VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

FUNDAMENTINIŲ MOKSLŲ FAKULTETAS

INFORMACINIŲ SISTEMŲ KATEDRA

Emilis Kaplunovas

SINTETINIŲ IR NESINTETINIŲ BALSO ĮRAŠŲ IDENTIFIKAVIMAS TAIKANT DIRBTINĮ INTELEKTĄ

IDENTIFICATION OF SYNTHETIC AND NON-SYNTHETIC VOICE RECORDINGS USING ARTIFICIAL INTELLIGENCE

Baigiamasis magistro darbas

Informacijos ir informacinių technologijų saugos studijų programa, valstybinis kodas 6211BX014

Informatikos inžinerijos studijų kryptis

Vilnius, 2024

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

FUNDAMENTINIŲ MOKSLŲ FAKULTETAS

INFORMACINIŲ SISTEMŲ KATEDRA

Emilis Kaplunovas

SINTETINIŲ IR NESINTETINIŲ BALSO ĮRAŠŲ IDENTIFIKAVIMAS TAIKANT DIRBTINĮ INTELEKTĄ

IDENTIFICATION OF SYNTHETIC AND NON-SYNTHETIC VOICE RECORDINGS USING ARTIFICIAL INTELLIGENCE

Baigiamasis magistro darbas

Informacijos ir informacinių technologijų saugos studijų programa, valstybinis kodas 6211BX014

Informatikos inžinerijos studijų kryptis

Vadovas	prof. dr. Nikolaj Goranin (Moksl. laipsnis/pedag. vardas, vardas, pavardė)
Konsultantas	_____ (Moksl. laipsnis/pedag. vardas, vardas, pavardė)
Konsultantas	_____ (Moksl. laipsnis/pedag. vardas, vardas, pavardė)
Lietuvių kalbos konsultantas	dr. Vaida Buivydienė (Moksl. laipsnis/pedag. vardas, vardas, pavardė)

Vilnius, 2024

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
INFORMACINIŲ SISTEMŲ KATEDRA

Studijų kryptis: Informatikos inžinerija

Studijų programa: Informacijos ir informacinių technologijų sauga, valstybinis kodas 6211BX014

Specializacija: Informacijos ir informacinių technologijų sauga

TVIRTINU
Katedros vedėjas
Nikolaj Goranin
2024-05-29

MAGISTRO BAIGIAMOJO DARBO UŽDUOTIS

Nr. ITSfm-22–8751

Vilnius

Studentas (-ė): Emilis Kaplunovas

Baigiamojo darbo tema: Sintetinių ir nesintetinių balso įrašų identifikavimas taikant DI

Baigiamojo darbo užbaigimo terminas pagal numatytą studijų kalendorinį grafiką.

BAIGIAMOJO DARBO UŽDUOTIS:

Tikslas: Padidinti sintetinių ir nesintetinių balso įrašų atpažinimo tikslumą taikant dirbtinio intelekto metodus.

Uždaviniai:

1. Atlikti sintetinių ir nesintetinių balso įrašų atpažinimo metodų bei duomenų rinkinių analizę.
2. Surinkti lietuvių kalbos sintetinių ir nesintetinių balso įrašų duomenų rinkinį.
3. Identifikuoti perspektyvius ML/DL metodus, juos pritaikyti sudarytam duomenų rinkiniui, atlikti jų parametrų optimizavimą, atlikti gautų metrikų analizę.

Planuojamas rezultatas: Lietuvių kalbos sintetinių ir nesintetinių balso įrašų duomenų rinkinys, apmokyti ir optimizuoti ML/DL modeliai skirti nurodytam uždaviniui, metodų testavimo rezultatai.

Vadovas profesorius Nikolaj Goranin

Anotacija

Vilniaus Gedimino technikos universitetas
Fundamentinių mokslų fakultetas
Informacinių sistemų katedra

ISBN	ISSN
Egz. sk.	
Data .....-.....-.....	

Antrosios pakopos studijų **Informacijos ir informacinių technologijų saugos** programos magistro baigiamasis darbas	
Pavadinimas	**Sintetinių ir nesintetinių balso įrašų identifikavimas taikant DI**
Autorius	**Emilis Kaplunovas**
Vadovas	**Nikolaj Goranin**

	Kalba: lietuvių

Anotacija

Didėjant poreikiui identifikuoti sintetinius balso įrašus tam tikri iššūkiai išlieka. Pažangūs gilaus mokymosi metodai, tokie kaip „WaveNet“„Tacotron“, GAN pagrįsti metodai paskatino sintetinės kalbos generavimą, kuri beveik nesiskiria nuo žmogaus kalbos. Didžioji dalis tyrimų yra atlikta su rinkiniais anglų kalba, todėl šiuo tyrimu bus prisidėta prie sintetinių balso įrašų identifikavimo lietuvių kalba. Tobulėjant garso manipuliacijų technologijai sintetiniai balso įrašai gali būti piktybiškai naudojami apsimetinėti tam tikrais asmenimis, skleisti dezinformaciją arba vykdyti nesąžiningą veiklą, todėl tai padidino patikimų identifikavimo metodų poreikį. Identifikavimo metodų kūrimas įvairioms kalboms, įskaitant mažiau išteklių turinčias, tokias kaip lietuvių, padeda užtikrinti identifikavimo sistemų patikimumą su įvairiomis kalbomis. Apžvelgus identifikavimo modelius išsiaiškinta, kad egzistuoja įvairios mašinio mokymosi ir gilaus mokymosi architektūros. Mašininio mokymosi algoritmai reikalauja rankinio požymių išskyrimo ir duomenų paruošimo. Bandymams su gilaus mokymosi algoritmais garso įrašams reikia specifinių transformacijų. Eksperimentams atlikti pasirinkti trys identifikavimo modeliai: „Inc-TSSDNet“, „Res-TSSDNet“, „2D-Res-TSSDNet“. Šių modelių pagrindu kaip patobulinimai buvo sukurti savo trys identifkavimo modeliai: „RawNet“, „2D-Inc-TSSDNet“, kombinuotas „RawNet“ ir „Inc-TSSDNet“. Surinktas savo duomenų rinkinys lietuvių kalba, iš viso apie 18 000 įrašų, sintetiniams balso įrašams sugeneruoti panaudotos tekstas į kalbą technikos „VITS“ ir „LIEPA TTS“, nesintetiniai balso įrašai surinkti iš audio knygų. Atlikus stratifikuotą kryžminį tikrinimą ir apskaičiavus visu modelių vidurkių rezultatus ir standartinius nuokrypius nustatyta, kad aukščiausi rezultatai buvo pasiekti su kombinuotu identifikavimo modeliu, tikslumas – 86,64%. Antrą geriausia tikslumą pademonstravo pritaikytas „2D-Inc-TSSDNet“ modelis – 84,84%, su „RawNet“ tikslumas buvo 84,12%.

Prasminiai žodžiai: Sintetiniai balso įrašai, tekstas į kalbą, gilusis mokymas, duomenų rinkinys, identifikavimo modeliai, modelių tikslumas.

Annotation

Vilnius Gediminas Technical University
Faculty of Fundamental Sciences
Department of Information Systems

ISBN	ISSN
Copies No.	
Date .....-.....-.....	

Master Degree Studies **Information and Information Technologies Security** study programme Master Graduation Thesis	
Title	**Identification of Synthetic and Non-Synthetic Voice Recordings Using AI**
Author	**Emilis Kaplunovas**
Academic supervisor	**Nikolaj Goranin**

Thesis language: Lithuanian

Annotation

As the need to identify synthetic voice recordings increases, certain challenges remain. Advanced deep learning techniques such as WaveNet, Tacotron, GAN-based techniques have led to the generation of synthetic language that is almost indistinguishable from human speech. Most of the research is done with datasets in English language, so this research will contribute to the identification of synthetic voice recordings in Lithuanian language. Advances in audio manipulation technology have made it possible for synthetic voice recordings to be maliciously used to impersonate certain individuals, spread misinformation or engage in fraudulent activity, which has increased the need for reliable identification methods. Developing identification methods for different languages, including less resourced ones such as Lithuanian, helps ensure the reliability of identification systems with different languages. A review of identification models reveals that there are various machine learning and deep learning architectures. Machine learning algorithms require manual feature extraction and data preparation. Testing deep learning algorithms on audio requires specific transformations. Three identification models were selected for experiments: Inc-TSSDNet, Res-TSSDNet, 2D-Res-TSSDNet. Based on these models, three identification models were developed as improvements: RawNet, 2D-Inc-TSSDNet, combined RawNet, and Inc-TSSDNet. A self- collected data set in Lithuanian language was used, about 18,000 recordings in total, text-to-speech techniques VITS and LIEPA TTS were used to generate synthetic voice recordings, non-synthetic voice recordings were collected from audio books. After performing stratified cross-validation and calculating the average results and standard deviations of all models, it was found that the highest results were achieved with the combined identification model, with an accuracy of 86.64%. The second best accuracy was demonstrated by the adapted 2D-Inc-TSSDNet model at 84.84%, with RawNet at 84.12%.

Keywords: Synthetic voice recordings, text-to-speech, deep learning, dataset, identification models, model accuracy.

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Emilis Kaplunovas, 20172290

(Studento vardas ir pavardė, studento pažymėjimo Nr.)

Fundamentinių mokslų fakultetas

(Fakultetas)

Informacijos ir informacinių technologijų sauga, ITSfm-22

(Studijų programa, akademinė grupė)

BAIGIAMOJO DARBO (PROJEKTO) SĄŽININGUMO DEKLARACIJA

2024 m. gegužės 30 d.

Patvirtinu, kad mano baigiamasis darbas tema „Sintetinių ir nesintetinių balso įrašų identifikavimas taikant DI“ yra savarankiškai parašytas. Šiame darbe pateikta medžiaga nėra plagijuota. Tiesiogiai ar netiesiogiai panaudotos kitų šaltinių citatos pažymėtos literatūros nuorodose.

Mano darbo vadovas profesorius daktaras Nikolaj Goranin.

Kitų asmenų indėlio į parengtą baigiamąjį darbą nėra. Jokių įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs (-usi).

Emilis Kaplunovas

(Parašas) (Vardas ir pavardė)

SUTIKIMAS DĖL ASMENS DUOMENŲ NAUDOJIMO

2024–05–30

(Data)

Šiuo sutikimu aš, Emilis Kaplunovas (toliau – Duomenų subjektas) sutinku, kad Vilniaus Gedimino technikos universitetas, juridinio asmens kodas 111950243, adresas Saulėtekio al. 11, LT-10223 Vilnius (toliau – Duomenų valdytojas), tvarkytų mano asmens duomenis kitų studentų mokymosi tikslu. T. y. tvarkytų (*pažymėkite tinkamą*):

☐ vardą, pavardę, bakalauro baigiamąjį darbą;
☐ bakalauro baigiamąjį darbą, nenurodant vardo, pavardės;
☒ vardą, pavardę, magistro baigiamąjį darbą;
☐ magistro baigiamąjį darbą nenurodant vardo, pavardės.

Šiuo tikslu tvarkomų asmens duomenų Duomenų valdytojas neperduos jokiems tretiesiems asmenims, studentams su baigiamaisiais darbais bus leidžiama susipažinti vidinėje informacinėje sistemoje. Duomenų subjekto asmens duomenys šiuo tikslu bus naudojami ne ilgiau nei 5 metai.

Šiuo sutikimu Duomenų subjektas patvirtina, kad yra supažindintas su šiomis teisėmis:

- Susipažinti su savo duomenimis ir kaip jie yra tvarkomi (teisė susipažinti);
- Reikalauti ištaisyti arba, atsižvelgiant į asmens duomenų tvarkymo tikslus papildyti asmens neišsamius asmens duomenis (teisė ištaisyti);
- Savo duomenis sunaikinti arba sustabdyti savo duomenų tvarkymo veiksmus (išskyrus saugojimą) (teisė sunaikinti ir teisė „būti pamirštam“);
- Reikalauti, kad asmens duomenų valdytojas apribotų asmens duomenų tvarkymą (teisė apriboti);
- Teise į duomenų perkėlimą (teisė perkelti);
- Nesutikti, kad būtų tvarkomi asmens duomenys, kai šie duomenys tvarkomi ar ketinami tvarkyti kitais tikslais;
- Pateikti skundą Valstybinei duomenų apsaugos inspekcijai;

Duomenų subjektas turi teisę bet kuriuo metu atšaukti savo sutikimą.

Emilis Kaplunovas

[Duomenų subjekto vardas, pavardė, parašas]

Turinys

Paveikslų sąrašas 9
Lentelių sąrašas 10
Santrumpos 11
Įvadas 12
1. Sintetinių balso įrašų identifikavimas 15
    1.1. Sintetinių balso įrašų generavimo ir jų identifikavimo problematika 15
    1.2. Sintetinių balso įrašų generavimo tipai 16
        1.2.1. Sintetiniu būdu paremtas arba tekstas į kalbą generavimo būdas 17
        1.2.2. Vieneto atranka paremtas generavimo būdas 18
        1.2.3. Parametrine sinteze paremtas generavimo būdas 18
        1.2.4. Statistine parametrine sinteze paremtas generavimo būdas 19
        1.2.5. Giliu neuroniniu tinklu (DNN) paremtas generavimo būdas 20
    1.3. Sintetinių ir nesintetinų balso įrašų identifikavimo metodai 21
    1.4. Sintetinių ir nesintetinių balso įrašų identifikavime naudojami duomenų rinkiniai 28
    1.5. Pirmo skyriaus apibendrinimas 32
    1.6. Pirmo skyriaus išvados ir tolimesni darbai 32
2. Siūlomas būdas 34
    2.1. Duomenų rinkinys 35
    2.2. Klasifikavimo modeliai 37
    2.3. Rezultatų įvertinimas 39
    2.4. Antro skyriaus apibendrinimas 40
    2.5. Antro skyriaus išvados 40
3. Eksperimentinė dalis 42
    3.1. Duomenų rinkinio surinkimas 42
    3.2. Modelių veikimo reguliavimo pritaikymas 43
    3.3. Naudojami identifikavimo modeliai eksperimente 44
        3.3.1. „Res-TSSDNet“ identifikavimo modelis 44
        3.3.2. „2D-Res-TSSDNet“ identifikavimo modelis 47
        3.3.3. „Inc-TSSDNet“ identifikavimo modelis 49
        3.3.4. „RawNet“ identifikavimo modelis 52
        3.3.5. „2D-Inc-TSSDNet“ identifikavimo modelis 54
        3.3.6. Kombinuotas „Inc-TSSDNet“ su „RawNet“ identifikavimo modelis 57
    3.4. Išmetimo (angl. *dropout*) sluoksnių pritaikymas identifikavimo modeliuose 59
    3.5. SMOTE technikos pritaikymas eksperimente 62
    3.6. Stratifikuotas kryžminis tikrinimas 65
    3.7. Klonuotų balso įrašų identifikavimas 68
    3.8. Trečio skyriaus apibendrinimas 69
    3.9. Trečio skyriaus išvados 69
Išvados 71
Literatūros sąrašas 74
Priedai 76
    1 Priedas. Pažymėjimas už dalyvavimą jaunųjų mokslininkų konferencijoje 77
    2 Priedas. Modelių tikslumo vidurko rezultatai stratifikuoto tikrinimo dalyse 78

Paveikslų sąrašas

1.1 pav. Tesktas į kalbą sistemų eigos diagrama. Parengta pagal (Masood et al., 2022) 18
1.2 pav. Tipinės HMM pagrįstos kalbos sintezės sistemos blokinė schema. Parengta pagal (Ling et al., 2015) 20
1.3 pav. Dviejų paslėptų sluoksnių DNN modelio reprezentacija. Parengta pagal (Ling et al., 2015) 21
1.4 pav. Siūlomos sistemos supaprastinta strukūra. Parengta pagal (P et al., 2020) 24
1.5 pav. Pasiūlyto metodo supaprastinta strukūra. Parengta pagal (Bartusiak & Delp, 2021) 26
1.6 pav. Siūlomos sistemos blokinė schema. Parengta pagal (Wu et al., 2020) 28
2.1 pav. Siūlomo būdo UML veiklos diagrama 35
2.2 pav. Garso įrašo sintezavimas pasinaudojant TTS (angl. *Text-to-speech*) technika 36
2.3 pav. „*Res-TSSDNet*“ modelio strukūra. Parengta pagal (Hua et al., 2021) 38
2.4 pav. „*RawBlock*“ bloko struktūra 39
3.1 pav. Gautos ROC kreivės su „*Res-TSSDNet*“ modeliu 46
3.2 pav. „*Res-TSSDNet*“ sutrikimo matricos 46
3.3 pav. Gautos ROC kreivės su „*2D-Res-TSSDNet*“ modeliu 48
3.4 pav. Gautos sutrikimo matricos su „*2D-Res-TSSDNet*“ modeliu 49
3.5 pav. „*Inc-TSSDNet*“ modelio ROC kreivės 51
3.6 pav. „*Inc-TSSDNet*“ sutrikimo matricos 51
3.7 pav. Gautos ROC kreivės su „*RawNet*“ modeliu 53
3.8 pav. „*RawNet*“ sutrikimo matricos 54
3.9 pav. Gautos ROC kreivės su „*2D-Inc-TSSDNet*“ modeliu 56
3.10 pav. Gautos sutrikimo matricos su „*2D-Inc-TSSDNet*“ modeliu 56
3.11 pav. „*2D-Res-TSSDNet*“ ir „*2D-Inc-TSSDNet*“ modelių palyginimas 57
3.12 pav. Gautos ROC kreivės su kombinuotu modeliu 58
3.13 pav. Gautos sutrikimo matricos su kombinuotu identifikavimo modeliu 59
3.14 pav. WCE ir išmetamųjų sluoksnių kombinavimo palyginimas 61
3.15 pav. Sumaišymo ir išmetamųjų sluoksnių kombinavimo palyginimas 61
3.16 pav. Etikečių išlyginimo ir išmetamųjų sluoksnių kombinavimo palyginimas 62
3.17 pav. Gauti tikslumo rezultatai su „*RawNet*“ identifikavimo modeliu pritaikius SMOTE techniką 64
3.18 pav. Gauti EER rezultatai su „*RawNet*“ identifikavimo modeliu pritaikius SMOTE techniką 65
3.19 pav. Identifikavimo modelių tikslumo rezultatų palyginimas po vidurkių apskaičiavimo 67
3.20 pav. Identifikavimo modelių EER rezultatų palyginimas po vidurkių apskaičiavimo 68

Lentelių sąrašas

1 lentelė. Tyrime naudoto „*ASVspoof 2019*“ duomenų rinkinio statistika pagal sintetinių ir tikrų įrašų skaičių kiekviename poaibyje, vyrų ir moterų kalbėtojų skaičių, klastojimo algoritmus ir diskretizavimo dažnį. Parengta pagal (Arif et al., 2021) 31
2 lentelė. Palyginamoji straipsnių ir jose taikytų metodų, duomenų rinkinių lentelė 33
3 lentelė. Surinkto duomenų rinkinio nesintetinių ir sintetinių balso įrašų skaičius 43
4 lentelė. „*Res-TSSDNet*“ modelio rezultatai priklausomai nuo modulių skaičiaus 45
5 lentelė. Gauti rezultatai su „*Res-TSSDNet*“ modeliu 45
6 lentelė. „*2D-Res-TSSDNet*“ modelio rezultatai priklausomai nuo modulių skaičiaus 47
7 lentelė. Gauti rezultatai su „*2D-Res-TSSDNet*“ modeliu 47
8 lentelė. „*Inc-TSSDNet*“ modelio rezultatai priklausomai nuo „*DilatedCovModule*“ modulių skaičiaus 50
9 lentelė. Gauti rezultatai su „*Inc-TSSDNet*“ modeliu 50
10 lentelė. „*RawNet*“ modelio rezultatai priklausomai nuo „*RawBlock*“ blokų skaičiaus 52
11 lentelė. Gauti rezultati su „*RawNet*“ modeliu 52
12 lentelė. „*2D-Inc-TSSDNet*“ modelio rezultatai priklausomai nuo „*DilatedCovModule*“ skaičiaus 55
13 lentelė. Gauti rezultatai su „*2D-Inc-TSSDNet*“ modeliu 55
14 lentelė. Gauti rezultatai su „*Inc-TSSDNet*“ ir „*RawNet*“ kombinuotu modeliu 58
15 lentelė. Gauti rezultatai su išmetimo sluoksniais 60
16 lentelė. Apmokymo poaibis prieš ir po SMOTE technikos pritaikymo 62
17 lentelė. Gauti rezultatai pritaikant SMOTE techniką 63
18 lentelė. Apskaičiuoti 5 dalių kryžminio tikrinimo vidurkiai pagal naudojamas reguliavimo technikas 66
19 lentelė. Identifikavimo modelių tikslumo ir EER rezultatų vidurkiai 67
20 lentelė. Klonuotų ir nesintetinių balso įrašų analizės rezultatai 68

Santrumpos

GAN (angl. *Generative Adversarial Networks*) – generatyvūs priešpriešiniai tinklai

XAI (angl. *Explainable artificial intelligence*) – aiškinamasis dirbtinis intelektas

LRP (angl. *Layer-wise relevance propagation*) – svarbos skleidimas per sluoksnius

Q-SVM (angl. *Quadratic Support Vector Machine*) – kvadratinių atraminių vektorių klasifikatorius

ML (angl. *Machine learning*) – mašininis mokymasis

TCN (angl. *Temporal Convolutional Network*) – laikinas konvoliucinis tinklas

STN (angl. *Spatial Transformer Network*) – erdvinio transformatoriaus tinklas

CNN (angl. *Convolutional Neural Network*) – konvoliucinis neuroninis tinklas

DNN (angl. *Deep Neural Network*) – gilusis neuroninis tinklas

FoR (angl. *Fake or Real*) – netikras arba tikras

TSSDNet (angl. *Time-domain Synthetic Speech Detection Net*) – laiko srities sintetinės kalbos aptikimo tinklas

FLAC (angl. *Free Lossless Audio Codec*) – nemokamas garso be nuostolių kodekas

EER (angl. *Equal Error Rate*) – vienodas klaidų dažnis

TPR (angl. *True Positive Rate*) – tikras teigiamas rodiklis

FPR (angl. *False Positive Rate*) – klaidingas teigiamas rodiklis

WCE (angl. *Weighted Cross-Entropy*) – svertinė kryžminė entropija

MFCC (angl. *Mel-frequency cepstral coefficients*) – Mel-dažnio kepstraliniai koeficientai

LFCC (angl. *Linear Frequency cepstral coefficients*) – tiesinio dažnio kepstraliniai koeficientai

CQT (angl. *Constant-Q tranform*) – konstanta Q transformacijos

SVM (angl. *Support Vector Machine*) – atraminių vektorių klasifikatorius

LGBM (angl. *Light Gradient Boosting Machine*) – šviesos gradiento didinimo klasifikatorius

XGBoost (angl. *eXtreme Gradient Boosting*) – ekstremalus gradiento didinimas

KNN (angl. *K-Nearest Neighbors*) – K-arčiausių kaimynų

RF (angl. *Random Forest*) – atsitiktinis miškas

Įvadas

Kalbos sintezė yra dirbtinis žmogaus kalbos sukūrimas. Šiam tikslui naudojama kompiuterinė sistema, vadinama kalbos sintezatoriumi, kuri gali būti įdiegta programinės ar techninės įrangos produktuose. Teksto į kalbą (TTS) sistema įprastos kalbos tekstą paverčia sintetine kalba. Kitos sistemos pateikia simbolines kalbines reprezentacijas, tokias kaip fonetinės transkripcijos į kalbą. Atvirkštinis procesas yra kalbos atpažinimas. Kalbos atpažinimas – tai tarpdisciplininis informatikos ir kompiuterinės lingvistikos pokrypis, kuriame kuriamos metodikos ir technologijos, leidžiančios kompiuteriu atpažinti ir paversti sakytinę kalbą į tekstą. Jis taip pat žinomas kaip automatinis kalbos atpažinimas (ASR), kompiuterio kalbos atpažinimas arba kalbos į tekstą (STT). Taip apt apima kompiuterių mokslo, lingvistikos ir kompiuterių inžinerijos sričių žinias ir tyrimus.

Sintetinis balsas – tai terminas, kuris atitinka tai, kas paprastai vadinama balso klastote. Dažnai jis pakaitomis vartojamas ir su balso klonavimu. Tačiau paprasčiau tariant, sintetinis balsas yra kompiuteriu sukurta kalba, dar vadinama kalbos sinteze, ir paprastai pasiekiama naudojant dirbtinį intelektą (DI) ir gilųjį mokymąsi.

Įvairiose srityse, tokiose kaip bankininkystė, telekomunikacijos ir autentifikavimo sistemos, balso įrašai naudojami tapatybei patvirtinti. Gebėjimas atskirti sintetinius ir nesintetinius balsus padeda išvengti sukčiavimo ir neteisėtos prieigos. Sintetiniai balsai gali būti generuojami siekiant apsimesti asmenimis, todėl labai svarbu aptikti tokius bandymus. Žiniasklaidoje ir žurnalistikoje balso įrašų autentiškumo patikrinimas yra labai svarbus siekiant užtikrinti naujienų pranešimų, interviu ir kito garso turinio patikimumą. Galimybė pažymėti sintetinius balsus padeda išlaikyti garso kanalais skleidžiamos informacijos vientisumą. Vis dažniau naudojami balso padėjėjai ir pokalbio agentai, todėl vartotojai pasitiki šiomis sistemomis, kad pateiktų tikslią ir patikimą informaciją. Sintetinių balsų aptikimas gali padėti vartotojams suprasti, ar atsakymą generuoja žmogus, ar dirbtinio intelekto sistema, o tai gali turėti įtakos jų pasitikėjimui asistentu. Turinio kūrėjams ir transliuotojams žinojimas, ar balso įrašas yra sintetinis, ar natūralus, gali turėti įtakos garso išteklių pasirinkimui ir naudojimui. Jie gali priimti pagrįstus sprendimus dėl sintetinių balsų naudojimo pasakojimui, įgarsinimui ar veikėjų dialogams.

Kai kuriose jurisdikcijose reglamentai ar pramonės standartai įpareigoja atskleisti sintetinio balso naudojimą tam tikrose situacijose. Pavyzdžiui, gali būti reikalaujama, kad automatizuotos klientų aptarnavimo sistemos informuotų skambinančius, jei jie bendrauja su sintetiniu balsu, o ne su žmogaus operatoriumi. Plačiai paplitusios sintetinės balso technologijos kelia etinių problemų, susijusių su apsimetinėjimu, privatumu, sutikimu ir galimu piktnaudžiavimu DI sukurtu turiniu. Sintetinių balsų nustatymas prisideda prie diskusijų

apie atsakingą DI naudojmą. Paprastai kalbant, tikslus sintetinių ir nesintetinių balso įrašų identifikavimas padeda išlaikyti pasitikėjimą, vientisumą ir saugumą įvairiose programose, susijusiose su balso sąveika ir turiniu.

Tyrimo objektas – dirbtiniu intelektu paremti sintetinių ir nesintentinių balso įrašų identifikavimo metodai.

Darbo tikslas: pagerinti sintetinių ir nesintentinių balso įrašų identifikavimo metodų efektyvumą taikant dirbtinio intelekto metodus.

Darbo uždaviniai:

1. Apžvelgti dirbtinio intelekto identifikavimo metodus, duomenų rinkinius, sintetinių balso įrašų kūrimo metodus atliekant literatūros analizę;
2. Pasiūlyti metodo patobulinimą sintetinių ir nesintetinių balso įrašų identifikavimui taikant dirbtinį intelektą;
3. Sukurti duomenų rinkinį, kuris panaudojamas dirbtiniam intelektui identifikuoti sintetinius ir nesintetinius balso įrašus;
4. Įgyvendinti pasiūlyto metodo patobulinimo dirbtinio intelekto pritaikymą sintetiniams ir nesintetiniams balso įrašams identifikuoti ir įvertinti eksperimentiškai.

Temos naujumas – didėjant poreikiui identifikuoti sintetinius balso įrašus tam tikri iššūkiai vis tiek išlieka. Pažangūs gilaus mokymosi metodai, tokie kaip „*WaveNet*“, „*Tacotron*“, GAN pagrįsti metodai paskatino sintetinės kalbos generavimą, kuri beveik nesiskiria nuo žmogaus kalbos. Tai kelia iššūkių identifikavimo sistemoms, reikalaujančių sudėtingų metodų, kurie leistų identifikuoti sintetinę kalbą. Realaus laiko programos išplėtė galimybes generuoti aukštos kokybės sintetinę kalbą realiu laiku, įskaitant virtualius asistentus, interaktyvias balso atsako sistemas, todėl sintetinės kalbos identifikavimas tapo dar svarbesnis. Didžioji dalis tyrimų yra atlikta su duomenų rinkiniais anglų kalba, lietuvių kalba duomenų rinkinių praktiškai neegzistuoja arba nėra viešai prieinami.

Temos aktualumas – tobulėjant garso manipuliacijų technologijai sintetiniai balso įrašai gali būti piktybiškai naudojami apsimetinėti tam tikrais asmenimis, skleisti dezinformaciją arba vykdyti nesažiningą veiklą, todėl tai padidino patikimų identifikavimo metodų poreikį. Daugelis saugos sistemų naudoja balso biometrinius duomenis naudotojo autentifikavimui. Norint išvengti neteisėtos prieigos ir užtikrinti šių sistemų vientisumą, labai svarbu nustatyti sintetinio balsus. Identifikavimo metodų kūrimas įvairioms kalboms, įskaitant mažiau išteklių turinčias, tokias kaip lietuvių, padeda užtikrinti identifikavimo sistemų patikimumą su įvairiomis kalbomis.

Tyrimo metodika – literatūros analizėje atlikti susijusių su magistrinio darbo tema mokslinių tyrimų analizę. Eksperimentiškai įvertinti identifikavimo metodų efektyvumą ir palyginti.

Mokslinio darbo vertė – įgyvendinus identifikavimo metodų patobulinimus ir surinkus savo duomenų rinkinį lietuvių kalba eksperimentiškai įvertinti šių metodų efektyvumą identifikuoti sintetinius ir nesintetinius balso įrašus. Tai parodys, ar patobulinti metodai sugebės pagerinti rezultatus lyginant su kitais metodais, taip pat prisidės prie jų pritaikymo realiuose sistemuose.

Darbo aprobacija – pagrindiniai darbo rezultatai pagal magistro darbo tema pristatyti pranešimuose, skaitytuose:

- 2024 m. balandžio 3 d., VILNIUS TECH (Vilniaus Gedimino technikos universitete) vykusioje 27-oji JM konferencija „Mokslas – Lietuvos ateitis“, sekcija „Informacinių technologijų sauga“.

Pažymėjimas už temos pristatymą įkeltas 1 priede (žr. 1 priedą).

1. Sintetinių balso įrašų identifikavimas

Šiame skyriuje apibrėžiami sintetinių balso įrašų generavimo būdai, problematika, aprašomas jų principas ir identifikavimas. Apžvelgiami dirbtinio intelekto metodai naudojami identifikuoti sintetiniams ir nesintetiniams balso įrašams, pateikiamas jų palyginimas, peržiūrimi taip pat ir duomenų rinkiniai, naudojami sintetinių balso įrašų identifikavimo tyrimuose.

1.1. Sintetinių balso įrašų generavimo ir jų identifikavimo problematika

Naujausi pasiekimai dirbtinio intelekto sintezės srityje kaip kalbos sintezės ir balso klonavimo algoritmai parodė, kad galima sukurti tikroviškus netikrus balsus, kurie beveik nesiskiria nuo tikros žmogaus kalbos. Sintetinės kalbos generavimas yra problema, kuri yra tyrinėjama daugelį metų ir sprendžiama keliais būdais. Dėl šios priežasties literatūroje pateikiama daug gerų rezultatų pasiekiančių metodų ir nėra vieno unikalaus sintetinių balso įrašų generavimo ir identifikavimo būdo.

Šie algoritmai gali sukurti sintetinę kalbą, kuri skamba kaip tikro žmogaus kalba, remiantis tekstu ar kalbėtojo kalbos įrašu, ir pasiekti labai įtikinamų rezultatų. Sintetinis balsas yra plačiai pritaikytas įvairioms programoms kurti, pavyzdžiui, automatizuotam dubliavimui televizijoje ir kine, dirbtinio intelekto asistentams, asmeniniai sintetiniai balsai gali būti panaudojami žmonėms turintiems balso negalią. Tačiau tapo ir didele grėsme, nes yra naudojama kenkėjiškiems tikslams, pavyzdžiui, politinei naudai, netikroms naujienoms skleisti, sukčiavimams. Sintetinio garso turinio derinimas su vaizdine manipuliacija gali padaryti vaizdo įrašus žymiai įtikinamesnius ir padidinti jų poveikį. Nepaisant didelės pažangos, sintetiniams kalbos įrašams vis dar trūksta kai kurių balso kokybės aspektų būdingų konkrečiam asmeniui, pavyzdžiui išraiškingumo, šiurkštumo, dusulio, streso ir emocijų.

Dauguma esamų identifikavimo metodų yra grįsti ranka koduotų (angl. *hand-coded*) požymių panaudojimu sintetinei kalba aptikti. Kai kuriuose darbuose naudojami ištisiniai (angl. *end-to-end*) modeliai, o kiti mokymui naudoja ir ranka koduotus ir giliuosius požymius (angl. *deep features*) sintetintai kalbai aptikti. Galiausiai, papildomas apribojimas esamų metodų yra identifikavimo našumas esant ribotiems apmokymų duomenis ir skaičiavimo ištekliams (Masood et al., 2022).

Sukurta nemažai mašininio mokymosi ir gilaus mokymosi modelių, kurie naudoja skirtingas strategijas netikram garsui identifikuoti. Sintetinės kalbos identifikavimas yra tapęs aktyvia mokslinių tyrimų sritimi, kurioje kuriama ir tobulinami pažangūs metodai ir technikos. Tačiau, esant tokiai pažangai, dabartiniai metodai vis tiek susiduria su sunkumais, todėl yra būtina atlikti tolesnius tyrimus, kad būtų galima suprasti, kuria kryptimi sintetinės kalbos identifikavimą reikia toliau tobulinti.

1.2. Sintetinių balso įrašų generavimo tipai

Svarbiausios kalbos sintezės sistemos savybės yra natūralumas ir suprantamumas. Natūralumas apibūdina, ar išvestis skamba kaip žmogaus kalba, o suprantamumas – kaip lengvai išvestis suprantama. Sintetiniai balso generavimo metodai apima dirbtinių kalbos signalų, imituojančių žmogaus kalbą, kūrimą, naudojant įvairius metodus.

Sintetinės kalbos generavimas yra problema, kuri yra daug metų studijuojama ir todėl egzistuoja įvairūs požiūriai. Dėl šios priežasties literatūroje yra daug gerų rezultatų pasiekiančių metodų ir nėra vieno unikalaus būdo sukurti sintetinį balso įrašą (Borrelli et al., 2021).

Kaip ir bet kurioje mokslo srityje, tiriant kalbą pirmiausia buvo siekiama nustatyti svarbius faktus apie sistemą. Šiuos „faktus“ įkūnija dirbtiniai kalbėtojai ir kalbos sintezatoriai, tačiau jie paprastai fiksuoja tik esminius kalbos aspektus. Dėl to sintezė dažnai pateikiama kaip garsinė karikatūra, kuri gali būti suvokiama kaip nenatūralus, o kartais ir juokingas norimo sakinio ar kalbos garso perteikimas (Story, 2019).

Tai pat svarbu paminėti garso gilių klastočių technologiją, kuri per pastaruosius penkerius metus tapo svarbia tema. Naujoji technologija užleido vietą naujam, kūrybiškam sprendimui, kuris turi įtakos daugeliui pramonės šakų. Tai taip pat sukėlė daug etinių klausimų, kuriuos daugiausia lėmė netinkamas technologijų naudojimas ir neigiama spauda.

Garso giliosios klastotės gali būti apibrėžiamos kaip kalbos signalas, kuris buvo pakeistas arba sukurtas pasinaudojant giliu neuroniniu tinklu, kuris vėliau gali būti panaudojamas tyčia ar netyčia apgauti klausytojui (Müller et al., 2021).

Pačios garso giliosios klastotės, neseniai pavadintos garso manipuliacijomis, tapo plačiai prieinamos naudojant paprastus mobiliuosius įrenginius ar asmeninius kompiuterius, todėl visame pasaulyje kilo didelis susirūpinimas dėl kibernetinio saugumo ir šių technologijų šalutinio poveikio. Dėl šių priežaščių tapo privaloma autentifikuoti bet kokį paskirstytą garso įrašą, kad būtų galima išvengti dezinformacijos skleidimo (Almutairi & Elgibreen, 2022).

Naujausi algoritmai gali generuoti sintetinę kalbą, kuri skamba kaip tikslinio kalbėtojo, remiantis tekstu arba tikslinio kalbėtojo kalbos pavyzdžiais, su labai įtikinamais rezultatais. Naujausia pažanga dirbtinio intelekto sintezuotų netikrų balsų srityje taip pat padeda sukurti itin tikroviškus suklastotus vaizdo įrašus. Sintetinio garso turinio derinimas su vaizdine manipuliacija gali padaryti padirbtus vaizdo įrašus žymiai įtikinamesnius ir padidinti jų poveikį. Nepaisant pažangos, sintetiniams balso įrašams vis dar trūksta kai kurių balso kokybės aspektų, būdingų žmogui (Masood et al., 2022).

Teksto keitimo į kalbą (TTS) sistemos tikslas yra sintezuoti suprantamą ir natūralią kalbą iš teksto, o pastaraisiais metais dėl gilaus mokymosi technologijų plėtros padaryta sparti pažanga. Neuroniniais tinklais pagrįsta TTS vystėsi nuo CNN ir RNN modelių iki transformatoriais pagrįstų modelių, nuo bazinių generatyvinių modelių (autoregresinių) iki galingesnių modelių, toki kaip variaciniai autoenkoderiai, nuo kaskadinių akustinių modelių/vokoderių iki visiškai ištisinių (angl. *end-to-end*) modelių (Tan et al., 2022).

Saudo Arabijos, Karaliaus Saudo universiteto mokslininkai Z. Almutairi ir H. Elgibreen (2022) remiantis egzistuojančia literatūra išskyrė sintetiniu būdu paremta (angl. *Synthetic-based or Text-To-Speech (TTS)*) garso manipuliacijų generavimo tipą, kuris gali būti panaudojamas sintetinio balso generavimui ir tikslinio kalbėtojo balso manipuliacijų kūrimui.

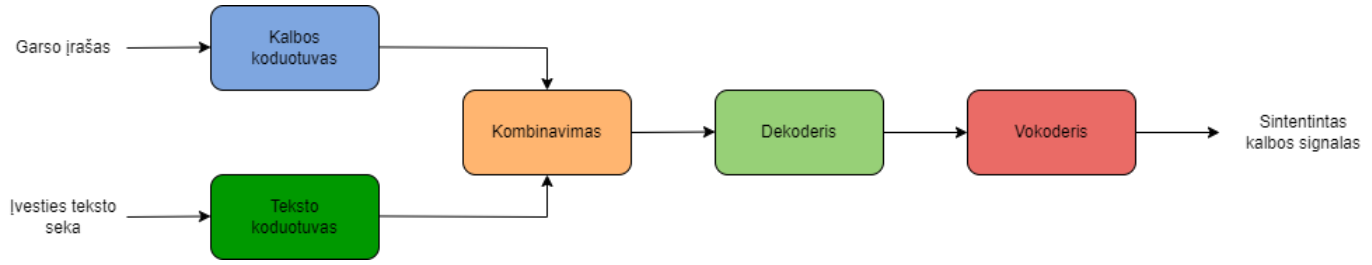
1.2.1. Sintetiniu būdu paremtas arba tekstas į kalbą generavimo būdas

Sintetinės arba tekstas į kalbą manipuliacijos tekstą transformuoja į priimtiną naturialią kalbą realiuoju laiku. Susideda iš trijų modulių: teksto, analizės modulio, akustinio modelio ir vokoderio. Norint sukurti tokio tipo garso įrašą reikia atlikti du esminius veiksmus. Pirmiausia, reikia surinkti tvarkingą ir struktūrizuotą, neapdorotą garso signalą su teksto stenograma. Antra, modulis turi būti apmokytas naudojant sintetinio garso generavimui surinktus duomenis. Siūlomi „*Tactoran 2*“, „*Deep Voice 3*“ ir „*FastSpeech 2*“ yra gerai žinomi modelių generavimo būdai ir gali sukurti aukštos kokybės naturaliai skambantį garsą. „*Tactoran 2*“ sukuria Mel spektrogramas su modifikuotu „*WaveNet*“ vokoderiu. *„Fast2Speech“* generuoja aukštos kokybės rezultatus per greičiausią apmokymo laiką. *„Deep Voice 3“* yra neuroninis teksto į kalbą modelis.

„*Deep Voice 3*“ karkasas susideda iš trijų dalių (Khanjani et al., n.d.):

- Kodavimo įrenginio. Liekamieji konvoliucijos sluoksniai naudojami tekstui koduoti į laiko žingsnio rakto ir reikšmės vektorius.
- Dekoderio. Dekoderis naudojamas Mel skalės spektrogramoms numatyti.
- Keitiklio. Visiškai konvoliucinis papildomo apdorojimo tinklas. Remdamasis pasirinktomis vokoderio ir dekoderio paslėptomis būsenomis, jis numato vokoderio parametrus.

Taikant šį sintetinį metodą, stenogramos tekstas su kalbėtojo balsu įtraukiamas į generavimo modelį. Tuomet teksto analizės modulis apdoroja gaunama tekstą ir paverčia jį lingvistiniu. Tada akustinis modulis ištraukia kalbėtojo parametrus iš duomenų rinkinio, priklausomai nuo kalbos ypatybių, sugeneruotų iš teksto analizės modulio. Galiausiai, vokoderis išmoksta kurti kalbos bangų formas.



1.1 pav. Tesktas į kalbą sistemų eigos diagrama. Parengta pagal (Masood et al., 2022)

Edinburgo ir Glazgo universiteto mokslininkai M. Aylett, A. Vinciarelli ir M. Wester (2017) savo tyrime „Kalbos sintezė dirbtinei asmenybei sukurti“ išskyrė du dominuojančius generavimo požiūrius:

- Vieneto atranka (angl. *Unit Selection*).
- Parametrinė sintezė (angl. *Parametric synthesis*).

1.2.2. Vieneto atranka paremtas generavimo būdas

Atrenkant vienetus, užuot modifikavus kalbą, kad būtų sukurtos skirtingos balso savybės, taikomas alternatyvus metodas – įrašomos skirtingos balso savybes ir tiesiogiai naudojamos atliekant konkatenatyvinę sintezę. Šis metodas taikytas difonų sintezei ir išplėstas vienetų atrankai sistemoje „CereVoice“, kurioje vienetų atrankai naudojami iš anksto įrašyti balso kokybės subkorpusai. Tai skiriasi nuo kitų vienetų atrankos metodų, kuriuose vietoj to buvo nagrinėjamas konkrečių emocijų naudojimas. Dėmesio sutelkimas į balso kokybę, o ne į konkrečias emocijas, leidžia derinti skaitmeninio signalų apdorojimo (angl. *Digital Signal Processing*) metodus ir padalinių atranką sukurti įvairesnį ir subtilesnį duomenų rinkinį (Aylett et al., 2020).

Apibendrinant vienetų atrankos sintezė sujungia konkatenatyvinės ir parametrinės sintezės elementus. Užuot tiesiogiai sujungę iš anksto įrašytus segmentus, ši technika atrenka ir sujungia mažesnius vienetus, fonemas ar difonus iš duomenų bazės pagal tokius kriterijus kaip perėjimų sklandumas ir natūralumas. Vienetų pasirinkimo sintezė gali sukurti labai natūraliai skambančią kalbą, tačiau tam reikia didelės ir įvairios duomenų bazės.

1.2.3. Parametrine sinteze paremtas generavimo būdas

Parametrinė sintezė generuoja kalbą pagal matematinius kalbos modelius. Šie modeliai paprastai apima kalbinių ypatybių analizę ir signalų apdorojimo metodų naudojimą kalbai sintezuoti.

Parametrinėje sintezėje kalba parametrizuojama, modifikuojama ir visiškai atkuriama naudojant vokodavimo būdus. Dabartinėse sistemose balso vokodavimo procesas yra nuostolingas, todėl pablogėja balso kokybė ir kai kuriais atvejais jaučiamas triukšmas. Be to, parametrai modeliuojami naudojant statistinius metodus arba giliuosius neuroninius tinklus. Tai paprastai sujungia kadrus su kintančiomis parametrų

reikšmėmis ir gaunama vidutinė vertė, kuri dažnai rodo mažiau variacijų, pašalindama fonetines ir prozodines detales. Dėl to kalba gali atrodyti nuobodi arba dusli. Tačiau pastaruoju metu buvo padaryta pažanga sprendžiant šias pagrindines problemas ir dabar jos tobulėja (Aylett et al., 2020).

1.2.4. Statistine parametrine sinteze paremtas generavimo būdas

Statistinis parametrinis kalbos generavimas (SPSG), kuris sujungia statistinius akustinius modelius ir balso kodavimo būdus, kad sukurtų kalbos bangų formas, buvo pagrindinis kalbos generavimo problemų sprendimo būdas. Šis metodas pirmiausia sukuria statistinius akustinius modelius, vaizduojančius išvesties akustinių savybių sąlyginės tikimybės tankio funkciją (PDF), atsižvelgiant į įvesties ypatybes. Modelio struktūra paprastai priklauso nuo užduoties, tačiau parametrai apskaičiuojami pagal mokymo duomenų bazę, kurią sudaro įvesties ir išvesties akustinių savybių poros (Ling et al., 2015).

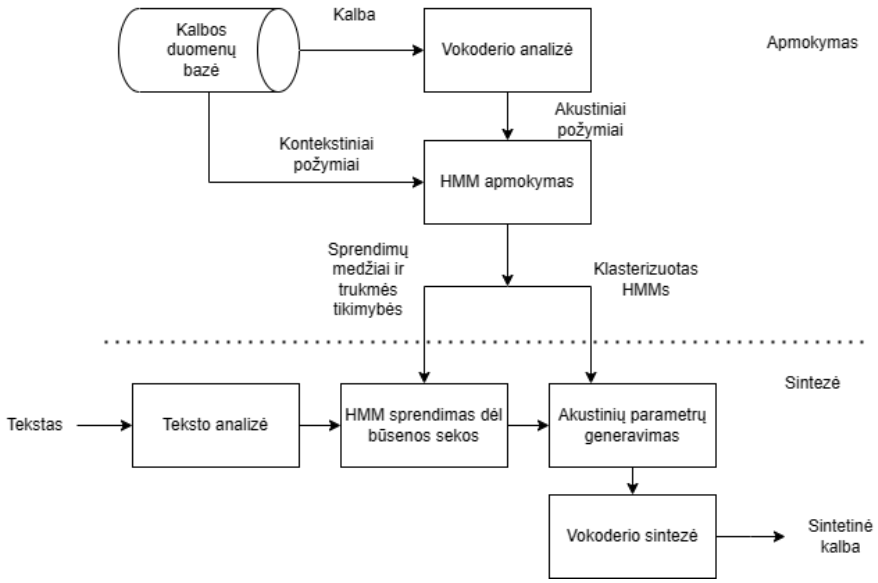
Taip pat populiariausi statistinės parametrinės kalbos generavimo modeliai pagal (Ling et al., 2015) yra:

- Paslėpti Markovo modeliai (HMM).
- Gauso mišinio modeliai (GMM).

1.2.4.1. ***HMM grįstas generavimas***

Paslėptaisiais Markovo modeliais (HMM) ir kitais statistiniais parametriniais metodais pagrįsta kalbos sintezė buvo aktuali tema kurį laiką. Naudojant šiuos metodus, kalbos sintezatoriai gali sukurti suprantamus balsus. Nepaisant pažangos, balsų kokybė sukurta naudojant statistinę parametrinę sintezę, dar nepasiekė šiuo metu dominuojančių vienetų atrankos metodų, pagal kuriuos atrenkami ir sujungiami kalbos įrašai. Dabar siekiama sukurti modelius, kurie tiksliau imituotų žmogaus balsus.

Taikant šį metodą, ryšys tarp teksto ir jo akustinės realizacijos modeliuojamas naudojant stochastinius generatyvinius akustinius modelius. Sprendimų medžių grupėmis suskirstyti, nuo konteksto priklausomi fonemų HMM su viengubais Gauso būsenų išvesties PDF yra populiariausias generatyvinis akustinis modelis, naudojamas SPSG. Šis metodas vadinamas HMM grįsta kalbos sinteze. HMM yra generatyvinis modelis, kuris generuoja stebėjimų seką naudodamas diskrečią ir paslėptą būsenų seką. Žemiau esančiame paveikslėlyje parodyta tipinės HMM grįstos kalbos sintezės sistema. Mokymo etape, akustinės kalbos savybės, įskaitant balso trakto ir vokalo šaltinio parametrai išgaunami iš bangų formų, esančių duomenų bazėje. Konteksto požymiai išvedami iš teksto segmentų ir prozodinių etikečių, atitinkančių bangų formas. Tuomet konteksto parametrų rinkinys naudojant mašininio mokymosi kriterijus. Paskutiniame etape akustiniai parametrai siunčiami į vokoderį kalbos bangų formų atkūrimui. (Ling et al., 2015). 1.2 pav. pateikta HMM grįstos sistemos blokinė diagrama.



1.2 pav. Tipinės HMM pagrįstos kalbos sintezės sistemos blokinė schema. Parengta pagal (Ling et al., 2015)

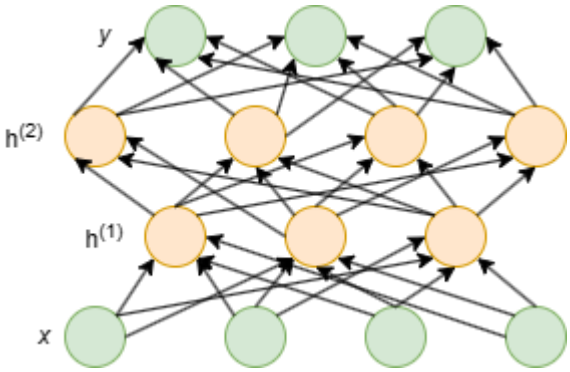
1.2.4.2. *GMM grįstas generavimas*

Gauso mišinio modeliu paremtas generavimas yra laikomas balso konertavimu, o ne kalbos sinteze kaip HMM generavimas. Šiuo atveju GMM paremtas balso konvertavimo tikslas – pakeisti įvesties kalbos nelingvistinę informaciją, tai gali būti kalbėtojo charakteristikos, o lingvistinę informaciją išlaikyti nepakitusią. Skirtingai nuo lingvistinių požymių, kurie naudojami kaip kalbos sintezės įvesties duomenys, balso konvertavimo įvesties požymiai paprastai yra ištisinės akustinės šaltinio balso reprezentacijos.

1.2.5. Giliu neuroniniu tinklu (DNN) paremtas generavimo būdas

Giliu neuroninio tinklo (DNN) pagrįsta kalbos sintezė, dar žinoma kaip neuroninių bangų formos sintezė, yra modernus būdas generuoti sintetinę kalbą naudojant gilaus mokymosi metodus. DNN pagrįsta kalbos sintezė pastaraisiais metais padarė didelę pažangą ir yra plačiai naudojama įvairiose programose, įskaitant virtualius asistentus, garso knygas ir pritaikymo neįgaliesiems įrankius.

DNN yra teisioginio sklidimo (angl. *feed-forward*) dirbtinis neuronų tinklas, turintis daugiau nei vieną paslėptą sluoksnį tarp įvesties ir išvesties sluoksnių. Kiekviename paslėptajame sluoksnyje kiekvienas paslėptasis vienetas paprastai atvaizduoja savo įėjimų iš žemiau esančio sluoksnio svertinę sumą į deterministinę reikšmę, naudojant netiesinę aktyvavimo funkciją ir perduoda ją į aukštesnį sluoksnį. DNN gali būti labai naudingas sudėtingų ir netiesinių modelių ryšiui tarp įėjimų ir išėjimų, tačiau yra sunku apmokyti tokį tinklą, su daug paslėptų sluoksnių, nes didelis jų kiekis sukelia nykstančio gradiento problemą. Taip pat prižiūrimas DNN mokymas gali sukelti per didelį prisitaikymą (angl. *overfitting*) (Ling et al., 2015). 1.3 pav. pateikta dviejų paslėptų sluoksnių DNN modelio reprezentacija.



1.3 pav. Dviejų paslėptų sluoksnių DNN modelio reprezentacija. Parengta pagal (Ling et al., 2015)

1.3. Sintetinių ir nesintetinų balso įrašų identifikavimo metodai

Egzistuoja įvairios priemonės ir metodai, kuriais galima sukurti netikrą garso įrašą, todėl pastaruoju metu daug dėmesio skiriama sintetinų balso įrašų identifikavimui skirtingoms kalboms. Apskritai dabartinius metodus galima suskirstyti į du pagrindinius tipus: mašininio mokymosi ir gilaus mokymosi metodus.

Iš mašininio mokymosi (ML) algoritmų galima būtų išskirti tiesinio atraminių vektorių klasifikavimo (Q-SVM) metodą. Mokslininkai pasiūlė šį metodą, kad galima būtų atskirti sintetinį garsą nuo naturalaus žmogaus balso. Naudojant šį modelį su dvejetainiu klasifikavimu, autoriai suskirstė garsą į dvi klases – žmogaus ir dirbtinio intelekto sukurtą. Šis metodas buvo lyginamas su kitais ML metodais, pavyzdžiui, tiesiniu diskriminantu, kvadratiniu diskriminatu, linijiniu SVM, svertiniu KNN, medžiu ansambliu ir logistine regresija. Jie nustatė, kad Q-SVM pranoko kitus klasikinius metodus, teisingo klasifikavimo rodiklis buvo 97,56 %. Taip pat šiam eksperimentui buvo panaudotas savo sukurtas duomenų rinkinys, šiam rinkiniui surinkti įrašai buvo paskirstyti atsitiktine tvarka į 622 mokymo ir 82 testavimo duomenis. Autoriai pasirenka neapdoroti duomenys, kad pritemptų eksperimentą prie realių salygų ir taip pat norėdami patikrinti, kiek jų funkcijos gali būti veiksmingos naudojant neapdorotus duomenis. Eksperimenti buvo parinkti du scenarijai. Pirmame kalbos įrašai klasifikuojami į dvi klases, žmogaus ir dirbtinio intelekto sintezuotų, šis klasifikavimas pavadinamas dvejetainiu klasifikavimu. Antrajame scenarijuje dirbtinio intelekto sintezuoti įrašai yra klasifikuojami atskirai, todėl iš viso gaunamos keturios klasės (žmonių, *„Natural Reader“*, *„Spik.AI“*, *„Replica“*). Taip pat tyrime atliktas 5 dalių kryžminis tikrinimas. Tyrimo rezultatai parodė, kad dvejetainio klasifikavimo rezultatai stipresni nei daugiaklasio klasifikavimo, nes daugiaklasiame klasifikavime buvo daug klaidingų teigiamų (angl. *false-positive*) rezultatų (Singh & Singh, 2020).

Hanyang universiteto Seule mokslininkai bandė aptikti sintetines balso giliąsias klastotes naudodami aiškinamuosius dirbtinio intelekto (XAI) metodus, kad galima būtų tinkamai perteikti jų veikimą, kurie

pagrinde yra skirti vaizdų klasifikavimui, tyrime daugiausia dėmesio skiriama intuityvumui ir aukštesnio lygio intepretavimui. Garso duomenų požymiams išgauti buvo naudojama spektrograma

Eksperimentui panaudoti konvoliucinio neuroninio tinklo (CNN), *„CNN-LSTM“* ir *„CNN-LSTM-perm“* modeliai. CNN modelio atveju dėl XAI metodų tendencijos rodyti priklausomybę nuo energijos formantų, didžioji dalis originalaus balso buvo atkurta iš atributinio balo, originalų balsą galima buvo atpažinti ir iš atkurto balso. *„Deep Taylor“* ir LRP priskyrimo balai *„CNN-LSTM“* pagrindu sukurtame modelyje rodė mažesnį jautrumą energijai, prarasdami formantų išryškinimus. Todėl transkripcija buvo neatpažįstama pagal perkomponuotą balsą. Eksperimente buvo panaudoti du duomenų rinkiniai su išskirtinėmis charakteristikomis, tai *„ASVspoof“* ir *„LJSpeech“*. Didžiausias tikslumas buvo pasiektas su *„CNN-LSTM-perm“* modeliu, testuojant su *„ASV Multiple TTS“* ir *„LJSpeech“* duomenų rinkiniais, su abiem buvo pasiektas 99,97% tikslumas, tačiau pagrindinis šio eksperimento tikslas buvo suteikti interpretaciją, o ne fokusuotis į tikslumą (Lim et al., 2022).

Mokslininkai iš Saudo Arabijos tyrė požymiais ir vaizdais grįstus metodus, skirtus sintezuotu būdu sukurtam suklastotam garsui klasifikuoti. Požymiais grįsti metodai apima garso duomenų konvertavimą į duomenų rinkinį, sudarytą iš įvairių garso įrašų spektrinių požymių, kurie toliau yra perduodami mašininio mokymosi algoritmams, kad būtų galima atitinkamai klasifikuoti garso įrašą kaip netikrą ar tikrą. O taikant vaizdais grįstus metodus garso įrašai konvertuojami į Mel spektrogramas, kurios įvedamos į giliojo mokymosi algoritmus. Šio darbo eksperimente yra panaudoti 5 mašininio mokymosi, požymiais grįsti algoritmai (Khochare et al., 2022):

- Atraminių vektorių klasifikatorius (SVM).
- Šviesos gradiento didinimo klasifikatorius (LGBM).
- Ekstremalus gradiento didinimas (XGBoost).
- K-arčiausias kaimynas (KNN).
- Atsitiktinis miškas (RF).

Šiame darbe buvo naudojami naujesni gilaus mokymosi modeliai, kurie prieš tai daugiau buvo išbandyti vaizdų klasifikavime, kaip laikinas konvoliucinis tinklas (TCN) ir erdvinis transformatoriaus tinklas (STN), konvoliucinis neuroninis tinklas (CNN). TCN sugebėjo pasiekti daug žadančių rezultatų atskiriant sintetinius ir tikrus garso įrašus 92 % tikslumu, STN – 80 % tikslumu. TCN yra nuoseklusis tinklas, todėl gali užfiksuoti duomenis, kurie taip pat yra nuoseklūs. STN yra modulis, kurį galima įterpti į konvoliucinį neuroninį tinklą (CNN), siekiant veiksmingai padidinti erdvinį nekintamumą. Mašininio mokymosi algoritmų tikslumas buvo nuo 60 % iki 70 % (Khochare et al., 2022). Šiame tyrime duomenys buvo iš anksto apdorojami, nes duomenų rinkinio duomenys negali būti pateikiami kaip įvestis klasifikavimo modeliams. Duomenys būtina

koveruoti į tinkamą formatą prieš pradedant klasifikavimą. Kaip jau prieš tai minėta, buvo taikomi du klasifikavimo būdai, požymiais grįstas ir vaizdu grįstas.

Taip pat mokslininkai iš Alborgo universiteto pasiūlė naują balų skaičiavimo metodą, pavadintą „Žmogus Žurnalas-Tikimybė“ (HLLs), remiantis giliojo neuroninio tinklo klasifikatoriumi (DNN), kad pagerinti identifikavimo rodiklį. Tai jie palygino su klasikiniu balų skaičiavimo metodu, vadinamu „Tikimybės koeficientai“ (LLR), kuris priklauso nuo Gauso mišinio modelio (GMM). Abu metodai buvo išbandyti su „*ASVSpoof 2015*“ duomenų rinkiniu. Bandymai patvirtino, kad DNN-HLLs sukūrė geresnius aptikimo rezultatus pasiekiant 0,045 % EER reikšmę. Tyrime buvo naudojami dinaminiai požymiai, spektru grįstas kepstralinis koeficientas (SBCC) ir konstantos Q kepstralinis koeficientai (CQCC) įvertinti DNN klasifikatoriaus veikimui. (Yu et al., 2018).

Mokslininkai iš Prancūzijos naudodami „*RawNet2*“ klasifikatorių pasiekė 0,1175 t-DCF rodiklį, šis rezultatas buvo prastesnis už pradinių metodų palyginimui. Tačiau, jie fokusavo į rezultatus susijusius su A17 ataka iš *„ASVSpoof2019“* rinkinio, palyginimui bazinis klasifikatorius pasiekė 0,3524 min t-DCF rodkilį, o „*RawNet2*“ sugebėjo pasiekti 0,181 minimalų t-DCF naudodami tiesinio mastelio „*sinc*“ filtrus, kaip požymius šiam klasifikatoriui. Tinklas buvo apmokytas naudojant ADAM optimizavimo metodą, taikant mokymosi greitį 0,0001, 100 epochų ir 32 mini partijos dydį (Tak et al., 2020).

Buvo pasiūlyta giliuoju mokymu pagrįsta sintetinio garso identifikavimo sistema. 60 dimensijų linijinių filtrų bankai (LFB) yra išgaunami iš kalbos įrašų ir vėliau panaudojami apmokyti modifikuotam „*ResNet*“ modelį. Šis darbas pagerino netikro garso identifikavimo tikslumą, EER reikšmė siekė 1,81 %, tačiau jam buvo būdingos didelės skaičiavimo sąnaudos (Chen et al., 2020). Įvade mokslininkai aprašo, kad didžiausia problema yra apibendrinimo gebėjimas, todėl yra stengiamasi padidinti modelio apibendrinimo gebėjimą, tam naudojama didelio skirtumo kosinuso nuostolių funkcija (LMCL) kartu su dažniu maskavimu (FM).

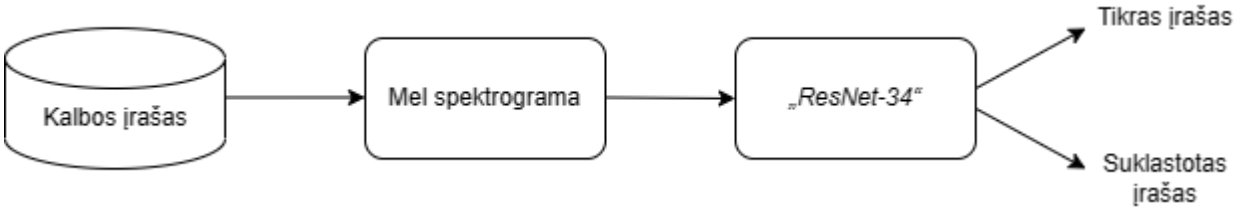
Mokslininkų grupė sukūrė klasifikavimo modelį, kurį pavadino „*Deep4Snet*“, kuris vizualizavo garso duomenų rinkinį pagal 2D CNN modelį, kad būtų galima atitinkamai klasifikuoti imitacinius ir sintetinius garso įrašus naudojant vaizdo padidinimą ir išbraukimą. „*Deep4SNet*“ pademonstravo 98,5 % tikslumą nustatant sintetinius garsus ir 99,7 % tikslumą nustatant imitacinius garso įrašus, tačiau jo našumui įtakos turėjo duomenų transformavimo procesas, eksperimente naudotas „*H-Voice*“ duomenų rinkinys. Šis pasiūlytas metodas apima apmokymo ir patvirtinimo etapus. Mokymo etape apmokymo histogramos skirstomos į originalias ir netikras. jos transformuojamos keičiant vaizdo dydį, mastelį ir horizontaliai apverčiant. Tuomet CNN apmokomas su naujais originaliais ir netikrais vaizdais. Patvrtinimo etape vaizdai transformuojami keičiant dydį ir mastelį, bet ne horizontaliai ir tada jie perduodami apmokytam modeliui, kuris klasifikuoja histogramas į originalias ir sintetines ir jų rezultatai lyginami su tikrosiomis vertėmis (Ballesteros et al., 2021).

Buvo sukurti nauji metodai sintentiniam garsui identifikuoti remiantis CNN modeliu (Subramani & Rao, 2020):

- „*EfficientCNN*“.
- „*RESEfficientCNN*“.

„*RESEfficientCNN*“ atliekant eksperimentą sugebėjo pasiekti 97,61 F1 balą, o „*EfficientCNN*“ pasiekė 94,14 F1 balą testuojant su „*ASVSpoof 2019*“ duomenų rinkiniu. „*EfficientCNN*“ modeliui kaip įvestis yra paduodama normalizuota logaritminė spektrograma, kuri apdorojama per įvesties apdorojimo bloką ir keturis konvoliucijos blokus. Išvestis po paskutinio konvoliucijos bloko perduodam į klasifikavimo bloką, kad būtų gautos klasifikatoriaus prognozės. „*RESEfficient*“ tai likutinė „*EfficientCNN*“ versija. Likutiniai ryšiai konvoliuciniuose tinkluose labai sėkmingai naudojami kompiuterinės regos srityse, siekiant pagerinti stabilumą ir tikslumą. Eksperimentui taip pat atliktas išankstinis apdorojimas, ilgos trukmės garso įrašai sutrumpinti iki 4 sekundžių, o jei garso trumpesnis, tai jis yra pailginamas. Požymių išgavimui kiekvienas garso fail‘as buvo konvertuojamas į spektrogramą naudojant 16 GHz diskretizavimo dažnį, 108 ms „*Hamming‘o*“ langą, 10 ms lango poslinkį. Galiausiai, paimamas spektrogramos reikšmių logaritmas ir Z-normalizavimas, kad sukurti kiekvieno garso įrašo įvestį.

Atsižvelgiant į apmokymo perkėlimą ir vienmodalius metodus. Buvo pasiūlyta nauja sistema, pagrįsta apmokymo perkėlimu ir „*ResNet-34*“ metodu, kuri buvo skirta aptikti sintezuotus angliškai kalbančius balsus, siūlomos sistemos supaprastinta struktūra pateikta 1.4 pav. Apmokymo perkėlimo modelis buvo iš anksto apmokytas CNN tinkle. „*ResNet-34*“ metodas buvo naudojamas nykstančio gradiento problemai spręsti, kuri visada iškyla gilaus mokymosi modeliuose. Eksperimento rezultatai parodė, kad pasiūlytas būdas pasiekė neblogų rezultatų vertinant pagal EER ir t-DCF rodiklius, atitinkamai 5,32% ir 0,1514%. Nors ir „*ResNet-34*“ išsprendė nykstančio gradiento problemą, tačiau jo apmokymas trunka ilgai dėl jo architektūros (P et al., 2020).



1.4 pav. Siūlomos sistemos supaprastinta strukūra. Parengta pagal (P et al., 2020)

Mokslininkai iš Ispanijos, Granados ir Malagos universitetų pasiūlė lengvų kovoliucinių vartų RNN (LCGRNN) (angl. *Light Convolutional Gated RNN*) modelį, kuris sudarytas iš hibridinio šviesos konvoliucinio neuroninio tinklo (LCNN) ir pasikartojančio neuroninio tinklo (RNN), skirtą apskaičiuoti giliuosius požymius ir klasifikuoti tikrai ir sintetinei kalbai. Šis modelis skaičiavimo požiūriu yra efektyvus, tačiau nėra gerai

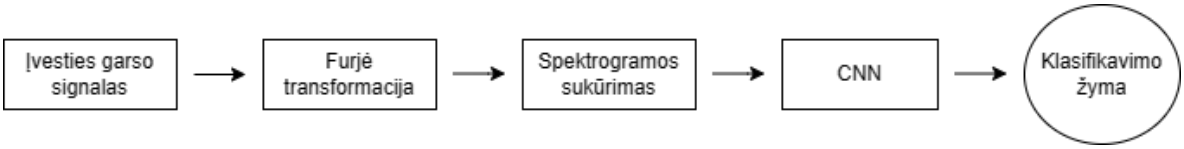
pritaikytas realaus pasaulio pavydžiams (Gomez-Alanis et al., 2019). Pasiūlyta architektūra sujungia šviesos konvoliucinių sluoksnių (angl. *light convolutional layer*) galimybes išgauti diskriminacinius požymius kadro lygmenyje su RNN. Darbe siūloma pakeisti visiškai sujungtus sluoksnius, esančius pasikartojančiose ląstelėse, LCNN sluoksniais, kad išgauti diskriminacinius požymius kadro lygmenyje, išmokti ilgalaikių priklausomybių ir integruoti giliųjų kadro lygmens požymių išskyrimą ir pasisakymo lygmens tapatybės vektorių į vieną tinklą. Eksperimente, po požymių išskyrimo testavimai buvo atlikti su trimis skirtingais klasifikatoriais: SVM, tiesine diskriminantine analize (LDA) ir jos tikimybine verisiją (PLDA). Kadangi šis modelis buvo siūlomas „*ASVspoof 2019 Challenge*“ iššūkiui, tai labiau telkiamas dėmesys „*ASVspoof 2019*“ rinkiniui, su šiuo rinkiniu buvo pasiektas 6,28% EER rodiklis su LC-GRNN + LDA technika prieš TTS ir VC technikas. Atliekant tyrimus prieš pakartojimo atakas pasiektas 2,23% EER rodiklis us LC-GRNN + PLDA technika, papildomai dar atliekant normalizavimą.

Vienas tyrimų apribojimų buvo, kad kai kur buvo naudojama daug sluoksnių ir konvoliucinių tinklų, todėl buvo sudėtinga juos valdyti. Siekiant pašalinti šį apribojimą, mokslininkai M. Shan ir T. Tsai (2020) pasiūlė derinimo metodą pagrįsta klasifikavimo modeliais: ilgalaikės trumpalaikės atminties (LSTM), dvikrypčiu LSTM ir transformatorių architektūra. Šis metodas klasifikuoja kiekvieną garso kadrą, kaip atitinkantį ar neatitinkantį, iš 50 įrašų. Tyrimo rezultatai parodė, kad dvikryptis LSTM pranko kitus modelius pasiekiant 99,7% tikslumą ir 0,43% EER rodiklį, tačiau apmokymo procesas užtruko ilgai ir duomenų rinkinys buvo labai mažas, sudarytas iš Donaldo Trumpo kalbos įrašų, iš viso buvo surinkta 50 įrašų, verta paminėti, kad pasiūlytas metodas nebuvo veiksmingas su 5 sekundžių trukmės garso įrašais, o su trumpesnės trukmės veikė efektyviau (Shan & Tsai, 2020).

Mokslininkų iš Kinijos buvo pasiūlytas savarankiškai prižiūrimas sintezuoto garso identifikavimo modelis (SSAD). Siūlmas modelis priklauso nuo daugiasluoksnių konvoliucinių blokų, skirtų išgauti požymius iš garso srauto. Jis buvo išbandytas su „*ASVspoof 2019*“ duomenų rinkiniu ir pasiekė 5,31% EER rodiklį. Lyginant su kitais metodais modelio našumas nebuvo toks geras kaip kitų. Pasiūlyta SSAD architektūra susideda iš aštuonių konvoliucinių blokų, vieno TCN sluoksnio ir netiesinės projekcijos sluoksnio. Konvoliuciniai blokai naudojami, kad atlikti neapdorotų įvesties fragmentų konvoliuciją. Naudojant TCN pasiūlytas modelis SSAD gali mokytis ilgalaikių priklausomybių veiksmingiau. Eksperimente buvo panaudoti trys klasifikatoriai: „*SENet12*“, „*LCNN-small*“, „*LCNN-big*“, su pastaruoju ir buvo pasiektas geriausias rodiklis (Jiang et al., 2020).

Tyrėjai E.R. Bartusiak ir E.J. Delp iš Purdue universiteto JAV lygino CNN modelio našumą su atsitiktinai parinktu metodu aptinkant sintetinius garso signalus. Nors CNN modelis pasiekė 85,99% tikslumą su „*ASVSpoof 2019*“ duomenų rinkiniu, jie taikė vaizdais grįstą metodą, kuriam reikėjo specialios Furjė

transformacijos, taip pat tyrėjai nustatė, kad pasiūlytas metodas nesugebėjo tesingai klasifikuoti naujų garso signalų, kurie nebuvo panaudoti apmokymui. Pasiūlytas metodas nagrinėjo garso signalus dažnių srityje spektrogramų pavidalu, supaprastinta struktūra pavaizduota 1.5 paveikslėlyje (Bartusiak & Delp, 2021).



1.5 pav. Pasiūlyto metodo supaprastinta strukūra. Parengta pagal **(Bartusiak & Delp, 2021)**

Vieni mokslininkai bandė pateikti nuolatiniu mokymusi pagrįstą metodą, skirtą sumanipuliuotos kalbos identifikavimo sistemoms pagerinti. Pasiūlytame metode buvo įvesta žinių distiliavimo nuostolių funkcija, kad būtų padidintas modelio mokymosi gebėjimas. Šis metodas skaičiavimo požiūriu efektyvus ir sugebėjo aptikti nematytas sintetines kalbas, tačiau jo našumas nebuvo įvertintas su triukšmingais pavyzdžiais (Ma et al., 2021). Eksperimente jie išgauna 60 dimensijų tiesinio dažnio kepstralinius koeficientus (LFCC), kaip įvesties požymius. Išskyrimo būdas yra toks pat, kaip ir „*ASVspoof 2019*“ oficiali bazinė sistema, naudojamas 25 ms langas, nustatomas 512 Greitosios Furjė transformacijos (FFT) binų skaičius, nustatomas 20 filtrų skaičius. Klasifikavimui panaudojamas lengvasis konvoliucinis neuroninis tinklas (LCNN). Pasiūlytas metodas sugebėjo pasiekti 9,25% EER rodiklį.

Buvo sukurtas garso požymių deskriptorius, kuris buvo pavadintas „*ELTP-LFCC*“, jis buvo pagrįstas vietiniu trigubu modeliu (ELTP) ir tiesiniai dažnio kepstraliniais koeficientais (LFCC). Klasifikavimui mokslininkai naudojo gilų dvikryptį ilgalaikės trumpalaikės atminties (DBiLSTM) tinklą, siekant identifikuoti sintetinį garsą įvairioms patalpų ir lauko aplinkos sąlygoms. Modelis išbandytas su *„ASVspoof 2019“* duomenų rinkiniu, su sintetiniu ir imituotu garsu. Po eskperimento paaiškėjo, kad modelis geriau veikė su sintetinio garso duomenu rinkiniu ir pasiekė 0,002% EER rodiklį, bet su imituotais duomenimis buvo pasiektas 33,28% EER rodiklis, su bendru šių dviejų tipų rinkinio dalimi pasiektas 0,74% EER rodiklis (Arif et al., 2021). Pasiūlytas ELTP metodas yra atsparus nevienodam triukšmui ir dinaminių modelių aptikimui ir gali užfiksuoti dinamiškai kintančius nesintetines kalbos požymius ir algoritminius artefaktus sintetiniame balse. Integruotas LFCC su pasiūlytu ELTP, kad dar labiau pagerinti identifikavimo našumą. Klasifikavimui buvo naudojama 10 dvikrypčių LSTM sluoksnių, išskirti ELTP-LFCC požymiai ir pirmasis BiLSTM sluoksnis. Pirmo BiLSTM sluoksnio išėjimai sujungiami ir perduodami kitam BiLSTM sluoksniui. Požymių vektorius iš 10-ojo sluoksnio perduodamas į visiškai sujungtą (FC) sluoksnį. FC sluoksnio išvestis perduodam į „*softmax*“ sluoksnį ir iš jo galiausiai į klasifikavimo sluoksnį, kuris priskira įvestį vienai ar kitai klasei.

Buvo pasiūlytas mokslininkų iš Milano politechnikos universiteto, Elektronikos, Informacijos ir Inžinerijos katedros SVM modelis su atsitiktinio miško (RF) algoritmu, skirtu sintetiniams balsams identifikuoti pagal naują garso požymį, kuris buvo pavadintas trumpalaikis-ilgalaikis (STLT). Eksperimente modeliams apmokyti panaudotas *„ASVspoof 2019“* duomenų rinkinys (Borrelli et al., 2021). Buvo tyriami kalbos garso įrašai siekiant nustatyti, ar jie yra tikri ar sintetiniai, todėl buvo svarstomi uždaraus ir atviro rinkinio scenarijai. Uždaro scenarijaus atveju siūlomas metodas aptinka, ar kalba yra tikra, ar sintetinė, taip pat nustatoma, koks algoritmas panaudotas sintetinei kalbai generuoti. Atviros rinkinio atveju metodas gali nustatyti, ar sintetinė kalba buvo sugeneruota naudojant algoritmą, kuris yra nežinomas. Taip pat dar palyginimui naudotas dvejetainis scenarijus, kuris tiesiog nustatydavo, ar įrašas tikras ar sintetinis. Eksperimentui panaudoti atsitiktinio miško (RF), tiesinis SVM , spindulinės bazinės funkcijos (RBF) SVM klasifikatoriai

Be to, eksperimentas atliktas apjungiant STLT požymius su dvirišamumu (angl. *bicoherence*) ir taip pat atskirai su šiais požymiais. Tyrime atskleidžiama, kad SVM klasifikatorius pranoko RF. Didžiausias tyrimo trūkumas, kad požymiai išskiriami rankiniu būdu. Taip pat sudėtinga nustatyti sintetinės kalbos generavimo alogritmus dėl jų didelės įvairovės atvirojo rinkinio scenarijaus atveju. Tyrime naudoti apdoroti įrašai, todėl neaišku, kaip jie būtų paveikti triukšmo ar perdavimo klaidų.

Mokslininkai iš Kinijos mokslų akademijos universiteto, Kibernetinio saugumo mokyklos pasiūlė sintetinės kalbos identifikavimo modelį pagal „*ResNet*“ modelį su transformatoriniu koduotuvu, kuris vadinosi „*TEResNet*“. Transformatorinis koduotuvas naudojamas, kad išgauti kontekstinius akustinius požymius, tuomet naudojamas liekanų tinklas giliosioms savybėms apdoroti ir apskaičiuoti balą, pagal kurį nustatoma, ar kalba yra sintetinė. Eksperimente naudojami „*ASVspoof 2019 LA*“ ir „*FoR-normal*“ duomenų rinkiniai, su pastaruoju pasiekiamas 3,99%, o su pirmu 5,89% EER rodikliai (Zhang et al., 2021).

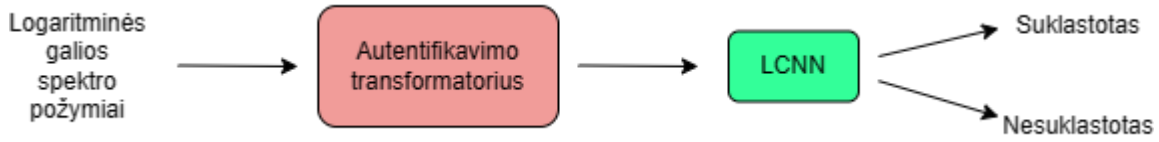
Viename tyrime buvo išbandyti keturi identifikavimo modeliai (Kawa et al., 2022):

- LCNN.
- „*XceptionNet*“.
- „*MesoInception-4*“.
- „*RawNet2*“.
- GMM.

LCNN pasiekia geriausią apibendrinimą, surinkdamas mažiausias EER per visas duomenų rinkinio dalis, atitinkamai 9,525%, 9,523% ir 2,370%. Iš visų modelių, LCNN pasižymi didžiausiu nuoseklumo balų skaičiumi mažiausiu standartiniu nuokrypiu, o tai rodo geriausią jo stabilumą. „*XceptionNet*“ pateikia antrą geriausią vidutinį EER (16,206 proc., 12,766 proc, 13,056 %), o tai rodo, kad kai kurie modeliai, paprastai

naudojami sumanipuliuoto vaizdo identifikavimui, gali būti sėkmingai taikomi ir garso srityje. Tačiau šie rezultatai yra mažiau stabilūs nei rezultatai, gauti su „*RawNet2*“. Nepaisant didesnio EER (19,443 %, 23,793 % ir 14,144 %), „*RawNet2*“ standartinio nuokrypio rodikliai yra mažesni: 1,358 %, 2,046 %, 0,891 %. priešingai nei „*XceptionNet*“ 1,476%, 1,171%, 3,224%. Tyrime taip pat naudojami MFCC ir LFCC, spektrogramų požymiai.

Buvo pasiūlytas naujas metodas susijęs su požymiu autentifikavimu, kurio metu konvoliucinis neuroninis tinklas (CNN) yra apmokomas naudojant nesintetinės kalbos požymius. Tuomet yra naudojamas autentiškumo transformatorius su lengvuoju konvoliuciniu neuroniniu tinklu (LCNN) klasifikatoriumi, „*ASVspoof 2019*“ loginės prieigos duomenų rinkinys buvo naudojamas įvertinti metodą, su šio rinkinio vertinimo dalimi buvo pasiektas 4,07 EER rodiklis. Šis pasiūlytas metodas yra patikimas su sintetinio tipo kalbos manipuliacijom, tačiau jis nėra efektyvus su balso klonavimo pasikartojimo atakomis (Wu et al., 2020). 1.6 pav. yra parodyta siūlomos sistemos blokinė schema, pagrįsta autentifikavimo transformatoriumi ir LCNN klasifikatoriumi.



1.6 pav. Siūlomos sistemos blokinė schema. Parengta pagal (Wu et al., 2020)

Taip pat mokslininkų buvo pasiūlyta „*ResNet*“ modelio „*Res2Net*“ modifikacija. Šis modelis modifikuoja „*ResNet*“ bloką, kad būtų galima naudoti kelių požymių skalę, jis suskaldo vieno bloko požymių žemėlapį į kanalų grupes ir sukuria likutinį ryšį tarp skirtingų kanalų grupių (Li et al., 2021). Eksperimento tyrimai parodė, kad modifikuotas „*Res2Net*“ modelis pademonstravo geresnius rodiklius nei „*ResNet34*“ ir „*ResNet50*“ su „*ASVspoof 2019*“ duomenų rinkinio loginė prieigos (LA) ir fizinės prieigos (PA) dalimis. Be to, integracija su išspaudimo ir iššaukimo (SE) blokas dar labiau pagerina modelio našumą, pasiektas buvo 2,502% EER rodiklis. Tyrime buvo panaudotos skirtingos akustinės savybės ir pastebėta, kad geriausi rezultatai pasiekiami su konstanta Q transformacija (CQT).

1.4. Sintetinių ir nesintetinių balso įrašų identifikavime naudojami duomenų rinkiniai

Garso duomenys – tai bet kokie skaitmeniniai duomenys, kuriais perteikiamas garsas. Tai gali būti kalba, muzika, aplinkos garsai ar bet kokio kito tipo garso signalas. Garso duomenys paprastai saugomi skaitmeniniais formatais, tokiais kaip WAV, MP3 ar AAC, ir gali būti apdorojami bei tvarkomi naudojant

įvairias programinės įrangos priemones ir metodus. Garso duomenys dažnai naudojami algoritmams, skirtiems tokioms užduotims, kaip kalbos atpažinimas, kalbėtojo identifikavimas ir emocijų nustatymas, apmokyti. Surinkti garso duomenys gali būti iš anksto apdorojami ir transformuojami į įvairius požymius, kurie vėliau įvedami į mašininio ar gilaus mokymosi modelius.

Kalbos duomenų rinkimas – tai kalbos įrašymo procesas, skirtas tolesniam naudojimui, pavyzdžiui, moksliniams tyrimams, kalbos atpažinimo mokymui ir kalbos sintezei. Duomenys gali būti renkami iš garso įrašų arba tekstinių korpusų, kuriuose yra kalbos pavyzdžių. Kalbos atpažinimo sistemoms ir taikomosioms programoms kalbos duomenų rinkiniai yra būtini norint sukurti tikslius ir patikimus modelius, kurie buvo apmokyti pagal natūralius pokalbius.

„*ASVspoof 2019*“ iššūkis tęsia tris specialiąsias sesijas, skirtas klastojimui ir atsakomosioms priemonėms, kurios buvo surengtos 2013, 2015, 2017 m. Pirmoji 2013 m. buvo skirta daugiau informuotumui didinti. 2015 m. leidimas apėmė pirmąjį iššūkį šia tema, prie kurio buvo pridėti bendrai apibrėžti vertinimo duomenys, metrikos, protokolai. Šių metų „*ASVspoof*“ užduotis buvo sukurti atsakomąsias priemones galinčias askirti tikrą kalbą nuo suklastotos ar sintetinės kalbos. „*ASVspoof 2017*“ užduotyje daugiausia dėmesio skirta kurti atsakomąsias priemones, skirtas aptikti pakartojimo atakas (*angl. replay spoofing*). „*ASVspoof2019*“ iššūkis praplėtė ankstesnius keliomis kryptimis, pirmiausia daugiausia dėmesio skiriamas visų trijų pagrindinių atakų tipams, t. y. kylančioms iš TTS, VC ir pakartojimo suklastojimo atakų. „*ASVspoof2019*“ loginės prieigos duomenų bazė yra pagrįsta standartine kelių garsiakalbių kalbos sintezės duomenų baze, vadinama „*VCTK2*“. Tikroji kalba surinkta iš 107 kalbėtojų ir be reikšmingo kanalo ar foninio triukšmo poveikio. Sintetinė kalba generuojama iš tikrų duomenų, naudojant keletą skirtingų algoritmų. Visas duomenų rinkinys suskirstytas į tris pogrupius: pirmasis skirtas mokymui, antrasis – vystymui, o trečiasis – vertinimui. Pagal fizinės prieigos (FA) sąlygą atsižvelgiama į jutiklio lygiu vykdomas atakas. Tai reiškia, kad ir tikri, ir sintetinei signalai sklinda per fizinę erdvę prieš gavimą. Todėl šiame scenarijuje atakos vadinamos pakartotinėmis atakomis, kai nesintetinis prieigos bandymo įrašas pirmiausia užfiksuojamas, tikėtina, slapta, o tik po to atkuriamas į mikrofoną (Wang et al., 2019).

Pirmasis rinkinys su kuriuo mokslininkai pasiekia 97,56 % tikslumą naudodami Q-SVM metodą yra surinktas iš kitų pavyzdžių ir iš įrašytų įrašų. Kalbos pavyzdžiai rinkti iš įvairių šaltinių, iš viso rinkinį sudarė 704 įrašai, taip pat pasirinkta duomenų neapdoroti.

Žmogaus kalbos įrašų iš viso 250, iš kurių 110 paimta iš *„Kaggle“* duomenų saugyklos, o 140 įrašyti naudojant mikrofoną. Siekiant įvairovės, paimta ir vyrų ir moterų kalbos pavyzdžių. Dirbtinio intelekto sintezuotai kalbai surinkti pavyzdžiai iš trijų šaltinių: „*Natural Reader*“, „*Spik.AI*“ ir „*Replica*“. Iš „*Natural Reader*“ paimti 79 įrašai, iš *„Spik.AI*“ 230 ir iš „*Replica*“ 126. Įrašai paimti iš šių šaltinių yra grįsti teksto

keitimo į kalbą būdu. Kiekvienas kalbos signalas buvo sutrumpinamas iki maždaug 5 sek. Taip pat visi kalbos pavyzdžiai buvo konvertuoti į mono tipą (Singh & Singh, 2020).

Tyrime, kuriame buvo naudojami aiškinamieji dirbtinio intelekto metodai (XAI), mokslininkai eksperimentui pasiteklia du duomenų rinkinius. Pirmasis rinkinys buvo „*ASVspoof 2021*“, jį sudarė 2580 nemodifikuotų įrašų, surinktų iš 107 kalbėtojų ir atitinkamai 22 800 sintezuotų kalbos duomenų, sukurtų naudojant 19 sintezatorių. Kitas duomenų rinkinys buvo „*LJSpeech*“, kurį sudaro 13 100 nemodifikuotų ir sintetinių duomenų, iš šio rinkinio eksperimentui buvo pasirinkta iš viso 8076 modifikuoti ir nemodifikuoti kalbos įrašai.

Mokslininkai iš Saudo Arabijos eksperimentui panaudojo „*Fake or Real*“ (FoR) duomenų rinkinį, sudaryta iš 195 000 tikrų žmonių pasisakymų ir dirbtinai sugeneruotų. Dirbtinai sugeneruoti įrašai sukurti taikant naujausius kalbos sintezės metodus. Tikri pasisakymai sudaryti iš įrašų, kuriems įrašyti naudoti skirtingi mikrofonai, taip pat yra skirtingų akcentų su skirtingų lyčių pasisakymais. Egzistuoja keturios šio duomenų rinkinio versijos:

- „*For-original*“.
- „*For-norm*“.
- „*For-2second*“.
- „*For-rerec*“.

„*For-original*“ duomenų rinkinys sudarytas iš įvairių kalbos garso įrašų. Šis rinkinys turi nevienodą skirtingų lyčių ir klasių įrašų skaičių. Jame yra apie 195 541 garso įrašas iš įvairių šaltinių ir jis buvo paskelbtas, kad neapdorotus duomenis galima būtų naudoti su įvairiomis išankstinio apdorojimo (angl. *pre-processing*) technikomis. „*For-norm*“ panašus į „*For-original*“ ir turi vienodą skirtingų lyčių ir klasių įrašų skaičių. Šis rinkinys yra normalizuotas pagal garsumą, imties dažnį ir kelis kanalus, sudaro maždaug 69 400 įrašai. „*For-2second*“ duomenų rinkinys sudarytas iš įrašų sutrumpintų iki maždaug 2 sekundžių, o toliau savybėmis nesiskiria nuo „*For-norm*“, turi 17 870 įrašų. „*For-rerec*“ rinkinį sudaro pakartotinai įrašyti įrašai iš „*For-2second*“ rinkinio.

Mokslininkų iš Alborgo universitete eksperimente yra naudoajmas „*ASVSpoof 2015*“ duomenų rinkinys. Jis sudarytas iš dar trijų smulkesnių rinkinių: apmokymo, vystymo, vertinimo. Apmokymo dalis yra sudaryta iš 3750 orignalių ir 12625 sintetinių įrašų. Vystymo iš 3497 originalių ir 49 875 sintetinių, ir vertinimo iš 9404 tikrų ir 184 000 sintetinių įrašų. Taip pat apmokymo ir vystymo rinkinių netikri įrašai yra sudaryti naudojant penkias žinomas VC ir TTS technikas ir penkias nežinomas šių tipų technikas.

Tyrime, kuriame panaudojama didelio skirtumo kosinuso nuostolių (LMCL) funkcija ir dažnio maskavimas (FM), mokslininkai naudoja trys duomenų rinkinius:

- „*ASVSpoof 2019 Challenge Dataset*“.
- „*Augmented ASVSpoof 2019*“.
- „*Logically-Replayed ASVSpoof 2019*“.

„*Augmented ASVSpoof 2019*“ duomenų rinkinys naudojamas siekiant patikrinti sistemą triukšmo salygomis, išplėčiant „*ASVSpoof 2019*“ rinkinį. Duomenims papildyti panaudoti dviejų tipų iškraipymai: aidas ir foninis triukšmas. Aidas buvo parinktas iš viešai prieinamų patalpos impulsų atsakų (RIR) duomenų rinkinių. Pasirinkti keturi skirtingi foniniai trukšmai: muzika, televizija, šneka, laisvas garsas. „*Logically-Replayed ASVSpoof 2019*“ rinkinys buvo panaudotas balso klastojimo imitavimui skambučių centro aplinkoje, taip pat buvo sumažintas diskretizavimo dažnis nuo 16 kHz iki 8 kHz. Šis duomenų rinkinys panaudotas įvertinimui.

Straipsnyje, kuriame yra pristatoma „*Deep4SNet*“ sistema buvo naudojamas „*H-Voice*“ duomenų rinkinys. Šis rinkinys yra sukurtas remiantis imitaciniai ir sintetiniais balsais, kalbančiais įvairiomis kalbomis, anglų, ispanų, portugalų, prancūzų, tegalų. Įrašai yra saugomi kaip histogramos, rinkinį sudaro 6672 įrašai, 6596 imitaciniai ir 76 sintetiniai. Sintetiniams garsams sukurti panaudotas „*Deep Voice 3*“ modelis.

Taip pat buvo pasiūlytas mokslininkų iš Vroclavo mokslo ir technologijų universiteto, Lenkijos savo duomenų rinkinys. Jie pasiūlė „Attack Agnostic“ duomenų rinkinį, kuriame sujungti 3 duomenų rinkiniai – „*ASVspoof (LA)*“, „*WaveFake*“ ir „*FakeAVCeleb*“ garso įrašų poaibis, kuriuose iš viso yra 31 083 tikri ir 222 035 netikrų pavyzdžių, apimančių 27 įvairius metodus. Jų tikslas – sukurti įrankį, kuris leistų analizuoti modelio apibendrinimą ir stabilumą. Todėl siūlomi trys skirtingi duomenų rinkinio skaidiniai, kuriuose įrašai skirtingai paskirstomi į apmokymo, testavimo ir įvertinimo poaibius. Modifikacijos, naudojamos iškraipymams sukurti, yra nenuosekliai suskirstytos į šiuos pogrupius. Šis metodas leidžia atkreipti dėmesį į atskirų rezultatų skirtumus (Kawa et al., 2022).

Mokslininkų iš Pakistano ir Saudo Arabijos tyrime, kuriame jie pasiūlo „*ELTP-LFCC*“ garso požymių deskriptorių eksperimente buvo panaudotas „*ASVspoof 2019*“ duomenų rinkinys, konkrečiai loginės prieigos poaibis. Jis sudarytas iš apmokymo, vertinimo ir vystymo poaibių, taip pat kiekvienas poaibis sudarytas iš tikrų ir netikrų įrašų, kur sintetiniai įrašai generuojami naudojant kelis algoritmus. Nesintetinės kalbos įrašai surinkti iš 107 kalbėtojų. Apmokymo poaibį sudaro 25 380 pavyzdžiai, vystymo 24 986, o vertinimo 71 933 įrašai. Kiekvieno įrašo trukmė yra nuo vieno iki dviejų sekundžių, o visi įrašai yra saugomi FLAC (angl. *Free Lossless Audio Codec*) formatu, kuris yra skirtas suspaust garso įrašą neprarandant kokybės.

1 lentelė. Tyrime naudoto „*ASVspoof 2019*“ duomenų rinkinio statistika pagal sintetinių ir tikrų įrašų skaičių kiekviename poaibyje, vyrų ir moterų kalbėtojų skaičių, klastojimo algoritmus ir diskretizavimo dažnį. Parengta pagal (Arif et al., 2021)

Poaibis	Garso įrašai	Kalbėtojai	Suklastojimo algoritmai			Diskretizavimo dažnis
	Loginės prieigos	Iš viso: 107	VC	TTS	VC ir TTS	16 kHz

	Tikri	Suklastoti	Vyrai	Moterys				
Apmokymo	2 580	22 800	8	12	2	4	x	
Vystymo	2 548	22 296	8	12				
Vertinimo	7 355	63 882	30	37	5	11	3	

1.5. Pirmo skyriaus apibendrinimas

Šiame skyriuje buvo apibrėžta sintetinių balso įrašų sąvoka, aprašyti jų tipai, generavimo principas. Paaiškinta sintetinių balso įrašų generavimo ir jų identifikavimo problematika. Apžvelgti sintetinių balso įrašų mašininio mokymosi ir gilaus mokymosi identifikavimo metodai, eksperimentuose naudojami duomenų rinkiniai, kaip buvo atliekami eksperimentai, jų salygos. Taip pat akcentuota balso gilių klastočių technologija, kuri pastaruoju metu yra labai aktuali.

1.6. Pirmo skyriaus išvados ir tolimesni darbai

Apibendrinus atliktą literatūros analizę suformuluojamos tokios išvados:

1. Apžvelgus sintetinių balso įrašų generavimo principus paaiškėjo, kad egzistuoja įvairūs jų tipai ir generavimo metodai, kurie nesustoja tobulėti.
2. Išsiaiškinta sintetinių balso įrašų generavimo ir jų identifikavimo problematika. Generuojant sintetinius balso įrašus vis dar trūksta kai kurių balso kokybės aspektų būdingų žmonėms, pavyzdžiui išraiškingumas, šiurkštumas, dusulys, stresas, emocijos. Identifikuojant sintetinius balso įrašus dabartiniai metodai vis tiek susiduria su sunkumais, todėl yra būtina atlikti tolesnius tyrimus, kad būtų galima suprasti, kuria kryptimi sintetiniu balso įrašų identifikavimą reikia toliau tobulinti.
3. Vertinant algoritmų našumą, jis pablogėdavo su nematytais duomenimis arba skirtingais generavimo tipais, papildomas apribojimas esamų metodų yra identfikavimo našumas esant ribotiems mokymų duomenims ir skaičiavimo ištekliams.
4. Pasiūlytiems metodams reikia specialaus duomenų apdorojimo, kad jie veiktų gerai, mašininio mokymosi algoritmai reikalauja didesnio laiko kiekio požymių išskyrimui ir duomenų paruošimui. Kalbant apie gilaus mokymosi algoritmus, garso įrašams reikėjo specifinių transformacijų, kad algoritmai veiktų tinkamai.
5. Beveik visuose tyrimuose daugiausia dėmesio skiriama identifikavimo metodams, skirtiems identifikuoti sintetinius balso įrašus anglų kalba.

Remiantis gautomis išvadomis numatyti tolimesni darbai:

1. Išsiaiškinti, ar galima būtų patobulinti jau esamą metodą panaudojant geresnį neuroninį tinklą, kombinuojant identifikavimo modelius ar naudojant papildomas reguliavimo, duomenų padidinimo priemones, kad pagerinti sintetinių balso įrašų identifikavimo metodų efektyvumą;
2. Sukurti duomenų rinkinį, kuris panaudojamas metodo patobulinimui identifikuoti sintetinius ir nesintetinius balso įrašus;
3. Suprogramuoti metodo patobulinimą;
4. Atlikti pasiūlyto metodo empirinį tyrimą;
5. Pateikti empirinio tyrimo išvadas ir rekomendacijas.

2 lentelė. Palyginamoji straipsnių ir jose taikytų metodų, duomenų rinkinių lentelė

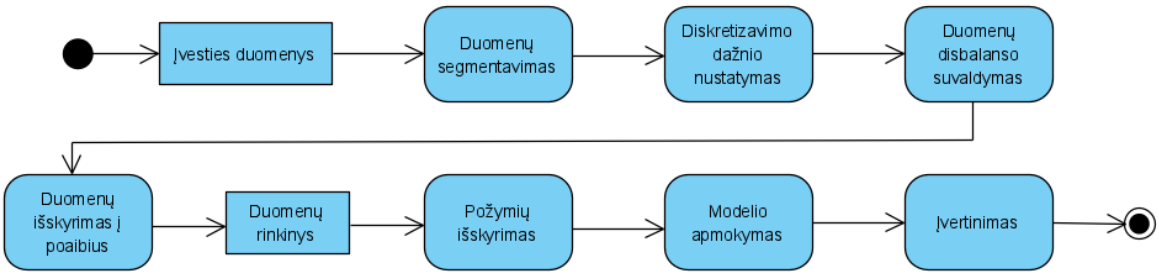
Eil. Nr	Nuoroda į šaltinį	Straipsnio publikavimo metai	Naudoti duomenų rinkiniai	Naudoti metodai	Geriausias našumo įvertinimas	Metodas arba modelis, su kuriuo pasiektas geriausias rezultatas
1	(Singh & Singh, 2020)	2020	Savo sukurtas duomenų rinkinys	Q-SVM, tiesinis diskriminantas, kvadratinis diskriminatas, linijinis SVM, svertinis KNN, medžių ansamblis ir logistinė regresija	Tikslumas: 97,56%	Q-SVM
2	(Lim et al., 2022)	2022	„*ASV Spoof 2021*“, „*LJSpeech*“	CNN, CNN-LSTM, CNN-LSTM-perm	Tikslumas: 99,97%	CNN-LSTM-perm
3	(Zhang et al., 2021)	2021	„*ASV Spoof 2019*“	„*TEResNet*“	EER: 3,99%	„*TEResNet*“
4	(Khochare et al., 2022)	2022	FoR	TCN, STN, CNN	Tikslumas: 92%	TCN
5	(Yu et al., 2018)	2018	„*ASV Spoof 2015*“	DNN-HLL, GMM-LLR	EER: 0,045%	DNN-HLL

2. Siūlomas būdas

Sintetinių ir nesintetinių balso įrašų identifikavimas yra sudėtingas uždavinys dėl gilaus mokymosi metodų, naudojamų labai tikroviškiems ir įtikinamiems netikriems, sintetiniams balso įrašams generuoti, pažangos. Siūlomą būdą iš esmės sudarys trys pagrindinės dalys:

- Duomenų rinkinys.
- Klasifikavimo modelis.
- Vertinimo dalis.

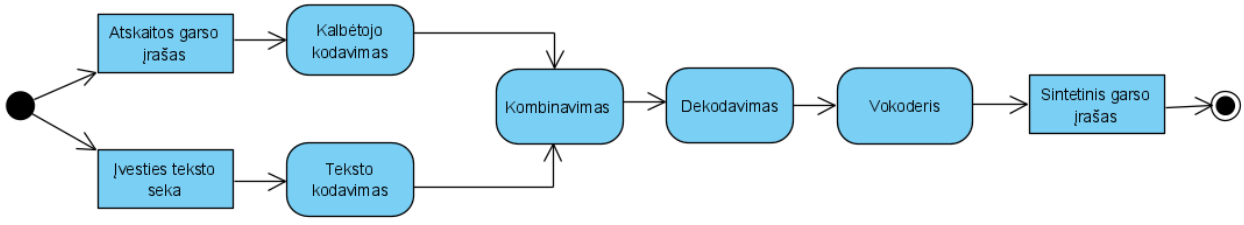
Svarbi dalis yra surinkti įvairų duomenų rinkinį, sudarytą iš nesintetinių ir sintetinių balso įrašų. Duomenų rinkinys turėtų apimti įvairius kalbėtojus, kalbas ir akustines sąlygas, kad būtų užtikrintas identifikavimo modelio patikimumas ir apibendrinimas. Sekantis dalykas yra pirminis išankstinis apdorojimas (angl. *pre-processing*). Pirminis balso duomenų apdorojimas, kad jie būtų paruošti analizei. Tai apima tokius etapus kaip garso normalizavimas, diskretizavimo dažnio (angl. *sampling rate*) nustatymas ir požymių išskyrimas. Konkretūs naudojami išankstinio apdorojimo metodai gali skirtis priklausomai nuo pasirinktų išskirti požymių ir pasrinkto identifikavimo modelio. Kita svarbi dalis yra modelio pasirinkimas. Pasirinkti tinkamą mašininio mokymosi arba gilaus mokymosi modelį sintetiniams ir nesintetiniams balso įrašams identifikuoti. Galima naudoti įvairius modelius, įskaitant tradicinius, giliuosius neuroninius tinklus, konvoliucinius neuroninius tinklus, pasikartojančius neuroninius tinklus arba pažangesnes architektūras, specialiai sukurtas sintetinių balso įrašų analizei. Tuomet pasirinktą modelį apmokyti naudojant iš anksto apdorotus balso duomenis. Tai reiškia, kad išskirtos savybės kartu su atitinkamomis etiketėmis (sintetinis arba nesintetinis) įvedamos į modelį ir optimizuojami jo parametrai. Mokymo proceso tikslas – sudaryti sąlygas modeliui išmokti modelius ir savybes, pagal kurias būtų galima atskirti sintetinius ir nesintetinius balso įrašus. Paskutinis etapas būtų įvertinimas, įvertinti apmokyto modelio veikimą naudojant vertinimo rodiklius, tokius kaip tikslumas, atšaukimas ir F1 balas, EER, AUC rodikliai. Modelis turėtų būti vertinamas naudojant atskirą patvirtinimo arba testavimo duomenų rinkinį, kuriame yra ir nesintetinių, ir sintetinių balso įrašų pavyzdžių. Siekiant patikimesnių rezultatų, taip pat galima taikyti kryžminio vertinimo arba kitus vertinimo metodus. 2.1 pav. galima matyti siūlomo būdo UML veiklos diagrama:



2.1 pav. Siūlomo būdo UML veiklos diagrama

2.1. Duomenų rinkinys

Surinkti patikimą ir įvairialypį sintetinių ir nesintetinų balso įrašų duomenų rinkinį gali būti sudėtinga. Renkant duomenų rinkinį gali kilti įvairūs iššūkiai. Viešai prieinamų sintetinių balso įrašų duomenų rinkinių yra nedaug, lietuvių kalba jų praktiškai neegzistuoja. Šiai problemai spręsti mokslinių tyrimų institucijos, organizacijos ir pramonės suinteresuotieji subjektai galėtų bendradarbiauti kurdami ir dalydamiesi lyginamaisiais duomenų rinkiniais, specialiai skirtais sintetiniams ir nesintetiniams balso įrašams identifikuoti. Dėl nedidelio duomenų rinkinų paplitimo nuspręsta savo sintetinių ir nesintetinių balso įrašų rinkinį surinkti rankiniu būdu. Vieni iš populiariausių garso duomenų rinkinių yra sudaryti iš trijų poaibių: testavimo, vertinimo, vystymo. Todėl nuspręsta savo duomenų rinkinį taip pat sudaryti iš trijų tokių pat dalių. Balso įrašų sintezei bus panaudotas „*VITS*“ (angl. *Variational Inference with adversarial learning for end-to-end Text-to-Speech*) – tai galutinis kalbos sintezės modelis, kuris numato kalbos bangos formą priklausomai nuo įvesties teksto sekos. Sąlyginis variacinis autokoderis (VAE), kurį sudaro posteriorinis koderis, dekoderis. Modelis mokomas nuo pastoviai, derinant nuostolius, gautus iš variacinės apatinės ribos ir priešingo mokymo. Siekiant pagerinti modelio išraiškingumą, sąlyginiam išankstiniam pasiskirstymui taikomi normalizuojantys srautai. Darant išvadą, teksto koduotės didinamos pagal trukmės prognozavimo modulį, o tada atvaizduojamos į bangos formą naudojant srauto modulio ir „*HiFi-GAN*“ dekoderio kaskadą. 2.2 esančiame paveikslėlyje galima matyti tekstas į kalbą (TTS) proceso diagrama:



2.2 pav. Garso įrašo sintezavimas pasinaudojant TTS (angl. *Text-to-speech*) technika

„*LIEPA TTS*“ (angl. *Text-To-Speech*) sintezės variklis lietuvių kalbai. Pagrindinis sintezatoriaus branduolys yra LIEPA kalbos sintezatorius. Šis teksto į kalbą sintezatorus padės susintezuoti balso įrašus lietuvių kalba iš įvesties teksto paduoto jam. Taip pat šis sintezatorius gali sintezuoti balso įrašus, tiek vyrų, tiek moterų.

Pirminiai išankstinio apdorojimo (angl. *preprocessing*) etapai padeda paruošti garso duomenis sintetinių įrašų identifikavimo modeliams, kad jie galėtų išmokti ir nustatyti modelius, susijusius su nesintetiniais garso įrašais ir sintetiniais. Konkretūs naudojami iš ankstinio apdorojimo metodai gali skirtis priklausomai nuo duomenų rinkinio pobūdžio ir naudojamo gilaus mokymosi modelio reikalavimų. Todėl svarbus dalykas yra požymių išskyrimas. Taip pat garso segmentavimas padės ilgus garso įrašus suskirstyti į mažesnes dalis ar kadrus, tai leis modeliui analizuoti ir apdoroti mažesnes garso įrašų atkarpas, išvengti atminties apbribojimų mokymo metu. Pasirinkta panaudoti šiuos požymius savo modeliui apmokyti:

- Laiko srities garso įrašai.
- Konstanta Q transformacijos (CQT) požymius.

Matematikoje ir signalų apdorojime pastoviosios Q transformacijos ir kintamosios Q transformacijos, paprasčiausiai vadinamos CQT ir VQT, transformuoja duomenų eilutę į dažnių sritį. Ji susijusi su Furjė transformacija ir labai glaudžiai susijusi su kompleksine Morleto bangine transformacija. Šios transformacijos pritaikytos muzikiniam vaizdavimui.

Laiko srities garso signalai bus sutrumpinti arba prailginti iki 6 sekundžių. Konstantos Q transformacijos požymiai ir kitų požymių ilgis bus 6,4 sekundės, diskretizavimo dažnis 16 kHz. Taip pat duomenų rinkinio klasių disbalanso tvarkymas identifikuojant sintetinius balso įrašus yra svarbus žingsnis siekiant užtikrinti, kad modelis nebūtų šališkas vieno klasės atžvilgiu ir galėtų veiksmingai identifikuoti ir sintetinius, ir nesintetinius balso pavyzdžius. Vienas iš būdų susitvarkyti su duomenų rinkinio disbalansu yra padidinant garso įrašų skaičių tam tikroje klasėje, pavyzdžiui, nesintetinių balso įrašų. Tai galima išspręsti dubliuojant arba generuojant naujus įrašus. Atitinkamai galima ir mažinti sintetinių garso įrašų skaičių, jeigu tam tikroje klasėje įrašų yra žymiai daugiau arba naudoti kartu šias abi metodikas siekiant subalansuoti balso įrašų duomenų rinkinį. Būtina suprasti, kad klasių disbalanso tvarkymo metodo pasirinkimas priklauso nuo duomenų rinkinio ir identifikavimo modelio. Norint nustatyti efektyviausią būdą reikia eksperimentuoti ir įvertinti skirtingus metodus konkrečiuose duomenų rinkiniuose. Taip pat sprendžiant duomenų disbalanso problemą atsižvelgiant į tai, kad bendruosiuose sintetinių įrašų identifikavimo uždaviniuose nesintetinių pavydžių skaičius yra daug mažesnis nei sintetinių bus pritaikoma svertinės kryžminės entropijos (WCE)

funkcija siekiant susidoroti su duomenų disbalansu rinkinyje. Jeigu $\{x_i,y_i\}$ sudaro pažymėtą apmokymo poaibį, kur $\forall i$ žymi $y_i \in \{0,1\}$, tuomet WCE nuostolio funkcija yra tokia:

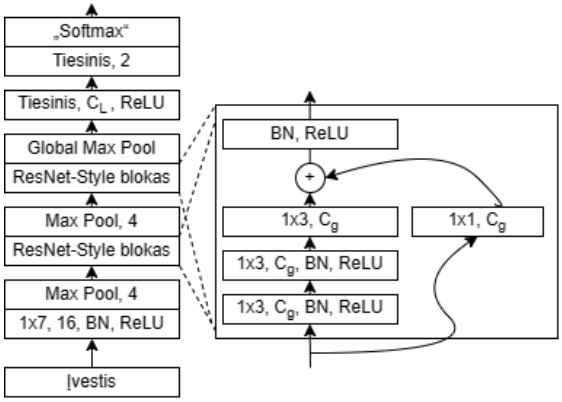
$$WCE(z, y_i) = -w_{y_i} \log(z_{y_i})$$

Kur $z = [z_0, z_1]$ – tai 2 „*softmax*“ klasių tikimybės, o w_{y_i} yra atvirkštinis y_i žymos santykis apmokymo poabyje. Visiems mokymo procesams naudojamas Adamo optimizatorius ir numatyti nustatymai. Eksponentinis mokymosi greičio mažėjimas su daugikliu 0,95, partijos dydis 32. Modelis duodantis mažiausią vienodą klaidų lygį (EER) atrenkamas vertinimui. Mokymo proceso metu naudojant Adamo optimizatorių, modelis gali veiksmingai optimizuoti savo parametrus, reguliuoti mokymosi greitį pagal kiekvieną parametrą ir konverguoti prie gero sintetinių ir nesintetinių balso įrašų identifikavimo užduoties sprendimo, tačiau tam, kad pasiekti optimalius rezultatus konkrečiuose identifikavimo uždaviniuose, gali prireikti eksperimentuoti ir tikslinti parametrų pasirinkimus.

Taip pat bus pritaikomas sumaišymo reguliavimas (angl. *Mixup Regularization*), kad apmokytas modelis geriau apibendrintų nematytas technikas. Sumaišymo reguliavimas turės pagerinti apibendrinimo galimybes. „*Mixup*“ yra populiarus giliųjų neuroninių tinklų mokymo reguliarizavimo metodas, kuris gali pagerinti apibendrinimą ir padidinti atsparumą priešiškiems veiksniams. Jis iškreipia įvesties mokymo duomenis kitų atsitiktinai parinktų mokymo aibės atvejų kryptimi.

2.2. Klasifikavimo modeliai

Renkantis identifikavimo modelius, svarbu atsižvelgti į konkrečius užduoties reikalavimus ir ypatybes, duomenų rinkinio struktūrą, įvairumą. Šiam tikslui yra pasirenkamas laiko srities sintetinės kalbos aptikimo tinklas (angl. *Time-domain Synthetic Speech Detection Net*), toliau (TSSDNet), kuris turi „*ResNet*“ arba „*Inception*“ stiliaus struktūras. Siūlomi modeliai turi geras apibendrinimo galimybes. „*Res-TSSDNet*“ ir „*Inc-TSSDNet*“, kurie atitinkamai turi tą patį pirmąjį sluoksnį, 3 galutiniai visiškai sujungts linijinius sluoksnius ir globalų maksimalų sujungimą prieš linijinius sluoksnius. „*ResNet*“ stiliaus ir „*Inception*“ stiliaus blokai kartojami atitinkamai M kartų, ir abiejuose tinkluose taikomas paketinis normalizavimas (BN). C_R ir C_I žymi kanalų skaičių atitinkamame modulyje, kurie įvairiuose sluoksniuose gali skirtis. Norint padidinti recepcinio lauko ir valdymo modelio sudėtingumą, reikia išplėsti konvoliuciją su dilatacija d yra įtraukus į „*Inc-TSSDNet*“. Visuose konvoliucijos sluoksniuose taikomas „*SAME*“ užpildymas su žingsniu = 1, o sujungimo sluoksniams žingsnio reikšmė lygi atitinkamam branduolio dydžiui. „*Res-TSSDNet*“ struktūra pavaizduota 2.3 esančiame pav.



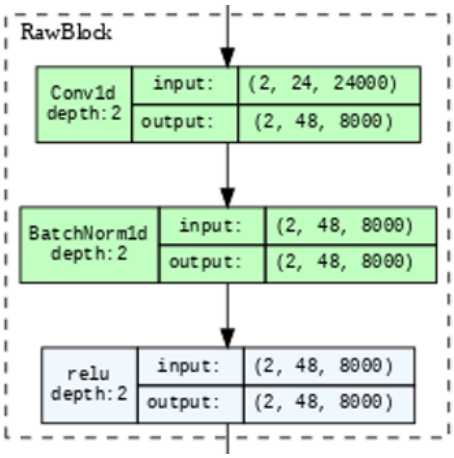
2.3 pav. „*Res-TSSDNet*“ modelio strukūra. Parengta pagal (Hua et al., 2021)

Trečias modelis „*2D-Res-TSSDNet*“ yra „*Res-TSSDNet*“ 2D versija, turi tą pačią architektūrą, išskyrus tai, kad visos konvoliucijos ir sujungimo (2 × 2 sujungimo) sluoksniai naudoja 2D branduolius.

Kaip identifikavimo modelių patobulinimai buvo sukurti trys identifikavimo modeliai:

- „*RawNet*“.
- „*2D-Inc-TSSDNet*“.
- Kombinuotas „*RawNet*“ ir „*Inc-TSSDNet*“.

Šioje „*RawNet*“ modifikuotoje versijoje taikomas sluoksnio normalizavimas, naudojami keturi „*RawBlock*“ blokai, šie blokai sudaryti iš vieno konvoliucijos ir rinkinio normalizavimo sluoksnio. Taip pat naudojami trys sujungimo sluoksniai su „*ReLU*“ aktyvacijos funkcija, paskutiniame sluoksnyje ji nenaudojama ir globalus vidurkio skaičiavimo sluoksnis. Taip pat tik šiame identifikavimo modelyje naudojamas „stride“ parametras. „Stride“ yra parametras, kuris diktuoja branduolio arba filtro judėjimą per įvesties duomenis. Atliekant konvoliucijos operaciją, „stride“ nustato, kiek vienetų filtras pasislenka kiekviename žingsnyje. Identifikavimo modelis bus apmokytas naudojant Adamo optimizavimą, taikant mokymosi greitį 0,001. 2.4 pav. galima matyti „*RawNet*“ modelio „*RawBlock*“ bloko strukūrą.



2.4 pav. „*RawBlock*“ bloko struktūra

Toliau nuspręsta sukurti savo „*2D-Inc-TSSDNet*“ identifikavimo modelį, šis modelis iš esmės yra 2D versija paprasto „*Inc-TSSDNet*“ modelio, todėl ir architektūra šio modelio nesiskiria labai stipriai, vietoj paprastų konvoliucijos sluoksnių naudojami 2D konvoliucijos sluoksniai.

Paskutinis identifikavimo modelis yra kombinacija „*RawNet*“ ir „*Inc-TSSDNet*“ identifikavimo modelių, tokiu būdų yra apjungiamos šių modelių ypatybės siekiant gauti geresnius rezultatus.

2.3. Rezultatų įvertinimas

Vertinant sintetinio ir nesintetinio balso identifikavimo modelius, reikia įvertinti jų našumą tiksliai atpažįstant tikrus ir sintetinius balso įrašų pavyzdžius. Naudojamos įprastos vertinimo priemonės ir metodai, skirti sintetinių ir nesintetinių balso įrašų identifikavimo modelių veiksmingumui įvertinti:

- Tikslumas, kuris apskaičiuojamas iš teisingai suklasifikuotų pavyzdžių ir bendro pavyzdžių skaičiaus santykio, įvertinamas bendras modelio prognozių teisingumas. Tačiau vien tik tikslumo gali nepakakti, kai susiduriama su nesubalansuotais duomenų rinkiniais, kuriuose dominuoja viena tam tikra klasė.
- Preciziškumas (angl. *Precision*) – tai teisingumo matas, pasiekiamas atliekant tikrąjį numatymą. Paprastais žodžiais tariant, ji mums nurodo, kiek prognozių iš tikrųjų yra teigiamų iš visų numatytų teigiamų pavyzdžių. Jis apibrėžiamas kaip bendro teisingai klasifikuotų teigiamų klasių skaičiaus ir bendro prognozuojamų teigiamų klasių skaičiaus santykis. Arba kiek iš visų nuspėjamų teigiamų klasių teisingai nuspėjome.
- Atšaukimas (angl. *Recall*) – Tai teisingai prognozuotų faktinių stebėjimų matas, t. y. kiek teigiamos klasės stebėjimų iš tikrųjų prognozuojama kaip teigiami. Jis taip pat vadinamas jautrumu. Atšaukimas yra tinkamas vertinimo metrikos pasirinkimas, kai norime užfiksuoti kuo daugiau teigiamų teiginių.
- F1 įvertis – tai skaičius nuo 0 iki 1 ir yra tikslumo ir atšaukimo harmoninis vidurkis. Šis vidurkis skirtingai nuo paprastų vidurkių, nėra jautrus labai didelėms reikšmėms. F1 rezultatas palaiko pusiausvyrą tarp klasifikatoriaus tikslumo ir atšaukimo. Jei tikslumas mažas, F1 rodiklis yra mažas, o jei atšaukimas mažas, F1 rodiklis taip pat yra mažas.
- AUC įvertis – tai plotas po ROC kreive. Tai yra sritis po ROC (angl. *Receiver Operating Characteristic*) kreive, kuri matuoja modelio gebėjimą atskirti teigiamus ir neigiamus atvejus.

AUC (angl. *Area Under the Curve*) vertė parodo, kaip gerai modelis skiria pavyzdžius tarp teigiamų ir neigiamų atvejų.

- Vienodas klaidų dažnis (angl. *Equal Error Rate*) – tai dažniausiai naudojamas našumo rodiklis, skirtas įvairių sistemų veiksmingumui vertinti. EER matuoja tašką, kuriame klaidingo priėmimo lygis (FAR) yra lygus klaidingo atmetimo lygiui (FRR). EER yra svarbus rodiklis, nes juo subalansuotai įvertinamas sistemos našumas. Mažesnės EER reikšmės rodo geresnį veikimą, nes jos rodo mažesnį bendrą klaidų lygį atskiriant tikrą ir sintetinį balso įrašą.
- Atvaizdavimui planuojama naudoti gavėjo operacinių charakteristikų (ROC) kreivę. Tai grafinis vaizdas, iliustruojantis dvejetainio klasifikavimo modelio našumą esant skirtingoms klasifikavimo slenkstinėms (angl. *threshold*) reikšmėm. ROC kreivė gali būti naudojama identifikavimo sistemos našumui įvertinti, nubrėžiant tikrojo teigiamo rodiklio (TPR) ir klaidingai teigiamo rodiklio (FPR) santykį esant skirtingoms klasifikavimo riboms.
- Sutrikimo matrica (angl. *Confusion matrix*) – yra klasifikatoriaus pateiktų teisingų ir neteisingų prognozių skaičiaus lentelės suvestinė. Jis gali būti naudojamas klasifikavimo modelio našumui įvertinti. Paprastai tariant, tai yra lentelė, kuri rodo klasifikavimo modelio veiksmingumą prognozuojant tikrus ir netikrus rezultatus.

2.4. Antro skyriaus apibendrinimas

Šiame skyriuje buvo aprašytas planuojamas surinkti duomenų rinkinys, taip pat iš ankstinis apdorojimas (angl. *pre-processing*), požymių išskyrimas, rinkinio struktūra. Apibrėžti eksperimentui naudojami identifikavimo modeliai sintetiniams ir nesintetiniams balso įrašams nustatyti. Nuspręsta naudoti Adamo optimizatorių apmokymo procesui. Pasirinktos technikos sugeneruoti sintetiniams balso įrašams. Išvardinti rezultatų įvertinimo žingsniai ir būdai.

2.5. Antro skyriaus išvados

Apibendrinus šį skyrių suformuluojamos tokios išvados:

1. Išanalizavus įvariuis duomenų rinkinius, jų strukūrą, nuspręsta surinkti savo duomenų rinknį sintetiniams ir nesintetiniams balso įrašams identifikuoti.
2. Nuspręsta naudot šiuos požymius klasifikavimo modeliams apmokyti:
    a. Laiko srities garso įrašai.
    b. Konstanta Q transformacijos (CQT) požymius.

3. Apžvelgus sintetiniams balso įrašams naudojamus identifikuoti klasifikavimo modelius nuspręsta naudoti šiuos modelius savo darbe:
    a. „*Res-TSSDNet*“.
    b. „*Inc-TSSDNet*“.
    c. „*2D-Res-TSSDNet*“.
    d. Pritaikytas „*RawNet*“.
    e. Pritaikytas „*2D-Inc-TSSDNet*“.
    f. Kombinuotas „*Inc-TSSDNet*“ ir „*RawNet*“ identifikavimo modelis.
4. Apmokymo procesui naudojamas Adamo optimizatorius.
5. Parinktos vertinimo priemonės gautiems rezultatms įvertinti.

3. Eksperimentinė dalis

Sintetinių balso įrašų identifikavimas yra esminis žingsnis į priekį siekiant pagerinti balso atpažinimo sistemas ir jų naudojimą įvairiose srityse, įskaitant kalbos technologijas, saugumą ir sąveiką su elektroniniais įrenginiais. Šiame eksperimente bus stengiamasi suprasti, kaip giliosios mokymosi technologijos gali būti pritaikytos sintetinių balsų identifikacijai ir kiek tai gali pagerinti balso identifikavimo tikslumą. Siekiama išbandyti pasirinktus modelius, kad nustatyti identifikavimo galimybes, taip pat jų pranašumus ir trūkumus, kaip teisingai modeliai sugeba nusatyti tikrus ir sintetinius balso įrašus, koks yra jų veikimas su savo surinktu duomenų rinkiniu, palyginti juos tarpusavyje ir pateikti savo įžvalgas apie gautus rezultatus šioje dalyje. Šiame skyriuje bus panaudoti šie gilaus mokymosi identifikavimo modeliai:

- „*Res-TSSDNet*“.
- „*Inc-TSSDNet*“.
- „*2D-Res-TSSDNet*“.
- Pritaikytas „*RawNet*“.
- Pritaikytas „*2D-Inc-TSSDNet*“.
- Kombinuotas „*Inc-TSSDNet*“ ir „*RawNet*“ modelis.

3.1. Duomenų rinkinio surinkimas

Duomenų rinkiniai yra iš esmės pagrindas, kuris leidžia giliuosius mokymosi modelius apmokyti, įvertinti ir tobulinti identifikavimo užduotims. Štai kodėl duomenų rinkiniai yra nepaprastai svarbūs tokio tipo užduotims. Duomenų rinkiniai suteikia modeliams autentiškų ir sintetinių balso įrašų, kurie reikalingi mokymui. Tinkamai parengtas duomenų rinkinys leidžia modeliams išmokti identifikuoti įvairias sintezavimo technikas ir skirtumus tarp autentiškų ir sintetinių balso įrašų. Duomenų rinkiniai taip pat naudojami modelių validacijai ir testavimui. Jie padeda įvertinti modelių tikslumą, jautrumą, specifiškumą ir kitus veiksmingumo rodiklius, naudojant atskirą duomenų rinkino dalį, kuri nebuvo naudojama apmokymui. Galiausiai, kuo didesnis ir įvairesnis duomenų rinkinys, tuo geriau modeliai gali įsisavinti ir identifikuoti skirtingas sintezavimo technikas. Dėl prieš tai išvardintų priežaščių nuspręsta surinkti savo balso įrašų duomenų rinkinį lietuvių kalba. Tam, kad išanalizuoti siūlomų metodų identifikavimo tikslumą ir kitas vertinimo metrikas labai svarbu turėti gerą ir reprezentatyvų duomenų rinkinį efektyvumo įvertinimui, kad būtų parodytas identifikavimo modelių gebėjimas apibendrinti. Savo duomenų rinkiniui tikri, nesintezuoti garso įrašai buvo paimt iš kelių audio knygų skaidant knygas dalimis į trumpus garso įrašus ir paskirstant juos per duomenų rinkinio apmokymo, vystymo, vertinimo dalis. Kiekvieno garso įrašo diskretizavimo dažnis duomenų rinkinyje

buvo konvertuotas į 16 000 Hz.. Šiuo metu turimų duomenų rinkinių diskretizavimo dažnis yra maždaug 16 kHz, todėl kalbos kokybė nėra visiškai tobula. Padidinus diskretizavimo dažnį, generavimas gali būti kokybiškesnis. Taip pat visi įrašai buvo konvertuoti į mono garsą, kuris yra vieno kanalo garsas, kai visi instrumentai sumaišomi į vieną signalą, skirtą girdėti tarsi sklindantį iš vienos padėties. Palyginimui, stereo (stereofoninis) garsas pasiekiamas naudojant du garso kanalus, maitinančius 2 atskirus garsiakalbius. Balso įrašų sintezavimui buvo panaudotis dvi teksto į kalbą technikos:

- „*VITS*“.
- „*LIEPA TTS*“.

Žymint garso įrašus protokolo faile buvo nurodomi kalbėtojo ID, garso failo pavadinimas, kalbėjimo suklastojimo technika ir raktas, „bonafide“, jei kalba yra nesintetinė, „spoof“, jei kalba yra sintetinė. Dalis sintezuotų garso įrašų buvo sukurti tiek vyro balsu, tiek moters balsu. Taip pat svarbu paminėti svarbią eksperimento sąlygą, kad identifikavimo modeliams apmokyti bus naudojami balso įrašai sukurti su „*VITS*“ technika, o vertinama bus su „*VITS*“ ir „*LIEPA TTS*“, kad nustatyti, kaip modelis identifikuos nematytus apmokymo metu balso įrašus.

3 lentelė. Surinkto duomenų rinkinio nesintetinių ir sintetinių balso įrašų skaičius

Poaibis	**Balso įrašai**		
	Nesintetiniai	**Sintetiniai**	**Iš viso įrašų**
Apmokymo	1 024	1 977	3 001
Vystymo	1 092	4 904	5 996
Vertinimo	3 030	6 161	9 191

3.2. Modelių veikimo reguliavimo pritaikymas

Eksperimente nuspręsta panaudoti įvarias reguliavimo priemones, kurios gerina apibendrinimo galimybes, pirmoji yra svertinės kryžminės entropijos (WCE) funkcija, taip pat nuspręsta panaudoti sumaišymo reguliavimą. Sumaišymo reguliavimas (angl. *Mixup reguliarization*) – tai sintetinių duomenų generavimo ir modelio apibendrinimo tobulinimo technika. 2018 m. pristatytas duomenų papildymo metodas, skirtas neuroniniams tinklams apmokyti, konstruojant virtualius mokymo pavyzdžius naudojant išgaubtus pavyzdžių porų ir jų žymėjimo derinius. Iš esmės šis metodas sureguliuoja neuroninį tinklą, kad tarp mokymo pavyzdžių būtų palankesnis paprastas tiesinis elgesys ir pagerina apibendrinimą. Sumaišymas veikia kaip tam tikra reguliacijos forma, neleidžianti modeliui per daug prisitaikyti prie mokymo duomenų ir taip pagerinanti

apibendrinimą nematytiems pavyzdžiams. Taip pat egzistuoja ir interpoliavimo nauda. Kuriant išgaubtus įvesties ir etikečių derinius, „*mixup*“ skatina modelį išmokti tiesinės interpoliacijos tarp duomenų taškų, todėl sprendimų ribos gali būti lygesnės ir patikimesnės. Ši technika padeda kovoti ir su duomenų disbalansu duomenų rinkinyje, gali padėti sumažinti su duomenų trūkumu ir disbalansu susijusias problemas. Paskutinė šios technikos nauda yra, kad taikant atsitiktinumą, mažiau tikėtina, kad modelis per daug prisitaikys prie triukšmingų ar nereikšmingų mokymo duomenų požymių. Tačiau, kaip ir bet kurio kito reguliavimo metodo atveju, sumaišymo veiksmingumas gali skirtis priklausomai nuo konkrečių duomenų rinkinio savybių ir modelio architektūros. Šioje vietoje labai svarbu eksperimentuoti. Trečias metodas yra etikečių išlyginimas (angl. *Label Smoothing*) – tai dar vienas klasifikavimo uždaviniuose dažnai naudojamas reguliarizavimo metodas. Jis apima vienkartinių užkoduotų tikslinių žymėjimų išlyginimą, pakeičiant reikšmes 0 ir 1 mažomis teigiamomis reikšmėmis, paprastai artimomis nuliui. Tai neleidžia modeliui pernelyg pasitikėti savo prognozėmis ir skatina jį išmokti patikimesnių sprendimo ribų. Etikečių išlyginimas gali padėti pagerinti modelio apibendrinimo rezultatus, ypač kai susiduriama su triukšmingomis ar neapibrėžtomis etiketėmis. Kiekvienas iš šių metodų atlieka svarbų vaidmenį gerinant modelio patikimumą ir apibendrinamąjį pobūdį, nes mokymo metu įvedami skirtingi reguliarumo tipai. Todėl galutinai nuspręsta naudoti šias trys reguliavimo technikas apmokymo metu:

- Svertinės kryžminės entropijos (WCE) funkcija.
- Sumaišymo reguliavimas (angl. *Mixup reguliarization*).
- Etikečių išlyginimas (angl. *Label Smoothing*).

3.3. Naudojami identifikavimo modeliai eksperimente

Šio eksperimento tikslas yra ištirti giliuosius mokymosi metodus, siekiant efektyviai identifikuoti sintetinius ir nesintetinius balso įrašus. Tikimasi, kad gilieji mokymosi modeliai gali sugebėti išgryninti svarbiausius balso bruožus ir identifikuoti gilesnes semantines struktūras, kurios yra reikšmingos balso atpažinimo uždaviniuose.

3.3.1. „Res-TSSDNet“ identifikavimo modelis

„*Res-TSSDNet*“ tai laiko srities sintetinės kalbos aptikimo tinklas (TSSDNet), kuris turi „*ResNet*“ architektūrą. Šį identifikavimo modelį sudaro 1D konvoliucijos sluoksnis „*conv1*“ su „*ReLU*“ aktyvacijos funkcija ir rinkinio normalizavimas, „*ResNet-Style*“ moduliai, kiekvienas su savo konvoliucijos ir rinkinio normalizavimo sluoksniais, taip pat trys pilnai sujungti sluoksniai (fc1, fc2, out), tarp kurių yra „*ReLU*“

aktyvacijos funkcijos, išskyrus paskutinį. Modelio apmokymui naudojami vienodos trukmės laiko srities neapdorotos bangų formos požymiai, kurios yra sugeneruojamos iš savo surinkto duomenų rinkinio.

Taip pat su šiuo modeliu nuspręsta keisti modelio „*ResNet-Style*“ modulių skaičių, kad parinkti geriausią jų kiekį:

4 lentelė. „*Res-TSSDNet*“ modelio rezultatai priklausomai nuo modulių skaičiaus

„*ResNet-Style*“ modulių skaičius	**Vystymo poaibio tikslumas**	**Vertinimo poaibio tikslumas**	**EER**	**Bendras modelio parametrų skaičius**
3	96,88%	88,55%	13,83%	183922
4	83,98%	79,14%	**8,51%**	348530
5	86,46%	82,84%	12,33%	513138
6	74,72%	73,59%	9,85%	677746

Iš gautų rezultatų nuspręsta naudoti 4 „*ResNet-Style*“ modulius, nes šiuo atveju svarbiau yra atsižvelgti į EER rodiklį nei į tikslumą rodiklį, nors šis ir yra pasiekiamas geresnis, bet rinkinyje vyrauja duomenų disbalansas, todėl svarbesnis yra EER rodiklis.

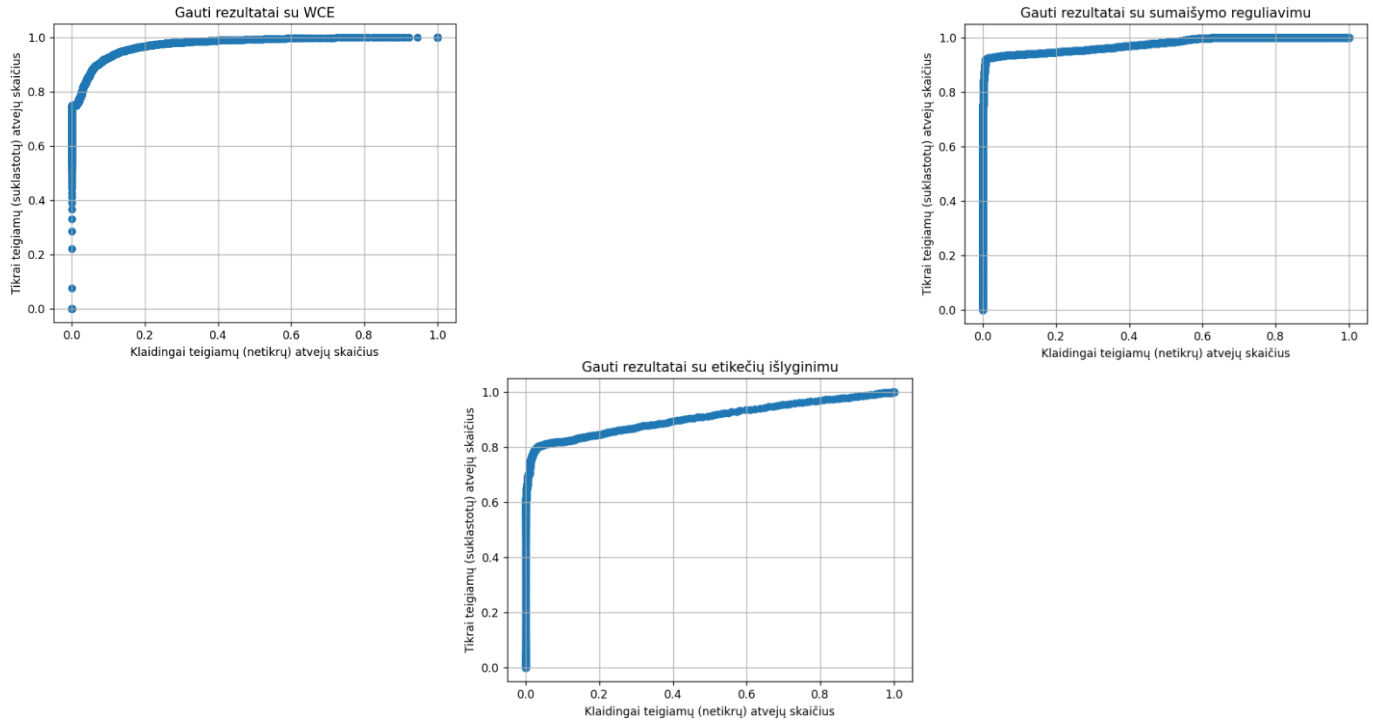
Modelio informacija:

- „*ResNet-Style*“ modulių skaičius: 4.
- Naudojami požymiai apmokymui: laiko srities garso signalai.
- Aktyvacijos funkcija: „*ReLU*“.

5 lentelė. Gauti rezultatai su „*Res-TSSDNet*“ modeliu

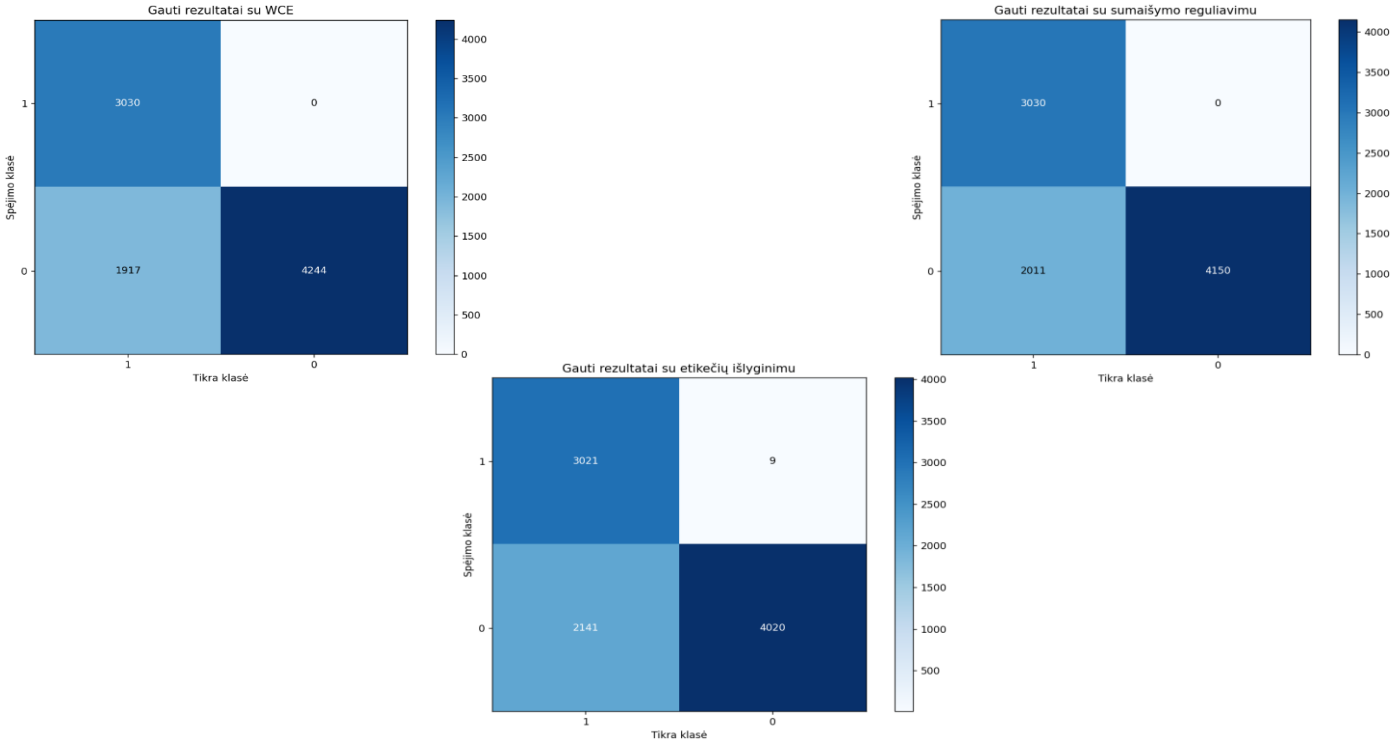
Reguliavimo technika	**Tikslumas**	**EER vertė**	**Preciziškumas**	**Atšaukimas**	**F1 įvertis**	**AUC įvertis**
WCE	79,14%	8,51%	1,0	0,68	0.81	0,84
Sumaišymo reguliavimas	78,11%	6,49%	1,0	0,67	0,8	0,83
Etikečių išlyginimo	76,6%	16,24%	0,99	0,65	0,78	0,82

Gautos ROC kreivės su „*Res-TSSDNet*“ modeliu:



3.1 pav. Gautos ROC kreivės su „*Res-TSSDNet*“ modeliu

Sutrikimo matricos:



3.2 pav. „*Res-TSSDNet*“ sutrikimo matricos

Iš gautų rezultatų galima matyti, kad aukščiausias tikslumas pasiekimas naudojant WCE reguliavimo techniką (5 lentelė), bet mažiausias EER rodiklis su sumaišymo reguliavmo technika 6,4971%, tai taip pat galima matyti vizualiai, nes ROC kreivė (paveikslėlis 3.1) vizuliai idealesnė yra naudojant sumaišymo reguliavimo techniką. Taip pat nustebino, kad naudojant WCE ir sumaišymo reguliavimą identifikavimo metodas neužfiksavo klaidingai teigiamų rezultatų, nes preciziškumo rodikliai lygūs 1, tai taip pat matoma ir iš sutrikimo matricų (paveikslėlis 3.2), dešiniųjų viršutinių kvadratų langeliuose reikšmės lygios 0. Prasčiausi rezultatai pasiekti su etikečių išlyginimo reguliavimo technika, tikslumas 76,6%, EER 16,24%.

3.3.2. „2D-Res-TSSDNet“ identifikavimo modelis

Šis naudojamas modelis „*2D-Res-TSSDNet*“ yra „*Res-TSSDNet*“ 2D versija, kuri turi tą pačią architektūrą, tačiau visos konvoliucijos ir sujungimo (2 × 2 sujungimo) sluoksniai naudoja 2D branduolius. Modeliui apmokyti naudojami konstanta Q transformacijos (CQT) požymiai išskirti iš savo duomenų rinkinio.

Kaip ir su „*Res-TSSDNet*“ identifikavimo modeliu nuspręsta keisti 2D „*ResNet*“ modulių skaičių siekiant nustatyti tinkamiausią jų kiekį.

6 lentelė. „*2D-Res-TSSDNet*“ modelio rezultatai priklausomai nuo modulių skaičiaus

2D „*ResNet-Style*“ modulių skaičius	Vystymo poaibio tikslumas	Vertinimo poaibio tikslumas	EER	Bendras modelio parametrų skaičius
2	74,45%	72,65%	25,43%	371602
3	78,64%	76,12%	**19,8%**	507154
4	72,93%	69,1%	38,71%	966674

Kaip parodė rezultatai aukščiau esančioje lentelėje nuspręsta naudoti 3 2D „*ResNet-Style*“ modulius, nes buvo gautas žemiausias EER rodiklis – 19,8%.

Modelio informacija:

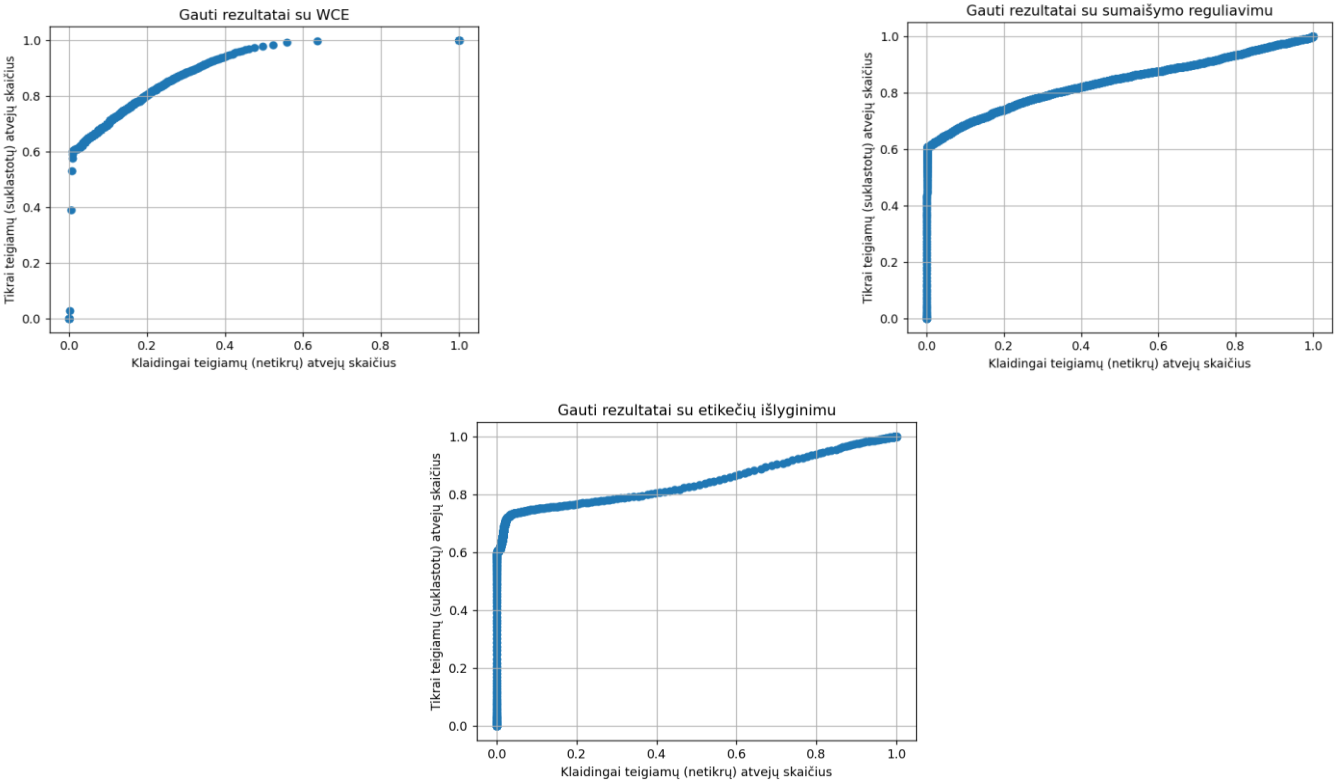
- 2D „*ResNet-Style*“ modulių skaičius: 3.
- Naudojami požymiai apmokymui: CQT požymiai.
- Aktyvacijos funkcija: „*ReLU*“.

7 lentelė. Gauti rezultatai su „*2D-Res-TSSDNet*“ modeliu

Reguliavimo technika	Tikslumas	EER vertė	Preciziškumas	Atšaukimas	F1 įvertis	AUC įvertis
WCE	76,12%	19,8%	0,94	0,68	0,79	0,8

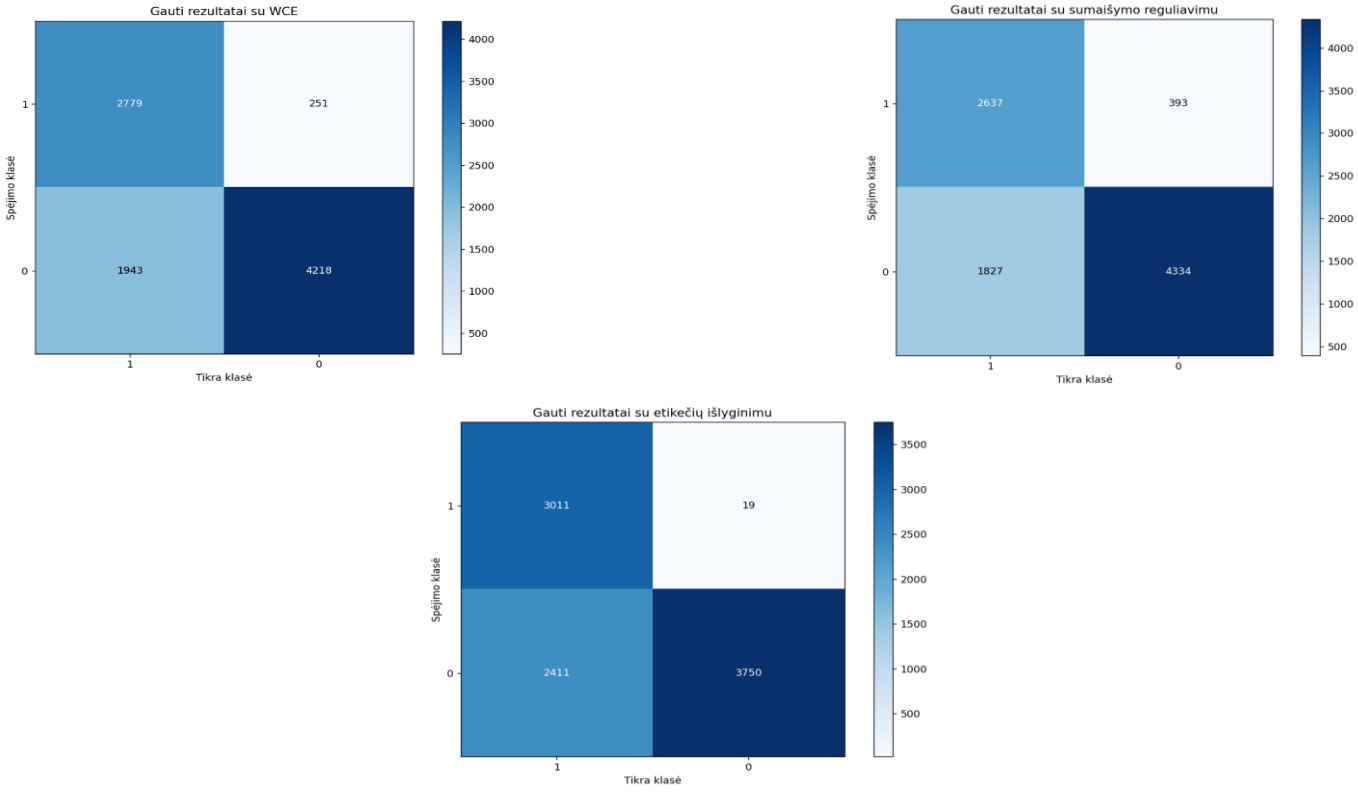
Sumaišymo reguliavimas	75,84%	23,84%	0,91	0,7	0,79	0,78
Etikečių išlyginimas	73,56%	22,83%	0,99	0,6	0,75	0,8

Gautos ROC kreivės:



3.3 pav. Gautos ROC kreivės su „*2D-Res-TSSDNet*“ modeliu

Gautos sutrikimo matricos:



3.4 pav. Gautos sutrikimo matricos su „*2D-Res-TSSDNet*“ modeliu

Iš gautų rezultatų galima matyti, kad aukčiausi rodikliai pasiekti su WCE reguliavimo technika, tikslumas 76,12% ir EER rodiklis 19,8% (7 lentelė). Lyginant sumaišymo reguliavimo ir etikečių išlyginimo technikas, tai pastaroji užfiksavo mažiau kladingai teigiamų atvejų – 19, bet daugiau klaidingai neigiamų – 2411 (paveikslėlis 3.4). Sumaišymo reguliavo technika nustatė 393 klaidingai teigiamus ir 1827 klaidingai neigiamus atvejus.

3.3.3. „Inc-TSSDNet“ identifikavimo modelis

„*Inc-TSSDNet*“ modelis tai taip vadinamas laiko srities sintetinės kalbos aptikimo tinklas (TSSDNet), turintis „Inception“ stiliaus architektūrą. padidintas recepcinio lauko ir valdymo modelio sudėtingumas, išsiplėtė konvoliucijos su dilatacija d yra įtrauktos į „*Inc-TSSDNet*“, kuris skiriasi nuo originalaus „*Inception*“ tinklo. Šis modelis naudoja „*DilatedNet*“ ir „*DilatedCovModule*“ klases. „*DilatedNet*“ naudoja 1D konvoliucijas ir yra paremtas gilaus mokymosi architektūra, kuri naudoja dilatuotas konvoliucijas. „*DilatedCovModule*“ taip pat naudoja dilatuotas konvoliucijas, tačiau tai yra pagalbinė klasė, kuri yra naudojama „*DilatedNet*“ viduje. „*DilatedNet*“ sudaro 1D konvoliucijos sluoksnis „*conv1*“ su ReLU aktyvacijos funkcija ir rinkinio normalizacija (angl. *Batch Normalization*). „*DilatedCovModule*“ moduliai,

kurių kiekvienas naudoja dilatuotas konvoliucijas ir rinkinio normalizacija. O taip pat ir pilnai sujungti sluoksniai tarp kurių yra „*ReLU*“ aktyvacijos funkcijos, išskyrus paskutinį, kuris nenaudoja jokios aktyvacijos funkcijos.

Šio identifikavimo modelio architektūroje buvo keičiamas „*DilatedCovModule*“ skaičius siekiant išsiaiškinti tinkamiausią jų kiekį.

8 lentelė. „*Inc-TSSDNet*“ modelio rezultatai priklausomai nuo „*DilatedCovModule*“ modulių skaičiaus

„*DilatedCovModule*“ modulių skaičius	**Vystymo poaibio tikslumas**	**Vertinimo poaibio tikslumas**	**EER**	**Bendras modelio parametrų skaičius**
2	83,02%	80,63%	10,89%	24690
3	83,27%	79,38%	12,62%	43250
4	89,5%	88,01%	**6,86%**	92658
5	77,79%	75,55%	14,88%	142066
6	77,69%	75,18%	12,68%	191474

Iš gautų rezultatų (9 lentelė) pasirinkta naudoti 4 „*DilatedCovModule*“ modulius.

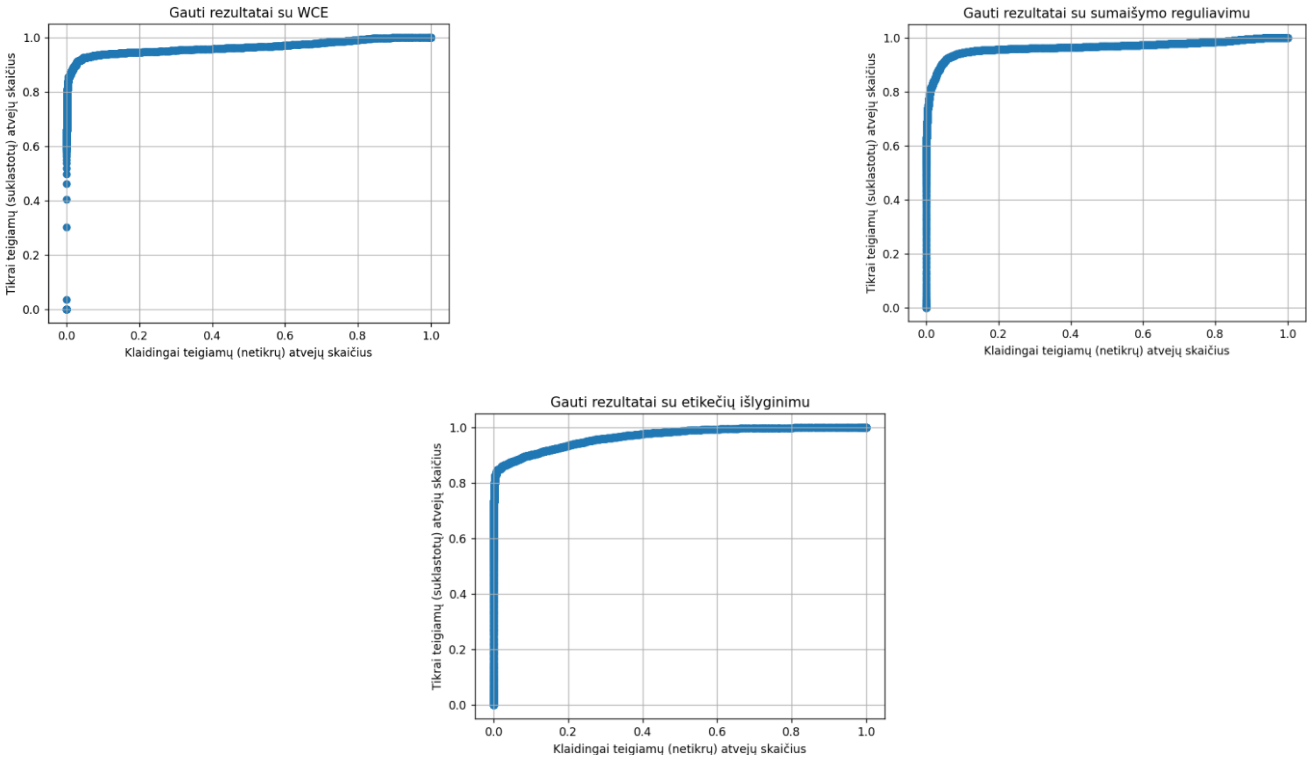
Modelio informacija:

- „*DilatedCovModule*“ modulių skaičius: 4.
- Naudojami požymiai apmokymui: laiko srities garso signalai.
- Aktyvacijos funkcija: „*ReLU*“.

9 lentelė. Gauti rezultatai su „*Inc-TSSDNet*“ modeliu

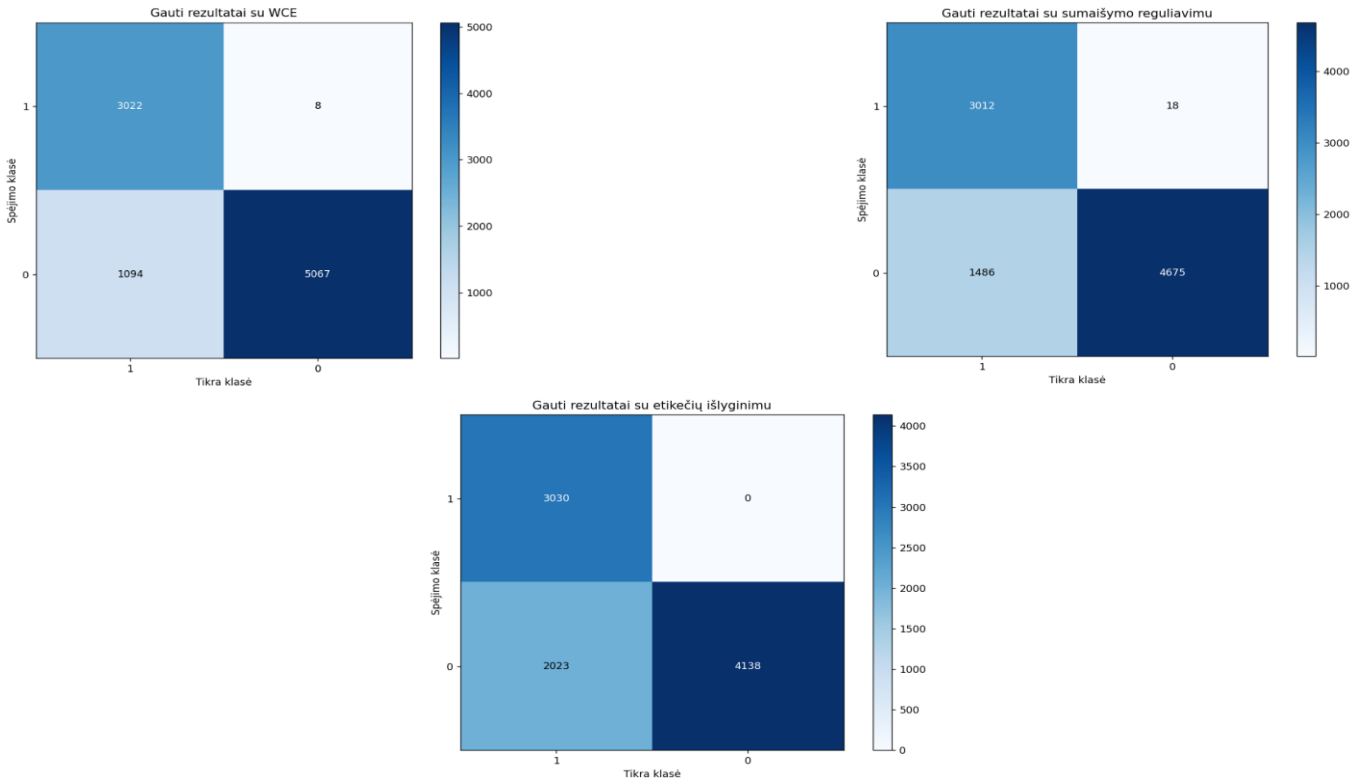
Reguliavimo technika	**Tikslumas**	**EER vertė**	**Preciziškumas**	**Atšaukimas**	**F1 įvertis**	**AUC įvertis**
WCE	88,01%	6,86%	0,99	0,82	0,9	0,9
Sumaišymo reguliavimas	83,63%	6,96%	0,99	0,75	0,86	0,87
Eitkečių išlyginimas	77,98%	9,92%	1,0	0,67	0,8	0,83

Gautos ROC kreivės:



3.5 pav. „*Inc-TSSDNet*“ modelio ROC kreivės

Gautos sutrikimo matricos:



3.6 pav. „*Inc-TSSDNet*“ sutrikimo matricos

Apibendrinant „*Inc-TSSDNet*“ identifikavimo modelio rezultatus galima išskirti, kad atliekant bandymus su etikečių išlyginimo reguliavimo technika, modelis neužfiksavo klaidingai teigiamų įrašų (paveikslėlis 3.6), bet užfiksavo daugiausia klaidingai neigiamų įrašų – 2023. Geriausi rezultatai buvo gauti naudojant WCE techniką, tikslumas 88,01%, o EER 6,86% (9 lentelė). Nedaug nuo WCE atsiliko ir sumaišymo reguliavimas, atitinkamai tikslumas 83,63%, EER 6,96% ir kladingai neigiamų atvejų buvo daugiau – 1486, kai WCE turėjo mažiausiai iš visų trijų reguliavimo technikų – 1094.

3.3.4. „RawNet“ identifikavimo modelis

Šis identifikavimo modelis skiriasi nuo kitų tuo, kad nenaudoja „*ResNet*“ stiliau modulių. Vietoj jų jis naudoja paparastą blokų strukūrą, kuri susideda iš daugybės sluoksnių, įskaitant konvoliucijos, normalizacijos ir pilnai sujungtus sluoksnius. Panašiai kaip ir prieš tai buvę modeliai, jis susideda iš 1D konvoliucios sluoksnio su „*ReLU*“ aktyvacijos funkcija ir rinkio normalizavimu, tuomet yra „*RawBlock*“ blokai, kurių kiekvienas naudoja savo konvoliucijos ir rinkinio normalizavimo sluoksnius. Taip pat ir globalus vidurkio sluoksnis, skirtas vidurkiui apskaičiuoti per visą kanalo ašį ir trys pilnai sujungti sluoksniai.

10 lentelė. „*RawNet*“ modelio rezultatai priklausomai nuo „*RawBlock*“ blokų skaičiaus

„*RawNet*“ blokų skaičius	**Vystymo poaibio tikslumas**	**Vertinimo poaibio tikslumas**	**EER**	**Bendras modelio parametrų skaičius**
1	94,26%	88,83%	7,76%	84122
2	96,05%	87,79%	11,68%	170618
3	88,56%	86,78%	12,33%	378170
4	74,92%	73,74%	**4,18%**	931514

Eksperimentai su „*RawNet*“ identifikavimo modelio blokų skaičiumi parodė skirtingus rezultatus su įvariu blokų skaičiumi (11 lentelė), todėl buvo nuspręsta naudoti 4 blokus tolimesniam ekspermentavimui.

Modelio informacija:

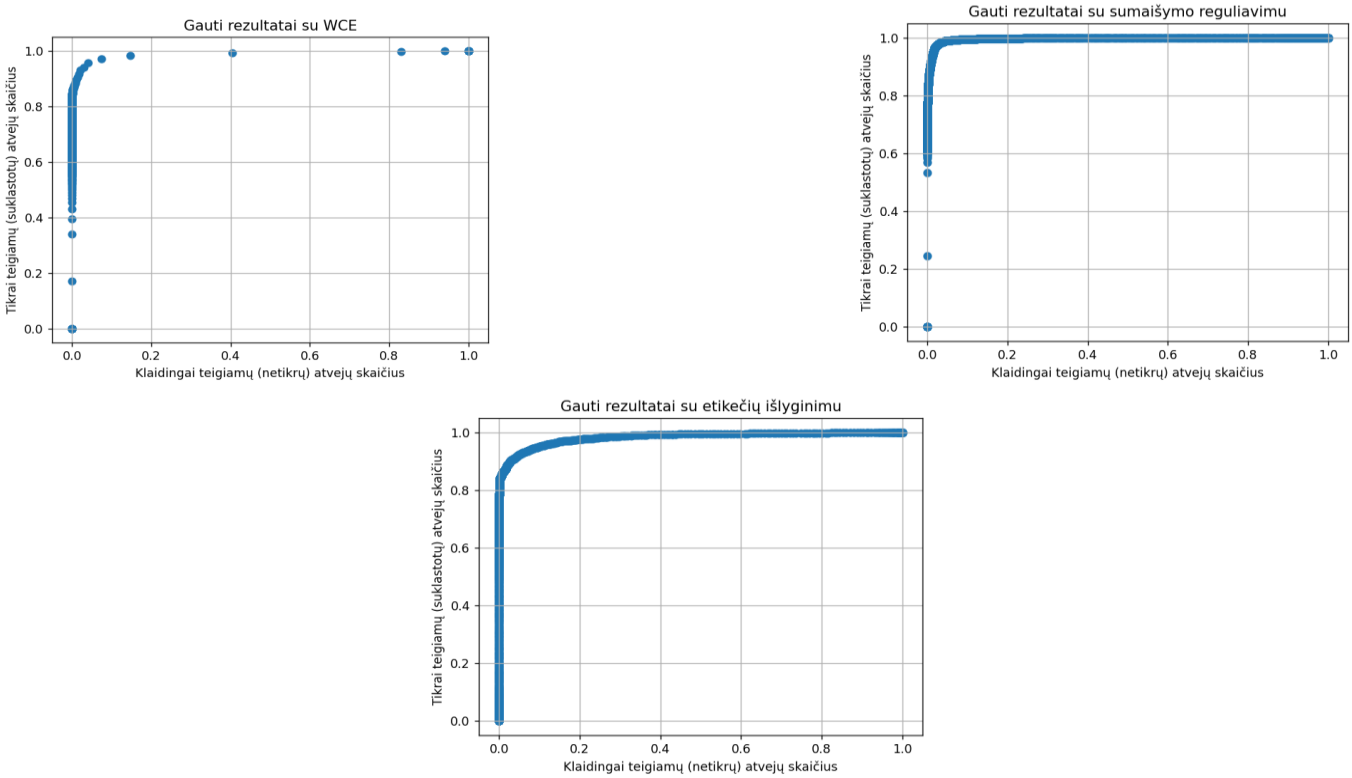
- „*RawNet*“ blokų skaičius: 4.
- Naudojami požymiai apmokymui: laiko srities garso signalai.
- Aktyvacijos funkcija: „*ReLU*“.

11 lentelė. Gauti rezultati su „*RawNet*“ modeliu

Reguliavimo technika	**Tikslumas**	**EER vertė**	**Preciziškumas**	**Atšaukimas**	**F1 įvertis**	**AUC įvertis**

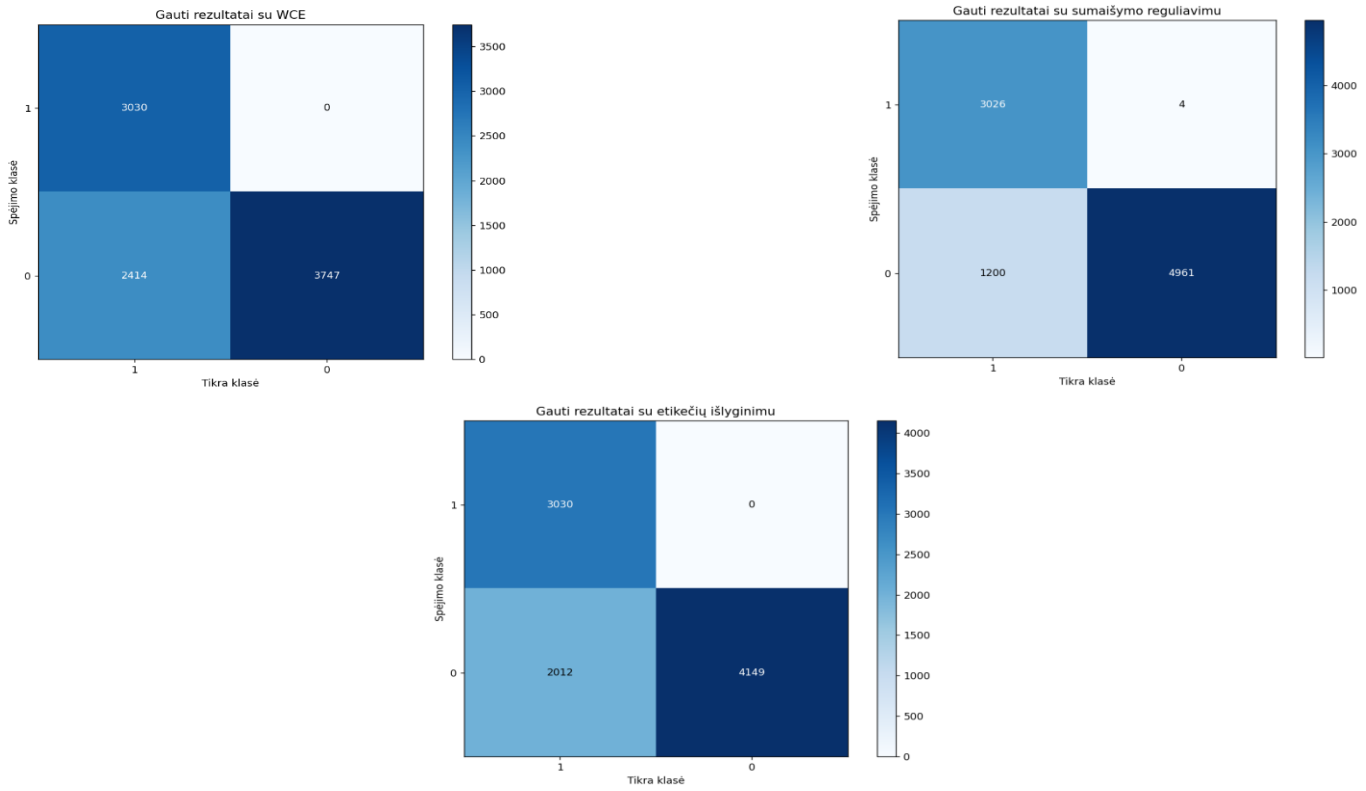
WCE	73,73%	4,18%	1,0	0,6	0,75	0,8
Sumaišymo reguliavimas	86,9%	2,47%	0,99	0,8	0,89	0,9
Etikečių išlyginimas	78,1%	6,61%	1,0	0,67	0,8	0,83

Gautos ROC kreivės:



3.7 pav. Gautos ROC kreivės su „*RawNet*“ modeliu

Sutrikimo matricos:



3.8 pav. „*RawNet*“ sutrikimo matricos

Apibendrinant „*RawNet*“ modelio rezultatus buvo gauti geriausi rezultatai naudojant sumaišymo reguliavimo techniką (11 lentelė), tikslumas 86,9% ir EER rodiklis 2,47%. Nors ir modelis užfiksavo 4 kladingai teigiamus atvejus, bet nustatė nedaug klaidingai neigiamų atvejų – 1200 (3.8 paveikslėlis). Palyginimui, naudojant WCE ir etikečių išlyginimo reguliavimo technikas modelis nustatė po 0 klaidingai teigiamų atvejų, bet daugiau kladingai neigamų atvejų, 2414 ir 2012, tai taip pat parodo ir gauti žemesni atšaukimo skaičiai, 0,6 ir 0,67.

3.3.5. „2D-Inc-TSSDNet“ identifikavimo modelis

Šis modelis buvo sukurtas kaip „*Inc-TSSDNet*“ pakeitimas, kad galėtų priimti CQT požymius, išskirtus iš garso įrašų, todėl ir architektūra šio modelio nedaug skiriasi. Naudojama „*DilatedCovModule*“ klasė, kurią sudaro keturi 2D konvoliucijos sluoksniai, kurių kiekvienas turi skirtingą dilatacijos lygį, taip pat yra keturi rinkinio normalizavimo sluoksniai, kiekvieną jų susietas su atitinkamu konvoliucijos sluoksniu. Tuomet yra „*DilatedNet*“ klasė, ją sudaro 2D konvoliucijos sluoksnis su „*ReLU*“ aktyvacijos funkcija, rinkinio normalizavimu, yra „*DilatedCovModule*“ moduliai ir trys pilnai sujungti sluoksniai. Modelio parametrų skaičius 256146.

Kaip ir su paprastu „*Inc-TSSDNet*“, taip ir su šiuo nuspręsta keisti „*DilatedCovModule*“ modulų skaičių:

12 lentelė. „*2D-Inc-TSSDNet*“ modelio rezultatai priklausomai nuo „*DilatedCovModule*“ skaičiaus

„*DilatedCovModule*“ modulių skaičius	**Vystymo poaibio tikslumas**	**Vertinimo poaibio tikslumas**	**EER**	**Bendras modelio parametrų skaičius**
2	71,61%	67,03%	37,58%	53010
3	74,33%	73,34%	16,91%	108434
4	81,49%	79,95%	17,72%	256146
5	94,91%	86,42%	**9,85%**	403858

Iš gautų rezultatų nuspręsta naudoti 5 „*DilatedCovModule*“ modulius savo modelio architektūroje.

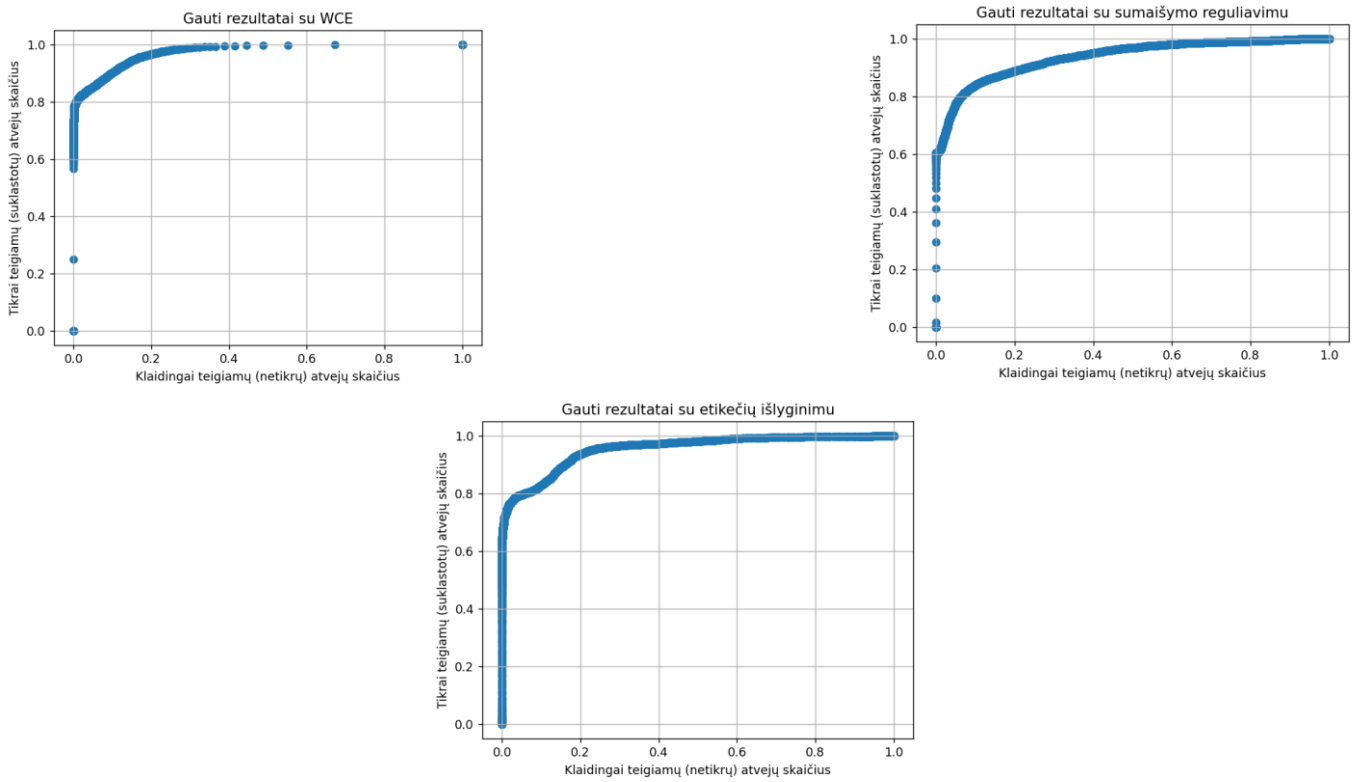
Modelio informacija:

- „*DilatedCovModule*“ modulių skaičius: 5.
- Naudojami požymiai apmokymui: CQT.
- Aktyvacijos funkcija: „*ReLU*“.

13 lentelė. Gauti rezultatai su „*2D-Inc-TSSDNet*“ modeliu

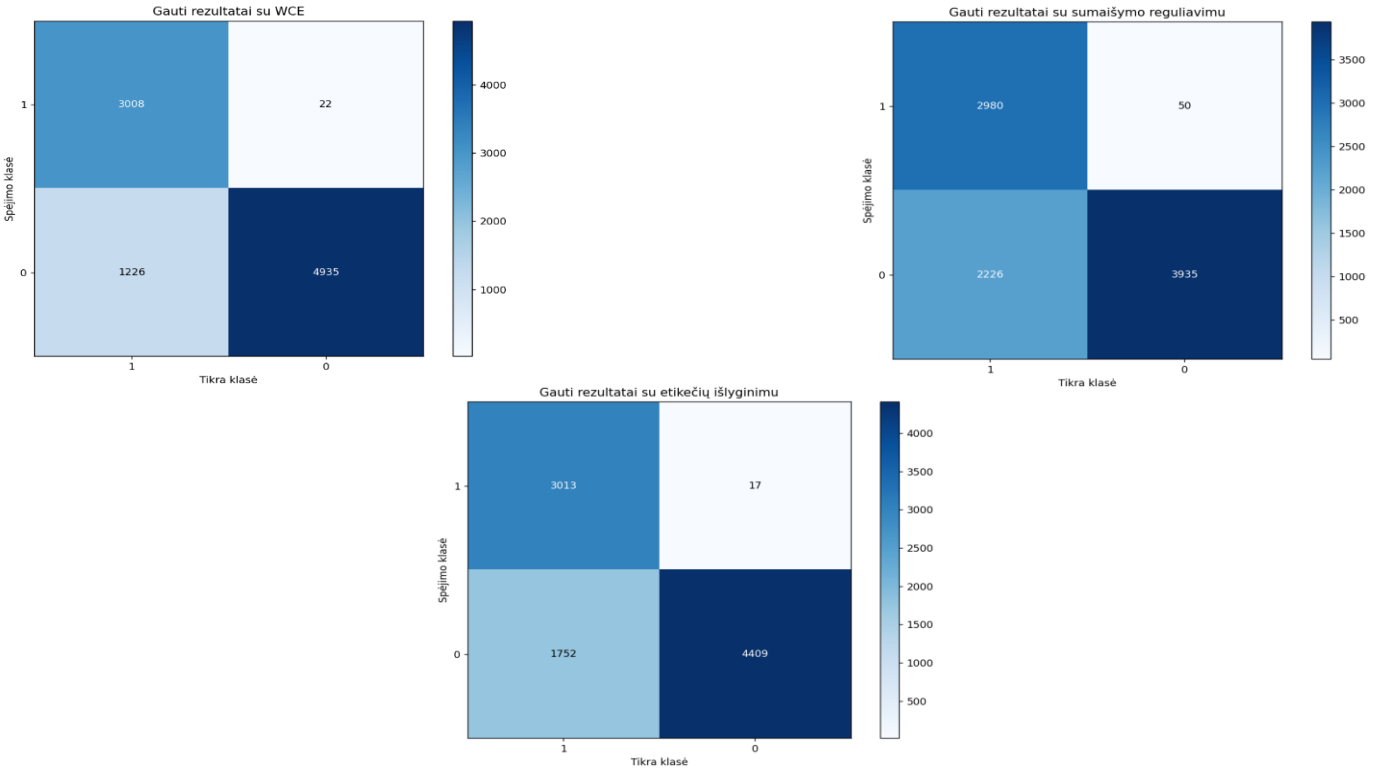
Reguliavimo technika	**Tikslumas**	**EER vertė**	**Preciziškumas**	**Atšaukimas**	**F1 įvertis**	**AUC įvertis**
WCE	86,42%	9,85%	0,99	0,8	0,88	0,89
Sumaišymo reguliavimas	75,23%	13,78%	0,98	0,63	0,77	0,81
Etikečių išlyginimas	80,75%	13,29%	0,99	0,71	0,83	0,85

Gautos ROC kreivės:



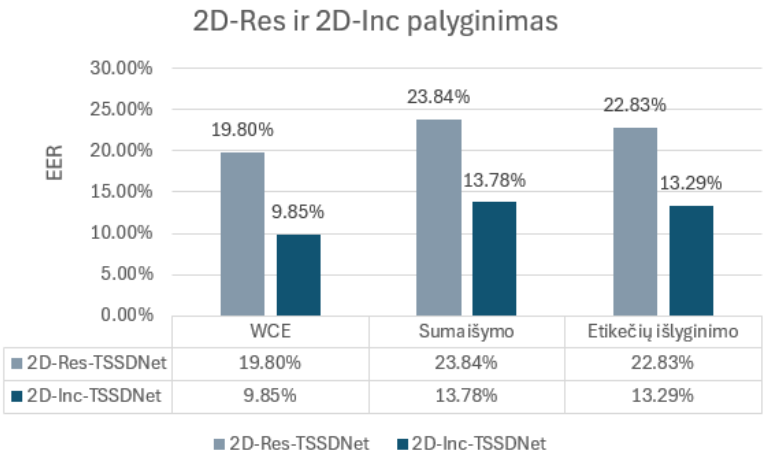
3.9 pav. Gautos ROC kreivės su „*2D-Inc-TSSDNet*“ modeliu

Gautos sutrikimo matricos:



3.10 pav. Gautos sutrikimo matricos su „*2D-Inc-TSSDNet*“ modeliu

Eksperimentuojant su šiuo modeliu geriausi rezultatai gauti su WCE reguliavimo technika, tikslumas 86,42%, EER 9,85% (13 lentelė). Sumaišymo reguliavimo technika pademonstravo kiek prastesnį tikslumą lyginant su etikečių išlyginimo 75,23% prieš 80,75% ir taip pat užfiksavo daugiau klaidingai neigiamų ir klaidingai tegiamų pavyzdžių (paveikslėlis 3.10). Taip pat šis modifikuotas identifikavimo modelis buvo palygintas su „*2D-Res-TSSDNet*“ modeliu (paveikslėlis 3.11), rezultatai parodė, kad „*2D-Inc-TSSDNet*“ užfiksavo žemesnį EER rodiklį su visom reguliavimo technikom.



3.11 pav. „*2D-Res-TSSDNet*“ ir „*2D-Inc-TSSDNet*“ modelių palyginimas

3.3.6. Kombinuotas „Inc-TSSDNet“ su *„RawNet“ identifikavimo modelis*

Siekiant pagerinti gautus rezultatus buvo sugalvota sujungti „*Inc-TSSDNet*“ ir „*RawNet*“ aptikimo modelius, nes, kaip parodė gauti eksperimentiniai rezultatai atliekant atskirai su šiais modeliais, šie modeliai parodė geriausius rezultatus. Iš tikrųjų kelių aptikimo metodų derinimas sintetinių balso įrašų identifikavimui gali būti naudingas, nes tai gali padėti pagerinti visos aptikimo ar identifikavimo sistemos patikimumą ir tikslumą. Suderinta sistema gali būti atsparesnė įvairioms sintezavimo technikoms ar vengimo metodams, kuriuos naudoja sintetinų įrašų kūrėjai. Jei vienu metodu nepavyksta identifikuoti sintetinio įrašo dėl vengimo, kitas metodas vis tiek gali būti veiksmingas ją nustatant. Šis kombinuotas aptikimo modelis yra kombinuotas modelis, jungiantis „*RawNet*“ ir „*DilatedNet*“ išvestis, o darydamas išvadas jis apskaičiuoja abiejų tinklų prognozes ir sujungia jas išilgai kanalo dimensijos, kad sudarytų galutinę kombinuotą išvestį. Nuspręsta pamažinti „*DilatedCovModule*“ modulių skaičių iki 4. Bendras modelio parametrų skaičius: 470762.

Modelio informacija:

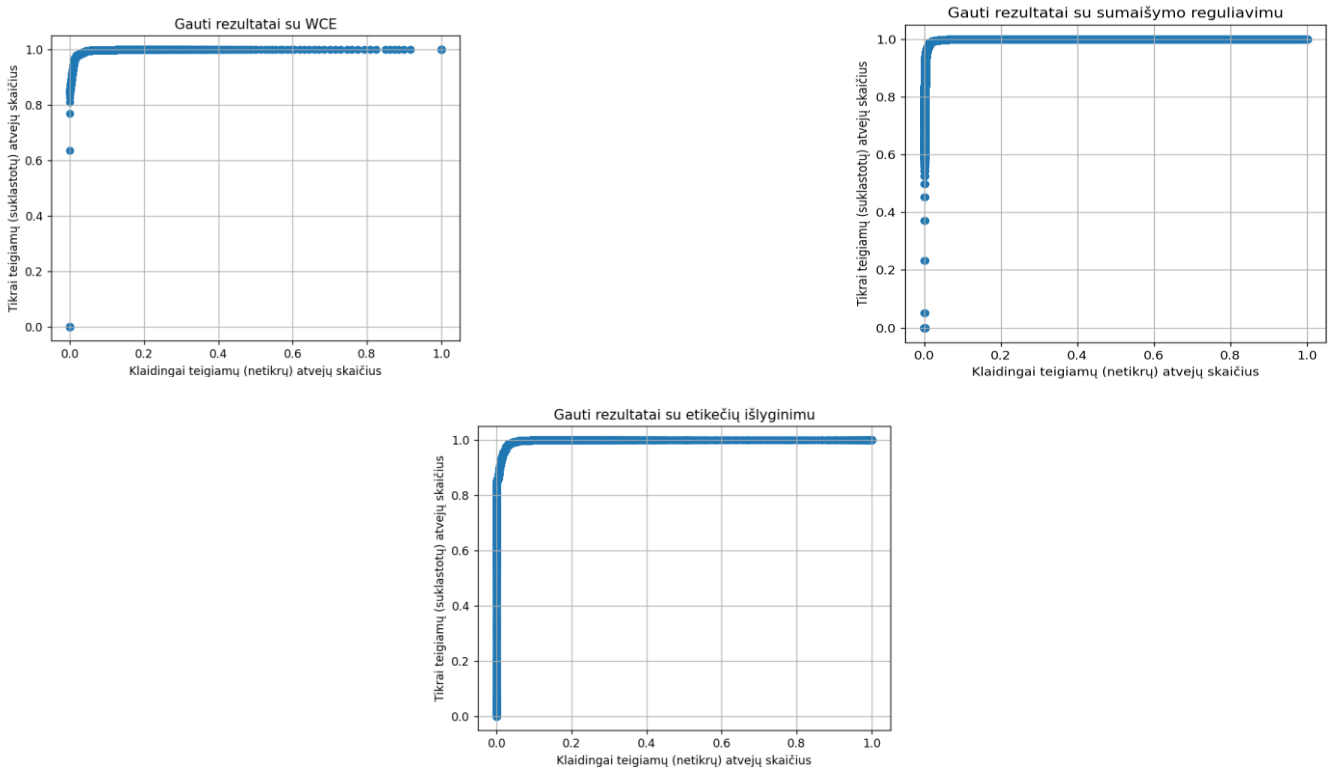
- „*DilatedCovModule*“ modulių skaičius: 4.
- „*RawNet*“ blokų skaičius: 3.

- Naudojami požymiai apmokymui: laiko srities signalai.
- Aktyvacijos funkcija: „*ReLU*“.

14 lentelė. Gauti rezultatai su „*Inc-TSSDNet*“ ir „*RawNet*“ kombinuotu modeliu

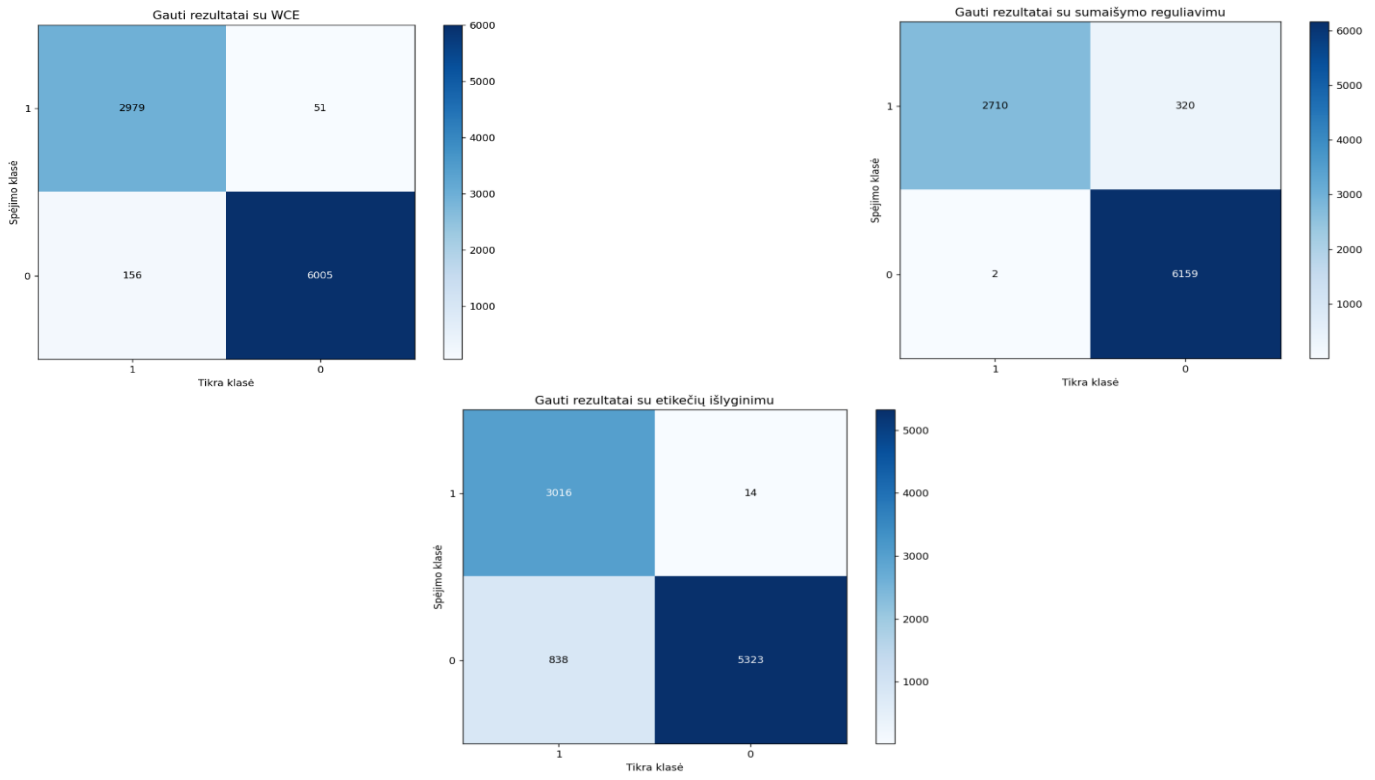
Reguliavimo technika	Tikslumas	EER vertė	Preciziškumas	Atšaukimas	F1 įvertis	AUC įvertis
WCE	97,74%	2,1111%	0,99	0,97	0,98	0,97
Sumaišymo reguliavimas	96,49%	1,35%	0,95	0,99	0,97	0,94
Etikečių išlyginimas	90,73%	2,51%	0,99	0,86	0,92	0,92

Gautos ROC kreivės:



3.12 pav. Gautos ROC kreivės su kombinuotu modeliu

Gautos sutrikimo matricos su kombinuotu identifikavimo modeliu:



3.13 pav. Gautos sutrikimo matricos su kombinuotu identifikavimo modeliu

Su šiuo identifikavimo modeliu buvo pasiekti geriausi ir aukščiausi rezultatai eksperimente. 96,49% tikslumo ir 1,35% EER rodikliai pasiekti naudojant sumašymo reguliavimo techniką, tai žemiausias EER rodiklis lyginant su kitais metodais. Aukščiausias tikslumas buvo 97,74% pasiektas su WCE technika (14 lentelė). Taip pat su sumaišymo reguliavimo technika pavyko užfiksuoti žemesnį kladingai negiamų atvejų skaičių, 2 prieš 320 klaidingai teigiamus (paveikslėlis 3.13).

3.4. Išmetimo (angl. *dropout*) sluoksnių pritaikymas identifikavimo modeliuose

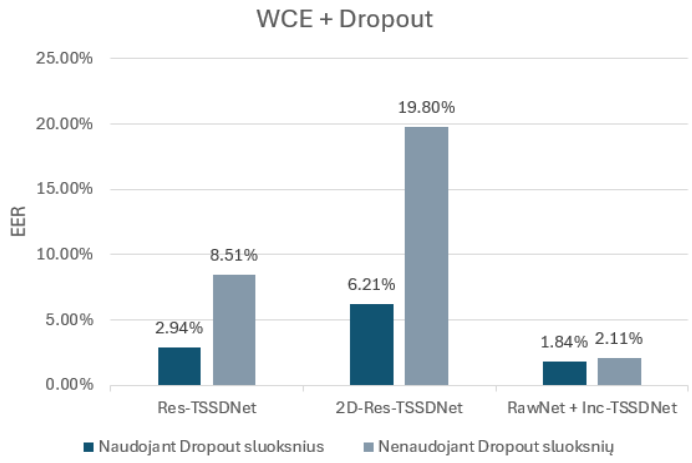
Išmetamieji sluoksniai gali būti naudingi sintetinių balso įrašų identifikavimo modeliuose, kaip ir kitų tipų neuroniniuose tinkluose. Tačiau jų veiksmingumas gali skirtis priklausomai nuo tokių veiksnių, kaip modelio architektūra, duomenų rinkinio sudėtingumas ir konkrečios analizuojamų garso signalų savybės. Išmetamieji sluoksniai veikia kaip reguliarizacijos metodas, padedantis išvengti per didelio prisitaikymo, nes mokymo metu atsitiktinai nustatoma nulinė įvesties vienetų dalis. Tai skatina tinklą išmokti patikimesnių funkcijų, kurios geriau apibendrinamos nematytiems duomenims. Tai gali padėti geriau apibendrinti išmoktus požymius, o tai labai svarbu norint identifikuoti sintetinius garso signalus, kurie gali pasižymėti įvairiomis sintezavimo technikomis duomenų rinkiniuose.

Giliuosiuose neuroniniuose tinkluose perteklinis pritaikymas (angl. *overfitting*) paprastai atsiranda dėl to, kad tam tikri skirtingų sluoksnių neuronai daro įtaką vienas kitam. Pritaikant išmetimo sluoksnį savo identifikavimo modeliuose, neuronai nebegali pasikliauti ankstesnių ar vėlesnių sluoksnių mazgais, nes jie negali manyti, kad jie egzistuoja toje konkrečioje mokymo eigoje. 15 lentelėje paryškinti rezultatai, kuriuos pavyko pagerinti su „*dropout*“ technika:

15 lentelė. Gauti rezultatai su išmetimo sluoksniais

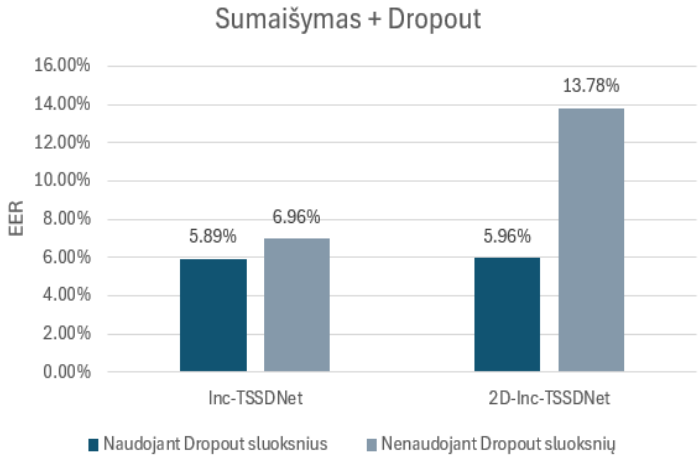
Identifikavimo modelis	**Reguliavimo technika**	**Tikslumas**	**EER vertė**	**Preciziškumas**	**Atšaukimas**	**F1 įvertis**	**AUC įvertis**
„*Inc-TSSDNet*“	WCE	**88,26%**	12,5%	0,99	0,82	0,9	**0,91**
	Sumaišymo reguliavimas	**93,7%**	**5,89%**	0,97	**0,92**	**0,95**	**0,94**
	Etikečių išlyginimas	**87,17%**	**7,19%**	0,98	**0,81**	**0,89**	**0,89**
„*2D-Inc-TSSDNet*“	WCE	85,42%	13,88%	0,95	**0,82**	0,88	0,86
	Sumaišymo reguliavimas	**81,22%**	**5,96%**	**0,99**	**0,72**	**0,83**	**0,85**
	Etikečių išlyginimas	79,51%	19,99%	0,89	**0,78**	0,83	0,79
„*Res-TSSDNet*“	WCE	**96,59%**	**2,94%**	0,99	**0,94**	**0,97**	**0,97**
	Sumaišymo reguliavimas	**86,79%**	12,04%	0,98	**0,81**	**0,89**	**0,89**
	Etikečių išlyginimas	**78,76%**	21,25%	0,88	**0,78**	**0,83**	0,78
„*2D-Res-TSSDNet*“	WCE	75,33%	**6,21%**	**1,0**	0,63	0,77	**0,81**
	Sumaišymo reguliavimas	74,18%	29,89%	0,75	**0,9**	**0,82**	0,65
	Etikečių išlyginimas	**82,45%**	17,55%	0,89	**0,83**	**0,86**	0,82
„*RawNet*“	WCE	73,51%	11,16%	1,0	0,6	0,75	0,8
	Sumaišymo reguliavimas	**89,31%**	2,53%	0,99	**0,84**	**0,91**	**0,91**
	Etikečių išlyginimas	**86,82%**	**4,58%**	0,99	**0,8**	**0,89**	**0,9**
Kombinuotas „*RawNet2*“ ir „*Inc-TSSDNet*“	WCE	88,93%	**1,84%**	**1,0**	0,83	0,91	0,91
	Sumaišymo reguliavimas	90,57%	2,07%	0,99	0,86	0,92	0,92
	Etikečių išlyginimas	74,94%	14,94%	0,99	0,62	0,77	0,81

Žemiau yra pateikta diagrama palygnimui, kiek skiriasi rezultatai naudojant išmetamuosius sluoksnius su WCE reguliavimo technika ir nenaudojant išmetamųjų sluoksnių, joje vaizduojama, kad pavyko pagerinti rezultatus su „*Res-TSSDNet*“, „*2D-Res-TSSDNet*“ ir kombinuotu „*RawNet*“ ir „*Inc-TSSDNet*“ modeliais ir pasiekti žemesnį EER rodiklį:

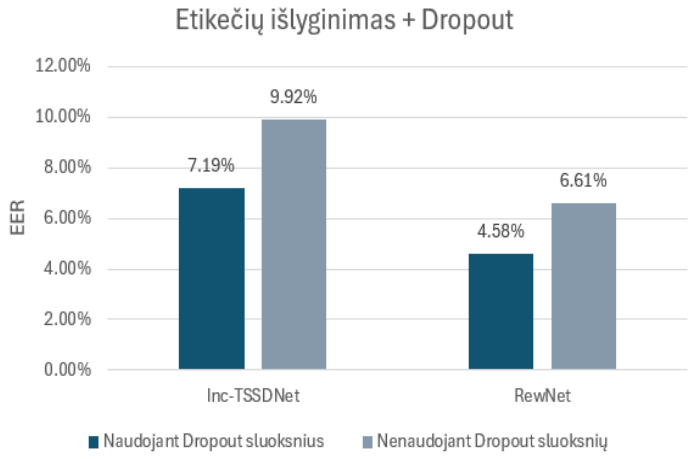


3.14 pav. WCE ir išmetamųjų sluoksnių kombinavimo palyginimas

Taip pat pateikti vizualiniai palyginimai ir naudojant sumaišymo reguliavimo (paveikslėlis 3.15) ir etikečių išlyginimo (paveikslėlis 3.16) reguliavimo technikas su išmetimo sluoksniais:



3.15 pav. Sumaišymo ir išmetamųjų sluoksnių kombinavimo palyginimas



3.16 pav. Etikečių išlyginimo ir išmetamųjų sluoksnių kombinavimo palyginimas

Pritaikius išmetamuosius sluoksnius ir pastebėjus pagerėjimą, galima daryti prielaidą, kad modelis galėjo būti per daug pritaikytas arba jautrus triukšmui, o išmetimo reguliavimas padėjo kažkiek išspręsti šias problemas, todėl garso sintetinių balso įrašų identifikavimo užduotis buvo atlikta geriau ir rezultatai tam tikrais atvejais pagerinti.

3.5. SMOTE technikos pritaikymas eksperimente

Sintetinės mažumos padidinimo technika (angl. *Synthetic Minority Over-sampling Technique*), sutumpintai (SMOTE) – tai išankstinio apdorojimo metodas, naudojamas siekiant pašalinti klasių disbalansą duomenų rinkinyje sugeneruojant sintetinius duomenis. Nors SMOTE dažnai naudojamas vaizdų ar lentelių duomenų klasifikavimo užduotims atlikti, jis taip pat gali būti taikomas garso duomenims, įskaitant sintetinio balso identifikavimo užduotis. Savo eksperimente nuspręsta SMOTE techniką panaudoti tik ant apmokymo dalies savo duomenų rinkinio, nes šios technikos panaudojimas su vystymo ir vertinimo rinkiniais gali iškreipti rezultatus. Šie rinkiniai turi atspindėti nematytus pasaulio duomenis, o sintetinės mažumos padidinimo technika gali iškrepti šią reprezentaciją ir lemti pernelyg optimistinius rezultatus.

16 lentelė. Apmokymo poaibis prieš ir po SMOTE technikos pritaikymo

Poaibis	**Balso įrašai**		
	Nesintetiniai	**Sintetiniai**	**Iš viso įrašų**
Apmokymo	1 024	1 977	3 001
Apmokymo po SMOTE pritaikymo	1977	1977	3954

Taikant SMOTE mano atveju buvo įkeltas apmokymo duomenų rinkinio poaibis, iš jo išrinkti garso mėginiai ir žymėjimai, tada inicializuotas SMOTE objektas. Tuomet pasinaudojus SMOTE objekto fit_resample() metodu buvo padidinta mažumos klasė, tai yra nesintetinių įrašų skaičius iki 1977, kad sulyginti jų skaičių (16 lentelė). SMOTE technika generuoja sintetninius pavyzdžius kaip plokščius požymių vektorius, remiantis požymių erdvės interpoliacija, todėl jie atspindi charakteristikas panašias į originalių įrašų.

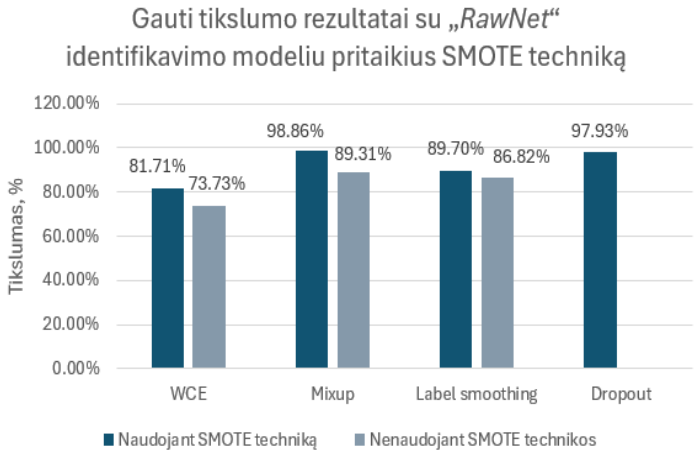
Pasirinkta apmokyti identifikavimo modelius, kurie kaip požymius naudoja laiko srities garso įrašus: „*Res-TSSDNet*“, „*Inc-TSSDNet*“, pritaikytas „*RawNet*“, kombinuotas „*Inc-TSSDNet*“ ir „*RawNet*“ modelis.

17 lentelė. Gauti rezultatai pritaikant SMOTE techniką

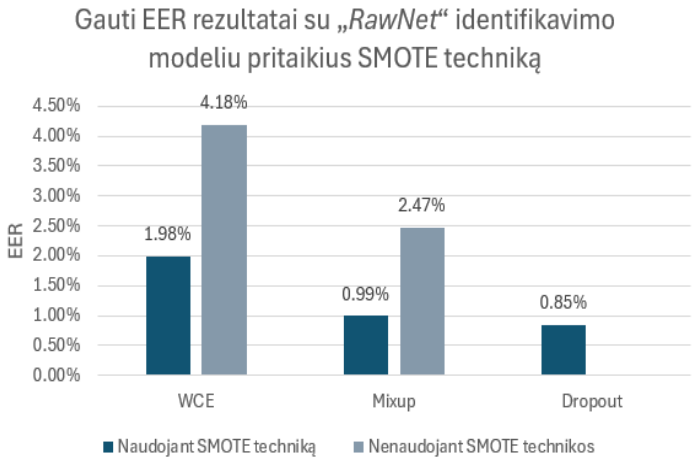
Identifikavimo modelis	**Reguliavimo technika**	**Tikslumas**	**EER vertė**	**Preciziškumas**	**Atšaukimas**	**F1 įvertis**	**AUC įvertis**
„*Inc-TSSDNet*“	WCE	**92,43%**	**4,65%**	0,99	**0,89**	**0,94**	**0,94**
	Sumaišymo reguliavimas	84,71%	9,2%	0,99	0,77	0,87	0,88
	Etikečių išlyginimas	85,11%	10,67%	0,99	0,78	0,87	0,88
	Nenaudojant reguliavimo priemonių	85,57%	5,69%	0,99	0,78	0,87	0,89
„*Res-TSSDNet*“	WCE	75,8%	9,32%	**1,0**	0,63	0,77	0,81
	Sumaišymo reguliavimas	85,13%	8,09%	0,99	0,77	0,87	0,88
	Etikečių išlyginimas	**79,4%**	**14,9%**	0,98	0,7	0,82	**0,84**
	Nenaudojant reguliavimo priemonių	84,32%	9,81%	0,99	0,76	0,86	0,88
„*RawNet*“	WCE	**81,71%**	**1,98%**	0,99	**0,72**	**0,84**	**0,86**
	Sumaišymo reguliavimas	**98,86%**	**0,99%**	0,99	**0,98**	**0,99**	**0,99**
	Etikečių išlyginimas	**89,7%**	9,82%	0,98	**0,86**	**0,91**	**0,91**
	Nenaudojant reguliavimo priemonių	97,93%	0,85%	0,97	0,99	0,98	0,96
Kombinuotas „*RawNet2*“ ir „*Inc-TSSDNet*“	WCE	78,18%	6,52%	0,99	0,67	0,8	0,83
	Sumaišymo reguliavimas	89,43%	2,47%	0,99	0,84	0,91	0,92
	Etikečių išlyginimas	**95,02%**	5,27%	0,94	**0,98**	**0,96**	**0,93**
	Nenaudojant reguliavimo priemonių	83,17%	5,02%	1,0	0,74	0,85	0,87

Pritaikius SMOTE techniką buvo gauti rezultatai (17 lentelė), paryškinti rezultatai, kuriuos pavyko pagerinti. Taip pat šiame eksperimente nuspręsta atlikti papildomai bandymus nenaudojant reguliavimo priemonių, bet modelių architektūroje palikti išmetamuosius sluoksnius. Su kiekvienu identifikavimo modeliu pavyko pagerinti nors vieną rezultatą lyginant su prieš tai gautais. Tačiau geriausius rezultatus pavyko pasiekti su „*RawNet*“, pavyko pagerinti prieš tai geriausią tikslumo rodiklį 97,74%, po SMOTE pritaikymo pavyko pasiekti 98,86% su sumaišymo reguliavimo technika. Tai pat pagerintas EER rodiklis, prieš SMOTE pritaikymą 1,35% su kombinuotu modeliu, pritaikius SMOTE pasiektas 0,85% su „*RawNet*“ modeliu naudojant tik išmetimo sluoksnius.

3.17 ir 3.18 pav. parodo, kiek pavyko pagerinti tikslumo ir EER rezultatus naudojant SMOTE techniką su „*RawNet*“ identifikavimo modeliu.



3.17 pav. Gauti tikslumo rezultatai su „*RawNet*“ identifikavimo modeliu pritaikius SMOTE techniką



3.18 pav. Gauti EER rezultatai su „*RawNet*“ identifikavimo modeliu pritaikius SMOTE techniką

Buvo patikrinta, ar SMOTE veiksmingai išsprendė klasių disbalanso problemą duomenų rinkinyje. Išnagrinėjus klasių pasiskirstymą prieš taikant SMOTE ir pritaikius SMOTE, nustatyta, kad disbalansas buvo sušvelnintas ir sintetinės imtys sukurtos tinkamai. Tai pat reikia suprasti, kad tai yra sintetiškai sukurti duomenys, pernelyg didelis pasikliovimas sintetiniais pavyzdžiais, neatsižvelgiant į jų įtaką apibendrinimui, gali lemti prastus rezultatus atliekant bandymus su realiais duomenimis.

3.6. Stratifikuotas kryžminis tikrinimas

Kryžminis tikrinimas (angl. *cross-validation*) yra mašininio mokymosi metodas, naudojamas modelio veikimui įvertinti. Tai ypač naudinga norint įvertinti, kaip gerai modelis apibendrins nematytus apmokymo metu duomenis. Pagrindinė kryžminio patvirtinimo idėja yra padalyti duomenų rinkinį į kelis poaibius, vieną naudoti kaip apmokymo poaibį, o kitą modelio vertinimui. Šis procesas yra kartojamas kelis kartus, kiekvieną kartą naudojant skirtingus poaibius. Kiekvieno vertinimo etapo rezultatai apskaičiuojami, kad būtų galima tiksliau įvertinti modelio našumą. Šis tikrinimas turi vieną pagrindinį apribojimą, jis gali sukurti mokymo ir vertinimo poaibius, kurie neatspindi bendro klasių pasiskirstymo, ypač tais atvejais, kai duomenys yra nesubalansuoti. Esant tokiems scenarijams gelbsti stratifikuotas kryžminis tikrinimas.

Stratifikuotas kryžminis tikrinimas (angl. *stratified cross-validation*) yra metodas, naudojamas modelio našumui įvertinti, kartu užtikrinant, kad kiekvienas kryžminio tikrinimo dalis išsaugo tą patį klasių pasiskirstymą kaip ir pradinis duomenų rinkinys. Tai ypač naudinga dirbant su nesubalansuotais duomenų rinkiniais, kai viena klasė gali būti daug labiau paplitusi nei kita.

Savo eksperimente nuspręsta stratifikuotą kryžminį tikrinimą atlikti su apmokymo ir vertinimo poaibiais. Taip pat nuspręsta sukurti papildomus keturis papildomus apmokymo ir vertinimo duomenų poaibius, kad bendroje sumoje su 3.3 skyriuje gautais rezultatai gautųsi penki skirtingi poaibių variantai, kas padidina eksperimentų skaičių ir suteikia daugiau įvairesnių įžvalgų, apmokymo dalyje taip pat išlaikant pirminę eksperimento sąlygą, kad apmokyti su „*VITS*“ teksto į kalbą technika, o vertinti su „*VITS*“ ir „*LIEPA TTS*“, kad nustatyti, kaip modeliai klasifikuos apmokymo metu nepanaudotus balso įrašus. 18 lentelėje pavaizduoti vidurkių rezultatai apskačiuoti iš 5 dalių rezultatų. Taip pat paskaičiuoti modelių tikslumo rezultatų vidurkiai kiekvienoje dalyje (žr. 2 priedą).

18 lentelė. Apskaičiuoti 5 dalių kryžminio tikrinimo vidurkiai pagal naudojamas reguliavimo technikas

Identifikavimo modelis	Reguliavimo technika	Tikslumas	EER vertė	Preciziškumas	Atšaukimas	F1 įvertis	AUC įvertis
„*Inc-TSSDNet*“	WCE	79,83%	12,6%	0,96	0,72	0,82	0,83
	Sumaišymo reguliavimas	84,53%	6,87%	0,96	0,8	0,87	0,86
	Etikečių išlyginimas	80,33%	11.12%	0,97	0,72	0,82	0,84
„*2D-Inc-TSSDNet*“	WCE	89,01%	7,63%	0,97	0,85	0,91	0,9
	Sumaišymo reguliavimas	81,8%	10,75%	0,96	0,75	0,84	0,85
	Etikečių išlyginimas	83,72%	10,16%	0,96	0,78	0.86	0,86
„*Res-TSSDNet*“	WCE	77,45%	16,9%	0,96	0,67	0,79	0,8
	Sumaišymo reguliavimas	87,86%	7,81%	0,95	0,87	0,9	0,88
	Etikečių išlyginimas	78,8%	14,25%	0,95	0,71	0,81	0,82
„*2D-Res-TSSDNet*“	WCE	86,2%	8,79%	0,93	0,85	0,88	0,86
	Sumaišymo reguliavimas	82,25%	13,23%	0,88	0.86	0,85	0,8
	Etikečių išlyginimas	82,02%	10,97%	0,92	0,82	0,85	0,82
„*RawNet*“	WCE	82,05%	8,42%	0,97	0,75	0,84	0,85
	Sumaišymo reguliavimas	85,43%	**4,29%**	0,96	0,81	0,87	0,87
	Etikečių išlyginimas	84,88%	7,72%	0,97	0,8	0,87	0,87
Kombinuotas „*RawNet*“ ir „*Inc-TSSDNet*“	WCE	84,57%	6,53%	0,98	0,78	0,86	0,87
	Sumaišymo reguliavimas	**91,99%**	4,9%	0,96	0,92	0,93	0,91
	Etikečių išlyginimas	83,39%	7,7%	0,97	0,77	0,85	0,86

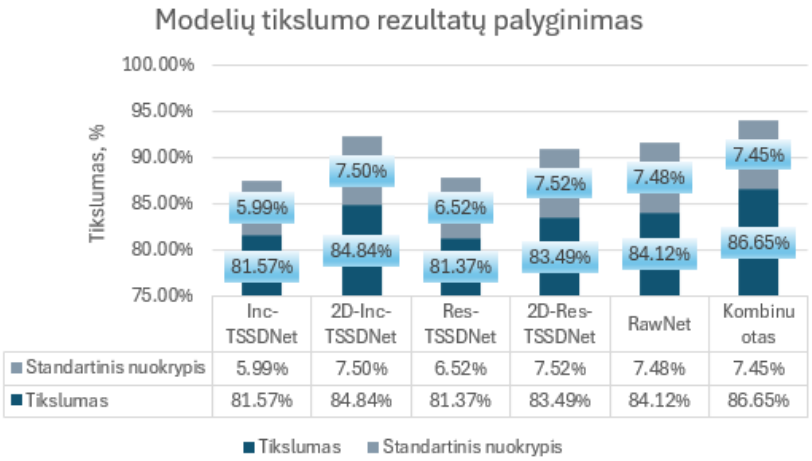
Gavus rezultatus nuspręsta apskaičiuoti standartinį nuokrypį. Atliekant stratifikuotą kryžminį tikrinimą standartinis nuokrypis gali suteikti vertingų įžvalgų apie modelio našumo nuoseklumą arba stabilumą. Standartinio nuokrypio formulė, pagal kurią atliekami skaičiavimai:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

Nuspręsta apskaičiuoti kiekvieno identifikavimo modelio vidurkio rezultatą nepriklausomai, kokia reguliavimo technika panaudota (19 lentelė), tikslumo ir EER su standartiniais nuokrypiais palyginimus galima matyti 3.19 ir 3.20 paveikslėliuose.

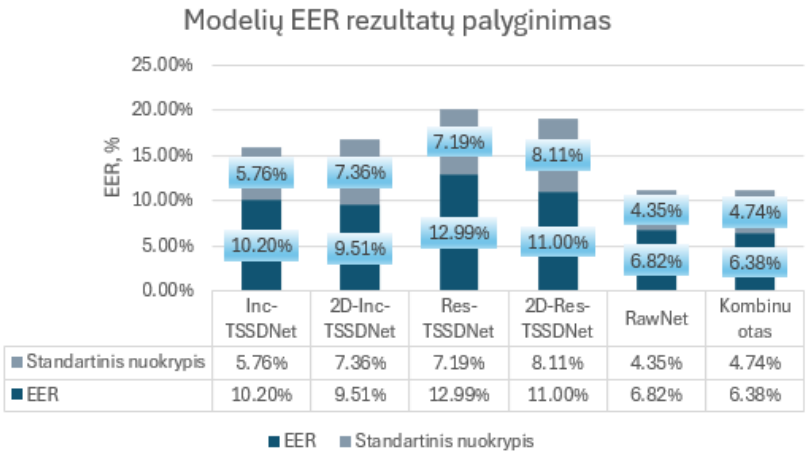
19 lentelė. Identifikavimo modelių tikslumo ir EER rezultatų vidurkiai

Identifikavimo modelis	Tikslumas	EER vertė
„*Inc-TSSDNet*“	81,57%	10,20%
„*2D-Inc-TSSDNet*“	84,84%	9,51%
„*Res-TSSDNet*“	81,37%	12,99%
„*2D-Res-TSSDNet*“	83,49%	11,00%
„*RawNet*“	84,12%	6,82%
Kombinuotas „*RawNet*“ ir „*Inc-TSSDNet*“	**86,65%**	**6,38%**



3.19 pav. Identifikavimo modelių tikslumo rezultatų palyginimas po vidurkių apskaičiavimo

Iš 3.19 ir 3.20 paveikslėlių galima matyti, kad aukčiausi tikslumo ir geriausi EER rezultatai buvo pasiekti su kombinuotu „*RawNet*“ ir „*Inc-TSSDNet*“ identifikavimo modeliu.



3.20 pav. Identifikavimo modelių EER rezultatų palyginimas po vidurkių apskaičiavimo

3.7. Klonuotų balso įrašų identifikavimas

Balso klonavimas reiškia sintetinės kalbos, imituojančios konkretaus asmens balsą, generavimo procesą. Tai pasiekiama naudojant įvairius metodus, įskaitant giliu mokymusi pagrįstus modelius. Balso klonų nustatymas yra svarbus dėl kelių priežasčių. Balso klonavimo technologija gali būti piktnaudžiaujama piktybiniais tikslais, pavyzdžiui, apsimetinėjimu, sukčiavimu arba netikrų garso įrašų kūrimu šantažo ar dezinformacijos tikslais. Balso klonų nustatymas padeda aptikti ir užkirsti kelią tokiai nesąžiningai veiklai. Taip pat autentifikavimui, programose, kuriose balsas naudojamas autentifikavimo arba patvirtinimo tikslais, labai svarbu užtikrinti, kad pateikiamas balsas būtų tikras, o ne sintetinis klonas. Nors balso klonavimo technologija suteikia daug privalumų, ji taip pat kelia iššūkių ir pavojų, kuriuos reikia spręsti taikant patikimus identifikavimo ir tikrinimo būdus. Dėl šios priežasties iš savo surinkto duomenų rinkinio buvo panaudotas balso įrašas ir pagal jį sukurta 15 klonuotų balso įrašų naudojantis „*DupDub*“ įrankiu. Didžiausias trūkumas, kad šis įrankis nepalaiko konkrečiai lietuvių kalbos ir girdimas nelietuviškas kirčiavimas, tačiau klonuotas balso įrašas išlaiko tembro savybes panašias į tikro balso įrašo. Bandymams panaudota minėtų 15 klonuotų balso įrašų ir 15 tikrų to paties balso įrašų. Šie 30 įrašų buvo tik vertinami, nepanaudojant jų modelių apmokymui, kad pamatyti, kaip modeliai suklasifikuos šiuos nematytus įrašus. O apmokyti modeliai ant surinkto savo duomenų rinkinio prieš tai atliktuose eksperimentuose.

20 lentelė. Klonuotų ir nesintetinių balso įrašų analizės rezultatai

Identifikavimo modelis	Tikrai nesintetiniai	Kladingai nesintetiniai	Tikrai sintetiniai	Klaidingai sintetiniai
„*Inc-TSSDNet*“	15	0	1	14

„*2D-Inc-TSSDNet*“	15	0	0	15
„*Res-TSSDNet*“	15	0	2	13
„*2D-Res-TSSDNet*“	14	1	6	9
„*RawNet*“	13	2	2	13
Kombinuotas „*RawNet*“ ir „*Inc-TSSDNet*“	12	3	7	8

Kaip ir galima buvo tikėtis, identifikavimo modeliai parodė neblogas apibendrinimas galimybes atskiriant nesintetinius garso įrašus, tačiau sunkiai susitvarkė su klonuotais sintetiniais balso įrašais, nes jie nebuvo panaudoti apmokymui. Mažiausiai kartų klydo „*2D-Res-TSSDNet*“ modelis – 10 kartų. Prieš tai geriausiu rezultatus rodęs kombinuotas „*RawNet*“ ir „*Inc-TSSDNet*“ identifikavimo modelis suklydo 11 kartų.

3.8. Trečio skyriaus apibendrinimas

Atlikti eksperimentai ir jų rezultatai parodė, kad panaudoti identifikavimo modeliai gali pasiekti neblogų rezultatų identifikuojant sintetinius balso įrašus su savo surinktu duomenų rinkiniu. Duomenų rinkinyje esantys balso įrašai buvo surinkti arba susintetinti lietuvių kalba. Eksperimentams atlikti panaudoti šeši identifikavimo modeliai: „*Res-TSSDNet*“, „*Inc-TSSDNet*“, „*2D-Res-TSSDNet*“, Pritaikytas „*RawNet*“, Pritaikytas „*2D-Inc-TSSDNet*“, Kombinuotas „*Inc-TSSDNet*“ ir „*RawNet*“ modelis. Atlikti bandymai su trimis skirtingos reguliavimo priemonėmis apmokymo metu siekiant pagerinti modelio apibendrinimą, sumažinti perteklinį pritaikymą ir padidinti našumą. Eksperimente panaudota SMOTE technika klasių disbalansui pašalinti, generuojant sintetines mažumos klasės pavyzdžius. Taikant SMOTE techniką duomenų rinkinio apmokymo poaibiui ir vėliau apmokant identifikavimo modelį šis tampa mažiau šališkas ir gali veiksmingiau atskirti nesintetinius ir sintetinius balso įrašus. Taip pat atliktas stratifikuotas kryžminis tikrinimas su savo duomenų rinkiniu padalijant duomenų rinkinį į kelis poaibius ir išlaikant tokį patį klasių pasiskirstymą kiekviename poaibyje, kaip ir pradiniame duomenų rinkinyje. Atlikta nedidelė analizė su klonuotais balso įrašais.

3.9. Trečio skyriaus išvados

Apibendrinus šį skyrių suformuluojamos tokios išvados:

1. Eksperimento pradžioje atlikus bandymus pritaikytas „*2D-Inc-TSSDNet*“ identifikavimo modelis, kuris naudojo konstanta Q transformacijos požymius užfiksavo apytiksliai 9,85% geresnius EER metrikos ir 5,63% tikslumo metrikos rezultatus nei „*2D-Res-TSSDNet*“.
2. Pritaikytas „*RawNet*“ identifikavimo modelis eksperimento pradžioje apėmus trys reguliavimo technikas (WCE, sumaišymo, etikečių išlyginimas) fiksavo apytiksliai 4,42% EER rodiklį, kuris buvo antras geriausias po kombinuoto „*RawNet*“ ir „*Inc-TSSDNet*“ identifikavimo modelio.
3. Kombinuotas „*RawNet*“ ir „*Inc-TSSDNet*“ identifikavimo modelis pirmoje eksperimento dalyje iki išmetimo sluoksnių ir SMOTE technikos pritaikymo užfiksavo geriausią 97,74% tikslumą su WCE reguliavimo technika ir 1,35% EER rodiklį su sumaišymo reguliavimu.
4. Pritaikius išmetimo sluoksnius identifikavimo modelių architektūroje ir kombinuojant kartu su WCE reguliavimo technika pavyko sumažinti EER rodiklį nuo 19,8% iki 6,21% su „*2D-Res-TSSDNet*“ identifikavimo modeliu. Su „*Res-TSSDNet*“ modeliu nuo 8,51% iki 2,94%.
5. Naudojant SMOTE technika pašalinti klasių disbalansui su kombinuotu „*RawNet*“ ir „*Inc-TSSDNet*“ identifikavimo modeliu pavyko užfiksuoti žemiausią ir geriausią per visą eksperimentą 0,85% EER rodiklį nenaudojant reguliavimo technikų (WCE, sumaišymo, etikečių išlyginimo), paliekant tik išmetimo sluoksnius. Taip pat su šiuo identifikavimo modeliu užfiksuotas geriausias tikslumas 98,86% su sumaišymo reguliavimu.
6. Siekiant gauti daugiau rezultatų ir įžvalgų buvo sukurti papildomi 4 duomenų rinkinio poaibiai, kad būtų paskaičiuoti vidurkio ir standartinio nuokrypio rodikliai. Stratifikuotas tikrinimas padeda išlaikyti tolygų pavyzdžių pasiskirstymą tarp klasių. Nustatyta, kad geriausi tiek tikslumo, tiek EER vidurkio rezultatai buvo pasiekti su kombinuotu „*Inc-TSSDNet*“ ir „*RawNet*“ identifikavimo modeliu, tikslumas – 86,64%, EER – 6,38%.

Išvados

Apibendrinant galima teigti, kad sintetinių ir nesintetinių balso įrašų identifikavimas naudojant dirbtinį intelektą yra svarbi tyrimų sritis, turinti plataus masto pasekmes įvairiose srityse, tokiose kaip saugumas, skaitmeninė kriminalistika ir žmogaus ir kompiuterio sąveika. Dirbtinio intelekto sintezuotų žmonių balsų pažanga sukėlė didėjančią apsimetinėjimo ir dezinformacijos grėsmę, todėl labai svarbu sukurti metodus sintetiniams žmogaus balsams identifikuoti, kad būtų sumažintas neigiamas jų poveikis.

Siekiant pagerinti sintetinių ir nesintetinių balso įrašų identifikavimą buvo apžvelgti metodai ir technikos, kurie yra skirti sintetiniams balso įrašams generuoti, aptartas generavimo principas, tam tikri trūkumai, privalumai. Taip pat apžvelgti identifikavimui skirti algoritmai, duomenų rinkiniai, nustatyta, kad egzistuoja įvairūs mašininio ir gilaus mokymosi metodai. Tuomet buvo nuspręsta surinkti savo duomenų rinkinį lietuvių kalba panaudojant kelias tekstas į kalbą (TTS) technikas. Atrinkti trys identifikavimo modeliai, pasiūlyti taip pat trys modeliai keičiant arba kombinuojant atrinktų modelių architektūras, kad būtų pagerinti identifikavimo rezultatai. Visi eksperimentai atlikti su savo surinktu duomenų rinkiniu panaudojant reguliavimo, duomenų padidinimo technikas. Atlikus eksperimentus buvo pasiekti tokie rezultatai ir išvados:

1. Sintetinės kalbos identifikavimas lietuvių kalba yra iššūkis dirbtinio intelekto ir kalbos apdorojimo srityje. Sintetinės kalbos technologijoms toliau tobulėjant, didėja metodų skaičius, skirtų atskirti natūralią kalbą nuo dirbtinai sukurtos, poreikis įvairiose srityse, įskaitant autentifikavimą, saugumą ir turinio moderavimą. Atlikus literatūros analizę nustatyta, kad egzistuoja įvairios sintetinių balso įrašų generavimo technikos, tačiau daugiausia dėmesio šiam momentui skiriama giluminio mokymosi generavimo modeliams. Apžvelgus identifikavimo modelius išsiaiškinta, kad egzistuoja įvairios mašinio mokymosi ir gilaus mokymosi architektūros. Mašininio mokymosi algoritmai reikalauja rankinio požymių išskyrimo ir duomenų paruošimo. Eksperimentuojant su gilaus mokymosi algoritmais garso įrašams reikia specifinių transformacijų. Taip pat gilaus mokymosi modeliai gali automatiškai išmokti hierarchines reprezentacijas iš neapdorotų garso signalų, todėl nebereikia rankinio požymių išskyrimo, tačiau reikia daug duomenų ir didelių skaičiavimo išteklių. Analizuotoje literatūroje tai pat dažniausiai pasitaikydavo duomenų rinkiniai anglų kalba. Lietuvių kalba duomenų rinkinių praktiškai neegzistuoja arba jie nėra viešai prieinami.
2. Eksperimentams atlikti pasirinkti trys identifikavimo modeliai: „*Inc-TSSDNet*“, „*Res-TSSDNet*“, „*2D-Res-TSSDNet*“. Šių modelių pagrindu kaip patobulinimai buvo sukurti savo trys identifkavimo modeliai: „*RawNet*“, „*2D-Inc-TSSDNet*“, kombinuotas „*RawNet*“ ir „*Inc-TSSDNet*“. Taip pat eksperimente buvo panaudotos 4 reguliarizavimo technikos: svertinės kryžinės entropijos, sumaišymo, etikečių išlyginimo ir išmetimo sluoksnių. Šios reguliarizavimo technikos padeda

pagerinti modelių apibendrinimą ir patikimumą, ypač kai eksperimentuojama su dideliais duomenų rinkiniais ir sudėtingomis modelių architektūromis. Be to, kadangi surinktame duomenų rinkinyje vyrauja duomenų disbalansas buvo panaudota SMOTE technika. Ši technika padidina mažumos klasę sukuriant sintetinius duomenis, nuspręsta SMOTE techniką panaudoti tik ant apmokymo dalies savo duomenų rinkinio, nes šios technikos panaudojimas su kitomis rinkinio dalimis gali iškreipti rezultatus. Paskutinėje eksperimento dalyje atliktas stratifikuotas kryžminis tikrinimas, kuris suteikė daugiau rezultatų ir įžvalgų.

3. Atliekant tyrimus labai svarbu pasirinkti tinkamą duomenų rinkinį, kuris būtų sudarytas iš įvarių sintetiniams balso įrašams sugeneruoti skirtų technikų. Norint parengti tikslius ir apibendrinamus modelius, būtina surinkti pakankamai ir įvairių duomenų, apimančių tiek natūralią, tiek sintetinę kalbą įvairiuose aplinkose ir scenarijuose. Šiame magistro darbe buvo surinktas savo duomenų rinkinys lietuvių kalba, pasirinktos „*LIEPA TTS*“ ir „*VITS*“ technikos, kurios panaudojamos sukurti balso įrašams lietuvių kalba. Sintetiniai balso įrašai buvo surinkti tiek moters, tiek vyro balsais, nesintetiniams įrašai surinkti iš kelių audioknygų, iš viso surinkta 18188 balso įrašai. Surinkus duomenų rinkinį taip pat svarbios yra iš ankstinio apdorojimo operacijos, požymių išskyrimas, kurie yra skirti identifikavimo modeliams apmokyti ir įvertinti jų veikimą. Kaip požymius modeliams apmokyti ir vertinti pasirinkta naudoti laiko srities garso signalus ir konstanta Q transformacijas išskirtus iš savo rinkinio balso įrašų.
4. Atlikus bandymus su pasirinktais identifikavimo modeliais pirmoje eksperimento dalyje geriausi rezultatai buvo pasiekti su kombinuotu „*RawNet*“ ir „*Inc-TSSDNet*“ identifikavimo modeliu, tikslumas buvo 97,74% su svertinės kryžminės entropijo reguliavimu, o EER 1,35% su sumaišymo reguliavimu. Savo pasiūlytas „*2D-Inc-TSSDNet*“ modelis šioje eksperimento dalyje pranoko „*2D-Res-TSSDNet*“, tikslumas buvo didesnis 5,63%, o EER 9,85% apskaičiavus bendrą reguliarizavimo technikų vidurkį. Antroje eksperimento dalyje panaudojus išmetimo sluoksnius (angl. *dropout layers*) identifikavimo modelių architektūroje ir kombinuojant kartu su prieš tai naudotom reguliavimo technikom tam tikrais atvejais pavyko pagerinti tikslumą ir sumažinti EER. Trečioje eksperimento dalyje panaudojus SMOTE techniką ant apmokymo poaibio savo rinkinyje siekiant sumažinti klasių disbalansą su „*RawNet*“ identifikavimo modeliu pavyko užfiksuoti žemiausią ir geriausią 0,85% EER rodiklį panaudojant tik išmetimo sluoksnius. Be to, su šiuo identifikavimo modeliu užfiksuotas geriausias tikslumas 98,86% su sumaišymo reguliavimu. Paskutinėje eksperimento dalyje buvo atliktas stratifikuotas kryžminis tikrinimas sukuriant 4 papildomus poaibius, kad bendroje sumoje būtų 5 poaibiai ir daugiau eksperimentų rezultatų pridedant

rezultatus gautus primoje dalyje. Apskaičiavus visu modelių vidurkių rezultatus ir standartinius nuokrypius nustatyta, kad aukščiausi rezultatai buvo pasiekti su kombinuotu identifikavimo modeliu, tikslumas – 86,64%, o EER – 6,38%. Antrą geriausia tikslumą pademonstravo pritaikytas „*2D-Inc-TSSDNet*“ modelis – 84,84%, o antras geriausias EER rodiklis pasiektas su pritaikytu „*RawNet*“ modeliu – 6,82%. Reziumuojant rezultatus galima teigti, kad kombinuotas modelis pasiekė geriausius rezultatus tyime, apjungdamas dviejų skirtingų identifikavimo modelių architekūras, kas parodo, kad sujungus kelis identifikavimo modelius sintetiniams balsui identifikuoti, gali padidėti visos sistemos patikimumas ir tikslumas.

Atlikus bandymus nustatyta, kad geriausius rezultatus pasiekė kombinuotas iš dviejų skirtingų architektūrų modelis. Nors ir pavyko pasiekti aukštus rezultatus su savo duomenų rinkiniu, bet rinkyje sintetiniams balso įrašams sugeneruoti buvo panaudotos tik dvi sintezavimo technikos. Taip pat svarbu paminėti, kad vertinant modelį su nematytai duomenimis, kurie nebuvo naudojami apmokymo metu, rezultatai gali būti nenuspėjami, todėl svarbu tęsti mokslinius tyrimus šioje kriptyje siekiant užtikrinti autentiškumą, skatinti pasitikėjimą, skaidrumą, spręsti saugumo ir privatumo problemas.

Literatūros sąrašas

Almutairi, Z., & Elgibreen, H. (2022). A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. In *Algorithms* (Vol. 15, Issue 5). MDPI. https://doi.org/10.3390/a15050155

Arif, T., Javed, A., Alhameed, M., Jeribi, F., & Tahir, A. (2021). Voice Spoofing Countermeasure for Logical Access Attacks Detection. *IEEE Access*, *9*, 162857–162868. https://doi.org/10.1109/ACCESS.2021.3133134

Aylett, M. P., Vinciarelli, A., & Wester, M. (2020). Speech synthesis for the generation of artificial personality. *IEEE Transactions on Affective Computing*, *11*(2), 361–372. https://doi.org/10.1109/TAFFC.2017.2763134

Ballesteros, D. M., Rodriguez-Ortega, Y., Renza, D., & Arce, G. (2021). Deep4SNet: deep learning for fake speech classification. *Expert Systems with Applications*, *184*. https://doi.org/10.1016/j.eswa.2021.115465

Bartusiak, E. R., & Delp, E. J. (2021). Frequency Domain-Based Detection of Generated Audio. *In Proceedings of the Electronic Imaging; Society for Imaging Science and Technology, New York, NY, USA, 11–15 January*, 273–281. http://arxiv.org/abs/2205.01806

Borrelli, C., Bestagini, P., Antonacci, F., Sarti, A., & Tubaro, S. (2021). Synthetic speech detection through short-term and long-term prediction traces. *Eurasip Journal on Information Security*, *2021*(1). https://doi.org/10.1186/s13635-021-00116-3

Chen, T., Kumar, A., Nagarsheth, P., Sivaraman, G., & Khoury, E. (2020). Generalization of Audio Deepfake Detection. *The Speaker and Language Recognition Workshop (Odyssey 2020)*, 132–137. https://doi.org/10.21437/odyssey.2020-19

Gomez-Alanis, A., Peinado, A. M., Gonzalez, J. A., & Gomez, A. M. (2019). A light convolutional Gru-RNN deep feature extractor for ASV spoofing detection. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2019-September*, 1068–1072. https://doi.org/10.21437/Interspeech.2019-2212

Hua, G., Teoh, A. B. J., & Zhang, H. (2021). Towards End-to-End Synthetic Speech Detection. *IEEE Signal Processing Letters*, *28*, 1265–1269. https://doi.org/10.1109/LSP.2021.3089437

Jiang, Z., Zhu, H., Peng, L., Ding, W., & Ren, Y. (2020). Self-supervised spoofing audio detection scheme. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2020-October*, 4223–4227. https://doi.org/10.21437/Interspeech.2020-1760

Kawa, P., Plata, M., & Syga, P. (2022). *Attack Agnostic Dataset: Towards Generalization and Stabilization of Audio DeepFake Detection*. https://doi.org/10.21437/Interspeech.2022-10078

Khanjani, Z., Watson, G., & Janeja, V. P. (n.d.). *How Deep Are the Fakes? Focusing on Audio Deepfake: A Survey*.

Khochare, J., Joshi, C., Yenarkar, B., Suratkar, S., & Kazi, F. (2022). A Deep Learning Framework for Audio Deepfake Detection. *Arabian Journal for Science and Engineering*, *47*(3), 3447–3458. https://doi.org/10.1007/s13369-021-06297-w

Li, X., Li, N., Weng, C., Liu, X., Su, D., Yu, D., & Meng, H. (2021). Replay and Synthetic Speech Detection with Res2net Architecture. *In 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 6354–6358. http://arxiv.org/abs/2010.15006

Lim, S. Y., Chae, D. K., & Lee, S. C. (2022). Detecting Deepfake Voice Using Explainable Deep Learning Techniques. *Applied Sciences (Switzerland)*, *12*(8). https://doi.org/10.3390/app12083926

Ling, Z. H., Kang, S. Y., Zen, H., Senior, A., Schuster, M., Qian, X. J., Meng, H., & Deng, L. (2015). Deep Learning for Acoustic Modeling in Parametric Speech Generation: A systematic review of existing techniques and future trends. In *IEEE Signal Processing Magazine* (Vol. 32, Issue 3, pp. 35–52). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/MSP.2014.2359987

Ma, H., Yi, J., Tao, J., Bai, Y., Tian, Z., & Wang, C. (2021). Continual Learning for Fake Audio Detection. *In: 22nd Annual Conference of the International Speech Communication Association. ISCA*, 886–890. https://doi.org/10.21437/Interspeech.2021-794

Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2022). Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*. https://doi.org/10.1007/s10489-022-03766-z

Müller, N. M., Pizzi, K., & Williams, J. (2021). Human Perception of Audio Deepfakes. *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia (DDAM '22), October 14, 2022, Lisabona, Portugal*, *1*, 85–91. https://doi.org/10.1145/3552466.3556531

P, R. T., Aravind, P. R., C, R., Nechiyil, U., & Paramparambath, N. (2020). *Audio Spoofing Verification using Deep Convolutional Neural Networks by Transfer Learning*. http://arxiv.org/abs/2008.03464

Shan, M., & Tsai, T. (2020). *A Cross-Verification Approach for Protecting World Leaders from Fake and Tampered Audio*. http://arxiv.org/abs/2010.12173

Singh, A. K., & Singh, P. (2020). Detection of AI-Synthesized Speech Using Cepstral & Bispectral Statistics. *IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR), 2021, Tokyo, Japan*, 412–417. https://arxiv.org/abs/2009.01934

Story, B. H. (2019). History of speech synthesis. In *The Routledge Handbook of Phonetics* (pp. 9–33). Taylor and Francis. https://doi.org/10.4324/9780429056253-2

Subramani, N., & Rao, D. (2020). Learning Efficient Representations for Fake Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence, 34(04)*, 5859–5866. https://doi.org/10.1609/aaai.v34i04.6044

Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N., & Larcher, A. (2020). *End-to-end anti-spoofing with RawNet2*. http://arxiv.org/abs/2011.01108

Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., Soong, F., Qin, T., Zhao, S., & Liu, T.-Y. (2022). *NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality*. http://arxiv.org/abs/2205.04421

Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., Sahidullah, M., Vestman, V., Kinnunen, T., Lee, K. A., Juvela, L., Alku, P., Peng, Y.-H., Hwang, H.-T., Tsao, Y., Wang, H.-M., Maguer, S. Le, Becker, M., Henderson, F., … Ling, Z.-H. (2019). *ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech*. http://arxiv.org/abs/1911.01601

Wu, Z., Das, R. K., Yang, J., & Li, H. (2020). Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2020-October*, 1101–1105. https://doi.org/10.21437/Interspeech.2020-1810

Yu, H., Tan, Z. H., Ma, Z., Martin, R., & Guo, J. (2018). Spoofing Detection in Automatic Speaker Verification Systems Using DNN Classifiers and Dynamic Acoustic Features. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(10), 4633–4644. https://doi.org/10.1109/TNNLS.2017.2771947

Zhang, Z., Yi, X., & Zhao, X. (2021). Fake Speech Detection Using Residual Network with Transformer Encoder. *IH and MMSec 2021 - Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, 13–22. https://doi.org/10.1145/3437880.3460408

Priedai

1 Priedas. Pažymėjimas už dalyvavimą jaunųjų mokslininkų konferencijoje



VILNIUS TECH
Fundamentinių mokslų fakultetas

PAŽYMĖJIMAS

Nr. 24-0403-23

Emilis Kaplunovas

2024 m. balandžio 3 d. dalyvavo
27-osios Lietuvos jaunųjų mokslininkų konferencijos
„Mokslas – Lietuvos ateitis“ teminėje konferencijoje
Informacinių technologijų sauga ir informacinės sistemos.
Sekcijoje „Informacinių technologijų sauga“ pristatė pranešimą
„Sintetinių ir nesintetinių balso įrašų identifikavimas taikant DI“
(„Identification of Synthetic and Non-Synthetic Voice Recordings Using AI“).

Mokslo komiteto pirmininkas prof. dr. Nikolaj Goranin

Organizacinio komiteto pirmininkas prof. dr. Diana Kalibatienė

Vilnius, 2024

2 Priedas. Modelių tikslumo vidurko rezultatai stratifikuoto tikrinimo dalyse

