
    #arr = np.empty((size,300), dtype='f')
        # add the vector for each of the words to the array
    #arr = np.array(freq_list)
    # find tsne coords for 2 dimensions
    tsne = TSNE(n_components=2, random_state=0)
    np.set_printoptions(suppress=True)
    Y = tsne.fit_transform(freq_list)
    x_coords = Y[:, 0]
    y_coords = Y[:, 1]
        return x_coords, y_coords
speakers audio
#splits: do it once.

spxtrain, spxtest, spytrain, spytest = train_test_split(sp_audio_labels, speakers_ids, test_size=0.2, shuffle= True)

#save to csv

import pandas as pd

df_spxtrain = []

df_spytrain = []

df_spxtest = []

df_spytest = []

for i, x in enumerate(spxtrain):
    df_spxtrain.append(x[0])
    df_spytrain.append(spytrain[i])
```

```
for j, x in enumerate(spxtest):
    df_spxtest.append(x[0])
    df_spytest.append(spytest[j])
    train_speakers_data = {"file":df_spxtrain,"speakerID":df_spytrain}
test_speakers_data = {"file":df_spxtest,"speakerID":df_spytest}
train_df = pd.DataFrame(train_speakers_data)
test_df = pd.DataFrame(test_speakers_data)
train_df.to_csv("train_speakers.csv")
test_df.to_csv("test_speakers.csv")
#read from csv
import pandas as pd
train_df = pd.read_csv("train_speakers.csv")
test_df = pd.read_csv("test_speakers.csv")
spxtrain = []
spytrain = []
spxtest = []
spytest = []
for test_file, spid in zip(test_df['file'],test_df['speakerID']):
    spxtest.append(test_file)
    spytest.append(spid)
for file, spid in zip(train_df['file'],train_df['speakerID']):
    spxtrain.append(file)
    spytrain.append(spid)
statistics
def show_statistics(audio_labels):
    print("number of audios",len(audio_labels))
    n_male = 0
    n_female = 0
    for f in audio_labels:
        if f[1] == 1:
```

```
            n_female += 1
        elif f[1] == 0:
            n_male += 1
    print("number of female",n_female)
    print("number of male",n_male)
print("------train-------")
show_statistics(tr_audio_labels)
print("speakers ", len(speakers_id))
print("------test-------")
show_statistics(te_audio_labels)
print("speakers ", len(speakers_id))
```

MFCC extraction

```
import warnings;
warnings.filterwarnings('ignore');
import os
import pickle
import numpy as np
from scipy.io.wavfile import read
import python_speech_features as mfcc
from sklearn import preprocessing
def get_MFCC(sr, audio):
    features = mfcc.mfcc(audio,sr, 0.025, 0.01, 13,appendEnergy = False)
    features = preprocessing.scale(features)
    return features
def feat_extraction(audio_labels):
    # Read data
    audios_mfcc = []
    labels = []
    maxLength = 0
    for f in audio_labels:
```

```
        sr,audio = read(f[0])
        vector = get_MFCC(sr,audio)
        if len(vector) > maxLength:
            maxLength = len(vector)
        audios_mfcc.append(vector[:,0])
        labels.append(f[1])
    #padding
    pad_audios_mfcc = []
    for tr_data in audios_mfcc:
        tr_data = np.pad(tr_data, (0, 3887-len(tr_data)), 'constant', constant_values=(0))
        pad_audios_mfcc.append(tr_data)
    print("end")
    return pad_audios_mfcc,labels
def feat_extraction_speakers(audio_labels):
    # Read data
    audios_mfcc = []
    labels = []
    maxLength = 0
    for f in audio_labels:
        sr,audio = read(f)
        vector = get_MFCC(sr,audio)
        if len(vector) > maxLength:
            maxLength = len(vector)
        audios_mfcc.append(vector[:,0])
        labels.append(f[1])

    #padding
    pad_audios_mfcc = []
    for tr_data in audios_mfcc:
        tr_data = np.pad(tr_data, (0, 3887-len(tr_data)), 'constant', constant_values=(0))
```

```
        pad_audios_mfcc.append(tr_data)
    print("end")
    return pad_audios_mfcc,labels
# #MLP, gender  array (3887,1)
# tr_mfcc, tr_labels = feat_extraction(tr_audio_labels)
# te_mfcc, te_labels = feat_extraction(te_audio_labels)
# #CNN, gender matrix 61x61
# tr_mfcc_m = []
# te_mfcc_m = []
# for mf in tr_mfcc:
#     tr_mfcc_m.append(mf.reshape(299,13))
# for mf in te_mfcc:
#     te_mfcc_m.append(mf.reshape(299,13))
##Speakers
tr_mfcc, tr_labels = feat_extraction_speakers(spxtrain)
te_mfcc, te_labels = feat_extraction_speakers(spxtest)
#CNN, gender matrix 61x61
tr_mfcc_m = []
te_mfcc_m = []
for mf in tr_mfcc:
    tr_mfcc_m.append(mf.reshape(299,13))
for mf in te_mfcc:
    te_mfcc_m.append(mf.reshape(299,13))
```

Input

```
from keras.models import Sequential
from keras.utils import np_utils
from keras.layers.core import Dense, Activation, Dropout
from keras.utils import to_categorical
import pandas as pd
import numpy as np
```

```
from keras.models import Model
from sklearn.model_selection import train_test_split
import warnings
warnings.filterwarnings('ignore')
def normalize_input(input_feat, labels):
    X = input_feat
    #this function turn the vector to 10000,01000,00100 format
    Y = to_categorical(labels)#speakers , labels
    # pre-processing: divide by max and substract mean
    scale = np.max(X)
    X /= scale
    mean = np.std(X)
    X -= mean
       input_dim = X.shape[1]
    nb_classes = Y.shape[1]
    return input_dim, nb_classes,X,Y
#gender indentification
#tr_labels, te_labels
# input_dim, num_classes, xtrain, ytrain = normalize_input(tr_mfcc, tr_labels)
# input_dim, num_classes, xtest, ytest = normalize_input(te_mfcc, te_labels)
# #gender data for CNN
# _, num_classes, xtrain, ytrain = normalize_input(tr_mfcc_m, tr_labels)
# _, num_classes, xtest, ytest = normalize_input(te_mfcc_m, te_labels)
# #speaker data for MLP
# input_dim, num_classes, xtrain, ytrain = normalize_input(tr_mfcc, spytrain)
# input_dim, num_classes, xtest, ytest = normalize_input(te_mfcc, spytest)
# #speaker data for CNN
_, num_classes, xtrain, ytrain = normalize_input(tr_mfcc_m, spytrain)
_, num_classes, xtest, ytest = normalize_input(te_mfcc_m, spytest)
import collections
```

```
train_speaker = collections.Counter()
test_speaker = collections.Counter()
for x in spytrain:
    train_speaker[x] += 1
    for x in spytest:
    test_speaker[x] += 1
sorted(test_speaker.items())
[(0, 12),
 (1, 12),
 (2, 19),
 (3, 18),
 (4, 17),
 (5, 15),
 (6, 12),
 (7, 9),
 (8, 16),
 (9, 16),
 (10, 12),
 (11, 14),
 (12, 10),
 (13, 14),
 (14, 15),
 (15, 18),
 (16, 21),
 (17, 16),
 (18, 19)]
sorted(train_speaker.items())
[(0, 63),
 (1, 63),
 (2, 56),
 (3, 57),
 (4, 58),
 (5, 60),
 (6, 63),
 (7, 66),
 (8, 59),
 (9, 59),
 (10, 63),
 (11, 61),
 (12, 65),
```

(13, 61),
(14, 60),
(15, 57),
(16, 54),
(17, 59),
(18, 56)]

MLP

```
%matplotlib notebook
plt.figure(figsize=(14, 5))
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from keras.models import Model
from keras.layers import Input, Dense, Dropout, Flatten
from sklearn.metrics import accuracy_score
from keras.callbacks import EarlyStopping
EPOCH = 100000
BATCHSIZE = 50
TASK_TYPE = "speaker" #gender, speaker
MODEL_SIZE = "small" #large, small
# Model
model = Sequential()
model.add(Dense(128, input_dim=input_dim))
model.add(Activation('relu'))
model.add(Dropout(0.15))
model.add(Dense(64))
model.add(Activation('relu'))
model.add(Dropout(0.15))
model.add(Dense(32,name='layer2'))
model.add(Activation('relu'))
model.add(Dropout(0.15))
model.add(Dense(num_classes))
```

```
model.add(Activation('softmax'))
# we'll use categorical xent for the loss, and RMSprop as the optimizer
model.compile(loss='categorical_crossentropy',
optimizer='rmsprop',metrics=['accuracy'])
print("Training...")
es = EarlyStopping(monitor='val_loss', mode='min', verbose=1, patience=1000)
#gender identification
history = model.fit(xtrain, ytrain, nb_epoch=EPOCH, validation_data=(xtest,ytest),
                batch_size=BATCHSIZE, verbose=1, callbacks=[es])
model.save("exp_results/mlp."+TASK_TYPE+"."+MODEL_SIZE+".model")
score = model.evaluate(xtest, ytest, verbose = 0)
print('Test loss:', score[0])
print('Test accuracy:', score[1])

predictions = model.predict_classes(xtest)
# prediction = pd.DataFrame({'predictions':predictions,
'gold':ytest}).to_csv('result_mlp.'+TASK_TYPE+"."+MODEL_SIZE+'.csv')
# Plot training & validation accuracy values
f1 = plt.figure(figsize = (14, 5))
plt.plot(history.epoch,history.history['acc'])
plt.plot(history.epoch,history.history['val_acc'])
plt.title('Model accuracy', fontsize=16)
plt.ylabel('Accuracy', fontsize=16)
plt.xlabel('Epoch', fontsize=16)
plt.legend(['Train', 'Test'], loc='upper left', fontsize=16)
# plt.xticks(history.epoch,fontsize = 12)
# plt.yticks(fontsize = 12)
plt.show()
f1.savefig("exp_results/acc.mlp."+TASK_TYPE+"."+MODEL_SIZE+".pdf",
bbox_inches='tight')
# Plot training & validation loss values
```

```
f2 = plt.figure(figsize = (14, 5))
plt.plot(history.epoch,history.history['loss'])
plt.plot(history.epoch,history.history['val_loss'])
plt.title('Model loss', fontsize=16)
plt.ylabel('Loss', fontsize=16)
plt.xlabel('Epoch', fontsize=16)
plt.legend(['Train', 'Test'], loc='upper left', fontsize=16)
# plt.xticks(history.epoch,fontsize = 12)
# plt.yticks(fontsize = 12)
plt.show()
f2.savefig("exp_results/loss.mlp."+TASK_TYPE+"."+MODEL_SIZE+".pdf", bbox_inches='tight')
len(predictions)
speakerIDs
[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18]
Precision, Recall, Fscore calculation
from sklearn.metrics import precision_recall_fscore_support as score
from sklearn.metrics import accuracy_score

#gender labels=[0,1], te_labels
#speaker labels = speakerIDs, spytest
y_gold = spytest
#speakers,
speakerIDs = []
for spid in speakers_ids:
    if spid not in speakerIDs:
        speakerIDs.append(spid)
precision, recall, fscore, support = score(y_gold, predictions,labels = speakerIDs)
print('precision: {}'.format(precision))
print('recall: {}'.format(recall))
print('fscore: {}'.format(fscore))
```

```
print('support: {}'.format(support))

print("")

precision, recall, fscore, support = score(y_gold, predictions,average="macro")

print('precision: {}'.format(precision))

print('recall: {}'.format(recall))

print('fscore: {}'.format(fscore))

print('support: {}'.format(support))

print('accuracy: {}'.format(accuracy_score(y_gold, predictions)))
```

```
precision: [0.16666667 0.10526316 0.11111111 0.30769231 0.09090909 0.28571429
 0.15       0.33333333 0.21621622 0.2        0.33333333 0.11111111
 0.         0.15789474 0.         0.18181818 0.125      0.2
 0.07692308]
recall: [0.16666667 0.16666667 0.05263158 0.22222222 0.05882353 0.13333333
 0.25       0.44444444 0.5        0.125      0.33333333 0.07142857
 0.         0.42857143 0.         0.11111111 0.04761905 0.25
 0.05263158]
fscore: [0.16666667 0.12903226 0.07142857 0.25806452 0.07142857 0.18181818
 0.1875     0.38095238 0.30188679 0.15384615 0.33333333 0.08695652
 0.         0.23076923 0.         0.13793103 0.06896552 0.22222222
 0.0625    ]
support: [12 12 19 18 17 15 12  9 16 16 12 14 10 14 15 18 21 16 19]

precision: 0.1659466636613451
recall: 0.179709658563333
fscore: 0.16027905013552418
support: None
accuracy: 0.16842105263157894
```

Load MLP Model and Test

```
from keras.models import load_model

TASK_TYPE = "speaker" #gender, speaker

MODEL_SIZE = "small" #large, small

# load model

model = load_model('exp_results/mlp.'+TASK_TYPE+'.'+MODEL_SIZE+'.model')

# summarize model.

model.summary()

score = model.evaluate(xtest, ytest, verbose = 0)
```

```
print('Test loss:', score[0])

print('Test accuracy:', score[1])

predictions = model.predict_classes(xtest)

prediction = pd.DataFrame({'predictions':predictions,
'gold':spytest}).to_csv('exp_results/result_mlp.'+TASK_TYPE+"."+MODEL_SIZE+'
.csv')
```

```
Layer (type)                 Output Shape              Param #
=================================================================
dense_51 (Dense)             (None, 128)               497664
_________________________________________________________________
activation_68 (Activation)   (None, 128)               0
_________________________________________________________________
dropout_52 (Dropout)         (None, 128)               0
_________________________________________________________________
dense_52 (Dense)             (None, 64)                8256
_________________________________________________________________
activation_69 (Activation)   (None, 64)                0
_________________________________________________________________
dropout_53 (Dropout)         (None, 64)                0
_________________________________________________________________
layer2 (Dense)               (None, 32)                2080
_________________________________________________________________
activation_70 (Activation)   (None, 32)                0
_________________________________________________________________
dropout_54 (Dropout)         (None, 32)                0
_________________________________________________________________
dense_53 (Dense)             (None, 19)                627
_________________________________________________________________
activation_71 (Activation)   (None, 19)                0
=================================================================
Total params: 508,627
Trainable params: 508,627
Non-trainable params: 0
_________________________________________________________________
Test loss: 10.876866323906079
Test accuracy: 0.1298245612989392
```

```
output of intermediate layer
layer_name = 'layer2'

intermediate_layer_model = Model(inputs=model.input,
```

```
                                 outputs=model.get_layer(layer_name).output)
intermediate_output = intermediate_layer_model.predict(xtrain)
print(len(x_test),len(y_test))
```

285 285

CNN

```
#gender identification
%matplotlib notebook
import matplotlib.pyplot as plt
from keras.layers import Conv2D, MaxPooling2D, BatchNormalization
from keras.models import Model
from keras.layers import Input, Dense, Dropout, Flatten
from sklearn.metrics import accuracy_score
from keras.callbacks import EarlyStopping
EPOCH = 2000
BATCHSIZE = 30
TASK_TYPE = "speaker" #gender, speaker
MODEL_SIZE = "large" #large, small
y_train = ytrain
x_train = xtrain
y_test = ytest
x_test = xtest
# input reshape dimensions
au_rows, au_cols = x_train[0].shape[0], x_train[0].shape[1]
x_train = x_train.reshape(x_train.shape[0], au_rows, au_cols, 1)
x_test = x_test.reshape(x_test.shape[0], au_rows, au_cols, 1)
input_shape = (au_rows, au_cols, 1)
print(input_shape)
x_train = x_train.astype('float32')
x_test = x_test.astype('float32')
```

```
print('x_train shape:', x_train.shape)
print(x_train.shape[0], 'train samples')
print(x_test.shape[0], 'test samples')
#model
model = Sequential()
model.add(Conv2D(512, kernel_size = (3, 3),
                 activation = 'relu',
                 input_shape = input_shape))
model.add(Conv2D(256, (3, 3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Dropout(0.15))
model.add(Flatten())
model.add(Dense(128, activation='relu', name='cnn_last_layer'))
model.add(Dropout(0.15))
model.add(Dense(num_classes, activation='softmax'))
model.compile(loss='categorical_crossentropy',
optimizer='rmsprop',metrics=['accuracy'])
es = EarlyStopping(monitor='val_loss', mode='min', verbose=1, patience=50)
#gender Identification
history = model.fit(x_train, y_train,
          batch_size = BATCHSIZE,
          epochs = EPOCH,
          verbose = 1,
          validation_data=(x_test, y_test),callbacks=[es])
score = model.evaluate(x_test, y_test, verbose = 0)
print('Test loss:', score[0])
print('Test accuracy:', score[1])
model.save("exp_results/CNN."+TASK_TYPE+"."+MODEL_SIZE+".model")
score = model.evaluate(x_test, y_test, verbose = 0)
print('Test loss:', score[0])
```

```
print('Test accuracy:', score[1])

predictions = model.predict_classes(x_test)

# prediction = pd.DataFrame({'predictions':predictions,
'gold':te_labels}).to_csv('exp_results/result_CNN.'+TASK_TYPE+"."+MODEL_SIZ
E+'.csv')

# Plot training & validation accuracy values

f1 = plt.figure(figsize=(14, 5))

plt.plot(history.epoch,history.history['acc'])

plt.plot(history.epoch,history.history['val_acc'])

plt.title('Model accuracy', fontsize=16)

plt.ylabel('Accuracy', fontsize=16)

plt.xlabel('Epoch', fontsize=16)

plt.legend(['Train', 'Test'], loc='upper left', fontsize=16)

# plt.xticks(history.epoch,fontsize = 12)

# plt.yticks(fontsize = 12)

plt.show()

f1.savefig("exp_results/acc.CNN."+TASK_TYPE+"."+MODEL_SIZE+".pdf",
bbox_inches='tight')

# Plot training & validation loss values

f2 = plt.figure(figsize=(14, 5))

plt.plot(history.epoch,history.history['loss'])

plt.plot(history.epoch,history.history['val_loss'])

plt.title('Model loss', fontsize=16)

plt.ylabel('Loss', fontsize=16)

plt.xlabel('Epoch', fontsize=16)

plt.legend(['Train', 'Test'], loc='upper left', fontsize=16)

# plt.xticks(history.epoch,fontsize = 12)

# plt.yticks(fontsize = 12)

plt.show()

f2.savefig("exp_results/loss.CNN."+TASK_TYPE+"."+MODEL_SIZE+".pdf",
bbox_inches='tight')
```

(299, 13, 1)
x_train shape: (1140, 299, 13, 1)
1140 train samples
285 test samples
Train on 1140 samples, validate on 285 samples
Epoch 1/2000
1140/1140 [==============================] - 71s 63ms/step - loss: 3.3506 - acc: 0.0693 - val_loss: 2.7204 - val_acc: 0.1298
Epoch 2/2000
1140/1140 [==============================] - 70s 62ms/step - loss: 2.6510 - acc: 0.1579 - val_loss: 2.6335 - val_acc: 0.1719
Epoch 3/2000
1140/1140 [==============================] - 70s 61ms/step - loss: 2.3544 - acc: 0.2377 - val_loss: 2.5244 - val_acc: 0.1825
Epoch 4/2000
1140/1140 [==============================] - 70s 61ms/step - loss: 1.9986 - acc: 0.3737 - val_loss: 2.5852 - val_acc: 0.1719
Epoch 5/2000
1140/1140 [==============================] - 70s 62ms/step - loss: 1.5114 - acc: 0.5140 - val_loss: 2.6832 - val_acc: 0.1614
Epoch 6/2000
1140/1140 [==============================] - 69s 61ms/step - loss: 1.0899 - acc: 0.6518 - val_loss: 2.8277 - val_acc: 0.1860
Epoch 7/2000
1140/1140 [==============================] - 70s 62ms/step - loss: 0.7133 - acc: 0.8035 - val_loss: 3.1454 - val_acc: 0.1719
Epoch 8/2000
1140/1140 [==============================] - 70s 61ms/step - loss: 0.4446 - acc: 0.8711 - val_loss: 3.8037 - val_acc: 0.1684
Epoch 9/2000
1140/1140 [==============================] - 70s 61ms/step - loss: 0.2803 - acc: 0.9132 - val_loss: 3.9153 - val_acc: 0.1614
Epoch 10/2000
1140/1140 [==============================] - 70s 61ms/step - loss: 0.1686 - acc: 0.9491 - val_loss: 4.8675 - val_acc: 0.1649
Epoch 11/2000
1140/1140 [==============================] - 70s 61ms/step - loss: 0.1543 - acc: 0.9482 - val_loss: 4.7976 - val_acc: 0.2070
Epoch 12/2000
1140/1140 [==============================] - 69s 61ms/step - loss: 0.0885 - acc: 0.9702 - val_loss: 5.0964 - val_acc: 0.1754
Epoch 13/2000
1140/1140 [==============================] - 69s 61ms/step - loss: 0.0825 - acc: 0.9746 - val_loss: 5.9523 - val_acc: 0.1439

Epoch 14/2000
1140/1140 [==============================] - 70s 61ms/step - loss: 0.0739 - acc: 0.9807 - val_loss: 6.0497 - val_acc: 0.1579
Epoch 15/2000
1140/1140 [==============================] - 69s 61ms/step - loss: 0.0585 - acc: 0.9816 - val_loss: 5.9576 - val_acc: 0.1614
Epoch 16/2000
1140/1140 [==============================] - 69s 61ms/step - loss: 0.0580 - acc: 0.9789 - val_loss: 5.8899 - val_acc: 0.1719
Epoch 17/2000
1140/1140 [==============================] - 69s 61ms/step - loss: 0.0382 - acc: 0.9868 - val_loss: 6.6115 - val_acc: 0.1860
Epoch 18/2000
1140/1140 [==============================] - 69s 61ms/step - loss: 0.0632 - acc: 0.9816 - val_loss: 6.1205 - val_acc: 0.1649
Epoch 19/2000
1140/1140 [==============================] - 70s 61ms/step - loss: 0.0371 - acc: 0.9886 - val_loss: 6.3892 - val_acc: 0.1684
Epoch 20/2000
1140/1140 [==============================] - 69s 61ms/step - loss: 0.0383 - acc: 0.9877 - val_loss: 6.7099 - val_acc: 0.1474
Epoch 21/2000
1140/1140 [==============================] - 69s 61ms/step - loss: 0.0147 - acc: 0.9965 - val_loss: 7.6075 - val_acc: 0.1544
Epoch 22/2000
1140/1140 [==============================] - 70s 61ms/step - loss: 0.0427 - acc: 0.9868 - val_loss: 7.4431 - val_acc: 0.1368
Epoch 23/2000
1140/1140 [==============================] - 69s 61ms/step - loss: 0.0271 - acc: 0.9912 - val_loss: 7.2977 - val_acc: 0.1404
Epoch 24/2000
1140/1140 [==============================] - 69s 61ms/step - loss: 0.0484 - acc: 0.9816 - val_loss: 6.7334 - val_acc: 0.1333
Epoch 25/2000
1140/1140 [==============================] - 70s 61ms/step - loss: 0.0116 - acc: 0.9965 - val_loss: 7.0421 - val_acc: 0.1754
Epoch 26/2000
1140/1140 [==============================] - 70s 61ms/step - loss: 0.0228 - acc: 0.9921 - val_loss: 7.5810 - val_acc: 0.1825
Epoch 27/2000
1140/1140 [==============================] - 70s 61ms/step - loss: 0.0341 - acc: 0.9904 - val_loss: 7.1730 - val_acc: 0.1509
Epoch 28/2000

```
1140/1140 [==============================] - 69s 61ms/step - loss: 0.0156
- acc: 0.9930 - val_loss: 7.7007 - val_acc: 0.1439
Epoch 29/2000
1140/1140 [==============================] - 70s 61ms/step - loss: 0.0261
- acc: 0.9912 - val_loss: 7.8211 - val_acc: 0.1509
Epoch 30/2000
1140/1140 [==============================] - 69s 60ms/step - loss: 0.0170
- acc: 0.9912 - val_loss: 7.5239 - val_acc: 0.1649
Epoch 31/2000
1140/1140 [==============================] - 69s 61ms/step - loss: 0.0220
- acc: 0.9930 - val_loss: 7.5171 - val_acc: 0.1789
Epoch 32/2000
1140/1140 [==============================] - 69s 61ms/step - loss: 0.0485
- acc: 0.9886 - val_loss: 7.7829 - val_acc: 0.1649
Epoch 33/2000
1140/1140 [==============================] - 69s 61ms/step - loss: 0.0199
- acc: 0.9956 - val_loss: 8.1941 - val_acc: 0.1579
Epoch 34/2000
1140/1140 [==============================] - 69s 60ms/step - loss: 0.0213
- acc: 0.9930 - val_loss: 7.9513 - val_acc: 0.1684
Epoch 35/2000
1140/1140 [==============================] - 69s 60ms/step - loss: 0.0303
- acc: 0.9939 - val_loss: 7.7482 - val_acc: 0.1579
Epoch 36/2000
1140/1140 [==============================] - 69s 61ms/step - loss: 0.0142
- acc: 0.9965 - val_loss: 8.0250 - val_acc: 0.1649
Epoch 37/2000
1140/1140 [==============================] - 69s 60ms/step - loss: 0.0127
- acc: 0.9956 - val_loss: 7.9286 - val_acc: 0.1684
Epoch 38/2000
 930/1140 [=======================>......] - ETA: 12s - loss: 0.0276 - acc:
0.9914
```

```
Load CNN model and Test
from keras.models import load_model(
TASK_TYPE = "speaker" #gender, speaker
MODEL_SIZE = "large" #large, small
# load model
model = load_model('exp_results/CNN.'+TASK_TYPE+'.'+MODEL_SIZE+'.model')
# summarize model.
model.summary()
score = model.evaluate(x_test, y_test, verbose = 0)
print('Test loss:', score[0])
print('Test accuracy:', score[1])
predictions = model.predict_classes(x_test)
```

```
prediction = pd.DataFrame({'predictions':predictions,
'gold':spytest}).to_csv('exp_results/result_CNN.'+TASK_TYPE+"."+MODEL_SIZE
+'.csv')
ayer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)           (None, 297, 11, 512)      5120
_________________________________________________________________
conv2d_2 (Conv2D)           (None, 295, 9, 256)       1179904
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 147, 4, 256)       0
_________________________________________________________________
dropout_1 (Dropout)         (None, 147, 4, 256)       0
_________________________________________________________________
flatten_1 (Flatten)         (None, 150528)            0
_________________________________________________________________
cnn_last_layer (Dense)      (None, 128)               19267712
_________________________________________________________________
dropout_2 (Dropout)         (None, 128)               0
_________________________________________________________________
dense_1 (Dense)             (None, 19)                2451
=================================================================
Total params: 20,455,187
Trainable params: 20,455,187
Non-trainable params: 0
_________________________________________________________________
Test loss: 9.222553898995383
Test accuracy: 0.16842105331128104
layer_name = 'cnn_last_layer'
intermediate_layer_model = Model(inputs=model.input,
                                 outputs=model.get_layer(layer_name).output)
intermediate_output = intermediate_layer_model.predict(x_train)
```

gender visual.

```
gender_x = x_coords
gender_y = y_coords
%matplotlib notebook

# gernder
# print("Plotting...")
# for i, l in enumerate(set(labels)):
#     if l == 1:
#         plt.scatter(gender_x[i], gender_y[i], c="red", marker="o", s=20,
cmap='viridis')
```

```
#     else:
#         plt.scatter(gender_x[i], gender_y[i], c="blue", marker=">", s=20,
cmap='viridis')
# plt.legend(loc="best")
# plt.show()

fig, ax = plt.subplots()
lb = np.asarray(tr_labels)
for gi in np.unique(tr_labels):
    i = np.where(lb == int(gi))
    if gi == 0:
        l = "male"
    else:
        l = "female"
    ax.scatter(gender_x[i], gender_y[i],s=10, label=l)
ax.legend(loc="best")
plt.show()
fig.savefig("exp_results/cnn.gender.small.plot.train.pdf", bbox_inches='tight')
```

speaker visual.

```
%matplotlib notebook
speaker_x = []
speaker_y = []

female_x = []
female_y = []
female_speaker = []
female_speaker_label = []
male_x = []
male_y = []
male_speaker = []
male_speaker_label = []
# speaker
for i, l in enumerate(spytrain):
#     if l == 1:
#         female_x.append(x_coords[i])
#         female_y.append(y_coords[i])
#         female_speaker.append(speakers[i])
#         female_speaker_label.append("speaker"+str(speakers[i]))
#         #plt.scatter(x_coords[i], y_coords[i], c=speakers[i], marker="o", s=20,
cmap='viridis')
#     else:
#         male_x.append(x_coords[i])
#         male_y.append(y_coords[i])
#         male_speaker.append(speakers[i])
```

```
#         male_speaker_label.append("speaker"+str(speakers[i]))
        #plt.scatter(x_coords[i], y_coords[i], c=speakers[i], marker=">", s=20,
cmap='viridis')
    speaker_x.append(x_coords[i])
    speaker_y.append(y_coords[i])
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt

#plt.scatter(speaker_x, speaker_y, c=speakers, s=7)
#plt.scatter(male_x, male_y, c=male_speaker, marker=">", s=7)
#plt.scatter(female_x, female_y, c=female_speaker, marker="o", s=7)

# plt.legend(loc='best')
# plt.show()

colors = ['blue', 'green', 'red',
          'cyan', 'magenta', 'yellow',
          'black', 'purple', 'pink',
          'brown', 'orange', 'teal',
          'coral', 'lightblue', 'lime',
          'lavender', 'turquoise', 'darkgreen',
          'tan']
# 'salmon', 'gold', 'lightpurple', 'darkred', 'darkblue'

# fig, ax = plt.subplots()
# sp_x = np.asarray(speaker_x)
# sp_y = np.asarray(speaker_y)
# sp = np.asarray(speakers)
# for gi in np.unique(sp):
#     i = np.where(sp == gi)
#     ax.scatter(sp_x[i], sp_y[i],s=10, label=gi, c=colors[gi])
# ax.legend(loc="best")
# # Add a colorbar
# plt.show()

fig, ax = plt.subplots()
sp_x = np.asarray(speaker_x)
sp_y = np.asarray(speaker_y)
sp = np.asarray(spytrain)
for gi in np.unique(sp):
    i = np.where(sp == gi)
    for z in i[0]:
        print(z)
```

```
#         if labels[z] == 1:
#             im = ax.scatter(sp_x[z], sp_y[z], marker="o" ,s=5, c=colors[gi])
#         else:
#             im = ax.scatter(sp_x[z], sp_y[z], marker=">" ,s=5, c=colors[gi])
        im = ax.scatter(sp_x[z], sp_y[z], marker="o" ,s=2, c=colors[gi])
    im.set_label(gi)
ax.legend(loc='center left', bbox_to_anchor=(1, 0.5))
plt.show()
fig.savefig("exp_results/mlp.speaker.small.plot.train.pdf", bbox_inches='tight')
```

example

```
%matplotlib notebook
import matplotlib.pyplot as plt
import numpy as  np
fig, ax = plt.subplots()
scatter_x = np.array([1,2,3,4,5])
scatter_y = np.array([5,4,3,2,1])
group = np.array([1,3,2,1,3])
for g in np.unique(group):
    j = np.where(group == g)
    ax.scatter(scatter_x[j], scatter_y[j], label=g)
ax.legend()
plt.show()
```