COMPUTER SCIENCE
КОМПЬЮТЕРЛІК ҒЫЛЫМДАР
КОМПЬЮТЕРНЫЕ НАУКИ

ҒТАМР 28.23.37
ӘОЖ 004.934

https://doi.org/10.55452/1998-6688-2025-22-2-10-23

[1]*Ахмедиярова А.Т.,
PhD, қауымдастырылған профессор, ORCID ID: 0000-0003-4439-7313,
*e-mail: a.akhmediyarova@satbayev.university
[1]Алибиева Ж.М.,
PhD, қауымдастырылған профессор, ORCID ID: 0000-0001-9565-5621,
e-mail: zh.alibiyeva@satbayev.university
[1]Мукажанов Н.К.,
PhD, қауымдастырылған профессор, ORCID ID: 0000-0003-4835-5751,
e-mail: n.mukazhanov@satbayev.university

[1]Сәтбаев университеті, Алматы қ., Қазақстан

ДИКТОРДЫ СӘЙКЕСТЕНДІРУ КЕЗІНДЕ СӨЙЛЕУДІ СЕГМЕНТАЦИЯЛАУ

Аңдатпа

Сөйлеуді сегментациялау – сөйлеу сигналдарын бөлшектерге бөлу процесі, ол диктордыны сәйкестендіру және сөйлеуді тану жүйелерінің маңызды аспектісі. Бұл процесс сөйлеудің басталуын және аяқталуын дәл анықтауға мүмкіндік беріп, жүйенің тиімділігін арттырады. Сегментациялау кезінде дауыстық белсенділік детекторларын (VAD) пайдалану маңызды рөл атқарады, өйткені олар сөйлеу мен тыныштық арасындағы шекараларды анықтауға көмектеседі. Алайда, сегментациялау барысында жиі кездесетін қателіктер – жалған оң және жалған теріс нәтижелер, олар жүйенің жалпы дәлдігіне теріс әсер етеді. Осыған байланысты, әртүрлі тәсілдер мен әдістер арқылы қателіктерді азайту қажет. Фондық шуды азайту, терең оқыту модельдерін қолдану, сондай-ақ деректерді аугментациялау секілді шаралар сегментациялау сапасын едәуір жақсарта алады. Спектралды талдау әдістері мен ерекшеліктерін пайдалану сөйлеу мен фондық шу арасындағы айырмашылықты айқын ажыратуға мүмкіндік береді. Бұл зерттеудің мақсаты – сегментациялау процесін оңтайландыру және қателіктердің ықтималдықтарын талдау, сөйлеуді тану жүйелерінің тиімділігін арттыру. Нәтижесінде, бұл жұмыс сөйлеуді тану саласындағы жаңа зерттеулер мен әзірлемелер үшін негіз болады. Мақалада диктордыны анықтау үшін ауызша сөйлеуін сегментациялау мәселесі қарастырылған. Жұмыста сегменттеудің мүмкін критерийлері сипатталған – дыбыстық сөйлеудің сапалық және сандық сипаттамалары, мысалы, сөйлеу кідірістері мен интонациясы, сондай-ақ олардың акустикалық корреляциясы. Бұл сарапшыға нақты сегменттік бірліктерді (буындар, сөздер және т.б.) анықтауға, олардың құрылымын жазуға, негізгі белгілерді бөліп көрсетуге мүмкіндік береді.

Тірек сөздер: сөйлеу сигналы, сегментациялау, дауыс пен сөйлеу, сөйлеу сигналының тірек фрагменттері, сәйкестендіру әдісі.

Кіріспе

Сөйлеушіні тану – сөйлеушіні сөз тіркесі арқылы тану үшін қолданылатын процесс. Бұл аудио немесе бейне құжаттарды іздеу сияқты кең қолданылатын пайдалы биометриялық құрал. Сөйлеушіні тануда екі процедура басым, атап айтқанда сегменттеу және жіктеу. Сегменттеу және жіктеу үшін қолданылатын жаңа алгоритмдерді әзірлеу немесе ескілерін жақсарту үшін зерттеулер мен әзірлемелер жалғасуда [1].

Сөйлеушіні анықтау кезінде дәстүрлі субъективті әдістер қолданылады: есту арқылы тану және акустикалық көріністерді салыстыру. Сөйлеушіні тану үшін қолданылатын машиналар сөйлеуді автоматты тану машиналары (ASR) деп аталады. ASR машиналары адамды анықтау үшін немесе адамның жеке басын растау үшін қолданылады. Төменде сөйлеушіні тану саласында ұсынылған әртүрлі жақсартулар туралы талқылау берілген [2].

Автоматтандырылған сәйкестендіру жүйелерін қолдану саласындағы алғашқы зерттеулер шетелде жүргізілді және АҚШ, Германия, Ұлыбританияда қабылданған сөйлеушіні криминалистикалық сәйкестендіру әдістеріне негізделген. Шетелдік әдістерде сөйлеу сигналын зерттеудің аппараттық құралдарына, сондай-ақ тілдің сегменттік және супрасегменттік деңгейлеріне қатысты талдауларына баса назар аударылды.

Сөйлеушіні тану үшін маңызды екі процесс – дыбысты жіктеу және сегментация. Бұл екі процесс компьютерлік алгоритмдер арқылы жүзеге асырылады. Дыбысты жіктеудің тамаша процедурасын жасау кезінде фондық шудың әсерін ескеру қажет. Осы факторға байланысты кейбір ғалымдар тіпті шулы фонда да жоғары өнімділікті көрсететін есту моделін ұсынды: шулы фонда тұрақтылыққа қол жеткізетін модель өзін-өзі қалыпқа келтіру механизміне ие. Есту моделінің қарапайым түрі үш сатылы өңдеу тізбегі ретінде көрінеді, оның барысында дыбыстық сигнал нейрондық иллюстрация ішіндегі модель болатын есту спектріне айналады. Бұл модельді қолданудың кемшіліктеріне сызықтық емес өңдеудің күрделілігі мен жоғары есептеу ресурстарына деген талаптар жатады.

Қазіргі уақытта дикторды тану мәселесін шешу үшін автоматты және сараптамалық әдістер, соның ішінде жартылай автоматты әдістер кеңінен қолданылады. Сөйлеушіні анықтау немесе тексеру мақсатында фоноскопиялық зерттеулер жүргізу процесінде сараптамалық әдістерді қолдану сөйлеу сигналдарын талдау мен салыстырудың автоматты құралдарының жұмысын нақтылауға, түзетуге мүмкіндік береді. Алайда, бұл әдістерді қолдану жоғары білікті сарапшыларды тарту қажеттілігімен шектеледі. Сараптамалық әдістерді қолданудың жалпы шешімі негізінен субъективті, өйткені ол сарапшының жеке тәжірибесіне байланысты [3].

Сөйлеуді өңдеудің заманауи әдістерін қарастырайық. Терең оқыту модельдерін, атап айтқанда нейрондық желілер мен трансформерлерді қолдану сөйлеуді тану жүйелерінің дәлдігін арттыруға мүмкіндік береді. Сонымен қатар, фондық шуды азайту әдістерін жетілдіру сөйлеу сигналының сапасын жақсартып, тану үдерісінің тиімділігін арттырады. Спектралды талдау мен акустикалық ерекшеліктерді пайдалану сөйлеуді жоғары дәлдікпен сегментациялауға және оның құрылымын тереңірек талдауға ықпал етеді. Деректерді аугментациялау әдістері модельдердің әртүрлі сөйлеу жағдайларына бейімделу қабілетін күшейтеді. Сондай-ақ, интерактивті және адаптивті сөйлеуді тану жүйелерін дамыту пайдаланушыға ыңғайлы әрі жоғары тиімді шешімдерді ұсынуға мүмкіндік береді [3].

Дауыс пен сөйлеу арқылы дикторларды анықтаудың қазіргі әдістерінің басым көпшілігі аудиторлық-лингвистикалық немесе акустикалық белгілердің таралуын статистикалық талдауға негізделген. Дикторларды сәйкестендіру мәселесі бойынша заманауи әдістемелік және ғылыми әдебиеттерді талдау отандық заманауи сараптамалық тәжірибеде әртүрлі жартылай автоматты акустикалық және аудиторлық-лингвистикалық әдістер қолданылатынын көрсетті [4].

Негізгі ережелер

Сөйлеу сигналының тірек фрагменттерінің әуендік құрылымдарын салыстыру негізінде сәйкестендіру әдісі автоматтандырылған әдістердің бірі және сарапшыға сөйлеушінің сөйлеуіндегі әуендік құрылымдардың негізгі сипаттамаларын талдауға және салыстыруға мүмкіндік береді. Сондықтан дыбыстық мәтіннің құрамдас бөліктерінің фонетикалық құрылымдылығын зерттеу осы мәселені шешу үшін қажетті сегменттеу принциптерін іздеуге байланысты өзектілікке ие болды. Кез-келген дыбыстық сөйлеу хабарламасын зерттеу осы хабарламаны алдын-ала сегменттеуді, негізгі сегменттер мен олардың белгілерін, олардың арасындағы қатынастарды, олардың құрылымдық ұйымын одан әрі тану және түсіну мақсатында бөлуді қамтиды. Сонымен қатар, сөйлеу мәлімдемесін сегменттеу бірліктерін, әдістемесін, оларды анықтау критерийлерін, өзара әрекеттесу ерекшеліктері мен формаларын анықтау мәселесі туындайды [5, 6].

Жазбаша мәтінді сегменттеу әріптердің, бос орындардың және тыныс белгілерінің дискреттілігінің арқасында қиындық тудырмайды. Тағы бір маңызды мәселе – дыбыстарды, буындарды және сөздерді біртұтас ағында біріктіретін дыбыстық сөйлеудің құрылымы, ол тек коммуниканттың алдыңғы тілдік тәжірибесінің негізінде ғана жеке бірліктерге бөлінеді.

Фонация және артикуляция процестері сөйлеу тұжырымының компоненттерге бөлінуін анықтайды: минималды (дыбыс, буын) және максималды (мәтін, дискурс). Буындар мен сөздер кейбір жағдайларда дискретті бірліктер ретінде әрекет етеді, кейде «ұйымдастырылмаған» ағын болады. Дискреттілік белгілі бір тілдің жүйесін білу арқылы жүзеге асырылады, сонымен қатар коммуниканттың лингвистикалық есту қабілетіне, оның сөйлеудің белгілі бір компоненттерін «тану» және барабар оқшаулау қабілетіне байланысты.

Сегменттеу дегеніміз сөйлеу ағынының (мәтіннің) құрамдас сегменттерге сызықтық бөлінуі: сөйлемдер, сөздер, морфемалар немесе елеусіз – силлабемалар, фонемалар [7].

Фонемалар – акустикалық сөйлеу сигналының ұзындығының өзгеруінің жоғары дәрежесі бар лингвистикалық абстракция, сондықтан оларды жеке сегменттерге бөлу қиын. Фонеманың акустикалық көрінісі контекстке, сондай-ақ жасаушыға байланысты өзгереді. Дауысты фонемалар кез-келген акустикалық сөйлеу сигналының бөлігі. Дауысты дыбыстар сөйлеуде жиі кездеседі. Демек, дауысты фонемалар акустикалық ақпарат шу арқылы бұрмаланған жағдайларда сөйлеушіні ажыратын ақпараттың әртүрлі мөлшерін алу үшін пайдаланылуы мүмкін.

Материалдар мен әдістер

Дауысты дыбыстарды сөйлеушіні анықтау үшін негіз ретінде пайдалануды Окленд университетінің сөйлеуді өңдеу тобы бұрыннан бастаған. Содан бері фонеманы тану алгоритмдері және онымен байланысты әдістер сөйлеушіні тану мәселесіне айтарлықтай назар аударды және тіпті лингвистикалық салаға таралды. Дауысты фонеманың рөлі сөйлеушіні тексеру немесе анықтау саласында әлі де ашық мәселе болып табылады. Себебі дауысты фонемаларға негізделген айтылым аймақтық және тілдік әртүрлілікке байланысты өзгереді. Демек, сегменттелген дауысты сөйлеу фрагменттерін сөйлеушінің осы тілде сөйлеуіндегі аймақтық айырмашылықтарды бақылау үшін пайдалануға болады. Бұл көбінесе қазақ тіліне қатысты. Қазақ тілінің диалектілері – белгілі бір аймақтарда таралған қазақ тілінің түрлері. Тілтанушылардың арасында кең таралған пікір бойынша, қазақ тілі батыс, оңтүстік және солтүстік-шығыс деген үш диалектіден тұрады. Осылайша, дауысты сегменттеудің тиімді әдісі сөйлеушіні анықтауда, сондай-ақ сөйлеуді мәтінге айналдыру және дауысты белсендіру жүйесі сияқты қосымшалар үшін тиімді болады [8].

Семантикалық жазықтықтағы әрбір семантикалық бірлік акустикалық жазықтықтағы белгілі бір бірлікке сәйкес келеді. Просодикалық тұрғыдан бұл бірлік оның шекараларымен анықталады және бір немесе бірнеше сөздерге немесе сөз тіркестеріне сәйкес келуі мүмкін.

Мұндай бірліктің акустикалық корреляцияларының бірі – негізгі тон жиілігінің көтерілуі, ол сөйлеушінің осы семантикалық бірлікті бөліп көрсететіндей фразадағы барлық тондардан өзгеше.

Үш просодикалық сипаттаманың ішінен (негізгі тон жиілігі, қарқындылығы және ұзақтығы) сегменттеу кезінде буынның акустикалық корреляциясын анықтаған кезде уақыт сипаттамасы (ұзақтығы) ең негізгі болып табылады. Кез-келген мәлімдеменің болуы шекаралық сигнал ретінде қарастырылатын және мәтінді семантикалық кодтау және декодтау актісімен байланысты үзілістің болуын алдын-ала анықтайды. Мәтіндегі үзілістерді оқшаулау оның коммуникативті бағытымен, сондай-ақ сөйлеушінің мәтінді кодтау қабілетімен, коммуниканттардың физиологиялық жағдайымен байланысты. Сөйлеудегі кідіріс қарқындылық деңгейінің минималды уақыт учаскесінде нөлге дейін төмендеуі ретінде түсіндіріледі, оның ұзақтығы орташа есеппен 10 мс-ге тең [9]. Сонымен қатар, дыбыстағы үзіліс дыбыстың ішінде де болуы интрасегменттік кідіріс деп аталады. Интрасегменттік кідірістер дыбыстарды артикуляциялау процесінде пайда болады, мысалы, дауыссыз дыбыстарды айту кезінде. Интерсегменттік үзілістер өз кезегінде синтаксистік және синтаксистік емес болып бөлінеді. Синтаксистік кідірістер сөйлемнің бөлінуін жүзеге асырады, сонымен қатар сөйлемнің бөліктерін біртұтас тұтастыққа біріктіреді.

Мәлімдеменің кез-келген жерінде кідірістердің болу мүмкіндігін ескере отырып, фразадағы әуезді контурды жүзеге асырудың әртүрлі түрлері туралы айтуға болады. Әуезді контурлар жүйесі, әсіресе кідіріспен үйлескенде, дыбыстық мәтінді толық және әртүрлі етіп бөлуге мүмкіндік береді. Акустикалық сөйлеу сигналының сегменттелу белгілері ретінде негізгі тон жиілігінің болуын/болмауын да ажыратуға болады; дауыссыздан дауыстыға дейінгі негізгі тон жиілігінің секірмелі жоғарылауы; дауыстыдан дауыссызға ауысудағы негізгі тон жиілігінің секірмелі төмендеуі; шудың болуы/болмауы, оның ұзақтығы және өсу қарқыны; төмен жиілікті/жоғары жиілікті энергияның болуы, сондай-ақ сегменттердің ұзақтығы [10, 11].

Осы сипаттамалардың барлығын білу және оларды қазақ тіліндегі әуезді құрылымдарды талдау кезінде практикалық іс-әрекетте қолдану сарапшыға нақты сегменттік бірліктерді (буындар, сөздер және т. б.) дәл анықтауға, олардың құрылымын сипаттауға, негізгі белгілерді бөліп көрсетуге мүмкіндік береді.

Сөйлеу сигналының әртүрлі фрагменттерінің әуезді дизайнын салыстыру мүмкіндігі олардың белгілі бір диктордың сөйлеуіндегі типтілігімен және қайталануымен қамтамасыз етіледі, оған контекстік және басқа дикторлық вариацияның өзіне тән ерекшелігін түзетеді.

Салыстыру үшін фрагменттерді таңдағанда келесі критерийлерді басшылыққа алу керек:

- айқын ерекшелігі бар контур элементтерін және фонограммалардың әрқайсысында іске асырудың бірнеше мысалдарын таңдаңыз. Мысалы, кейбір дикторлардың сөйлеуінде көбінесе төмен немесе жоғары-төмен контуры бар ядролық әуезді модельдер кездеседі;
- бір типті әуезді құрылымдардың іске асырылуын тудыратын сыртқы факторларды шектеңіз – сөйлеу фрагменттері стилистикалық және эмоционалды бояумен мүмкіндігінше сәйкес келуі керек (бұл бүкіл сөйлеу фрагментіне де, оның жеке бөліктеріне де қатысты болуы мүмкін);
- бір типті әуезді құрылымдардың іске асырылу вариациясын тудыратын сегменттік және құрылымдық факторларды шектеу. Атап айтқанда, әуенді талдау және салыстыру кезінде дауыстық учаскенің ұзындығы бойынша, атап айтқанда, ядролық буындардың болуы/болмауы бойынша, ал ядролық учаске болмаған кезде – ядролық буынның сегменттік құрамын ескере отырып, салыстырылатын учаскелерді таңдау керек. Сонымен, қысқа дауыстық бөлімде (мысалы, егер ядролық буын дауыссыз дыбыспен аяқталса) қарапайым, ерте немесе орташа уақыт және тонның біркелкі қозғалысы болуы мүмкін, ал көптеген параметрлердің өзгеруі мүмкін емес. Ядролық буынның тональды бөлігінің ұзақтығымен – мысалы, егер ядролық буынның соңында сонант немесе ұзын дауысты дыбыс болса, онда уақыттың өзгеруі және тондың өзгеру жылдамдығының біркелкі болмауы мүмкін. Ядролық буындар пайда болған кезде іске асыру параметрлерінің ықтимал өзгеру аймағы одан әрі ұлғаюы мүмкін, бұл интонацияның

жеке ерекшеліктерін зерттеу мүмкіндіктерін кеңейтеді. Тональды аймақтың ұзақтығының жоғарылауы/төмендеуі сөйлеу жылдамдығының артуының / баяулауының әсерінен де болуы мүмкін [12].

Сонымен қатар, ядролық буынның сегменттік құрамының өзі тональды конфигурация сипатына және ЧОТ мәндеріне белгілі бір әсер етуі мүмкін. Мәселен, мысалы, жоғары көтерілген дауысты дыбыстардың төмен көтерілген дауыстыларға қарағанда өзіндік жиілігі жоғары екендігі белгілі; дауыссыз дыбыстар келесі дауысты дыбыстың басында ұяшықтардың аздап көтерілуіне, ал дауысты дыбыстар, керісінше, оның төмендеуіне әкеледі. Сондай-ақ, ядролық учаскеден кейінгі синтагматикалық шекараның сипатына назар аудару керек, өйткені егер шекара кідіріспен белгіленбесе, бірақ басқа просодикалық құралдармен жүзеге асырылса, ядролық тонустың соңғы бөлімі келесі синтагманың тональды басталуының әсерінен біршама өзгеруі мүмкін.

Таңдалған контур фрагменті үшін келесі параметрлер жиынтығы есептеледі [13–14]:

- бастапқы жиілік – бірінші санақтың мәні (Гц-те) – таңдалған контур фрагментінің бастапқы нүктесінің бастапқы жиілігі;
- соңғы жиілік – соңғы санақтың мәні (Гц) таңдалған тізбек фрагментінің соңғы нүктесіндегі соңғы жиілік;
- максималды жиілік – жиіліктің максималды мәні максималды жиілік бастап (Гц) таңдалған тізбек фрагменті шегінде;
- максимум уақыты – максималды мәннің координаты бөлінген фрагменттің жалпы ұзақтығының пайызымен максимум уақыты;
- минималды жиілік – таңдалған тізбек фрагменті шегінде (Гц-те) жиіліктің минималды мәні;
- минимум уақыты – минималды мәннің координаты бөлінген фрагменттің жалпы ұзақтығының пайызымен минимум уақыты;
- аралық – максималды және минималды жиілік мәні арасындағы айырмашылық;
- жарты жиілік уақыты – таңдалған фрагменттің жалпы ұзақтығының пайызымен жарты жиілік мәнінің координаты (максимум мен минимум арасындағы интервалдан);
- орташа жиілік – таңдалған контур фрагменті шегінде шоттардың орташа мәні (Гц)
- тонды өзгерту жылдамдығы;
- қиғаштық (асимметрия) орташа мәнге қатысты мәндердің таралуының асимметриялық дәрежесін сипаттайды;
- Эксцесс – қалыпты үлестіріммен салыстырғанда мәндердің таралуының салыстырмалы өткірлігін немесе тегістігін сипаттайды;
- контурдың бұзылу коэффициенті – әуендік контурдың ұзартылған учаскелерінің (синтагмалар, шкалалар, ұзақ сөйлеу учаскелері) бұзылу дәрежесін бағалау үшін қолданылады;
- ұзақтығы–бөлінген фрагменттің ұзақтығы миллисекундтарда.

Осылайша, осы әдіспен фонограммаларды салыстырудың алдында сөйлеу материалын мұқият тыңдау керек, оның нәтижелері бойынша осы нақты жағдай үшін оңтайлы салыстырмалы талдау стратегиясы, яғни салыстыру үшін контур бөлімдерінің түрлері мен құрамы, оларды фонограммада таңдау критерийлері, анықталуы мүмкін. Талдаудың бұл түрінің тиімділігі көбінесе сарапшының құзыретіне, тәжірибесі мен түйсігіне, зерттеуге түскен нақты материалмен жұмыс істеу кезінде «фокустық» аймақтарды дәл анықтау қабілетіне байланысты [15, 16, 17].

Сөйлеуді тану жүйелерінің тиімділігін арттыру мақсатында біздің зерттеуде әртүрлі терең оқыту модельдері қолданылды. Эксперименттер барысында гибридті нейрондық желілер (CNN+LSTM) және трансформерлер (BERT, Wav2Vec 2.0) негізінде екі түрлі архитектура зерттелді. Әрбір модель шулы және таза сөйлеу мәліметтерінде сынақтан өткізілді.

Эксперименттік орта:

- Деректер жиынтығы: Mozilla Common Voice, TED-LIUM3, және арнайы жиналған қазақ тіліндегі деректер;

- Алдын ала өңдеу: VAD алгоритмдері арқылы фондық шудан тазарту;
- Гиперпараметрлер: CNN+LSTM, BERT, Wav2Vec 2.0;
- Бағалау метрикалары: Сөйлеуді тану дәлдігі (WER), сөйлеушіні сәйкестендіру дәлдігі, AUC-ROC, F1-score.

Нәтижелер мен талқылау

Дикторды сәйкестендіру кезінде сөйлеуді сегментациялаудың нақты қадамдарын қарастырайық. Алдымен дауыстық белсенділік детекторына (VAD) тоқталайық. VAD сөйлеу сигналын тыныштықтан немесе шудан ажырату үшін қолданылады. VAD алгоритмдері сигналдың энергиясын немесе сигналдың спектрлік сипаттамаларын негізге алады. Негізгі идея – сигналдың энергия деңгейіне қарап, сөйлеу және тыныштық (немесе шу) арасындағы шектерді анықтау. VAD көмегімен сигнал энергиясы есептеледі:

$$E(n) = \sum_{k=0}^{N-1} |x(n+k)|^2 \quad (1)$$

мұндағы:

– $E(n)$ – n фреймдегі энергия,

– $x(n+k)$ – n фреймдегі k үлгісі,

– N – фрейм ұзындығы.

Егер $E(n)$ мәні алдын-ала белгіленген шектен (T) жоғары болса, онда фрейм сөйлеуді қамтиды, ал төмен болса – тыныштық немесе шу.

Келесі қадамда сөйлеу фрагменттері анықталады. Сөйлеушілердің ауысу нүктелерін анықтау үшін сөйлеу сигналдарының ұқсастығын өлшейтін BIC (Bayesian Information Criterion) сияқты әдістер қолданылады.

BIC формуласы келесі түрде жазылады [18]:

$$BIC(X,Y) = \log\big(L(\theta_X|X)\big) + log\big(L(\theta_Y|Y)\big) - log\big(L(\theta_{XY}|X \cup Y)\big) - \lambda \cdot P \quad (2)$$

мұндағы:

– X және Y – екі түрлі фреймдер жиыны,

– θ_X және θ_Y – сәйкесінше фреймдерге қатысты модель параметрлері,

– $L(\theta_X|X)$ – ықтималдық функциясы (көбінесе Гаусс үлестірімдері қолданылады),

– λ – модельдің күрделілігін анықтайтын тұрақты (гиперпараметр),

– P – параметрлер саны.

BIC-тің оң мәні болса, онда фреймдер арасындағы ауысу нүктесі бар, ал теріс мән болса – жоқ.

Енді дикторды кластерлеуді қарастырайық. Сөйлеу фрагменттері арасындағы ұқсастықты бағалау үшін оларды векторлық кеңістікке проекциялау қажет. Мұнда MFCC (Mel-frequency cepstral coefficients) сияқты ерекшелік векторлары қолданылады.

MFCC есептеу:

1. Сигналды фреймдерге бөлу: сигналды X әрбір фреймге бөлу формуласы:

$$X = \{x_1, x_2, \dots, x_N\} \quad (3)$$

мұндағы x_i – і-ші фрейм.

2. Фурье түрлендіру: әрбір фрейм үшін сигналдың спектрін есептеу:

$$X_k = \sum_{n=0}^{N-1} x(n) e^{-j 2\pi \kappa n / N} \quad (4)$$

3. Мел-фильтрлерді қолдану: спектрді Мел шкаласы бойынша өзгерту.

4. Дискретті косинустық түрлендіру (DCT): Фильтрленген энергияларды косинустық түрлендіру арқылы MFCC коэффициенттерін алу:

$$MFCC(m) = \sum_{n=0}^{N-1} \log\big(S(n)\big) \cos\left(\frac{\pi m(n+0.5)}{N}\right) \quad (5)$$

Келесі қадам – сәйкестендіру (i-vector және x-vector). i-vector әдісінде сөйлеу сигналдарының ерекшеліктерін сипаттау үшін Гаусстық ықтималдық модельдер қолданылады:

$$M = m + T\omega$$

мұндағы:

– M – супер-вектор (сөйлеу сигналының ерекшеліктерін көрсетеді),

– m – орташа супер-вектор (әмбебап фондық модельден (UBM)),

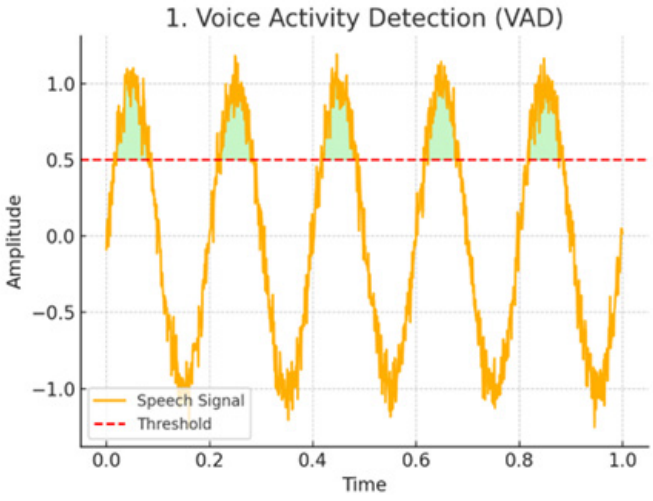
– T – жалпы фактор кеңістігі матрицасы,

– ω – і-вектор (төмен өлшемді ерекшелік векторы).

x-vector – нейрондық желілер негізінде құрылатын әдіс. Бұл әдісте нейрондық желінің аралық қабаттарының шығысын ерекшелік ретінде пайдаланады, нәтижесінде жоғары дәлдіктегі ерекшелік векторлары алынады.

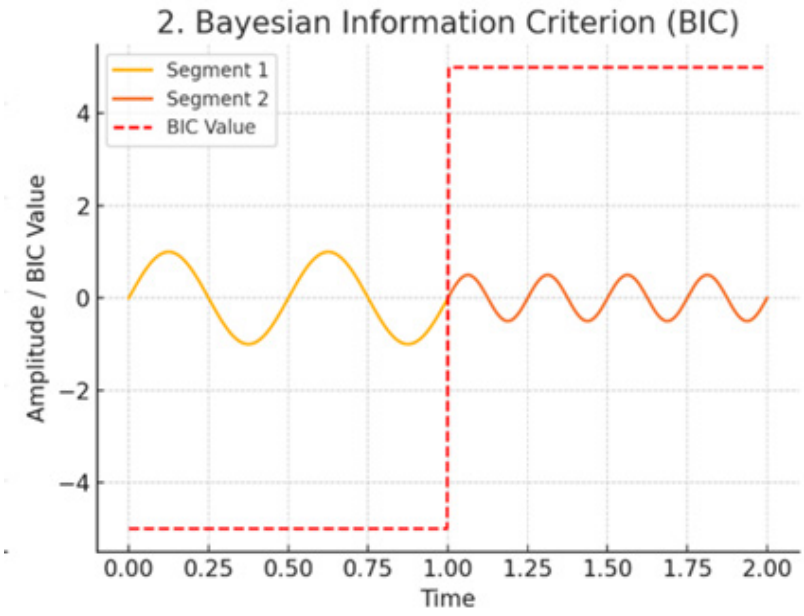
Әрбір жоғарыдағы қадамның сызбасын беру үшін келесі бейнелерді сипаттай аламыз:

1. VAD диаграммасы: Сөйлеу сигналының энергиясы және белгілі бір шек мәні арқылы сөйлеу фрагменттерін анықтау (1-сурет)



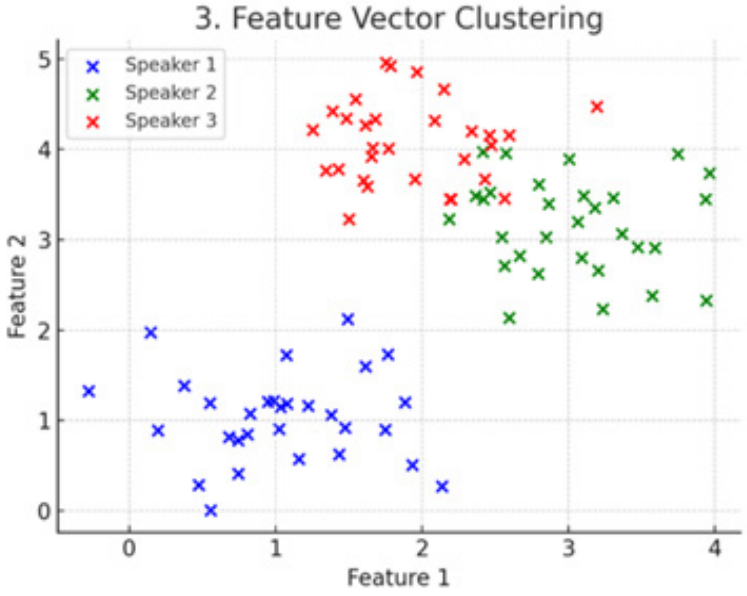
Сурет 1 – VAD диаграммасы

2. BIC диаграммасы: Әртүрлі фреймдердің BIC мәндерін көрсететін график, сөйлеушілердің ауысу нүктелерін айқындайды (2-сурет).



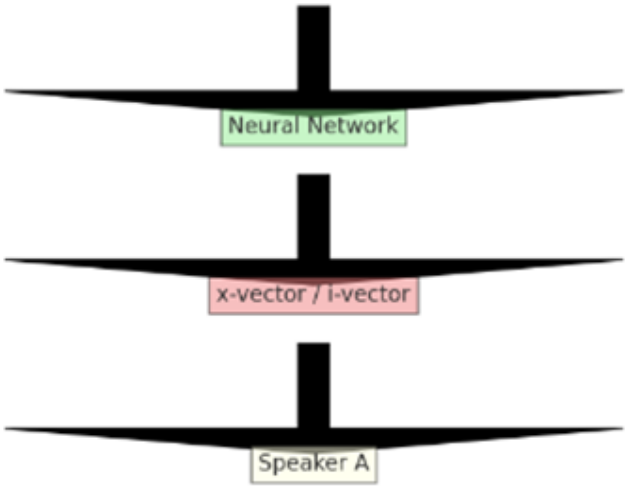
Сурет 2 – BIC диаграммасы

3. Кластерлеу нәтижесі: Сөйлеушілердің ерекшелік векторларын көрсету арқылы олардың қалай кластерленетінін көрсететін диаграмма (3-сурет).



Сурет 3 – Кластерлеу нәтижесі

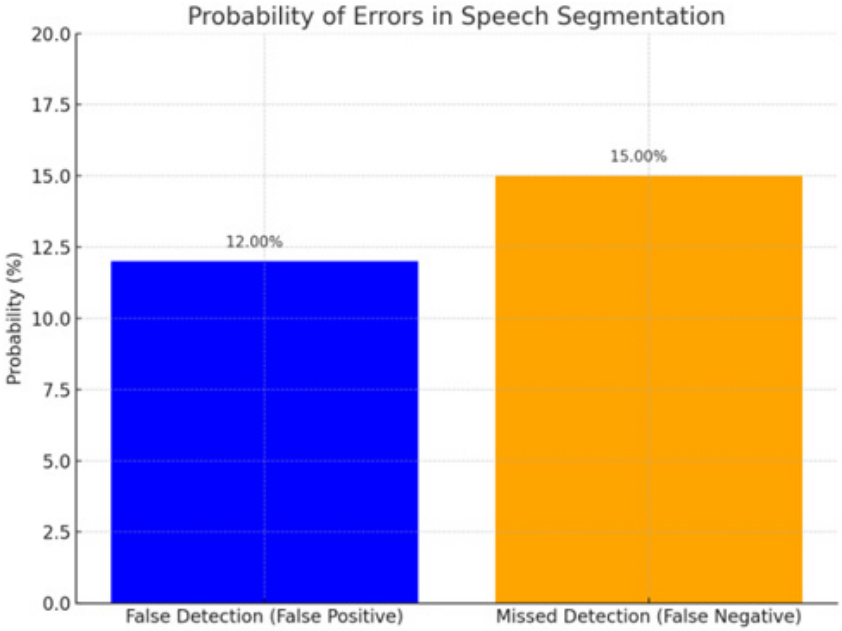
4. Сәйкестендіру схемасы: i-vector немесе x-vector арқылы диктордыны қалай сәйкестендіруге болатынын бейнелейтін сызба (4-сурет).



Сурет 4 – Сәйкестендіру схемасы

Сөйлеуді сегментациялау кезіндегі қателіктердің ықтималдық диаграммасы төменде келтірілген (5-сурет):

- Жалған анықтау ықтималдығы (False Detection): Сөйлеудің басталуын қате анықтау ықтималдығы 12%.
- Басталуды өткізіп алу ықтималдығы (Missed Detection): Сөйлеудің басталуын анықтай алмау ықтималдығы 15%.



Сурет 5 – Сөйлеуді сегментациялау кезіндегі
қателіктердің ықтималдық диаграммасы

Нәтижелер көрсеткендей, Wav2Vec 2.0 моделі сөйлеуді тану дәлдігі бойынша ең жоғары көрсеткішке ие болды, ал CNN+LSTM архитектурасы шулы ортада жақсы бейімделгенімен, таза сөйлеуге қарағанда 8,7%-ға төмен нәтижеге ие болды (1-кесте).

Кесте 1 – Салыстырмалы талдау

Модель	Шулы ортадағы дәлдік (%)	Таза ортадағы дәлдік (%)	WER (%)
CNN+LSTM	82,5	91,2	10,8
BERT	88,1	94,5	7,3
Wav2Vec 2.0	92,7	97,8	4,1

Модельдердің өнімділігін толық сипаттау үшін AUC-ROC қисықтарын қарастырдық. AUC 0.9-ден жоғары болған модельдер сенімді сөйлеуді тану жүйелері ретінде қарастырылуы мүмкін (2-кесте).

Кесте 2 – Модельдердің өнімділігі

Модель	AUC-ROC
CNN+LSTM	0.87
BERT	0.92
Wav2Vec 2.0	0.96

Бұл нәтижелер нейрондық желілерді дұрыс таңдаудың маңыздылығын көрсетеді. Сонымен қатар, деректерді аугментациялау (дыбыс өзгерістері, жылдамдық модификациясы, шуды қосу) модельдердің бейімделу қабілетін жақсартты.

Сөйлеуді сегментациялаудағы қателіктерді азайту үшін бірнеше әдісті қолдануға болады. Алдымен, VAD (дауыстық белсенділік детекторы) параметрлерін дұрыстап реттеу арқылы жалған оң және жалған теріс қателіктер арасындағы тепе-теңдікті табу маңызды. Фондық шуды азайту үшін сүзгілерді қолдану сөйлеу сигналдарын тазартуға және анықтауды дәлдеуге көмектеседі. Сонымен қатар, терең оқыту модельдері, мысалы, нейрондық желілер арқылы күрделі сөйлеу үлгілерін жақсы тануға болады. Спектралды талдау, мысалы, MFCC сияқты ерекшеліктер, сөйлеуді шудан ажыратуда тиімділік береді. Деректерді көбейту және аугментация жасау арқылы модельдерді әртүрлі жағдайларға бейімдеуге болады, бұл жалпы қателік мөлшерін азайтуға ықпал етеді [19–20].

Қорытынды

Қорытындылай келе, сөйлеуді сегментациялау–диктордыны сәйкестендіру және сөйлеуді тану жүйелерінің маңызды бөлігі болып табылады. Сегментация сапасы жалпы тану нәтижелерінің дәлдігіне тікелей әсер етеді, сондықтан қателіктерді азайту үшін әртүрлі әдістерді пайдалану өте маңызды. Жалған оң және жалған теріс қателіктер арасындағы тепе-теңдікті табу арқылы жүйенің сенімділігін арттыруға болады. Фондық шуды азайту және сүзгілеу әдістері сигнал сапасын жақсартуға көмектесіп, дәл сегментация жасауға мүмкіндік береді. Терең оқыту модельдері күрделі сөйлеу үлгілерін анықтауға бейімделіп, танудың жалпы нәтижелерін жақсартуға ықпал етеді.

Сонымен қатар, спектралды талдау әдістерін қолдану сөйлеу мен шу арасындағы айырмашылықты дәл анықтауға көмектеседі. Деректерді көбейту және әртүрлі жағдайларға бейімделу арқылы модельдер нақтырақ жұмыс істей алады. Осы тәсілдерді тиімді қолдану сөйлеуді тану жүйелерінің дәлдігі мен тұрақтылығын арттыруға мүмкіндік береді. Нәтижесінде, сөйлеуді сегментациялау процесін жақсарту арқылы сөйлеушіні сәйкестендірудің сенімділігін жоғарылату мүмкін болады. Бұл тәсілдер заманауи сөйлеуді тану жүйелерін жетілдіруге және олардың қолдану аясын кеңейтуге ықпал етеді.

Жүргізілген зерттеулер негізінде Wav2Vec 2.0 моделі қазақ тіліндегі сөйлеуді тану және сәйкестендіру бойынша ең жоғары нәтижелер көрсетті. Сонымен қатар, BERT негізіндегі

трансформерлік модельдер жақсы дәлдік көрсетті, бірақ есептеу ресурстарын көбірек қажет етті.

Бұл нәтижелер сөйлеуді тану жүйелерін жетілдіруде деректерді өңдеу әдістерінің және терең оқыту модельдерінің маңыздылығын растайды. Алдағы зерттеулерде қосымша тілдік модельдерді енгізу, казіргі модельдерді оңтайландыру және әртүрлі сөйлеу акценттері мен диалектілерді қамту мәселелерін қарастыру ұсынылады.

Қаржыландыру туралы ақпарат: Бұл зерттеуді Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым комитеті қаржыландырды (Грант № AP19678995).

ӘДЕБИЕТТЕР

1 Sujatha C. Vibration, Acoustics and Strain Measurement: Theory and Experiments. – 2023. – 722 p. https://doi.org/10.1007/978-3-031-03968-3_4.

2 Sudeep S. V. N. V. S., Venkata Kiran S., Nandan D., and Kumar S. An Overview of Biometrics and Face Spoofing Detection. – 2021. – ICCCE 2020. – P. 871–881.

3 Златоустова Л.В., Потапова Р.К., Потапов В.В, и Трунин-Донской В.Н. Общая и прикладная фонетика: Учеб. пособие. – 2-е изд., перераб. и доп. - М. : Изд-во Моск. гос. ун-та, 1997. – 416 с.

4 Mukazhanov N., Alibiyeva Zh., Yerimbetova A., Kassymova A., Alibiyeva N. Development of an augmented damerau–levenshtein method for correcting spelling errors in kazakh texts // Eastern-European Journal of Enter-prise Technologies. – 2023. – Vol. 5. – No. 2(125). – P. 23–33. https://doi.org/10.15587/1729-4061.2023.289187.

5 Амаан Ризви, Анупам Джаматия, Двиджен Рудрапал, Кунал Чакма и Бьёрн Гамбек. Cross-Lingual Speaker Identification for Indian Languages : Материалы 14-й Международной конференции по последним достижениям в обработке естественного языка, Варна, Болгария, 2023. – С. 979–987.

6 Pati D., and Prasanna S.R.M. Speaker verification using excitation source information // International Journal of Speech Technology. – 2012. – Vol. 15. – No. 3. – P. 241–257.

7 Zeinali H., Sameti H., and Burget L. HMM-based phrase-independent i-vector extractor for text-dependent speaker verification : IEEE/ACM Transactions on Audio Speech and Language Processing. – 2017. – Vol. 25. – P. 1421–1435.

8 Meftah A.H., Mathkour H., Kerrache S., and Alotaibi Y.A. Speaker Identification in Different Emotional States in Arabic and English. – 2020. – IEEE Access. – Vol. 8. – P. 60070–60083.

9 Гуртуева И.А., Бжихатлов К.Ч. Аналитический обзор и классификация методов выделения признаков акустического сигнала в речевых системах // Известия Кабардино-Балкарского научного центра РАН. – 2022. – Вып. 1. – С. 41–58. https://doi.org/10.35330/1991-6639-2022-1-105-41-58

10 Белов Ю.С., Нифонтов С.В., Азаренко К.А. Применение вейвлет-фильтрации для шумоподавления в речевых сигналах // Фундаментальные исследования. – 2017. – № 4 (часть 1) – С. 29–33.

11 Huang X., Acero A., and Hon H.W. Spoken Language Processing: A Guide to Theory, Algorithms, and Applications. – Prentice Hall, 2001.

12 Deng L., and Yu D. Deep Learning for Speech Recognition // IEEE Signal Processing Magazine. – 2012. – Vol. 29. – No. 6. – P. 82–97.

13 Zhang Y., and Wu Y. Robust Speech Segmentation using Spectral Clustering : IEEE Transactions on Audio, Speech, and Language Processing. – 2017. – Vol. 25. – No. – P. 91815–1827.

14 Hazen T.J., and Reddy R. Voice Activity Detection: A Review of the Literature // Journal of the Acoustical Society of America. – 2012. – Vol. 132. – No. 5. – P. 2994–3005.

15 Нефедов Н.Н., Алимурадов А.К. Краткий обзор способов обнаружения речевой активности // Инжиниринг и технологии. – 2024. – Т. 9. – № 2. – С. 1–6. https://doi.org/10.21685/2587- 7704-2024-9-2-9.

16 Sharma, S., and Goyal, N. A Review of Speech Segmentation Techniques // International Journal of Computer Applications. – 2016. – Vol. 134. – No. 11. – P. 10–14.

17 Sak H., Senior A., and Beaufays F. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. Proceedings of the 15th International Conference on Speech and Language Processing (INTERSPEECH), 2014. – pp. 338–342, (2014).

18 Gonzalez J., and Rios M.A Comprehensive Study on Speech Segmentation Approaches // Journal of Signal Processing Systems. – 2014. – Vol. 76. – No. 2. – P. 171–182.

19 Li X., and Zhao C. Improved Speech Segmentation Algorithm Based on GMM and VAD // Journal of the Acoustical Society of America. – 2015. – Vol. 137. – No. 3. – P. 1355–1363.

20 Yin J., and Wang Y. A Novel Method for Speech Segmentation Based on Wavelet Transform // International Journal of Speech Technology. – 2019. – Vol. 22. – No. 3. – P. 483–493.

REFERENCES

1 Sujatha C. Vibration, Acoustics and Strain Measurement: Theory and Experiments, 722 p. (2023). https://doi.org/10.1007/978-3-031-03968-3_4.

2 Sudeep S.V.N.V.S., Venkata Kiran S., Nandan D., and Kumar S. An Overview of Biometrics and Face Spoofing Detection, ICCCE 2020, 871–881, (2021).

3 Zlatoustova L.V., Potapova R.K., Potapov V.V., and Trunin-Donskoy V.N. General and Applied phonetics: textbook. stipend (Moscow, Publishing house of Moscow State University, 1997), 416 p. [In Russian].

4 Mukazhanov N., Alibiyeva Z., Yerimbetova A., Kassymova A., Alibiyeva N. Development of an augmented damerau–levenshtein method for correcting spelling errors in Kazakh texts, Eastern-European Journal of Enter-prise Technologies, 5 (2 (125)), 23–33 (2023). https://doi.org/10.15587/1729-4061.2023.289187.

5 Amaan Rizvi, Anupam Jamatia, Dwijen Rudrapal, Kunal Chakma and Bjorn Gambek. Cross-Lingual Speaker Identification for Indian Languages. In Proceedings of the 14th International Conference on the Latest Advances in Natural Language Processing (Varna, Bulgaria, 2023), pp. 979–987.

6 Pati D., and Prasanna S.R.M. Speaker verification using excitation source information. International Journal of Speech Technology, 15(3), 241–257 (2012).

7 Zeinali H., Sameti H., and Burget L. HMM-based phrase-independent i-vector extractor for text-dependent speaker verification. IEEE/ACM Transactions on Audio Speech and Language Processing, 25, 1421–1435 (2017).

8 Meftah A.H., Mathkour H., Kerrache S., and Alotaibi Y.A. Speaker Identification in Different Emotional States in Arabic and English. IEEE Access, 8, 60070–60083 (2020).

9 Gurtueva I.A., Brzikhatlov K.Ch. Analytical review and classification of methods for identifying acoustic signal features in speech systems, Izvestiya Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences, 1, 41–58 (2022). https://doi.org/10.35330/1991-6639-2022-1-105-41-58. [In Russian].

10 Belov Yu.S., Nifontov S.V., Azarenko K.A. Application of wavelet filtering for noise reduction in speech signals, Basic research, 4 (part 1), 29–33 (2017). [In Russian].

11 Huang X., Acero A., and Hon H.W. Spoken Language Processing: A Guide to Theory, Algorithms, and Applications. Prentice Hall, 2001.

12 Deng L., and Yu D. Deep Learning for Speech Recognition. IEEE Signal Processing Magazine, 29(6), 82–97 (2012).

13. Zhang Y., and Wu Y. Robust Speech Segmentation using Spectral Clustering. IEEE Transactions on Audio, Speech, and Language Processing, 25(9), 1815–1827 (2017).

14 Hazen T.J., and Reddy R. Voice Activity Detection: A Review of the Literature. Journal of the Acoustical Society of America, 132(5), 2994–3005 (2012).

15 Nefedov N.N., Alimuradov A.K. A brief overview of ways to detect speech activity. Engineering and technology, 9 (2), 1–6 (2024). https://doi.org/10.21685/2587- 7704-2024-9-2-9. [In Russian].

16 Sharma S., and Goyal N. A Review of Speech Segmentation Techniques. International Journal of Computer Applications, 134(11), 10–14 (2016).

17 Sak H., Senior A., and Beaufays F. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. Proceedings of the 15th International Conference on Speech and Language Processing (INTERSPEECH, 2014), pp. 338–342.

18 Gonzalez J., and Rios M. A Comprehensive Study on Speech Segmentation Approaches. Journal of Signal Processing Systems, 76(2), 171–182 (2014).

19 Li X., and Zhao C. Improved Speech Segmentation Algorithm Based on GMM and VAD. Journal of the Acoustical Society of America, 137(3), 1355–1363 (2015).

20 Yin J., and Wang Y. A Novel Method for Speech Segmentation Based on Wavelet Transform. International Journal of Speech Technology, 22(3), 483–493 (2019).

[1]*Ахмедиярова А.Т.,
PhD, ассоц. профессор, ORCID ID: 0000-0003-4439-7313,
*e-mail: a.akhmediyarova@satbayev.university
[1]Алибиева Ж.М.,
PhD, ассоц. профессор, ORCID ID: 0000-0001-9565-5621,
e-mail: zh.alibiyeva@satbayev.university
[1]Мукажанов Н.К.,
PhD, ассоц. профессор, ORCID ID: 0000-0003-4835-5751,
e-mail: n.mukazhanov@satbayev.university

[1]Сатбаев университеті, Алматы қ., Қазақстан

СЕГМЕНТАЦИЯ РЕЧИ ВО ВРЕМЯ СООТВЕТСТВИЯ ДИКТОРА

Аннотация

Сегментация речи – это процесс разделения речевых сигналов на части, который является важным аспектом систем идентификации говорящего и распознавания речи. Этот процесс повышает эффективность системы, позволяя точно определять начало и конец речи. Использование детекторов речевой активности (VAD) играет важную роль в сегментации, поскольку они помогают определить границы между речью и тишиной. Однако наиболее распространенными ошибками при сегментации являются ложноположительные и ложноотрицательные результаты, которые негативно влияют на общую точность системы. В связи с этим необходимо снижать ошибки за счет различных подходов и методов. Такие меры, как снижение фонового шума, использование моделей глубокого обучения и увеличение данных, могут значительно улучшить качество сегментации. Использование методов и особенностей спектрального анализа позволяет четко различать речь и фоновый шум. Целью данного исследования является оптимизация процесса сегментации и анализ вероятности ошибок, повышение эффективности систем распознавания речи. В результате эта работа является основой для новых исследований и разработок в области распознавания речи. В статье рассматривается проблема сегментации речи для идентификации говорящего. В работе описаны возможные критерии сегментации – качественные и количественные характеристики звуковой речи, например, речевые задержки и интонация, а также их акустическое соотношение. Это позволяет специалисту выделить конкретные сегментные единицы (слоги, слова и т. д.), записать их структуру, выделить основные признаки.

Ключевые слова: речевой сигнал, сегментация, голос и речь, поддержка фрагментов речевого сигнала, метод идентификации.

[1]*Akhmediyarova A.T.,
PhD, Associate Professor, ORCID ID: 0000-0003-4439-7313,
*e-mail: a.akhmediyarova@satbayev.university
[1]Alibiyeva Zh.M.,
PhD, Associate Professor, orcid.org/0000-0001-9565-5621,
e-mail: zh.alibiyeva@satbayev.university
[1]Mukazhanov N.K.,
PhD, Associate Professor, orcid.org/ 0000-0003-4835-5751,
e-mail: n.mukazhanov@satbayev.university

[1]Satbaev University, Almaty, Kazakhstan

SPEECH SEGMENTATION DURING SPEAKER MATCHING

Abstract

Speech segmentation is the process of dividing speech signals into parts, which is an important aspect of speaker identification and speech recognition systems. This process improves the efficiency of the system by accurately detecting the beginning and end of speech. The use of voice activity detectors (VADs) plays an important

role in segmentation, as they help to determine the boundaries between speech and silence. However, the most common errors in segmentation are false positives and false negatives, which negatively affect the overall accuracy of the system. In this regard, it is necessary to reduce errors through various approaches and methods. Measures such as reducing background noise, using deep learning models, and data augmentation can significantly improve the quality of segmentation. Using spectral analysis methods and features allows you to clearly distinguish between speech and background noise. The purpose of this study is to optimize the segmentation process and analyze the probability of errors, improve the efficiency of speech recognition systems. As a result, this work provides a basis for new research and development in the field of speech recognition. The article considers the problem of speech segmentation for speaker identification. The paper describes possible segmentation criteria - qualitative and quantitative characteristics of sound speech, such as speech delays and intonation, as well as their acoustic relationship. This allows a specialist to identify specific segment units (syllables, words, etc.), record their structure, and identify the main features.

Keywords: speech signal, segmentation, voice and speech, support for speech signal fragments, identification method.

Мақаланың редакцияға түскен күні:14.10.2024