

**Rīgas Tehniskās universitātes**

**Inženierzinātņu vidusskola**

**Balss stila pārnese, izmantojot konvolūcijas neironu tīklus**

Zinātniskās pētniecības darbs dabaszinātņu nozaru grupas datorzinātnes un informācijas zinātnes nozarē

**Darba autors:**

Adriāns Piliksers

11.klases skolnieks

**Darba vadītāji:**

Ēvalds Urtāns

*Ph.D. Comp. Sci.*

Laura Fjodorova

RTU IZV ķīmijas skolotāja

Arturs Ķempelis

*Ph.D. Comp. Sci.*, RTU IZV informātikas skolotājs

**Rīga, 2022**

**Anotācija**

Pēdējo gadu laikā balss stila pārnese kļuvusi par aktuālu tēmu mašīnapmācības nozarē daudzo dažādos sektoros (digitālā drošība, medicīna, izklaide) iespējamo pielietojumu dēļ. Taču vispārējai sabiedrībai šīs tehnoloģijas paliek lielā mērā nepieejamas dēļ augstā nepieciešamo zināšanu sliekšņa, lai nepieciešamos neironu tīklu modeļus operētu un pielāgotu individuālajām vajadzībām.

Šajā darbā tiks izveidota un analizēta relatīvi vienkārša balss stila pārneses metode.

Darbs rakstīts latviešu valodā, satur 11 lapas, 5 attēlus. un tajā izmantoti 12 literatūras avoti.

**Atslēgas vārdi:** balss stila pārnese, neirālā stila pārnese, dzimuma identifikācija, neironu tīkli, konvolūcijas neironu tīkli, spektrogrammas

**Abstract**

In recent years, voice style transfer has quickly gained traction in the machine learning industry due to its potential applications in many different sectors (digital security, medicine, entertainment). However, the knowledge required to operate and adapt neural network models to individual needs remains largely inaccessible to the general public.

In this work, a relatively simple voice style transfer method has been developed and analyzed.

The thesis is written in Latvian and consists of 11 pages, 5 images, and 12 sources of literature are used.

**Keywords:** speech-to-speech audio synthesis,neural style transfer,voice style transfer, speaker identification, neural networks, spectrograms

**Saturs**

[Ievads 4](#_Toc125441992)

[1. Literatūras apskats 5](#_Toc125441993)

[1.1. Balss pārneses metodes 5](#_Toc125441994)

[1.2. Balss dzimuma klasifikācijas metodes 5](#_Toc125441995)

[1.3. Datu kopas 6](#_Toc125441996)

[2. Metožu apraksts 7](#_Toc125441997)

[2.1. Spektrogrammas 7](#_Toc125441998)

[2.2. Neirālā stila pārnese 8](#_Toc125441999)

[2.3. Modelis 10](#_Toc125442000)

[2.4. Datu kopa 10](#_Toc125442001)

[2.5. Metrikas 11](#_Toc125442002)

[Pētījuma rezultāti 11](#_Toc125442003)

[Pētījuma rezultātu analīze 11](#_Toc125442004)

[Secinājumi 12](#_Toc125442005)

[Izmantotās literatūras un avotu saraksts 13](#_Toc125442006)

# Ievads

Kopš pirmajiem eksperimentiem 20. gadsimta beigās, balss pārnese ir bijusi plaši pētīta dažādās industrijās potenciālo pielietojumu dēļ. Pirmajos mēģinājumos tika izmantotas paralelizētas datu kopas ar vienādu lingvistisko saturu, un runas lokālās nianses tika pārveidotas ar vektoru kvantizāciju. Salīdzinājumam, nesenākie pētījumi izmanto Ģeneratīvi Adversiālos Tīklus(GAN) ar neparalelizētām datu kopām, iegūstot kvalitatīvus rezultātus.

Balss-uz-balsi audio sintēze rada labākus rezultātus par teksts-uz-balsi audio sintēzi balss audio ģenerēšanā, jo tiek saglabāts emocionālais saturs, kā arī efektīvāk apstrādāti neparasti vārdi, nedzirdētas valodas un lingvistiskās nianses, piemēram, ķiķināšana un dungošana. Šo iemeslu dēļ balss-uz-balsi audio sintēze ir piemērotāka daudziem pielietojumiem vairākās nozarēs: kino nozarē(mirušo aktieru balss klonēšana), medicīnas nozarē(cilvēki ar balss implantiem atgūst savu balsi) un kiberdrošības nozarē(balss modificēšana anonimitātei).

Taču tiem, kuri nav aktīvi iesaistīti audio sintēzes jomā, šī tehnoloģija paliek nepieejama un sveša publiski pieejamu materiālu un iepriekš apmācītu modeļu trūkuma, kā arī augstā nepieciešamo zināšanu līmeņa šādu modeļu operēšanai dēļ.

Rodas jautājums – vai audio sintēzi tādā pašā līmenī var veikt kādā citā, vienkāršākā veidā? Autors šajā darbā izvirza priekšlikumu jaunai balss audio sintēzes metodei, izmantojot neirālās stila pārneses metodi.

**Pētījuma problēma:** Cik lingvistiski un stilistiski kvalitatīvi ir neirālās pārneses modeļa audio rezultāti salīdzinājumā ar sarežģītākām alternatīvām?

**Darba mērķis:** Izveidot neirālās stila pārneses modeli balss stila pārneses veikšanai.

**Darba uzdevumi:**

1. Veikt izpēti par balss dzimuma klasifikāciju, dažādām balss stila pārneses metodēm un iespējamām datu kopām.
2. Apmācīt balss dzimuma klasifikācijas modeli.
3. Apmācīt balss stila pārneses neironu tīkla modeli, izmantojot neirālo stila pārnesi.
4. Ģenerēt balss pārneses audio failus, iegūtos rezultātus analizēt un savstarpēji salīdzināt.

# 1. Literatūras apskats

Literatūras apskatā tiks aprakstīta balss pārnese un iepriekš izmantotas metodes tās sasniegšanai, runātāja identifikācija un iepriekš izmantotas metodes tās sasniegšanai, un potenciālās izmantojamās datu kopas eksperimentu veikšanai.

## 1.1. Balss pārneses metodes

Visos izpētītajos literatūras avotos datu kopas audio faili tika pārveidoti uz Mel-spektrogrammām – audio spektrogrammu tipu, kur frekvence tiek attēlota logaritmiskā skalā un katra pikseļa gaišums norāda noteiktas frekvences skaļumu noteiktā laika brīdī. Šādā veidā Mel-spektrogramas pēc iespējas simulē cilvēka auss dzirdēto. Pēc Mel-spektrogrammas modificēšanas un/vai sintezēšanas to konvertēja atpakaļ uz audio faila formātu.

Vairākos no izpētītajiem pētījumiem jaunā balss audio faila sintezēšanai tika izmantots Ģeneratīvais Adversiālais Neironu tīkls (GAN). Lyu et. al. [1] to apvienoja ar WaveNet vokoderi – specializētu audio sintezētāju, savukārt Zhou et. al. [2] to apvienoja ar WORLD vokoderi un Balss Emociju Atpazīšanas modeli (SER). Pasini [3] ievades Mel-spektrogramma sadalīja vairākās daļās gar laika asi, lai panāktu pēc patikas gara balss audio faila balss pārnesi.

Radzikowski et. al. [4] izmantoja konvolūcijas neironu tīklu – rekurento neironu tīklu(CNN-RNN) arhitektūru, lai ar runātāju identifikācijas modeļa aprēķinātu kļūdu atrastu optimālos parametrus balss akcentu maiņai. Rebryk un Beliaev [5] izmantoja kodēšanas/atkodēšanas modeļa struktūru, lai balss īpatnības un definējošās vērtības modificētu uz reducētu N-dimensiju vektoru. Wang et. al. [6] izmantoja “bottleneck” metodi kopā ar šķautņu regularizāciju, lai garantētu, ka modelī netiek saglabāta cita informācija bez runātāja balss īpatnībām.

[5] un [6] tiek izmantota “zero-shot” un “one-shot” apmācība, kuras rezultātā modelis var sintezēt modificētus balss failus no iepriekš neredzētiem vai vienu reizi redzētiem paraugiem. Šis nevajadzīgi samazina rezultātu kvalitāti, ja mērķis ir balss pārnese kādā modeļa iepriekš redzētā un sagatavotā datu kopā.

[2] tika izmantota “parallel” datu kopa – katru kopas frāzi izrunā katrs kopas runātājs. Šis atvieglotu modeļa trenēšanu, ļaujot “tieši” salīdzināt Mel-spektrogrammas. Taču šāds datu tips ievērojami samazina potenciālo datu kopas lielumu un šādus datus ir neērti un ilgi ģenerēt.

Katrā izpētītajā metodē kā divas no metrikām izmantoja sintezētā audio lingvistiskā satura līdzību ar ievades audio un sintezētā audio balss līdzību ar mērķaudio. Lingvistiskā satura noteikšanai tika izmantoti iepriekš apmācīti balss-uz-tekstu neironu tīkla modeļi, savukārt audio balss salīdzināšanai izmantoja runātāju identifikācijas neironu tīklu modeļus.

## 1.2. Balss dzimuma klasifikācijas metodes

Izpētes laikā tika noskaidrotas arī iespējamās runātāju identifikācijas metodes.

Soleymanpour un Marvi [7] balss audio failu modificēja un reducēja uz 39 dimensiju mel-frekvenču cepstrālo koeficientu(MFCC) vektoru. Pēc tam ar šiem vektoriem tika apmācīts vispārīgs neironu tīkla klasifikators, kas vēlāk ar 90% precizitāti balss audio failus pareizi klasificēja starp 22 iespējamajiem runātājiem.

Ye un Yang [8] balss audio failu konvertēja uz Mel-spektrogrammu, no kuras ar 2-D CNN un vārtoto rekurento vienību(GRU) tika atdalītas balsi definējošās īpatnības. Pēc tam ar diviem FC slāņiem un vienu SoftMax slāni tika izveidots klasifikators, kuru apmācīja ar iepriekšējā solī iegūtajiem datiem.

Ravanelli un Bengio [9] ar CNN un SincNet modeli no nemodificētā audio viļņformas filtrēja iegūtos datus, ko pēc tam izmanto neironu tīklu klasifikatora apmācībā.

Minētās metodes var apvienot ar tabulu, kurā apvienotas runātāju identitātes ar to dzimumiem, lai iegūtu, lai gan komplicētu, taču efektīvu balss dzimuma klasifikatoru.

## 1.3. Datu kopas

Izpētes laikā tika apskatītas vairākas potenciālās datu kopas eksperimentu veikšanai.

Pirmā no šīm datu kopām ir ESD – emocionālās runas datu kopa. Datu kopa sastāv no 350 frāzēm, kuras izrunājuši 10 angļu valodā runājoši un 10 mandarīnu valodā runājoši cilvēki, tās kopējais audio ilgums ir 29 stundas. Katram audio failam pievienots runātāja identifikācijas numurs un emocijas tips. Datu kopa vērtīga ar to, ka katru frāzi katrs cilvēks izrunājis 5 reizes (neitrālā, priecīgā, bēdīgā, dusmīgā, pārsteigtā noskaņā), kā arī katru frāzi ir izrunājuši visi 20 runātāji – datu kopa var izmantot ar paralēlus-datus izmantojošām metodēm. Liels trūkums šai datu kopai ir tās nelielā datu dažādība – balss pārneses modeļi varētu tikt pārpielāgoti.

Vēl tika apskatīta TIMIT datu kopa. Tā sastāv no 630 runātājiem, kas katrs izrunājuši 10 fonētiski bagātīgas frāzes angļu valodā, tās kopējais audio ilgums ir 5.4 stundas. Katram audio failam pievienota attiecīgā frāze, frāzē esošo vārdu saraksts un frāzes fonētiskais pieraksts. Viena no datu kopas priekšrocībām ir lielā runātāju dažādība un frāžu fonētiskais bagātīgums – tas ļaus modeļus apmācīt arī lingvistiski negaidītām situācijām. Viens trūkums šai datu kopai ir tās nelielais ilgums – balss pārneses modeļi var tikt nepietiekami apmācīti.

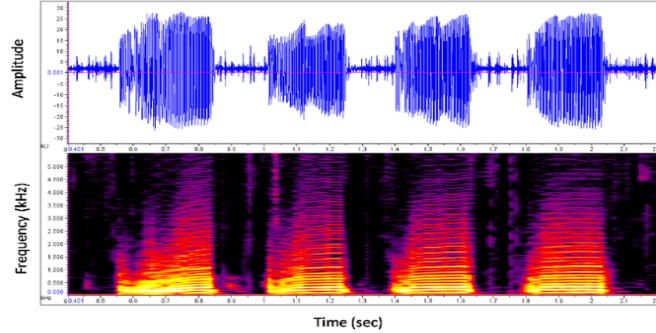
Visbeidzot tika apskatīta LibriSpeech datu kopa. Tā sastāv no audiogrāmatu fragmentiem izlasītiem angļu valodā, tās kopējais audio ilgums ir ~500 stundas. Katram audio failam pievienots runātāja identifikācijas numurs un izrunātais teksts. Lielākā datu kopas priekšrocība ir ievērojamā runātāju dažādība. Turklāt datu kopas lielums ļaus balss pārneses modeļos izslēgt jebkādus artefaktus, kādus varētu radīt neliels datu kopas izmērs. Taču lielais datu izmērs arī ievērojami paildzinās modeļa apmācības ilgumu.

# 2. Metožu apraksts

Eksperimentālajā daļā balss pārnesei tiks izmantota neirālā stila pārnesi, kā ievaddatus iestatot ievadaudio spektrogrammas.

## 2.1. Spektrogrammas

Spektrogrammas ir audio vizuāla reprezentācija divdimensiju attēla veidā. Attēla pikseļu gaišums norāda frekvences amplitūdu noteiktos laika posmos. Izšķir divu tipu spektrogrammas – lineārās spektrogrammas un Mel-spektrogrammas. Lineārām spektrogrammām frekvenču ass ir lineāra, dodot vienādu nozīmīgumu gan augstām, gan zemām frekvencēm, savukārt Mel-spektrogrammu frekvenču ass ir logaritmiska, lielāku nozīmi piešķirot zemajām frekvencēm, tādējādi simulējot cilvēka auss uztverto informāciju.

Chart

Description automatically generated

*1. attēls*. Spektrogrammas un fāžudiagrammas salīdzinājums[14].

*2. attēls*. Spektrogrammas trīsdimensionāla vizualizācija[15].

Audio lineāro spektrogrammu iegūst, veicot STFT(Short-Time Fourier Transform), kuras procesā tiek aprēķinātas frekvenču amplitūdas nelielos, savstarpēji pārklājošos laika posmos, kurus pēc tam apvieno. Atšķirībā no SFT(Standard Fourier Transform), kas aprēķina frekvenču sadalījumu nemainīgam audio signālam, STFT frekvenču sadalījumu aprēķina audio signālam, kas laika gaitā mainās, piemēram, mūzikai, runai vai citām dabā sastopamām skaņām. Mel-spektrogrammu iegūst, piemērojot logaritmizēšanu lineārās spektrogrammas frekvenču asij. Jāņem vērā, ka, spektrogrammas frekvenču asi logaritmizējot, tiek zaudēti dati par spektrogrammas augstāko frekvenču amplitūdām.

Graphical user interface, application, table

Description automatically generated

*3. attēls*. Lineārās un Mel-spektrogrammas salīdzinājums[16].

Audio no spektrogrammas iegūst, veicot Griffin-Lim audio rekonstrukcijas algoritmu. Tiek iteratīvi veikta ISTFT(Inverse Short-Time Fourier Transform), iegūtajiem datiem veikta STFT un iegūtajam rezultātam amplitūdas matrica aizstāta ar jau doto spektrogrammu. Pietiekami lielam iterāciju skaitam audio fāzu diagramma atbildīs patiesajam audio signālam.

Spektrogrammu frekvenču amplitūdu var pārveidot uz decibelu skalu, šādi precīzāk vizualizējot cilvēka auss dzirdēto. To panāk, normalizējot un logaritmizējot visus spektrogrammas datus. Balansētu informācijas daudzumu var iegūt, pirms logaritmizēšanas datiem pieskaitot noteiktu konstanti, tādējādi ierobežojot mazāko dB vērtību pēc logaritmizēšanas.

Spektrogrammas var arī ģenerēt, izmantojot vokoderus – kompleksus neironu tīklus ar mērķi maksimāli ātri un kvalitatīvi ģenerēt audio viļņformu no dotas Mel-spektrogrammas. Piemēram, HiFi-GAN[10] vokoderis izmanto CNN(Convolutional Neural Network) un multi-periodu diskrimināciju, lai reģistrētu dažādas audio satura sakarības, kuras izmanto vēlāk, audio fāžu diagrammas ģenerēšanā.

## 2.2. Neirālā stila pārnese

Neirālās stila pārneses mērķis ir no dotiem satura un stila attēliem ģenerēt saplūdinātu attēlu tā, lai tas kopumā izskatītos pēc satura attēla, bet būtu “uzzīmēts” stila attēla stilā. Piemēram, kā satura un stila attēlus izmantojot attiecīgi kādas pilsētas attēlu un dažādu mākslinieku gleznu attēlus, ar neirālo stila pārnesi iegūst attiecīgo izvadi:



*4. attēls*. Neirālā stila pārnese dažādiem stila ievaddatiem un vienādiem satura ievaddatiem[17].

Vispirms tiek izveidots CNN modelis ar nejauši iestatītiem filtriem, kā arī randomizēts izvadattēls. Vienu reizi CNN modelī tiek apstrādāti satura un stila ievadattēli, visu CNN slāņu aktivizācijas tiek saglabātas.

Iteratīvi CNN modelī tiek apstrādāts izvadattēls. Tiek saglabātas katra CNN slāņa aktivizācijas. Tiek aprēķinātas trīs kļūdas funkcijas.

* Satura kļūdas funkcija tiek iegūta, aprēķinot MSE(Mean Square Error) starp attiecīgiem CNN slāņu rezultātiem saturattēlam un izvadattēlam. Funkcija tuvojas nullei, kad CNN slāņu rezultāti abiem attēliem kļūst vienādi(spektrogrammu saturi kļūst vienādi).

1. ,

kur ir i-tais rezultāts CNN slānī, kuru ieguva, apstrādājot saturattēlu, un ir i-tais rezultāts CNN slānī, kuru ieguva, apstrādājot izvadattēlu.

* Stila kļūdas funkcija tiek iegūta, aprēķinot MSE(Mean Square Error) starp attiecīgo CNN slāņu rezultātu Gram-matricām saturattēlam un izvadattēlam. Funkcija tuvojas nullei, kad CNN slāņu rezultātu savstarpējās korelācijas abiem attēliem kļūst vienādas(spektrogrammu iekšējās sakarības kļūst vienādas).

1. ,

kur ir Gram-matrica(), un ir CNN l-tā slāņa i-tais rezultāts, kuru ieguva, apstrādājot kādu no attēliem.

* Gala variācijas zuduma funkcija tiek iegūta, katru no iepriekšminētajām funkcijām pareizinot ar iepriekšnoteiktu koeficientu.

Izmantojot iegūto kļūdas funkcija, tiek aprēķināti tās parciālatvasinājums attiecībā pret izvadattēlu. Visbeidzot, izvadattēls tiek atjaunināts parciālatvasinājuma virzienā par kādu iepriekšnoteiktu lielumu.

1. ,

kur x ir izvadattēls, ir apmācības ātruma parametrs un Q ir kļūdas funkcijas parciālatvasinājums attiecībā pret izvadattēlu.

Ievērojam, ka modeļa optimizācijas laikā CNN filtri nemainās, mainās tikai izvades spektrogramma.

Kad ir sasniegts noteiktais iterāciju skaits, izvadattēls tiek atgriezts kā izvade.

## 2.3. Modelis

Abi audio faili ar STFT vispirms tiek pārveidoti uz lineārajām spektrogrammām. Tiek inicializēts CNN modelis un izvadattēls no satura spektrogrammas. Abus ievadattēlus un izvadattēlu atkārtoti apstrādā CNN. Tiek aprēķināta gala variācijas zaudējuma funkcija, izmantojot satura un stila zaudējuma funkcijas. Pēc tam aprēķina gala variācijas zaudējuma funkcijas parciālatvasinājumu attiecībā pret izvadattēlu. Izvadattēls tiek atjaunināts parciālatvasinājuma virzienā. Šo procesu atkārto noteiktu iterāciju skaitu. Visbeidzot izvadattēls kā spektrogramma tiek pārveidots uz audio signālu, izmantojot Griffin-Lim algoritmu.

Autora izveidotais modelis no daudziem citiem modeļiem atšķiras ar jebkurš-uz-jebkuru balss pārnesi, t.i., modelis var apstrādāt vēl nedzirdētas balsis un neredzētu lingvistisko saturu. Tāpāt modelī bija iekļauti vairāki parametri runas līdzības lingvistiskajam saturam un balss stilam regulēšanai.

Diagram, schematic

Description automatically generated

*5. attēls*. Modeļa arhitektūras diagramma.

## 2.4. Datu kopa

Eksperimenta nolūkos izmantos LibriSpeech[12] datu kopu. Šī kopa tika izvēlēta dēļ lielā ierakstu apjoma katram runātājam. Tādējādi iespējams kā stila audio izmantot vairākus viena runātāja runātos ierakstus. Katru audio ierakstu vispirms atsevišķi apstrādāja, no tā izņemot visus klusuma(<-45dB) brīžus, kas bija ilgāki par 100ms. Pēc tam audio ierakstus sagrupēja pēc runātājiem. Visbeidzot katram audio ierakstam tika piešķirta bināra vērtība, kas raksturo balss dzimumu.

## 2.5. Metrikas

Lai noteiktu ģenerēto balss failu kvalitāti, tika apmācīts atsevišķs balss dzimuma klasifikācijas modelis. No katra LibriSpeech datu kopas audio ieraksta tika iegūtas balss audio pirmās 40 MFCC koeficientu rindas, šo rindu deltas un rindu deltu deltas. Šīs vērtības apvienoja 120-dimensiju vektorā. Pēc tam ar visiem iegūtajiem vektoriem tika apmācīts SVM(Support Vector Machine) modelis ar radiālo bāzes kodola funkciju, kas pēc 200 iterācijām sasniedza 98.5% precizitāti iepriekš dzirdētām balsīm un 89.7% precizitāti iepriekš nedzirdētām balsīm.

Ģenerēto balss ierakstu kvalitāti noteica, līdzīgu procesu piemērojot arī šo failu saturam, pēc tam arī visus iegūtos vektorus klasificējot SVM modelī un kā galveno dzimumu piešķirot biežāk izvēlēto.

# Pētījuma rezultāti

Darba ietvaros tika ģenerēti vairāki audio faili, par satura un stila ievadaudiofailiem izvēloties vairākas LibriSpeech datu kopas audiofailu kombinācijas. Daļa no ģenerētajiem audiofailiem ir pieejama tīmekļvietnē GitHub­[[1]](#footnote-1).

Darba gaitā izveidotais pirmkods audiofailu ģenerēšanai un dzimuma klasifikācijai ir pieejams platformā GitHub[1. pielikums].

# Pētījuma rezultātu analīze

Klausoties ģenerētos audio failus, balsī var novērot ievērojamu līdzību ievaddatu stila balsij. Tāpat, salīdzinot tos ar satura ievadaudio, sākotnēji runātais teksts lielākoties ir atpazīstams.

Apstrādājot iegūtos audiofailus ar apmācīto dzimuma klasifikatoru, ģenerētie faili vidēji ieguva 99% precizitāti attiecībā pret sagaidīto dzimumu ar 95% pārliecību.

Datorprogramma tika izpildīta ar divām dažādām tehnoloģijām – AMD Ryzen 5 5600H datorprocesoru un RTX 3050 Mobile videokarti. Abos gadījumos rezultātu kvalitāte neatšķiras, taču rezultātu ģenerēšana ar videokarti aizņēma būtiski mazāk laika(vidēji 100 reizes ilgāk par satura ievadaudio garumu) par uz procesora aizņemto laiku(vidēji 2000 reizes ilgāk par satura ievadaudio garumu).

Kritiski izvērtējot ģenerētos audiofailus, autors formulēja divas galvenās rezultātu nepilnības un izvirzīja priekšlikumus turpmākām rīcībām nepilnību novēršanai.

* Ģenerētajos audiofailus bija novērojams atšķirīgo balss nianšu trūkums. To var izskaidrot ar nepietiekamu konvolūcijas neironu tīkla slāņu daudzumu un/vai nepietiekamu starprezultātu skaitu katrā slānī. Šos parametrus palielināt nebija iespējams videokartes atmiņas trūkuma dēļ. Šo nepilnību varētu novērst ar lielāku videokartes atmiņas daudzumu, vairāku videokaršu izmantošanu paralēlai koda izpildei, vai ar iepriekš apmācīta advancēta konvolūcijas neironu tīkla izmantošanu, piemēram, VGG16[13].
* Ģenerētajos audiofailos tika pamanītas arī nevajadzīgi paaugstinātas amplitūdas visām frekvencēm, kas bija dzirdamas kā sīkšana audio. Šo nepilnību varētu novērst ar atsevišķu minimālās frekvences amplitūdas kļūdas funkcijas izmantošanu, lai samazinātu nevajadzīgu frekvences amplitūdas palielināšanos nevajadzīgās audiofaila daļās.

# Secinājumi

1. Darba mērķis ir sasniegts, jo tika izveidots konvolūcijas neironu tīkls, kas spēj pārnest viena audiofaila balss stilu uz cita audiofaila runas saturu. Tika sasniegti arī visi izvirzītie darba uzdevumi.
2. Tika salīdzināta vairāku hiperparametru kombināciju radīto rezultātu kvalitāte. Viskvalitatīvākie rezultāti tika sasniegti ar 3 slāņu konvolūcijas neironu tīklu, 1/1000 stila un satura koeficientu attiecību, un pēc 2500 modeļa iterācijām.
3. Konvolūcijas neironu tīkla ģenerētos rezultātus autors uzskata par apmierinošiem, jo tajos lielākoties atpazīstams sākotnēji teiktais saturs, savukārt dzirdamas balss stila iezīmes, kas atšķir dzirdamo runātāju no sākotnējā satura ievadaudio runātāja, tas ir, veiksmīgi tikusi veikta balss stila pārnese.
4. Ģenerēto balsu dzimumi atbilst sagaidītajam dzimumam(stila balss dzimumam) ar 94% pārliecību. Tas liecina, ka tikusi veikta arī balss dzimuma pārnese.

# Izmantotās literatūras un avotu saraksts

1. AlBadawy, E.A., & Lyu, S. (2020). Voice Conversion Using Speech-to-Speech Neuro-Style Transfer. INTERSPEECH.
2. Zhou, K., Sisman, B., Liu, R., & Li, H. (2021). Seen and Unseen Emotional Style Transfer for Voice Conversion with A New Emotional Speech Dataset. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 920-924.
3. Pasini, M. (2019). MelGAN-VC: Voice Conversion and Audio Style Transfer on arbitrarily long samples using Spectrograms. ArXiv, abs/1910.03713.
4. Radzikowski, K., Wang, L., Yoshie, O., & Nowak, R. (2021). Accent modification for speech recognition of non-native speakers using neural style transfer. EURASIP Journal on Audio, Speech, and Music Processing, 2021, 1-10.
5. Rebryk, Y., & Beliaev, S. (2020). ConVoice: Real-Time Zero-Shot Voice Style Transfer with Convolutional Network. ArXiv, abs/2005.07815.
6. Wang, Z., Xie, Q., Li, T., Du, H., Xie, L., Zhu, P., & Bi, M. (2022). One-Shot Voice Conversion For Style Transfer Based On Speaker Adaptation. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6792-6796.
7. Soleymanpour, M., & Marvi, H. (2017). Text-independent speaker identification based on selection of the most similar feature vectors. International Journal of Speech Technology, 20, 99-108.
8. Ye, F., & Yang, J. (2021). A Deep Neural Network Model for Speaker Identification. Applied Sciences, 11, 3603.
9. Ravanelli, M., & Bengio, Y. (2018). Speaker Recognition from Raw Waveform with SincNet. 2018 IEEE Spoken Language Technology Workshop (SLT), 1021-1028.
10. Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. Neural Information Processing Systems, 33, 17022–17033
11. Lawlor, Bob. “A Novel Efficient Algorithm for Voice Gender Conversion.” (1999).
12. Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5206-5210.
13. Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, abs/1409.1556.
14. Kovitvongsa, Ke and Ps Lobel. “Convenient Fish Acoustic Data Collection in the Digital Age.” (2009).
15. Pieejams: <https://physics.stackexchange.com/questions/632697/are-soundproof-foam-shapes-reflective-of-audio-frequencies> [Skatīts 09.01.2023.]
16. Pieejams: <https://ai.googleblog.com/2018/10/acoustic-detection-of-humpback-whales.html> [Skatīts 09.01.2023.]
17. Gatys, Leon A. et al. “A Neural Algorithm of Artistic Style.” ArXiv abs/1508.06576 (2015): n. pag.

**Pielikums**

1. Pielikums – izstrādātās programmatūras pirmkods, balss dzimuma klasifikatora pirmkods un apstrādātā datu kopa: <https://github.com/ASomniphobeHere/Classifier> un <https://github.com/ASomniphobeHere/NeuralStyleTransferVoiceConversion> (šeit arī pieejami ģenerētie audiofaili)

1. <https://github.com/ASomniphobeHere/NeuralStyleTransferVoiceConversion/blob/main/README.md> [↑](#footnote-ref-1)