RĪGAS TEHNISKĀ UNIVERSITĀTE
Datorzinātnes un informācijas tehnoloģijas fakultāte
Lietišķo datorsistēmu institūts
Mākslīgā intelekta un sistēmu inženierijas katedra

Betija Rudzīte
Akadēmiskā bakalaura studiju programma
„Intelektuālas robotizētas sistēmas"
(stud. apl. nr. 201RDB305)

Runas atpazīšanas priekšapstrādes metožu salīdzināšana

Bakalaura darbs

Zinātniskais vadītājs Phd. Sc.Comp., pētnieks
Ēvalds Urtāns

Rīga - 2023

RĪGAS TEHNISKĀ UNIVERSITĀTE

DATORZINĀTNES UN INFORMĀCIJAS TEHNOLOĢIJAS FAKULTĀTE

Lietišķo datorsistēmu institūts

Mākslīgā intelekta un sistēmu inženierijas katedra

Bakalaura darba izpildes lapa

Noslēguma darba autors:

students(-e) Betija Rudzīte _____

(paraksts, datums)

Noslēguma darbs ieteikts aizstāvēšanai:

Zinātniskais vadītājs:

Phd. Sc.Comp., pētnieks Ēvalds Urtāns _____

(paraksts, datums)

ANOTĀCIJA

RUNAS ATPAZĪŠANA, PRIEKŠAPSTRĀDES METODES, RUNAS STILA PĀRNESE, RUNAS ATTĪRĪŠANA

Bakalaura darba tips:
1. tips: Moderno risinājumu izpēte

Runas atpazīšana atvieglo pierakstu veikšanu, kā arī tā tiek izmantota tādās tehnoloģijās kā balss asistenti tālruņos un palīdz cilvēkiem ar redzes traucējumiem. Tomēr neskatoties uz to augsto pieprasījumu, automātiskajā teksta izvadē joprojām ir sastopamas kļūdas.

Šajā darbā tika veikta salīdzināšana runas atpazīšanas metriku rezultātiem izmantojot priekšapstrādes metodes jeb precīzāk - vai ir iespējams uzlabot metriku rezultātus, kas iegūti no runas atpazīšanas, ja tam pirms tam veic priekšapstrādi, runas stilu pārnesot uz kādu konkrētu runātāju uz kuru ir trenēts runas atpazīšanas modelis.

Kopumā darbā tika noskaidrots, ka apmācot Whisper (Radford, Kim et al., 2022) runas atpazīšanas modeli uz konkrētu runātāju no VCTK (Veaux, Yamagishi et al., 2017) datu kopas, pēc tam veicot runas stila pārnesi ar FreeVC (Li, Tu et al., 2022) uz to pašu konkrēto runātāju, ir iespējams iegūt labākus rezultātus nekā tad, ja runas atpazīšanas modelis ir trenēts uz visas datu kopas un tam pirms tam nav veikta nekāda priekšapstrāde. Rezultātu metriku uzlabojums ir sākot ar 34% no datu kopas runātāju līdz pat 91% no datu kopas runātāju. Vislabāko sniegumu parāda runas atpazīšanas modelis, kas trenēts uz p254 runātāju, kur, pārveidojot visus datu kopas runātājus uz p254 runātāja stilu, parāda 74% un 73% runātāju uzlabojumu vārdu kļūdas līmeņa un normalizētās vārdu kļūdas līmeņa metrikām un 86% un 91% runātāju uzlabojumu rakstzīmes kļūdas līmeņa un normalizētajai rakstzīmes kļūdas līmeņa metrikām.

Darba pamattekstā ir 68 lappuses, 9 attēli, 38 tabulas, 10 pielikumi un 63 informācijas avoti.

ABSTRACT

SPEECH RECOGNITION, PRE-PROCESSING METHODS, SPEECH STYLE TRANSFER, SPEECH ENHANCEMENT

Bachelor thesis type:
Type 1: Research on Modern Solutions

Speech recognition makes it easier to take notes as well as it is used in technology such as voice assistants in phones and it is helpful for visually impaired people. However, despite their high demand, there are still errors in the output of automatic speech recognition systems.

In this work, a comparison was made using metric results from a speech recognition system using pre-processing methods, specifically - whether it is possible to improve metric results, that have been obtained from speech recognition with a pre-processing method of speech style transfer to a specific speaker on which the speech recognition model has been trained.

Overall, it has been concluded in this work that training Whisper's (Radford, Kim et al., 2022) speech recognition model on a specific speaker from the VCTK (Veaux, Yamagishi et al., 2017) dataset, then performing speech style transfer with FreeVC (Li, Tu et al., 2022) on the same specific speaker, can produce better results than training the speech recognition model on the entire dataset and using no pre-processing methods. The improvement in outcome metrics ranges from improving 34% of the speakers to as high as improving 91% of the speakers in the dataset. The best performance is shown by the speech recognition model trained on the p254 speaker, where transforming all speakers in the dataset to the p254 speaker style shows 74% and 73% speaker improvement for the word error rate and normalized word error rate metrics, and 86% and 91% speaker improvement for character error rate and normalized character error rate metrics.

The main text of the bachelor's thesis consists of 68 pages, 9 pictures, 38 tables, 10 appendices and 63 sources of information.

SATURS

IEVADS **6**

1. Saistītie pētījumi **7**
1.1. Dziļā mašīnmācīšanās 12
1.2. Runas ierakstu apstrāde 18

2. Sistemātiskā literatūras analīze **25**

3. Metodoloģija **48**
3.1. Datu kopa 50
3.2. Metrikas 50
3.3. Modeļu arhitektūras 51
3.3.1. Runas atpazīšanas modelis 51
3.3.2. Runas stila pārneses modelis 53
3.4. Apmācību un testēšanas protokols 55

4. Rezultāti **57**

5. Tālākie pētījumi **71**

SECINĀJUMI **72**

IZMANTOTĀ LITERATŪRA **74**

PIELIKUMI **80**
1. pielikums. Vispārēja informācija par runas atpazīšanas modeļu publikācijām. 80
2. pielikums. Vispārēja informācija par runas attīrīšanas modeļu publikācijām. 83
3. pielikums. Vispārēja informācija par runas stila pārneses modeļu publikācijām. 85
4. pielikums. Vispārēja informācija par kombinēto modeļu publikācijām. 88
5. pielikums. Metriku rezultāti, kas iegūti no VCTK modeļa. 90
6. pielikums. Metriku rezultāti, kas iegūti no p304 modeļa. 93
7. pielikums. Metriku rezultāti, kas iegūti no p317 modeļa. 96

8. pielikums. Metriku rezultāti, kas iegūti no p363 modeļa. 99
9. pielikums. Metriku rezultāti, kas iegūti no p287 modeļa. 102
10. pielikums. Metriku rezultāti, kas iegūti no p254 modeļa. 106

IEVADS

Runas atpazīšana, izmantojot mākslīgo neironu tīklus, joprojām pieļauj kļūdas rezultātu ieguvē un snieguma precizitāte nav absolūta. Nav skaidri zināms, vai rezultātus ietekmē audio kvalitāte, apmācībā esošie dati vai kāda specifiska runātāja balss īpatnības, kuras ir grūtāk atpazīt. Tāpēc ir nozīmīgi pārbaudīt kā izmainās rezultāti izmantojot dažādas priekšapstrādes metodes, kā piemēram runas attīrīšana vai balss stila pārnese.

Darba mērķis ir salīdzināt un atrast piemērotāko metodi balss audio ierakstu priekšapstrādei, lai iegūtu augstāko precizitāti angļu valodas runas atpazīšanas modeļiem. Lai sasniegtu darba mērķi ir izvirzīti šādi **darba uzdevumi**:

- Apgūt klasiskās skaņas apstrādes metodes.
- Apgūt dziļajā mašīnmācīšanā balstītās skaņas apstrādes metodes (runas attīrīšana, tempa maiņa, runātāja tembra maiņa, utt.).
- Apgūt runas atpazīšanas modeļu metodes (Viendaļīgie un trīsdaļīgie modeļi).
- Apgūt metrikas, runas atpazīšanas modeļu salīdzināšanai.
- Eksperimentāli salīdzināt priekšapstrādes metodes, lai uzlabotu runas atpazīšanas modeļu rezultātus.
- Publicēt rezultātus zinātniskā publikācijā.

Saskaņā ar darba mērķi un darba uzdevumiem darba struktūra ir šāda: pirmajā nodaļā tiek apkopota teorētiskā daļa par mākslīgo intelektu, kā arī specifiskas detaļas tā izmantošanā, otrajā nodaļā tiek veikta literatūras sistemātiskā analīze, lai izvērtētu šobrīd modernos risinājumus runas atpazīšanas priekšapstrādei, trešajā nodaļā tiek aprakstīti eksperimentāli izmantotās metodes un beigās tiek veikts apkopojums par iegūtajiem rezultātiem, norādīti tālākie pētījumi un veikti secinājumi.

1. SAISTĪTIE PĒTĪJUMI

Runas atpazīšana ir svarīgs uzdevums, kur pielietojumu radušas jau vairākas tehnoloģijas, un pieprasījums pēc tā ir augošs. Tā ir spējīga radīt alternatīvu rakstīšanai uz papīra vai izmantojot klaviatūru, tādējādi ietaupot laiku, kā arī ir pretimnākoša tehnoloģija cilvēkiem ar redzes traucējumiem. Taču, lai panāktu augstu runas atpazīšanas precizitāti, nepieciešama efektīva priekšapstrāde, kas var ievērojami uzlabot runas kvalitāti, samazinot trokšņus un citus traucējumus. Līdz ar to, šajā pētījumā tiks apskatītas, salīdzinātas un novērtētas dažādas priekšapstrādes metodes, lai atrastu labāko risinājumu runas atpazīšanas uzdevumiem.

Nepieciešamība pēc runas atpazīšanas sistēmām var rasties no tā, ka cilvēki brauc vai nav spējīgi izmantot rokas, jo tās ir netīras vai aizņemtas, kā arī šādas sistēmas ir noderīgas cilvēkiem ar redzes traucējumiem, kur vizuālā maņa nav tik ērti izmantojama. Bieži tiek izmantoti balss asistenti, kas ir pieejami viedtālruņos vai citās ierīcēs. Tas padara ikdienas dzīvi nedaudz vienkāršāku.

Balss asistenti gūst arvien augstāku popularitāti, kaut arī tā jau ir bijusi izmantota ilgu laiku, piemēram, kā Apple produktos izmantotais Siri, kas atpazīst balss komandas, vai Android produktos izmantotais Google assistant, vai arī "Alexa", kas ir Amaozon radīts produkts, kas nodrošina gudrās mājas funkcionalitāti. "Siri", "Google Assistant" un "Alexa" ir 3 vadošie balss asistentu jomas produkti (Thompson & Munster, 2019) un tos cilvēki savā ikdienā izmanto visbiežāk. Pētījumā (Thompson & Munster, 2019), ko 2019. gadā veica "Deepwater Asset Management" tika pārbaudīts šo asistentu IQ jeb viņiem tika uzdoti 800 balss jautājumi vai sniegtas komandas 5 kategorijās un tika izvērtēts, cik daudz viņi spēj pilnvērtīgi izpildīt. Pētījuma rezultāti norāda, ka visefektīvāk strādājošā sistēma ir "Google Assistant", kas pareizi izpildīja 92.9 %, tam sekoja "Alexa" ar 83.1%, un beigās bija "Siri" ar 79.8 % pareizi izpildīto komandu. Pētījumā arī norāda, ka salīdzinot ar iepriekšējo gadu, visas sistēmas ir pilnveidotas un rādījušas labāku sniegumu nekā iepriekš, līdz ar to, pat pie iegūtajiem augstajiem rezultātiem, notiek nepārtraukta šo modeļu attīstība.

Turpmāk tika aplūkoti 3 tirgus pētījumi, ko izveidojuši "Fortune Business Insights", "Grand View Research" un "MarketsandMarkets Research" un izvērtēti to rezultāti, lai noskaidrotu šobrīd esošās tendences runas un balss atpazīšanas jomā. Pēc "Fortune Business Insights" (Fortune Business Insights, 2022) datiem, globālā tirgus vērtība balss un runas atpa-

zīšanai 2021. gadā bija 9.56 miljardi ASV dolāri un pētījumā tiek izteikts, ka līdz 2029. gadam tas pieaugs līdz pat 49.79 miljardiem ASV dolāru, paredzot 23.7 % CAGR (compound annual growth rate (kopējais gada izaugsmes temps)). "Grand View Research" (Grand View Research, 2021) norāda, ka tirgus vērtība 2021. gadā ir novērtēta uz 14.42 miljardiem ASV dolāru un izsaka prognozi, ka no 2022. gada līdz 2030. gadam CAGR pieaugums būs 15.3 %. "MarketsandMarkets Research" (MarketsandMarkets Research, 2022) norāda, ka 2022. gadā tirgus vērtība bija 9.4 miljardi ASV dolāri un sniedz minējumu, ka līdz 2027. gadam sasniegs 28.1 miljardu dolāru, kas tiktu vērtēts ar 24.4 % CAGR pieaugumu. Neskatoties uz šiem atšķirīgajiem rezultātiem, tiek paredzēts vismaz 15.3 % CAGR pieaugums šajā nozarē, kas nozīmē, ka tehnoloģijas tiks attīstītas un ir pieprasījums pēc attīstītākām sistēmām.

Kā norāda "Fortune Business Insights" (Fortune Business Insights, 2022), balss atpazīšanu izmanto un attīsta tādas zināmas un lielas kompānijas kā Amazon, IBM, Apple, Google, Microsoft un vēl daudzas citas, kas iegulda daudz līdzekļus savos produktos un cenšas sniegt labākos rezultātus saviem klientiem. Jāņem vērā, ka jebkāda veida balss ieraksta sistēmām ir jānodrošina lietotāju privātums un datu aizsardzība, kas rada nepieciešamību ieguldīt vēl līdzekļus efektīvās drošības sistēmās. Šis ir viens no lielākajiem iemesliem, kāpēc izaugsme nav spējīga sasniegt savu maksimālo pieaugumu, jo ir papildus līdzekļi, kas ir jātērē šīm sistēmām un nav iespējams visu finansējumu ieguldīt tikai zinātnes attīstībai. Viena no zināmākajām opcijām, ko šī tehnoloģija būtu spējīga pilnībā aizstāt, ir subtitru nodrošināšana, ko līdz šim ir veikuši cilvēki, kas arī pētījumā norāda, ka šis sastāda aptuveni 44 % no visa tirgus.

Balss atpazīšana tiek izmantota daudzās nozarēs. Protams visvairāk pielietota tā ir IT nozarē (Fortune Business Insights, 2022), taču to izmanto arī tādās nozarēs, kā automobiļu rūpniecībā, veselības aprūpe, īpaši pēc Covid un Covid laikā, kad ārstu pieprasījums un noslogojums ir bijis augsts, bankas, finanšu pakalpojumi un apdrošināšana, un vēl arī citās.

Pētījumā, ko veicis ir "Grand View Research" (Grand View Research, 2021) viņi paredz, ka runas atpazīšanas sistēmas gūs savu izaugsmi tehnoloģiju attīstības rezultātā, kā arī iespējams, ka balss atpazīšanas sistēmas tiks plašāk pielietotas drošības sistēmas kā biometrijas rīks, tādējādi pārliecinoties par cilvēka identitātes patiesumu pēc viņa balss. Arvien pieaug pieprasījums pēc balss vadītām sistēmām gan programmatūrai, gan arī aparatūras pielietošanai. Šobrīd sistēmas plaši tiek izmantotas automašīnā,

kur ir aizliegums izmantot telefonu kamēr autovadītājs brauc, jo tas novērš uzmanību. Veselības aprūpes industrijā ir bijis īpašs pieaugums runas atpazīšanas sistēmas pielietojumā, jo tas atvieglo medicīnas personālu darbu, kad tiem nepieciešams veikt pierakstus par pacientu, viņa stāvokli, simptomiem un iespējamajām diagnozēm vai kādu svarīgu informāciju procedūras laikā, tādējādi ietaupot speciālista laiku, ļaujot tam vairāk laika veltīt tieši pašam pacientam un paaugstinot personāla produktīvo laiku. Kā arī iepriekšējā pētījumā minēts, sevišķi Covid laikā, ir radies augstais pieprasījums šādām tehnoloģijām, jo pacientu skaita palielinājuma rezultātā, medicīnas darbiniekiem bija augsta nepieciešamība apkalpot maksimāli daudz pacientus pēc iespējas īsākā laikā.

Arī pētījumā ko veicis "MarketsandMarkets Research" (MarketsandMarkets Research, 2022) tiek paredzēts, ka būs pieaugums pēc runas atpazīšanas tehnoloģijām, jo veidosies lielāks pieprasījums pēc veselības aprūpes efektivitātes uzlabošanas, kā arī viedās jeb gudrās ierīces tiek izmantotas arvien vairāk. Pēc pētījuma rezultātiem tika arī secināts, ka iepērkoties tiešsaistē, 41 % cilvēku labprāt vēlētos iespēju nodrošināt iepirkšanos ar balss asistenta palīdzību, jo tas atvieglojot viņu darbības, padarot šo procedūru automatizētu.

Tomēr, kā minēts pētījumā (MarketsandMarkets Research, 2022), fona trokšņi var ietekmēt audio kvalitāti, kas var ietekmēt balss atpazīšanas modeļa veiktspēju un samazināt iespēju, ka balss asistents spēs atpazīt sniegto komandu vai saprast uzdoto jautājumu. Šī problēma ir īpaši izaicinoša āra un biroja telpās, kur fona trokšņu līmenis ir visaugstākais. Tiek paredzēts, ka uzņēmumi, kas pievērsīs vairāk uzmanības tam, kā risināt šo fona trokšņu problēmu, lai uzlabotu runas atpazīšanas sistēmu, būs tie, kuri spēs piedāvāt klientiem labākus produktus un apmierināt viņu vēlmes.

Vēl "MarketsandMarkets Research" (MarketsandMarkets Research, 2022) norāda, ka runas un balss atpazīšanas tirgus izaugsmi nodrošinās pieprasījums pēc virtuālajiem asistentiem, brīvroku saskarnes sistēmas un pieejamība cilvēkiem ar invaliditāti, kā arī gudro skaļruņu popularitātes pieaugums. Galvenokārt peļņu radīs ieguldījumi pētniecībā un attīstībā, uzmanību vēršot precizitātei un uzticamībai un balss-vadītu saskarņu paplašināšana tādās jaunās nozarēs, kā veselības aprūpe, izglītība un komercija.

Neskatoties uz jau esošajiem pieejamajiem risinājumiem, ir vairākas problēmas, kā piemēram, runas atpazīšanas modeļus apmāca ar datiem, kuri satur limitētu skaitu runātāju, taču tos pielieto uz plašāku skaitu runātāju. Lai runas atpazīšanas modeli apmācot iegūtu pēc iespējas augstāku

rezultātu, apmācība tiek izmantoti dati ar nelielu skaitu runātājiem, taču pēc apmācības, dati, ko saņems modelis nebūs tikai no tiem runātājiem, kas tika izmantoti apmācībā. Mēdz būt tā, ka modelis ir trenēts uz vienu konkrētu runātāju, taču pēc tam ir kāds pilnībā cits, un, līdz ar to, rezultāti var iznākt ne tik veiksmīgi.

Informācija par to, kāda ietekme ir runātāju īpašībām uz rezultātu precizitāti nav zināma. Iespējams, ka šo runātāja personības ietekmes samazināšana spēs uzlabot rezultātu precizitāti. Cilvēki mēdz runāt dažādi – skaļāk, klusāk, ātrāk, lēnāk – viņu balss stili un tembri ir atšķirīgi. Dabīgi runājot, cilvēki mēdz pa vidu ieturēt pauzes, vai norādīt savu viedokli, izdvešot skaņas, kuras netiek pārveidotas vārdos, kas atrodas vārdnīcā (aha, mhm, hmm, u.t.t.), cilvēki runājot mēdz pārteikties vai pateikt kādu vārdu nepareizi, viņi var runāt ar dažādiem akcentiem, izmantot atšķirīgus dialektus, likt uzsvaru uz citām vietām vārdos, veidot gramatiski nepareizus teikumus. Visas šīs īpašības var ietekmēt to, kā modelis mācās. Neskaitot tikai to, fonā var būt kādi trokšņi no dabas, vai vienlaicīgi runāt vairāki cilvēki. Kā arī ir nepieciešams nodrošināt, ka apmācība izmantotie dati nekāda veidā neaizskar cilvēku privātumu un ar tiem nebūs iespējams norādīt vai atpazīt konkrēto cilvēku.

Modeļa precizitāti spēj ietekmēt ne tikai runātāja īpašības, bet arī audio kvalitāte. Trenētie modeļi bieži vien ir uz datu kopām, kuras ir attīrītas, jeb ierakstītā runa ir bijusi veikta no klusas telpas vai audio pēc tam tika pārstrādāts noņemot trokšņus, līdz ar to, apmācības dati satur tīrus audio ierakstus, bet inferencē izmantojot dabiskus audio failus, tie var saturēt troksni un nebūt tik skaidri saprotami.

Līdz šim nav zināms, vai ir iespējams uzlabot precizitāti, attīrot runu ar runas kvalitātes uzlabošanas modeļiem vai arī pārveidojot runas stilu uz tādu, kāds tika izmantots apmācībā. Iespējams, iegūt augsta rezultāta runas apmācības modeli uz vienu runātāju ir ļoti vērtīgi, ja pēc tam visus pārējos datus, kas iet inferencē ar runas stila pārveidošanas modeli var nomainīt tādus, kā runātāja datus, uz kura ir apmācīts modelis, tādējādi normalizējot datus uz vienu runātāju.

Lai varētu nodrošināt lielu runātāju skaitu apmācību un dažādību var būt nepieciešams ilgs laiks un iegūtie rezultāti var arī nesasniegt iecerētās vēlmes. Tikai dati nav spējīgi nodrošināt, ka pēc tam būs iespējams labi atpazīt jebkura veida runātāju ar visām viņa īpašībām.

Izmantotajām datu kopām arī ir liela nozīme tajā, cik labus rezultātus modelis spēs producēt. Lai labi varētu apmācīt modeli ir nepieciešams

izmantot lielas datu kopas, kas nodrošina augstu dažādību, taču, jo lielāka datu kopa tiek izmantota, jo lielāka ir iespēja, ka tajā kaut kur ir kāda kļūda. Kā arī, piemēram, trenējot datus uz tīras datu kopas un tādas, kurā ir trokšņi var sniegt dažādus rezultātus. Tas pats ir ja piemēram modelis ir trenēts tikai uz tīriem datiem un testēšanai tiek izmantoti dati ar trokšņiem. Tas uzreiz sniegs zemākus rezultātus. Bet cenšoties uztrenēt datu kopu uz trokšņiem, tas centīsies iemācīties trokšņus un tos atpazīt vai sasaistīt ar vārdiem, kā arī tas negarantē to, ka tas spēs noteikt iepriekš nedzirdētus trokšņus.

Ņemot vērā tirgus pētījumus un pieejamos rezultātus par tiem, ir skaidrs, ka šī joma šobrīd ir aktuāla un arvien vairāk tiek attīstīta, un tiek meklēti risinājumi, lai iegūtu visaugstākās kvalitātes runas atpazīšanas sistēmas. Lai būtu iespējams uzlabot pašreizējos rezultātus un veicināt attīstību, viena no prognozēm, kā to panākt, ir izmantot priekšapstrādes metodes, līdz ar to, šī darba mērķis ir salīdzināt un atrast piemērotāko metodi balss audio ierakstu priekšapstrādei, lai iegūtu augstāko precizitāti angļu valodas runas atpazīšanas modeļiem.

Dotajā darbā ir plāns izskatīt šobrīd pieejamos modeļus un to sniegtos rezultātus runas atpazīšanā (speech to text), runas attīrīšanā (speech enhancement), runas stila pārveidošanā (speech style transfer) un šo modeļu kombinācijas. Sākotnējais darba pieņēmums ir, ka izmantojot runas atpazīšanas modeļus, kas ir trenēti uz konkrētiem runātājiem, pēc trenēšanas procesiem tiem veicot audio pārveidošanu uz tekstu, kur šim audio ir veikta priekšapstrāde, tā stilu pārnesot uz zināmu runātāju, attiecīgais modelis būs spējīgs sniegt labākus rezultātus nekā tad, ja tas veiks runas atpazīšanas uzdevumu uz datiem, kuri iepriekš nav bijuši apmācības procesā un kuriem nav bijusi veikta nekāda apstrāde. Lai būtu iespējams pilnvērtīgi veikt visas šīs darbības, tiek izvirzīti sekojoši darba uzdevumi:

1.) Apgūt klasiskās skaņas apstrādes metodes.
2.) Apgūt dziļajā mašīnmācīšanā balstītās skaņas apstrādes metodes (runas attīrīšana, tempa maiņa, runātāja tembra maiņa, utt.).
3.) Apgūt runas atpazīšanas modeļu metodes (Viendaļīgie un trīsdaļīgie modeļi).
4.) Apgūt metrikas, runas atpazīšanas modeļu salīdzināšanai.
5.) Eksperimentāli salīdzināt priekšapstrādes metodes, lai uzlabotu runas atpazīšanas modeļu rezultātus.

Praktiskajā daļa ir paredzēts izvēlēties vienu runas atpazīšanas modeli un tam veikt apmācību uz vienu iepriekš izvēlētu datu kopu. Iegūt teksta rezultātus no apmācītā runas atpazīšanas modeļa un no tiem iegūt vārda kļūdas līmeņa un rakstzīmes kļūdas līmeņa metrikas. Tālāk izvēlēties piecus runātājus no visas datu kopas un apmācīt runas atpazīšanas modeļus uz šiem runātājiem. Iegūt teksta rezultātus no apmācītajiem runātāju modeļiem, visas datu kopas runātāju runas stilu pārveidojot uz piecu izvēlēto runātāju stilu, izmantojot runas stila pārveidošanas modeli. No teksta rezultātiem iegūt vārda kļūdas līmeņa un rakstzīmes kļūdas līmeņa metrikas. Abos gadījumos iegūtās metrikas salīdzināt. Mērķis ir samazināt vārda kļūdas līmeņa un rakstzīmes kļūdas līmeņa rezultātus tekstam, kurš iegūts audio izejot cauri runas stila pārveidošanas modelim un pēc tam runas atpazīšanas modelim attiecībā pret to tekstu, kurš iegūts no audio izejot cauri tikai runas atpazīšanas modelim.

1.1. Dziļā mašīnmācīšanās

Mākslīgajam intelektam, tāpat kā cilvēka dabīgajam intelektam, nav viennozīmīga un konkrēta definīcija, jo mākslīgais intelekts, kā jau pēc tā nosaukuma var spriest, ir dabīgā intelekta atdarinājums. Mākslīgo intelektu var uztvert kā cilvēku darbību vai domāšanas veida automatizāciju jeb, kā jau iepriekš minēts, cilvēka intelektuālu darbību atdarinājumu.

Mašīnmācīšanās (Chollet, 2017) ir mākslīgā intelekta apakš nozare, kas, kā jau pēc vārda var noprast, pati mācās. Tas, kā notiek mācīšanās ir sekojoši: modelim tiek padots liels daudzums datu, kuriem tas veic matemātiskus aprēķinus, lai atrastu līdzīgās un atšķirīgās iezīmes starp visiem datiem. Tālāk dati tiek sašķiroti balstoties uz šīm iezīmēm. Jau iepriekš definētas atbildes par kategorijām, kurās datiem būtu jābūt sašķirotiem un to, kurai kategorijai kas pieder, var būt, bet var arī nebūt. Mācīšanās rezultātā modelis ir saglabājis likumu kopu, pēc kuras šos datus var sašķirot un, ja apmācība ir bijusi veiksmīga un rezultāti ir pietiekami precīzi, tad, izmantojot šo likumu kopu, modelis spēs precīzi sašķirot nezināmus, iepriekš apmācības procesā neizmantotus datus.

Dziļā mācīšanās (Chollet, 2017) ir mašīnmācīšanās specifiska apakšnozare. Lai arī tāpat kā mašīnmācīšanās tā savus rezultātus iegūst no ievades datiem apmācības rezultātā, dziļajā mašīnmācīšanās lielākais uzsvars ir uz to, ka ir daudz neironu slāņi, caur kuriem iziet cauri dati, un tāpēc to sauc tieši par dziļo mācīšanos. Jaunākie modeļi var saturēt pat vairāk kā simts

slāņus, līdz ar to ir liels apmācāmo parametru skaits. Ejot cauri slāņiem, ievades datu informācija tiek sadalīta vairākās iezīmēs, tādējādi ir iespējams aprakstīt daudz vairāk svarīgu informāciju, kas tiek izsecināta apmācības procesā, kas beigās tiek izmantota rezultāta noteikšanai. Būtībā dziļie neironu tīkli ir daudz slāņu sistēma, lai iegūtu informāciju par svarīgām ievades datu iezīmēm un attēlojumiem.

Mācīšanās procesā tiek iziets cauri vairākām iterācijām atkārtojot konkrētas darbības. Trenējot modeli, sekojošie soļi tiek atkārtoti tik daudz, cik nepieciešams, lai iegūtie rezultāti no kļūdas funkcijas būtu pietiekami apmierinoši vai arī ir sasniegts iepriekš noteiktais iterāciju skaita maksimums. Apmācības process var izskatīties sekojoši (Chollet, 2017):

1. Paraugu grupa ar ievades datiem un to attiecīgās reālās rezultātu vērtībām tiek padotas modelim apmācībai.

2. Uz paņemtās paraugu grupas tiek palaists neironu tīkls, kā rezultātā tiek iegūta kāda vērtība, kas ir prognozētais rezultāts.

3. Tiek aprēķināta kļūdas vērtība, kas norāda, cik daudz reālā vērtība atšķiras no prognozētās vērtības.

4. Attiecīgie algoritmi nomaina svara vērtības uz citiem parametriem, tādējādi cerams samazinot kļūdas funkcijas rezultāta vērtību nākamajā solī.

Kad apmācības process ir pabeigts un ir iegūta pietiekami zema kļūdas funkcijas vērtība vai arī sasniegts maksimālais iterāciju skaits, saglabāto modeli var pārbaudīt uz datiem, kuri nav parādījušies apmācības procesā. Rezultātā, ar modeļa apmācībā iegūtajām labākajām parametru vērtībām, tam tiks izlaisti cauri dati, kuriem viņš aprēķinās kādu gala vērtību.

Mašīnmācīšanos var pielietot dažādās jomās un atšķirīgiem uzdevumiem, tomēr atkarībā no pieejamajiem ieejas datiem un informāciju par izejas vērtībām, mašīnmācīšanos var iedalīt trijās kategorijās (Vasilev, Slater et al., 2019): pārraudzītā, nepārraudzītā un stimulētā. Pārraudzītā mašīnmācīšanās kopā ar ievades datiem modelim ir arī zināmi izvades dati un apmācības procesā tas iemācās ievades datu iezīmes, lai pareizi spētu noteikt izvades datus. Pārraudzīto mašīnmācīšanos visbiežāk izmanto klasifikācijas vai regresijas uzdevumos. Nepārraudzītā mašīnmācīšanās neizmanto jau zināmas gala vērtības, bet visbiežāk atrod kādas kopīgas iezīmes datos un pēc

tām atdala datus vairākās kategorijas pēc to kopējām iezīmēm. Bieži vien dabīgās valodas apstrādē izmanto nepārraudzīto mašīnmācīšanos, kā arī ģeneratīvajos modeļos. Stimulētā mašīnmācīšanās izmanto vidi, kura apmācībā esošajam modelim sniedz atgriezenisko saiti par to, vai modeļa lēmumi bija pareizi vai nē, un balstoties uz šo atbildes reakciju, modelis cenšas veikt tādus lēmumus, kas tam dos vislabāko rezultātu.

Bieži vien arī atsevišķi izdala vēl ceturto apmācības veidu - pašpārraudzīto apmācību (Chollet, 2017). Šis ir pārraudzītās apmācības veids, kuram sākotnēji apmācībai netiek doti jau iepriekš definēti izvades dati jeb kādām ir jābūt gala vērtībām. Modelis apmācības procesā no ieejas datiem iemācās un izveido izejas datus, kurus tad izmanto pārraudzītās apmācības veidā. Šādu apmācības pieeju parasti izmanto autoenkoderi.

Viena no zināmākajām metodēm, kā izveidot mašīnmācīšanās modeli (Aggarwal, 2018) ir izmantojot mākslīgo neironu tīklus. Nosaukums tiem ir piešķirts pateicoties to shematiskajam izskatam, kas līdzinās bioloģiskiem neironiem.

Katrs neirons atbilst kādai iezīmei, ko modelis ir atdalījis no ieejas datiem (Aggarwal, 2018), un katrs savienojums starp neironiem ir aprakstīts ar skaitlisku vērtību, ko sauc par svariem, tādā veidā piešķirot iezīmēm skaitlisku vērtību. Visas iezīmes ir savstarpēji viena ar otru savienotas. Ar funkciju palīdzību tiek aprēķināts skaitlisks iznākums tā brīža svaru vērtībām, kuras beigās tiek salīdzinātas ar reālo vērtību. Ja reālā vērtība no aprēķinātās atšķirsies, kas sākumā notiks gandrīz garantēti, tad svaru vērtības tiek mainītas, tādējādi cenšoties aprēķināto vērtību iegūt pēc iespējas tuvāk sagaidāmajai. Šo darbību uzskata par mācīšanās procesu. Ir vairākas funkcijas, kas nosaka atšķirību starp reālo un aprēķināto vērtību sauktas par kļūdas funkcijām, kā arī funkcijas, kas nomaina svaru vērtības pareizā veidā sauktas par optimizācijas funkcijām. Mācīšanās procesa mērķis ir iegūt tādas svaru vērtības, lai starpība starp reālo un mācīšanās procesā aprēķināto gala vērtību būtu minimāla, kā arī, tālāk izmantojot šīs svaru vērtības citiem datiem, uz kuriem modelis nav trenējies, atšķirība joprojām būtu minimāla.

Dziļā mašīnmācīšanaās balstās uz vairākām funkcijām, lai tā spētu pilnvērtīgi darboties. Tiek izmantoti modeļi, kas būtībā balstās uz matemātiku. Ievades dati tiek pārveidoti skaitļos, un iznākums arī tiek attēlots skaitļos. Tiesa gan bieži gala vērtību skaitļi tiek pārveidoti par kādām klasēm vai ko vairāk, lai lietotājam būtu vieglāk uztvert to nozīmi. Apmācības procesā izmanto vairākas matemātiskas funkcijas, kā piemēram kļūdas funkcija, kas nosaka atšķirību starp minēto un reālo rezultātu un uz kuras modelis

balsta savu apmācību, optimizācijas funkcija, kas nosaka to, kā jāizmaina modeļa neironu tīklu iekšējie svari . Arī svarīgi ir izprast to, kas ir atpakaļizplatīšanās algoritms.

Kā jau iepriekš minēts, lai noteiktu atšķirību starp apmācības aprēķināto vērtību un reālo vērtību ir jāizmanto kāda funkcija. Šī funkcija ir kļūdas funkcija (Chollet, 2017), kas pēc kādas matemātiskas formulas, kas ir izvēlēta piemērota konkrētai situācijai, aprēķina cik tuvu prognozētais rezultāts ir reālajai vērtībai. Kļūdas funkcijas vērtības rezultātā tiek iegūts skaitlis, kas parāda precizitāti konkrētajā gadījumā, un šo rezultātu izmanto tālāk, lai veicinātu mācīšanos. Tā ir kā atgriezeniskā saite par to, cik vērtīgas ir bijušas izmaiņas un vai tās ir bijušas pareizajā virzienā. Uzsākot apmācību svaru vērtības tiek izvēlētas kā nejauši gadījuma skaitļi, tāpēc kļūdas funkcija varētu iznākt samērā liela, taču mainot šīs vērtības ar optimizācijas algoritmu, mērķis ir iegūt kļūdas funkcijas vērtību pēc iespējas tuvāku nullei, tādējādi panākot, ka prognozētā vērtība gandrīz vai sakrīt ar reālo.

Ir vairākas kļūdas funkcijas, kuras izmanto dažādām situācijām. Turpmāk tiks aprakstītas dažās no zināmākajām un to pielietojums.

Pirmā un vienkāršākā funkcija ir vidējā absolūtā kļūda (Mean absolute error (MAE)), attēlota 1.1. vienādojumā (Chollet, 2017), kas ir modulis starpībai starp reālo un aprēķināto vērtību. Šī funkcija ir tiešs attēlojums aprēķinātās un reālās vērtības starpībai.

$$L(MAE) = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{1.1}$$

1.1. vienādojumā un arī turpmākajos vienādojumos n apzīmē datu kopas lielumu, ŷ apzīmē aprēķināto izejas vērtību un y apzīmē reālo vērtību.

Līdzīgi kā iepriekšējā, tikai vairāk akcentējot pareizo un nepareizo, ir vidējās kvadrātiskās kļūdas (Mean squared error(MSE)) funkcija, parādīta 1.2. vienādojumā (Yathish, 2022).Funkcijas rezultātā vērtības, kuras ir bijušas zem viens jeb tuvāk nullei, kas nozīmē, ka starp aprēķināto un reālo vērtību ir mazāka atšķirība, tiek padarītas vēl mazākas, tādējādi veicinot šādu rezultātu. Taču vērtības, kuru rezultāts ir bijis augstāks par viens, tiek palielināts, tādējādi cenšoties šādu rezultātu parādīt kā sliktu.

$$L(MSE) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{1.2}$$

Funkcija, kuru izmanto ar bināriem datiem ir loģistikas kļūdas(logistic loss) funkcija jeb, kā to biežāk pazīst, krustentropijas zuduma (cross-entropy (CE) loss) funkcija, parādīta 1.3. vienādojumā (Yathish, 2022). Šī funkcija ir vairāk domāta varbūtības noteikšanai. Tā kā tiek izmantots logaritms, tas dod eksponenciāli zemākus rezultātus, ja vērtība ir tuvāk nullei un otrādi, ja tā ir tālāk no nulles. Salīdzinot ar iepriekšējo kļūdas funkciju, kura ir ļoti labi klasifikācijas uzdevumiem, šī funkcija ir labāk uzdevumiem, kur svarīgāka ir pareizās atbildes varbūtība.

$$L(CE) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \times \log(\hat{y}_i) + (1 - y_i) \times \log(1 - \hat{y}_i)] \tag{1.3}$$

Ja izvades dati vairs nav binārā vērtību kopā, bet ir vairākas klases, un joprojām ir aktuāli noteikt klases iestāšanās varbūtību, tad izmanto uz krustentropijas zuduma balstītu funkciju, kuru sauc par kategorijas krustentropijas zuduma (categorical cross-entropy (CCE) loss) funkciju, attēlota 1.4. vienādojumā (Yathish, 2022).

$$L(CCE) = -\sum_{i=1}^{n}\sum_{j=1}^{k}[y_{ij} \times \log(\hat{y}_{ij})] \tag{1.4}$$

Šīs ir zināmākās un biežāk izmantotās kļūdas funkcijas, taču to ir vēl ļoti daudz. Neskatoties uz to, ja ir nepieciešams, ir iespējams arī manuāli izveidot jaunu kļūdas funkciju, kuru izmantot modeļa apmācībā. Šāda situācija var rasties, ja neironu tīkls paliek ļoti sarežģīts, vai ir kādi specifiski nosacījumi. Tomēr ir nepieciešams izprast kā strādā kļūdas funkcijas un kādu rezultātu tās dod, lai būtu iespējams saprast, kā labāk modeli trenēt un līdz ar to, izveidot jaunu kļūdas funkciju, kas to spētu darīt pilnvērtīgi un spētu palīdzēt modeļa trenēšanā, un būtu iespējams iegūt precīzākus rezultātus.

Lai kļūdas funkcijas vērtību samazinātu ir nepieciešams pareizi izmainīt svaru vērtības, taču sākumā ir jāzina, cik lielu ietekmi gala rezultātam ir devuši kuri svari, lai zinātu, kas un par cik daudz ir jāmaina. Tas var arī būt ļoti sarežģīts process palielinoties neironu slāņu skaitam, jo svaru skaits palielinās par n + m ar katru pieliktu neironu (n - iepriekšējā slāņa neironu skaits, m - nākamā slāņa neironu skaits) un tas palielinās par k * (n + m) ar katru pielikto slāni (k - esošā slāņa neironu skaits). Tāpēc tiek izmantots atpakaļizplatīšanās algoritms (Aggarwal, 2018), kas risina tieši šo problēmu.

Atpakaļizplatīšanās algoritms (Chollet, 2017) izmanto pēdējo kļūdas vērtību un ar to iziet cauri visam neironu tīklam, katrai funkcijai, kas

izmantota uz neirona, pielietojot ķēdes likumu, lai aprēķinātu katra parametra ietekmi kļūdas funkcijas rezultāta iegūšanā un saprastu, uz kuru pusi - paaugstināt vai pazemināt - ir nepieciešams izmainīt attiecīgo vērtību. Ķēdes likums atvasināšanā izpaužas tā, ka, ja jāatvasina funkciju, kura sevī satur citu funkciju, tad rezultāts ir ārējās funkcijas atvasinājums sareizināts ar tās funkcijas atvasinājumu, kas atrodas iekšā. Labāk attēlots piemērs ir 1.5. vienādojumā.

$$\frac{\partial f(g(x))}{\partial x} = \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial x} \tag{1.5}$$

Lai algoritms būtu pilnvērtīgs, tas sastāv no divām fāzēm (Aggarwal, 2018): tā, kas iziet visu no sākuma līdz beigām un atpakaļgaitas, jo, sākumā, ar apmācībā iegūtajām svaru vērtībām, ir nepieciešams iziet cauri tīklam, lai iegūtu rezultātus, kurus pēc tam izmantotu atpakaļgaitas fāzē. Izejot visu no sākuma līdz galam, beigās tiek iegūta prognozētā vērtība, kuru izmanto kļūdas funkcijā kopā ar reālo vērtību, lai iegūtu kļūdas funkcijas vērtību.

Kopumā atpakaļizplatīšanās algoritms (Vasilev, Slater et al., 2019) aprēķina kļūdu pēdējā apslēptajā slāni un cenšas noteikt kāda tā būtu varējusi būt iepriekšējā slāni ar attiecīgajā piemērā izmantotajām vērtībām. Kļūdas vērtība tiek padota no pēdējā slāņa atpakaļgaitā līdz sākumam, tāpēc to sauc par atpakaļizplatīšanās algoritmu. Taču, lai veiksmīgāk spētu izmainīt svaru vērtības, tiek pielietots optimizācijas algoritms, kas aprakstīts tālāk.

Optimizācijas algoritmu izmanto (Chollet, 2017) lai pareizi izmainītu svaru vērtības, kas piesaistītas pie neironiem, tādā veidā, lai iegūtu mazāku kļūdas funkciju un labāk apmācītu modeli. Šis algoritms izmanto funkcijas, kas nosaka, kādā veidā tiks izmainītas svaru vērtības. Tāpat kā kļūdas funkcijas, arī optimizācijas algoritmi ir vairāki. Tālāk aprakstīti daži no tiem.

Optimizācijas algoritmi izmanto atpakaļizplatīšanās algoritmu, lai iegūtu svaru vērtības. Gradienta nolaišanas (Gradient descent) algoritmi (Ruder, 2017) ir visbiežāk izmantotie. Tie ir daudz un dažādi. Gradienta nolaišanas algoritmi tālāk iedalās trijās lielās kategorijās. Pirmā ir partijas gradienta nolaišana (Batch gradient descent), aplūkojams 1.6. vienādojumā, kas pēc atpakaļizplatīšanās algoritma aprēķina svaru vērtību izmaiņas uzreiz visai trenētajai partijai.

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta) \tag{1.6}$$

1.6. vienādojumā θ ir parametri, kuri jāizmaina, η ir mācīšanās solis un $\nabla_\theta J(\theta)$ ir aprēķinātās izmaināmās vērtības.

Otrais algoritms ir Stohastiskā gradienta nolaišana (Stochastic Gradient Descent (SGD)) (Ruder, 2017) kas ir visbiežāk izmantotais optimizācijas algoritms, redzams 1.7. vienādojumā. Atšķirībā no iepriekšējā, šis algoritms nevis uzreiz aprēķina visai partijai, bet gan aprēķina vērtības katram ievades ($x^{(i)}$) un izvades ($y^{(i)}$) lielumam.

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x^{(i)}; y^{(i)}) \tag{1.7}$$

Trešais gradienta nolaišanas algoritms ir mini-partijas gradienta nolaišanas (Mini-batch gradient descent) algoritms, aplūkojams 1.8. vienādojumā (Ruder, 2017). Šis algoritms ir kā apvienojums abiem iepriekšējiem, jo tas paņem nelielu partijas daļu un tai aprēķina svaru vērtību izmaiņas un tās implementē. Šis algoritms ir tas, ko visbiežāk izvēlās lietot apmācot neironu tīklus. Kaut arī visbiežāk kā optimizācijas algoritms tiek pieminēts SGD, parasti ar to tiek domāts šis algoritms.

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x^{(i:i+n)}; y^{(i:i+n)}) \tag{1.8}$$

Tomēr, arī mini-partijas gradienta nolaišanas algoritms saskaras ar problēmām, kā piemēram mācīšanās soļa pareiza izvēle, jo pārāk mazs mācīšanās solis radīs ļoti lēnu apmācību, taču pārāk liels var nekad nesasniegt vēlamos rezultātus. Tāpēc ir arī vairāki citi algoritmi, kas spēj sniegt nozīmīgākus rezultātus un ir kā risinājums problēmām.

1.2. Runas ierakstu apstrāde

Skaņa (Vasilev, Slater et al., 2019) būtībā ir vibrācijas, kas iet cauri kādai videi, un šo vibrāciju ir iespējams saglabāt pārveidojot uz digitālu signālu, ko tālāk izmantot audio apstrādē. Parasti kontinuālais signāls tiek diskretizēts, saglabājot tā amplitūdas izmaiņas laika gaitā jeb to mēdz arī saukt par attēlojumu laika domēnā. Bieži vien, pirms audio failu izmanto tālākai apmācībai, to kādā veidā pārveido, piemēram, par mela spektogrammu, kas parāda frekvences izmaiņu signālā laika gaitā. Mela spegktogramma tiek iegūta sadalot to audio signālu, kas ir attēlots laika domēnā, vairākās daļās, kas savstarpēji pārklājās, piemēram, daļās, kura katra ir 25 milisekundes un šo daļu paņem ik pa 10 milisekundēm, un tālāk šim daļām veic

Furjē transformāciju. Furjē transformācijas nodrošina signāla ieguvi frekvences domēnā. Iegūtie rezultāti tiek sadalīti pa 40 frekvenču kategorijām pēc logaritmiskā jeb Mela mēroga un to attēlojot ir iegūta Mela spektogramma.

Pēc audio priekšapstrādes procedūrām, lai sagatavotu to tālākai apstrādei, ir iespējams iegūt tekstu no audio. Runas atpazīšana (Vasilev, Slater et al., 2019) pēc audio skaņām cenšas atrast tos vārdu salikumus, kas audio skaņām atbildīs ar vislielāko varbūtību. Runas atpazīšanu veic ar trīsdaļīgiem vai viendaļīgiem modeļiem.

Trīsdaļīgie modeļi (Vasilev, Slater et al., 2019) sastāv no trim galvenajām daļām: audio signāla pazīmju atdalīšanas, fonēmu atdalīšanas un dekodēšanas. Fonēmas ir izrunātās skaņas. Angļu valodā ir iespējamas 44 dažādas fonēmas un to salikumi veido izrunātos vārdus.

Parasti audio sākotnēji tiek veikta pirmapstrāde (Vasilev, Slater et al., 2019), kur audio failam tiek izvilktas audio iezīmes. Tālāk caur akustisko modeli tiek atdalītas fonēmas. Atdalītās fonēmas tālāk ar dekodētāja palīdzību tiek sasaistītas kopā ar kādu rakstzīmi vai vārdu. Ja apmācības procesā tiek izmantots arī valodas modelis - atsevišķi trenējams modelis tieši uz valodu, kas var uzlabot rezultātus -, tad tālāk iegūtās rakstzīmes un vārdi tiek salīdzināti ar vārdu salikumu varbūtību no valodas modeļa, lai uzlabotu rezultātus. Un gala rezultātā tiek iegūti vārdi vai teikumi, kas ir atbilstoši attiecīgi ievadītā audio faila anotējums.

Viendaļīgie modeļi (Vasilev, Slater et al., 2019) saukti arī par modeļiem no viena līdz otram galam (end-to-end) veic to pašu uzdevumu ko trīsdaļīgie, tikai tas viss ir apvienots vienā modelī, nevis vairākos atsevišķos, līdz ar to, šāda veida modeļiem nav atsevišķi nodalīta fonēmu atdalīšana, un tie iemācās akustiskās audio iezīmes un uzreiz izvadē dod atbilstošās teksta daļas.

Zinātnē šobrīd ir vairāki runas atpazīšanas modeļi, kas efektīvi spēj pārveidot runu uz tekstu. Metrika ko izmanto, lai novērtētu šo modeļu izpildījumu ir WER (Word error rate (vārda kļūdas līmenis)) un CER (Character error rate (rakstzīmes kļūdas līmenis), kas salīdzina reālo teksta anotējumu ar to, kuru izdod modelis, izvērtējot vārdu vai simbolu kļūdas. Precīzāk – kļūdas tiek aprēķinātas (Leung, 2021) kā aizvietošanas, pievienošanas un dzēšanas kļūdu summa attiecībā pret vārdu vai simbolu skaitu teikumā. Aizvietošanas kļūda ir ja viens simbols vai vārds ir aizvietots ar citu. Ievietošanas kļūda ir, kad modeļa izdotajā tekstā ir ievietots kāds simbols vai vārds, kas nav bijis oriģinālā, un dzēšanas kļūda ir pretēja ievietošanas kļūdai, kad modeļa producētajā tekstā iztrūkst kāds vārds vai simbols, kurš ir bijis oriģinālajā.

Taču izmantojot šīs metrikas ir jāņem vērā, ka ar CER ir svarīgi vārdos lielie sākumburti, jo tās tiktu uzskatītas par atšķirīgām rakstzīmēm, vai arī tad ir jāapstrādā abi teksta fragmenti, visas rakstzīmes padarot par, piemēram, mazajiem burtiem. Taču vārdu kļūdas metrika ņem vērā tikai pareizi uzrakstītos vārdus, un, pat ja vārdā ir kāda kļūda vienā burtā, tas uzreiz tiek uzskatīts par nepareizu. It sevišķi svarīgi ir ņemt vērā to, ka mēdz būt vārdi, īpaši angļu valodā, kuri pēc skaņas izklausās vienādi, taču to rakstība ir atšķirīga, saukti arī par homofoniem, un tiklīdz šis vārds ir nepareizi detektēts un pierakstīts, WER metrika palielinās, jo kaut arī vārds ir bijis saprasts pareizi, tas tiek pierakstīts nepareizi. Kā arī WER neņem vērā pieturzīmes. Līdz ar to, pat ja pēc metrikām iegūtie rezultāti ir zem kāda zināma sliekšņa, tas nebūt uzreiz nenozīmē, ka modelis ir labs, un ka tas vienmēr uz visām atšķirīgajām situācijām spēs atkārtot šo labo rezultātu. Bet, neatkarīgi no tā, zems WER vai CER rezultāts ir nozīmīgs modeļa izvērtēšanai.

Vislabākais veids, kā salīdzināt esošos modeļus būtu izmantot objektīvas metrikas, kā piemēram iepriekš minētās WER vai CER un veikt eksperimentus uz kādas konkrētas datu kopas, taču diemžēl bieži vien tas netiek realizēts, jo ir daudzas datu kopas, katra ar savām īpatnībām, un katrs zinātniskais pētījums, cenšoties pierādīt arvien labākus rezultātus, izvēlās sev piemērotāko datu kopu. "Papers With Code" sniedz iespēju aplūkot iegūtos rezultātus uz konkrētām datu kopām uzreiz arī norādot attiecīgās publikācijas, kas šos rezultātus ir sasniegušas, kā arī to pieejamos modeļus. Ir iespējams pēc uzdevuma veida veikt filtrāciju un attiecīgi tika atlasīti "Speech recognition" jeb runas atpazīšanas uzdevumi[1]. Jāņem vērā, ka ne visas publikācijas ir ievietotas šajā vietnē, un ir iespējams, ka ir pētījumi ar augstākiem rezultātiem nekā attiecīgajā lapā ir norādīts. Modeļi, kas parādās visbiežāk ar vislabāko rezultātu ir wav2vec 2.0 (Baevski, Zhou et al., 2020), Whisper (Radford, Kim et al., 2022), Conformer (Gulati, Qin et al., 2020) un SpeechStew (Chan, Park et al., 2021). Kopīgā datu kopa kas ir visiem šiem modeļiem ir LibriSpeech, un iegūtie rezultāti ir attēloti 1.1. tabulā. Rezultāti, kas ir attēloti iekavās, izmanto papildus valodas modeli.

[1]https://paperswithcode.com/task/speech-recognition/latest

1.1. tabula.
Runas atpazīšanas modeļu WER rezultāti uz LibriSpeech datu kopas.

	LibriSpeech Clean	LibriSpeech Other
wav2vec 2.0	(2.0)	(4.0)
Whisper	2.7	5.2
Conformer	2.1(1.9)	4.3(3.9)
SpeechStew	1.7	3.3

SpeechStew[1] un Whisper[2] savus eksperimentus ir veicis arī uz vēl citām datu kopām, kuru rezultāti arī ir aplūkojami 1.2. tabulā. Ailīte "vieta" norāda uz "Papers With Code" pieejamo publikācijas WER rezultātu reitinga vietu.

1.2. tabula.
Whisper un SpeechStew WER salīdzinājums dažādās datu kopās.

Datu kopa	Whisper		SpeechStew	
	WER	Vieta	WER	Vieta
AMI IMH	16.4	2	9	1
AMI SDM1	36.9	2	21.7	1
Artie Bias Corpus	6.7	1	-	-
CALLHOME	15.8	1	-	-
CHiME6	25.6	1	-	-
CHiME-6 dev_gss12	-	-	31.9	1
CHiME-6 eval	-	-	38.9	1
Common Voice	9.5	1	10.8	2
CORAAL	19.4	1	-	-
Fleurs (English)	4.6	1	-	-
LibriSpeech test-clean	2.7	30	1.7/ 2.0	4 / 12
LibriSpeech test-other	5.6	25	4.0 / 3.3	11 / 5
Switchboard corpus	13.1	1	-	-
Switchboard CallHome	-	-	8.3	1
Switchboard SWBD	-	-	4.7	1
Tedlium	4	1	5.3	2
Vox Populi	7.3	1	-	-

[1]https://paperswithcode.com/paper/speechstew-simply-mix-all-available-speech
[2]https://paperswithcode.com/paper/robust-speech-recognition-via-large-scale-1

Datu kopa	Whisper		SpeechStew	
	WER	Vieta	WER	Vieta
WSJ	3.1	1	-	-
WSJ eval92	-	-	1.3	1

Kā redzams 1.1. tabulā, kurā iegūtie rezultāti ir ņemti no LibriSpeech datu kopas, augstāko rezultātu sniedz SpeechStew, kam tālāk seko wav2vec 2.0, tad Conformer, un beigās Whisper. Šeit gan jāņem vērā, ka LibriSpeech ir datu kopa, kura nesatur trokšņainus audio failus, kā arī pēc 1.2. tabulas, kurās ir arī norādīti šo modeļu vietas, var redzēt, ka ir vairāki citi augsti rezultāti, kas ir iegūti tieši uz šīs datu kopas. 1.2. tabulā ir arī aplūkojams, tas, ka WER ir virs 10 %, taču modeļu sniegtais rezultāts ir pirmajā vietā pēc globālā reitinga. Tas nozīmē, ka ir iespējami uzlabojumi.

Variants, kas varētu uzlabot runas atpazīšanas dotos rezultātus ir izmantot runas attīrīšanu (speech enhancement), kas ir kad audio failam, kuram ir kādi trokšņi, tie tiek noņemti ar audio apstrādes palīdzību, tādējādi visdrīzāk padarot audio failus tuvāk tam formātam, kādā tie tiki izmantoti apmācībā.

Taču šāda veida uzdevumiem ir grūtāk nomērīt rezultātus nekā runas atpazīšanas modeļiem, jo nav konkrēts rezultāts ar ko var salīdzināt to, cik labi ir noņemti trokšņi vai fona skaņas. Runas attīrīšanas rezultātiem izmanto vairākas metrikas, piemēram PESQ (perceptual evaluation of speech quality) (Rix, Beerends et al., 2001), kas noteiktā veidā salīdzina abus audio failus un norāda rezultātu audio faila kvalitātei diapazonā no -0.5 līdz 4.5, kur augstāks rezultāt nozīmē labāku kvalitāti, STOI (short-time objective intelligibility) (Lightning-AI, 2022), kas mēra specifiski attīrīta audio faila saprotamību diapazonā no 0 līdz 1 jeb dažreiz tas tiek attēlots kā no 0 % līdz 100 %, kur augstāks rezultāts nozīme labāk saprotamu audio failu, kā arī ir vēl daudzas citas metrikas, kuras ir iespējams izmantot runas attīrīšanas uzdevumos, jo ir vairāki veidi un īpašības kurām var pievērst uzmanību iegūstot rezultātus. Piemēram ir atšķirīgi veidi kā var uzlabot runas kvalitāti jeb attīrīt audio failus - visbiežāk tiek noņemti specifiski trokšņi, kas varētu radīt traucējumus, bet iespējams arī noņemt atbalsojumu, kas sadzirdams audio failā vai uzlabot audio kvalitāti, ja tas nav bijis pārāk labi saprotams.

Zināmi modeļi ar augstiem rezultātiem iepriekš norādītajās metrikās ir DEMUCS (Defossez, Synnaeve et al., 2020), TridentSE (Yin, Zhao et al., 2022) un CMGAN (Abdulatif, Cao et al., 2022). Visi modeļi ir veikuši savus rezultātus uz Voice Bank+DEMAND datu kopas (DEMUCS tā ir pie-

minēta kā Valentini, bet tas ir šīs datu kopas veidotājs[12]) DEMUCS labākie rezultāti ir 3.07 PESQ un 95 % (0.95) STOI. TridentSE labākie rezultāti ir 3.47 PESQ un 96 % (0.96) STOI. CMGAN modeļa labākie rezultāti ir 3.41 PESQ un 96 % (0.96) STOI. No šiem modeļiem, pēc PESQ metrikas labākais ir TridentSE, kam tālāk seko CMGAN un beigās ir DEMUCS, taču pēc STOI metrikas rezultāti ir ļoti līdzīgi, jo TridentSE un CMGAN abi ieguva 96 %, bet DEMUCS tikai par 1 % zemāku rezultātu.

Tomēr mēdz būt arī cilvēki, kas runā ar stipru akcentu vai izrunā vārdus nedaudz atšķirīgi un tas var sagādāt problēmas runas atpazīšanas modeļiem un šo problēmu runas attīrīšana nav spējīga atrisināt, tāpēc šim var būt labāk noderīgi runas stila pārveidošanas modeļi (speech style transfer), kad runātāja balss kaut kādā veidā tiek nomainīta uz citu. Šeit rezultātu efektivitāti nomērīt ir vēl sarežģītāk nekā runas attīrīšanā, jo rezultātu galamērķis ir vēl neskaidrāks un grūtāk nomērāms, kā arī ir ļoti daudz veidi, kā izmanto šos modeļus. Ir piemērs, kas pārveido parastu ierunātu balsi par dziedošu balsi (Agarwal, Ganapathy et al., 2022), vai pārnes emocijas (Zhou, Sisman et al., 2021), kur uzdevums nav objektīvi vērtējams, jo cilvēkiem dažreiz ir samērā grūti noteikt, ar kādu emociju otrs runā, kā arī cilvēki ir dažādi un atšķirīgi varētu izpaust prieku, bēdas, dusmas runājot. Plašāk zināms variants runas stila pārveidošanai ir, ka pārnes viena cilvēka balss stilu uz otru, liekot otrajam runātājam izklausīties kā pirmajam.

Bieži šādos eksperimentos izmanto subjektīvas metrikas, piemēram veicot aptauju noskaidrojot cilvēku viedokli par iegūtajiem rezultātiem un to, cik, viņuprāt, labi modelis ir izpildījis uzdevumu. No šī var izmantot metriku MOS (Mean Opinion Score), kas liekot cilvēkiem novērtēt iegūto rezultātu konkrētā skalā, bieži no 1 līdz 5, un beigās izrēķina vidējo vērtību. Objektīva metrika, kura ir izmantota varētu būt EER (Equal Error Rate) (Innovatrics, 2023), kas norāda cik daudz tiek nepareizi pieņemta pārveidotā runa īstās runas vietā, un noraidīta īstā runu pār pārveidoto. Šī metrika parasti tiek mērīta no 0 % līdz 50 %, kur zemāka vērtība norāda labāku rezultātu.

Esoši modeļi, kas ir rādījuši augstus sasniegumus runas stila pārveidošanā ir FreeVC (Li, Tu et al., 2022) vai DISSC (Maimon & Adi, 2022). FreeVC pārveidojot runas stilu izmantojot tikai apmācībā lietots datus sasniedza MOS 4.01, izmantojot tikai iepriekš apmācībā neizmantotus datus sasniedza MOS 4.06, un pārveidojot no datiem kas nebija apmācība uz tiem,

[1]https://paperswithcode.com/dataset/voice-bank-demand
[2]https://datashare.ed.ac.uk/handle/10283/2791

kas bija, sasniedza MOS 4.08. DISSC labākais rezultāts uz VCTK datu kopas ir 1.7 % EER, un uz ESD datu kopas ir 2.6 % EER.

2. SISTEMĀTISKĀ LITERATŪRAS ANALĪZE

Lai izvērtētu šobrīd esošos risinājumus, tika veikta sistemātiskā literatūras analīze. Tika apskatīti šobrīd esošie aktuālie audio apstrādes modeļi. Analīze tika veikta trijās kategorijās apskatot šobrīd aktuālo: audio pārveidošanai uz tekstu jeb runas atpazīšana, runas attīrīšana un runas stila pārnesei, kā arī vēl šo modeļu kombinācijas.

Sākumā tika apskatīti runas atpazīšanas modeļi, kas izmanto audio failu, kurā kāds runā, un cenšas iegūt tekstu no runātā. 1. pielikumā ir aplūkojama vispārēja informācija par aplūkoto modeļu publikācijām. 1. pielikums ietver informāciju par publikācijas nosaukumu, saiti uz pašu publikāciju, kā arī uz kodu, ja tāds ir pieejams, publikācijas izdošanas gadu, organizāciju kopā ar valsti, kā arī pēc "Semantic Scholars"[1] datiem citos darbos izmantotās atsauces uz šo publikāciju. Turpmāk aplūkojot detalizētāku informāciju par publikācijām jāņem vērā tabulas ailīte "Nr", jo tā katrai publikācijai ir unikāla.

Kā redzams pēc 1. pielikuma tika aplūkotas 10 runas atpazīšanas publikācijas, kuras publicētas starp 2019. un 2022. gadu. Četrām no publikācijām ir publiski pieejams kods.

[1] `https://www.semanticscholar.org/`

Tālāk 2.1. tabulā var aplūkot to, kādu uzdevumu veic modelis, kā arī galvenās metodes apkopojumu, jeb vispārēju informāciju par to, kas aprakstīts runas atpazīšanas publikācijā.

2.1. tabula.
Kopsavilkums par runas atpazīšanas publikācijām.

Nr	Uzdevums	Metode
1	Runas atpazīšana ar pieturzīmēm	Neizmanto atsevišķus modeļus lai veiktu atsevišķi runas atpazīšanu un pēc tam paredzētu pieturzīmes, bet to visu dara uzreiz vienā modelī, ņemot vērā runā ieturētās pauzes, toņu maiņas un tamlīdzīgas akustiskas iezīmes.
2	Runas atpazīšana, tulkošana un valodas noteikšana	Modelis, kas trenēts uz lielu daudzumu dažādu uzraudzīto datu, koncentrējoties uz tā sniegumiem nulles šāviena pārveidošanā, var ievērojami uzlabot pielāgojamību runas atpazīšanas sistēmās.
18	Runas atpazīšana	Jauna veida arhitektūras modelis, kas kombinē jau esošās transformatoru un konvolūcijas (CNN) arhitektūras.
19	Runas atpazīšana	Jaunāko sasniegumu modeļu arhitektūras kombinācijas, kā arī pašpārraudzītās apmācības izmantošana, cenšoties sasniegt pēc iespējas augstākus rezultātus uz LibriSpeech datu kopas.
20	Runas atpazīšana	Izveido un notestē divas dažādas modeļu pieejas runas atpazīšanai uz LibriSpeech datu kopas.
21	Reāllaika runas atpazīšana (Runas atpazīšanas straumēšana)	Uzlabo beigu aiztures un ģenerētā teksta kvalitāti reāllaika atkārtoto neironu tīklu pārveidotāja (RNN-T) balstītajā no viena līdz otram galam reāllaika runas atpazīšanas modelim.
22	Iezīmju izvilkšanas izmantošana kādam uzdevumam, piemēram, runas atpazīšanai	Vairāku uzdevumu pašpārraudzītās apmācības pieeja runas atpazīšanas modeļu veiktspējas uzlabošanai.
42	Runas atpazīšana	Izmanto vairākas populāras runas atpazīšanas datu kopas, sajaucot tās kopā, lai uztrenētu vienu lielu neironu tīklu priekš runas atpazīšanas modeļa.

Nr	Uzdevums	Metode
43	Runas atpazīšana	Izstrādāts jauns modelis HuBERT, kas veic runas atpazīšanu cenšoties paredzēt turpmākos ievades apslēptos segmentus ar K-tuvāko kaimiņu metodes grupēšanu.
44	Runas atpazīšana	Nepārraudzītā apmācība runas attēlojumam, kas maskē latentos attēlojumus viļņformā un izpilda kontrasta uzdevumus pār vairākiem runas attēlojumiem.

2.2. tabulā ir norādīta informācija par paša modeļa uzbūvi - tiek pieminēta modeļa arhitektūra vai arī kāda zināma arhitektūra uz kuru balstās modeļa arhitektūra, kā arī apmācībā izmantotā kļūdas funkcija, ar kuras palīdzību tiek trenēts modelis. Visbiežāk redzamās kļūdas funkcijas ir konnekcionistiskā laika klasifikācijas (CTC) un krosentropijas (CE) funkcijas, kā arī visbiežāk redzamās arhitektūras ir transformatora, vai konformera.

2.2. tabula.
Modeļa arhitektūra un kļūdas funkcijas runas atpazīšanas publikācijās.

Nr	Modeļa arhitektūra	Kļūdas funkcija
1	Saliktais transformatora iekodētājs	Konnekcionistiskā laika klasifikācijas (CTC) kļūda
2	Iekodētāja-dekodētāja transformators	Kategorijas krosentropijas (CCE) kļūda
18	Konformers: Transformators un konvolūciju neironu tīklu (CNN)	Nav pieminēta
19	Konformers	Starpība starp konteksta vektoriem no maskētajām un mērķa pozīcijām.
20	Hibrīda dziļo neironu tīkls (DNN), slēptā Markova modelis (HMM) un uzmanībā balstīts iekodētājs-dekodētājs.	Krosentropijas (CE) kļūda
21	Konformera atkārtoto neironu tīklu pārveidotājs (RNN-T)	Nav pieminēta

Nr	Modeļa arhitektūra	Kļūdas funkcija
22	PASE+: Sinc konvolūciju neironu tīkls (SincNet), konvolūciju neironu tīkls (CNN), Kvazi atkārtoto neironu tīkls (QRNN)	Vidējā katra darba vienības (worker) kļūda, vidējā kvadrātiskā kļūda (MSE).
42	Konformera atkārtoto neironu tīklu pārveidotājs (RNN-T)	Nav pieminēta
43	Vairāku slāņu konvolūciju viļņformu iezīmju iekodētājs, transformators	Krosentropijas (CE) kļūdas funkcija. Konnekcionistiskā laika klasifikācijas (CTC) kļūdas funkcija.
44	Vairāku slāņu konvolūciju iezīmju iekodētājs, transformators	Kontrasta kļūda, dažādības/daudzveidības (diversity) kļūda, konnekcionistiskā laika klasifikācijas (CTC) kļūda

2.3. tabulā aplūkojamas izmantotas datu kopas, kā arī iegūtie rezultāti no veiktajiem uzdevumiem saistībā ar runas atpazīšanu. Ja rezultāti ir ierakstīti iekavās, tad tiek izmantots papildus valodas modelis, ja vienīgi nav norādīts citādi. Tiek pieminēta arī metrika PER, kas iepriekš nav apskatīta, kas ir fonēmu kļūdas daudzums (Phone error rate). Visos pētījumos tika izmantota WER metrika un viszemākais tās rezultāts ir 1.3 % un tas ir sastopams 19. publikācijā ar LibriSpeech dev datu kopu bez valodas modeļa un LibriSpeech dev un test datu kopās izmantojot valodas modeli, kā arī 42. publikācijā uz WSJ eval92 datu kopas.

2.3. tabula.
Runas atpazīšanas modeļiem izmantotās datu kopas un iegūtie rezultāti.

Nr	Datu kopas	Metrikas un Rezultāti
1	MuST-C angļu valodai un JCALL japāņu.	JCALL CER: 14.3, MuST-C WER: 27.4, JCALL Punctuation F1: 67.4, MuST-C Punctuation F1: 75.9%

Nr	Datu kopas	Metrikas un Rezultāti
2	680 000 stundas ar audio no dažādām datu kopām kopā ar atbilstošo tekstu, kas paņemts no interneta.	WER LibriSpeech test-clean : 2.7; Artie: 6.7; Fleurs (English): 4.6; Common Voice: 9.5; Tedlium: 4.0; CHiME6: 25.6; WSJ: 3.1; VoxPopuli (English): 7.3; AMI-IHM: 16.4; CallHome: 15.8; Switchboard: 13.1; CORAAL: 19.4; AMI-SDM1: 36.9; LibriSpeech test-other: 5.6
18	LibriSpeech	WER LibriSpeech testclean/ testother S modelis: 2.7 / 6.3 (2.1 / 5.0); M modelis: 2.3 / 5.0 (2.0 / 4.3); L modelis: 2.1 /4.3 (1.9 / 3.9)
19	LibriSpeech, Libri-Light	LibriSpeech WER dev / dev-other / test / test-other : 1.3 / 2.7 / 1.5 / 2.7 (1.3 / 2.6 / 1.3 / 2.6)
20	LibriSpeech	Labākais LibriSpeech WER dev / dev-other / test / test-other : 1.9 / 4.5 / 2.3 / 5.0
21	Vairāku veidu balss izteikumi iegūti no balss meklēšanas, telefonsarunām un YouTube, kopā ar atbilstošo tekstu.	Labākais WER: 4.8; Labākais WER kopā ar beigu aizturi(EP): 5.3 / 290ms (EP50)/ 660ms (EP90)
22	LibriSpeech, DIRHA, CHiME-5, TIMIT, WSJ	TIMT PER clean / rev+noise: 17.2 / 33.8; DIRHA WER rev+noise: 27.4; CHiME-5 WER dev/ eval: 73.3 / 65.0
42	AMI, Broadcast News, Common Voice, LibriSpeech, Switchboard/Fisher, Tedlium, Wall Street Journal, CHiME-6.	Labākie WER : AMI HM / SDM1: 9.0 / 21.7; Common Voice (noņemot pieturzīmes): 10.8(8.4); LibriSpeech clean / other: 1.7 / 3.3; Switchboard/Fisher SWBD / CH: 4.7 / 8.3; Tedlium 5.3; WSJ eval92: 1.3
43	LibriSpeech, Libri-Light	Ar 100 stundu marķējumiem valodas modelim LibriSpeech WER dev-clean/ dev-other/ test-clean/ test-other: 1.7 / 3.0 / 1.9 / 3.5; Ar visām 960 stundām LibriSpeech : 1.5 / 2.5 / 1.8 / 2.9

Nr	Datu kopas	Metrikas un Rezultāti
44	LibriSpeech, Libri-Light, TIMT	WER LibriSpeech dev-clean / dev-other / test-clean / test-other: 1.6 / 3.0 / 1.8 / 3.3

2.4. tabula apraksta norādītos tālāk veicamos pētījumus, kurus ir nepieciešams pārbaudīt vai arī kas iespējams spēs uzlabot iegūtos rezultātus. Uz attiecīgā modeļa izmēģināt vēl citus uzdevumus norāda 2., 22. un 43. publikācijā, kā arī 1. un 21. publikācijā piemin to, ka nepieciešams sistēmu ieviest reāllaika runas atpazīšanā. Vēl 19. un 21. publikācijā iesaka koncentrēties uz skaitļošanas resursu samazināšanu, bet izmēģināt citas modeļa arhitektūras iesaka 1. un 44. publikācijā.

2.4. tabula.
Runas atpazīšanas publikācijās pieminētie vai ieteicamie turpmākie pētījumi.

Nr	Tālākie pētījumi
1	Izpētīt iespējamos uzlabojumus piedāvātajam modelim izmantojot citas modeļa arhitektūras, kā arī pielietojumu reāllaika runas atpazīšanai.
2	Uzlabot dekodēšanas stratēģiju; iegūt vairāk datu, lai varētu trenēt uz mazāk zināmām valodām; izpētīt precizēšanu(fine-tuning); izpētīt valodas modeļa ietekmi attiecībā uz noturību; pievienot papildu mācību uzdevumus.
18	-
19	Izpētīt konformera veiktspējas uzlabojumus saistībā ar kvantēšanas un dažādības kļūdas funkcijas uzstādījumiem.
20	-
21	Samazināt nepieciešamos skaitļošanas izmaksas, lai spētu arī labāko modeļa variantu pielietot reāllaikā.
22	Izpētīt pielietojumu citos uzdevumos, kā piemēram, runātāju, emociju un valodas atpazīšanu, kā arī secība-uz-secība (sequence-to-sequence) runas atpazīšanu.
42	Turpināt izpētīt pārneses mācīšanos runas atpazīšanā.
43	Uzlabot modeli, lai tā apmācības daļa sastāvētu no vienas fāzes. Kā arī izmantot šo modeli citiem atpazīšanas un ģenerēšanas uzdevumiem arī ārpus runas atpazīšanas.

Nr	Tālākie pētījumi
44	Iegūt augstākus rezultātus arhitektūru nomainot uz secība-uz-secība, kā arī izmantot vārdu krājumu ar daļējiem nevis pilniem vārdu gabaliem.

Tālāk tiek aprakstīti runas attīrīšanas modeļi, kas veic izmaiņas audio failā, tam noņemot fona trokšņus vai citus traucējumus, tādā veidā padarot šo failu tīrāku. 2. pielikumā ir aplūkojama vispārēja informācija par aplūkoto modeļu publikācijām. 2. pielikums satur 9 publikācijas, 6 no tām ir publiski pieejams kods, laika posmā no 2019. līdz 2022. gadam.

2.5. tabulā var aplūkot galvenās metodes apkopojumu, kā arī veicamo uzdevumu, bet tas pārsvarā ir runas attīrīšanas uzdevums, izņemot 7. publikācijā, kur tas ir norādīts kā trokšņu samazināšana.

2.5. tabula.
Kopsavilkums par runas attīrīšanas publikācijām.

Nr	Uzdevums	Metode
7	Trokšņu samazināšana	Jauna metode, kurā izmanto video informāciju ar redzamu trokšņu avotu bez runātāja, klasificējot troksni, lai samazinātu trokšņu esamību audio failā.
4	Runas attīrīšana	Izmanto DEMUCS arhitektūru, pielietojot to runas uzlabošanai, ir iespējams iegūt labus rezultātus reāllaikā neizmantojot lielus skaitļošanas resursus.
10	Runas attīrīšana	Jauna sistēma, kas veic runas attīrīšanu kompleksajā domēnā. Sastāv no "glance" un "gaze" modeļiem, kur "glance" noslāpē troksni amplitūdas (magnitude) domēnā, un "gaze" cenšās atgūt zaudētās spektra detaļas kompleksajā domēnā.
14	Runas attīrīšana	Kolaboratīvs modelis, izmantojot secība-uz-secību kartēšanu un piedāvā jaunu runas uzlabošanas sistēmu kompleksajā domēnā, izmantojot paralēlo topoloģiju kopā ar maskas bloku.

Nr	Uzdevums	Metode
15	Runas attīrīšana	Veic runas uzlabošanu, izmantojot jaunu arhitektūru TridentSE, kas izmanto struktūru ar vienu galveno un 2 pavadoņa zariem. Galvenajā zarā tiek saglabāta pilna spektogrammas izšķirtspēja, lai iegūtu sīkas detaļas katrā T-F (laika-frekvences) apgabalā, savukārt abi pavadoņa atzari izmanto kopā 32 marķierus lai efektīvi apstrādātu globālo informāciju.
24	Runas attīrīšana	Ģeneratīva pieeja bojātu audio atjaunošanai tīrā versijā izmantojot ģeneratīvo pretrunu tīklu (GAN) uz neapstrādātā audio.
29	Runas attīrīšana	Piedāvā jaunu modeli TSEGAN, kas ir kā ģeneratīvo pretrunu tīklu (GAN) paplašinājums laika domēnā ar metriku novērtēšanu, lai mazinātu mērogošanas problēmu, kā arī nodrošina modeļa apmācību stabilitāti, iegūstot uzlabojumus tā veiktspējā. Kā arī izpēta kā mainās modeļu rezultāti izmantojot dažādas kļūdas funkcijas.
30	Runas attīrīšana	DB-AIAT divzaru transformatorā balstīts modelis, kas paralēli nosaka amplitūdas spektru tīrajai runai un atjauno pazaudētās spektra daļas kompleksajā spektrā.
31	Runas attīrīšana	Jauns CMGAN modelis runas uzlabošanai vairākos uzdevumos: trokšņu noņemšana, atbalss noņemšana un izšķirtspējas uzlabošana.

2.6. tabulā ir norādīta informācija par modeļa arhitektūru, kā arī apmācības procesā izmantotās kļūdas funkcijas. Vairākas reizes tiek pieminēta vidējā kvadrātiskā kļūda (MSE). Arhitektūru ziņā bieži tiek izmantoti ģeneratīvo pretrunu tīklu (GAN) arhitektūrām līdzīgi modeļi, kā arī modeļi ar iekodētāju un dekodētāju.

2.6. tabula.

Modeļa arhitektūra un kļūdas funkcijas runas attīrīšanas publikācijās.

Nr	Modeļa arhitektūra	Kļūdas funkcija
7	Konvolūciju atkārtots tīkls (CRN)	Svērtā īstermiņa Furjē transformācijas (STFT) kļudas funkcija, mēroga invariantu signāla kropļojumu attiecības (SI-SDR) kļudas funkcija, binārā krosentropijas kļūdas (BCE) funkcija.
4	DEMUCS arhitektūra: daudzslāņu konvolūciju iekodētājs un dekodētājs ar U-neta izlaišanas savienojumiem, adaptētiem runas uzlabošanai	Regresijas kļūdas funkcija, spektogrammas domēna kļūdas funkcija, daudz izšķirtspējas svērtā īstermiņa Furjē transformācijas (multi-resolution STFT) kļūdas funkcija
10	GaGNet: iezīmju atdalīšanas modulis, kuram iekšā atrodas pārkalibrēti iekodētāja slāņi (REL) un beigās ir konvolūcijas slānis, kā arī sakrautie glance-gaze moduļi, kuriem iekšā ir sadarbības rekonstrukcijas modelis (CRM)	Svērtā daudz mērķu vidējā kvadrātiskā (MSE) kļūda
14	Iezīmju izvilkšanas bloks (FEB), kas izmanto U-Net bloku, lai izvilktu abstraktās iezīmes, komplekso vērtību uzlabojuma bloks (ComEB), kompensācijas bloks (CB) un maskas bloks (MB), kas sastāv no iekodētāja, dekodētāja un slēgtajām atlikuma vienībām (GRU).	Vidējā absolūtā kļūda (MAE) un mērogā nemainīga signāla-trokšņu attiecība (SI-SNR)
15	TridentSE: iekodētājs, dekodētājs, tridenta bloki.	Jaudā saspiestā amplitūdas vidējā kvadrātiskā kļūda (power-compressed amplitude MSE) un fāzē apzināta vidējā kvadrātiskā kļūda (phase-aware MSE), vidējā kvadrātiskā kļūda (MSE), diskriminatora kļūda, kas paredz PESQ rezultātus.

Nr	Modeļa arhitektūra	Kļūdas funkcija
24	SEGAN, kas ir balstīts uz ģeneratīvo pretrunu tīklu (GAN) arhitektūras.	Mazākā kvadrāta funkciju un izvades lineārā vienība.
29	TSEGAN: TasNet kā ģenerators un jaunizveidots metriku novērtēšanas diskriminators.	Mērogā nemainīga signāla-trokšņu attiecības (SI-SNR) kļūdas funkcija, vidējā kvadrātiskā kļūda (MSE), metrikas novērtējumā balstītā kļūda.
30	DB-AIAT: balstīts uz transformatora arhitektūru un sastāv no diviem atzariem - amplitūdas spektra maskēšanas zars (MMB) un kompleksais spektra atjaunošanas zars (CRB).	kļūdas funkcija attiecībā pret amplitūdas un komplekso spektru.
31	CMGAN: konformerā balstīts metrikas ģeneratīvo pretrunu tīkls (GAN) priekš runas uzlabošanas laika-frekvences domēnā.	Lineāra kombinācija amplitūdas kļūdai un kompleksajai kļūdai T-F (laika-frekvences) domēnā. Diskriminatora un ģeneratora vidējā kvadrātiskā kļūda (MSE), Vidējā absolūtā kļūda (MAE) laikam.

2.7. tabulā aplūkojami rezultāti pēc runas attīrīšanas uzdevumiem, kā arī publikācijās izmantotās datu kopas. Visvairāk izmantotās datu kopas šajā kategorijā ir DNS un VoiceBank + DEMAND (VBD). Visās publikācijās rezultāts tiek mērīts ar PESQ metriku, un vislabākais rezultāts tai ir 10. publikācijai uz DNS datu kopas 3.56, bet uz VBD datu kopas vislabākais ir 15. publikācijai 3.47. Otra metrika, kas arī atkārtojas lielākajā daļā publikāciju ir STOI, tā vislabākā ir 97.86 15.publikācijā uz DNS datu kopas. Uz VBD datu kopas labākais rezultāts ir 96 STOI un tāds rezultāts ir gan 15. publikācijai, gan arī 31. publikācijai.

2.7. tabula.
Runas attīrīšanas modeļiem izmantotās datu kopas un iegūtie rezultāti.

Nr	Datu kopas	Metrikas un Rezultāti
7	Audioset, DNS	Audioset PESQ: 2.6, STOI: 0.92, VISQOL-A: 4.45, VISQOL-S: 2.95

Nr	Datu kopas	Metrikas un Rezultāti
4	DNS, Voice Bank+DEMAND (VBD)	Labākie VBD PESQ: 3.07, STOI: 95, CSIG: 4.31, CBAK: 3.4, COVL: 3.63, MOS SIG: 4.18, MOS BAK: 3.69, MOS OVL: 3.72
10	WSJ0-SI84, DNS-Challenge, Voicebank+Demand (VBD).	WSJ0-SI84 avg PESQ: 2.93, avg ESTOI: 78.37 %, avg SDR: 12.08 dB; DNS-Challenge WB-PESQ: 3.17, PESQ: 3.56, STOI: 97.13 %, SI-SDR: 18.91 dB; VBD WB-PESQ: 2.94, CSIG: 4.26, CBAK: 3.45, COVL: 3.59
14	Librispeech, DNS-Challenge	LibriSpeech avg PESQ: 2.79, avg ESTOI: 76.29
15	VoiceBank+DEMAND (VBD), large-scale DNS	VBD PESQ 3.47, CSIG: 4.7, CBAK: 3.81, COVL: 4.1, STOI: 96 DNS no-reverb/with_reverb test set PESQ: 3.44 / 3.5, STOI: 97.86 / 95.22
24	VCTK Corpus, Demand	VCTK + Demand CSIG: 3.66, CBAK: 2.97, COVL: 3.00, PESQ: 2.42, SSNR: 7.05, STOI: 0.75
29	Voice-Bank + DEMAND (VBD)	VBD SI-SNR: 19.26, SSNR: 10.16, COVL: 3.1, CBAK: 3.27, CSIG: 3.73, PESQ: 2.52
30	Voice-Bank + DEMAND (VBD)	VBD PESQ: 3.31, STOI: 95.6, CSIG: 4.61, CBAK: 3.75, COVL: 3.96, SSNR: 10.79
31	Voice Bank+DEMAND (VBD), REVERB challenge dataset, VCTK	VBD PESQ: 3.41, CSIG: 4.63, STOI: 0.96; REVERB avg. tuvu mikrofons / tālu mikrofons CD: 1.96 / 2.54; LLR: 0.22 / 0.37, FWSegSNR: 13.18 / 10.3, SRMR: 5.77/ 5.55, SRMR-real: 7.71 / 7.62; VCTK viens / vairāki runātāji (pārtveršanas mērogs = 2) LSDe: 1.4 / 1.3, LSD10: 0.6 / 0.6, SNR: 24.7 / 24.4

2.8. tabula norāda publikācijās pieminētos iespējamos uzlabojumus, kur tādi tika norādīti tikai četrās publikācijās.

2.8. tabula.

Runas attīrīšanas publikācijās pieminētie vai ieteicamie turpmākie pētījumi.

Nr	Tālākie pētījumi
7	-
4	-
10	-
14	Padarīt modeļa darbību ātrāku, kā arī samazināt izmantojamo parametru skaitu. Izmantot paralēlo daudzpakāpju stratēģiju dažādiem blokiem, lai atvieglotu spektra attīrīšanu.
15	Izstrādāt parasto versiju šai arhitektūrai, lai to varētu izmantot zemas aizkaves reāllaika uzdevumiem.
24	Uzlabot modeļa arhitektūru, it īpaši iekodētāja struktūru, lai iegūtu labāku desmitdaļas noņemšanas shēmu.
29	Izpētīt citus ģeneratīvo pretrunu tīklos (GAN) balstītus modeļus pamatojoties uz šo pētījumu.
30	-
31	-

Turpmāk aprakstīti runas stila pārneses modeļi, kas veic izmaiņas audio failā, kaut kādā veidā mainot runātāja balss stilu tam uzliekot cita runātāja balsi, citas emocijas vai citādā veidā to izmainot. 3. pielikumā ir aplūkojama vispārēja informācija par 15 runas stila pārneses publikācijām no 2018. līdz 2022. gadam. 11 no šīm publikācijām ir publiski pieejams kods.

2.9. tabulā var aplūkot to, kādu stila pārneses uzdevumu veic balss stila pārneses modelis, kā arī vispārēju informāciju par to, ko publikācijā aprakstītais modelis dara. Visvairāk atkārtojas viena runātāja balss stila pārnese uz cita, kas tiek veikta 10 publikācijās.

2.9. tabula.
Kopsavilkums par runas stila pārneses publikācijām.

Nr	Uzdevums	Metode
5	Runas stila pārveidošana no parastas ierunātas balss par dziedošu balsi	Jauna modeļa arhitektūra, nosaukta par SymNet, kas apvieno ievades datus, kas ir parastā balss, ar mērķa melodiju, vienlaicīgi saglabājot runātāja identitāti, beigās izveidojot audio fragmentu, kurā izklausās, ka ievades runātājs būtu dziedājis pēc mērķa melodijas.
13	Dziedošas balss pārnese, tembra pārnešana	WORLD sintezators, kas pārveido audio stilu, piemēram, dziedošas balss pārveidošana un diferencējama digitālā signāla apstrādes (DDSP) tembra pārnešana.
25	Balss pārveidošana no čukstiem uz runāšanu	Pielāgojot runas uzlabošanas modeli ar ģeneratīvo pretrunu tīklu (GAN) arhitektūru, izstrādāja jaunu modeli, kurš, neatkarīgi no runātāja pārveido audio no čukstiem uz parastu normāla skaļuma balsi.
28	Globālās stila iezīmes, kas palīdz kontrolēt runas stilu vai ātrumu, neatkarīgi no teksta	Izveidoja globālā stila iezīmes (GST), kā palīglīdzekli izmantojot runas stila modelēšanai, runu ģenerējot no teksta. Spēj arī pārnest audio stilu uz garāku teksta fragmentu. Trenējot uz trokšņainiem audio, GST iemācās atpazīt runātāju un trokšņus.
32	Emociju pārnese	No vienas uz vairākām emocijas stila pārneses ietvars, kas balstīts uz VAW-GAN bez paralēlo datu nepieciešamības.
33	Viena runātāja balss pārnese uz otru runātāju vai mūzikas stila pārnese uz citu mūzikas fragmentu.	Audio stila pārnese, neizmantojot paralēlus datus.
34	Runas stila pārnese un runas ģenerēšana	Jauna runas stila pārneses metode, kas balstās uz neironu stila pārneses modeli, kas izmanto mela spektogrammas.
35	Runas stila pārnese, pielāgojoties arī mērķa runātāja runāšanas stilam	Vienkāršs un efektīvs veids runas stila pārnesei balstīts uz diskrētiem pašpārraudzītās apmācības runas attēlojumiem.

Nr	Uzdevums	Metode
36	Runas stila pārnese	Veic runas stila pārnesi, pārveidojot runātāju iegultnes vektoru un balss augstumu, tā, ka beigās ir balss, kurai nevar noteikt dzimumu.
37	Runas stila un runātāja identitātes pārnese	Viena šāviena balss stila pārneses modelis, tikai audio formātā, bez teksta.
38	Runas stila pārnese	Izmanto mīkstās runas vienības (soft speech units), lai uzlabotu nepārraudzīto balss stila pārnesi.
39	Audio stila pārnese	Uzliek mērķa stilu no viena audio uz cita, tādējādi automatizējot audio veidošanu, kā arī radot iepriekš definētus audio sākotnējos iestatījumus turpmākai izmantošanai. Iespējams izmantot gan runas fragmentiem, gan arī mūzikas.
40	Runas stila pārnese	Izpēta vairākas pašpārraudzītās apmācības metodes, lai uzlabotu balss pārnesi un iegūtu balss pārneses modeli, kas to spēs darīt, gan uz apmācībā izmantotiem, gan neizmantotiem datiem.
41	Runas stila pārnese	AutoVC: modelis, kas iegūst konkurētspējīgus rezultātus gan uz apmācībā izmantotiem, gan neizmantotiem datiem runas stila pārnesē, neizmantojot paralēlus datus, kā arī pirmais modelis, kas piedāvā nulles šāviena balss stila pārnesi.
45	Runas stila pārnese	Runas stila pārnese, kas neizmanto paralēlus datus, un spēj pārnest runas stilu neatkarīgi no tā vai runātāja balss ir bijusi vai nav bijusi apmācības datos.

2.10. tabulā ir norādīta informācija par modeļa arhitektūru un kļūdas funkcijām, kuras modelis izmanto apmācības procesā. Tā kā stila pārneses uzdevumi cenšas izveidot jaunu stilu, visbiežāk izmantota tiek ģeneratīvo pretrunu tīklu (GAN) arhitektūra, vai arhitektūra, kas uz to balstās, līdz ar to, arī kļūdas funkcija attiecīgi tad būs pretrunīgās kļūdas funkcija vai rekonstrukcijas kļūdas funkcija, kā arī vairākas reizes tiek pieminēta vidējā kvadrātiskā kļūda (MSE).

2.10. tabula.
Modeļa arhitektūra un kļūdas funkcijas runas stila pārneses publikācijās.

Nr	Modeļa arhitektūra	Kļūdas funkcija
5	SymNet: konvolūciju neironu tīkls (CNN), transformators un pašuzmanības (self-attention) slānis	Pretrunīgā kļūdas funkcija (adversarial loss), BEGAN kļūdas funkcija, vidējā kvadrātiskā kļūda (MSE).
13	Iekodētājs, dekodētājs, vokoderis	Vidējā kvadrātiskā kļūda (MSE), multi spektogrammas kļūda (MSL), pretrunīgā kļūda (adversarial loss (AL))
25	Pielāgots SEGAN (runas attīrīšanas ģeneratīvo pretrunu tīkls (GAN))	Pretrunīgā kļūdas funkcija (AL) un vidējā kvadrātiskā kļūda (MSE).
28	Balstīts uz Tacotron. References enkoderis, stila uzmanības, stila iegulšanas un secība-uz-secību (Tacatron) modelis.	Rekonstrukcijas kļūda (reconstruction loss (RL))
32	VAW-GAN ar dekoderi uz emociju iezīmēm.	Pretrunīgā kļūdas funkcija (AL)
33	Ģeneratīvo pretrunu tīkls (GAN)	Pretrunīgā kļūdas funkcija (AL), transformatora vektora mācīšanās (TraVeL) kļūda
34	Variācijas auto iekodētājs (VAE) un ģeneratīvo pretrunu tīklu (GAN) kombinācija, kam seko WaveNet bāzēts vokoderis.	VAE kļūda, GAN kļūda, latentā telpas kļūda, cikla konsistences (CC) kļūda
35	HiFi-GAN: ģenerators un diskriminatori.	Vidējā kvadrātiskā kļūda (MSE), binārā krosentropijas kļūda (BCE).
36	HiFiGAN, HuBERT, ECAPA-TDNN.	Nav pieminēta.

Nr	Modeļa arhitektūra	Kļūdas funkcija
37	Nosacījumu Variācijas auto iekodētājs (CVAE) apvienojums ar ģeneratīvo pretrunu tīklu (GAN) apmācību.	Rekonstrukcijas kļūda (RL), pretrunīgā kļūda (AL), iezīmju atbilstības kļūda.
38	Kontrastīvi paredzamā kodēšana (CPC) un HuBERT.	Pretrunīgā kļūda (AL).
39	WaveNet	Vidējā absolūtā kļūda (MAE), vairāku izšķirtspēju īsā laika Furjē transformācijas kļūda (MR-STFT).
40	FragmentVC kopā ar vairākām pašpārraudzītās apmācības metodēm (APC, CPC, wav2vec 2.0).	Pretrunīgā kļūda (AL).
41	AutoVC: runātāja iekodētājs, satura iekodētājs un dekoderis.	Pašrekonstrukcijas kļūda (self-RL).
45	StarGAN: ģeneratīvo pretrunu tīklu (GAN) arhitektūras paveids, izmantojot konvolūcijas neironu tīklu (CNN).	Pretrunīgā kļūdas funkcija (AL), domēna klasifikācijas kļūda, cikla konsistences kļūda, identitātes kartēšanas kļūda.

2.11. tabulā aplūkojamas izmantotas datu kopas, kā arī iegūtie rezultāti no veiktajiem uzdevumiem saistībā ar runas stila pārnesi. Šīm publikācijām rezultātu pierakstīšana ir samērā atšķirīga, kā arī izvēlētas datu kopas ir dažādas, jo veicamie uzdevumi ne vienmēr ir vienādi, jo runas stilu var censties pārveidot dažādos veidos. Metrikas, kas atkārtojas vairākās publikācijās ir EER un MOS un to labākie rezultāti ir 1.7 EER 35. publikācijā uz VCTK datu kopas un 4.15 MOS 38. publikācijā. 2.11. tabulā izmantoti vairāki saīsinājumi, lai rezultātu pieraksts būtu īsāks: pie emociju pārneses uzdevumiem izmanto N - neitrāla, A - dusmīga, S - bēdīga, citiem uzdevumiem F - sievietes balss, M - vīrieša balss, kā arī u - apmācība neizmantoti dati/balss, s - apmācībā izmantoti dati/balss.

2.11. tabula.
Runas stila pārnese modeļiem izmantotās datu kopas un iegūtie rezultāti.

Nr	Datu kopas	Metrikas un Rezultāti
5	NUS, NHSS, LibriSpeech	NUS LSD: 8.63, NHSS LSD: 8.67, izveidoto datu MOS: 3.08 +/- 0.09
13	TIMT	TIMT Error SVC / DDSP-TT: 0.0421 / 0.2465
25	Pašu veidots izmantojot CMU Artic datus	Histogrammas ar balss augstuma (pitch) vērtībām, subjektīva audio klausīšanās novērtēšana
28	147 stundas audio grāmatu datu angļu valodā. 439 Ted YouTube kanāla video. Mākslīgais troksnis.	MOS 50% / 75 % / 90 % / 95 % troksnim: 4.08 / 3.993 / 4.031 / 3.997; WER: 18.68
32	Jauna daudzvalodu un vairāku runātāju emociju datu kopa (ESD)	Objektīvie: MCD vīrietim N2H: 4.569, N2S: 4.127, N2A: 4.564; sievietei N2H: 4.26, N2S: 4.916, N2A: 4.451; Subjektīvie: MOS N2H / N2S / N2A: 3.24 / 2.94 / 3.15
33	ARCTIC, Donalda Trampa runas no YouTube, GTZAN	Rezultāti nav pierakstīti mērāmā veidā, ir tikai audio paraugi
34	Flickr8k Audio Caption Corpus, ASVspoof 2019	ASVspoof 2019 EER: 9.57, Flickr8k test split EER: 38.89; Oriģinālo piemēru WER: 2.01, mākslīgi radīto WER: 10.36
35	VCTK, ESD, Syn_VCTK	VCTK WER: 11.7, CER: 5.9, EMD: 10.53, TLE: 0.832, WLE: 0.056, PLE: 0.023, EER: 1.7; ESD WER: 23.2, CER: 10.3, EMD: 24.8, TLE: 0.35, WLE: 0.076, PLE: 0.037, EER: 2.6

Nr	Datu kopas	Metrikas un Rezultāti
36	LibriTTS, VoxCeleb2	Dzimuma noteikšana neinformēts EER: 28.99 %, daļēji-informēts EER: 24.13 %, WER: 4.81 %; runātāja noteikšana F / M / FM EER: 13.22 % / 9.85 % / 11.55 %
37	VCTK, LibriTTS	s2s MOS / SMOS: 4.01 / 3.8, u2s MOS / SMOS: 4.08 / 3.77, u2u MOS / SMOS: 4.06 / 2.83; WER: 4.23, CER: 1.46, F0-PCC: 0.778
38	LJSpeech, LibriSpeech, CSS10 (franču), South African languages corpus (afrikāņu)	PER: 7.8, WER: 2.6, EER: 45.6, MOS: 4.15; Franču valodai WER / EER: 28.2 / 33.9; Afrikāņu valodai WER / EER: 12.9 / 28.2
39	LibriTTS, MTG-Jamendo, MUSDB18, VCTK, DAPS	LibriTTS PESQ: 4.31, STFT: 0.388, MSD: 0.833, SCE: 111.5, RMS: 1.828, LUFS: 0.823; DAPS PESQ: 4.224, STFT: 0.391, MSD: 0.841, SCE: 109.0, RMS: 1.758, LUFS: 0.799; VCTK PESQ: 4.218, STFT: 0.441, MSD: 0.856, SCE: 152.7, RMS: 2.317, LUFS: 1.006
40	CSTR VCTK, LJSPeech, LibriTTS, CMU Arctic	Avg MOSNet pred. s2s / u2u Mel: 3.12 / 2.83, PPG: 2.93 / 2.95, APC: 3.01 / 3.02, CPC: 3.32 / 3.07, W2V: 3.09 / 2.77; runātāja noteikšana Mel: 87.3 / 96.1, PPG: 12.8 / 15.7, APC: 69.0 / 70.2, CPC: 74.0 / 77.9, W2V: 40.4 / 32.1
41	VCTK	(Rezultāti ir aptuveni ņemti no grafika) MOS M2M / F2F / M2F / F2M: 3.4 / 3.2 / 2.9 / 3.1, līdzība M2M / F2F / M2F / F2M: 3.6 / 3.7 / 3.4 / 3.2; nulles šāviena MOS M2M / F2F / M2F / F2M: 3.6 / 3.5 / 3.3 / 3.2, līdzība M2M / F2F / M2F / F2M: 3.5 / 2.8 / 3.4 / 2.7
45	Voice Conversion Challenge (VCC)	Izvēle (starp pārveidoto un reālo) pēc runas kvalitātes : 83%; Izvēlē pēc runātāja līdzības: 70%.

2.12. tabula apraksta norādītos publikāciju autoru ieteiktos turp-

māk veicamos pētījumus. Visvairāk iesaka tālāk izpētīt modeļa pielietojumu plašākiem uzdevumiem , kā arī vairāk izpētīt modeļu īpatnības un to pielietojumu.

2.12. tabula.
Runas stila pārneses publikācijās pieminētie vai ieteicamie turpmākie pētījumi.

Nr	Tālākie pētījumi
5	-
13	Attīstīt vieglas, zema latentuma dziedošas balss pārneses sistēmas. Izmantot citas sintezatora īpašības, lai iegūtu jaunus audio stila pārneses risinājumus ar citiem audio manipulācijas slāņiem.
25	Vēl lielāka pieeja no viena gala līdz otram izmantojot pārveidošanu uz runu. Kā arī mazināt augstās frekvences izvades kļūdas, kuras radās izmantotā modeļa arhitektūras rezultātā.
28	Uzlabot to, kā modelis iemācās globālās stila iezīmes (GST), un izmantot GST svarus kā mērķi, paredzot audio no teksta. Iezīmes līdzīgā veidā izmantot iegūstot attēlus no teksta vai tulkošanas modeļos. Pielietot uz citiem modeļiem, kas iegūst runu no teksta.
32	-
33	Attīstīt metodes, kas spēj atšķirt mākslīgi veidotos audio no dabīgajiem.
34	Izpētīt iespējamās modeļa modifikācijas, lai vispārinātu izpildījumu uz vairākiem trokšņainiem audio, kā arī stila pārnesi starp dažādām valodām.
35	Uzlabot noturību un atdalīšanu runas attēlojumiem ņemot vērā runātāja informāciju - ritmu un balss augstumu.
36	Izmantot šo darbu arī citām identitātes noteicošām īpatnībām, kā piemēram, akcentiem.
37	Izpētīt runātāju pielāgošanas metodes, lai uzlabotu balss līdzību, pārnesot balss stilu uz tādu runātāju, kurš nav bijis apmācībā.
38	Izpētīt mīkstās runas vienības balss stila pārnesei uz datiem, kas nav izmantoti apmācības procesā.
39	-
40	Izpētīt, kā dažādi attēlojumi palīdz iegūt vairāk informāciju par saturu un runātāju, tādējādi, uzlabojot rezultātus.

Nr	Tālākie pētījumi
41	-
45	-

Turpmāk tiks aprakstīti iepriekš minēto 3 kategoriju modeļu kombinācijas. 4. pielikumā ir vispārēja informācija par 7 modeļu publikācijām starp 2018. un 2022. gadu, kur tikai vienai ir publiski pieejams kods.

2.13. tabulā var aplūkot to, kādu uzdevumu veic kombinētie modeļi, kā arī galvenās metodes, kas aprakstītas publikācijā. Gandrīz vai katrā no publikācijām ir runas atpazīšana, un visbiežāk tā ir kombinācijā ar runas attīrīšanu.

2.13. tabula.
Kopsavilkums par kombinēto modeļu publikācijām.

Nr	Uzdevums	Metode
3	Runas attīrīšana, runas atpazīšana	Apvieno runas attīrīšanu ar runas atpazīšanas modeli, kas trenēts uz tīriem audio datiem.
6	Audio vizuālā runas atpazīšana, runas attīrīšana	Pret trokšņiem noturīgs audio vizuālās runas atpazīšanas modelis, kas ņem runātāja lūpu kustības vizuālo informāciju, labāk nosakot potenciāli atpazīto runu.
12	Audio emociju pārnese, runas ģenerēšana no teksta	Apskata divas pieejas ģenerējot audio no teksta ar konkrētām emocijām (prieku, bēdas, dusmas).
17	Vairāku valodu runas atpazīšana, runas tulkošana, runas ģenerēšana no teksta	Transformatora un RNN arhitektūras modeļu salīdzinājums dažādos runas uzdevumos dažādās datu kopās.
23	Runas attīrīšana, runas atpazīšana	Metode, kas balstīta uz GAN arhitektūras, lai veiktu runas attīrīšanas uzdevumu frekvences domēnā, kā arī iegūtu uzlabojumus runas atpazīšanā.
26	Runas stila pārnese, runas atpazīšana	Pārveido runu, noņemot akcentu, tādējādi, iegūstot augstākus runas atpazīšanas rezultātus, kad to izmanto cilvēks, kura pirmā valoda nav runātā valoda.

Nr	Uzdevums	Metode
27	Runas stila pārnese, runas atpazīšana, runas ģenerēšana no teksta	Jauns nepārraudzītās apmācības modelis stila pārnesei izmantojot runas ģenerēšanu no teksta.

2.14. tabulā ir norādīta informācija par izmantotajām kļūdas funkcijām un paša modeļa arhitektūru. Arhitektūras, kuras atkārtojas ir konformera un CNN, kur konformera arhitektūra tiek izmantota 3. un 6. publikācijā, un CNN arhitektūra tiek izmantota 12. un 26. publikācijā. Kļūdas funkcija kura atkārtojas ir krosentropijas kļūdas funkcija, kura tiek izmantota 27. publikācijā un binārā krosentropijas kļūdas funkcija 3., 17. publikācijā.

2.14. tabula.
Modeļa arhitektūra un kļūdas funkcijas modeļu kombinācijas publikācijās.

Nr	Modeļa arhitektūra	Kļūdas funkcija
3	Konformeris	Fāzē ierobežota amplitūdas (PCM) kļūda, krosentropijas kļūda (CE)
6	Konformeris un transformators	Apvienotā konnekcionistiskā laika klasifikācijas (CTC) un uzmanības kļūda
12	Konvolūciju neironu tīkls (CNN), MelGAN-VC	Satura un stila kļūdas funkcijas
17	Transformators, atkārtojošs neironu tīkls (RNN)	L1 kļūda mērķa iezīmēm, binārā krosentropijas (BCE) kļūda, vadītā uzmanības kļūda.
23	FSEGAN: frekvences domēna runas attīrīšanas GAN	Pretrunīgā kļūdas funkcija (adversarial loss), rekonstrukcijas kļūda (reconstruction loss)
26	CNN, RNN	Satura un stila kļūdas funkcijas
27	GST-Tacotron	Uzmanības konsekventā kļūda, krosentropijas (CE) kļūdas funkcija

2.15. tabulā aplūkojamas izmantotas datu kopas, kā arī iegūtie rezultāti no veiktajiem uzdevumiem saistībā ar kombinētajiem modeļiem. Gandrīz visās publikācijās kā metrika modeļu precizitātes noteikšanai ir izmantota WER, kur tā vislabākā ir 2.2 % 17. publikācijā ar transformatoru arhitektūru uz LibriSpeech datu kopas.

2.15. tabula.
Kombinētajiem modeļiem izmantotās datu kopas un iegūtie rezultāti.

Nr	Datu kopas	Metrikas un Rezultāti
3	CHiME-2	CHiME-2 ar modeļa uzlabojumiem WER: 6,28%, bez uzlabojumiem WER: 8.49%
6	LRS2 and LRS3	LRS2 WER [SNR] -5 / 0 / 5 / 10 / 15 / clean : 22.43 / 11.02 / 6.4 / 5.52 / 4.69 / 4.3; LRS3 WER [SNR] -5 / 0 / 5 / 10 / 15 / clean : 25.15 / 10.88 / 5.73 / 4.09 / 3.37 / 2.94;
12	CREMA-D, RAVDESS, SAVEE, TESS	Acc / F1 H2S: 43.8 / 10.3, H2A: 46.7 / 12.7, A2S: 32.7 / 8.22, A2H: 16.95 / 4.83, S2H: 33.33 / 8.33, S2A: 17.64 / 4.9
17	15 publiski pieejamas datu kopas	LibriSpeech WER RNN: 3.1/ 9.9/ 3.3/ 10.8; WER transformators: 2.2/ 5.6/ 2.6/ 5.7
23	WSJ un papildus dati no YouTube	WER 17.1, ASR-Clean WER 33.3, ASR-MTR WER 25.4
26	UME-ERJ, LibriSpeech	UME-ERJ bez modifikācijas / auto iekodētājs / stila pārnese CER: 46.3% / 36.1% / 31.7%; WER: 56.8% / 43.2% / 34.9%
27	VCTK, LibriSpeech	CER ar stila iezīmju slāni Google / Sphinx : 0.36 / 0.39; Bez stila iezīmju slāņa: 0.61 / 0.69

2.16. tabula apraksta norādītos tālāk veicamos pētījumus pēc publikāciju autoru ieskatiem. Visvairāk iesaka veikt kaut kāda veida paplašinājumus, vai nu uzdevumu vai datu kopu ziņā.

2.16. tabula.
Kombinēto modeļu publikācijās pieminētie vai ieteicamie turpmākie pētījumi.

Nr	Tālākie pētījumi
3	Veikt vairāk eksperimentus uz citām datu kopām. Izmantot citiem runas atpazīšanas uzdevumiem.
6	-

Nr	Tālākie pētījumi
12	Paplašināt šo pētījumu uz vēl citām emocijām, ne tikai prieku, bēdas un dusmas.
17	Izpētīt labāko kompromisu starp apmācības ātrumu un kvalitāti, kā arī izpētīt citas runas atpazīšanā izmantotās metodes, kas ģenerēta no teksta.
23	-
26	Uzlabot sākotnējo metodoloģijas kvalitāti. Veikt papildus eksperimentus, kā piemēram, izmantot vairākas tautības cilvēkus, izmantot citas datu kopas un citas valodas.
27	Izmantot neironu tīklos balstīts vokoderi, lai uzlabotu balss kvalitāti. Izmēģināt no gala līdz galam runas atpazīšanu kā diskriminatoru.

Kopumā sistemātiskajā literatūras analīzē tika aplūkota 41 publikācija, kur 10 bija par runas atpazīšanu, 9 par runas attīrīšanu, 15 par runas stila pārnesi un 7 par kombinētajiem uzdevumiem.

3. METODOLOĢIJA

Lai salīdzinātu un atrastu piemērotāko metodi balss audio ierakstu priekšapstrādei nepieciešams salīdzināt rezultātus, kas iegūti tikai no runas atpazīšanas modeļa ar tiem rezultātiem, kas iegūti no runas uzlabošanas modeļa.

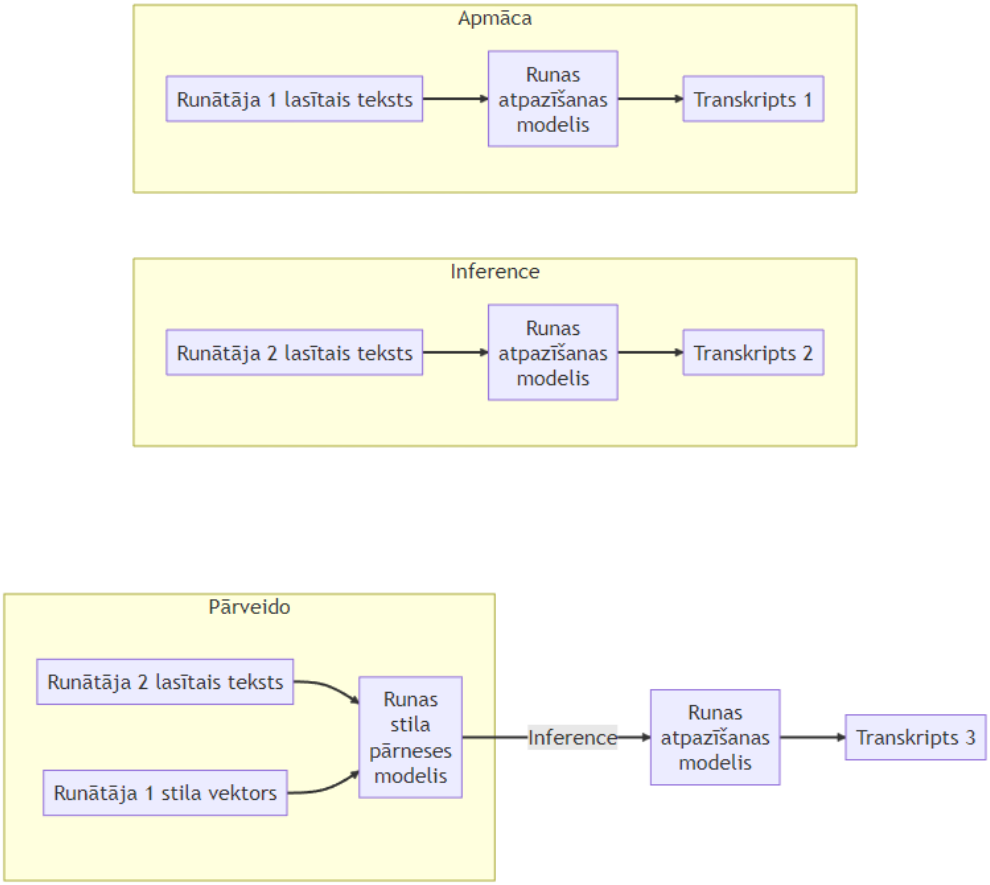
Pirmais uzdevums būtu sākotnēji veikt apmācību izvēlētajam runas atpazīšanas modelim. Kā ievades dati tiks padoti audio ar runu. Kā izvades vērtības būs modeļa ģenerēts audio attēlojums tekstā.

Nākamais uzdevums ir apmācīt runas uzlabošanas modeli, kas ir izvēlēts balstoties uz literatūras analīzi. Šim modelim, tāpat kā runas atpazīšanas, ievades dati būs audio ar runu, taču, atšķirībā no runas atpazīšanas modeļa, arī izvades dati būs audi ar runu, kura ir uzlabota vai izmainīta.

Tālāk salīdzināšanai tiks izmantots viens audio, kuru izlaiž cauri apmācītajam runas atpazīšanas modelim, rezultātā iegūstot tekstu. Šim gala rezultātam noteikt vārda kļūdas līmeni (WER) vai rakstzīmju kļūdas līmeni (CER), tādējādi ar WER vai CER iegūstot objektīvu rezultātu. Tālāk to pašu audio izlaiž cauri apmācītajam runas uzlabošanas modelim, un pēc tam runas atpazīšanas modelim. Ar WER vai CER iegūst objektīvu rezultātu. Abus iegūtos rezultātus salīdzina.

Mērķis ir samazināt WER un CER ģenerētajam tekstam, kurš iegūts no audio, kas izgājis cauri runas uzlabošanas modelim un pēc tam runas atpazīšanas modelim attiecībā pret to ģenerēto tekstu, kurš iegūts no audio, kas izgājis cauri tikai runas atpazīšanas modelim.

Lai pilnveidotu iepriekš aprakstītās metodes, praktiskās daļas izpildījuma vizuāls plāns ir aplūkojams 3.1. attēlā. Mērķis ir panākt, ka transkripts 3 galā dod labākus rezultātus nekā transkripts 2.



3.1. att. Praktiskās daļas plāna vizuāla diagramma.

Kopumā, **praktiskās daļas plāns** ir šāds:

1. Balstoties uz sistemātisko literatūras analīzi izvēlēties vienu runas uzlabošanas metodi.
2. Balstoties uz sistemātisko literatūras analīzi izvēlēties vienu datu kopu un runas atpazīšanas modeli.
3. Iegūt rezultātus un tos salīdzināt ar priekš-apmācītiem modeļiem, izmantojot runas atpazīšanas uzdevumu ar un bez runas uzlabošanas metodēm.
4. Veikt statistisko analīzi un aprakstīt rezultātus.

3.1. Datu kopa

Praktiskajā daļa izmantotā datu kopa ir VCTK (Veaux, Yamagishi et al., 2017) Tā sevī ietver audio failus, kas ierakstīti slēgtā telpā, kurus runā 110 runātāji, kur katrs vidēji ierunājis 400 audio failus. Runātāju apzīmējumi ir no p225 līdz p376 ar pa vidu izlaistiem skaitļiem, kā arī ir atsevišķi pieejami teksti tam, kas tiek pateikts katrā audio failā. Runātājs, ar apzīmējumu p315 netiek ņemts apmācība vai inferencē, jo tam nav pieejami atbilstošo audio teksta fragmenti. Datu kopā ir divu veidu faili - mic1 un mic2 -, šajā pētījumā tiek izmantoti mic1 faili, jo tie ieraksti ir ar augstāku kvalitāti, kā arī uz mic2 nav p280 runātāja faili.

Šī datu kopa tika izvēlēta, jo ir nepieciešams liels skaits ar runātāju, kā arī, lai katram runātājam būtu pietiekami daudz audio failu uz kuriem būtu iespējams apmācīt valodas atpazīšanas modeli, kā arī, lai tomēr runātāju un audio skaits nebūtu pārāk liels, jo ir pieejami ierobežoti atmiņas resursi.

3.2. Metrikas

Tā kā ar runas atpazīšanas modeļa palīdzību no audio tiek iegūts teksts, ir nepieciešams salīdzināt iegūto rezultātu ar reālo tekstu, ko runātājs ir teicis. Šim nolūkam izmanto WER jeb vārda kļūdas līmeņa metriku un CER jeb rakstzīmes kļūdas līmeņa metriku. Abām šīm metrikām formula ir vienāda un tā attēlota 3.1. vienādojumā (Leung, 2021), kur S ir aizvietošanas kļūdu skaits teikumā, jeb cik vārdi WER metrikai vai simboli CER metrikai teikumā ir aizvietoti ar citu vārdu vai simbolu, D ir dzēšanas kļūdu skaits teikumā, jeb tas, cik vārdi vai simboli no oriģinālā teksta neparādās ģenerētajā tekstā, I ir pievienošanas kļūdu skaits teikumā, jeb tas, cik jauni vārdi vai simboli parādās ģenerētajā tekstā, kuri nav bijuši oriģinālajā un N ir vārdu vai simbolu skaits teikumā. Aprēķinot WER un CER tiek pieņemts, ja summēto kļūdu skaits pārsniedz teikumā esošo vārdu vai simbolu skaits, tad kļūda ir 1 jeb 100%.

$$Error = \frac{S + D + I}{N} \tag{3.1}$$

3.2. vienādojumā (Leung, 2021) redzama WER un CER normalizētās metrikas formula, kura izmanto nevis vārdu vai simbolu skaitu teikumā kā lielumu uz kuru skatīties pēc precizitātes, bet gan ņem summu aizvietošanas, dzēšanas un pievienošanas kļūdu skaitu un tam vēl pieskaita pareizi

noteiktos vārdus vai rakstzīmes, kas vienādojumā apzīmētas ar C.

$$Error_{Norm.} = \frac{S + D + I}{S + D + I + C} \tag{3.2}$$

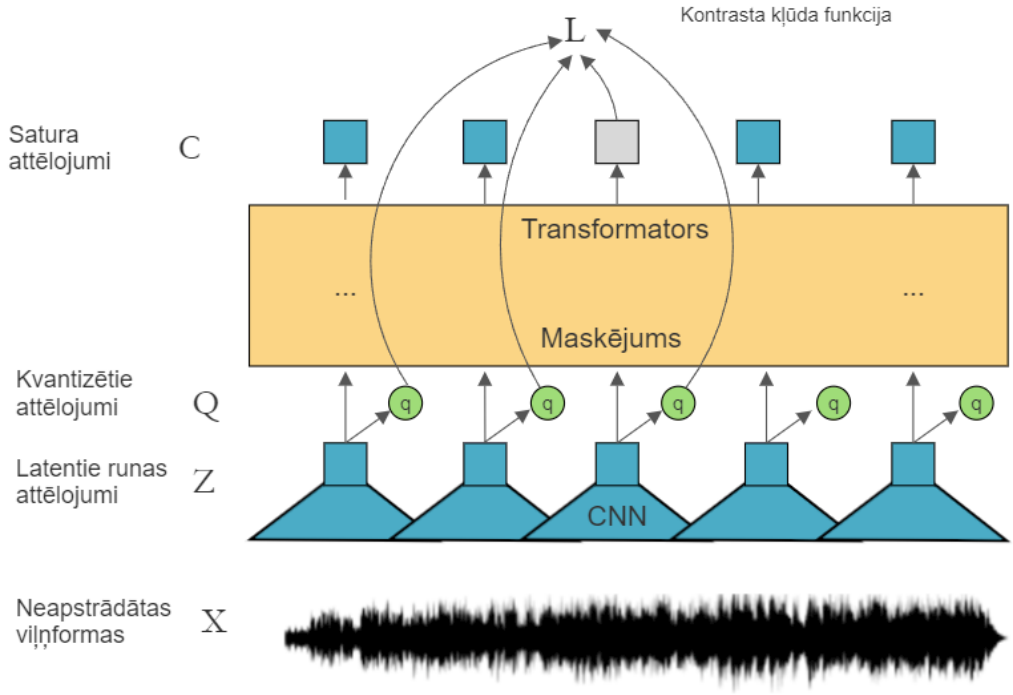
Kā jau pēc metriku aprēķināšanas formulām var noteikt, tās norāda uz kļūdu daudzumu ģenerētajā tekstā, līdz ar to šo metriku mazāka vērtība nozīmēs labāku rezultātu jeb to, ka ģenerētajā tekstā ir pieļautas mazāk kļūdas. Tāpat arī otrādi, lielākas metriku skaitliskās vērtības norāda uz sliktāku rezultātu jeb to, ka ģenerētajā tekstā ir pieļautas vairāk kļūdas.

3.3. Modeļu arhitektūras

3.3.1. Runas atpazīšanas modelis

Lai veiktu runas atpazīšanas uzdevumus, dotajā darbā tiks apmācīti un iegūti rezultāti no Whisper (Radford, Kim et al., 2022) modeļa.

2. publikācijas autori (Radford, Kim et al., 2022), savā darbā norāda, ka, tā kā viņi cenšas vairāk pierādīt savu rezultātu uzlabojumus saistībā ar datu apstrādi un to efektīvu izmantošanu, nevis modeļa attīstību, savu arhitektūru viņi balsta uz (Vaswani, Shazeer et al., 2017), kas būtība ir iekodētāja - dekodētāja transformators. Tāpat arī wav2vec 2.0 (Baevski, Zhou et al., 2020) publikācijā izmanto šo pašu arhitektūru, redzama 3.2. attēlā, tikai wav2vec 2.0 modelis, tāpat kā Whisper, ir runas atpazīšanas modelis un tā arhitektūras iezīmes būs nozīmīgākas, otra publikācija savus rezutltātus iegūst uz tulkošanu un Whisper modelī ir gan tulkošana, gan arī runas atpazīšana. Detalizēta Whisper arhitektūras darbība ir aprakstīta tālāk (balstoties uz whisper, wav2vec 2.0 un (Vaswani, Shazeer et al., 2017) publikācijām).



3.2. att. Wav2vec 2.0 arhitektūra (modificēts no (Baevski, Zhou et al., 2020))

Būtībā apmācība sastāv no 2 daļām (Baevski, Zhou et al., 2020), pašpārraudzītās daļas, kas notiek uz lielu daudzumu datu un pārraudzītās daļas, kas notiek uz mazāku skaitu datu. Pašpārraudzītā daļa cenšas saprast, kā pareizi paredzēt nākamo audio fragmentu un pārraudzītā daļa iemācās sakarības starp audio un teksta fragmentiem, jeb specifiski, kā pareizi ģenerēt tekstu no audio.

Sākotnēji audio tiek pārveidoti uz 16 kHz frekvenci (Radford, Kim et al., 2022) un attēloti 80 kanālu logaritmiskajā amplitūdu mela spektogrammās, katra no spektogrammām atbilst 25 ms audio, un spektogrammas tiek veidotas ik pa 10 ms katram audio, kā arī dati tiek globāli mērogoti vērtībās no -1 līdz 1, kur 0 atbilst visu datu vidējai vērtībai.

Pēc audio datu apstrādes (Radford, Kim et al., 2022; Baevski, Zhou et al., 2020) tie tiek padoti kā ievades dati modelim, kas, izmantojot divus konvolūciju slāņus, ar filtra lielumiem 3 un 2 un sekojošu GELU aktivizācijas funkciju, iegūst viendimensionālu vektoru ar audio attēlojuma iezīmēm laikā. Pēc šī datiem tiek veikta sinusoidāla pozīcijas iegulšana (embedding).

Tālāk esošie dati tiek padoti transformatora iekodētājam, kas, izmantojot pirmsaktivizācijas atlikuma blokus kopā ar normalizācijas slāni beigās, tos saglabā kā svarīgāko informāciju par attiecīgā segmenta satura attēlojumu kopā ar informāciju par secību. Šiem datiem tiek veikta diskretizācija ar kvantēšanas modeli, attēlojot to atsevišķās vērtības, kur tās nepārklājas citos fragmentos. Šīs vērtības tiks izmantotas apmācībā pēc dekodēšanas, kad nepieciešams veikt paredzējumu nākamajai vienībai. Kā arī vienlaicīgi ar kvantēšanu, datiem dažas vērtības tiek aizklātas jeb maskētas.

Transformatora dekodētājs (Radford, Kim et al., 2022; Baevski, Zhou et al., 2020) izmanto apgūtās pozīciju iegultnes kopā ar ievades-izvades iezīmju (token) attēlojumiem jeb dekodētāja atrasto izvades vērtību no iepriekšējās ievades. Tas cenšas pēc iekodētāja maskētajām izvades vērtībām atrast pareizo vērtību no kvantizētajām, kura iederētos maskētajā vietā.

Beigās notiek pārraudzītās apmācības daļa (Radford, Kim et al., 2022), kas starp paredzētajām vērtībām no dekodētāja izvades atrod sakarību ar teksta iezīmēm, lai pareizi no audio spētu paredzēt tekstu.

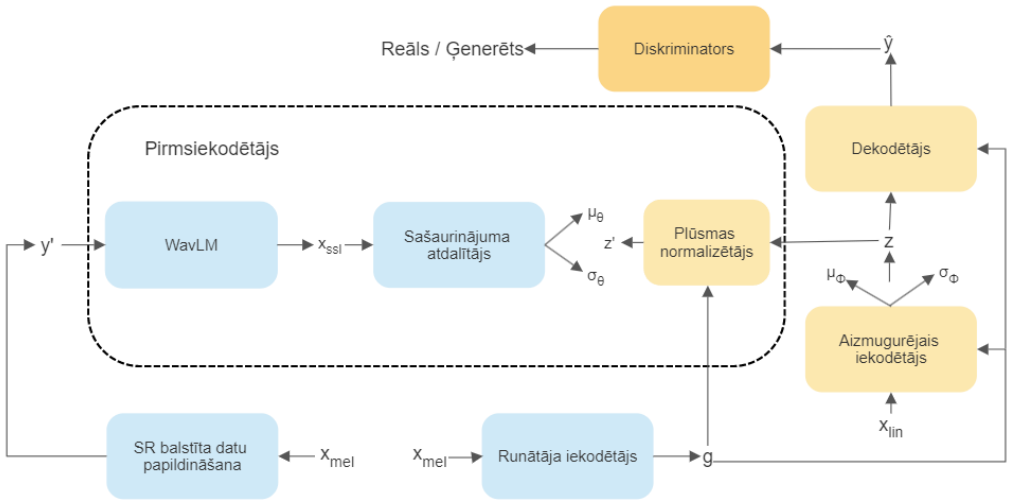
Īsāk aprakstot arhitektūru iznāk sekojoši: Audio failu attēlo secīgās mela spektogrammās. Mela spektogrammām caur iekodētāju tiek apstrādātas par audio attēlojuma iezīmēm, tālāk caur transformatora iekodētāju atrod svarīgākās informācijas iezīmes, kā arī vienlaicīgi vērtības diskretizē. Dažās no svarīgākajām informāciju iezīmēm aizklāj jeb maskē. Transformatora dekodētājs cenšas atrast pareizās diskretizētās vērtības, kuras atbilst maskētajām iezīmēm, kā arī vienlaicīgi ņemot vērā iepriekš paredzēto izvades vērtību, lai precīzāk noteiktu nākamo. Izvades vērtības no dekodētāja tiek attēlotas kā teksts, ņemot vērā teksta iezīmes (tokens).

3.3.2. Runas stila pārneses modelis

Lai veiktu runas stila pārneses uzdevumus, dotajā darbā tiks iegūti rezultāti no FreeVC (Li, Tu et al., 2022) modeļa, jo šis modelis ir uzrādījis augstus rezultātus runas stila pārnesē, tam ir pieejams kods un modeļi, kā arī tas ir ļoti adaptīvs jo spēj veiksmīgi pārnest runas stilu no un uz datiem, kuri iepriekš nav bijuši apmācībā.

FreeVC arhitektūra ir balstīta uz VITS (Kim, Kong et al., 2021), kas ir nosacījumu variācijas auto iekodētājs (CVAE) papildināts ar ģeneratīvo pretstatu tīklu (GAN) apmācību, taču atšķirībā no VITS, ieejas dati ir nevis teksts, bet audio viļņformas. Apmācības arhitektūra kopumā sastāv no pirmsiekodētāja, aizmugurējā iekodētāja, dekodētāja, diskriminatora

un runātāja iekodētāja. 3.3. attēlā redzams apmācības procesa arhitektūras vizuāls attēlojums. (Li, Tu et al., 2022)



3.3. att. FreeVC arhitektūra (modificēts no (Li, Tu et al., 2022))

Pirmsiekodētājā (Li, Tu et al., 2022) ir trīs sastāvdaļās WavLM modelis, sašaurinājuma atdalītājs un plūsmas normalizētājs. WavLM modelis pārvērš neapstrādātās viļņformas par 1024 dimensionālu īpašību vektoru priekš stimulētās apmācības, kas satur informāciju gan par saturu,gan runātāju. Lai iegūtu tikai satura informāciju un noņemtu lieko runātāja informāciju, iegūtais vektors tiek izlaists cauri sašaurinājuma atdalītājam, kas saspiež vektoru daudz mazākā dimensiju skaitā, tādējādi piespiežot atbrīvoties no liekās informācijas. Kopumā WavLM modelis un sašaurinājuma atdalītājs iegūst satura informāciju sadalījuma modelēšanas formā. Normalizējošā plūsma nodrošina to, ka iepriekš iegūtā sadalījuma sarežģītība tiek uzlabota.

Aizmugurējais iekodētājs (Kim, Kong et al., 2021) izmanto 16 WaveNet atlikuma blokus ar lineārajām spektogrammām kā ieejas datiem, lai iegūtu latento vektoru z. Tālāk tas tiek padots kā ieejas dati dekodētājam, kas ģenerē izejas datus jeb pārveidoto audio failu. Dekodētāja izejas dati tiek padoti diskriminatoram, kas cenšas noteikt, vai tas ir reāls vai ģenerēts audio, tādējādi arī trenējot šo modeli.

Runātāja iekodētājs (Li, Tu et al., 2022) no audio faila saglabā

runātāja iezīmes iegūstot balss latento vektoru, lai tās vēlāk izmantotu, pārveidojot audio uz vēlamo runātāju.

Kopumā pirmsiekodētājs izvelk visu saturisko informāciju ignorējot informāciju par pašu runātāju, taču runātāja iekodētājs izvelk informāciju tieši par mērķa runātāju. Aizmugurējais iekodētājs, līdzīgi kā pirmsiekodētājs, atdala satura informāciju no oriģinālā runātāja informācijas, un tālāk dekodētājs apvieno mērķa runātāju ar saturisko informāciju, iegūstot audio mērķa runātāja stilā, kas pēc tam tiek padots diskriminatoram, kurš cenšas noteikt, vai tas ir reāls vai ģenerēts audio.

3.4. Apmācību un testēšanas protokols

Pašā sākumā tiks veikta apmācība Whisper (Radford, Kim et al., 2022) modelim uz VCTK (Veaux, Yamagishi et al., 2017) datu kopas. Pēc šīs apmācības tiks noskaidroti WER un CER ar šo apmācīto modeli VCTK datu kopai.

Tad, lai noskaidrotu kuras balsis no VCTK datu kopas visvairāk atšķiras no citām balsīm jeb no vidēji visām balsīm datu kopā, tiks veikti balss id latento vektoru aprēķini. Ar (Ravanelli, Parcollet et al., 2021) runātāja atpazīšanas modeli (Desplanques, Thienpondt et al., 2020) tika iegūts balss latentais vektors, kuram tālāk tika iegūta vektora vidējā vērtība. Tas tika veikts reālajai balsij, mērķa balsij un balsij, kas pārveidota no reālās uz mērķa ar FreeVC (Li, Tu et al., 2022) modeli. Kad tiks iegūti katras balss latentie vektori, tiks noskaidrota kosinusa līdzība starp to, cik balss ir atšķirīga no vidējās balss datu kopā un cik labi strādā FreeVC konkrētajai balsij. Un ņemot vērā abus šos rezultātus, tika aprēķināts rezultāts katram runātājam pēc kura turpmāk 5 augstākie rezultāti tiek izmantoti kā tie, uz kuriem tiks veikti balss stila pārneses uzdevumi. 3.3. vienādojumā redzamas aprēķinu formulas, kur Z_real ir reālās balss latentais vektors, Z_oth ir mērķa balss latentais vektors un Z_conv ir latentais vektors balsij, kas pārveidota no reālās uz mērķa ar FreeVC, kā arī tad dist_cos_other ir kosinusa līdzība starp to, cik balss ir atšķirīga no vidējās balss datu kopā, dist_cos_conv ir kosinusa līdzība tam, cik labi strādā FreeVC konkrētajai balsij, un score_p ir beigu rezultāts.

$$\begin{aligned} dist\_cos\_other &= 1 + cos\_sim(Z\_real, Z\_oth) \\ dist\_cos\_conv &= 1 + cos\_sim(Z\_real, Z\_conv) \\ score\_p &= 0.5 * dist\_cos\_other + 2 - dist\_cos\_conv \end{aligned} \tag{3.3}$$

Ņemot vērā 5 augstākos rezultātus, kas iegūti no 3.3. vienādojuma, tiks apmācīti 5 citi Whisper modeļi specifiski uz tiem runātājiem. Tālāk ar FreeVC modeli tiks veikta balss stila pārnese visai VCTK datu kopai uz 5 iepriekš atrastajiem runātājiem, lai pēc tam uz katra runātāja apmācītā modeļa varētu vēlreiz iegūt WER un CER.

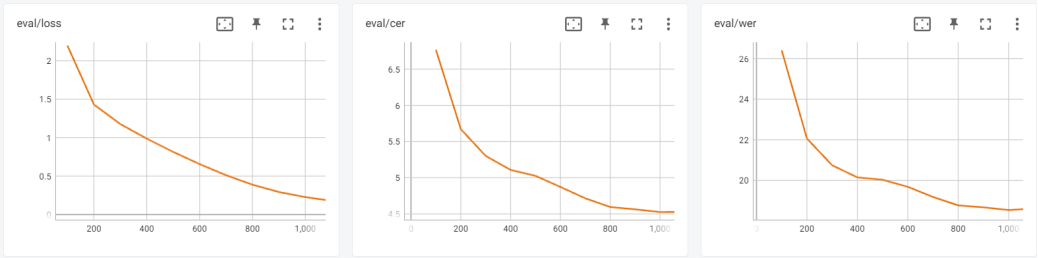
Iegūtie rezultāti tiks salīdzināti un tiks noskaidrots, vai izmantojot balss stila pārnesi un apmācot runas atpazīšanas modeli uz specifisku runātāju, ir iespējams iegūt augstākus rezultātus nekā bez balss stila pārneses.

4. REZULTĀTI

Metodoloģijā norādīto uzdevumu realizēšanai izmantotais kods un visi eksperimentāli iegūtie rezultāti ir pieejami šeit: `https://github.com/Betija13/Bakalaura_darbs_kods`

Pēc runātāju latento vektoru aprēķiniem, kas iegūti uz VCTK runātājiem, sekojošie runātāji, ar visaugstākajiem rezultātiem un aprēķinu rezultāts, ir šādi: p287 (0.844), p254 (0.834), p317 (0.815), p363 (0.812), p304 (0.807). Līdz ar to, audio VCTK datu kopā ar FreeVC tiks pārveidoti uz šiem mērķa runātājiem, kā arī Whisper tiks apmācīts uz šiem mērķa runātājiem.

4.1. attēlā var aplūkot apmācības procesa grafikus apmācot Whisper modelim uz visas datu kopas. 4.1. attēlā pa kreisi redzama kļūdas funkcijas vērtība testēšanas daļā, kā arī CER (pa vidu) un WER (labā mala) metriku rezultāti testēšanas daļā. Kā redzams, visas vērtības sāk tiekties uz kādu skaitli un īpaši vairs neizmainās, līdz ar to, modelis ir iemācījies visu, ko tas spēj iemācīties.



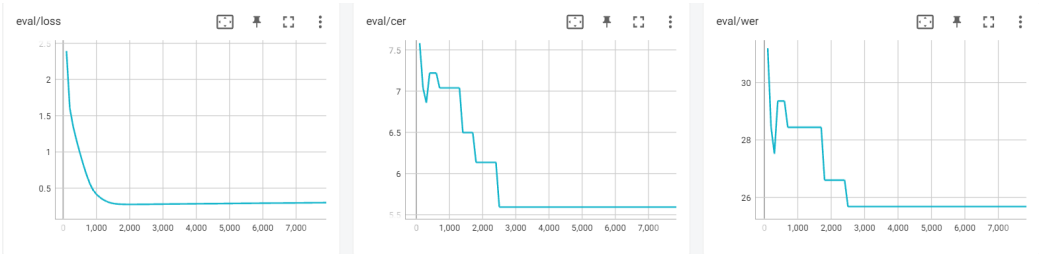
4.1. att. Grafiki runas atpazīšanas modeļa trenēšanai uz visas VCTK datu kopas.

4.1. tabulā redzami WER un CER parasto un normalizēto metriku rezultāti mērķa runātājiem uz runas atpazīšanas modeļa, kas trenēts uz visas VCTK datu kopas. No balss stila mērķa runātājiem vislabākais rezultāts rezultāts ir p287 runātājam, bet vissliktākais p254 runātājam. Pilnie rezultāti ir aplūkojami 5. pielikumā. Šie iegūtie rezultāti ir pamata rezultāti, kas tiks izmantoti tālākai salīdzināšanai ar rezultātiem no runas stila pārneses modeļiem.

4.1. tabula.
Metriku rezultāti uz modeļa, kas apmācīts uz visas VCTK datu kopas.

	WER	CER	WER Norm.	CER Norm.
p254	8.70	11.57	8.34	10.69
p287	**5.73**	**9.65**	**5.51**	**8.98**
p304	8.18	11.25	7.93	10.42
p317	7.52	9.97	7.21	9.32
p363	7.52	9.97	7.21	9.32
Vidējais uz visiem	8.33	11.37	8.00	10.50

4.2. attēlā var aplūkot apmācības procesa grafikus trenējot modeli uz p304 runātāju, kur pa kreisi ir kļūdas funkcijas vērtība, vidū ir CER un pa labi ir WER apmācības testēšanas daļā. Kā redzams visas iegūtās paliek nemainīgas, līdz ar to, modelis ir apmācīts.



4.2. att. Grafiki runas atpazīšanas modeļa trenēšanai uz p304 runātāju.

4.2. tabulā redzami metriku rezultāti mērķa runātājiem uz runas atpazīšanas modeļa, kas apmācīts uz p304 runātāju, datu kopai veicot runas stila pārnesi uz p304 runātāju. Pilnie rezultāti ir aplūkojami 6. pielikumā. Pēc 4.2. tabulas redzams, ka vislabākie rezultāti ir tieši uz p304 runātāju, kas arī ir sagaidāmi, jo tieši uz šo runātāju modelis ir trenēts, bet vissliktākie rādītāji starp mērķa runātājiem ir p254 runātājam.

4.2. tabula.
Metriku rezultāti uz modeļa, kas apmācīts uz p304.

	WER	CER	WER Norm.	CER Norm.
p254	7.94	9.69	7.52	8.68
p287	7.08	8.58	6.86	7.83
p304	**3.88**	**5.58**	**3.85**	**5.30**
p317	7.39	8.69	7.12	7.91
p363	4.23	7.22	4.12	6.61
Vidējais uz visiem	8.69	10.30	8.28	9.21

Salīdzinājums starp VCTK modeļa un p304 modeļa rezultātiem aplūkojams 4.3. tabulā. Pēc rezultātiem var redzēt, ka četriem no pieciem mērķa runātājiem ir uzlabots WER rezultāts un CER rezultāts ir uzlabots visiem mērķa runātājiem, un vidējais rezultāts uz visiem runātājiem arī ir uzlabots par 1.07.

4.3. tabula.
Salīdzinājums WER un CER starp visas VCTK datu kopas un p304 modeļiem.

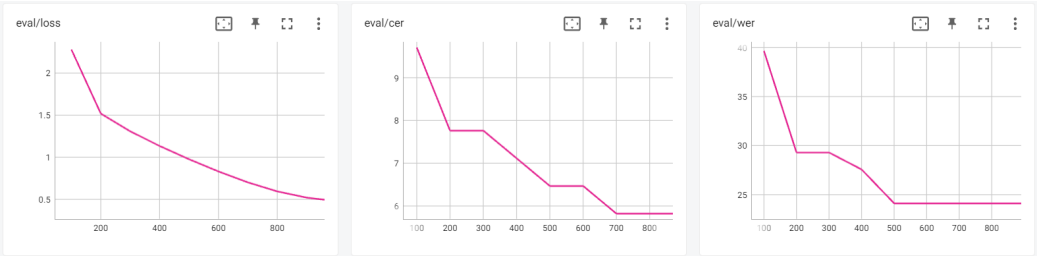
	WER			CER		
	VCTK	**p304**	**Uzlabojums**	**VCTK**	**p304**	**Uzlabojums**
p254	8.70	7.94	0.76	11.57	9.69	1.88
p287	5.73	7.08	-1.35	9.65	8.58	1.06
p304	8.18	3.88	4.29	11.25	5.58	5.67
p317	7.52	7.39	0.13	9.97	8.69	1.27
p363	7.52	4.23	3.29	9.97	7.22	2.74
Vidējais	8.33	8.69	-0.36	11.37	10.30	1.07

Salīdzinājums normalizētajām WER un CER metrikām starp VCTK modeļa un p304 modeļa rezultātiem aplūkojams 4.4. tabulā. Pēc rezultātiem var redzēt, ka četriem no pieciem mērķa runātājiem ir uzlabots normalizētais WER rezultāts un normalizētais CER rezultāts ir uzlabots visiem mērķa runātājiem, un vidējais rezultāts uz visiem runātājiem arī ir uzlabots par 1.29.

4.4. tabula.
Salīdzinājums normalizētajiem WER un CER starp visas VCTK datu kopas un p304 modeļiem.

	WER Norm.			CER Norm.		
	VCTK	**p304**	**Uzlabojums**	**VCTK**	**p304**	**Uzlabojums**
p254	8.34	7.52	0.83	10.69	8.68	2.02
p287	5.51	6.86	-1.35	8.98	7.83	1.15
p304	7.93	3.85	4.08	10.42	5.30	5.13
p317	7.21	7.12	0.10	9.32	7.91	1.41
p363	7.21	4.12	3.09	9.32	6.61	2.71
Vidējais	8.00	8.28	-0.28	10.50	9.21	1.29

4.3. attēlā var aplūkot apmācības procesa grafikus trenējot modeli uz p317 runātāju, kur pa labi ir kļūdas funkcijas vērtība, pa vidu ir CER un pa labi ir WER vērtības apmācības testēšanas daļā. Kā redzams WER un CER rezultāti ir nostabilizējušies un vairs neizmainās un arī kļūdas funkcijas vērtība tiecās uz nemainīgu vērtību, līdz ar to, modelis ir apmācīts.



4.3. att. Grafiki runas atpazīšanas modeļa trenēšanai uz p317 runātāju.

4.5. tabulā redzami metriku rezultāti mērķa runātājiem uz runas atpazīšanas modeļa, kas apmācīts uz p317 runātāju, datu kopai veicot runas stila pārnesi uz p317 runātāju. Pilnie rezultāti ir aplūkojami 7. pielikumā. Kā redzams pēc 4.5. tabulas, vislabākie rezultāti ir uz p317 runātāja, bet vissliktākie uz p304 runātāju.

4.5. tabula.
WER un CER uz modeļa, kas apmācīts uz p317.

	WER	CER	WER Norm.	CER Norm.
p254	9.90	12.08	9.25	10.59
p287	8.06	10.91	7.66	9.74
p304	11.29	13.64	10.90	12.26
p317	**2.52**	**6.47**	**2.42**	**6.04**
p363	4.58	9.42	4.38	8.52
Vidējais uz visiem	7.85	10.93	7.44	9.79

Salīdzinājums starp VCTK modeļa un p317 modeļa WER un CER rezultātiem aplūkojams 4.6. tabulā. Pēc rezultātiem var redzēt, ka diviem no pieciem mērķa runātājiem ir uzlabots gan WER, gan CER rezultāts, kā arī vidējais WER rezultāts ir uzlabots par 0.48 un vidējais CER rezultāts ir uzlabots par 0.43.

4.6. tabula.
Salīdzinājums WER un CER starp visas VCTK datu kopas un p317 modeļiem.

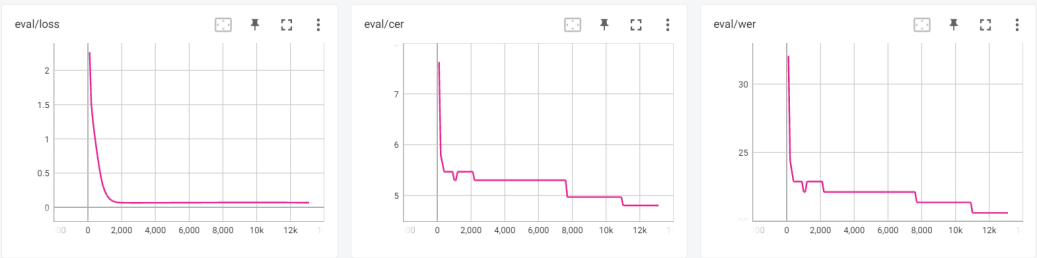
	WER			CER		
	VCTK	**p317**	**Uzlabojums**	**VCTK**	**p317**	**Uzlabojums**
p254	8.70	9.90	-1.20	11.57	12.08	-0.51
p287	5.73	8.06	-2.33	9.65	10.91	-1.27
p304	8.18	11.29	-3.12	11.25	13.64	-2.39
p317	7.52	2.52	5.01	9.97	6.47	3.50
p363	7.52	4.58	2.94	9.97	9.42	0.54
Vidējais	8.33	7.85	0.48	11.37	10.93	0.43

Salīdzinājums normalizētajām WER un CER metrikām starp VCTK modeļa un p317 modeļa rezultātiem aplūkojams 4.7. tabulā. Pēc rezultātiem var redzēt, ka normalizētais WER ir uzlabots diviem no pieciem mērķa runātājiem un vidējais rezultāts ir uzlabots par 0.56, un normalizētais CER ir uzlabots trijiem no pieciem mērķa runātājiem, kā arī vidējais rezultāts ir uzlabots par 0.71.

4.7. tabula.
Salīdzinājums normalizētajiem WER un CER starp visas VCTK datu kopas un p317 modeļiem.

	WER Norm.			CER Norm.		
	VCTK	**p317**	**Uzlabojums**	**VCTK**	**p317**	**Uzlabojums**
p254	8.34	9.25	-0.91	10.69	10.59	0.10
p287	5.51	7.66	-2.14	8.98	9.74	-0.75
p304	7.93	10.90	-2.97	10.42	12.26	-1.84
p317	7.21	2.42	4.79	9.32	6.04	3.28
p363	7.21	4.38	2.83	9.32	8.52	0.79
Vidējais	8.00	7.44	0.56	10.50	9.79	0.71

4.4. attēlā var aplūkot apmācības procesa grafikus trenējot modeli uz p363 runātāju. 4.4. attēlā pa labi ir kļūdas funkcijas vērtības, pa vidu CER un pa labi WER metriku vērtības apmācības procesa testēšanas daļā. Kā redzams visas vērtības nostabilizējas un paliek nemainīgas, līdz ar to, apmācības process ir izpildīts.



4.4. att. Grafiki runas atpazīšanas modeļa trenēšanai uz p363 runātāju.

4.8. tabulā redzami metriku rezultāti mērķa runātājiem uz runas atpazīšanas modeļa, kas apmācīts uz p363 runātāju, datu kopai veicot runas stila pārnesi uz p363 runātāju. Pilnie rezultāti ir aplūkojami 8. pielikumā. Pēc 4.8. tabulas var redzēt, ka vislabākais rezultāts ir uz p363 runātāju, bet vissliktākais uz p304 runātāju.

4.8. tabula.
WER un CER uz modeļa, kas apmācīts uz p363.

	WER	CER	WER Norm.	CER Norm.
p254	9.47	10.78	8.97	9.54
p287	6.58	8.30	6.36	7.45
p304	13.57	13.92	12.95	12.23
p317	6.84	8.07	6.47	7.36
p363	**3.00**	**5.67**	**2.94**	**5.25**
Vidējais uz visiem	10.00	11.19	9.43	9.86

Salīdzinājums starp VCTK modeļa un p363 modeļa rezultātiem aplūkojams 4.9. tabulā. Pēc rezultātiem var redzēt, ka WER metrika ir uzlabota diviem no pieciem mērķa runātājiem, bet CER metrika ir uzlabota četriem no pieciem mērķa runātājiem, kā arī vidējā CER vērtība ir uzlabota par 0.17.

4.9. tabula.
Salīdzinājums WER un CER starp visas VCTK datu kopas un p363 modeļiem.

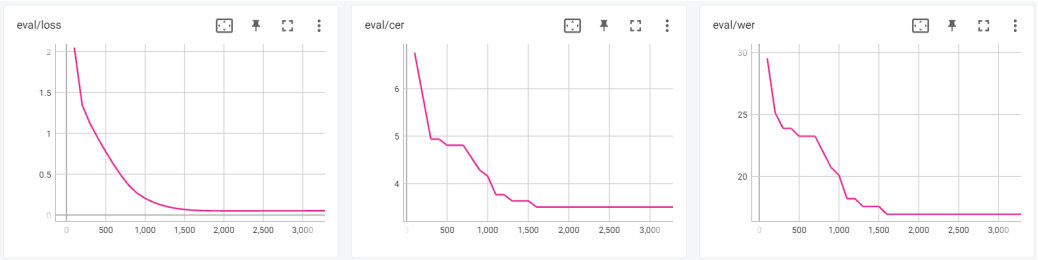
	WER			CER		
	VCTK	**p363**	**Uzlabojums**	**VCTK**	**p363**	**Uzlabojums**
p254	8.70	9.47	-0.77	11.57	10.78	0.79
p287	5.73	6.58	-0.85	9.65	8.30	1.35
p304	8.18	13.57	-5.39	11.25	13.92	-2.67
p317	7.52	6.84	0.68	9.97	8.07	1.89
p363	7.52	3.00	4.52	9.97	5.67	4.29
Vidējais	8.33	10.00	-1.66	11.37	11.19	0.17

Salīdzinājums normalizētajām WER un CER metrikām starp VCTK modeļa un p363 modeļa rezultātiem aplūkojams 4.10. tabulā. Pēc rezultātiem var redzēt, ka normalizētā WER metrika ir uzlabota diviem no pieciem mērķa runātājiem, bet normalizētā CER metrika ir uzlabota četriem no pieciem mērķa runātājiem, kā arī vidējā normalizētā CER vērtība ir uzlabota par 0.64.

4.10. tabula.
Salīdzinājums normalizētajām WER un CER starp visas datu kopas un p363 modeļiem.

	WER Norm.			CER Norm.		
	VCTK	**p363**	**Uzlabojums**	**VCTK**	**p363**	**Uzlabojums**
p254	8.34	8.97	-0.63	10.69	9.54	1.15
p287	5.51	6.36	-0.85	8.98	7.45	1.53
p304	7.93	12.95	-5.02	10.42	12.23	-1.81
p317	7.21	6.47	0.74	9.32	7.36	1.96
p363	7.21	2.94	4.27	9.32	5.25	4.06
Vidējais	8.00	9.43	-1.44	10.50	9.86	0.64

4.5. attēlā var aplūkot apmācības procesa grafikus trenējot modeli uz p287 runātāju. 4.5. attēlā pa kreisi ir kļūdas funkcijas rezultāti, pa vidu CER un pa labi ir WER metriku rezultāti apmācības procesa testa daļā. Kā redzams rezultātiem vairs nav izmaiņu vai uzlabojumi un to vērtība ir nemainīga, līdz ar to modeļa apmācība ir noslēgusies.



4.5. att. Grafiki runas atpazīšanas modeļa trenēšanai uz p287 runātāju.

4.11. tabulā redzami metriku rezultāti mērķa runātājiem uz runas atpazīšanas modeļa, kas apmācīts uz p287 runātāju, datu kopai veicot runas stila pārnesi uz p287 runātāju. Pilnie rezultāti ir aplūkojami 9. pielikumā. Kā 4.11. tabulā ir aplūkojams, vislabākais rezultāts ir p287 runātājam, bet vissliktākais ir p304 runātājam.

4.11. tabula.
Metriku rezultāti uz modeļa, kas apmācīts uz p287.

	WER	CER	WER Norm.	CER Norm.
p254	7.68	9.63	7.26	8.59
p287	**3.17**	**5.97**	**3.06**	**5.47**
p304	11.03	11.97	10.51	10.70
p317	8.07	9.58	7.71	8.59
p363	4.51	7.46	4.42	6.86
Vidējais uz visiem	9.66	11.24	9.11	9.97

Salīdzinājums starp VCTK modeļa un p287 modeļa WER un CER rezultātiem aplūkojams 4.12. tabulā. Pēc 4.12. tabulas var redzēt, ka WER ir uzlabots trim no pieciem mērķa runātājiem un CER ir uzlabots četriem no pieciem mērķa runātājiem, kā arī vidējais CER ir uzlabots par 0.13.

4.12. tabula.
Salīdzinājums WER un CER starp visas VCTK datu kopas un p287 modeļiem.

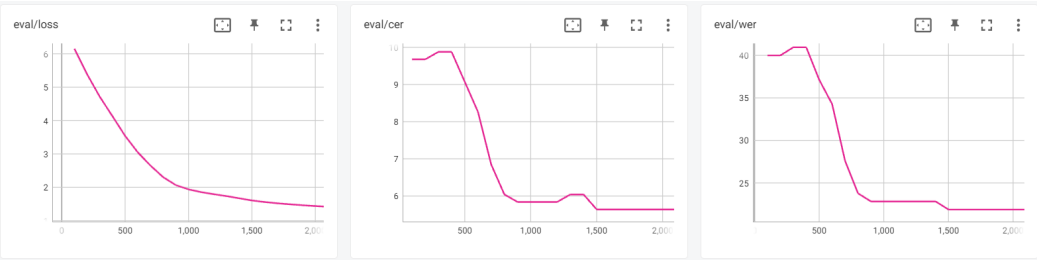
	WER			CER		
	VCTK	**p287**	**Uzlabojums**	**VCTK**	**p287**	**Uzlabojums**
p254	8.70	7.68	1.03	11.57	9.63	1.95
p287	5.73	3.17	2.56	9.65	5.97	3.68
p304	8.18	11.03	-2.85	11.25	11.97	-0.72
p317	7.52	8.07	-0.55	9.97	9.58	0.39
p363	7.52	4.51	3.01	9.97	7.46	2.50
Vidējais	8.33	9.66	-1.33	11.37	11.24	0.13

Salīdzinājums normalizētajām WER un CER metrikām starp VCTK modeļa un p287 modeļa rezultātiem aplūkojams 4.13. tabulā. Pēc 4.13. tabulas rezultātiem var redzēt, ka normalizētā WER metrika ir uzlabota trim no pieciem mērķa runātājeim un normalizētā CER metrika ir uzlabota četriem no pieciem mērķa runātājiem, kā arī vidējais normalizētais CER rezultāts ir uzlabots par 0.53.

4.13. tabula.
Salīdzinājums normalizētajām WER un CER metrikām starp visas VCTK datu kopas un p287 modeļiem.

	WER Norm.			CER Norm.		
	VCTK	**p287**	**Uzlabojums**	**VCTK**	**p287**	**Uzlabojums**
p254	8.34	7.26	1.08	10.69	8.59	2.10
p287	5.51	3.06	2.46	8.98	5.47	3.51
p304	7.93	10.51	-2.58	10.42	10.70	-0.28
p317	7.21	7.71	-0.50	9.32	8.59	0.73
p363	7.21	4.42	2.80	9.32	6.86	2.46
Vidējais	8.00	9.11	-1.11	10.50	9.97	0.53

4.6. attēlā var aplūkot apmācības procesa grafikus trenējot modeli uz p254 runātāju. Kā redzams 4.6. attēlā, kurā pa kreisi atrodas kļūdas funkcijas rezultāti, pa vidu CER un pa labi WER metriku rezultāti apmācības precesa testēšanas daļā, visi rezultāti paliek nemainīgi un tiecās uz vienu konkrētu vērtību, tādējādi tas parāda, ka apmācības process ir galā.



4.6. att. Grafiki runas atpazīšanas modeļa trenēšanai uz p254 runātāju.

4.14. tabulā redzami metriku rezultāti mērķa runātājiem uz runas atpazīšanas modeļa, kas apmācīts uz p254 runātāju, datu kopai veicot runas stila pārnesi uz p254 runātāju. Pilnie rezultāti ir aplūkojami 10. pielikumā. Kā pēc 4.14. tabulas redzams, vislabākais rezultāts no mērķa runātājiem ir p254 runātājam, bet vissliktākais p304 runātājam.

4.14. tabula.
Metriku rezultāti uz modeļa, kas apmācīts uz p254.

	WER	CER	WER Norm.	CER Norm.
p254	**2.51**	**5.27**	**2.47**	**4.87**
p287	6.63	8.19	6.46	7.35
p304	9.47	10.71	9.35	9.63
p317	6.43	7.82	6.19	7.12
p363	4.20	6.95	4.10	6.28
Vidējais uz visiem	7.46	9.17	7.18	8.19

Salīdzinājums starp VCTK modeļa un p254 modeļa WER un CER rezultātiem aplūkojams 4.15. tabulā. Pēc rezultātiem var redzēt, ka WER metrikas vērtība ir uzlabota trim no pieciem mērķa runātājiem un vidējais WER ir paaugstinājies par 0.87, toties CER metrikas vērtība ir uzlabota visiem pieciem mērķa runātājiem un vidējā CER vērtība ir paaugstināta par 2.19.

4.15. tabula.
Salīdzinājums WER un CER starp visas VCTK datu kopas un p254 modeļiem.

	WER			CER		
	VCTK	**p254**	**Uzlabojums**	**VCTK**	**p254**	**Uzlabojums**
p254	8.70	2.51	6.19	11.57	5.27	6.31
p287	5.73	6.63	-0.90	9.65	8.19	1.46
p304	8.18	9.47	-1.29	11.25	10.71	0.53
p317	7.52	6.43	1.09	9.97	7.82	2.14
p363	7.52	4.20	3.32	9.97	6.95	3.02
Vidējais	8.33	7.46	0.87	11.37	9.17	2.19

Salīdzinājums normalizētajām metrikām WER un CER metrikām starp VCTK modeļa un p254 modeļa rezultātiem aplūkojams 4.16. tabulā. Pēc rezultātiem var redzēt, ka normalizētā WER metrikas vērtība ir uzlabota trim no pieciem mērķa runātājiem un vidējais normalizētais WER ir paaugstinājies par 0.81, toties normalizētā CER metrikas vērtība ir uzlabota visiem pieciem mērķa runātājiem un vidējā normalizētā CER vērtība ir paaugstināta par 2.31.

4.16. tabula.
Salīdzinājums normalizētajiem WER un CER starp visas VCTK datu kopas un p254 modeļiem.

	WER Norm.			CER Norm.		
	VCTK	**p254**	**Uzlabojums**	**VCTK**	**p254**	**Uzlabojums**
p254	8.34	2.47	5.88	10.69	4.87	5.82
p287	5.51	6.46	-0.95	8.98	7.35	1.64
p304	7.93	9.35	-1.42	10.42	9.63	0.80
p317	7.21	6.19	1.02	9.32	7.12	2.20
p363	7.21	4.10	3.11	9.32	6.28	3.04
Vidējais	8.00	7.18	0.81	10.50	8.19	2.31

Vārdu kļūdas metrikas rezultātu apkopojumus var redzēt 4.17. tabulā. Pēc rezultātiem var redzēt, ka uzlaboti rezultāti ir kā minimums 37 runātājiem, kas sastāda 34% runātāju no visas VCTK datu kopas, ko ir spējīgs veikt modelis, kas trenēts uz p363 runātāju, un ka p254 spēja uzlabot WER pat 81 runātājam rezultātus, kas ir 74% no datu kopas runātājiem, kā arī maksimālais uzlabojums bija 6.19 uzlabojums WER metrikā, ko spēja panākt modelis, kas trenēts uz p254 runātāju.

4.17. tabula.
Salīdzinājums WER starp visiem runātāju modeļiem.

	p254	**p287**	**p304**	**p317**	**p363**
Uzlabojums (skaits)	81	38	60	74	37
Uzlabojums (%)	74%	35%	55%	67%	34%
Vislielākais uzlabojums	6.19	2.75	4.29	5.01	3.61

Rakstzīmes kļūdas metrikas rezultātu apkopojumus var redzēt 4.18. tabulā. Pēc rezultātiem var redzēt, ka uzlaboti rezultāti ir kā minimums 70 runātājiem, kas sastāda 64% runātāju no visas VCTK datu kopas, ko ir spējīgs veikt modelis, kas trenēts uz p363 runātāju, un ka p254 spēja uzlabot CER pat 95 runātājiem rezultātus, kas ir 86% no datu kopas runātājiem, kā arī maksimālais uzlabojums bija 6.31 uzlabojums CER metrikā, ko spēja panākt modelis, kas trenēts uz p254 runātāju.

4.18. tabula.
Salīdzinājums CER starp visiem runātāju modeļiem.

	p254	**p287**	**p304**	**p317**	**p363**
Uzlabojums (skaits)	95	72	84	78	70
Uzlabojums (%)	86%	65%	76%	71%	64%
Vislielākais uzlabojums	6.31	3.68	5.67	4.00	4.90

Normalizētās vārdu kļūdas metrikas rezultātu apkopojumus var redzēt 4.19. tabulā. Pēc rezultātiem var redzēt, ka uzlaboti rezultāti ir kā minimums 39 runātājiem, kas sastāda 35% runātāju no visas VCTK datu kopas, ko ir spējīgs veikt modelis, kas trenēts uz p363 runātāju, un ka p254 spēja uzlabot normalizēto WER pat 80 runātājiem rezultātus, kas ir 73% no datu kopas runātājiem, kā arī maksimālais uzlabojums bija 5.88 uzlabojums normalizētajā WER metrikā, ko spēja panākt modelis, kas trenēts uz p254 runātāju.

4.19. tabula.
Salīdzinājums normalizētajiem WER rezultātiem starp visiem runātāju modeļiem.

	p254	**p287**	**p304**	**p317**	**p363**
Uzlabojums (skaits)	80	41	60	74	39
Uzlabojums (%)	73%	37%	55%	67%	35%
Vislielākais uzlabojums	5.88	2.64	4.08	4.81	3.53

Normalizētās rakstzīmes kļūdas metrikas rezultātu apkopojumus var redzēt 4.20. tabulā. Pēc rezultātiem var redzēt, ka ... ka uzlaboti rezultāti ir kā minimums 79 runātājiem, kas sastāda 72% runātāju no visas VCTK datu kopas, ko ir spējīgs veikt modelis, kas trenēts uz p363 runātāju, un ka p254 spēja uzlabot normalizēto CER pat 100 runātājiem rezultātus,

kas ir 91% no datu kopas runātājiem, kā arī maksimālais uzlabojums bija 5.82 uzlabojums normalizētajā CER metrikā, ko spēja panākt modelis, kas trenēts uz p254 runātāju.

4.20. tabula.
Salīdzinājums CER normalizētajiem rezultātiem starp visiem runātāju modeļiem.

	p254	**p287**	**p304**	**p317**	**p363**
Uzlabojums (skaits)	100	81	89	81	79
Uzlabojums (%)	91%	74%	81%	74%	72%
Vislielākais uzlabojums	5.82	3.51	5.13	3.76	4.60

Kopumā pēc rezultātiem ir redzams, ka daļa rezultātu ir uzlabojumi. Īpaši labus rezultātus ir parādījis modelis uz p254 runātāju, kurš normalizēto rakstzīmes kļūdas metriku spēja uzlabot pat 91 % no VCTK datu kopas runātājiem, taču arī pat p363 modelis, kurš veicis vismazāk uzlabojumus starp citiem modeļiem, šo metriku spēja uzlabot 72% no visas VCTK datu kopas runātājiem. Vismazāk uzlabojumi ir novērojami WER metrikai, kur p363 veicis uzlabojumus tikai 34% no runātājiem, toties p254 joprojām rāda augstu rezultātu uzlabojot to 80% runātāju no visas VCTK datu kopas.

Pēc rezultātiem var arī redzēt, ka rakstzīmes kļūdas metrikas uzlabojums ir vairāk lietotājiem, nekā vārdu kļūdas metrikas uzlabojumi, kas nozīmē, ka modeļi pieļauj kļūdas cenšoties paredzēt pareizo vārdu, taču spēj noprast, burtus no kuriem sastāv vārds.

5. TĀLĀKIE PĒTĪJUMI

Ņemot vērā esošā pētījuma darbību un rezultātus, lai dziļāk izpētītu priekšapstrādes metodes runas atpazīšanai un kā tās izmaina rezultātu efektivitāti ir nepieciešams pārbaudīt vai attiecīgie rezultāti ir līdzīgi izmantojot citus jau esošas runas atpazīšanas sistēmas un trenētos modeļus.

Darbā izmantotais modelis bija Whisper, kas balstās uz transformatora arhitektūru, tālāk varētu izpētīt, kā runas atpazīšanas priekšapstrādes ietekmē atkārtojošo neironu tīklu (RNN) tipa modeļus, jo šī tipa modeļi arī bieži tiek izmantoti runas atpazīšanā. Pie arhitektūras izmaiņām derētu arī izpētīt, kā papildus valodas modeļa izmantošana spēj izmainīt rezultātus.

Noteikti ir nepieciešams arī izmēģināt rezultātus uz citām datu kopām. Kā arī iespējams pārbaudīt rezultātus modeļiem uz citām valodām. Varbūt cita veida stila pārneses pārveidojumiem būtu atšķirīgi rezultāti, līdz ar to, ir noderīgi arī izpētīt cita stila pārneses uzdevumus, piemēram, uzliekot visiem runātājiem vienu ātrumu vai arī vienu emocijas stilu vai vēl ko citu. Kā arī nepieciešams izpētīt vai un kā rezultātu precizitāti ietekmē runas uzlabošanas un trokšņu noņemšanas modeļi.

Visbeidzot ir noderīgi izpētīt citas iespējamos veidus, kā iegūt skaitliskus rezultātus un pārbaudīt cik labus rezultātus uzrāda citas metrikas, piemēram, fonēmu kļūdas daudzums, kā tas tika pieminēts 2.3. tabulas 22. pētījumā.

SECINĀJUMI

Darbā tika veikta salīdzināšana runas atpazīšanas priekšapstrādes metodēm jeb precīzāk, kā izmainās runas atpazīšanas rezultāti veicot runas stila pārnesi. Dotajā darbā tika iegūti vārdu kļūdas un rakstzīmju kļūdas metriku rezultāti audio failiem no runas atpazīšanas modeļa.

Tika izvēlēta VCTK datu kopa, kas satur vairākus angļu valodas runātājus, kur katram runātājam ir vairāki audio faili. Datu kopa nodrošināja runātāju daudzveidību, kā arī daudzie viena runātāja faili nodrošināja veiksmīgu iespēju apmācīt atsevišķos modeļus.

Salīdzināšanas process tika veikts sākotnēji apmācot Whisper (Radford, Kim et al., 2022) runas atpazīšanas modeli uz visas datu kopas un iegūstot metriku rezultātus, kas tālāk tika izmantoti kā atskaites punkts, lai salīdzinātu uzlabojumus. No datu kopas runātājiem tika atrasti pieci, kas bija ar visaugstāko kosinusu līdzības atšķirību no pārējās datu kopas, un kas tika izmantoti tālāk kā mērķa runātāji. Uz katru no šiem pieciem mērķa runātājiem (p254, p287, p304, p317 un 0363) tika apmācīts runas atpazīšanas modelis, kā arī visi audio faili ar runas stila pārneses FreeVC (Li, Tu et al., 2022) modeļa palīdzību pārveidoti uz šī apmācītā runātāja stilu. Beigās tika iegūti runas atpazīšanas rezultāti katram no mērķa runātājiem ar runas stila pārnesi uz šo runātāju un salīdzināti ar visas datu kopas rezultātiem.

Vislabāko uzlabojumu rādīja modelis, kas apmācīts uz p254 runātāju kopā ar datu kopas stila pārnesi uz šo runātāju. Tas spēja uzlabot WER metrikas rezultātus 73% no datu kopas runātājiem un normalizētās CER metrikas rezultātus pat 91% no datu kopas runātājiem. Arī p317 runas atpazīšanas modelis parādīja sākot no 67% līdz pat 74% runātāju rezultātu uzlabojumus, ar stila pārnesi uz šo mērķa runātāju. p304 runas atpazīšanas modelis uzlaboja no 55% datu kopas runātāju rezultātus līdz 81% pārveidojot datu kopas runāšanas stilu uz p304 runātāju. Pārnesot runas stilu uz p363 un izmantojot runas atpazīšanas modeli, kas apmācīts uz šo runātāju, metriku rezultāti tika uzlaboti starp 34% un 72% no runātājiem. Visbeidzot p287 runas atpazīšanas modelis ar runas stila pārnesi uz p287 uzlaboja metriku rezultātus no 35% runātāju līdz 74% runātāju.

Ir noderīgi arī piebilst, ka neviens no mērķa runātāju runas atpazīšanas modeļiem izņemot pašu p304 neuzlaboja WER rezultātus p304 runātājam, kā arī vienīgais, kas CER rezultātus spēja uzlabot, neskaitot pašu p304, bija p287. Kā arī p254 un p317 runātāju modeļi spēja uzlabot vidējo datu kopas rezultātu visās metrikās, kamēr pārējie to spēja uzlabot tikai rakstzīmes

metrikām. Vēl p304 un p254 runātāji spēja uzlabot CER un normalizētos CER visiem mērķa runātājiem.

Pēc veiktajiem uzdevumiem un iegūtajiem rezultātiem ir iespējams secināt, ka izmantojot runas stila pārnesi kā priekšapstrādes metodi runas atpazīšanai, rezultātus ir iespējams uzlabot kā minimums 34% runātāju un maksimāli pat 90% runātājiem no visas datu kopas.

Salīdzinot iegūtos rezultātus ar rezultātiem no sistemātiskās literatūras analīzes, kuros izmanto VCTK datu kopu un WER un CER metrikas, 35. pētījumā ((Maimon & Adi, 2022), atrodams 2.11. tabulā) WER bija 11.7, un šajā darbā visi mērķa runātāju modeļu vidējais datu kopas rezultāts 35. pētījuma rezultātu pārspēj. 35. pētījumā tika iegūts arī 5.9 CER, kur šī darba mērķa runātāju modeļi ar datu kopas vidējo vērtību nav spējīgi pārspēt, taču četri no pieciem runātājiem šo rezultātu tieši savam runātājam ir ieguvuši labāku. 37. pētījumā ((Li, Tu et al., 2022), atrodams 2.11. tabulā) norādīts 4.23 WER un 1.46 CER, kur CER diemžēl nevienam no šī darba runātāju modeļiem nav izdevies pārspēt, toties p317 runa stila pārnese un runas atpazīšanas modelis ir ieguvis 2.52 WER uz p317 runātāja failiem. Jāņem gan vērā, ka 37. pētījumā tika izmantotas nejauši izvēlēti 200 audio no VCTK datu kopas un 200 no LibriTTS. 27. pētījumā ((Liu, Yang et al., 2018), atrodams 2.15. tabulā) tika iegūts 0.36, ko izsakot procentuāli kā tiek darīts šajā darbā būtu 36 CER, tādējādi, visi šajā darbā iegūtie rezultāti ir labāki par 27. pētījuma rezultātiem.

IZMANTOTĀ LITERATŪRA

Abdulatif, S., Cao, R. & Yang, B. *CMGAN: Conformer-Based Metric-GAN for Monaural Speech Enhancement*. 2022. arXiv: 2209.11112 [cs.SD]. Pieejams: https://arxiv.org/pdf/2209.11112.pdf.

Agarwal, S., Ganapathy, S. & Takahashi, N. *Leveraging Symmetrical Convolutional Transformer Networks for Speech to Singing Voice Style Transfer*. 2022. arXiv: 2208.12410 [cs.SD]. Pieejams: https://arxiv.org/pdf/2208.12410.pdf.

Aggarwal, C. C. *Neural Networks and Deep Learning. A Textbook*. Cham: Springer, 2018, 497. lpp. ISBN: 978-3-319-94462-3.

AlBadawy, E. A. & Lyu, S. *Voice Conversion Using Speech-to-Speech Neuro-Style Transfer*. 2020. Pieejams: http://www.interspeech2020.org/uploadfile/pdf/Thu-3-4-11.pdf.

Baevski, A., Zhou, H., Mohamed, A. & Auli, M. *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. 2020. arXiv: 2006.11477 [cs.CL]. Pieejams: https://arxiv.org/pdf/2006.11477.pdf.

Bansal, R. *Read it to me: An emotionally aware Speech Narration Application*. 2022. arXiv: 2209.02785 [cs.SD].

Chan, W., Park, D., Lee, C., Zhang, Y., Le, Q. & Norouzi, M. *SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network*. 2021. arXiv: 2104.02133 [cs.CL]. Pieejams: https://arxiv.org/pdf/2005.08100.pdf.

Chollet, F. *Deep Learning with Python*. 2017. ISBN: 9781617294433.

Cui, J. & Bleeck, S. *Parallel Gated Neural Network With Attention Mechanism For Speech Enhancement*. 2022. arXiv: 2210.14509 [cs.SD].

Defossez, A., Synnaeve, G. & Adi, Y. *Real Time Speech Enhancement in the Waveform Domain*. 2020. arXiv: 2006.12847 [eess.AS]. Pieejams: https://arxiv.org/pdf/2006.12847.pdf.

Desplanques, B., Thienpondt, J. & Demuynck, K. "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification". *Interspeech 2020*. ISCA, 2020, 3830.—3834. lpp.

Donahue, C., Li, B. & Prabhavalkar, R. *Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition*. 2018. arXiv: 1711.05747 [cs.SD].

Fortune Business Insights. *Speech and Voice Recognition Market Size, Share COVID-19 Impact Analysis, By Technology (Voice Recognition and Speech Recognition), By Deployment (Cloud and On-Premise), By End-user (Healthcare, IT and Telecommunications, Automotive, BFSI, Government, Legal, Retail, Travel and Hospitality, and Others), and Regional Forecast, 2022-2029*. en. 2022. Pieejams: https://www.fortunebusinessinsights.com/industry-reports/speech-and-voice-recognition-market-101382.

Grand View Research. *Voice And Speech Recognition Market Size, Share Trends Analysis Report By Function (Speech, Voice Recognition), By Technology (Artificial Intelligence Based, Non-Artificial Intelligence Based), By Vertical (Healthcare, BFSI), And Segment Forecasts, 2022 – 2030*. en. 2021. Pieejams: https://www.grandviewresearch.com/industry-analysis/voice-recognition-market.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. & Pang, R. *Conformer: Convolution-augmented Transformer for Speech Recognition*. 2020. arXiv: 2005.08100 [eess.AS]. Pieejams: https://arxiv.org/pdf/2005.08100.pdf.

Hong, J., Kim, M., Yoo, D. & Ro, Y. M. *Visual Context-driven Audio Feature Enhancement for Robust End-to-End Audio-Visual Speech Recognition*. 2022. arXiv: 2207.06020 [cs.SD].

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R. & Mohamed, A. *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units*. 2021. arXiv: 2106.07447 [cs.CL].

Innovatrics. *Equal Error Rate (EER)*. en-US. 2023. Pieejams: https://www.innovatrics.com/glossary/equal-error-rate-eer/.

Jiang, H., Murdock, C. & Ithapu, V. K. *Egocentric Deep Multi-Channel Audio-Visual Active Speaker Localization*. 2022. arXiv: 2201.01928 [cs.CV].

Kameoka, H., Kaneko, T., Tanaka, K. & Hojo, N. *StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks*. 2018. arXiv: 1806.02169 [cs.SD].

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N. E. Y., Yamamoto, R., Wang, X., Watanabe, S., Yoshimura, T. & Zhang, W. *A Comparative Study on Transfor-*

mer vs RNN in Speech Applications. 2019. Pieejams: https://ieeexplore.ieee.org/document/9003750.

Kim, J., Kong, J. & Son, J. *Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech*. 2021. arXiv: 2106.06103 [cs.SD].

Leung, K. *Evaluate OCR Output Quality with Character Error Rate (CER) and Word Error Rate (WER)*. en. 2021. Pieejams: https://towardsdatascience.com/evaluating-ocr-output-quality-with-character-error-rate-cer-and-word-error-rate-wer-853175297510.

Li, A., Zheng, C., Zhang, L. & Li, X. *Glance and Gaze: A Collaborative Learning Framework for Single-channel Speech Enhancement*. 2021. arXiv: 2106.11789 [cs.SD].

Li, B., Gulati, A., Yu, J., Sainath, T. N., Chiu, C.-C., Narayanan, A., Chang, S.-Y., Pang, R., He, Y., Qin, J., Han, W., Liang, Q., Zhang, Y., Strohman, T. & Wu, Y. *A Better and Faster End-to-End Model for Streaming ASR*. 2021. arXiv: 2011.10798 [eess.AS].

Li, J., Tu, W. & Xiao, L. *FreeVC: Towards High-Quality Text-Free One-Shot Voice Conversion*. 2022. arXiv: 2210.15418 [cs.SD]. Pieejams: https://arxiv.org/pdf/2210.15418.pdf.

Lightning-AI. *Short-Time Objective Intelligibility (STOI)*. 2022. Pieejams: https://torchmetrics.readthedocs.io/en/stable/audio/short_time_objective_intelligibility.html.

Lin, J.-h., Lin, Y. Y., Chien, C.-M. & Lee, H.-y. *S2VC: A Framework for Any-to-Any Voice Conversion with Self-Supervised Pretrained Representations*. 2021. arXiv: 2104.02901 [eess.AS].

Liu, D.-R., Yang, C.-Y., Wu, S.-L. & Lee, H.-Y. "Improving Unsupervised Style Transfer in end-to-end Speech Synthesis with end-to-end Speech Recognition". *2018 IEEE Spoken Language Technology Workshop (SLT)*. 2018, 640.—647. lpp. Pieejams: doi: 10.1109/SLT.2018.8639672.

Lüscher, C., Beck, E., Irie, K., Kitza, M., Michel, W., Zeyer, A., Schlüter, R. & Ney, H. *RWTH ASR Systems for LibriSpeech: Hybrid vs Attention*. 2019. arXiv: 1905.03072 [cs.CL].

Maimon, G. & Adi, Y. *Speaking Style Conversion With Discrete Self-Supervised Units*. 2022. arXiv: 2212.09730 [cs.SD]. Pieejams: https://arxiv.org/pdf/2212.09730.pdf.

MarketsandMarkets Research. *Speech and Voice Recognition Market by Deployment Mode (On-Cloud, On-Premises/Embedded), Technology (Speech*

Recognition, Voice Recognition), Vertical and Geography (Americas, Europe, APAC, Rest of the World) - Global Forecast to 2027. 2022. Pieejams: https://www.marketsandmarkets.com/Market-Reports/speech-voice-recognition-market-202401714.html.

Nercessian, S. *Differentiable WORLD Synthesizer-based Neural Vocoder With Application To End-To-End Audio Style Transfer*. 2022. arXiv: 2208.07282 [eess.AS].

Niekerk, B. van, Carbonneau, M.-A., Zaidi, J., Baas, M., Seute, H. & Kamper, H. *A Comparison of Discrete and Soft Speech Units for Improved Voice Conversion*. 2022. arXiv: 2111.02392 [eess.AS].

Noé, P.-G., Miao, X., Wang, X., Yamagishi, J., Bonastre, J.-F. & Matrouf, D. *Hiding speaker's sex in speech using zero-evidence speaker representation in an analysis/synthesis pipeline*. 2023. arXiv: 2211.16065 [eess.AS].

Nozaki, J., Kawahara, T., Ishizuka, K. & Hashimoto, T. *End-to-end Speech-to-Punctuated-Text Recognition*. 2022. arXiv: 2207.03169 [eess.AS].

Pascual, S., Bonafonte, A., Serrà, J. & Gonzalez, J. A. *Whispered-to-voiced Alaryngeal Speech Conversion with Generative Adversarial Networks*. 2018. arXiv: 1808.10687 [cs.SD].

Pascual, S., Serrà, J. & Bonafonte, A. *Time-domain speech enhancement using generative adversarial networks*. 2019. Pieejams: https://www.sciencedirect.com/science/article/pii/S0167639319301359.

Pasini, M. *MelGAN-VC: Voice Conversion and Audio Style Transfer on arbitrarily long samples using Spectrograms*. 2019. arXiv: 1910.03713 [eess.AS].

Qian, K., Zhang, Y., Chang, S., Yang, X. & Hasegawa-Johnson, M. *AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss*. 2019. arXiv: 1905.05879 [eess.AS].

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. & Sutskever, I. *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022. arXiv: 2212.04356 [eess.AS]. Pieejams: https://cdn.openai.com/papers/whisper.pdf.

Radzikowski, K., Wang, L., Yoshie, O. & Nowak, R. *Accent modification for speech recognition of non-native speakers using neural style transfer*. 2021. Pieejams: https://doi.org/10.1186/s13636-021-00199-3.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C.,

Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D. & Bengio, Y. *SpeechBrain: A General-Purpose Speech Toolkit*. arXiv:2106.04624. 2021. arXiv: 2106.04624 [eess.AS].

Ravanelli, M., Zhong, J., Pascual, S., Swietojanski, P., Monteiro, J., Trmal, J. & Bengio, Y. *Multi-task self-supervised learning for Robust Speech Recognition*. 2020. arXiv: 2001.09239 [eess.AS].

Rix, A., Beerends, J., Hollier, M. & Hekstra, A. *Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs*. 2001. Pieejams: http://www.recursosvoip.com/docs/english/pap465.pdf.

Ruder, S. *An overview of gradient descent optimization algorithms*. 2017. arXiv: 1609.04747 [cs.LG].

Sharma, R., He, W., Lin, J., Lakomkin, E., Liu, Y. & Kalgaonkar, K. *Ego-centric Audio-Visual Noise Suppression*. 2022. arXiv: 2211.03643 [cs.SD].

Steinmetz, C. J., Bryan, N. J. & Reiss, J. D. *Style Transfer of Audio Effects with Differentiable Signal Processing*. 2022. arXiv: 2207.08759 [cs.SD].

Tengan, E., Dietzen, T., Ruiz, S., Alkmim, M., Cardenuto, J. & Waterschoot, T. van. *Speech enhancement using ego-noise references with a microphone array embedded in an unmanned aerial vehicle*. 2022. arXiv: 2211.02690 [eess.AS].

Thompson, W. & Munster, G. *Annual Digital Assistant IQ Test*. en-US. 2019. Pieejams: https://deepwatermgmt.com/annual-digital-assistant-iq-test/.

Vasilev, I., Slater, D., Spacagna, G., Roelants, P. & Zocca, V. *Python Deep Learning: Exploring deep learning techniques and neural network architectures with PyTorch, Keras, and TensorFlow, 2nd Edition*. Packt Publishing, 2019. ISBN: 9781789349702. Pieejams: https://books.google.lv/books?id=ESKEDwAAQBAJ.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

Veaux, C., Yamagishi, J. & MacDonald, K. *SUPERSEDED - CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit*. 2017. Pieejams: https://datashare.ed.ac.uk/handle/10283/2651.

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y. & Saurous, R. A. *Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis*. 2018. arXiv: 1803.09017 [cs.CL].

Xiao, F., Guan, J., Kong, Q. & Wang, W. *Time-domain Speech Enhancement with Generative Adversarial Learning*. 2021. arXiv: 2103.16149 [cs.SD].

Xu, C., Hu, B., Li, Y., Zhang, Y., huang, shen, Ju, Q., Xiao, T. & Zhu, J. *Stacked Acoustic-and-Textual Encoding: Integrating the Pre-trained Models into Speech Translation Encoders*. 2021. arXiv: 2105.05752 [cs.CL].

Yang, Y., Pandey, A. & Wang, D. *Time-Domain Speech Enhancement for Robust Automatic Speech Recognition*. 2022. arXiv: 2210.13318 [eess.AS].

Yathish, V. *Loss Functions and Their Use In Neural Networks*. en. 2022. Pieejams: https://towardsdatascience.com/loss-functions-and-their-use-in-neural-networks-a470e703f1e9.

Yin, D., Zhao, Z., Tang, C., Xiong, Z. & Luo, C. *TridentSE: Guiding Speech Enhancement with 32 Global Tokens*. 2022. arXiv: 2210.12995 [eess.AS]. Pieejams: https://arxiv.org/pdf/2210.12995.pdf.

Yu, G., Li, A., Zheng, C., Guo, Y., Wang, Y. & Wang, H. *Dual-branch Attention-In-Attention Transformer for single-channel speech enhancement*. 2022. arXiv: 2110.06467 [cs.SD].

Zhang, Y., Qin, J., Park, D. S., Han, W., Chiu, C.-C., Pang, R., Le, Q. V. & Wu, Y. *Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition*. 2022. arXiv: 2010.10504 [eess.AS].

Zhao, J., Yang, H., Shareghi, E. & Haffari, G. *M-Adapter: Modality Adaptation for End-to-End Speech-to-Text Translation*. 2022. arXiv: 2207.00952 [cs.CL].

Zhou, K., Sisman, B., Liu, R. & Li, H. *Seen and Unseen emotional style transfer for voice conversion with a new emotional speech dataset*. 2021. arXiv: 2010.14794 [cs.SD]. Pieejams: https://arxiv.org/pdf/2010.14794.pdf.

PIELIKUMI

1. pielikums. Vispārēja informācija par runas atpazīšanas modeļu publikācijām.

Nr	Nosaukums	Links	Gads	Organizācija, valsts	Citātu skaits	Git code links
1	End-to-end Speech-to-Punctuated-Text Recognition (Nozaki, Kawahara et al., 2022)	https://arxiv.org/pdf/2207.03169.pdf	2022	Kioto universitāte, Japāna	0	-
2	Robust Speech Recognition via Large-Scale Weak Supervision (Radford, Kim et al., 2022)	https://cdn.openai.com/papers/whisper.pdf	2022	OpenAI	63	https://github.com/openai/whisper
18	Conformer: Convolution-augmented Transformer for Speech Recognition (Gulati, Qin et al., 2020)	https://arxiv.org/pdf/2005.08100.pdf	2020	Google Inc.	1144	-
19	Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition (Zhang, Qin et al., 2022)	https://arxiv.org/pdf/2010.10504.pdf	2022	Google Research, Brain Team	182	-
20	RWTH ASR Systems for LibriSpeech: Hybrid vs Attention - w/o Data Augmentation (Lüscher, Beck et al., 2019)	https://arxiv.org/pdf/1905.03072.pdf	2019	RWTH Aachen universitāte, AppTek GmbH, Vācija	204	-
21	A better and faster end-to-end model for streaming ASR (Li, Gulati et al., 2021)	https://arxiv.org/pdf/2011.10798.pdf	2021	Google LLC, ASV	71	-

Nr	Nosaukums	Links	Gads	Organizācija, valsts	Citātu skaits	Git code links
22	Multi-task self-supervised learning for robust speech recognition (Ravanelli, Zhong et al., 2020)	https://arxiv.org/pdf/2001.09239.pdf	2020	Monreālas universitāte, INRS/CRIM, Kanāda; Ročesteras Universitāte, Džona Hopkinsa universitāte, ASV; Katalonijas politehniskā universitāte, Spānija; Jaundienvidvelsas Universitāte, Austrālija; CIFAR	184	https://github.com/santi-pdp/pase
42	SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network (Chan, Park et al., 2021)	https://arxiv.org/pdf/2104.02133.pdf	2021	Google Research, Brain Team	55	-
43	HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units (Hsu, Bolte et al., 2021)	https://arxiv.org/pdf/2106.07447.pdf	2021	-	545	https://github.com/facebookresearch/fairseq/tree/main/examples/hubert
44	wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations (Baevski, Zhou et al., 2020)	https://arxiv.org/pdf/2006.11477.pdf	2020	FaceBook AI	1754	https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec

2. pielikums. Vispārēja informācija par runas attīrīšanas modeļu publikācijām.

Nr	Nosaukums	Links	Gads	Organizācija, valsts	Citātu skaits	Git code links
7	Egocentric Audio-Visual Noise Suppression (Sharma, He et al., 2022)	https://arxiv.org/pdf/2211.03643.pdf	2022	Kārnegija Melona universitāte, ASV	0	-
4	Real Time Speech Enhancement in the Waveform Domain (Defossez, Synnaeve et al., 2020)	https://arxiv.org/pdf/2006.12847.pdf	2020	Facebook AI Research, INRIA, PSL pētniecības universitāte, Francija	173	https://github.com/facebookresearch/denoiser
10	Glance and Gaze: A Collaborative Learning Framework for Single-channel Speech Enhancement (Li, Zheng et al., 2021)	https://arxiv.org/pdf/2106.11789.pdf	2021	Ķīnas Zinātņu akadēmija, Ķīnas Zinātņu akadēmijas universitāte, Harbinas Tehnoloģiju institūts, Ķīna	43	https://github.com/Andong-Li-speech/GaGNet
14	Parallel Gated Neural Network With Attention Mechanism For Speech Enhancement (Cui & Bleeck, 2022)	https://arxiv.org/ftp/arxiv/papers/2210/2210.14509.pdf	2022	Sauthemptonas universitāte, Apvienotā Karaliste	0	-
15	TridentSE: Guiding Speech Enhancement with 32 Global Tokens (Yin, Zhao et al., 2022)	https://arxiv.org/pdf/2210.12995.pdf	2022	Ķīnas Zinātnes un tehnoloģiju universitāte, Microsoft Research Asia, Ķīna	2	-

Nr	Nosaukums	Links	Gads	Organizācija, valsts	Citātu skaits	Git code links
24	Time-domain speech enhancement using generative adversarial networks (Pascual, Serrà et al., 2019)	https://www-sciencedirect-com.resursi.rtu.lv/science/article/pii/S0167639319301359	2019	Katalonija politehniskā universitāte, Telefónica Research, Spānija	18	https://github.com/santi-pdp/segan_pytorch
29	Time-domain Speech Enhancement with Generative Adversarial Learning (Xiao, Guan et al., 2021)	https://arxiv.org/pdf/2103.16149.pdf	2021	Harbinas Inženierzinātņu universitāte, ByteDance, Ķīna; Surejas universitāte, Apvienotā Karaliste	4	https://github.com/littleflyingsheep/tsegan
30	Dual-branch Attention-In-Attention Transformer for single-channel speech enhancement (Yu, Li et al., 2022)	https://arxiv.org/pdf/2110.06467.pdf	2022	Ķīnas Komunikācijas universitāte, Ķīnas Zinātņu akadēmija, Ķīna	8	https://github.com/yuguochencuc/db-aiat
31	CMGAN: Conformer-Based Metric-GAN for Monaural Speech Enhancement (Abdulatif, Cao et al., 2022)	https://arxiv.org/pdf/2209.11112.pdf	2022	IEEE	4	https://github.com/ruizhecao96/cmgan

3. pielikums. Vispārēja informācija par runas stila pārneses modeļu publikācijām.

Nr	Nosaukums	Links	Gads	Organizācija, valsts	Citātu skaits	Git code links
5	Leveraging Symmetrical Convolutional Transformer Networks for Speech to Singing Voice Style Transfer (Agarwal, Ganapathy et al., 2022)	https://arxiv.org/pdf/2208.12410.pdf	2022	Indijas Zinātņu institūts, Indija; Sony Group Corporation, Japāna	1	-
13	Differentiable WORLD Synthesizer-based Neural Vocoder With Application To End-To-End Audio Style Transfer (Nercessian, 2022)	https://arxiv.org/pdf/2208.07282.pdf	2022	iZotope, Inc.	1	-
25	Whispered-to-voiced Alaryngeal Speech Conversion with Generative Adversarial Networks (Pascual, Bonafonte et al., 2018)	https://arxiv.org/pdf/1808.10687.pdf	2018	Katalonijas politehniskā universitāte, Telefonica Research, Malagas universitāte, Spānija	12	-
28	Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis (Wang, Stanton et al., 2018)	https://arxiv.org/pdf/1803.09017.pdf	2018	Google, Inc.	552	-
32	Seen and Unseen emotional style transfer for voice conversion with a new emotional speech dataset (Zhou, Sisman et al., 2021)	https://arxiv.org/pdf/2010.14794.pdf	2021	Singapūras Nacionālā universitāte, Singapūras Tehnoloģiju un dizaina universitāte, Singapūra	70	https://github.com/KunZhou9646/controllable_evc_code
33	MelGAN-VC: Voice Conversion and Audio Style Transfer on arbitrarily long samples using Spectrograms (Pasini, 2019)	https://arxiv.org/pdf/1910.03713.pdf	2019	Alma Mater Studiorum, Itālija	25	https://github.com/marcoppasini/MelGAN-VC

Nr	Nosaukums	Links	Gads	Organizācija, valsts	Citātu skaits	Git code links
34	Voice Conversion Using Speech-to-Speech Neuro-Style Transfer (AlBadawy & Lyu, 2020)	http://www.interspeech2020.org/uploadfile/pdf/Thu-3-4-11.pdf	2020	Albānijas universitāte, ASV	5	https://github.com/ebadawy/voice_conversion
35	Speaking Style Conversion With Discrete Self-Supervised Units (Maimon & Adi, 2022)	https://arxiv.org/pdf/2212.09730.pdf	2022	Jeruzalemes Ebreju universitāte, Jeruzaleme	1	https://github.com/gallilmaimon/DISSC
36	Hiding speaker's sex in speech using zero-evidence speaker representation in an analysis/synthesis pipeline (Noé, Miao et al., 2023)	https://arxiv.org/pdf/2211.16065.pdf	2022	Aviņonas datoru laboratorija, Francija; Nacionālais informātikas institūts, Japāna	0	https://github.com/nii-yamagishilab/speaker_sex_attribute_privacy
37	FreeVC: Towards High-Quality Text-Free One-Shot Voice Conversion (Li, Tu et al., 2022)	https://arxiv.org/pdf/2210.15418.pdf	2022	Uhaņas universitāte, Ķīna	2	https://github.com/olawod/freevc
38	A Comparison of Discrete and Soft Speech Units for Improved Voice Conversion (Niekerk, Carbonneau et al., 2022)	https://ieeexplore-ieee-org.resursi.rtu.lv/document/9746484	2022	Stellenbošas universitāte, Dienvidāfrika; Ubisoft La Forge, Kanāda	10	https://github.com/bshall/soft-vc
39	Style Transfer of Audio Effects with Differentiable Signal Processing (Steinmetz, Bryan et al., 2022)	https://arxiv.org/pdf/2207.08759.pdf	2022	Londonas Karalienes Marijas universitāte, ASV	4	https://github.com/adobe-research/DeepAFx-ST
40	S2VC: A Framework for Any-to-Any Voice Conversion with Self-Supervised Pretrained Representations (Lin, Lin et al., 2021)	https://arxiv.org/pdf/2104.02901.pdf	2021	Elektrotehnikas un datorzinātņu koledža, Taivānas Nacionālā universitāte, Taivāna	25	https://github.com/howard1337/S2VC

Nr	Nosaukums	Links	Gads	Organizācija, valsts	Citātu skaits	Git code links
41	AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss (Qian, Zhang et al., 2019)	https://arxiv.org/pdf/1905.05879.pdf	2019	Ilinoisas Universitāte, MIT-IBM Watson AI Lab, IBM Research, ASV	259	https://github.com/auspicious3000/autovc
45	StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks (Kameoka, Kaneko et al., 2018)	https://arxiv.org/pdf/1806.02169.pdf	2018	NTT Corporation, Japāna	236	https://github.com/kamepong/StarGAN-VC

4. pielikums. Vispārēja informācija par kombinēto modeļu publikācijām.

Nr	Nosaukums	Links	Gads	Organizācija, valsts	Citātu skaits	Git code links
3	Time-Domain Speech Enhancement for Robust Automatic Speech Recognition (Yang, Pandey et al., 2022)	https://arxiv.org/abs/2210.13318v2	2022	Ohaio štata universitāte, ASV	0	-
6	Visual Context-driven Audio Feature Enhancement for Robust End-to-End Audio-Visual Speech Recognition (Hong, Kim et al., 2022)	https://arxiv.org/pdf/2207.06020.pdf	2022	KAIST, Genesis Lab Inc., Dienvidkoreja	5	-
12	Read it to me: An emotionally aware Speech Narration Application (Bansal, 2022)	https://arxiv.org/pdf/2209.02785.pdf	2022	Dienvidkalifornijas Universitāte, ASV	0	-
17	A Comparative study on Transformer vs RNN in speech applications (Karita, Chen et al., 2019	https://arxiv.org/pdf/1909.06317.pdf	2019	NTT Komunikācijas zinātnes laboratorijas, Vasedas universitāte, LINE korporācija, Nagojas universitāte, Human Dataware Lab. Co., Ltd., Kioto universitāte, Japāna; Džona Hopkinsa universitāte, Mitsubishi Electric Research Laboratories, ASV; Šanhajas Jiao Tong universitāte, Ķīna	474	https://github.com/espnet/espnet

Nr	Nosaukums	Links	Gads	Organizācija, valsts	Citātu skaits	Git code links
23	Exploring speech enhancement with generative adversarial networks for robust speech recognition (Donahue, Li et al., 2018)	https://arxiv.org/pdf/1711.05747.pdf	2018	UC Sandjego Mūzikas un Google departaments, ASV	166	-
26	Accent modification for speech recognition of non-native speakers using neural style transfer (Radzikowski, Wang et al., 2021	https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-021-00199-3	2021	Vasedas Universitāte, Japāna; Varšavas Tehnoloģiju universitāte, Polija	7	-
27	Improving Unsupervised Style Transfer in end-to-end Speech Synthesis with end-to-end Speech Recognition (Liu, Yang et al., 2018)	https://ieeexplore-ieee-org.resursi.rtu.lv/document/8639672	2019	Taivānas Nacionālā universitāte, Taivāna	18	-

5. pielikums. Metriku rezultāti, kas iegūti no VCTK modeļa.

runātājs	WER	WER Norm.	CER	CER Norm.
p225	8.42	7.98	10.71	9.85
p226	7.89	7.41	11.27	10.39
p227	5.60	5.42	9.42	8.89
p228	7.32	7.13	11.00	10.14
p229	8.04	7.77	11.12	10.32
p230	6.40	6.20	10.34	9.63
p231	8.63	8.32	11.46	10.68
p232	8.31	7.89	11.39	10.52
p233	7.97	7.69	10.75	10.05
p234	4.06	3.98	8.13	7.68
p236	9.82	9.48	12.83	11.75
p237	7.76	7.45	11.22	10.42
p238	7.23	6.87	10.25	9.50
p239	10.17	9.66	12.58	11.45
p240	5.47	5.30	9.59	9.06
p241	7.48	7.22	11.42	10.57
p243	7.05	6.80	9.54	8.82
p244	8.58	8.40	12.23	11.15
p245	5.98	5.89	9.56	9.01
p246	6.79	6.51	10.80	9.95
p247	9.13	8.73	11.86	10.80
p248	10.38	9.96	13.39	12.21
p249	7.23	6.92	10.51	9.71
p250	9.18	8.75	11.89	11.00
p251	11.65	11.23	13.92	12.72
p252	9.08	8.67	11.22	10.41
p253	10.67	10.41	14.02	12.75
p254	8.70	8.34	11.57	10.69
p255	8.27	7.87	11.41	10.55
p256	7.76	7.46	11.47	10.64
p257	11.66	11.21	14.32	13.09
p258	9.20	9.02	12.31	11.49
p259	5.68	5.51	8.62	8.20
p260	8.92	8.59	12.01	10.99

runātājs	WER	WER Norm.	CER	CER Norm.
p261	6.72	6.49	10.52	9.72
p262	8.80	8.47	11.58	10.70
p263	10.62	10.15	12.94	12.00
p264	10.58	10.05	12.96	11.88
p265	8.00	7.82	11.43	10.62
p266	11.05	10.62	14.05	12.85
p267	6.93	6.77	10.71	9.97
p268	6.34	6.06	9.05	8.50
p269	7.54	7.27	10.97	10.16
p270	7.84	7.49	11.06	10.13
p271	5.63	5.47	9.96	9.30
p272	9.50	9.15	12.55	11.63
p273	8.97	8.56	11.81	10.82
p274	7.00	6.69	9.68	9.06
p275	9.93	9.52	12.82	11.81
p276	6.53	6.19	9.54	8.90
p277	13.52	12.72	15.89	14.44
p278	7.67	7.37	10.71	9.97
p279	11.52	10.96	13.61	12.47
p280	7.24	6.99	9.53	8.85
p281	8.19	7.88	11.34	10.51
p282	10.56	10.30	13.15	12.23
p283	9.13	8.72	12.22	11.20
p284	8.93	8.64	12.14	11.22
p285	12.61	12.16	15.67	14.19
p286	7.28	6.96	10.68	9.75
p287	5.73	5.51	9.65	8.98
p288	8.16	7.90	11.39	10.61
p292	7.25	7.03	10.25	9.64
p293	11.17	10.55	13.27	12.11
p294	6.82	6.58	9.70	9.13
p295	12.63	12.08	14.33	13.01
p297	9.90	9.51	11.75	10.92
p298	9.37	8.97	11.46	10.50
p299	6.70	6.42	9.85	9.19

runātājs	WER	WER Norm.	CER	CER Norm.
p300	6.65	6.27	9.79	9.01
p301	6.58	6.29	9.73	9.01
p302	10.00	9.62	12.14	11.19
p303	9.43	8.96	11.82	10.77
p304	8.18	7.93	11.25	10.42
p305	8.30	7.89	10.74	9.88
p306	9.59	9.08	12.28	11.17
p307	6.50	6.17	9.92	9.15
p308	9.28	8.86	11.26	10.50
p310	9.29	8.83	11.96	11.02
p311	8.34	7.96	10.96	10.16
p312	9.66	9.32	12.45	11.55
p313	9.79	9.31	12.44	11.38
p314	7.03	6.71	10.48	9.70
p316	6.81	6.57	10.13	9.52
p317	7.52	7.21	9.97	9.32
p318	5.62	5.40	9.57	8.92
p323	6.70	6.51	9.71	9.03
p326	11.73	11.27	13.60	12.45
p329	6.77	6.61	10.33	9.61
p330	7.11	6.75	9.93	9.23
p333	7.50	7.16	10.30	9.59
p334	9.57	9.13	12.39	11.35
p335	11.47	10.99	13.00	12.01
p336	7.66	7.47	10.72	10.06
p339	5.91	5.62	9.86	9.08
p340	8.78	8.48	11.73	10.93
p341	6.69	6.43	9.77	9.16
p343	4.83	4.67	9.38	8.77
p345	9.06	8.74	11.80	11.03
p347	12.79	11.88	15.07	13.34
p351	10.99	10.44	13.42	12.24
p360	8.67	8.43	12.21	11.25
p361	7.03	6.92	11.15	10.31
p362	6.43	6.25	10.30	9.53

runātājs	WER	WER Norm.	CER	CER Norm.
p363	6.58	6.39	10.57	9.86
p364	7.29	6.92	11.74	10.73
p374	8.08	7.73	11.26	10.43
p376	6.81	6.50	10.02	9.36
Vidējais uz visiem	8.33	8.00	11.37	10.50

6. pielikums. Metriku rezultāti, kas iegūti no p304 modeļa.

runātājs	WER	WER Norm.	CER	CER Norm.
p225	11.77	11.26	12.24	10.88
p226	6.62	6.33	8.72	7.81
p227	5.93	5.65	7.67	6.99
p228	6.01	5.87	7.66	7.08
p229	8.23	7.84	9.48	8.52
p230	10.50	9.71	12.12	10.60
p231	12.37	11.81	13.85	12.21
p232	7.54	7.27	9.58	8.69
p233	9.14	8.77	10.19	9.14
p234	3.06	2.98	5.93	5.53
p236	10.16	9.51	11.85	10.38
p237	6.07	5.88	8.52	7.80
p238	5.74	5.52	7.92	7.15
p239	15.68	15.16	15.84	14.04
p240	5.12	4.99	7.33	6.79
p241	5.77	5.49	7.96	7.28
p243	7.00	6.78	8.05	7.29
p244	10.94	10.71	12.84	11.37
p245	5.77	5.73	7.20	6.72
p246	4.31	4.02	6.56	5.99
p247	8.73	8.26	10.13	8.90
p248	13.06	12.43	13.45	11.96
p249	9.71	9.01	11.29	9.94
p250	12.90	12.32	13.78	12.23
p251	21.10	19.92	21.00	18.27
p252	8.72	8.30	9.67	8.60

runātājs	WER	WER Norm.	CER	CER Norm.
p253	10.96	10.79	12.96	11.55
p254	7.94	7.52	9.69	8.68
p255	8.22	7.87	9.55	8.63
p256	5.87	5.62	8.24	7.42
p257	13.86	13.35	15.13	13.47
p258	11.67	11.11	12.94	11.53
p259	5.49	5.08	7.10	6.48
p260	7.64	7.36	9.48	8.48
p261	7.61	7.23	9.58	8.61
p262	9.06	8.57	10.77	9.56
p263	13.59	12.92	14.22	12.72
p264	13.60	12.75	13.71	12.13
p265	7.39	6.92	9.44	8.42
p266	18.32	17.46	18.92	16.48
p267	9.82	9.46	11.25	10.07
p268	4.78	4.50	6.76	6.21
p269	7.39	7.01	9.25	8.39
p270	6.90	6.50	8.14	7.34
p271	4.17	3.87	7.23	6.59
p272	7.32	6.88	9.13	8.23
p273	9.63	9.17	10.56	9.56
p274	6.47	6.09	7.76	6.95
p275	9.96	9.63	11.00	9.97
p276	7.75	7.38	9.06	8.23
p277	18.61	16.59	18.90	15.91
p278	7.61	7.29	9.04	8.17
p279	13.79	13.17	14.62	12.94
p280	7.45	7.15	8.82	7.96
p281	10.12	9.71	11.21	10.01
p282	11.22	10.74	12.92	11.61
p283	19.77	18.59	19.27	16.51
p284	8.25	7.93	10.13	9.07
p285	14.59	13.91	15.93	14.04
p286	5.65	5.44	7.37	6.68
p287	7.08	6.86	8.58	7.83

runātājs	WER	WER Norm.	CER	CER Norm.
p288	7.10	6.99	8.38	7.77
p292	5.53	5.33	7.33	6.68
p293	19.44	17.87	19.47	16.59
p294	5.15	4.91	7.32	6.64
p295	15.40	14.57	16.35	14.46
p297	8.69	8.30	9.32	8.49
p298	8.48	8.11	10.15	9.00
p299	4.81	4.69	7.28	6.59
p300	6.46	6.20	8.34	7.44
p301	4.29	4.14	6.94	6.31
p302	10.71	10.23	11.64	10.50
p303	7.53	7.02	9.59	8.58
p304	3.88	3.85	5.58	5.30
p305	10.14	9.42	11.00	9.85
p306	9.74	9.27	10.87	9.59
p307	4.22	4.05	6.81	6.25
p308	8.90	8.58	10.26	9.35
p310	8.19	7.87	10.07	9.08
p311	6.79	6.41	8.24	7.48
p312	7.55	7.31	9.79	8.86
p313	11.70	11.17	14.14	12.35
p314	8.43	8.01	9.82	8.87
p316	4.83	4.54	6.95	6.39
p317	7.39	7.12	8.69	7.91
p318	3.52	3.41	6.13	5.64
p323	6.76	6.57	8.72	7.82
p326	14.49	13.98	15.72	13.68
p329	5.33	5.14	7.58	6.92
p330	4.08	3.89	6.79	6.15
p333	5.78	5.54	7.85	7.19
p334	9.22	8.70	10.59	9.41
p335	12.58	11.90	12.94	11.53
p336	8.75	8.43	10.26	9.31
p339	5.13	4.95	8.05	7.19
p340	10.81	10.13	11.99	10.69

runātājs	WER	WER Norm.	CER	CER Norm.
p341	3.46	3.38	6.30	5.84
p343	4.16	3.84	7.42	6.68
p345	7.85	7.47	9.86	8.93
p347	13.63	12.95	14.37	12.62
p351	12.95	12.29	13.87	12.23
p360	9.65	9.31	10.77	9.74
p361	6.06	5.99	9.35	8.43
p362	4.94	4.66	7.41	6.76
p363	4.23	4.12	7.22	6.61
p364	6.57	6.30	9.18	8.26
p374	6.67	6.31	9.09	8.14
p376	4.88	4.77	6.99	6.42
vidējais uz visiem	8.69	8.28	10.30	9.21

7. pielikums. Metriku rezultāti, kas iegūti no p317 modeļa.

runātājs	WER	WER Norm.	CER	CER Norm.
p225	11.69	11.16	12.71	11.07
p226	7.52	6.99	11.06	9.72
p227	6.62	6.41	10.16	9.11
p228	5.17	5.07	8.51	7.80
p229	6.82	6.37	10.80	9.64
p230	7.31	6.91	11.29	10.19
p231	12.34	11.66	15.00	13.15
p232	8.37	7.98	11.01	9.94
p233	6.81	6.44	9.14	8.32
p234	3.97	3.87	7.43	6.91
p236	11.40	10.62	12.83	11.17
p237	7.09	6.84	10.49	9.53
p238	7.47	6.95	8.87	7.85
p239	13.34	12.57	14.12	12.38
p240	3.74	3.63	7.25	6.67
p241	7.16	6.56	11.24	10.04
p243	7.61	7.30	9.20	8.42
p244	10.79	10.42	13.10	11.67

runātājs	WER	WER Norm.	CER	CER Norm.
p245	5.38	5.28	8.64	8.02
p246	3.67	3.44	7.78	7.06
p247	8.83	8.15	11.37	9.98
p248	10.30	9.95	11.98	10.58
p249	8.45	8.01	12.48	11.25
p250	11.64	11.20	13.19	11.66
p251	22.10	20.81	24.05	20.65
p252	10.56	9.88	12.70	11.22
p253	8.61	8.44	12.04	10.93
p254	9.90	9.25	12.08	10.59
p255	7.48	6.97	10.94	9.65
p256	5.40	5.16	8.63	7.92
p257	12.03	11.30	14.04	12.50
p258	11.61	11.17	14.12	12.62
p259	5.19	4.97	7.81	7.30
p260	7.25	7.01	10.69	9.65
p261	6.18	5.85	9.20	8.29
p262	9.29	8.82	12.46	11.01
p263	15.76	14.75	17.26	15.21
p264	15.35	14.25	16.56	14.40
p265	5.94	5.47	9.93	8.83
p266	13.69	13.04	16.53	14.75
p267	9.45	8.95	12.65	11.15
p268	4.12	3.82	6.39	5.85
p269	6.70	6.37	10.52	9.45
p270	8.42	7.82	10.82	9.73
p271	4.84	4.53	9.94	9.02
p272	8.90	8.49	12.05	10.77
p273	13.75	12.81	15.50	13.58
p274	6.82	6.44	8.87	7.95
p275	10.40	9.91	13.54	12.10
p276	6.12	5.83	9.69	8.81
p277	15.09	13.88	18.42	15.97
p278	8.69	8.33	11.10	10.01
p279	10.41	10.06	12.56	11.38

runātājs	WER	WER Norm.	CER	CER Norm.
p280	6.88	6.53	8.41	7.62
p281	13.23	12.60	16.16	14.34
p282	8.70	8.35	11.91	10.80
p283	14.28	13.42	16.95	14.86
p284	7.64	7.26	10.48	9.29
p285	15.97	15.18	17.64	15.47
p286	6.65	6.28	8.87	7.92
p287	8.06	7.66	10.91	9.74
p288	5.38	5.27	8.47	7.70
p292	5.59	5.43	9.05	8.33
p293	16.15	15.03	18.93	16.28
p294	2.74	2.63	7.22	6.69
p295	13.36	12.52	16.33	14.32
p297	5.23	5.05	8.48	7.80
p298	8.84	8.39	10.91	9.65
p299	3.84	3.70	6.86	6.21
p300	4.61	4.25	7.72	6.76
p301	2.83	2.70	6.90	6.36
p302	8.71	8.22	11.18	10.17
p303	5.01	4.70	8.57	7.81
p304	11.29	10.90	13.64	12.26
p305	5.95	5.56	9.14	8.36
p306	6.80	6.43	10.02	9.07
p307	2.90	2.80	6.42	5.93
p308	7.90	7.56	10.69	9.60
p310	7.21	6.85	10.73	9.66
p311	4.96	4.66	9.00	8.34
p312	4.69	4.51	8.84	8.10
p313	8.98	8.49	13.11	11.60
p314	6.23	6.00	9.02	8.28
p316	4.68	4.49	8.82	8.23
p317	2.52	2.42	6.47	6.04
p318	1.93	1.85	5.88	5.44
p323	5.25	5.03	7.54	6.80
p326	15.83	14.99	18.27	15.80

runātājs	WER	WER Norm.	CER	CER Norm.
p329	4.22	4.06	8.21	7.50
p330	2.92	2.88	7.00	6.51
p333	3.35	3.20	6.30	5.83
p334	7.70	7.23	11.63	10.34
p335	9.86	9.41	12.23	10.97
p336	6.75	6.49	9.55	8.64
p339	3.18	3.02	7.61	6.97
p340	7.47	7.13	11.12	10.06
p341	2.29	2.23	6.10	5.64
p343	2.33	2.16	7.39	6.75
p345	5.63	5.23	10.00	9.20
p347	13.18	12.30	14.77	12.83
p351	8.77	8.35	12.62	11.36
p360	7.93	7.69	12.25	10.99
p361	4.39	4.37	9.46	8.63
p362	3.70	3.47	7.49	6.88
p363	4.58	4.38	9.42	8.52
p364	6.51	6.00	11.69	10.38
p374	5.63	5.21	9.41	8.40
p376	2.98	2.90	6.38	6.00
Vidējais uz visiem	7.85	7.44	10.93	9.79

8. pielikums. Metriku rezultāti, kas iegūti no p363 modeļa.

runātājs	WER	WER Norm.	CER	CER Norm.
p225	11.59	10.97	11.06	9.94
p226	8.52	8.06	9.87	8.77
p227	8.13	7.61	9.22	8.13
p228	7.07	6.88	8.81	8.01
p229	8.89	8.41	9.97	8.88
p230	11.28	10.45	12.79	11.11
p231	12.57	11.93	13.33	11.68
p232	8.67	8.30	10.72	9.62
p233	9.62	9.20	10.78	9.54
p234	5.22	5.06	6.83	6.24

runātājs	WER	WER Norm.	CER	CER Norm.
p236	14.93	13.98	15.40	13.26
p237	6.75	6.59	8.79	8.03
p238	9.71	8.98	10.71	9.39
p239	18.97	17.94	18.56	16.08
p240	5.60	5.32	7.74	7.00
p241	7.36	6.97	8.75	7.90
p243	8.56	8.25	9.22	8.24
p244	14.30	13.81	15.06	13.18
p245	6.67	6.44	8.13	7.38
p246	5.23	4.87	7.75	6.83
p247	12.46	11.57	12.47	10.89
p248	17.46	16.30	18.02	15.29
p249	12.06	11.20	12.90	11.27
p250	12.80	12.15	13.84	12.21
p251	24.04	22.38	23.94	20.30
p252	12.09	11.41	12.27	10.69
p253	12.33	11.85	13.94	12.20
p254	9.47	8.97	10.78	9.54
p255	10.38	9.69	11.22	9.79
p256	6.59	6.37	8.43	7.63
p257	17.37	16.20	17.69	15.23
p258	11.58	11.01	12.96	11.38
p259	6.89	6.50	8.44	7.54
p260	7.55	7.44	9.36	8.45
p261	9.72	9.17	10.74	9.60
p262	13.01	12.24	13.60	11.79
p263	20.60	19.14	20.12	17.33
p264	19.54	18.23	19.01	16.38
p265	10.31	9.69	11.77	10.31
p266	21.41	19.98	21.14	18.15
p267	12.82	12.17	13.52	11.75
p268	4.98	4.73	6.59	6.07
p269	8.62	8.18	9.87	8.90
p270	8.56	7.90	9.58	8.44
p271	5.17	4.81	7.55	6.81

runātājs	WER	WER Norm.	CER	CER Norm.
p272	9.16	8.76	10.78	9.61
p273	12.19	11.46	12.06	10.58
p274	8.12	7.62	9.04	7.96
p275	11.48	11.02	12.42	11.12
p276	9.80	9.26	10.87	9.67
p277	20.31	18.55	20.69	17.42
p278	8.14	7.73	8.72	7.89
p279	13.45	12.92	13.83	12.27
p280	9.11	8.61	9.89	8.75
p281	13.97	13.27	14.91	12.90
p282	11.58	10.97	12.20	10.84
p283	19.49	18.34	19.15	16.39
p284	8.98	8.60	10.75	9.38
p285	15.18	14.46	15.25	13.49
p286	7.02	6.64	8.47	7.59
p287	6.58	6.36	8.30	7.45
p288	8.63	8.33	10.19	9.09
p292	7.15	6.76	8.65	7.70
p293	21.11	19.34	20.94	17.55
p294	5.24	4.94	7.54	6.74
p295	19.20	17.76	18.97	16.33
p297	8.22	7.86	9.19	8.21
p298	9.26	8.69	10.70	9.36
p299	4.81	4.65	6.94	6.28
p300	5.91	5.70	7.94	7.03
p301	3.92	3.76	6.27	5.68
p302	11.44	10.79	11.39	10.26
p303	7.56	7.10	8.98	8.00
p304	13.57	12.95	13.92	12.23
p305	9.26	8.48	10.02	8.86
p306	10.07	9.28	10.94	9.55
p307	3.92	3.75	6.17	5.61
p308	9.57	9.01	9.89	8.84
p310	8.71	8.23	9.72	8.69
p311	7.11	6.62	8.47	7.67

runātājs	WER	WER Norm.	CER	CER Norm.
p312	8.10	7.77	9.62	8.67
p313	13.94	13.17	14.86	12.96
p314	9.30	8.74	10.34	9.20
p316	5.29	5.00	7.19	6.56
p317	6.84	6.47	8.07	7.36
p318	3.08	2.94	6.24	5.62
p323	8.36	7.98	9.54	8.49
p326	15.57	14.65	17.03	14.51
p329	5.53	5.33	7.24	6.57
p330	4.73	4.43	6.83	6.12
p333	5.52	5.25	7.20	6.51
p334	9.00	8.44	10.11	8.92
p335	14.36	13.63	14.00	12.39
p336	8.87	8.39	10.02	8.88
p339	4.59	4.36	7.27	6.45
p340	11.68	11.10	13.57	11.79
p341	3.52	3.36	5.74	5.29
p343	4.03	3.79	6.96	6.27
p345	5.45	5.21	7.58	6.97
p347	17.03	15.97	17.26	14.71
p351	13.72	12.94	14.42	12.55
p360	9.67	9.42	11.20	9.91
p361	5.91	5.76	9.05	8.06
p362	6.57	5.97	7.99	7.17
p363	3.00	2.94	5.67	5.25
p364	9.15	8.34	11.78	10.04
p374	6.11	5.74	8.45	7.46
p376	3.93	3.87	6.25	5.72
Vidējais uz visiem	10.00	9.43	11.19	9.86

9. pielikums. Metriku rezultāti, kas iegūti no p287 modeļa.

runātājs	WER	WER Norm.	CER	CER Norm.
p225	9.97	9.73	10.29	9.26
p226	6.66	6.28	8.88	7.93

runātājs	WER	WER Norm.	CER	CER Norm.
p227	5.45	5.28	7.61	6.98
p228	7.20	6.96	9.03	8.22
p229	6.94	6.73	8.57	7.84
p230	10.20	9.35	11.96	10.46
p231	10.19	9.74	11.66	10.42
p232	9.00	8.58	10.47	9.48
p233	8.24	7.77	9.54	8.52
p234	5.12	4.78	7.46	6.77
p236	10.01	9.30	12.27	10.75
p237	8.74	8.30	10.52	9.50
p238	7.66	7.07	9.74	8.60
p239	13.99	13.17	14.46	12.74
p240	4.73	4.56	6.82	6.36
p241	8.14	7.48	10.04	8.93
p243	6.76	6.46	7.78	7.06
p244	11.69	11.26	13.34	11.75
p245	6.79	6.58	8.24	7.55
p246	5.31	4.91	8.30	7.44
p247	12.39	11.38	12.78	11.13
p248	15.01	14.24	15.13	13.37
p249	14.06	12.89	14.79	12.77
p250	11.97	11.28	13.21	11.64
p251	24.94	22.98	23.95	20.32
p252	12.52	11.69	13.02	11.30
p253	9.24	8.89	11.56	10.30
p254	7.68	7.26	9.63	8.59
p255	9.31	8.87	10.20	9.22
p256	5.69	5.40	7.92	7.22
p257	14.72	13.78	15.97	13.96
p258	10.60	10.09	12.32	11.02
p259	5.56	5.19	7.42	6.72
p260	8.58	8.41	10.62	9.59
p261	9.90	9.38	11.42	10.17
p262	10.78	10.19	12.11	10.68
p263	18.41	17.41	17.95	15.71

runātājs	WER	WER Norm.	CER	CER Norm.
p264	16.50	15.48	16.92	14.81
p265	9.03	8.55	10.81	9.56
p266	22.30	21.27	22.52	19.50
p267	9.61	9.18	11.06	9.85
p268	3.59	3.42	6.08	5.59
p269	9.01	8.53	10.77	9.64
p270	9.09	8.40	9.91	8.83
p271	5.75	5.28	8.52	7.63
p272	9.23	8.58	10.79	9.60
p273	10.37	9.58	11.03	9.79
p274	5.72	5.38	7.16	6.49
p275	12.22	11.62	12.84	11.48
p276	6.44	6.10	8.94	7.97
p277	16.94	15.52	17.80	15.33
p278	8.51	7.90	9.97	8.87
p279	11.49	11.04	12.31	10.97
p280	7.53	7.16	8.64	7.80
p281	14.01	13.19	14.86	13.01
p282	10.83	10.33	12.15	10.88
p283	19.44	18.16	19.18	16.41
p284	9.90	9.46	11.62	10.25
p285	12.38	11.82	13.54	12.13
p286	5.32	5.07	7.54	6.78
p287	3.17	3.06	5.97	5.47
p288	9.28	8.99	10.66	9.63
p292	5.70	5.45	8.08	7.30
p293	22.98	21.16	23.00	19.35
p294	5.90	5.57	8.63	7.77
p295	18.58	17.08	19.10	16.36
p297	10.82	10.29	12.00	10.72
p298	10.09	9.46	11.38	10.04
p299	5.12	4.96	7.71	7.04
p300	7.01	6.65	9.21	8.16
p301	4.98	4.73	7.96	7.14
p302	13.61	12.86	14.38	12.75

runātājs	WER	WER Norm.	CER	CER Norm.
p303	8.42	8.01	10.71	9.47
p304	11.03	10.51	11.97	10.70
p305	12.41	11.35	12.78	11.30
p306	11.58	10.94	12.59	11.06
p307	5.07	4.85	7.57	6.90
p308	11.65	11.02	13.15	11.59
p310	8.56	8.19	10.59	9.52
p311	8.30	7.82	9.96	8.92
p312	9.00	8.66	11.45	10.27
p313	12.73	12.02	14.05	12.29
p314	8.60	8.13	9.72	8.73
p316	6.84	6.45	8.75	7.96
p317	8.07	7.71	9.58	8.59
p318	4.39	4.16	7.32	6.57
p323	7.06	6.68	8.86	7.94
p326	13.28	12.68	14.41	12.60
p329	6.45	6.22	8.56	7.78
p330	6.69	6.18	8.69	7.74
p333	8.14	7.59	9.55	8.56
p334	10.76	9.96	12.63	10.97
p335	12.12	11.34	13.02	11.53
p336	7.99	7.57	9.64	8.67
p339	5.32	5.10	8.24	7.38
p340	12.46	11.83	13.84	12.29
p341	4.42	4.22	7.49	6.82
p343	5.93	5.42	8.72	7.78
p345	7.65	7.16	10.04	9.03
p347	15.00	14.00	15.52	13.52
p351	15.88	15.17	16.24	14.16
p360	10.46	10.15	12.11	10.83
p361	7.88	7.65	10.86	9.61
p362	6.65	6.18	8.65	7.76
p363	4.51	4.42	7.46	6.86
p364	10.40	9.48	13.01	11.22
p374	5.55	5.20	8.14	7.36

runātājs	WER	WER Norm.	CER	CER Norm.
p376	5.18	5.01	7.56	6.90
Vidējais uz visiem	9.66	9.11	11.24	9.97

10. pielikums. Metriku rezultāti, kas iegūti no p254 modeļa.

runātājs	WER	WER Norm.	CER	CER Norm.
p225	8.40	8.11	9.35	8.31
p226	5.19	4.97	6.97	6.32
p227	4.33	4.21	6.42	5.84
p228	5.39	5.37	7.14	6.53
p229	6.33	6.11	8.11	7.24
p230	7.23	6.89	9.26	8.29
p231	10.22	9.86	11.52	10.22
p232	6.97	6.80	8.93	8.01
p233	6.82	6.47	8.31	7.40
p234	3.77	3.69	5.74	5.36
p236	10.45	9.94	11.92	10.44
p237	5.13	5.04	7.66	7.04
p238	6.61	6.28	8.03	7.21
p239	12.44	12.04	12.91	11.47
p240	3.56	3.49	5.94	5.47
p241	4.76	4.50	7.29	6.51
p243	7.12	6.97	7.64	6.95
p244	10.68	10.35	12.49	10.87
p245	4.81	4.72	6.41	5.94
p246	2.37	2.22	5.81	5.25
p247	8.46	8.04	9.02	8.04
p248	11.60	11.22	12.89	11.34
p249	7.98	7.58	9.57	8.45
p250	10.60	10.27	12.25	10.82
p251	19.34	18.16	19.35	16.83
p252	7.79	7.55	8.95	8.08
p253	8.10	7.95	10.27	9.17
p254	2.51	2.47	5.27	4.87
p255	6.68	6.28	8.13	7.29

runātājs	WER	WER Norm.	CER	CER Norm.
p256	3.52	3.47	5.71	5.31
p257	10.72	10.30	12.74	11.20
p258	7.47	7.33	9.52	8.65
p259	3.85	3.77	5.69	5.28
p260	7.36	7.14	9.16	8.27
p261	7.90	7.57	9.43	8.44
p262	7.78	7.60	9.25	8.37
p263	14.94	14.34	15.09	13.37
p264	14.59	13.98	14.67	12.88
p265	6.60	6.23	8.63	7.66
p266	17.91	17.12	18.54	16.02
p267	7.44	7.19	9.15	8.22
p268	3.38	3.26	5.13	4.74
p269	6.65	6.40	8.46	7.66
p270	5.61	5.40	7.04	6.47
p271	3.34	3.20	6.04	5.57
p272	6.28	6.18	8.15	7.41
p273	8.42	8.01	9.48	8.40
p274	3.99	3.89	5.63	5.17
p275	8.07	7.87	9.49	8.60
p276	6.34	6.06	7.87	7.12
p277	14.64	13.72	15.66	13.57
p278	6.80	6.50	8.17	7.35
p279	12.11	11.81	12.88	11.48
p280	6.14	5.95	6.84	6.24
p281	10.91	10.53	11.69	10.41
p282	9.24	8.99	10.47	9.52
p283	17.64	16.76	17.72	15.20
p284	7.70	7.56	9.70	8.65
p285	11.55	11.17	13.16	11.57
p286	5.26	5.13	7.14	6.42
p287	6.63	6.46	8.19	7.35
p288	6.62	6.43	8.41	7.65
p292	4.87	4.83	7.02	6.38
p293	16.46	15.43	16.96	14.55

runātājs	WER	WER Norm.	CER	CER Norm.
p294	4.16	4.03	6.75	6.14
p295	15.70	14.87	16.58	14.43
p297	7.42	7.27	8.01	7.38
p298	7.05	6.81	8.50	7.52
p299	3.54	3.46	6.05	5.50
p300	5.11	4.92	6.73	5.99
p301	3.05	2.98	5.28	4.78
p302	10.25	9.96	11.34	10.24
p303	5.83	5.60	7.75	6.91
p304	9.47	9.35	10.71	9.63
p305	7.21	6.74	8.75	7.88
p306	8.02	7.61	9.27	8.20
p307	3.21	3.18	5.58	5.08
p308	8.68	8.31	9.96	8.91
p310	6.91	6.60	8.94	8.00
p311	6.90	6.59	8.31	7.47
p312	6.61	6.46	9.08	8.14
p313	10.86	10.35	12.54	10.94
p314	6.66	6.42	8.60	7.75
p316	4.13	4.02	6.12	5.66
p317	6.43	6.19	7.82	7.12
p318	2.35	2.25	5.43	4.94
p323	5.09	4.93	7.21	6.46
p326	12.56	12.15	14.64	12.62
p329	4.17	4.10	6.67	6.05
p330	3.38	3.21	5.88	5.28
p333	4.93	4.72	7.29	6.66
p334	7.71	7.38	10.20	8.93
p335	8.59	8.26	9.48	8.45
p336	6.60	6.37	8.30	7.49
p339	4.57	4.37	7.57	6.69
p340	10.17	9.91	11.56	10.32
p341	2.35	2.29	5.02	4.59
p343	2.99	2.83	6.15	5.60
p345	6.05	5.83	8.29	7.56

runātājs	WER	WER Norm.	CER	CER Norm.
p347	10.99	10.52	11.75	10.47
p351	12.65	12.16	14.08	12.28
p360	7.34	7.21	9.09	8.22
p361	5.78	5.75	9.09	8.06
p362	5.16	5.05	6.89	6.29
p363	4.20	4.10	6.95	6.28
p364	7.45	6.92	10.53	9.04
p374	5.23	4.96	7.62	6.85
p376	3.73	3.69	5.62	5.22
Vidējais uz visiem	7.46	7.18	9.17	8.19