LATVIJAS UNIVERSITĀTE
DATORIKAS FAKULTĀTE

ANGĻU VALODAS RUNAS SINTĒZES MODEĻU SALĪDZINĀJUMS

KURSA DARBS

Autors: **Krišs Saulītis**
Studentu apliecības Nr.: ks18108
Darba vadītājs: Phd. Comp. Sc. Ēvalds Urtāns

RĪGA, 2024

ANOTĀCIJA

Šajā darbā tiek veikts angļu valodas runas sintēzes modeļu un sistēmu salīdzinājums. Šis pētījums koncentrējas uz sistemātisku literatūras apskatu un objektīvu runas sintēzes sistēmu salīdzinājumu, izmantojot testa teksta datu kopu modeļu novērtēšanai. Pētījuma metodoloģija ietver runas sintēzes modeļu konfigurāciju audio paraugu ģenerēšanai, ko tālāk izmanto modeļu salīdzināšanai, balstoties uz noteiktajiem kvalitātes un precizitātes kritērijiem. Vislabākos kvalitātes rādītājus uzrādīja CoMoSpeech modelis (MOS - 3.85), savukārt VITS modelis uzrādīja visaugstāko precizitāti (CER - 1.48%). Pētījumā padziļināti tiek apskatīts arī dažādu modeļu stipro un vājo pušu izvērtējums.

Darba kopējais apjoms ir 29 lappuses.

Atslēgvārdi: runas sintēze, dziļā mašīnmācīšanās, difonu apvienošana, literatūras analīze, objektīvs salīdzinājums

ABSTRACT

In this work, a comparison of English language speech synthesis models and systems is conducted. This study focuses on a systematic literature review and an objective comparison of speech synthesis systems, using a test text data set for model evaluation. The research methodology includes the configuration of speech synthesis models for generating audio samples, which are then used for comparing models based on established quality and precision criteria. The CoMoSpeech model showed the best quality indicators (MOS - 3.85), while the VITS model demonstrated the highest precision (CER - 1.48%). The study includes a detailed assessment of the strengths and weaknesses of various models.

The total amount of this work is 29 pages.

Keywords: speech synthesis, deep-learning, diphone concatenation, literature analysis, objective comparison

Saturs

1. Ievads **6**

2. Saistītie pētījumi **7**

2.1. Difonu apvienošana 7

2.2. Dziļā mašīnmācīšanās 8

2.3. Modeļi 9

2.4. Rādītāji 12

3. Sistemātiskā literatūras analīze **14**

3.1. Meklēšanas protokols 14

3.2. Kvalitātes kritēriji 17

4. Metodoloģija **19**

4.1. Datu kopas 19

4.2. Testa kopas sintēze 21

4.3. Rādītāji 21

4.3.1. WER un CER 21

4.3.2. NISQA 22

5. Rezultāti **23**

6. Secinājumi **25**

Bibliogrāfija **26**

Apzīmējumu saraksts

AI (Artificial Intelligence) - Mākslīgais intelekts

API (Application Programming Interfac) - Lietojumprogrammas saskarne

ASR (Automatic Speech Recognition) - Automātiskā runas atpazīšana

CER (Character Error Rate) - Simbolu kļūdas biežums

LSTM (Long Short-Term Memory Network) - Garās īslaicīgās atmiņas tīkls

MOS (Mean Opinion Score) - Vidējā viedokļa vērtējums

NHMM (Neural Hidden Markov Model) - Neirālais, apslēptais "Markova" modelis

NISQA (Non-Intrusive Speech Quality Assessment) - Neuzbāzīgs runas kvalitātes novērtējums

RNN (Recurrent Neural Network) - rekurentais neironu tīkls

RVW (Residual Vector Quantazion) - Atlikuma vektora kvantēšana

VQ (Vector Quantizing) - Vektora kvantēšana

WER (Word Error Rate) - Vārdu kļūdas biežums

1 Ievads

Runas sintēze ir viena no aktuālākajām un strauji augošajām tehnoloģiskajām jomām mūsdienu informācijas sabiedrībā. Veicot sīkāku izpēti *Semantic Scholar* datubāzē, izmantojot atslēgas vārdu "TTS", var secināt, ka ar katru gadu pieaug zinātnisko publikāciju skaits, kurās tiek pētīta runas sintēze. Pēdējo 10 gadu laikā par šo tēmu tika atrastas 3 060 publikācijas, no kurām 475 ir tikušas publicētas pēdējā gada laikā.

Runas sintēze ir process, kurā datorizētas sistēmas ģenerē cilvēka balss skanējumam līdzīgu runu. Šī tehnoloģija tiek plaši izmantota dažādās nozarēs, piemēram, virtuālo asistentu izstrādē, automātiskās runas atpazīšanas sistēmās un ekrāna lasītāju tehnoloģijās, lai palīdzētu cilvēkiem ar redzes traucējumiem. Turklāt, runas sintēzes modeļu attīstība ir cieši saistīta ar mākslīgā intelekta un mašīnmācīšanās progresu, īpaši ņemot vērā tās saikni ar dziļo mašīnmācīšanu un neironu tīklu tehnoloģijām. [23]

Šajā darbā tika atlasīti un apskatīti 19 nozīmīgi un aktuāli runas sintēzes modeļi angļu valodā, izvirzīta testa datu kopa, kvalitātes kritēriji un veikts objektīvs savstarpējais 9 modeļu salīdzinājums. Izmantojot iegūtos rezultātus, iespējams atlasīt labāko modeli, lai veiktu jaunu modeļu trenēšanu citās valodās.

Šī kursa darba mērķis ir veikt runas sintēzes sistēmu sistemātisko zinātniskās literatūras pārskatu un eksperimentālo salīdzināšanu starp angļu valodas runas sintēzes sistēmām.

Darba uzdevumi:

1. Izpētīt un apkopot dažādas runas sintēzes sistēmas un modeļus, izvirzot arī atlases kvalitātes kritērijus.
2. Izvirzīt testa tekstuālo datu kopu un tās paraugu atlases kritērijus.
3. Uzstādīt atlasītos runas sintēzes modeļus un ģenerēt testa audio kopas, izmantojot izvirzīto tekstuālo testa kopu.
4. Izvirzīt kvalitātes un precizitātes rādītājus atlasīto modeļu salīdzināšanai.
5. Veikt atlasīto modeļu salīdzināšanu un apkopot iegūtos rezultātus.
6. Izdarīt secinājumus un izvirzīt priekšlikumus un rekomendācijas, balstoties uz iegūtajiem rezultātiem.

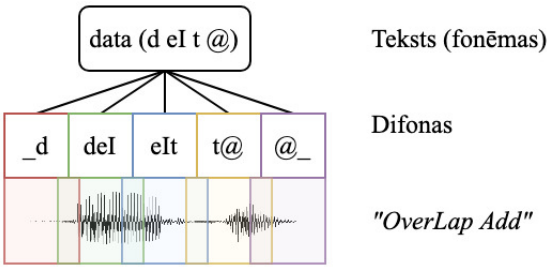
2 Saistītie pētījumi

Mūsdienu tehnoloģiju attīstība un zinātniskās izpratnes paplašināšana ir sekmējusi nozīmīgu progresu runas sintēzes jomā. Šī nodaļa koncentrējas uz divām galvenajām metodēm - difonu apvienošanas bāzētu pieeju un dziļās mašīnmācīšanās tehniku. Abas no tām ir ieguvušas ievērojamu popularitāti un veicinājušas ievērojamas izmaiņas mākslīgā intelekta, datu analīzes un runas tehnoloģiju jomās. Šīs metodes piedāvā jaunas perspektīvas un izaicinājumus runas sintēzes attīstībā, iezīmējot jaunas iespējas un pievienotu vērtību šajā strauji augošajā zinātnes nozarē.

2.1 Difonu apvienošana

Difona balss sintēze ir viena no galvenajām pieejām mākslīgās balss ģenerēšanai, tā lieto difonus kā pamatelementus un izmantojot datorizētas tehnoloģijas. Difons ir runas segments, kur katrs segment sastāv no diviem fonētiskajiem elementiem. Tas sākas fonēmas stabilajā vidusdaļā un beidzas nākamās fonēmas stabilajā vidusdaļā. Izmantojot difonus kā pamatelementus, sintēzē apvienošanas punkti tiek novietoti fonēmu stabilās daļās, kas atvieglo izlīdzināšanas operāciju veikšanu sintēzes laikā un samazina iespējamās nepārtrauktības apvienošanas punktos.

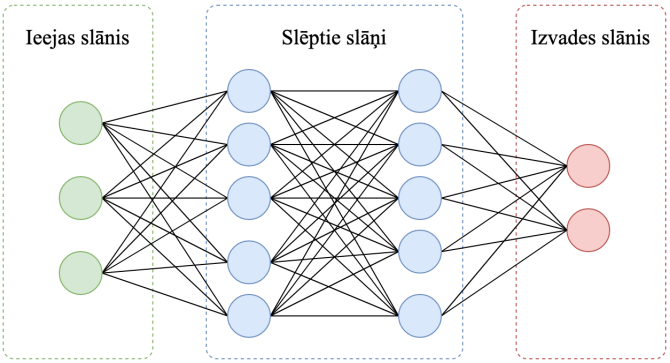
Viena no plašāk izmantotajām sistēmām ir MBROLA (*Multi-Band Resynthesis OverLap Add*), jeb daudzjoslu atkārtotā sintēze izmantojot pārklāšanās apvienojumu. Šī sistēma izmanto noteiktas valodas viena runātāja difonu ierakstu datubāzi, kas pirms tam tiek sagatavota un normalizēta. Runas sintēze notiek atlasot sagatavotos difonu ierakstus un veicot to apvienošanos izmantojot pārklāšanas apvienošanas algoritmu (*OverLap Add*), kā parādīts 1. attēlā. MBROLA sintezētās runas kvalitāte bieži tiek vērtēta kā augsti saprotama, bet diezgan datorizēta, jo difonu datu bāze nesastāv no visiem cilvēka runas kombināciju variantiem [6].



1. att. Vienkāršots difonu apvienošanas modeļa process

2.2 Dziļā mašīnmācīšanās

Dziļā mašīnmācīšanās, kas balstās uz vairāku slāņu neironu tīkliem ir spērusi nozīmīgus soļus datu reprezentāciju apgūšanā. Ar šīs metodes palīdzību ir strauji uzlabojies patreizējais tehnoloģiju stāvoklis dažādās jomās, tai skaitā skaitā runas sintēzē.



2. att. Vienkāršots neironu tīkla modelis

Dziļā mašīnmācīšanās balstās uz neironu tīkliem, kuru pamatā ir savstarpēji savienotas mezglu grupas, kas imitē neironus un ir sakārtotas slāņos: ieejas slānis, viens vai vairāki slēptie slāņi un izvades slānis (skatīt 2. attēlu). Katrs mezgls vai mākslīgais neirons saņem ieejas signālus, apstrādā tos, izmantojot **aktivizācijas funkcijas**, un nodod rezultātu tālāk nākamajam slānim. Dažas no izplatītākajām aktivizācijas funkcijām ir [44]:

- **ReLU** (*Rectified linear unit*) - saglabā tikai pozitīvās vērtības un negatīvās vērtības pārvērš par 0. $f(x) = max(x, 0)$.
- **Sigmoid** - "saspiešanas funkcija", kas visas vērtības pārveido, lai tās būtu robežās no 0 līdz 1. $f(x) = \frac{1}{1+e^{-x}}$.
- **Tanh** - arī šī funkcija līdzīgi kā Sigmoid "saspiež" visas padotās vērtības, bet to dara robežās no -1 līdz 1. $f(x) = \frac{1-e^{-2x}}{1+e^{-2x}}$

Nākamais solis ir tīkla **kļūdas funkcijas** aprēķins, kas nosaka to, cik liela ir atšķirība starp modeļa prognozēto izvadu un reālo vērtību. Atkarībā no uzdevuma tipa, modeļa mērķiem un pieejamajiem datiem, var tikt izvēlēta atbilstošāka kļūdas funkcijas. Divi galvenie kļūdas funkcijas veidi ir:

- **MSE** (*Mean Squared Error*) - dispersija, jeb vidējā kvadrātiskā kļūda, ko izmanto regresijas uzdevumos. $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_i)^2$
- **CE** (*Cross-Entropy*) - krustentropija, ko klasifikācijas uzdevumos, kur modeļa mērķis ir klasificēt ieejas datus konkrētās kategorijās. Šī funkcija tiek izmantota gan binārajā, gan vairāku kategoriju klasifikācija. $CE = -\sum_x p(x) \cdot log\ q(x)$

Atpakaļizplatīšanās algoritms ir viens no svarīgākajiem pamatelementiem neironu tīklā. Tas veic neironu tīkla parametru krituma aprēķināšanu, izmantojot tīklu pretējā virzienā - no izvades slāņa līdz ieejas slānim. Algoritms izmanto atvasinājumu ķēdes likumu, lai aprēķinātu nepieciešamos parciālos atvasinājumus. Ķēdes likums ļauj aprēķināt atvasinājumu vienas mainīgās attiecībā pret otru, izmantojot funkciju kompozīciju, tādējādi vienkāršojot sarežģītus atvasinājumu aprēķinus. Piemēram, ja mums ir funkcijas $Y = f(X)$ un $Z = g(Y)$, tad, izmantojot ķēdes noteikumu, mēs varam aprēķināt Z atvasinājumu attiecībā pret X - $\frac{\partial Z}{\partial X} = \frac{\partial Z}{\partial Y} \cdot \frac{\partial Y}{\partial X}$

Tālāk tiek izmantoti **optimizācijas algoritmi**, jeb optimizatori un iegūtie atvasinājuma funkcijas gradienti, lai atjaunotu katra svara mezglus. Optimizatori virza modeli tā, lai kļūdas funkcija samazinātos, bet precizitāte pieaugtu. Darbs pie optimizatoriem turpinās, un tiek izstrādātas dažādas metodes, kas ņem vērā vairākus aprēķinātos gradientus. Divi no visplašāk izmantotajiem optimizatoriem ir SGD (*Stochastic Gradient Descent*) un Adam algoritmi [18][9].

2.3 Modeļi

Šajā nodaļā tiks aplūkoti dažādi mašīnmācīšanās tīklu veidi, kas ir būtiski dziļās mācīšanās jomā un runas sintezēšanā. Katrs tīkls piedāvā unikālas iespējas datu apstrādē un analīzē, sniedzot vērtīgu ieguldījumu dažādās jomās un uzdevumos. Apskatītie tīklu veidi ir: konvolūciju neironu tīkli (CNN), rekurentie neironu tīkli (RNN), variacionālie autoenkoderi (VAE), ģeneratīvi konkurējošie tīkli (GAN), plūsmas modeļi, difūzijas modeļi un transformeri.

CNN *(Convolution Neural Network)* jeb konvolūciju neironu tīkli ir neironu tīklu klase, kas specializējas datu apstrādē, kam ir režģa līdzīga topoloģija, piemēram, attēli vai audio signāli. Šāda veida tīkli ļauj datoriem "redzēt", apstrādājot attēlus, izmantojot vairākus slāņus, kuri pakāpeniski atpazīst vienkāršākus un sarežģītākus modeļus. CNN strādā, izmantojot slīdošā loga principu, kur vieni un tie paši parametri tiek pielietoti dažādām attēla vietām, tādējādi samazinot nepieciešamību pēc atsevišķiem neironiem katrā no tām. CNN tipiski veido trīs galvenie slāņi:

- Konvolūcijas slānis - veic galveno aprēķinu daļu, izmantojot kodolu vai filtru, kas pārvietojas pāri attēlam, lai izveidotu aktivācijas karti. Šis process ļauj CNN efektīvi uztvert un analizēt lokālās iezīmes, piemēram, līnijas un leņķus.

- Apvienošanas (*pooling*) slānis - samazina attēla izmēru, saglabājot būtiskāko informāciju, un palīdz padarīt modeli noturīgu pret nelielām attēla izmaiņām.

- Pilnībā savienots (*fully connected*) slānis - kalpo kā neironu tīkla "domāšanas" daļa, kas analizē iepriekšējos slāņos iegūto informāciju un veic gala klasifikāciju vai citu

uzdevumu izmantojot datu saspiešanas metodi lineārajā slānī.

Šāda veida arhitektūra ļauj CNN efektīvi tikt galā ar dažādiem attēlu apstrādes uzdevumiem, piemēram, objektu atpazīšanu, semantisko segmentāciju un attēlu aprakstu un ne tikai. [25]

RNN *(Recurrent Neural Network)*, jeb rekurentie neironu tīkli ir mākslīgo neironu tīklu veids, kas ir izstrādāts, lai apstrādātu datu secības, piemēram, laika rindas datus, balss ierakstus un dabisko valodu. Atšķirībā no tiešās padeves (*Feed Forward*) neironu tīkliem, RNN saglabā iepriekšējā slāņa izvadi un izmanto to kā ievadi nākamajam slānim. Šī īpašība ļauj RNN efektīvi uztvert un apstrādāt laikā mainīgus atkarības datus, kas noder tādās jomās, kā: laika rindu prognozēšana, mašīntulkošana, dabiskās valodas apstrāde un sintezēšana.

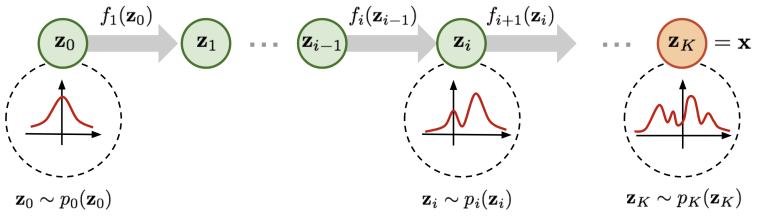
Viena no RNN būtiskākajām priekšrocībām ir spēja apstrādāt mainīga garuma secības, saglabājot informāciju par iepriekšējām ievadēm. Tas ir izšķiroši uzdevumos, kuros ir svarīga konteksta uztvere. Tomēr RNN struktūra rada tādas problēmas kā pazūdošie un sprāgstošie gradienti, kas apgrūtina tīkla efektīvu apmācību un mācīšanos no datiem ar ilgtermiņa atkarībām. Lai risinātu šīs problēmas, tika izstrādātas RNN variācijas, piemēram, ilgtermiņa atmiņas (LSTM) tīkls un vārtu rekurento vienību (GRU) tīkls, kas ļauj RNN efektīvāk apstrādāt ilgtermiņa atkarības un palielina to spēju mācīties no sarežģītākām datu kopām [12].

VAE *(Variational Autoencoder)* - variacionālie autoenkoderi ir neironu tīklu metode, kas tiek izmantota dziļās mācīšanās jomā. Tie izmanto ģeneratīvu pieeju, lai radītu jaunus datus. VAE darbības pamatā ir divi galvenie komponenti - kodētājs un dekodētājs. Kodētājs saspiež datus slēptajā vektorā, savukārt dekodētājs to atkal izpleš uz orģinālo datu formātu. Lai apmācītu šo modeli, tiek izmantota KL (*Kullback-Leibler*) novirze, kas optimizē kodētāja izvadīto slēptā vektora sadalījumu, lai tas tuvinātos normālajam sadalījumam. Tas nodrošina, ka izmantojot modeli, var vieglāk ģenerēt paredzamus un saturīgus jaunus datus. Kā alternatīvu nepārtrauktajai slēptajai telpai, var pielietot **VQ** (*Vector Quantization*), jeb vektoru kvantēšanu, kas nozīmē diskrētu telpas reprezentāciju izmantojot kodu grāmatu (*codebook*), kas sastāv no vektoru saraksta. Katram kodētāja izvadītajam elementam tiek piemeklēts tuvākais vektors no kodu grāmatas, un tas tālāk tiek nodots dekodētājam [31].

GAN (*Generative Adversarial Networks*) - ģeneratīvi konkurējošo tīklu ietekme uz teksta pārvēršanu runā ir nozīmīga. Izmantojot GANs, ir iespējams uzlabot sistēmu spēju radīt runu, kas daudz precīzāk atspoguļo cilvēka balss intonācijas, akcentus un pat emocijas. Šo modeļu pamatā ir divu konkurējošu tīklu – ģeneratora (G) un diskriminatora (D) – sacenšanās. Ģenerators mēģina radīt datus, kas līdzinās patiesiem datiem, bet diskriminators mēģina atšķirt ģenerētos datus no īstajiem. Šajā procesā tiek izmantota *minmax* spēles teorija (*game theory*), kur G mēģina maksimizēt D kļūdas varbūtību, bet

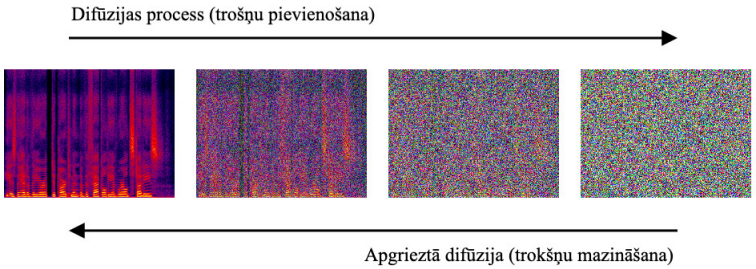
D mēģina to minimizēt. GAN arhitektūra sniedz iespēju trenēt abus tīklus vienlaikus, izmantojot atpakaļizplatīšanas algoritmu. Lai arī šī arhitektūra ir pierādījusi sevi gan reālistisku bilžu, gan runas audio sintēze, šo modeļu trenēšana nav viegla, tā bieži ir lēna un nestabila [10].

Plūsmas *(flow)* modeļi ir īpaši noderīgi dažādu datu, piemēram, attēlu vai runas sintēzē un ģenerēšanā. Šo modeļu pamatā ir spēja apgūt sarežģītas datu sadalījumu struktūras, pārveidojot vienkāršus, labi zināmus sadalījumus (piemēram, normālo sadalījumu) par daudz sarežģītākiem sadalījumiem (skatīt 3. attēlu). Tas nodrošina labāku mērķa datu varbūtisko sadalījumu atspoguļojumu. Plūsmas modeļi izmanto apgriežamas transformācijas funkcijas, kuru parametri tiek apmācīti, lai optimizētu logaritmisko ticamību (*log-likelihood*), tādējādi precīzi modelējot datu sadalījumu.



3. att. Plūsmas modeļu process vienkārša sadalījuma pārveidošanai sarežģītākā [39]

Difūzijas *(diffusion)* modeļi ir viena no jaunākajām metodēm ģeneratīvo modeļu jomā, kas pamatojas uz Markova ķēdes principiem. Šie modeļi galvenokārt izmanto divus būtiskus procesus: uz priekšu vērsto difūziju un apgriezto difūziju (skatīt 4. attēlu). Uz priekšu vērsta difūzija pakāpeniski pievieno mērenu daudzumu Gausa trokšņa datu paraugam, tādējādi padarot paraugu trokšņaināku. Turpretī apgrieztā difūzija ir process, kura laikā tiek mēģināts atjaunot sākotnējo paraugu no trokšņainā parauga, izmantojot apmācītu modeli. Šāda pieeja ļauj radīt detalizētus un precīzus datu paraugus, kas ir īpaši svarīgi, piemēram, attēlu vai skaņas ģenerēšanā. Tomēr, neskatoties uz to spēju radīt augstas kvalitātes paraugus, difūzijas modeļiem ir savi trūkumi. Viens no galvenajiem ir paraugu ģenerēšanas laikietilpīgums [40].



4. att. Difūzijas modeļa process izmantojot mel spektrogrammu

Transformeri *(Transformers)* - viena no galvenajām dziļās mašīnmācīšanās arhitektūrām, kas ieviesta 2017. gadā. Šī arhitektūra izmanto pašuzmanības (*self-attention*) mehānismu, ļaujot tai analizēt secīgus datus kopumā, nevis pēc kārtas. Tas palīdz modelim labāk saprast kontekstuālās nozīmes.

Transformeri sastāv no kodētāja un dekodētāja. Kodētājs apstrādā ieejas datus, kas iepriekš pārvērsti priekšapmācītos kartējumos. Pozicionālā kodēšana pievieno informāciju par vārdu secību teikumā. Kodētājs rada abstraktu, nepārtrauktu ieejas secības reprezentāciju, kas iever informāciju par vārdu savstarpējām attiecībām. Dekodētājs, izmantojot līdzīgus slāņus kā kodētājs, bet ar papildus soļiem, ģenerē tekstu, nosakot nākamā vārda varbūtību secībā. [22]

Arhitektūras pamatā ir vairāki identiski kodētāju un dekodētāju slāņi, kas ir savstarpēji saistīti ar uzmanības mehānismiem un tiešās padeves blokiem. Tās galvenā priekšrocība ir spēja vienlaikus apstrādāt vairākus vārdus, kas nodrošina ievējojamu efektivitātes uzlabojumu salīdzinājumā ar secīgajiem modeļiem, piemēram, RNN un LSTM. [38]

2.4 Rādītāji

WER un CER rādītāji tiek plaši izmantoti dažādās sitēmās, sākot no teksta atpazīšanas sistēmām līdz ASR sistēmām. Šie rādītāji palīdz objektīvi novērtēt, cik atbilstoši un precīzi sistēma ir ģenerējusi un atpazinusi tekstu salīdzinājumā ar orģinālo versiju. Ar šo rādītāju palīdzību tiek mērīta vārdu izlaišanas, iespraušanas un aizvietošanas skaita vērtība. Abi no rādītājiem tiek aprēķināti pēc sekojošās formulas:

$$CER\ vai\ WER = \frac{S + D + I}{N}$$

kur S - substitūciju skaits, D - dzēšanu skaits, I - iespraušanu skaits un N - kopējais simbolu vai vārdu skaits.

Lai veiktu šo rādītāju aprēķināšanu, runas sintēzes sistēmās nepieciešams sintezētos audio failus pārvērst atpakaļ teksta formā, izmantojot kādu ASR sistēmu. Kad šis process ir paveikts, tiek veikta rādītāju aprēķināšana, salīdzinot ievadtekstu ar ASR sistēmas atpazīto tekstu. [5] [19]

MOS ir plaši izmantots rādītājs telekomunikāciju un runas sintēzes jomās, lai novērtētu audio kvalitāti. Šis rādītājs balstās uz cilvēku subjektīviem vērtējumiem. Respondenti novērtē audio kvalitāti Likerta skalā no 1 līdz 5, kur 1 nozīmē zemu kvalitāti, bet 5 - augstu kvalitāti.

MOS rādītājam ir arī dažādi paveidi, kas tāpat kā MOS rādītājs, balstās uz subjektīvu vērtējumu pamata. Viens no populārākajiem paveidiem ir CMOS (Comparative MOS). Tas ir salīdzinošais vidējais viedokļa rādītājs, kas tiek noteikts, liekot responden-

tiem vērtēt kvalitātes atšķirību starp diviem paraugiem. Arī šī rādītāja novērtēšanai bieži tiek izmantota Likerta skala no -3 līdz +3, kur negatīvie skaitļi nozīmē, ka pirmais paraugs ir sliktāks, savukārt pozitīvie, ka labāks.

Dažādās situācijas ir iespējams šos rādītājus pielāgot, liekot respondentiem vērtēt nevis tikai kvalitāti, bet kādas citas pazīmes, piemēram, dabīgumu, trokšņus vai skaļumu. [35] [11]

3 Sistemātiskā literatūras analīze

Šajā nodaļā tiek aprakstīta sistemātiskā literatūras analīzes metodoloģija, kas nepieciešama zinātnisko darbu meklēšanai un izvēlēto modeļu un sistēmu tālākai izpētei un atlasei.

3.1 Meklēšanas protokols

Modeļu, sistēmu un to zinātnisko darbu meklēšanai tika izmantoti sekojošie rīki un paņēmieni:

1. Meklēšana tika veikta, izmantojot *Papers With Code, Semantic Scholar, IEEE, ArXiv* un *Google Scholar* datubāzes un rīkus, kā arī lietojot šādas frāzes un atslēgvārdus:

   - Text to speech systems;
   - TTS model comparison;
   - Speech synthesis models;
   - SOTA TTS models;
   - Diphone based TTS systems.

   Tika apskatīti gan zinātniskie raksti, gan dažāda veida apkopojumi un pirmkoda respozitoriji.

2. Pēc pirmo darbu atrašanas tika veikta darbu atsauču pārbaude, kuras rezultātā tika identificēti citi ar runas sintēzi saistītie modeļi.

3. Tālāk tika meklētas atsauces, kas norādītas jau atrastajos dabos, izmantojot *Google Scholar* meklēšanas rīku.

4. Visiem modeļiem, kas datēti ar 2021. gadu vai agrākiem datumiem, nepieciešami vismaz 100 citāti.

Kopumā tika atlasītas 19 runas sintēzes sistēmas, kurām tika veikta padziļināta analīze kā arī izvērtēti izvirzītie kvalitātes kritēriji. Šīs sistēmas redzamas 1. tabulā. Šajā tabulā arī tiek nodefinēti modeļu un sistēmu saīsinātie nosaukumi, kas tiks izmantoti turpmākajā darbā.

1. tabula. Modeļu un sistēmu pārskats

Nr	Nosaukums	Pilnais nosaukums	Citāti	Iesniegts	Autori
1	**CoMoSpeech** [42]	CoMoSpeech: One-Step Speech and Singing Voice Synthesis via Consistency Model	4	2023.10	Hong Kong University of Science and Technology; Microsoft Research Asia; Hong Kong Baptist University; Hong Kong Institute of Science Innovation; Chinese Academy of Sciences
2	**VITS 2** [17]	VITS2: Improving Quality and Efficiency of Single-Stage Text-to-Speech with Adversarial Learning and Architecture Design	2	2023.07	SK Telecom, South Korea
3	**NaturalSpeech 2** [35]	NaturalSpeech 2: Latent Diffusion Models are Natural and Zero-Shot Speech and Singing Synthesizers	36	2023.04	Microsoft Research Asia; Microsoft Azure Speech
4	**FoundationTTS** [41]	FoundationTTS: Text-to-Speech for ASR Customization with Generative Language Model	4	2023.03	Microsoft
5	**MQTTS** [4]	A Vector Quantized Approach for Text to Speech Synthesis on Real-World Spontaneous Speech	12	2023.02	Language Technologies Institute, Carnegie Mellon University
6	**OverFlow** [24]	OverFlow: Putting flows on top of neural transducers for better TTS	7	2022.11	Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden
7	**DelightfulTTS 2** [21]	DelightfulTTS 2: End-to-End Speech Synthesis with Adversarial Vector-Quantized Auto-Encoders	18	2022.06	Microsoft Azure Speech; Microsoft Research Asia
8	**Guided-TTS 2** [15]	Guided-TTS 2: A Diffusion Model for High-quality Adaptive Text-to-Speech with Untranscribed Data	28	2022.05	Seoul National University
9	**NaturalSpeech** [36]	NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality	85	2022.05	Microsoft Research Asia; Microsoft Azure Speech
10	**VQTTS** [7]	VQTTS: High-Fidelity Text-to-Speech Synthesis with Self-Supervised VQ Acoustic Feature	41	2022.04	MoE Key Lab of Artificial Intelligence, AI Institute; X-LANCE Lab, Department of Computer Science and Engineering; Shanghai Jiao Tong University, Shanghai, China
11	**JETS** [20]	JETS: Jointly Training FastSpeech2 and HiFi-GAN for End to End Text to Speech	31	2022.03	Kakao Enterprise Corporation, Seongnam, Republic of Korea
12	**YourTTS** [3]	YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone	152	2021.12	Universidade de Sao Paulo, Brazil; Sopra Banking Software, France; defined.ai, United States of America; Coqui, Germany; Federal University of Technology - Parana, Brazil;
13	**SpeechT5** [1]	SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing	121	2021.10	Department of Computer Science and Engineering, Southern University of Science and Technology; Department of Computing, The Hong Kong Polytechnic University; Department of Computer Science and Technology, Tongji University; Microsoft; Peng Cheng Laboratory;
14	**VITS** [13]	Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech	447	2021.06	Kakao Enterprise, Republic of Korea; School of Computing, KAIST, Daejeon, Republic of Korea;
15	**Grad-TTS** [28]	Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech	273	2021.05	Huawei Noah's Ark Lab, Moscow, Russia; Higher School of Economics, Moscow, Russia;
16	**FastSpeech 2** [30]	FastSpeech 2: Fast and High-Quality End-to-End Text to Speech	1042	2020.06	Zhejiang University; Microsoft Research Asia; Microsoft Azure Speech;
17	**Glow-TTS** [14]	Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search	360	2020.05	Kakao Enterprise; Data Science & AI Lab, Seoul National University;
18	**Tacotron 2** [34]	Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions	2810	2017.12	Google, Inc.; University of California, Berkeley
19	**MaryTTS** [33]	The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching	610	2003.10	German Research Center for Artificial Intelligence; Saarland University, Institute of Phonetics;

Katram no atlasītajiem darbiem tika apkopotas būtiskākās metodes; arhitektūras iezīmes, uz ko šīs sistēmas balstās, lai gūtu vispārīgu priekšstatu par aplūkojamo modeļu darbību, uzbūvi un saistību ar citiem modeļiem; kā arī izvērtējums, vai sistēma atbalsta dažādu balsu sintezēšanas iespējas. Šis apkopojums redzams 2. tabulā.

2. tabula. Sistēmās izmantoto metožu apraksts

Nr	Nosaukums	Apraksts	Dažādu balsu atbalsts
1	**CoMoSpeech**	Destilēts transformera + difūzijas (Unet [32] bāzēts) modelis no Grad-TTS + balss kodētājs (HiFiGAN [16])	
2	**VITS 2**	Transformer + Flow + HiFiGAN [16] tipa apvienotais modelis	+
3	**NaturalSpeech 2**	Transformer + difūzijas + RVQ [43] apvienotais modelis	+
4	**FoundationTTS**	Transformer valodas modelis + smalka un rupja veida RVQ [43] apvienotais modelis	+
5	**MQTTS**	VQ, kas balstās Transformer tipa arhitektūrā + HiFiGAN [16] apvienotais modelis	
6	**OverFlow**	Transformer + Flow (NHMM [37]) + LSTM modelis + balss kodētājs (HiFiGAN [16])	+
7	**DelightfulTTS 2**	Transformer + uz VQ balstīta balss kodētāja (VQ-GAN, [8]) apvienotais modelis	
8	**Guided-TTS 2**	Transformer + difūzijas (Unet [32] bāzēts) modelis + balss kodētājs (HiFiGAN [16])	+
9	**NaturalSpeech**	Transformer + Flow + RCNN apvienotais modelis	
10	**VQTTS**	Conformer + RNN (LSTM) + HiFiGAN [16] VQ apvienotais modelis	
11	**JETS**	Transformer (FastSpeech 2) + HiFiGAN [16] apvienotais modelis	
12	**YourTTS**	Transformer + Flow + HiFiGAN [16] tipa apvienotais modelis	+
13	**SpeechT5**	Multi modāls VQ modelis, kas balstās Transformer tipa arhitektūra un izmanto HiFiGAN [16], wav2vec [2] modeļus	+
14	**VITS**	Transformer + Flow + HiFiGAN [16] tipa apvienotais modelis	+
15	**Grad-TTS**	difūzijas (Unet [32] bāzēts) modelis + balss kodētājs (HiFiGAN [16])	
16	**FastSpeech 2**	Transformer tipa modelis + balss kodētājs (HiFiGAN [16])	
17	**Glow-TTS**	Transformer + Flow + WaveGlow [29] tipa apvienotais modelis	+
18	**Tacotron 2**	LSTM modelis + balss kodētājs (WaveNet [27])	
19	**MaryTTS**	Difonu apvienošanas sistēma	+

Tālāk tika veikts arī katra atlasītā darba rezultātu un sistēmu rādītāju informācijas apkopojums, lai nākamajos soļos varētu izvērtēt to, cik labi salīdzināmi šie rezultāti ir savā starpā. 3. tabulā redzams rezultātu apkopojums, kur iekļauti tieši tie modeļi un to

rādītāji, kuriem varēja veikt salīdzinājumu ar vismaz vēl vienu citu modeli.

3. tabula. Dokumentos norādītās metrikas

Nr	Nosaukums	LJSpeech (MOS)	LJSpeech (CER)	VCTK (MOS)	LibriTTS (MOS)
1	**CoMoSpeech**	4.24			
2	**VITS 2**	4.47	3.92	3.99	
6	**OverFlow**	3.43			
8	**Guided-TTS 2**	4.23	**0.89**	4.23	
9	**NaturalSpeech**	4.56			
10	**VQTTS**	**4.71**			
11	**JETS**	4.02			
12	**YourTTS**			4.24	**4.25**
13	**SpeechT5**				3.65
14	**VITS**	4.43		**4.38**	
15	**Grad-TTS**	4.44			
16	**FastSpeech 2**	3.83			
17	**Glow-TTS**	4.01			3.45

3.2 Kvalitātes kritēriji

Atlasīto darbu filtrēšanai, pēc meklēšanas protokolā paveiktā darba, tika izmantoti sekojošie kvalitātes kritēriji, kas palīdzēja sašaurināt tālāko modeļu un sistēmu salīdzināšanas apjomu:

1. Modeļa trenēšanai izmantota publiski pieejama datu kopa;
2. Modelim vai sistēmai pieejami sintezētās runas piemēri;
3. Modelim ir brīvi pieejams tā pirmkods;
4. Modelim ir brīvi pieejami tā trenēšanas rezultātu svari;
5. Nepieciešama vismaz viena runas sintēzes sistēma, kas nebalstās uz dziļo mašīnmācīšanos;
6. Sistēma izveduma režīmā sintezē līdzvērtīgas kvalitātes audio fragmentus, kā publicētajos piemēros.

Izvirzītie kvalitātes kritēriji tika izskatīti katram no atlasītajiem modeļiem un apkopoti 4. tabulā. Katrai sistēmai tika dots kopējais vērtējums, pēc kura vadoties, sistēmas tika virzītas tālākai izpētei un salīdzināšanai.

4. tabula. Modeļu un sistēmu kvalitātes rezultāti

Nr	Nosaukums	K1	K2	K3	K4	K5	K6	Vērtējums
1	**CoMoSpeech**	+	+	+	+		+	**5**
2	**VITS 2**	+	+	+				3
3	**NaturalSpeech 2**	+	+	+				3
4	**FoundationTTS**	+						1
5	**MQTTS**	+	+	+	+		+	**5**
6	**OverFlow**	+	+	+	+		+	**5**
7	**DelightfulTTS 2**	+	+					2
8	**Guided-TTS 2**	+	+	+				3
9	**NaturalSpeech**	+	+					2
10	**VQTTS**	+	+					2
11	**JETS**	+	+	+	+			4
12	**YourTTS**	+	+	+	+		+	**5**
13	**SpeechT5**	+		+	+		+	4
14	**VITS**	+	+	+	+		+	**5**
15	**Grad-TTS**	+	+	+	+		+	**5**
16	**FastSpeech 2**	+	+	+	+		+	**5**
17	**Glow-TTS**	+	+	+	+		+	**5**
18	**Tacotron 2**		+	+	+		+	4
19	***MaryTTS***					+	+	**2**

4 Metodoloģija

Darba galvenais mērķis ir veikt objektīvu runas sintēzes sistēmu salīdzināšanu. Lai to panāktu, nepieciešams veikt sekojošus soļus:

1. Sistēmu publikāciju atlasi un izpēti, izmantojot atlases un kvalitātes kritērijus.
2. Modeļu pirmkoda un svaru uzstādīšanu runas paraugu sintezēšanai, papildus pārliecinoties, ka tie ir līdzīgā kvalitātē, kā autoru piedāvātajos paraugos.
3. Runas paraugu sintezēšanu, kā arī to apstrādi, izmantojot ASR sistēmas API.
4. Rādītāju kvalitātes aprēķinu, izmantojot NISQA modeli; un precizitātes aprēķinu, izmantojot CER un WER rādītājus, balstoties uz ASR sistēmas iegūtajiem datiem.

4.1 Datu kopas

Lai iegūtu validācijas datu kopas katrai runas sintēzes sistēmai un modelim, vispirms nepieciešams iegūt vienotu teksuālo datu kopu, ar kuras palīdzību tiks ģenērēti audio faili. Šādas datu kopas izvēlei tika izvirzīti vairāki kritēriji, lai nodrošinātu pietiekamu tekstu dažādību, kā arī pietiekamu apjomu objektīvu rezultātu iegūšanai.

Galvenie izvirzītie kritēriji šādai datu kopai ir:

1. Datu kopa nav izmantota neviena modeļa trenēšanā;
2. Datu kopas piemēru sintēze audio failos rezultējas runā, kas nav garāka par 10 sekundēm;
3. Datu kopā ir vismaz 4 000 piemēri un tie ir ar lielu dažādību, lai varētu veikt visaptverošu sistēmu un modeļu kvalitātes analīzi;
4. Datu kopas licence atļauj to izmantot zinātniskās pētniecības darbos.

Vispirms tika apkopotas izmantotās datu kopas no izvēlētajām runas sintēzes sistēmām un modeļiem. Šis apkopojums redzams 5. tabulā. No modeļu aprakstiem tika apkopoti arī populārāko datu kopu parametri:

- **LJSpeech** - viena runātāja, angļu valodas datu kopa, kas satur aptuveni 13 100 audio paraugus.
- **VCTK** (*Voice Cloning Toolkit*) - dažādu akcentu angļu valodas datu kopa, kurā ir 109 dažādi runātāji, un kas satur aptuveni 44 000 audio paraugus.

- **LibriSpeech** - angļu valodas datu kopa, kas satur 2 484 dažādus runātājus un aptuveni 1 000 stundas ar ierunātu tekstu.
- **LibriTTS** - angļu valodas datu kopa, kas satur 2 456 dažādus runātājus un ir aptuveni 1 000 stundas ar ierunātu tekstu.

5. tabula. Modeļu datu kopas

Nr	Nosaukums	Datu kopas
1	**CoMoSpeech**	LJSpeech
2	**VITS 2**	LJSpeech; VCTK
3	**EfficientSpeech**	LJSpeech
4	**NaturalSpeech 2**	VCTK; LibriSpeech
5	**FoundationTTS**	VCTK; LibriTTS; Proprietary
6	**MQTTS**	VoxCeleb; GigaSpeech
7	**OverFlow**	LJSpeech
8	**Estonian TTS**	Proprietary
9	**DelightfulTTS 2**	Proprietary
10	**Guided-TTS 2**	LJSpeech; VCTK; LibriSpeech; LibriTTS; VoxCeleb; LibriLight
11	**NaturalSpeech**	LJSpeech; Proprietary
12	**VQTTS**	LJSpeech
13	**JETS**	LJSpeech
14	**YourTTS**	VCTK; LibriTTS; TTS-Portuguese; M-AILABS (Franču); MLS (Portugāļu)
15	**EdiTTS**	LJSpeech
16	**SpeechT5**	LibriSpeech; LibriTTS
17	**VITS**	LJSpeech; VCTK
18	**Grad-TTS**	LJSpeech
19	**Parallel Tacotron 2**	Proprietary
20	**Apple-TTS**	Proprietary
21	**FastSpeech 2**	LJSpeech
22	**Glow-TTS**	LJSpeech; LibriTTS
23	**Tacotron 2**	Proprietary
24	**MaryTTS**	Proprietary

Tālāk tika veikta citu datu kopu meklēšana, kā rezultātā tika izvēlēta "*Mozzila Common Voice*" datu kopa. Šī datu kopa aptver dažādu valodu teksta - audio pārus. Šī darba ietvaros tika izvēlēta angļu valodas datu kopa, kas sastāv no 1 752 390 ierakstiem.

Lai iegūtu pēc iespējas kvalitatīvākus datus un sašaurinātu sintezējamo un pārbaudāmo failu apjomu, tie tika atfiltrēti, izmantojot sekojošos kritērijus:

1. Vismaz 7 vērtētāji audio ierakstu ir novērtējuši pozitīvi;
2. Ne vairāk kā 10% no vērtētājiem audio ierakstu ir novērtējuši negatīvi;
3. Tika izņemti ieraksti ar marķējumu "*Benchmark*";
4. Duplikātu gadījumā tika izvēlēts ieraksts ar visaugstāko pozitīvo vērtējumu skaitu.

Rezultātā tika iegūti 4 677 ieraksti, kas tālāk tika izmantoti modeļu testa kopas sintēzei.

4.2 Testa kopas sintēze

Balstoties uz izvirzītajiem sistēmu kvalitātes kritērijiem, lokāli tika uzstādītas visas 9 atlasītās sistēmas. Uzstādīšana notika balstoties uz publiski pieejamajiem pirmkoda avotiem, bet modeļu gadījumā - publiski pieejamajiem svariem. Gadījumā, ja modelis atbalstīja dažādu balsu sintēzi, tika izvēlēts pirmais autoru rekomendētais balss sintēzes vektors. Katra no uzstādītajām sistēmām tika darbināta izvešanas (*inference*) režīmā, tika padoti visi testa kopas tekstu ieraksti, kā rezultātā tika saglabāti sintezētie audio failu paraugi. Sintezētie rezultāti tālāk tika izmantoti katra modeļa rādītāju aprēķināšanai.

4.3 Rādītāji

Izvēlēto runas sintēzes modeļu un sistēmu testa kopas salīdzināšanai tika izvēlēti vairāki rādītāji, kas tālāk nodrošināja objektīvu sistēmu salīdzīnājumu.

4.3.1. WER un CER

Pirmie no izvēlētajiem rādītājiem ir - WER un CER. Šīe rādītāji palīdz objektīvi novērtēt to, cik precīzi, atbilstoši ievadītajam tekstam, runas sintēzes sistēma ir sintezējusi tekstu.

Lai veiktu šo rādītāju aprēķināšanu, vispirms nepieciešams iegūtos testa audio kopas datus pārvērst tekstuālā formā, kam tika izmantota Asya.ai readītā ASR sistēma. Visi testa audio kopas ieraksti tika nosūtīti uz šo sistēmu un saglabāti tekstuālie rezultāti. Tālāk tika veikta katra ieraksta rādītāju aprēķināšana un vidējā rezultāta iegūšana konkrētajai runas sintēzes sistēmai.

Šī pati WER un CER rādītāju aprēķināšanas procedūra tika veikta arī izvēlētās datu kopas audio failiem. Iegūtie rezultāti tika pievienoti pārējiem sistēmu un modeļu rezultātiem, un tika izveidota 7. tabula, kurā redzami katras sistēmas un datu kopas iegūtie rezultāti.

4.3.2. NISQA

Pārējie rādītāji tika iegūtas izmantojot NISQA modeli, kas ir automatizēta sistēma, kas paredzēta runas kvalitātes novērtēšanai. Tā izmanto mašīnmācīšanās tehnoloģijas, lai analizētu un vērtētu runas signālu, nodrošinot objektīvu kvalitātes mērījumu. NISQA piedāvā vairākus atšķirīgus novērtējumus:

- Kopējā kvalitāte (MOS/quality) - vispārējs vērtējums par runas uztveramo kvalitāti. Imitē manuālo MOS vērtējumu, kur vairāki klausītāji manuāli vērtē runas kvalitāti.
- Dabiskums (naturalness) - novērtē, cik dabiska un cilvēka balsij līdzīga ir runas sintēzes sistēmas radītā runa. Tas ir svarīgs rādītājs, lai noteiktu, cik efektīvi sistēma var imitēt cilvēka runu.
- Krāsojums (coloration) - mēra nevēlamo skaņu vai frekvenču klātbūtni runā, kas kropļo skaņu.
- Trokšņi (noisiness) - nosaka fona trokšņu līmeņa pakāpi. Augsts trokšņa līmenis var ievērojami samazināt runas saprotamību.
- Pārtraukumi (discontinuity) - meklē un novērtē jebkādus nepārtrauktības traucējumus runā, piemēram, pārtraukumus vai skaņas defektus.
- Skaļums (loudness) - mēra to cik optimāls ir runas skaļuma līmenis, lai runa būtu komfortabli un skaidri saklausāma, pārāk kluss vai pārāk skaļš līmenis nozīmē zemāku vērtējumu.

Visiem mērījumiem augstāks vērtējums nozīmē labāku runas kvalitāti. [26]

Šo rādītāju aprēķināšanai tika izmantota runas sintezētās testa datu kopa, ko katra modeļa sintezētajam ierakstam aprēķināja izmantojot NISQA modeli. Tālāk tika veikta vidējā rezultāta aprēķināšana katra modeļa katram rādītājam. 6. tabulā redzami visu modeļu rezultāti.

5 Rezultāti

Kopumā tika apskatītas 19 sistēmas, no kurām tika atlasītas 9, kas atbilda izvirzītajiem kvalitātes kritērijiem. Tika veikta katras sistēmas uzstādīšana uz lokālās vides un audio testa kopas sintezēšana.

Izveidotajām testa kopām tika pielietota NISQA rādītāju aprēķināšanas metode, bet pēc tam sekoja datu kopas vidējo rezultātu aprēķināšana. Šī metrika norāda uz kopējo runas sintēzes kvalitāti vai dabīgumu. 6. tabulā redzams, ka vislabākos vērtējumus ieguvis CoMoSpeech modelis, kas ir par vidēji 22% (skatīt 5. attēlu) labāks nekā izvēlētā datu kopa. Augstu vērtējumu ieguva arī MQTTS un Grad-TTS modeļi. Interesanti ir tas, ka pašas izvēlētās datu kopas (*Common Voice*) runas paraugi uzstādīja vienus no zemākajiem rezultātiem, neskatoties uz to, ka tika izvēlēti paraugi ar visaugstākajiem vērtējumiem. Ļoti iespējams, ka rezultāti ir šādi tieši tādēļ, ka šo datu kopu ierunājuši cilvēki mājas apstākļos, ar dažādas kvalitātes mikrofoniem un fona trokšņiem. Vissliktāko vērtējumu, par vidēji 7%, ieguva MaryTTS difonu apvienošanas bāzētā sistēma, kas izklausījās pārlieku robotizēta. Spilgtākie piemēri, balstoties uz šo rādītāju, apkopoti publiski pieejamā vietnē. [1]

6. tabula. Sistēmu un datu kopas NISQA metrikas

Nr	Nosaukums	Kvalitāte	Dabiskums	Krāsojums	Trokšņi	Pārtraukumi	Skaļums
1	**CoMoSpeech**	**3.85**	4.41	**4.41**	**4.67**	**4.58**	3.97
5	**MQTTS**	3.54	**4.53**	4.24	4.61	4.33	**4.12**
6	**OverFlow**	3.08	3.93	4.17	4.50	4.02	3.55
12	**YourTTS**	3.24	4.02	3.73	4.30	4.09	4.04
14	**VITS**	3.07	4.29	4.10	4.43	4.27	3.66
15	**Grad-TTS**	3.64	4.36	4.38	4.65	4.57	3.89
16	**FastSpeech 2**	2.87	3.44	3.72	4.04	3.62	3.40
17	**Glow-TTS**	3.04	3.85	4.16	4.52	4.03	3.78
19	***MaryTTS***	*2.38*	*3.35*	*3.43*	*3.17*	*3.73*	*3.72*
	Common Voice	*3.25*	*3.38*	*3.36*	*3.87*	*3.76*	*3.52*

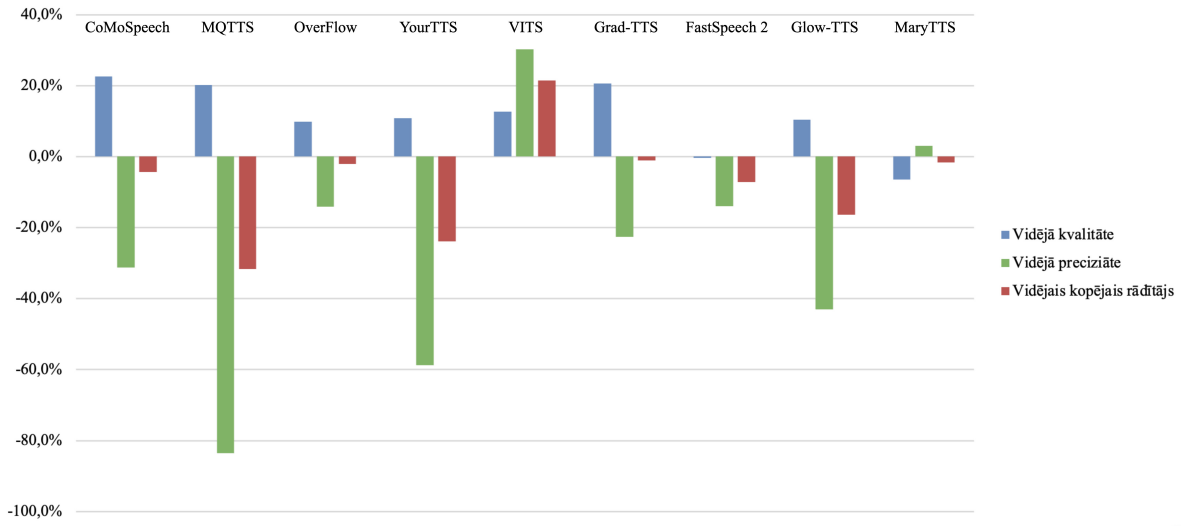
Visas testa kopas tika apstrādātas, izmantojot ASR sistēmu un tika iegūti katra ieraksta noteiktie teksti. Šī metrika norāda uz runas sintēzes precizitāti. Teksti tika salīdzināti ar orģināliem izmantojot CER un WER rādītājus un rezultāti apkopoti 7. tabulā. Šajā tabulā redzams, ka visprecīzāk tekstu sintezēja VITS modelis, vidēji par 30% labāk (skatīt 5. attēlu) nekā šajā pašā rādītājā ieguva izvēlētās datu kopas (*Common Voice*) runas paraugi. Pārējās sistēmas, izņemot MaryTTS, kas bija līdzvērtīga, ieguva sliktākus šajā rādītājā. Arī šajiem rādītājiem spilgtākie piemēri ir apkopoti publiski pieejamā vietnē. [2]

[1] `https://research.saulitis.dev/english-speech-synthesis-comparison-2024#NISQA`
[2] `https://research.saulitis.dev/english-speech-synthesis-comparison-2024#CER-WER`

7. tabula. Sistēmu un datu kopas CER, WER metrikas

Nr	Nosaukums	WER	CER
1	**CoMoSpeech**	0.07	0.03
5	**MQTTS**	0.21	0.18
6	**OverFlow**	0.05	0.02
12	**YourTTS**	0.11	0.05
14	**VITS**	**0.04**	**0.01**
15	**Grad-TTS**	0.06	0.03
16	**FastSpeech 2**	0.06	0.02
17	**Glow-TTS**	0.08	0.04
19	***MaryTTS***	*0.05*	*0.02*
	Common Voice	*0.05*	*0.02*

Kopumā, aprēķinot vidējos rezultātus abu metriku grupām, izteikti vadībā bija VITS runas sintēzes modelis, kas ir ļoti precīzs, un arī runas kvalitātes ziņā ieguvis augstu vērtējumu. Par nākamo labāko modeli var atzīt *Grad-TTS*, kas sasniedza nedaudz zemāku vidējo vērtību kā *Common Voice* datu kopa.



5. att. Sistēmu kvalitātes, precizitātes un kopējie vidējie rādītāji

Viss datu apstrādes un runas sintēzes algoritmu pirmkods pieejams projekta repozitorijā [1]. Tur atrodami arī visu modeļu iegūto datu un rādītāju neapstrādātās versijas.

[1] `https://github.com/krsaulitis/course-project`

6 Secinājumi

Šī kursa darba mērķis bija veikt runas sintēzes sistēmu literatūras atlasi un analīzi, un sniegt praktisku un objektīvu salīdzinājumu starp angļu valodas runas sintēzes sistēmām. Lai to izdarītu, tika izpildīti izvirzītie uzdevumi, no kuriem izriet sekojoši secinājumi:

1. Runas sintēzes sistēmu attīstība turpinās, un ir pieejamas daudz dažādas runas sintēzes sistēmas. Diemžēl vairākas no jaunākajām sistēmām nav izmantojamas un prasītu daudz laika, lai atkārtotu publicētos rezultātus, jo to pirmkods un svari nav publiski pieejami.

2. Modeļi izmanto dažādas datu kopas, kas apgrūtina modeļu savstarpēju salīdzināšanu, bet izmantotajām datu kopām visizplatītākā ir LJSpeech.

3. Runas sintēzes sistēmu uzstādīšanas sarežģītība ir ļoti dažāda un vairākiem modeļiem to apgrūtina novecojušu pakotņu kļūdas un nepietiekoši sniegtā dokumentācija.

4. Audio failu kvalitātes objektīvai noteikšai ir maz sistēmu, un autori pārsvarā paļaujas uz subjektīviem rādītājiem, kas savā starpā nav salīdzināmi.

5. Vislabākos kvalitātes rādītājus uzrādīja CoMoSpeech modelis (MOS - 3.85), savukārt VITS modelis uzrādīja visaugstāko precizitāti (CER - 1.48%). Svarīgi ir pārbaudīt gan runas sintēzes kvalitāti, gan precizitāti, jo šie rādītāji ne vienmēr savstarpēji korelē.

6. Balstoties uz apkopotajiem secinājumiem, pētījumos būtu ieteicams izmantot objektīvus rādītājus kā NISQA un WER vai CER, lai sistēmas būtu vieglāk savstarpēji salīdzināt.

Bibliogrāfija

[1] Junyi Ao u. c. "SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing". *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022, 5723.—5738. lpp. DOI: 10.18653/v1/2022.acl-long.393. URL: https://aclanthology.org/2022.acl-long.393.

[2] Alexei Baevski u. c. "wav2vec 2.0: A framework for self-supervised learning of speech representations". *Advances in neural information processing systems* 33 (2020), 12449.—12460. lpp.

[3] Edresson Casanova u. c. "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone". *International Conference on Machine Learning*. PMLR. 2022, 2709.—2720. lpp.

[4] Li-Wei Chen, Shinji Watanabe un Alexander I. Rudnicky. "A Vector Quantized Approach for Text to Speech Synthesis on Real-World Spontaneous Speech". *ArXiv* abs/2302.04215 (2023). URL: https://api.semanticscholar.org/CorpusID:256662411.

[5] Fabio Chiusano. *Two minutes NLP-intro to word error rate (wer) for speech-to-text*. 2022. g. febr. URL: https://medium.com/nlplanet/two-minutes-nlp-intro-to-word-error-rate-wer-for-speech-to-text-fc17a98003ea.

[6] Nicolas D'Alessandro u. c. "MaxMBROLA: A Max/MSP MBROLA-based tool for real-time voice synthesis". *2005 13th European Signal Processing Conference*. IEEE. 2005, 1.—4. lpp.

[7] Chenpeng Du u. c. "VQTTS: High-Fidelity Text-to-Speech Synthesis with Self-Supervised VQ Acoustic Feature". *ArXiv* abs/2204.00768 (2022). URL: https://api.semanticscholar.org/CorpusID:247939783.

[8] Patrick Esser, Robin Rombach un Bjorn Ommer. "Taming transformers for high-resolution image synthesis". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, 12873.—12883. lpp.

[9] PhD Everton Gomede. *Backpropagation*. 2023. g. maijs. URL: https://medium.com/@evertongomede/backpropagation-90736abaf1ca.

[10] Ian Goodfellow u. c. "Generative adversarial nets". *Advances in neural information processing systems* 27 (2014).

[11] ITU. "Vocabulary for performance, quality of service and quality of experience". (2017).

[12] Adrey Karpathy. URL: http://karpathy.github.io/2015/05/21/rnn-effectiveness/.

[13] Jaehyeon Kim, Jungil Kong un Juhee Son. “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech”. *International Conference on Machine Learning*. PMLR. 2021, 5530.—5540. lpp. URL: https://api.semanticscholar.org/CorpusID:235417304.

[14] Jaehyeon Kim u. c. “Glow-tts: A generative flow for text-to-speech via monotonic alignment search”. *Advances in Neural Information Processing Systems* 33 (2020), 8067.—8077. lpp.

[15] Sungwon Kim, Heeseung Kim un Sung-Hoon Yoon. “Guided-TTS 2: A Diffusion Model for High-quality Adaptive Text-to-Speech with Untranscribed Data”. *ArXiv* abs/2205.15370 (2022). URL: https://api.semanticscholar.org/CorpusID:249209915.

[16] Jungil Kong, Jaehyeon Kim un Jaekyoung Bae. “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis”. *Advances in Neural Information Processing Systems* 33 (2020), 17022.—17033. lpp.

[17] Jungil Kong u. c. “VITS2: Improving Quality and Efficiency of Single-Stage Text-to-Speech with Adversarial Learning and Architecture Design”. *arXiv preprint arXiv:2307.16430* (2023).

[18] Simeon Kostadinov. *Understanding backpropagation algorithm*. 2019. g. aug. URL: https://towardsdatascience.com/understanding-backpropagation-algorithm-7bb3aa2f95fd.

[19] Kenneth Leung. *Evaluate OCR output quality with character error rate (CER) and word error rate (WER)*. 2021. g. sept. URL: https://towardsdatascience.com/evaluating-ocr-output-quality-with-character-error-rate-cer-and-word-error-rate-wer-853175297510.

[20] Dan Lim, Sunghee Jung un Eesung Kim. “JETS: Jointly Training FastSpeech2 and HiFi-GAN for End to End Text to Speech”. 2022. g. sept., 21.—25. lpp. DOI: 10.21437/Interspeech.2022-10294.

[21] Yanqing Liu u. c. “DelightfulTTS 2: End-to-End Speech Synthesis with Adversarial Vector-Quantized Auto-Encoders”. *Interspeech*. 2022. URL: https://api.semanticscholar.org/CorpusID:250425685.

[22] Cory Maklin. *Transformers explained*. 2022. g. aug. URL: https://medium.com/@corymaklin/transformers-explained-610b2f749f43.

[23] Erik J. Martin. *The 2023 state of speech engines*. 2023. g. febr. URL: https://www.speechtechmag.com/Articles/ReadArticle.aspx?ArticleID=156995.

[24] Shivam Mehta u. c. “OverFlow: Putting flows on top of neural transducers for better TTS”. *arXiv preprint arXiv:2211.06892* (2022).

[25] Mayank Mishra. *Convolutional Neural Networks, explained*. 2020. g. sept. URL: https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939.

[26] Gabriel Mittag u. c. "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets". *arXiv preprint arXiv:2104.09494* (2021).

[27] Aaron van den Oord u. c. "Wavenet: A generative model for raw audio". *arXiv preprint arXiv:1609.03499* (2016).

[28] Vadim Popov u. c. "Grad-tts: A diffusion probabilistic model for text-to-speech". *International Conference on Machine Learning*. PMLR. 2021, 8599.—8608. lpp.

[29] Ryan Prenger, Rafael Valle un Bryan Catanzaro. "Waveglow: A flow-based generative network for speech synthesis". *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, 3617.—3621. lpp.

[30] Yi Ren u. c. "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech". *ArXiv* abs/2006.04558 (2020). URL: https://api.semanticscholar.org/CorpusID:219531522.

[31] Joseph Rocca. *Understanding variational autoencoders (VAES)*. 2021. g. marts. URL: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73.

[32] Olaf Ronneberger, Philipp Fischer un Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, 234.—241. lpp.

[33] Marc Schröder un Jürgen Trouvain. "The German text-to-speech synthesis system MARY: A tool for research, development and teaching". *International Journal of Speech Technology* 6 (2003), 365.—377. lpp.

[34] Jonathan Shen u. c. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions". *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2018, 4779.—4783. lpp.

[35] Kai Shen u. c. "NaturalSpeech 2: Latent Diffusion Models are Natural and Zero-Shot Speech and Singing Synthesizers". *ArXiv* abs/2304.09116 (2023). URL: https://api.semanticscholar.org/CorpusID:258187322.

[36] Xu Tan u. c. "NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality". *ArXiv* abs/2205.04421 (2022). URL: https://api.semanticscholar.org/CorpusID:248572487.

[37] Ke Tran u. c. “Unsupervised neural hidden Markov models”. *arXiv preprint arXiv:1609.09007* (2016).

[38] Turing. 2022. g. febr. URL: https://www.turing.com/kb/brief-introduction-to-transformers-and-their-power.

[39] Lilian Weng. *Flow-based deep generative models*. 2018. g. okt. URL: https://lilianweng.github.io/posts/2018-10-13-flow-models/.

[40] Lilian Weng. “What are diffusion models?”: *lilianweng.github.io* (2021. g. jūl.). URL: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/.

[41] Ruiqing Xue u. c. “FoundationTTS: Text-to-Speech for ASR Custmization with Generative Language Model”. *arXiv preprint arXiv:2303.02939* (2023).

[42] Zhe Ye u. c. “CoMoSpeech: One-Step Speech and Singing Voice Synthesis via Consistency Model”. *Proceedings of the 31st ACM International Conference on Multimedia* (2023). URL: https://api.semanticscholar.org/CorpusID:258615270.

[43] Neil Zeghidour u. c. *SoundStream: An End-to-End Neural Audio Codec*. 2021. eprint: arXiv:2107.03312.

[44] Aston Zhang u. c. *Dive into deep learning*. Cambridge University Press, 2023.

Kursa darbs ”Angļu valodas runas sintēzes modeļu salīdzinājums” izstrādāts LU Datorikas fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts pastāvīgi, izmantoti tikai tajā norādītie informācijas avoti un iesniegtā darba elektroniskā kopija atbilst izdrukai:

Autors: _____ Krišs Saulītis

Rekomendēju/nerekomendēju darbu aizstāvēšanai

Vadītājs/a: Darba vadītājs: Phd. Comp. Sc. Ēvalds Urtāns _____

Janvāris 2024

Darbs aizstāvēts kursa darba komisijas sēdē

____.____._____ .prot. Nr. ____

Komisijas sekretāre: _____