VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

INFORMATIKOS STUDIJŲ PROGRAMA

Baigiamasis magistro darbas

Emocionalios lietuvių kalbos sintezė naudojant neuroninius tinklus

Text-To-Speech Synthesis Of Emotional Lithuanian Language Based On Neural Networks

Gintaras Marčiukonis

Darbo vadovas : Doc. Dr. Pijus Kasparaitis

Recenzentas : Prof. Dr. Povilas Treigys

Vilnius
2026

Santrauka

Šiame darbe tiriamas papildomo mokymo metodas emocinės lietuviškos kalbos sintezei esant ribotam duomenų kiekiui. Darbo metu sukurtas originalus lietuviškos emocinės kalbos duomenų rinkinys LESS, sudarytas iš 960 garso įrašų trijose emocinėse kategorijose: linksma, neutrali ir liūdna. Naudojant VITS architektūrą apmokyti keturi baziniai modeliai — trys naudojant CVL ir LIEPA duomenų rinkinius, vienas naudojant studijos kokybės ELKG duomenų rinkinį. Remiantis kiekvienu baziniu modeliu, papildomo mokymo metodu sukurti dvylika emocinę kalbą sintetizuojančių modelių. Modelių kokybė įvertinta dviem metodais: objektyvia akustine analize, lyginant sintetizuotos ir realios kalbos akustinius požymius, bei subjektyviu MOS tyrimu, kuriame dalyvavo 15 lietuvių kalbos gimtakalbių. Rezultatai parodė, kad studijos kokybės duomenimis apmokyti S grupės modeliai pasiekė žymiai geresnius rezultatus nei CVL ir LIEPA duomenimis apmokyti B grupės modeliai tiek akustinėje analizėje, tiek subjektyviame vertinime. Nustatyta, kad bazinio modelio ir papildomo mokymo duomenų diktorių akustinis panašumas yra svarbus veiksnys sėkmingam emocinių savybių perėmimui.

Raktiniai žodžiai: teksto į garsą sintezė, neuroniniai tinklai, papildomas mokymas, akustinė analizė, riboti duomenys, VITS, emocijos

Summary

This paper investigates fine-tuning methods for emotional Lithuanian speech synthesis under limited data conditions. An original Lithuanian emotional speech dataset LESS was created, consisting of 960 recordings across three emotional categories: happy, neutral, and sad. Using the VITS architecture, four base models were trained — three using the CVL and LIEPA datasets, and one using the studio-quality ELKG dataset. Based on each base model, twelve emotional speech synthesis models were created using the fine-tuning method. Model quality was evaluated using two methods: objective acoustic analysis comparing synthesized and real speech features, and a subjective MOS study involving 15 native Lithuanian speakers. Results showed that S-group models trained on studio-quality data significantly outperformed B-group models trained on CVL and LIEPA data in both acoustic analysis and subjective evaluation. It was found that acoustic similarity between base model training data and fine-tuning data speakers is an important factor for successful emotional feature transfer.

Keywords: text-to-speech synthesis, neural networks, fine-tuning, acoustic analysis, low-resource, VITS, emotions

Turinys

Įvadas

Pastaraisiais metais dirbtinis intelektas ir giliojo mokymosi metodai sparčiai keičia kalbos technologijų sritį. Neuroniniai tinklai leidžia analizuoti, generuoti ir suprasti kalbą vis sudėtingesniais lygmenimis, todėl vis daugiau dėmesio skiriama natūralios kalbos apdorojimui. Vienas iš įdomiausių ir sudėtingiausių uždavinių šioje srityje – emocinės kalbos sintezė.

Lietuvių kalba dėl savo sudėtingos gramatinės struktūros, turtingo žodyno ir palyginti mažo skaitmeninio duomenų kiekio kelia ypatingus iššūkius, norint sukurti kokybiškus kalbos sintezatorius. Emocinės kalbos sintezės srityje šie iššūkiai dar labiau išryškėja – viešai prieinamų lietuviškų emocinės kalbos duomenų rinkinių praktiškai nėra. Aukštos kokybės emocinių duomenų rinkimas ir anotavimas yra brangus ir daug laiko reikalaujantis procesas, todėl itin svarbu ieškoti metodų, leidžiančių sumažinti reikiamų duomenų kiekį išlaikant sintezės kokybę. Šiame darbe tiriamas papildomo mokymosi metodas (angl. *fine-tuning*) emocinės lietuviškos kalbos sintezei esant ribotiems duomenų ištekliams. Papildomas mokymasis leidžia apmokyti jau egzistuojantį neutralios kalbos modelį naudojant nedidelį emocinių įrašų kiekį, tokiu būdu išsaugant bazinio modelio išmoktas kalbines charakteristikas ir pritaikant jas emocinei raiškai.

Darbo tikslas

Ištirti papildomo mokymosi metodo tinkamumą emocinės lietuvių kalbos sintezei esant ribotiems duomenų ištekliams ir įvertinti sukurtų modelių gebėjimą perteikti emocines akustines savybes.

Darbo uždaviniai:

1. Sudaryti originalų lietuviškos emocinės kalbos duomenų rinkinį LESS, skirtą modelių mokymui ir testavimui.
2. Apmokyti bazinius neutralios kalbos sintezės modelius naudojant skirtingus duomenų rinkinius ir mokymo strategijas.
3. Pritaikyti papildomo mokymosi metodą ir apmokyti emocinius modelius naudojant LESS ir ELKG duomenų rinkinius.
4. Įvertinti apmokytų modelių gebėjimą perteikti emocines savybes atliekant akustinę analizę.
5. Atlikti subjektyvų vertinimą, naudojant MOS testą su lietuvių kalbos gimtakalbiais.

Metodologija

Mokymo tipas

Aukštos kokybės emocinės kalbos duomenų rinkinių sudarymas yra brangus ir daug laiko reikalaujantis procesas, todėl mokymas nuo nulio praktiškai neįgyvendinamas esant ribotiems duomenų ištekliams [THD19]. Šiame darbe pasirinktas papildomo mokymosi metodas, leidžiantis papildomai

apmokyti jau egzistuojantį neutralios kalbos TTS modelį naudojant nedidelį emocinių įrašų kiekį, tokiu būdu išsaugant bazinio modelio išmoktas kalbines charakteristikas ir pritaikant jas emocinei raiškai.

Modelio architektūra

Kalbos sintezės modelių mokymas yra sudėtingas uždavinys, reikalaujantis ištirti skirtingas architektūras, mokymo parametrus ir duomenų rinkinius. Emocinės kalbos generavimo atveju šis uždavinys įgauna papildomą sudėtingumo dimensiją. Siekiant supaprastinti sprendžiamą uždavinį, šiame darbe pasirinkta naudoti VITS architektūrą [KKS21] dėl šių priežasčių:

1. Remiantis ankstesniais tyrimais, VITS architektūros modeliai geba išmokti generuoti aukštos kokybės kalbą nuo nulio, naudojant santykinai nedidelius duomenų kiekius [KKS21; Rad22].
2. Yra įrodymų, kad papildomai mokant VITS architektūros modelius nedideliais emocinių įrašų kiekiais, galima perteikti emocines akustines savybes sintetizuojamoje kalboje [THD19].
3. VITS yra end-to-end sintezės modelis [MYD21], kuris tiesiogiai iš teksto generuoja garso bangą, taip supaprastinant sintezės procesą. Dviejų etapų architektūrose, tokiose kaip Tacotron 2 [SPW+18], tekstas pirmiausia paverčiamas mel spektrogramu, o tuomet atskiras vokoderių modelis mel spektrogramą paverčia garso banga. Toks atskiras apdorojimas gali lemti klaidų kaupimąsi tarp etapų, o VITS šią problemą išsprendžia sujungdamas abu etapus į vieną modelį.

Modelių įvertinimas

Objektyvus vertinimas Šiame darbe papildomai apmokytų modelių gebėjimas sintezuoti emocinę kalbą vertinamas remiantis normalizuotu akustiniu atstumu tarp sintetizuotos ir realios kalbos akustinių savybių vidurkių. Analizei pasirinkti keturi akustiniai požymiai: pagrindinio tono vidurkis, pagrindinio tono standartinis nuokrypis, pagrindinio tono diapazonas ir įgarsinimo koeficientas. Šie požymiai leidžia akustiškai apibūdinti ir atpažinti emocijas kalboje – linksma kalba pasižymi aukštesniu pagrindinio tono vidurkiu, didesniu kintamumu ir mažiau pauzių, tuo tarpu liūdna kalba – žemesniu tonu, monotoniškesniu skambesiu ir dažnesnėmis pauzėmis [JL03].

Subjektyvus vertinimas Subjektyviam vertinimui paruošta internetinė apklausa, kurios metu lietuvių kalbos gimtakalbiai klausytojai klausėsi darbe sukurtų modelių sintezuotos kalbos ir vertino kiekvieną sakinį atskirai pasitelkdami MOS būdą. Be MOS papildomai dalyviai turėjo nurodyti kurią emocinę raišką labiausiai girdi kiekviename įraše.

Naudojami instrumentai

Modelių mokymas atliktas naudojant Coqui TTS biblioteką ir VITS architektūrą. Mokymo aplinka – Google Colab Pro ir Pro+ su NVIDIA A100 bei NVIDIA RTX PRO 6000 Blackwell Server Edition grafikos procesoriais. Akustinė analizė atlikta naudojant Python bibliotekas `parselmouth` ir `librosa`. Subjektyvus vertinimas atliktas per specialiai sukurtą internetinę apklausos platformą.

Siekiami rezultatai

1. Sudaryti originalų lietuviškos emocinės kalbos duomenų rinkinį LESS.
2. Sukurti veikiančius emocinius lietuvių kalbos sintezės modelius naudojant ribotus duomenų išteklius.
3. Nustatyti, kokie veiksniai lemia sėkmingą emocinių akustinių savybių perėmimą papildomo mokymosi metu.
4. Palyginti skirtingų duomenų rinkinių ir mokymo strategijų įtaką sukurtų modelių kokybei.
5. Pasitelkus objektyvius ir subjektyvius vertinimo metodus, įvertinti ir palyginti kiekvieną sukurtą modelį.

1. Literatūros analizė

1.1. Emocijos ir jų akustinės charakteristikos

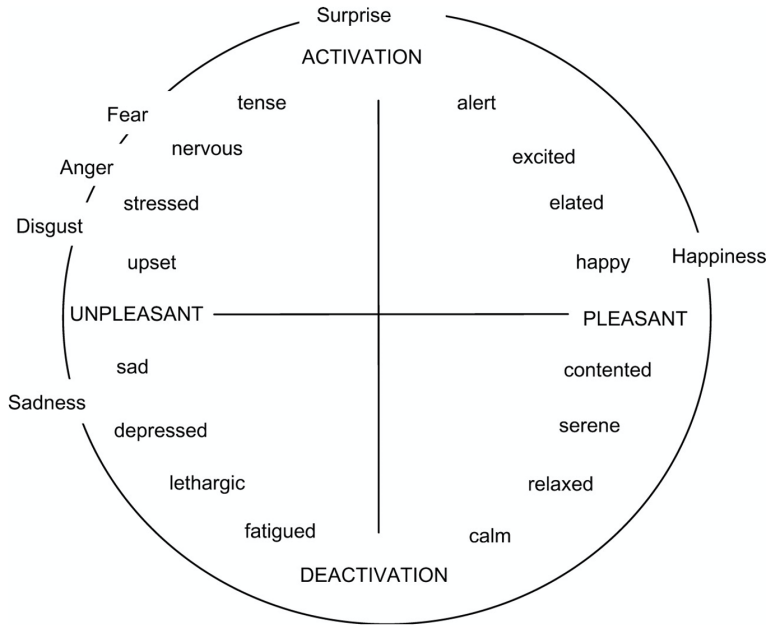
Emocijos yra sudėtingi psichofiziologiniai procesai, kurie atspindi žmogaus subjektyvų požiūrį į tam tikrus reiškinius, situacijas ar įvykius. Emocijos apima įvairius komponentus: fiziologines reakcijas (širdies plakimo dažnio pokytis), elgesio tendencijas (polinkis verkti ar juoktis) ir subjektyvią patirtį (jausmas liūdesio ar džiaugsmo) [Sch05]. Vokalinėje raiškoje emocijos pirmiausia pasireiškia per akustinių ir prozodinių ypatybių pokyčius. Pavyzdžiui, laimę dažnai apibūdina aukštesnis tonas, padidėjęs intensyvumas ir greitesnis kalbos greitis, o liūdesys yra susijęs su žemesniu tonu, sumažėjusiu intensyvumu ir lėtesniu kalbėjimu. Kita vertus, pyktis pasižymi įtempta balso kokybe, didesniu tono kintamumu ir staigiais ritmo pokyčiais [Sto15].

1.2. Emocijų klasifikacija

Emocijų klasifikavimas yra esminis aspektas norint suprasti, kaip emocijos išreiškiamos ir suvokiamos kalboje. Nagrinėjant literatūrą, galima rasti įvairių pasiūlytų būdų emocijoms klasifikuoti. Vieni iš labiausiai paplitusių požiūrių yra kategoriniai ir dimensiniai emocijų modeliai [PM17].

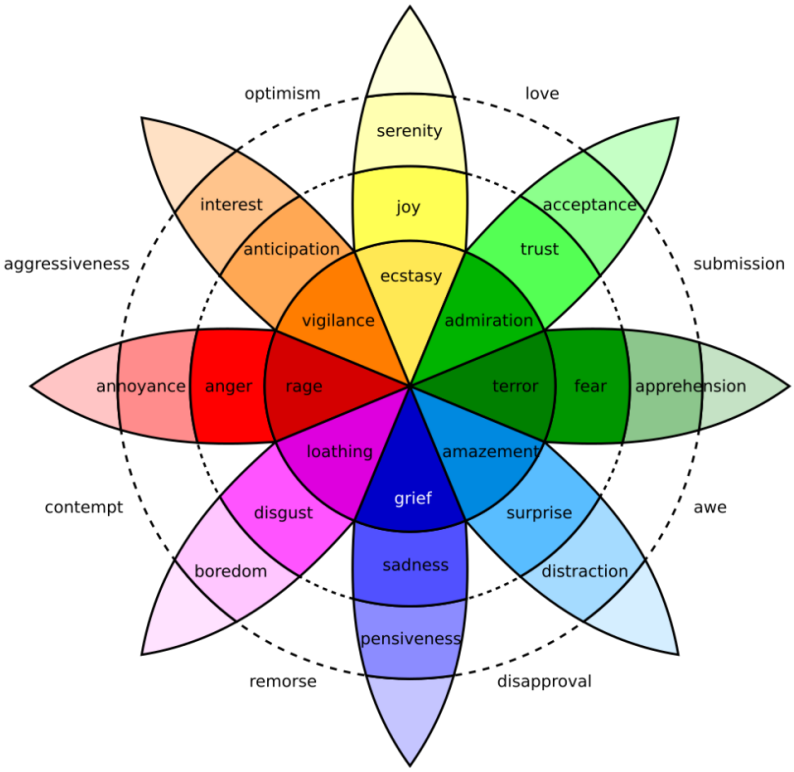
Kategoriniai modeliai emocijas skirsto į atskiras, aiškiai apibrėžtas kategorijas. Vienas žymiausių pavyzdžių – Paulo Ekmano pagrindinių emocijų teorija, kuri išskiria šešias bazines emocijas: džiaugsmą, liūdesį, pyktį, baimę, pasibjaurėjimą ir nuostabą. Šios emocijos, pasak Ekmano, yra universalios ir atpažįstamos visose kultūrose [Ekm92].

Dimensiniai modeliai emocijas vaizduoja kaip kontinuumą tam tikroje erdvėje, o ne kaip atskiras kategorijas. Pavyzdžiui, Jameso A. Russello emocijų ratas, kuris pavaizduotas 1 pav. naudoja dvi pagrindines dimensijas: malonumo lygį (nuo teigiamų iki neigiamų emocijų) ir sužadinimo intensyvumą (nuo ramybės iki didelio susijaudinimo). Šis modelis leidžia matyti emocijas kaip nuoseklų spektrą, kuriame, pavyzdžiui, džiaugsmas būtų teigiama ir aukšto sužadinimo emocija, o liūdesys – neigiama ir žemo sužadinimo emocija [PRP05].



1 pav. *Jameso A. Russello emocijų ratas*

Kitas reikšmingas dimensinis modelis yra Plutchiko emocijų ratas, kuriame emocijos išdėstytos pagal jų intensyvumą ir tarpusavio santykį, leidžiantį vizualiai pamatyti, kaip viena emocija gali pereiti į kitą. Plutchiko emocijų ratas pavaizduotas 2 pav. Plutchikas taip pat pabrėžė emocijų sudėtingumą, siūlydamas, kad emocijos gali būti derinamos tarpusavyje, sukuriant sudėtines emocijas, pavyzdžiui, meilė kaip džiaugsmo ir pasitikėjimo derinys [Plu80].



2 pav. *Robert Plutchiko emocijų ratas*

Šiame darbe naudojamas kategorinis požiūris – pasirinktos trys emocinės kategorijos: linksma, neutrali ir liūdna. Šios kategorijos pasirinktos dėl ryškaus akustinio kontrasto tarp linksmos ir liūdnos emocijų bei neutralios emocijos kaip bazinės lyginamosios kategorijos.

1.3. Akustinės emocijų charakteristikos

Kiekvienai emocinei kategorijai būdingi skirtingi akustiniai požymiai [ESS+15; JL03]:

- **Linksma emocija** pasižymi aukštesniu pagrindinio tono vidurkiu, didesniu tono kintamumu ir platesniu diapazonu. Kalbos tempas greitesnis, pauzės trumpesnės, o įgarsinimo koeficientas aukštesnis.
- **Neutrali emocija** pasižymi vidutinėmis visų akustinių požymių reikšmėmis – nėra ryškių nukrypimų nei į vieną, nei į kitą pusę.
- **Liūdna emocija** pasižymi žemesniu pagrindinio tono vidurkiu, mažesniu kintamumu ir siauresniu diapazonu. Kalbos tempas lėtesnis, pauzės ilgesnės, o įgarsinimo koeficientas žemesnis.

Šie akustiniai skirtumai sudaro pagrindą objektyviam modelių vertinimui – lyginant sintetizuotos kalbos akustines savybes su realių įrašų charakteristikomis galima įvertinti, ar modelis sėkmingai perėmė tikslinės emocijos akustines ypatybes.

1.4. Kalbos sintezė

1.4.1. Tradiciniai kalbos sintezės metodai

Kalbos sintezė – tai procesas, kurio metu sistemos generuoja žmogaus balsą imituojančią kalbą. Šiam procesui įgyvendinti egzistuoja įvairūs tradiciniai metodai, tokie kaip konkatenacinė sintezė, formantinė sintezė, artikuliacinė sintezė ir hibridinė sintezė, kurių kiekvienas turi savitus principus ir taikymo ypatumus [BAD13].

1.4.2. Konkatenacinė sintezė

Konkatenacinė sintezė (angl. *Concatenative Synthesis*) – tai kalbos sintezės metodas, kai kalba generuojama sujungiant iš anksto įrašytus mažus kalbos segmentus, tokius kaip fonemos, skiemenys, žodžiai ar frazės. Šie segmentai saugomi didelėje duomenų bazėje, o sintezės metu parenkami tinkamiausi vienetai ir sujungiami į vientisą kalbos signalą. Šis metodas pasižymi gana aukšta garso kokybe, nes naudoja realaus žmogaus balso įrašus, tačiau gali kilti problemų, kai jungiamų segmentų ribose atsiranda girdimų perėjimų ar nesuderinamumų [BAD13].

1.4.3. Formantų sintezė

Formantinė kalbos sintezė – tai kalbos gaminimo metodas, kuris remiasi formantų, t.y. specifinių aukšto dažnio garsų juostų, modeliavimu. Formantai yra svarbūs, nes jie padeda atskirti įvairius balsių garsus ir sudaro žmogaus kalbos garsų struktūrą.

Šio tipo sintezėje, norint sukurti kalbos garsus, imituojami formantų dažniai ir jų stiprumas. Tai nereiškia, kad kalba kuriama naudojant tikrą žmogaus balsą, o vietoj to taikomi matematika ir akustikos principai, kad būtų atkurtos balsių ir kitų kalbos garsų savybės [Kel95].

1.4.4. Artikuliacinė sintezė

Artikuliacinė kalbos sintezė yra pagrįsta fiziniais žmogaus kalbos aparato modeliais. Ji apima akustinių funkcijų modeliavimą ir dinaminį balso takų judėjimo atkūrimą. Artikuliacinis modelis atkuria balso takų formą, priklausomai nuo fonavimo organų (lūpų, žandikaulio, liežuvio) padėties. Garso signalas generuojamas atliekant matematinę oro srauto per balso takus simuliaciją. Tokio sintezatoriaus valdymo parametrai apima balso stygų įtempimą ir skirtingų artikuliacinių organų tarpusavio padėtį. Artikuliacinis modelis atkuriamas taip, kad kuo tiksliau atitiktų natūralią balso takų formą ir jų veikimą [BAD13].

1.5. Kalbos sintezė naudojant neuroninius tinklus

Neuroniniai tinklai yra modeliai, kurie gali analizuoti ir generuoti duomenis pagal ankstesnės patirties pavyzdžius. Kalbos sintezėje jie naudojami mokant sistemas kurti garso signalus, kurie atkartotų žmogaus kalbos savybes.

Tradicinės neuroninių tinklų pagrindu veikiančios kalbos sintezės sistemos veikia dviem pagrindiniais etapais:

1. Akustinio modelio etapas – neuroninis tinklas iš teksto generuoja tarpinį akustinį atvaizdą, dažniausiai mel spektrogramą, kurioje užkoduota informacija apie kalbos intonaciją, ritmą ir pauzes. Šiame etape dažnai naudojami neuroniniai tinklai, transformerių modeliai arba konvoliuciniai neuroniniai tinklai.

2. Vokoderio etapas – atskiras modelis mel spektrogramą paverčia garso banga. Šiuolaikiniuose sistemose dažnai naudojami neuroniniai vokoderiai, tokie kaip WaveNet, HiFi-GAN ar WaveGlow, kurie generuoja itin realistišką balsą.

1.6. Neuroninių tinklų mokymo iššūkiai

Neuroninių tinklų mokymas kalbos sintezėje yra sudėtingas procesas, reikalaujantis didelio duomenų kiekio ir skaičiavimo išteklių. Toliau pateikiami pagrindiniai iššūkiai: [TQS+21]

- **Didelis duomenų poreikis** – neuroniniai tinklai efektyviai mokosi tik tada, kai jie apdoroja didelį kalbos įrašų kiekį kartu su atitinkamais tekstais.
- **Dideli skaičiavimo resursų reikalavimai** – norint efektyviai apmokyti tokius modelius, reikia naudoti aukštos našumo kompiuterines sistemas su specializuota aparatine įranga.
- **Duomenų kokybės svarba** – net jei turima daug duomenų, jie turi būti aukštos kokybės. Triukšmingi ar prastai įrašyti duomenys gali neigiamai paveikti modelio veikimą.

- **Perteklinis pritaikymas** (angl. *overfitting*) – kai modelis per daug prisitaiko prie mokymosi duomenų ir tampa mažiau lankstus naujiems duomenims.
- **Intonacijos ir emocijų perteikimas** – nors neuroniniai tinklai sugeba generuoti realistiškai skambančią kalbą, intonacijų, emocijų ir subtilių kalbos niuansų atkūrimas išlieka vienu didžiausių iššūkių.

Ypatingas dėmesys duomenų kiekiui ir kokybei aktualus ir šiame darbe – lietuvių kalbos emocinių duomenų trūkumas buvo vienas pagrindinių iššūkių kuriant sintezės modelius.

1.7. End-to-end sintezė

End-to-end kalbos sintezė – tai modernus požiūris į kalbos sintezę, kuriame visas procesas – nuo teksto iki garso generavimo – vyksta viename neuroniniame modelyje, be atskirų tarpinės informacijos lygmenų [MYD21]. Toliau aptariami pagrindiniai aspektai, dėl kurių end-to-end sintezė yra itin svarbi.

- **Supaprastintas sintezės procesas** – tradiciniai metodai reikalavo daug atskirų žingsnių, o end-to-end sintezė viską atlieka vienu neuroniniu tinklu.
- **Geriau skambantis balsas** – nauji modeliai gali sukurti natūralesnį ir sklandesnį balsą.
- **Lengviau pritaikyti naujoms kalboms ir balsams** – galima lengvai išmokyti modelį kalbėti nauja kalba arba mėgdžioti konkretų balsą.
- **Mažiau žmogaus įsikišimo** – end-to-end sintezės modeliai gali būti mokomi tiesiogiai iš didelių tekstas–garsas duomenų rinkinių [MYD21; NHW+19].

Vienas pažangiausių end-to-end sintezės modelių yra VITS [KKS21], kuris tiesiogiai generuoja garso bangas iš teksto naudodamas giliojo mokymosi metodus. Pagrindiniai VITS architektūros komponentai:

- **Variacinis autokoderis** (VAE) – koduoja tekstą į latentinę erdvę ir iš jos generuoja garso atvaizdą.
- **Normalizuojantys srautai** – pagerina latentinės erdvės atvaizdavimą, leidžiant modeliui tiksliau atkurti garso savybes.
- **HiFi-GAN vokoderis** – iš latentinio atvaizdavimo generuoja aukštos kokybės garso bangą. Naudoja GAN (angl. *Generative Adversarial Network*) principą – generatorius ir diskriminatorius mokosi vienu metu: generatorius bando sukurti kuo realistiškesnį garsą, o diskriminatorius bando atskirti sintetinį garsą nuo realaus.
- **Trukmės prediktorius** – nustato kiek laiko kiekvienas tekstinis vienetas turi trukti sintezuojamoje kalboje.

Skirtingai nei ankstesni metodai, VITS nereikalauja tarpinės spektrogramos, o visą sintezės procesą optimizuoja viename neuroniniame modelyje, užtikrindamas itin natūralų balso skambesį [ZY23].

1.8. Emocinės kalbos sintezės mokymo metodai

Emocinės kalbos sintezės modeliai skirstomi į dvi pagrindines kategorijas pagal mokymo tipą: prižiūrimą ir neprižiūrimą mokymą [BTD24]. Prižiūrimas mokymas (angl. *supervised learning*) – tai metodas, kuomet modelis mokosi iš duomenų rinkinių, kuriuose kiekvienas pavyzdys turi aiškiai apibrėžtą kategoriją ar etiketę. Šiame darbe naudojamas prižiūrimas mokymas, o konkreti strategija – atskiri modeliai kiekvienai emocinei kategorijai. Šis metodas vykdomas dviem etapais. Pirmiausia mokomas neutralaus kalbėjimo bazinis modelis, naudojant didelį kiekį duomenų. Tada sukuriamos šio modelio kopijos, kurios papildomai apmokomos su mažesniais, specifinės emocijos duomenimis. Tokia strategija leidžia išsaugoti bazinio modelio išmoktas kalbines charakteristikas ir pritaikyti jas emocinei raiškai [BTD24].

1.9. Emocinės kalbos sintezės tyrimų apžvalga

Emocinė kalbos sintezė yra svarbi kalbos apdorojimo technologija, kuri siekia ne tik generuoti fonetiškai teisingą kalbą, bet ir perteikti įvairias emocines būsenas. Toks gebėjimas suteikia kompiuteriams ir virtualiems asistentams galimybę bendrauti su žmonėmis natūraliau ir empatiškiau. Emocinės kalbos sintezės tyrimai sparčiai vystosi, o moksliniai straipsniai siūlo įvairius požiūrius ir metodus, kaip pasiekti geriausią emocijų perteikimo kokybę. Šiame skyriuje bus analizuojami atlikti tyrimai, kurių metu buvo bandoma sukurti emocinės kalbos sintezatorius. Akcentas bus dedamas į įvairius aspektus, tokius kaip modelių architektūra, pasitelkti duomenų rinkiniai, gauti rezultatai bei rezultatų įvertinimo metodai.

„Emotional Vietnamese Speech Synthesis Using Style-Transfer Learning“ [Tha23]

Naudojami modeliai. „Tacotron 2” kartu su „Flowtron”.

Mokymo duomenys. Šiame darbe autoriai pasiūlė naujus metodus emocinės kalbos sintezavimui. Norint sukurti aukštos kokybės kalbos sintezatorių, svarbu turėti aukštos kokybės duomenų rinkinį iš vieno kalbėtojo. Mokydami modelį sintezuoti vietnamiečių kalbą darbo autoriai pateikė argumentus, kodėl yra svarbu sudaryti savo duomenų rinkinį. Visų pirma tai užtikrina stabilesnę kokybę ir sumažina triukšmą sintezės metu. Vienodas ryšys tarp teksto ir atitinkamo garso padeda išlaikyti nuoseklumą, net kai sintezuojami skirtingi sakiniai. Naudojant vieno balso duomenų rinkinį, galima sukaupti daugiau duomenų iš vieno kalbėtojo, o tai leidžia pagerinti balso natūralumą, lyginant su ribotu duomenų rinkiniu iš daugelio balsų. Be to, platus žodynas, įtraukiantis tiek kalbinę, tiek rašytinę kalbą, leidžia modeliui įvaldyti įvairius žodžius ir frazes. Modelio mokymui buvo naudojami tokie duomenys: 30 sakinių perskaitytų su linksma emocija, 30 sakinių perskaitytų su liūdna emocija ir 10 000 sugeneruotų sakinių su neutralia emocija. Sukurtam modeliui įvertinti buvo naudojamas subjektyvus garso kokybės įvertinimo metodas MOS, kuris bus aptartas sekančiame skyriuje.

Tyrimo rezultatų įvertinimas. Tyrimo rezultatai parodė, kad sintezatoriaus sugeneruota kalba buvo gana natūrali ir lengvai suprantama – dauguma įvertinimų viršijo 4 balus. Vertintojų nuomonės

tarp vyrų ir moterų šiek tiek skyrėsi, tačiau bendrai rezultatai buvo panašūs, o moterys dažniau suteikė aukštesnius balus. Kalbant apie emocijų perteikimą, modelis geriau atkartojo laimingą emociją (vidutinis įvertinimas 3,95) nei liūdną (vidutinis įvertinimas 3,61). Be to, modelis veikė stabiliai tiek su sakiniais, kurie buvo naudojami mokymui, tiek su visiškai naujais, kas rodo jo gebėjimą apibendrinti duomenis. Visgi, autoriai teigė jog liūdesio emocijos perteikimas vis dar nėra idealus ir galėtų būti patobulintas.

„MIST-Tacotron: End-to-End Emotional Speech Synthesis Using Mel-Spectrogram Image Style Transfer“ [MKC22]

Naudojami modeliai. Straipsnyje „MIST-Tacotron: End-to-End Emotional Speech Synthesis Using Mel-Spectrogram Image Style Transfer“ pristatomas „MIST-Tacotron“ – „Tacotron 2“ pagrindu sukurtas kalbos sintezės modelis, kuris naudoja mel-spektrogramų stiliaus perkėlimo techniką, siekiant pagerinti sintezuoto balso kokybę ir išraiškingumą.

Mokymo duomenys. Eksperimentui buvo naudojami du atskiri duomenų rinkiniai: vieno diktoriaus ir daugelio diktorių duomenų rinkiniai. Vieno diktoriaus duomenys: Šis duomenų rinkinys apima 3 000 sakinių, kuriuos įrašė viena profesionali moteris, vaidinanti septynias emocijas: neutralią, pyktį, baimę, pasibjaurėjimą, džiaugsmą, liūdesį ir nuostabą. Daugelio diktorių duomenys: Šis duomenų rinkinys buvo pateiktas iš Pitchtron duomenų rinkinio ir apima 24 000 ištarimų, kurių bendras įrašymo laikas siekia apie 34 valandas. Šie duomenys buvo žymimi pagal emocijos tipą ir pagal tai ar tai buvo vieno diktoriaus duomenys ar daugelio.

Tyrimo rezultatų įvertinimas. Rezultatai buvo įvertinti objektyviais rodikliais GPE, MCD, FFE bei subjektyviu MOS vertinimu. Eksperimentai parodė, kad MIST-Tacotron ženkliai pagerino balso sintezės kokybę, ypač kalbant apie aiškesnę emocinę raišką ir mažesnį klaidų skaičių.

„End-to-End Emotional Speech Synthesis Using Style Tokens and Semi-Supervised Training“ [WLL+19]

Naudojami modeliai. Straipsnyje pristatomas pusiau prižiūrimo mokymo metodas, skirtas emocinei kalbai sintezuoti, naudojant stiliaus žetonus ir „Tacotron“ pagrindu sukurtą modelį. Modelis sujungia stiliaus žetonų metodą, kuris padeda perteikti emocinę kalbą, ir pusiau prižiūrimą mokymą, kuris leidžia modeliams veikti net ir esant mažiems emocijų žymių kiekio duomenims. Autoriai pristato įvairius modelius, tokius kaip EI (Emotion-Infused) modelis, Semi-EI modelis ir Semi-GST modelis, siekdami pagerinti emocijų raišką ir kalbos natūralumą, net jei yra naudojama tik dalis emocijų žymių.

Mokymo duomenys. Šiame tyrime buvo naudojami skirtingi emocijų rinkiniai, apimantys keturias pagrindines emocijas: neutralų, laimingą, liūdną ir pyktį. Iš viso galutinis duomenų rinkinys susidarė iš 2 937 ištarimų, kurie buvo įrašyti vienos diktorės. Įrašai truko apie 4,48 valandos, jų diskretizavimo dažnis buvo 16 kHz, o kvantavimas – 16 bitų. Modeliui apmokyti buvo naudojami duomenys, kurie apėmė įvairias emocijas, tačiau ypatingas dėmesys buvo skiriamas mažesniam duomenų

kiekiui (5% emocijų žymių), kad parodytų, kaip pusiau prižiūrimas metodas gali veikti su ribotais duomenimis. Buvo atlikti skirtingi eksperimentai su įvairiais emocijų žymių kiekiais (2%, 5%, 10%, 20%) ir palyginti rezultatai su įprastais modeliais.

Tyrimo rezultatų įvertinimas. Objektyvūs įvertinimai parodė, kad Semi-GST modelis, naudodamas tik 5% emocijų žymių, pasiekė labai panašius rezultatus į tradicinį EI modelį, kuris naudojo visus emocijų duomenis. Pagrindiniai vertinimo rodikliai buvo MCD, RMSE, FFE. Semi-EI modelis, kuris naudojo daugiau emocijų žymių, pasiekė prastesnius rezultatus, o Semi-GST modelis pranoko Semi-EI modelį visose kategorijose.

Subjektyvūs įvertinimai taip pat patvirtino šiuos rezultatus. Pirmiausia, atlikus ABX preferencijų testus, Semi-GST modelis buvo žymiai pranašesnis už Semi-EI modelį ir beveik pasiekė EI modelio natūralumo lygį. Neatsirado reikšmingų skirtumų tarp Semi-GST ir EI modelių, tai rodo, kad Semi-GST modelis, turintis tik 5% emocijų žymių, sugebėjo sugeneruoti tokią pačią natūralią kalbą.

„FastSpeech2 Based Japanese Emotional Speech Synthesis“ [IM24]

Naudojami modeliai. Šiame straipsnyje apžvelgiama japoniškų emocinių kalbos sintezatorių kūrimo proceso eiga. Tyrėjai naudojo FastSpeech2 modelį kaip pagrindą ir jį papildė dviem naujais blokais, skirtais emocijų kodavimui. Šie blokai, kurie dalijasi parametrais, buvo įterpti prieš ir po FastSpeech2 variacijos adapterio, siekiant užtikrinti tvirtą emocijų valdymą net ir esant ribotam mokymo duomenų kiekiui. Darbo metu buvo sukurti trys modeliai:

- bazinis modelis, kuriame nėra emocijų valdymo.
- Emo-FT modelis, kuriame emocijos yra kondicionuojamos per atskirus emocinius modelius.
- Autorių siūlomos architektūros modelis, kuriame naudojama modifikuota FastSpeech2 architektūra su emocijų blokais.

Mokymo duomenys. Tyrimo metu buvo naudojami du pagrindiniai duomenų rinkiniai: JSUT duomenų rinkinys ir STUDIES emocinių kalbos įrašų rinkinys.

JSUT (basic5000 dalis) - tai pagrindinis duomenų rinkinys, kuris susidarė iš 5 000 tekstų ir kalbos įrašų, kurie buvo neutralūs, be emocijų. Įrašus atliko viena moteris kalbėtoja, kuri skaitė tekstus, kuriuose buvo naudojami visi dažniausiai pasitaikantys japonų kalbos kandži simboliai.

STUDIES (ITA dalis) - tai papildomas rinkinys, skirtas modelio tobulinimui, kad būtų galima sukurti emocinę kalbos sintezę. Duomenų rinkinys apima 400 tekstų ir kalbos įrašų, kuriuos atliko viena moteris kalbėtoja. Kiekvienas iš 100 sakinių buvo įrašytas keturiose skirtingose emocinėse būsenose: neutralioje, pykčio, liūdesio ir laimės.

Abejuose duomenų rinkiniuose buvo įrašai su žmogaus kalba, atitinkamas tekstas ir kontekstinės žymos.

Tyrimo rezultatų įvertinimas. Modelių įvertinimui buvo pasitelktos MCD ir MOS metrikos. Pagal MCD metrikas bazinis modelis pasirodė prasčiausiai (MCD reikšmė viršijo 5), o autorių pasiūlytos architektūros modelis geriausiai (MCD reikšmė buvo mažesnė nei 4,3). Siekiant įvertinti modelių tikslumus pasitelkiant MOS metriką, autoriai surado 10 klausytojų, kurie dalyvavo dvejose klausymosi sesijose. MOS balai buvo gauti naudojant kalbos aiškumo, natūralumo, ir emocijų intensyvumo kriterijus. Geriausiai pagal visus kriterijus MOS metrikos vertinime pasirodė autorių siūlytas modelis, kuris emocijų intensyvumo kriterijumi vidutiniškai gavo 4,01 balo iš 5.

„Rasa: Building Expressive Speech Synthesis Systems for Indian Languages in Lowresource Settings“ [VSR^{+}24b]

Naudojami modeliai. Tyrime buvo naudojamas FastPitch modelis – tai lygiagrečiai veikiantis teksto į kalbą sintezės modelis, pagrįstas Transformer architektūra. Galutiniams garso bangų pavidalo įrašams generuoti iš mel-spektrogramų buvo naudojamas HiFiGAN-V1, kuris buvo apmokytas nuo nulio.

Mokymo duomenys. Tyrimas pristato „Rasa“ – pirmąjį viešai prieinamą ekspresyvų duomenų rinkinį trims Indijos kalboms: asamų, bengalų ir tamilų. Jame yra 10 valandų neutralaus ir 1 – 3 valandos ekspresyvaus kalbėjimo šešioms pagrindinėms Ekmano emocijoms. Eksperimentai parodė, kad net su 1 valanda neutralaus ir 30 minučių ekspresyvaus balso galima pasiekti „pakankamą“ kokybę, o padidinus neutralių duomenų kiekį iki 10 valandų, emocinė raiška ženkliai pagerėja. Autoriai pabrėžia, kad dirbant su limituotais duomenų kiekiais yra ypač svarbu, kad tie duomenys būtų kokybiški su subalansuota fonotaktine aprėptimi.

Tyrimo rezultatų įvertinimas. Modelių vertinimui naudotos metrikos: CER kalbos suprantamumui įvertinti bei MUSHRA skalė emocinės raiškos kokybei įvertinti. Nustatyta, kad bent 1 valandos neutralios kalbos duomenų reikia suprantamai sintezei, o geresni rezultatai pasiekiami optimizuojant skiemenų aprėptį. Emocinė sintezė tapo „priimtina“ su 15 minučių emocinių duomenų, o su 30 minučių pasiekė „gerą“ lygį. Modeliai, iš anksto mokyti su daugiau neutralios kalbos pavyzdžiui, 10 valandų neutralios ir 30 minučių emocinės, pasiekė geresnius rezultatus nei tie, kurie turėjo mažiau neutralių duomenų. Vieną emociją sintezuojantys modeliai veikė geriau su itin mažai duomenų mokymo metu, tačiau esant didesniam duomenų kiekiui, kelių emocijų modeliai buvo pranašesni.

Apžvalgos apibendrinimas. Toliau pateikiamos lentelės, kuriose aprašyta apžvelgtų tyrimų santrauka. 1 lentelėje pateikiami naudojami modeliai ir duomenų rinkiniai, o 2 lentelėje išdėstyti tyrimų rezultatai ir naudojami vertinimo metodai.

1 lentelė. *Tyrimų apžvalga: tyrimo pavadinimas, modeliai ir duomenų rinkiniai*

Straipsnio pavadinimas	**Naudojami modeliai**	**Duomenų rinkiniai**
Emotional Vietnamese Speech Synthesis Using Style-Transfer Learning	Tacotron 2, Flowtron	30 emocinių sakinių (liūdna, linksma), 10000 neutralios emocijos sakinių
MIST-Tacotron: End-to-End Emotional Speech Synthesis	Tacotron 2, Mel-spektrogramų stiliaus perkėlimas	3000 sakinių vieno kalbėtojo, 24000 sakinių daugelio kalbėtojų
End-to-End Emotional Speech Synthesis Using Style Tokens	Tacotron, stiliaus žetonai	2937 ištarimai vienos diktorės su skirtingomis emocijomis
FastSpeech2 Based Japanese Emotional Speech Synthesis	FastSpeech2, papildomi emocijų blokai	JSUT 5000 sakinių be emocijų, STUDIES 400 emocinių sakinių
Rasa: Building Expressive Speech Synthesis Systems	FastPitch, vokoderis HiFiGAN-V1	10 valandų neutralių sakinių, 1–3 valandos emocinės kalbos

2 lentelė. *Tyrimų apžvalga: tyrimo rezultatai ir vertinimo metodai*

Straipsnio pavadinimas	**Tyrimo rezultatai**	**Vertinimo metodai**
Emotional Vietnamese Speech Synthesis Using Style-Transfer Learning	Laimingos ir liūdnos emocijų MOS 3,95 ir 3,61	MOS
MIST-Tacotron: End-to-End Emotional Speech Synthesis	Pagerinta emocijų raiška, mažesnės klaidos	GPE, MCD, FFE, MOS
End-to-End Emotional Speech Synthesis Using Style Tokens	Pusiau prižiūrimas mokymas su 5% pažymėtų duomenų pasiekė gerus rezultatus	MCD, RMSE, FFE, ABX
FastSpeech2 Based Japanese Emotional Speech Synthesis	Autorių modelis surinko vidutiniškai 4,01 MOS balus	MCD, MOS
Rasa: Building Expressive Speech Synthesis Systems	1 valandos neutralių duomenų ir 30 minučių emocinių duomenų užtenka „geram“ lygiu	CER, MUSHRA

1.10. Modelių vertinimo metodai

Modelių įvertinimas yra esminė giliųjų neuroninių tinklų mokymo dalis, nes jis leidžia įvertinti sukurtų modelių kokybę ir jų gebėjimą atlikti užduotis. Tačiau kalbos sintezės, ypač emocinės kalbos sintezės, srityje vien objektyvūs vertinimo metodai ne visada tiksliai atspindi modelio veikimą. Dėl to autoriai savo darbuose dažnai naudoja tiek subjektyvius, tiek objektyvius vertinimo metodus, siekdami išsamiau įvertinti sintezės kokybę. Šiame skyriuje bus apžvelgti pagrindiniai objektyvūs ir subjektyvūs vertinimo metodai, kurie dažniausiai naudojami emocinės kalbos sintezės tyrimuose.

1.10.1. Objektyvūs vertinimo metodai

Objektyvūs vertinimo metodai remiasi kiekybiniais rodikliais ir algoritminiais įvertinimais, kurie leidžia nepriklausomai nuo žmogaus subjektyvumo įvertinti modelių atlikimą. Šie metodai, kaip taisyklė, yra pagrįsti matematiniais, statistiniais arba signalų apdorojimo algoritmais, kurie užtikrina nuoseklumą ir pakartojamumą vertinant kalbos sintezės kokybę. Pagrindiniai objektyvių metodų privalumai yra jų greitis, tikslumas ir galimybė palyginti įvairius modelius bei algoritmus, nes juos galima naudoti automatizuotai ir nesukelia vertinimo klaidų, susijusių su žmogaus nuotaikomis ar nuomonėmis. Toliau pateikiami pagrindiniai objektyvūs sintezės modelių vertinimo metodai [BTD24; MYD21]:

- **Vidutinė kvadratinė paklaida (RMSE)** – matuoja skirtumus tarp tikrojo ir sintezuoto garso akustinių parametrų.
- **Mel-cepstrumo iškraipymas (MCD)** – matuoja skirtumą tarp Mel-cepstrum koeficientų sintezuoto ir realaus garso. Kuo mažesnė MCD reikšmė, tuo geresnė kalbos kokybė.
- **Bendroji tonacijos paklaida (GPE)** – matuoja skirtumą tarp tikrosios ir sintezuotos garso tonacijos.
- **Simbolių klaidų dažnis (CER)**, žodžių klaidų dažnis (WER) – matuoja klaidų dažnį sintezuotame tekste.
- **f0 kadrų klaida (FFE)** – rodiklis, apjungiantis GPE ir VDE metrikas, kuris matuoja bendrą klaidų kiekį, susijusį su pagrindinio tono (f0) reikšmių nukrypimus ir skardumo nustatymo klaidomis.

Vis dėlto šie metodai turi ir trūkumų – jie nesugeba atskleisti kalbos emocinės raiškos subtilumų, kuriuos tiksliau gali įvertinti klausytojas žmogus. Todėl objektyvūs vertinimai dažnai papildo subjektyvius metodus, kad suteiktų pilnesnį vaizdą apie modelio veikimą.

1.10.2. Subjektyvūs vertinimo metodai

Subjektyvūs vertinimo metodai yra svarbūs, nes jie leidžia įvertinti kalbos sintezės kokybę iš vartotojo perspektyvos, atsižvelgiant į tokius aspektus kaip natūralumas, emocinė raiška ir bendras suvokimo kokybės jausmas. Kadangi kalba yra labai priklausoma nuo žmogaus interpretacijos ir nuotaikos, subjektyvūs vertinimai suteikia papildomos vertės, ypač vertinant emocinę kalbos sintezę, kurios negalima pilnai įvertinti tik objektyviais metodais. Pavyzdžiui, emocinės raiškos tikslumas arba natūralumo pojūtis gali būti tinkamai vertinami tik žmogaus klausytojo.

Tačiau pritaikant šiuos metodus kyla ir tam tikrų iššūkių. Pirma, vertinimas užtrunka ilgiau nei automatizuoti objektyvūs metodai, nes reikia pasitelkti žmogaus vertinimą, kuris gali būti laiko ir resursų reikalaujantis procesas. Antra, subjektyvūs vertinimai gali skirtis priklausomai nuo vertintojų, nes skirtingi žmonės gali turėti skirtingus požiūrius, nuotaikas ar nuostatas, o tai gali sukelti nuoseklumo problemų ir sumažinti vertinimo tikslumą. Be to, kai vertinimo procesas apima daug žmonių, kyla papildomų logistinių iššūkių, tokių kaip vertintojų parinkimas, jų instruktavimas ir laiko koordinavimas.

Nepaisant šių iššūkių, subjektyvūs vertinimo metodai išlieka būtini, nes jie suteikia gilų supratimą apie tai, kaip žmogus suvokia ir interpretuoja kalbos sintezę, ypač kai kalbama apie sudėtingas, emocijas perteikiančias sistemas. Toliau pateikiami pagrindiniai subjektyvūs modelių vertinimo metodai [BTD24; MYD21]:

- Vidutinio įvertinimo balas (MOS) – klausytojai įvertina sintezuotą kalbą pagal natūralumą ir suprantamumą, naudodami penkių balų skalę.
- Palyginamasis vidutinio įvertinimo balas (CMOS) – palygina skirtingų modelių MOS vertes su pagrindiniu modeliu, lygindamas tikrąją ir sintezuotą kalbą iš kiekvieno modelio.
- Diferencialinis vidutinio įvertinimo balas (DMOS) – klausytojai įvertina sintezuotus pavyzdžius pagal jų panašumą į tam tikrą emociją ar stilių, naudojant vieną iki penkių balų skalę.
- AB pirmumo testas (angl. *AB Preference Test*) – klausytojai įvertina tuos pačius sakinius,kuriuos sukuria du modeliai, ir pasirenka tą, kuris geriau atitinka nustatytą sąlygą.
- ABX pirmumo testas (angl. *ABX Preference Test*) – klausytojai klauso trijų pavyzdžių: A, B ir X, kur X yra tikrasis kalbos pavyzdys, ir pasirenka tą, kuris labiau atitinka tikrąjį pavyzdį.

2. Tyrimo duomenys

2.1. Common Voice Lithuanian

„Common Voice“ yra „Mozilla“ sukurtas kalbos korpusas [ABD+19], kuriame bendraautoriai įrašo savo balsą skaitydami ekrane rodomus sakinius. Įrašų kokybę patvirtina kiti bendraautoriai, naudodami balsų sistemą. Balsai yra loginiai pasirinkimai: „už“ arba „prieš“. Visas duomenų rinkinys sudarytas į kelis sub–rinkinius, kurie buvo sudaryti atsižvelgiat į balsų skaičius ir panaudojamumą. Įrašas laikomas tinkamu, kai gauna du „už” balsus iš kitų bendraautorių, o du „prieš” balsai jį atmeta.

„Common Voice Lithuanian” pagrindiniai sub–rinkiniai:

- validated – visi bendruomenės patvirtinti įrašai (apima train ir test)
- train – validated rinkinio poaibis, skirtas dirbtinio intelekto modelių mokymui.
- test – validated rinkinio poaibis skirtas automatinio kalbos atpažinimo modelių vertinimui
- invalidated – įrašai, atmesti dėl blogos kokybės.

Šiame darbe naudotas išskirtinai „Common Voice Lithuanian” train rinkinys. Toliau jis bus žymimas santrumpa CVL.

CVL duomenų rinkinys sudarytas iš 8640 garso įrašų, kuriuos pateikė 8 skirtingi diktoriai. Diktoriai suskirstyti pagal lytį: vyriškoji, moteriškoji ir nenustatyta. Šio rinkinio struktūra – nedidelis diktorių skaičius ir didelis įrašų kiekis vienam diktoriui, tai yra ypač tinkama teksto į kalbą modelių mokymui, nes TTS sistemoms reikalingi kuo išsamesni atskiro diktoriaus balso pavyzdžiai. Priešingai, validated rinkinį sudarė įrašai iš daugiau nei 300 skirtingų diktorių, tačiau vidutiniškai tik apie 10 įrašų vienam diktoriui, todėl jo duomenys nebuvo pasirinkti šiam tyrimui.

CVL duomenų rinkinio tekstai parašyti lietuviškomis raidėmis. Naudoti 38 unikalūs simboliai: standartinės lotynų abėcėlės raidės, lietuviškos diakritinės raidės (ą, č, ė, ę, į, š, ū, ų, ž) bei skyrybos ženklai.

2.2. LIEPA duomenų rinkinys

Šiame darbe naudotas iš projekto LIEPA[LTK+18] paimtas duomenų rinkinys, sudarytas iš 5000 garso įrašų, atliktų vienos diktorės neutralia emocija. Rinkinys suskirstytas į tris dalis: 4700 įrašų skirta mokymui, 100 – validavimui ir 200 – testavimui. Įrašų trukmė svyruoja nuo trumpų žodžių ir frazių iki vidutinio ilgio sakinių. Tekstai užrašyti naudojant lotynų abėcėlės raides bei lietuviškas diakritines raides, o tekstas sukirčiuotas naudojant specialius simbolius: `, ^ ir ~.

2.3. Lithuanian Emotional Speech Set

Šiam darbui modelių apmokymui buvo sukurtas originalus lietuviškos emocinės kalbos duomenų rinkinys LESS. Rinkinį sudaro 960 garso įrašų – 320 sakinių, įrašytų trimis emocinėmis intonacijomis: linksma, liūdna ir neutrali.

2.3.1. LESS sudarymo motyvai

Viešai prieinamų lietuvių kalbos garso duomenų rinkinių skaičius išlieka itin ribotas, o ši problema dar labiau išryškėja kalbant apie emocinius garso duomenis. Dėl tokio duomenų stygiaus šiame darbe buvo nuspręsta sukurti originalų duomenų rinkinį LESS. Emocijoms atvaizduoti pasirinktos trys kategorijos: linksma, liūdna ir neutrali. Linksma ir liūdna emocijos pasirinktos dėl ryškaus akustinio kontrasto ir palyginti nesudėtingo atkartojimo, o neutrali emocija įtraukta kaip bazinė lyginamoji kategorija akustinėje analizėje.

2.3.2. Sakinių įvairovė

Dirbant su ribotais duomenų kiekiais ypač svarbu užtikrinti jų kokybę ir akustinę įvairovę, kad modelis galėtų išmokti kuo platesnį lietuvių kalbos garsų spektrą [VSR+24b]. Sudarant duomenų rinkinį buvo siekiama atrinkti sakinius, kurių ilgis svyruoja nuo 4 iki 10 žodžių ir užtikrinti akustinę įvairovę. Sakiniai apima temines įvairoves tokias kaip orai, profesijos, kasdieninė veikla, vietos.

2.3.3. Diktoriaus informacija

Visi įrašai buvo atlikti vieno vyriškos lyties diktoriaus, priklausančio 25 – 30 metų amžiaus grupei. Visų trijų emocinių kategorijų įrašai buvo atliekami to paties diktoriaus, kas leidžia išvengti tarpasmeninių akustinių skirtumų ir užtikrina, kad pastebėti emociniai skirtumai atspindi išskirtinai emocijų įtaką kalbai, o ne skirtingų balsų ypatybes.

2.3.4. Įrašymo sąlygos

Įrašai atlikti vidutiniame (apie 20 m^2) darbo kambaryje, kurio durys ir langai įrašymo metu buvo uždaryti, siekiant sumažinti aplinkos triukšmą. Įrašymo įrangos charakteristikos pateiktos 3 lentelėje.

3 lentelė. *Įrašymo įrangos charakteristikos*

Parametras	**Reikšmė**
Mikrofonas	HyperX Cloud III
Įrašymo programinė įranga	Audacity
Diskretizavimo dažnis	44 100 Hz
Kodavimas	16 bitų PCM
Kanalų skaičius	Mono
Failų formatas	WAV

2.3.5. Rinkinio statistika

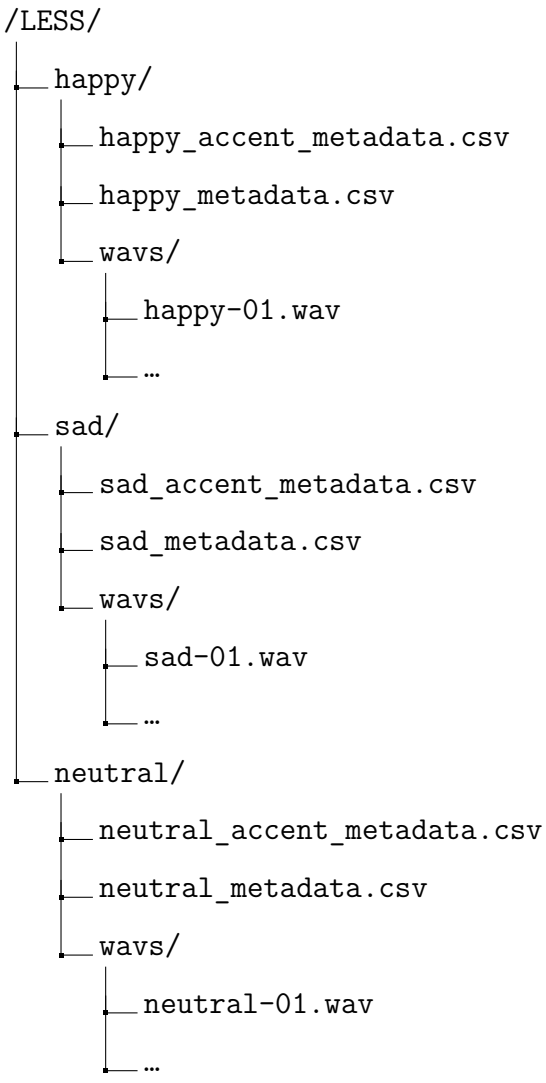
Duomenų rinkinio statistika pateikta 4 lentelėje. Ankstesni tyrimai rodo, kad panašūs duomenų kiekiai yra pakankami sėkmingam perdavimo mokymuisi [IM24; Tha23].

4 lentelė. *LESS duomenų rinkinio statistika*

Parametras	**Reikšmė**
Įrašų skaičius	320 sakinių $\times$3 emocijos = 960 įrašų
Bendra įrašų trukmė	1h 9min
Vidutinė įrašo trukmė	4,32s

2.3.6. Rinkinio struktūra

Duomenų rinkinys buvo sudarytas naudojant LJ Speech formatą. Rinkinio katalogų struktūra atvaizduota 3 pav.



3 pav. *LESS duomenų rinkinio katalogų struktūra*

Metadata failai sudaryti iš stulpelių, kuriuose kiekvienas įrašas pateikia informaciją apie atitinkamą garso įrašą. Garso įrašo informacija susideda iš trijų dalių, atskirtų „|“ simboliu. Toliau aprašomos minėtos dalys:

- Failo identifikatorius – garso įrašo pavadinimas be .wav plėtinio.

- Originalus tekstas – tekstas, kuris buvo skaitomas įrašinėjant failą.
- Normalizuotas tekstas – tekstas, kuris pateikiamas neuroniniam tinklui modelio apmokymo metu.

Šiame darbe visi modeliai buvo apmokyti pasitelkiant end-to-end sistemas, kurios pačios atlieka teksto normalizavimą, todėl LESS duomenų rinkinyje originalus ir normalizuotas tekstas sutampa. Žemiau pateikiami kelių įrašų pavyzdžiai:

5 lentelė. *LJ Speech formato pavyzdys (`happy_metadata.csv`)*

`happy-01|Šiandien oras yra labai gražus.|Šiandien oras yra labai gražus.`
`happy-02|Ji padėjo knygą ant stalo.|Ji padėjo knygą ant stalo.`
`happy-03|Jis atidarė langą ir pažvelgė laukan.|Jis atidarė langą ir pažvelgė laukan.`

2.3.7. Teksto kirčiavimas

Kuriant duomenų rinkinį, buvo parengti dviejų tipų anotacijų metaduomenų failai. Pirmojo tipo failus sudarė lotyniškos abėcėlės raidės, lietuviški diakritiniai ženklai ir skyrybos simboliai. Antrojo tipo failai buvo pažymėti papildomu žodžiu accent pavadinime ir juose kirčiavimas žymėtas naudojant simbolius `` ` ``, `^`, `~`. Žodžių kirčiavimas buvo atliktas automatiškai pasitelkiant žodžių kirčiavimo procesorių iš „LIEPA" kalbos sintezatoriaus. Žemiau pateikiami kelių sukirčiuotų įrašų pavyzdžiai:

6 lentelė. *LJ Speech formato pavyzdys (`happy_accent_metadata.csv`)*

`happy-01|Šian~dien o^ras yra`labai~gražu`s.|Šian~dien o^ras yra`labai~gražu`s.`
`happy-02|Ji`padė^jo kny~gą ant sta~lo.|Ji`padė^jo kny~gą ant sta~lo.`
`happy-03|Ji`s atida~rė la^ngą ir pa`žvelgė laukan~.|Ji`s atida~rė la^ngą ir pa`žvelgė laukan~.`

2.4. Emocinės lietuvių kalbos garsynas

Šiame darbe modelių mokymui taip pat naudotas emocinės lietuvių kalbos garsynas (toliau ELKG) [Vil26], sudarytas kalbos sintezės, emocijų atpažinimo ir emocijų raiškos tyrimų tikslams. Garsyną sudaro 699 392 garso įrašai, kurių bendra trukmė siekia 301 valandą. Įrašai atlikti garso studijoje 89 diktorių, iš kurių 41 vyriškos ir 48 moteriškos lyties. Garso įrašai pateikiami WAV formatu, 96 kHz diskretizavimo dažniu, 24 bitų tikslumu, mono kanalu.

Garsynas apima penkias emocines kategorijas: neutralią, džiaugsmą, liūdesį, nuostabą ir pyktį. Kiekvienas garso įrašas pažymėtas viena emocija. Garsyno diktoriai suskirstyti į tris amžiaus grupes: 13 – 17 metų, 18 – 60 metų ir virš 60 metų.

Šiame darbe naudoti išskirtinai vyriškos lyties 18 – 60 metų amžiaus grupės (M3) diktorių įrašai. Papildomų modelių mokymui pasirinktas vienas diktorius, turintis daugiausiai džiaugsmo ir liūdesio emocijų įrašų. Garsyno įrašai anotuoti frazių, žodžių ir fonemų lygiais, pasitelkiant lietuvių šnekos atpažintuvą ir vėliau tikslinant rankiniu būdu. Garsyno statistika pateikta 7 lentelėje.

7 lentelė. *ELKG garsyno statistika*

Parametras	**Reikšmė**
Garso įrašų skaičius	699 392
Bendra trukmė	301h 9min
Diktorių skaičius	89 (41 vyras, 48 moterys)
Emocinių kategorijų skaičius	5
Diskretizavimo dažnis	96 kHz
Kodavimas	24 bitų PCM, mono
Vidutinė įrašo trukmė	3,1 s

3. Modelių mokymas

Iš viso šiame darbe buvo apmokyti keturi baziniai modeliai bei dvylika papildomo mokymo metodu adaptuotų modelių. Siekiant sistemiškai apibūdinti kiekvieną jų, taikoma vieninga žymėjimo sistema. Modelių pavadinimai sudaromi iš bazinio modelio identifikatoriaus (**B1**, **B2**, **B3**, arba **S**) ir emocijos priesagos:

- **–NEU** – papildomai apmokytas modelis neutralios emocijos duomenimis
- **–LIN** – papildomai apmokytas modelis linksmos emocijos duomenimis
- **–LIU** – papildomai apmokytas modelis liūdnos emocijos duomenimis

Pavyzdžiui, **B1–LIN** žymi modelį, kuris buvo sukurtas naudojant pirmąjį bazinį modelį ir papildomai apmokytas generuoti linksmą emociją.

3.1. Mokymo aplinka

Modelių B1, B2, B3 mokymai ir testavimai buvo atlikti Google Colab Pro aplinkoje, naudojant NVIDIA A100 Tensor Core vaizdo plokštę, kurios parametrai atvaizduoti 8 lentelėje.

8 lentelė. *Pagrindinės eksperimentuose naudotos NVIDIA A100 specifikacijos*

Parametras	**Reikšmė**
Architektūra	Ampere
Atmintis	40 GB HBM2
Atminties pralaidumas	1 555 GB/s
FP32 našumas	19,5 TFLOPS
Tenzorinių branduolių našumas	156 TFLOPS
Mišraus tikslumo palaikymas	FP16, BF16, TF32

Bazinio modelio S mokymai ir papildomi mokymai buvo atlikti Google Colab Pro+ aplinkoje, naudojant NVIDIA RTX PRO 6000 Blackwell grafikos procesorių (Google Colab aplinkoje žymimą G4), kurio pagrindinės specifikacijos pateiktos 9 lentelėje.

9 lentelė. *Pagrindinės eksperimentuose naudotos NVIDIA RTX PRO 6000 Blackwell Server Edition specifikacijos*

Parametras	**Reikšmė**
Architektūra	Blackwell
Atmintis	96 GB GDDR7 ECC
Atminties pralaidumas	1 597 GB/s
FP32 našumas	120 TFLOPS
Tenzorinių branduolių našumas	4 000 AI TOPS (FP4)
Mišraus tikslumo palaikymas	FP4, FP8, FP16, BF16, TF32

3.2. Modelių ir mokymų parametrai

Šiame darbe visi modeliai yra sukurti naudojant VITS architektūrą ir priklauso grafemų į garsą modelių grupei, kurie generuoja garsą iš pateiktos simbolinės įvesties. Toliau pateikiami bendri modelių ir mokymų parametrai, kurie buvo taikomi visiems šiame darbe sukurtiems modeliams.

10 lentelė. *Bendri modelių parametrai.*

Parametras	**Reikšmė**
Modelio architektūra	VITS
Modelio įvesties vaizdavimas	Grafemos
Optimizatorius	AdamW
Mokymosi greitis	2×10^{-4}
Diskretizavimo dažnis	22050 Hz

3.3. Bazinių modelių mokymas

Siekiant apmokyti emocinę kalbą sintezuojančius modelius, visų pirma svarbu parengti kokybiškus bazinius modelius, kurie sintezuoja neutralią kalbą. Šiame darbe buvo apmokyti keturi baziniai modeliai, naudojant skirtingus duomenų rinkinius ir metodus. Bazinių modelių mokymas reikalauja didelių laiko ir skaičiavimo resursų, todėl duomenų kokybės užtikrinimas bei tinkamas modelių architektūros ir parametrų pasirinkimas yra kritiškai svarbūs siekiant užtikrinti efektyvų resursų naudojimą. Baziniai B1, B2, B3 modeliai buvo mokomi 1000 epochų, partijos dydį nustačius 32. Bazinis S modelis buvo mokomas 600 epochų, partijos dydį nustačius 32.

11 lentelė. *Bazinių modelių mokymų hiperparametrai*

Parametras	**Reikšmė**
Epochų skaičius (B1, B2, B3 modeliai)	1000
Epochų skaičius S modelis	600
Partijos dydis	32

3.4. Papildomas modelių mokymas

B1, B2, B3 baziniai modeliai buvo papildomai apmokyti naudojant LESS duomenų rinkinį. S bazinis modelis buvo papildomai mokomas naudojant ELKG rinkinio atrinktus emocinius duomenis. Visi papildomi mokymai buvo atliekami nustačius epochų skaičių 500, o partijos dydį parinkus 16. B1, B2, B3 papildomi mokymai užtruko apie 2 valandas, S modelio papildomi mokymai – apie 4 valandas.

12 lentelė. *Papildomų mokymų hiperparametrai*

Parametras	**Reikšmė**
Epochų skaičius	500
Partijos dydis	16

3.5. B1 modelio mokymas

B1 modelis buvo apmokytas naudojant CVL duomenų rinkinį, į kurį atrinkti išskirtinai vyriškos lyties diktorių įrašai. Toks sprendimas priimtas siekiant užtikrinti garso įrašų vienarūšiškumą tiek bazinio modelio mokymo, tiek vėlesnio papildomo mokymo etapuose naudojant LESS duomenų rinkinį. Modelio mokymas truko 22 valandas. Nors mokymo duomenis įrašė keli skirtingi diktoriai, visi įrašai buvo traktuojami kaip vieno diktoriaus įrašai – tai įprasta strategija, kai tolesnis mokymas planuojamas su vieno diktoriaus duomenimis. Tokiu atveju modelis išmoksta apibendrintą visų mokymo diktorių balsų reprezentaciją. Validacijos metu mel spektrogramos nuostolių reikšmės (loss_mel) stabilizavosi ties 20.

3.6. B1 modelio papildomi mokymai

B1 modelis buvo papildomai apmokomas tris kartus, kiekvieną kartą naudojant skirtingą LESS duomenų rinkinio emocinę kategoriją. Papildomo mokymo metu pastebėta, kad mel spektrogramos nuostolių reikšmė (loss_mel) toliau mažėjo, stabilizuodamasi ties 14. Tai rodo, kad modelis sėkmingai perėmė LESS duomenų rinkinio akustines savybes. Visi trys papildomai apmokyti modeliai įgijo LESS diktoriaus balso charakteristikas su akustiškai pastebimomis emocinėmis intonacijomis.

3.7. B2 modelio mokymas

B2 modelis buvo apmokytas naudojant visą CVL duomenų rinkinį, sudarytą iš 8640 garso įrašų. Priešingai nei B1 modelis, B2 modelio mokymo metu kiekvieno diktoriaus balsas buvo traktuojamas kaip atskiras – modelis išmoko aštuonių skirtingų diktorių balsų reprezentacijas. Sintezės metu pageidaujamas balsas parenkamas pateikiant atitinkamą diktoriaus identifikatorių. Kadangi šiam modeliui apmokyti buvo naudojama daugiausia duomenų iš visų bazinių modelių, mokymas užtruko ilgiausiai – apie 32 valandas. Mokymas buvo sustabdytas ties 800 epocha, pastebėjus, kad nuostolių funkcijų reikšmės stabilizavosi, taip siekiant efektyviau naudoti skaičiavimo resursus.

3.8. B2 modelio papildomi mokymai

Remiantis B2 modeliu, buvo sudaryti trys papildomi modeliai, kiekvienam naudojant skirtingą LESS duomenų rinkinio emocinę kategoriją. Papildomo mokymo procesas buvo suskirstytas į dvi fazes. Pirmosios fazės metu buvo siekiama išmokyti modelį generuoti naujo diktoriaus balsą, keičiant tik nedidelę modelio parametrų dalį – išskirtinai naujo diktoriaus įterpimo vektorių. Pirmoji fazė buvo mokoma 20 epochų. Tačiau visais trimis atvejais pirmoji fazė nepasiekė laukiamo rezultato: modeliai nesugebėjo įsisavinti LESS diktoriaus balso savybių. Tikėtina, kad tai lėmė nedidelis LESS duomenų rinkinio kiekis, kuris buvo nepakankamas efektyviam įterpimo vektoriaus apmokymui. Pirmosios fazės metu mel spektrogramos nuostolių reikšmės (loss_mel) nepakito ir išliko ties 20 – tokiame pat lygyje kaip ir bazinio B2 modelio. Antrosios fazės metu buvo keičiami visi modelio parametrai, įskaitant ir jau išmoktus skirtingų diktorių balsų atvaizdavimus. Ši fazė truko 300 epochų. Visi trys modeliai sėk-

mingai įsisavino LESS diktoriaus balso savybes, o mel spektrogramos nuostolių reikšmė stabilizavosi ties 14. Antrosios fazės rezultatai buvo panašūs į B1 modelio papildomo mokymo rezultatus.

3.9. B3 modelio mokymas

B3 modelis buvo apmokytas naudojant LIEPA duomenų rinkinį. Ankstesni tyrimai rodo, kad sukirčiuotų tekstų naudojimas lietuviškų TTS modelių mokymui leidžia pasiekti aukštesnę sintezės kokybę ir natūralesnį kalbos skambesį [Rad22]. B3 modelio mokymas truko apie 20 valandų, o mel spektrogramos nuostolių reikšmė stabilizavosi ties 17. B3 modelis įsisavino LIEPA rinkinio diktorės balso savybes ir sugebėjo generuoti kalbą tiek iš sukirčiuoto, tiek iš nekirčiuoto teksto. Vis dėlto modeliui pateikiant sukirčiuotą tekstą generuojamas balsas buvo pastebimai kokybiškesnis, kas patvirtina ankstesnių tyrimų rezultatus. Lyginant su B1 ir B2 modeliais, B3 modelis sintezavo kokybiškiausią kalbą. Tikėtina, kad tai lėmė duomenų rinkinio kokybė – aukštos kokybės vienos diktorės įrašai – bei sukirčiuotų tekstų anotavimas.

3.10. B3 modelio papildomi mokymai

Remiantis B3 baziniu modeliu buvo apmokyti trys papildomi modeliai, naudojant LESS duomenų rinkinį. Nors B3 bazinis modelis sintezavo kokybiškiausią kalbą iš visų bazinių modelių, papildomai apmokyti modeliai neįsisavino ryškių emocinių savybių. Visi papildomi modeliai perėmė LESS diktoriaus kalbos akustines charakteristikas, tačiau klausantis atskirti emocijas nepavyko – visi trys modeliai skambėjo panašiai ir labiausiai priminė LESS džiaugsmo emocijos įrašus. Verta paminėti, jog modeliams pateikiant sukirčiuotą tekstą rezultatai buvo pastebimai geresni nei pateikiant nekirčiuotą tekstą.

3.11. S modelio mokymas

S bazinis modelis apmokytas naudojant ELKG duomenų rinkinį. Mokymo duomenims atrinkti naudoti M3 amžiaus grupės (18–60 m. vyriškos lyties) diktorių neutralios emocijos įrašai, išskyrus 26 diktorių. Šis diktorius pašalintas iš bazinio mokymo duomenų, kadangi turėjo daugiausiai emocinių įrašų iš visų M3 grupės diktorių ir buvo paskirtas papildomam mokymui – taip užtikrinama, kad bazinis modelis nebuvo matęs šio diktoriaus balso prieš papildomą mokymą. S modeliui panaudota žymiai daugiau mokymo duomenų nei kitiems baziniams modeliams – apie 3,25 karto daugiau nei B2 modeliui ir apie 6 kartus daugiau nei B1 bei B3 modeliams. Mokymo tekstai buvo pateikti sukirčiuota forma, analogiškai kaip B3 modelio atveju. S modelio mokymas truko apie 30 valandų, mel spektogramos nuostolių funkcijos reikšmė stabilizavosi ties 17. Subjektyvaus klausymo metu nustatyta, kad S bazinis modelis generavo natūraliausią neutralią kalbą iš visų šiame darbe sukurtų bazinių modelių.

3.12. S modelio papildomi mokymai

Remiantis S baziniu modeliu buvo apmokyti trys papildomi modeliai: S-LIN, S-NEU ir S-LIU, naudojant ELKG duomenų rinkinio 26 diktoriaus emocinius įrašus. Lyginant su LESS duomenų rinkiniu,

kiekviena emocinė kategorija turėjo apie 7,5 karto daugiau įrašų, kas leido tikėtis geresnio emocinių savybių perėmimo. Kiekvieno papildomo modelio mokymas truko apie 4 valandas. Subjektyvaus klausymo metu nustatyta, kad S papildomi modeliai generavo natūraliausiai skambančią kalbą iš visų šiame darbe sukurtų emocinių modelių.

3.13. Modelių apibendrinimas

Šiame darbe iš viso buvo apmokyti šešiolika modelių — keturi baziniai ir dvylika papildomai apmokytų modelių. Kiekvienas bazinis modelis apmokytas naudojant skirtingą duomenų rinkinį ir mokymo strategiją. Remiantis B1, B2 ir B3 baziniais modeliais, apmokyti devyni papildomi modeliai naudojant LESS duomenų rinkinį, o remiantis S baziniu modeliu — trys papildomi modeliai naudojant ELKG duomenų rinkinį. Bazinių modelių apibendrinimas pateikiamas 13 lentelėje, papildomai apmokytų modelių — 14 lentelėje.

13 lentelė. *Bazinių modelių apibendrinimas*

Modelis	**Duomenų rinkinys**	**Mokymo strategija**
B1	CVL (vyriškos lyties diktoriai)	Vieno diktoriaus
B2	CVL (visi diktoriai)	Kelių diktorių
B3	LIEPA	Vieno diktoriaus (sukirčiuotas tekstas)
S	ELKG (M3 neutrali, be 26 diktoriaus)	Vieno diktoriaus (sukirčiuotas tekstas)

14 lentelė. *Papildomai apmokytų modelių apibendrinimas*

Modelis	**Bazinis modelis**	**Duomenų rinkinys**	**Emocija**
B1-LIN	B1	LESS	Linksma
B1-NEU	B1	LESS	Neutrali
B1-LIU	B1	LESS	Liūdna
B2-LIN	B2	LESS	Linksma
B2-NEU	B2	LESS	Neutrali
B2-LIU	B2	LESS	Liūdna
B3-LIN	B3	LESS	Linksma
B3-NEU	B3	LESS	Neutrali
B3-LIU	B3	LESS	Liūdna
S-LIN	S	ELKG (26 diktoriaus)	Linksma
S-NEU	S	ELKG (26 diktoriaus)	Neutrali
S-LIU	S	ELKG (26 diktoriaus)	Liūdna

4. Modelių emocijų perteikimo įvertinimas

Siekiant įvertinti apmokytų modelių gebėjimą perteikti emocijas, buvo atlikta sintetizuotos kalbos akustinė analizė. Analizei naudotas specialiai sudarytas LESS testavimo rinkinys, kurį sudarė 20 sakinių, įrašytų trimis emocinėmis intonacijomis – linksma, neutrali ir liūdna – iš viso 60 garso įrašų. Tie patys 20 sakinių buvo sintetizuoti kiekvienu iš devynių papildomai apmokytų modelių, todėl analizei iš viso buvo sugeneruota 180 sintetizuotų garso įrašų.

4.1. Įvertinimo planas

Modelių emocijų perteikimo įvertinimas atliktas šiais etapais:

1. Sudarytas LESS testavimo rinkinys – įrašyti 20 sakinių trimis emocinėmis intonacijomis (linksma, neutrali, liūdna), iš viso 60 garso įrašų. Testavimo rinkinys nebuvo naudojamas modelių mokymui.
2. Iš kiekvieno realaus testavimo įrašo gauti keturi akustiniai požymiai naudojant Python biblioteką `parselmouth`.
3. Devyni papildomai apmokyti modeliai sintetizavo tuos pačius 20 sakinių, iš viso sugeneruojant 180 sintetizuotų garso įrašų.
4. Iš kiekvieno sintetizuoto įrašo ištraukti tie patys keturi akustiniai požymiai.
5. Kiekvienam modeliui apskaičiuotas normalizuotas akustinis atstumas tarp sintetizuotos kalbos ir realių testavimo įrašų kiekvienai iš trijų emocinių kategorijų.
6. Rezultatai palyginti ir atvaizduoti naudojant akustinio atstumo matricas.

4.2. Akustinė analizė

Šiame skyriuje pateikiami eksperimentų rezultatai, gauti atlikus akustinę sintetizuotos ir realios kalbos analizę. Vertinimui naudotas normalizuotas akustinis atstumas – kiekvieno modelio sintetizuotos kalbos akustinių savybių palyginimas su realių testavimo įrašų akustinėmis savybėmis. Analizei pasirinkti keturi akustiniai požymiai [ESS+15]:

- pagrindinio tono vidurkis – vidutinis balso aukštis sakinyje, matuojamas Hz, linksma kalba pasižymi aukštesniu tonu nei liūdna ar neutrali.
- pagrindinio tono standartinis nuokrypis – balso aukščio kintamumas sakinyje (Hz), linksma kalba melodingesnė ir išraiškingesnė, liūdna monotoniškesnė ir stabilesnė.
- pagrindinio tono diapazonas – skirtumas tarp aukščiausio ir žemiausio balso taško sakinyje (Hz), linksma kalba pasižymi platesniu diapazonu, liūdna ir neutrali – siauresniu.

- įgarsinimo koeficientas – atspindi įgarsintos kalbos dalį viso įrašo atžvilgiu. Aukštesnė reikšmė rodo mažiau pauzių ir sklandesnę kalbą, žemesnė – dažnesnes pauzes ir fragmentiškesnę kalbą, būdingą liūdnai emocinei raiškai.

Kiekvienam modeliui buvo apskaičiuotas normalizuotas akustinis atstumas lyginant su realiais linksmos, neutralios ir liūdnos emocijų testavimo įrašais. Mažesnė atstumo reikšmė rodo didesnį akustinį panašumą, todėl pagal šiuos duomenis galima įvertinti ar kiekvienas modelis akustiškai artimiausias savo tikslinei emocijai.

Kiekvieno požymio normalizuotas skirtumas skaičiuojamas pagal formulę:

$$d_i = \frac{|s_i - r_i|}{|r_i|} \tag{1}$$

kur s_i – sintetizuoto modelio i-ojo požymio vidurkis, r_i – realių testavimo įrašų i-ojo požymio vidurkis.

Galutinis normalizuotas akustinis atstumas apskaičiuojamas kaip visų požymių normalizuotų skirtumų aritmetinis vidurkis:

$$D = \frac{1}{N}\sum_{i=1}^{N} d_i = \frac{1}{N}\sum_{i=1}^{N} \frac{|s_i - r_i|}{|r_i|} \tag{2}$$

kur $N = 4$ – požymių skaičius (pagrindinio tono vidurkis, standartinis nuokrypis, diapazonas ir įgarsinimo koeficientas).

4.3. B modelių akustinė analizė

4.3.1. Testavimo duomenų akustiniai požymiai

Toliau 15 lentelėje pateikiami testavimo LESS rinkinio realių įrašų akustinių savybių vidurkiai. Iš šių duomenų galima įvertinti testavimo duomenų rinkinio kokybę. Linksmos emocijos įrašai turėjo didžiausias reikšmes pagal visus požymius, kas atitinka linksmos emocijos balso akustines charakteristikas. Neutralios ir liūdnos emocijos reikšmės buvo panašios, išskyrus įgarsinimo koeficientą. Testavimo rinkinio sudarymo metu buvo skiriamas didesnis dėmesys pauzėms tarp žodžių, lyginant su mokymo duomenų rinkinio sudarymu. Šis skirtumas atsispindės vėliau lyginant modelių įgarsinimo koeficientų atstumus.

15 lentelė. *Testinio rinkinio realių įrašų akustinių savybių vidurkiai*

Emocija	F0 vidurkis (Hz)	F0 std (Hz)	F0 diapazonas (Hz)	Įgarsinimo koeficientas
Linksma	134,9	28,7	132,0	0,433
Neutrali	114,1	15,0	64,1	0,352
Liūdna	118,3	12,8	65,5	0,270

4.3.2. Sintezuotų garsų akustiniai požymiai

Naudojant kiekvieną papildomai apmokytą modelį, buvo sugeneruoti tie patys sakiniai, kurie sudaro LESS testavimo duomenų rinkinį. Iš viso buvo sugeneruota $9 \times 20 = 180$ sintetizuotų garso įrašų. Pagal sugeneruotus įrašus apskaičiuotos akustinės savybės kiekvienam modeliui atskirai, rezultatai pateikti 16 lentelėje.

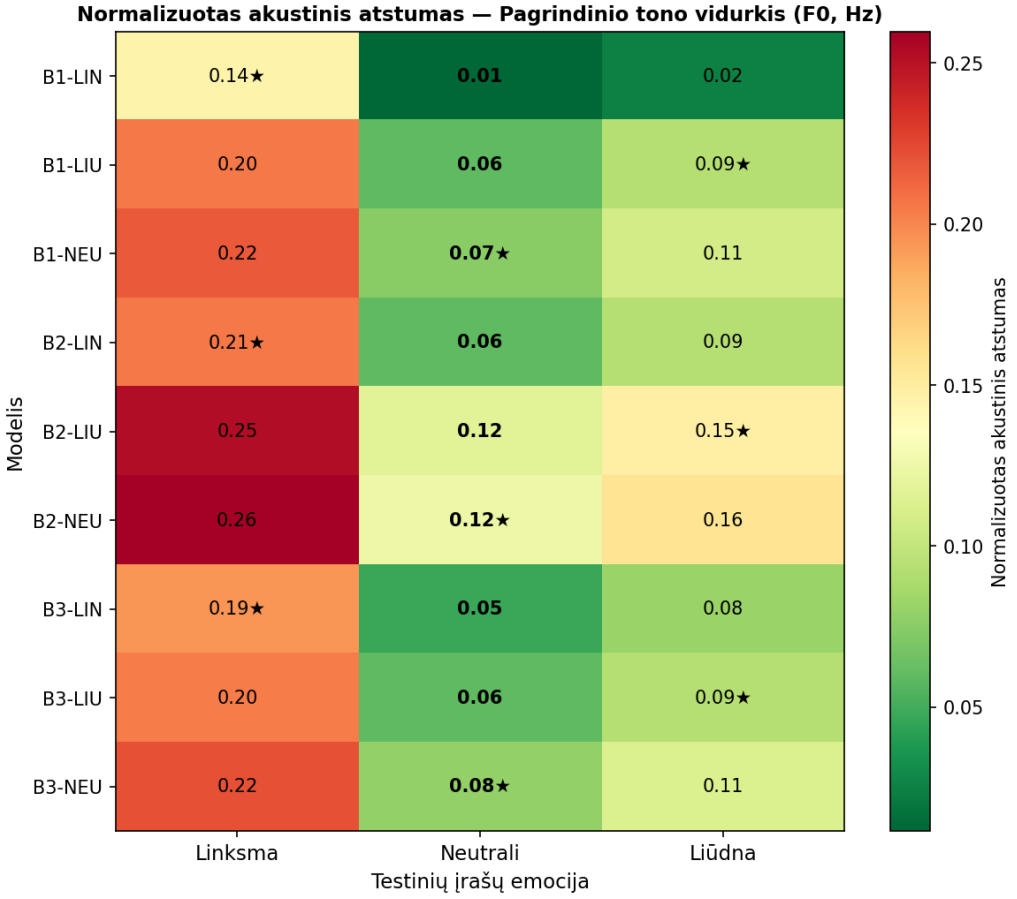
16 lentelė. *Sintetizuotų modelių akustinių savybių vidurkiai*

Modelis	F0 vidurkis (Hz)	F0 std (Hz)	F0 diapazonas (Hz)	Įgarsinimo koeficientas
B1-LIN	115,4	14,7	77,9	0,441
B1-LIU	107,3	11,4	80,9	0,380
B1-NEU	105,5	13,6	84,4	0,380
B2-LIN	107,2	13,9	56,6	0,472
B2-LIU	100,8	12,8	64,6	0,391
B2-NEU	99,8	12,8	64,6	0,362
B3-LIN	108,7	12,5	66,4	0,460
B3-LIU	107,3	6,0	37,3	0,404
B3-NEU	105,0	9,1	52,1	0,421

Pagal duomenis galima matyti, jog akustiškai geriausiai linksmą emociją sintezavo B1-LIN modelis – tai rodo aukštas pagrindinio tono vidurkis ir aukštas įgarsinimo koeficientas. Autoriaus nuomone šios dvi savybės patikimiausiai atspindi emociją, kadangi pagrindinio tono standartinis nuokrypis ir diapazonas ne visada vienareikšmiškai skiria emocijas – jų reikšmės gali sutapti skirtingų emocijų įrašuose.

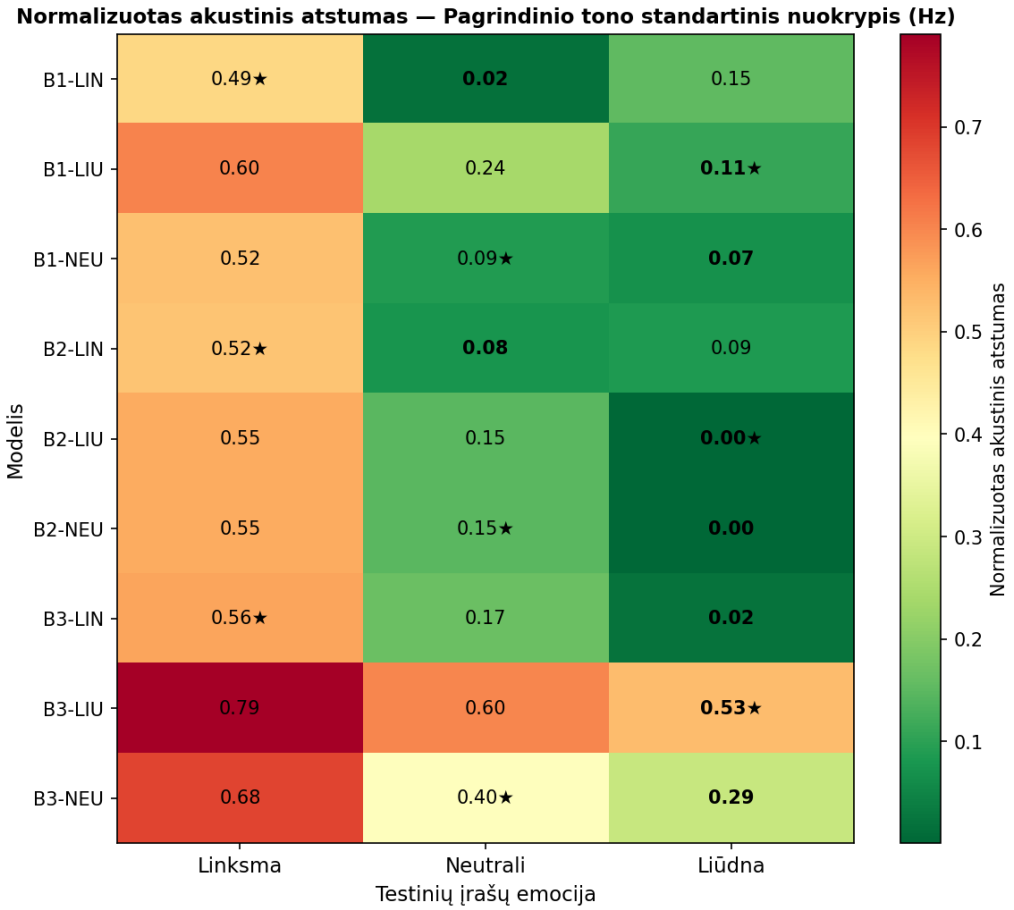
4.3.3. Akustinių savybių atstumų analizė

Toliau pateikiamos normalizuoto akustinio atstumo matricos kiekvienam iš keturių akustinių požymių atskirai, bei apibendrinta matrica, apimanti visų požymių vidurkį.



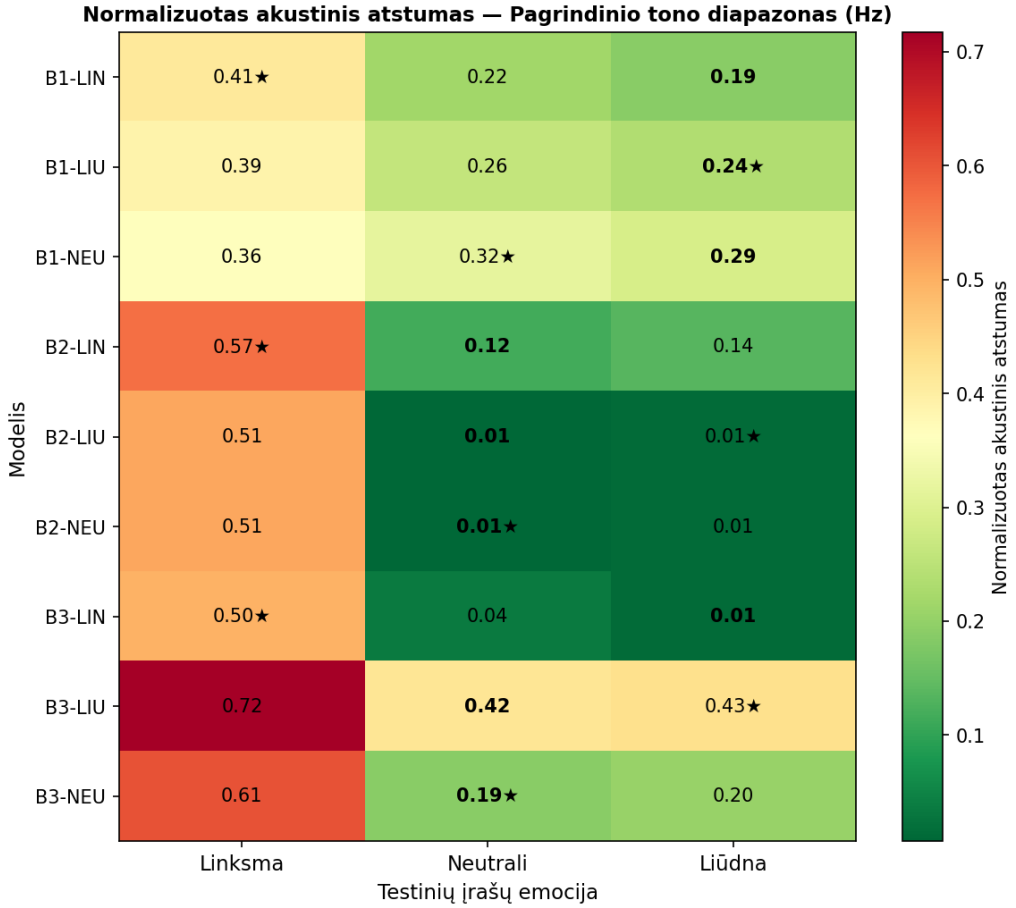
4 pav. *Pagrindinio tono vidurkio atstumų matrica*

4 pav. pateikiama pagrindinio tono vidurkio normalizuoto atstumo matrica. Pagal pateiktus duomenis matyti, kad visi modeliai generavo įrašus, kurių pagrindinio tono vidurkis buvo akustiškai artimesnis neutralios ir liūdnos emocijos testavimo įrašams, nei linksmos. Tai galėjo lemti bazinių modelių mokymo duomenys – CVL ir LIEPA rinkiniai apima skirtingus diktorius, kurių balsų aukštis skiriasi nuo LESS testavimo rinkinio diktoriaus. Vis dėlto, B1-LIN modelis iš visų modelių buvo arčiausiai linksmos emocijos pagrindinio tono vidurkio, kas rodo kad papildomas mokymas linksmos emocijos duomenimis šiek tiek pakėlė sintetizuojamo balso aukštį tikėtina kryptimi. Neutralios emocijos modeliai (B1-NEU, B2-NEU, B3-NEU) pasiekė mažiausią atstumą iki neutralios emocijos testavimo įrašų. Tai yra laukiamas rezultatas, kadangi tiek baziniai modeliai, tiek neutralios emocijos testavimo įrašai pasižymi panašiomis akustinėmis charakteristikomis. Taip pat pastebimas nedidelis skirtumas tarp neutralios ir liūdnos emocijos reikšmių – šios dvi emocijos pagal pagrindinio tono vidurkį akustiškai mažiausiai skyrėsi viena nuo kitos, kas atsispindi ir gautuose rezultatuose.



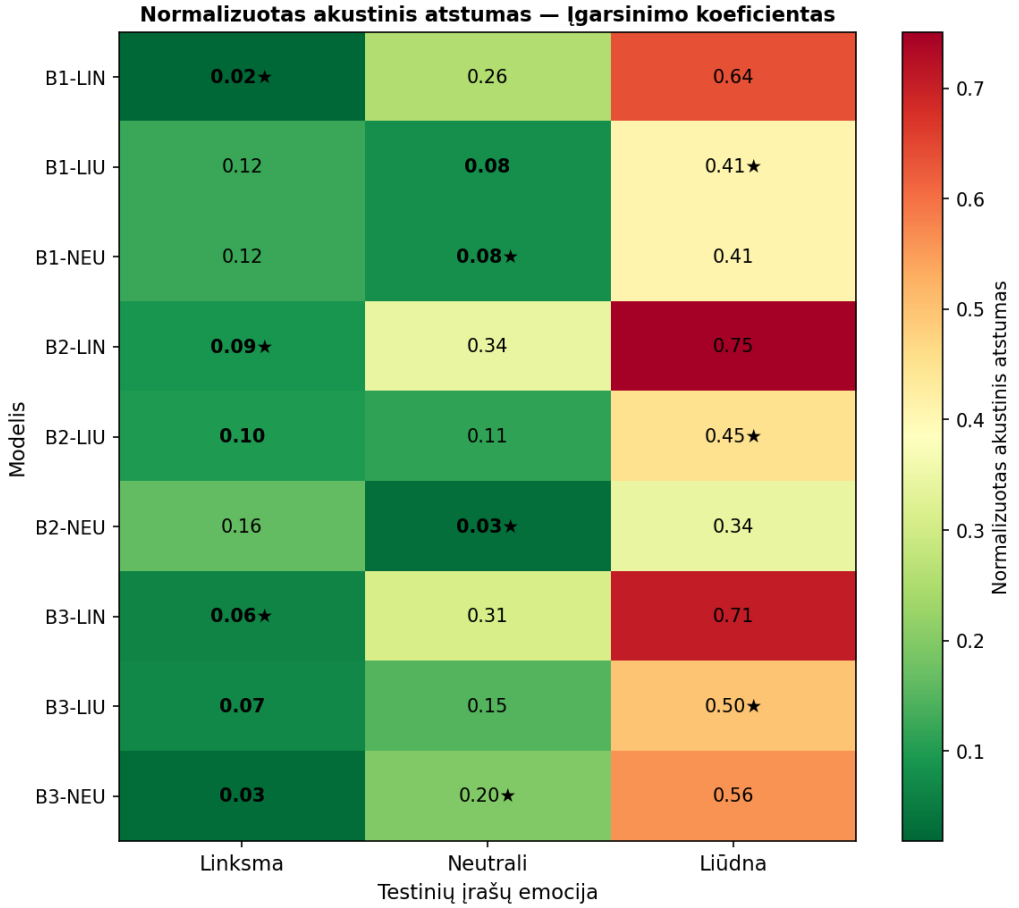
5 pav. *B modelių pagrindinio tono standartinio nuokrypio atstumų matrica*

5 pav. pateikiama pagrindinio tono standartinio nuokrypio normalizuoto atstumo matrica. Panašiai kaip ir 4 pav. galima pastebėti, jog B1-LIN modelis iš visų modelių buvo arčiausiai linksmos emocijos standartinio nuokrypio charakteristikų. Taip pat matyti, jog B1 ir B2 bazinių modelių papildomai apmokyti modeliai buvo akustiškai artimesni testavimo įrašams nei B3 pagrindu apmokyti modeliai. Tai galėjo lemti skirtingas mokymo duomenų pobūdis – B3 bazinis modelis buvo apmokytas naudojant moteriško balso įrašus (LIEPA rinkinys), tuo tarpu papildomo mokymo duomenys (LESS rinkinys) sudaryti išskirtinai iš vyriško balso įrašų. Šis lyties skirtumas tarp bazinio modelio ir papildomo mokymo duomenų galėjo apsunkinti akustinių emocijos savybių perėmimą.



6 pav. *B modelių pagrindinio tono diapazono atstumų matrica*

6 pav. pateikiama pagrindinio tono diapazono normalizuotų atstumų matrica. Iš pateiktų rezultatų matomos panašios tendencijos kaip ankstesniuose paveikslėliuose – B1 ir B2 bazinių modelių papildomai apmokyti modeliai akustiškai buvo artimesni testavimo įrašams nei B3 pagrindu apmokyti modeliai. Autoriaus nuomone, pagrindinio tono diapazonas yra silpniausias iš naudotų rodiklių emocijų perteikimui įvertinti. Kaip minėta anksčiau, net ir emociškai tiksliai sintezuoti įrašai gali turėti didelį diapazono atstumą dėl pavienių aukšto dažnio garsų, kurie reikšmingai iškreipia šią metriką. Diapazonas šiame darbe naudojamas kaip papildomas patvirtinimo rodiklis – siekiant patikrinti ar gauti rezultatai sutampa su pagrindinio tono vidurkio ir įgarsinimo koeficiento atstumų tendencijomis. Šio darbo atveju rezultatai yra nuoseklūs – panašios tendencijos matomos visuose trijuose F0 požymių paveikslėliuose.



7 pav. *B modelių įgarsinimo koeficiento atstumų matrica*

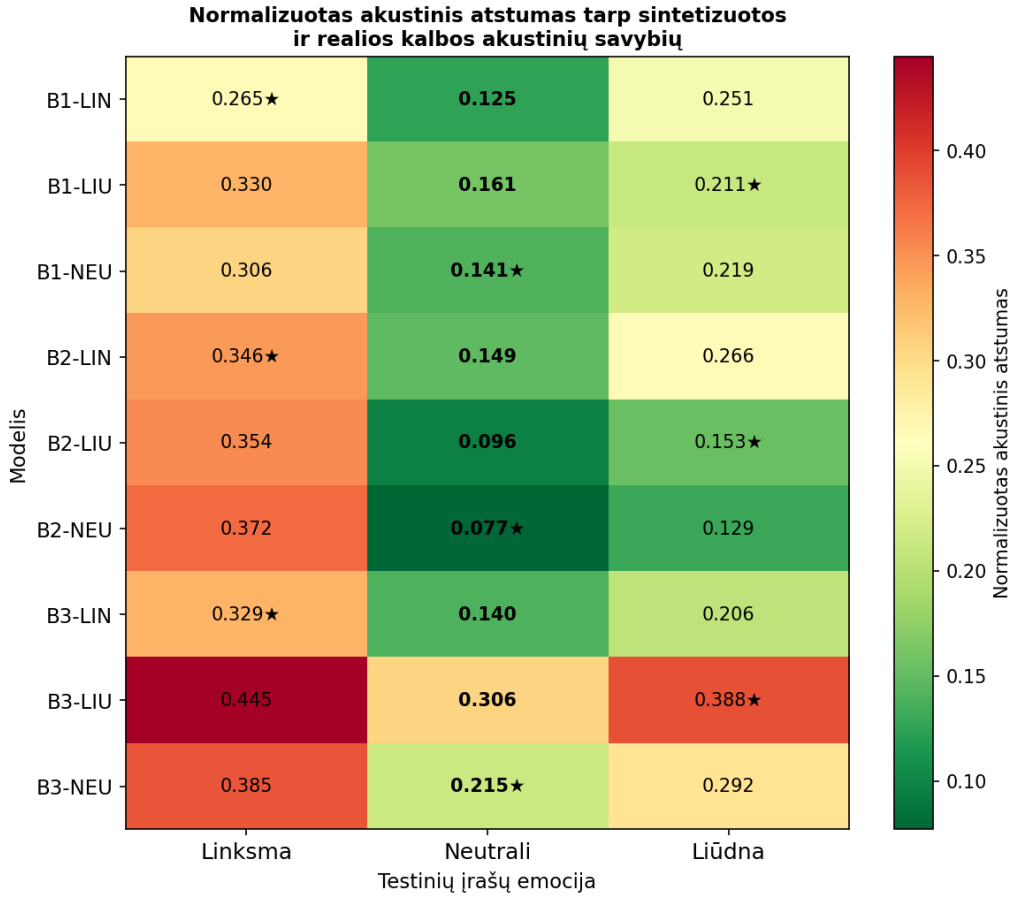
7 pav. pateikiama įgarsinimo koeficiento normalizuotų atstumų matrica. Iš pateiktų rezultatų galima pastebėti, kad linksmos ir neutralios emocijų papildomai apmokyti modeliai sintezavo kalbą panašiu įgarsinimo koeficientu kaip ir testavimo įrašai. Tai rodo, kad papildomo mokymo metu modeliai sugebėjo iš dalies perteikti emocijoms būdingas pauzes – linksmos emocijos kalboje pauzės trumpesnės, liūdnos – ilgesnės. Vis dėlto liūdnos emocijos modeliams nepavyko atkurti ilgų pauzių, būdingų testavimo liūdnos emocijos įrašams. Tai galėjo lemti skirtingas įrašymo metodas – testavimo rinkinio sudarymo metu liūdnos emocijos įrašuose buvo sąmoningai skiriamas didesnis dėmesys pauzėms, nei mokymo duomenų rinkinio įrašymo metu. Turėdami LESS mokymo rinkinio įgarsinimo koeficiento vidurkius, būtų galima patikrinti šią prielaidą.

4.3.4. Rezultatų apibendrinimas

8 pav. pateikiamas visų keturių aptartų akustinių požymių normalizuotų atstumų vidurkis. Iš pateiktos apibendrintos matricos ir anksčiau aptartų atskirų požymių rezultatų matyti, kad B1 ir B2 baziniais modeliais paremti papildomi modeliai sintezavo akustiškai panašesnę kalbą į testavimo įrašų emocijas nei B3 pagrindu apmokyti modeliai. Manoma, kad tai lėmė mokymo duomenų lyties sudėtis – B1 modelis buvo apmokytas naudojant vyriškos lyties diktorių įrašus, B2 – tiek vyriškos, tiek moteriškos lyties diktorių įrašus, tuo tarpu B3 apmokytas išskirtinai naudojant vienos moteriškos lyties diktorės įrašus. Kadangi LESS duomenų rinkinys sudarytas iš vyriškos lyties diktoriaus įrašų, B3

modelio mokymo ir papildomo mokymo duomenų akustinis kontrastas buvo didžiausias, kas galėjo neigiamai paveikti emocinių savybių perėmimą papildomo mokymo metu.

Didžiausias akustinis atstumas tarp papildomai apmokytų modelių ir testavimo įrašų buvo nustatytas lyginant linksmos emocijos rezultatus. Geriausiai linksmos emocijos akustines savybes perteikė B1-LIN modelis, kas sutapo ir su subjektyviais pastebėjimais klausantis sintetizuotų įrašų. Manoma, kad gerą rezultatą lėmė B1 bazinio modelio mokymo strategija – atrenkant išskirtinai vyriškos lyties diktorių įrašus, bazinio modelio balsas akustiškai buvo arčiausiai LESS diktoriaus balso. Nors B3 bazinis modelis mokymo metu pasiekė žemiausią mel spektrogramos nuostolių reikšmę ir generavo natūraliausią neutralią kalbą iš visų bazinių modelių, B3 pagrindu apmokyti papildomi modeliai perėmė mažiausiai emocinių akustinių savybių. Tai rodo, kad bazinio modelio ir papildomo mokymo duomenų diktorių akustinis panašumas yra svarbus veiksnys – kuo panašesni diktoriai, tuo sėkmingiau modelis perima tikslinės emocijos akustines charakteristikas.



8 pav. *B modelių normalizuoto akustinio atstumo matrica tarp sintezuotos ir realios kalbos akustinių savybių*

Galiausiai pažymėtina, kad nors normalizuotas akustinis atstumas yra naudingas rodiklis vertinant emocinių akustinių savybių perėmimą, jis ne visada gali objektyviai atspindėti šį procesą. Siekiant užtikrinti patikimesnius rezultatus, svarbu kiek įmanoma labiau užtikrinti mokymo ir testavimo duomenų rinkinių akustinę kokybę ir pastovumą. Emocinės kalbos įrašymas yra sudėtingas procesas dėl subjektyvaus emocijų raiškos pobūdžio – skirtingi žmonės tą pačią emociją gali išreikšti skirtingai, o tas pats žmogus skirtingų įrašymo seansų metu gali nenuosekliai perteikti emocijas. Dėl to sudarant

mokymo ir testavimo rinkinius rekomenduojama tikrinti kiekvieno įrašo akustinių savybių reikšmes ir įrašus, kurių reikšmės nepatenka į iš anksto nustatytas ribas, atmesti arba perrašyti iš naujo. Toks automatizuotas kokybės filtravimas leistų užtikrinti didesnį duomenų rinkinių vientisumą ir patikimumą.

4.4. S modelių akustinė analizė

4.4.1. Testavimo duomenų akustiniai požymiai

S modelių vertinimui naudoti ELKG duomenų rinkinio 26 diktoriaus testavimo įrašai – kiekvienai emocinei kategorijai atrinkta po 10% mokymo duomenų, iš viso 717 garso įrašų. Testavimo rinkinio realių įrašų akustinių savybių vidurkiai pateikti 17 lentelėje.

17 lentelė. *S modelių testavimo rinkinio realių įrašų akustinių savybių vidurkiai*

Emocija	F0 vidurkis (Hz)	F0 std (Hz)	F0 diapazonas (Hz)	Įgarsinimo koeficientas
Linksma	169,4	36,8	142,8	0,687
Neutrali	111,0	15,4	78,7	0,636
Liūdna	106,4	10,4	62,7	0,647

Lyginant su LESS testavimo rinkiniu, ELKG testavimo įrašų linksmos emocijos pagrindinio tono vidurkis yra žymiai aukštesnis (169,4 Hz palyginti su 134,9 Hz), standartinis nuokrypis taip pat didesnis, o įgarsinimo koeficientas aukštesnis. Tai rodo intensyvesnę emocinę raišką studijos sąlygomis įrašytuose garsyno duomenyse. Neutralios ir liūdnos emocijų reikšmės yra panašios tarpusavyje, tačiau skirtingai nei LESS rinkinyje, įgarsinimo koeficientai beveik nesiskiria.

4.4.2. Sintezuotų garsų akustiniai požymiai

Naudojant kiekvieną S papildomai apmokytą modelį, buvo sintetizuoti tie patys sakiniai, kurie sudaro S modelių testavimo rinkinį. Sintetizuotų įrašų akustinių savybių vidurkiai pateikti 18 lentelėje.

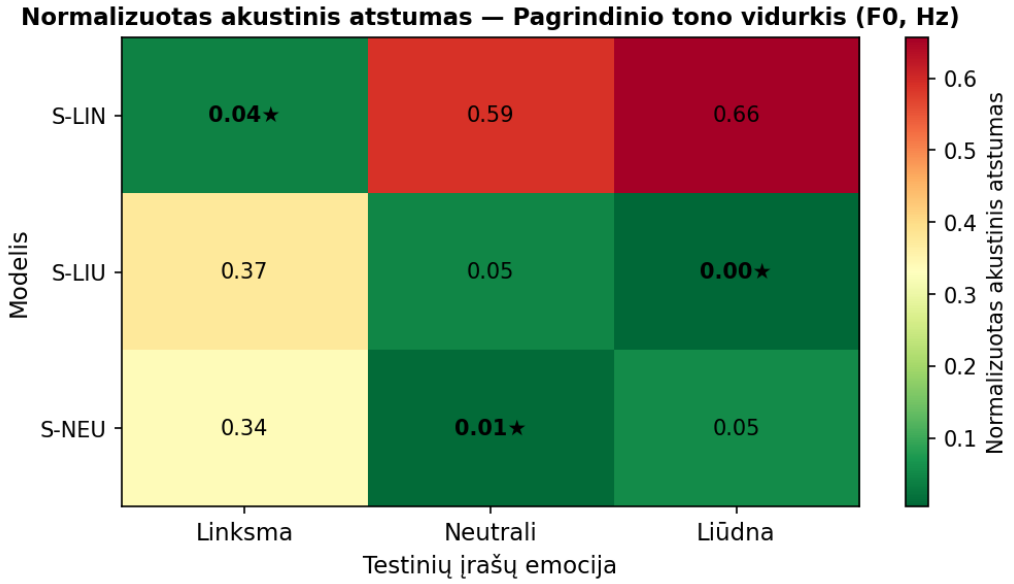
18 lentelė. *S modelių sintetizuotų įrašų akustinių savybių vidurkiai*

Modelis	F0 vidurkis (Hz)	F0 std (Hz)	F0 diapazonas (Hz)	Įgarsinimo koeficientas
S-LIN	176,3	35,9	143,7	0,709
S-NEU	112,1	16,0	76,8	0,649
S-LIU	105,9	8,4	48,0	0,666

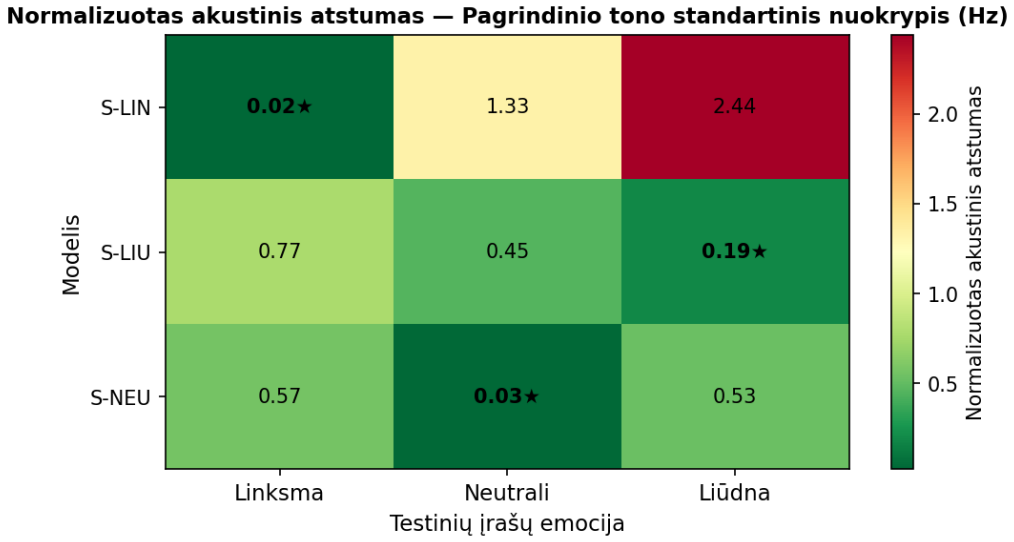
Lyginant su B modelių sintetizuotais įrašais, S modelių akustinės savybės labiau skiriasi tarp emocinių kategorijų. S-LIN modelio pagrindinio tono vidurkis (176,3 Hz) viršija net realių linksmos emocijos įrašų vidurkį (169,4 Hz), o standartinis nuokrypis (35,9 Hz) yra artimas realių įrašų reikšmei (36,8 Hz). S-NEU ir S-LIU modelių reikšmės taip pat glaudžiai atitinka testavimo įrašų charakteristikas.

4.4.3. Akustinių savybių atstumų analizė

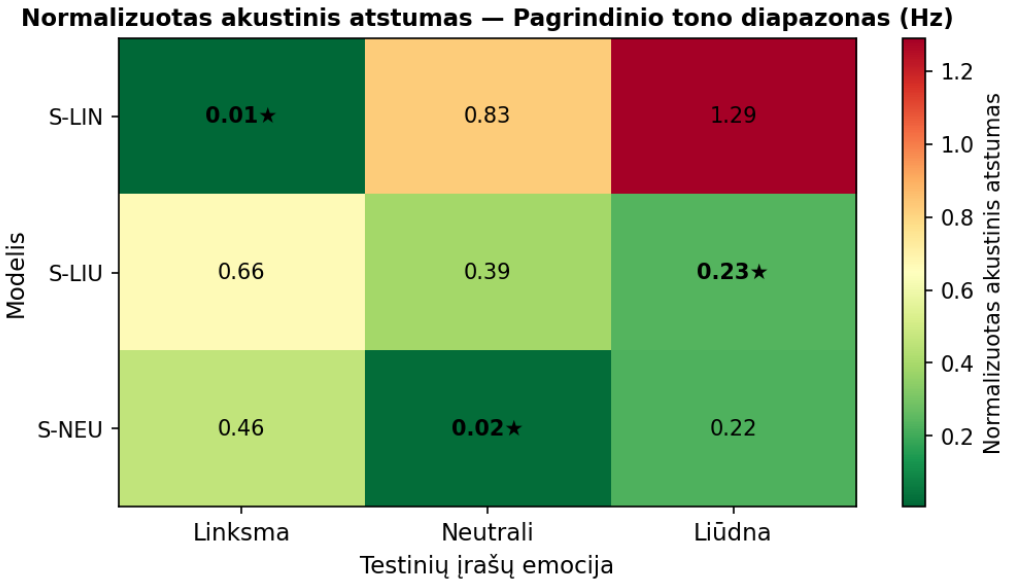
Analogiškai B grupės modelių vertinimui, S grupės papildomai apmokytų modelių akustiniai atstumai apskaičiuoti kiekvienam iš keturių akustinių požymių. Gauti rezultatai pateikiami žemiau esančiose atstumų matricose.



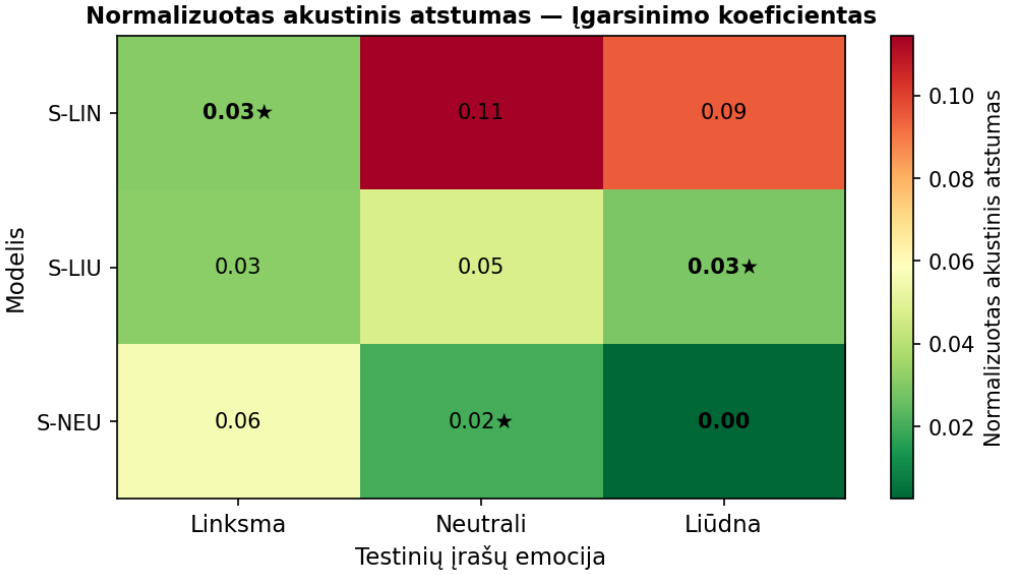
9 pav. *S modelių pagrindinio tono vidurkio atstumų matrica*



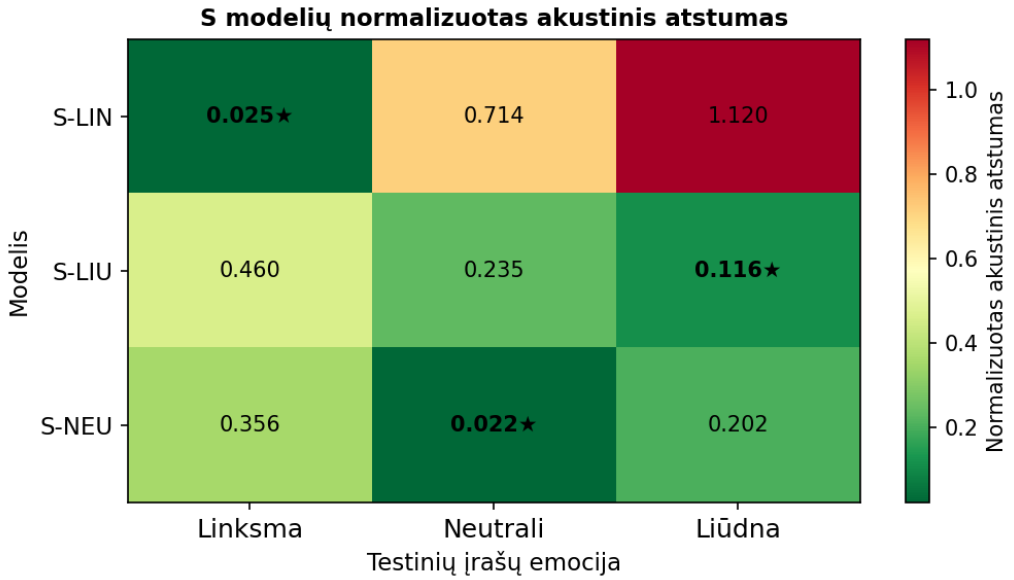
10 pav. *S modelių pagrindinio tono standartinio nuokrypio atstumų matrica*



11 pav. *S modelių pagrindinio tono diapazono atstumų matrica*



12 pav. *S modelių įgarsinimo koeficiento atstumų matrica*



13 pav. *S modelių normalizuoto akustinio atstumo matrica tarp sintezuotos ir realios kalbos akustinių savybių*

4.4.4. Rezultatų apibendrinimas

Pateiktose atstumų matricose matyti, kad visi trys S grupės modeliai akustiškai artimiausi savo tikslinei emocijai pagal visus keturis analizuotus požymius – tai patvirtina ir apibendrintoji matrica. S-LIN modelis pasiekė mažiausią atstumą iki linksmos emocijos testavimo įrašų, S-NEU – iki neutralios, S-LIU – iki liūdnos. Šis rezultatų nuoseklumas išlaikomas visuose atskiruose požymių matricose, kas rodo patikimą emocinių akustinių savybių perėmimą papildomo mokymo metu. Tiek iš akustinių požymių reikšmių, tiek iš gautų atstumų matyti, kad liūdna ir neutrali emocijos akustiškai yra panašesnės viena į kitą nei į linksmą emociją. Tai atsispindi ir atstumų matricose – S-LIN modelio atstumas iki liūdnos ir neutralios emocijos yra žymiai didesnis nei iki tikslinės linksmos emocijos. Analogiškai, S-LIU ir S-NEU modelių atstumai tarp liūdnos ir neutralios emocijų yra palyginti maži, tačiau abiejų modelių atstumas iki linksmos emocijos – žymiai didesnis. Tai rodo, kad S grupės modeliai ne tik sėkmingai perėmė tikslinės emocijos akustines savybes, bet ir aiškiai atskyrė ją nuo kitų emocinių kategorijų.

4.5. S ir B modelių rezultatų palyginimas

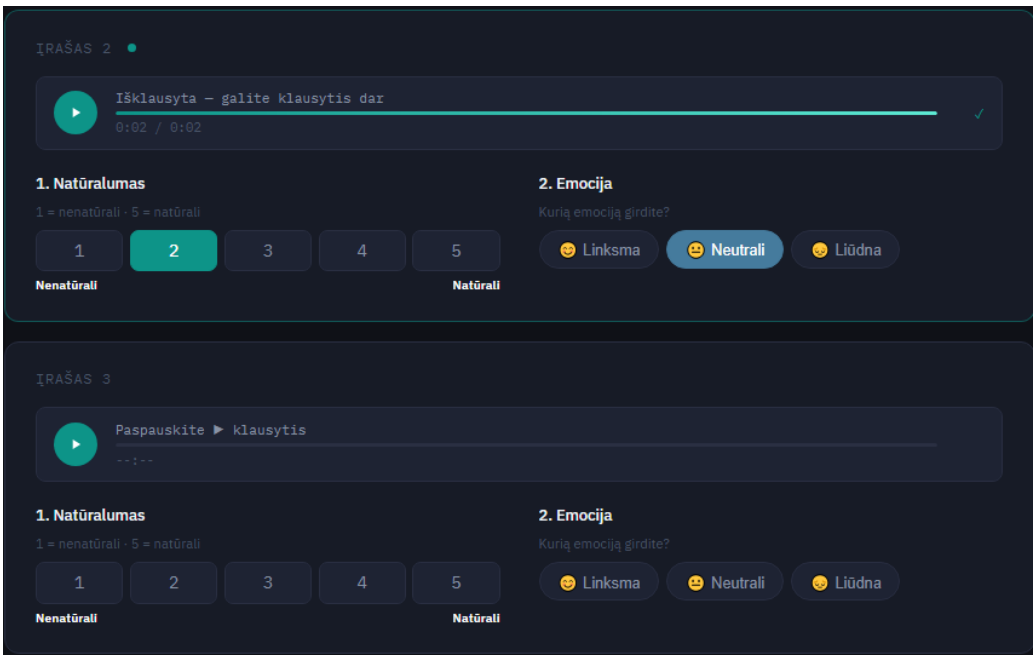
S grupės modelių rezultatai žymiai skiriasi nuo B grupės modelių rezultatų. B grupės modeliai ne visada buvo akustiškai artimiausi savo tikslinei emocijai, o atstumai tarp skirtingų emocinių kategorijų buvo gerokai mažesni – tai rodo ribotą emocinių savybių diferenciaciją. Tuo tarpu S grupės modeliai pasiekė aiškų ir nuoseklų emocinių kategorijų atskyrimą pagal visus keturis akustinius požymius. Manoma, kad šį skirtumą lėmė keli veiksniai. Pirma, ELKG duomenų rinkinys yra studijos kokybės – įrašytas profesionalioje garso studijoje su triukšmo lygiu žemiau 15 dB, kas užtikrina aukštesnę akustinę kokybę nei LESS, LIEPA ar CVL duomenų rinkiniai. Antra, tiek bazinio, tiek papildomų modelių mokymui S grupė naudojo kelis kartus daugiau duomenų nei B grupė. Trečia, kokybiškas S

bazinio modelio rezultatas sudarė tvirtą pagrindą papildomam mokymui – geresnis bazinis modelis leido efektyviau perteikti tikslinės emocijos akustines savybes. Šie rezultatai patvirtina, kad duomenų kokybė ir kiekis yra svarbūs veiksniai emocinės kalbos sintezės kokybei. Taip pat jie leidžia papildyti ankstesnę išvadą apie akustinės analizės ribotumą – B grupės modelių atveju akustinė analizė nebuvo pakankama aiškioms išvadoms daryti dėl menkos emocinių savybių diferenciacijos. Tačiau S grupės modelių atveju akustinė analizė parodė pakankamai stiprius ir nuoseklius rezultatus, leidžiančius daryti patikimesnes išvadas apie modelių gebėjimą perteikti emocijas. Tai rodo, kad akustinės analizės patikimumas priklauso nuo modelių kokybės – kuo geriau modelis perima emocines savybes, tuo informatyvesnė ir patikimesnė tampa akustinė analizė kaip vertinimo priemonė.

4.6. Subjektyvus modelių vertinimas

Siekiant įvertinti sukurtų modelių kokybę subjektyviu klausytojo požiūriu, buvo atliktas MOS tyrimas. Apklausoje dalyvavo 15 lietuvių kalbos gimtakalbių. Kiekvienas dalyvis įvertino 150 garso įrašų – po 10 įrašų iš kiekvieno iš 12 sintetizuotų modelių grupių ir po 10 realių testinių įrašų kiekvienai tiriamajai emocijai, iš viso 15 grupių. Testiniai sakiniai paimti iš LESS duomenų rinkinio testavimo poaibio. Visi modeliai sintezavo tuos pačius 10 sakinių, siekiant užtikrinti tyrimo pastovumą. Siekiant sumažinti nuovargio poveikį, apklausa padalinta į tris seansus po 50 įrašų. Kiekvienas įrašas vertintas dviem kriterijais: kalbos natūralumas pagal penkiabalę MOS skalę (1 – nenatūralu, 5 – visiškai natūralu) ir emocijų atpažinimas – dalyviai turėjo nurodyti kurią emociją atpažįsta (linksma, neutrali ar liūdna). Garso įrašai pateikti atsitiktine tvarka, dalyviai neturėjo jokios informacijos apie tiriamus modelius, jų skaičių, realius įrašus ir pan.

Apklausai buvo sukurta specializuota internetinė platforma, leidžianti dalyviams patogiai klausytis garso įrašų ir teikti vertinimus. Kiekvienas įrašas pateiktas atskirame bloke su garso grotuvu, natūralumo vertinimo skale ir emocijų pasirinkimo mygtukais. Vertinimo lango vaizdas pateiktas 14 pav.

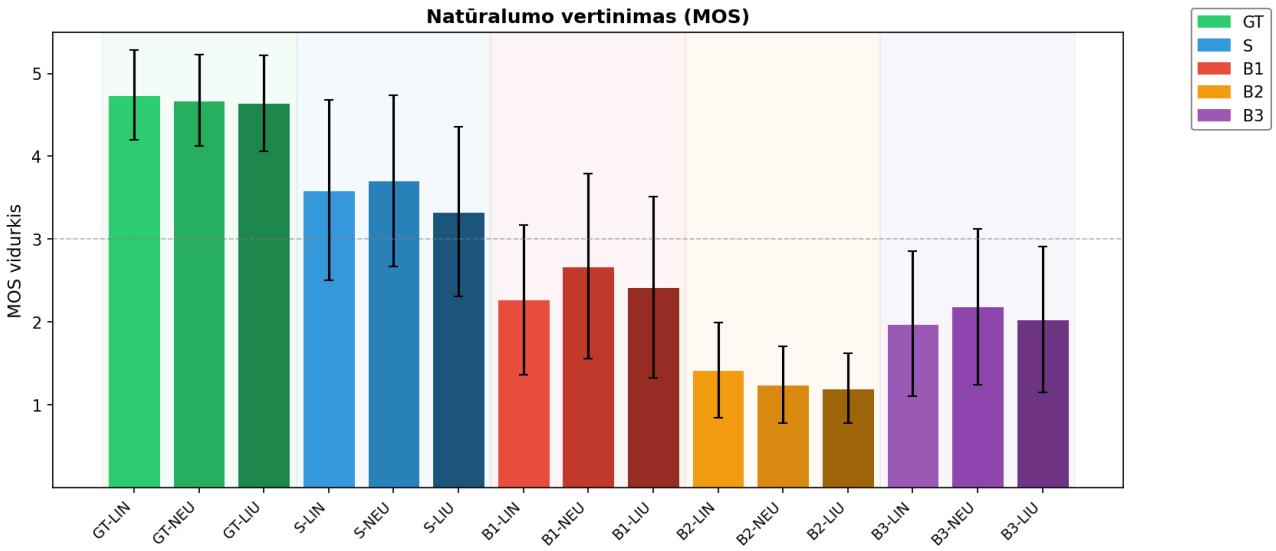


14 pav. *Subjektyvaus vertinimo platformos vertinimo langas*

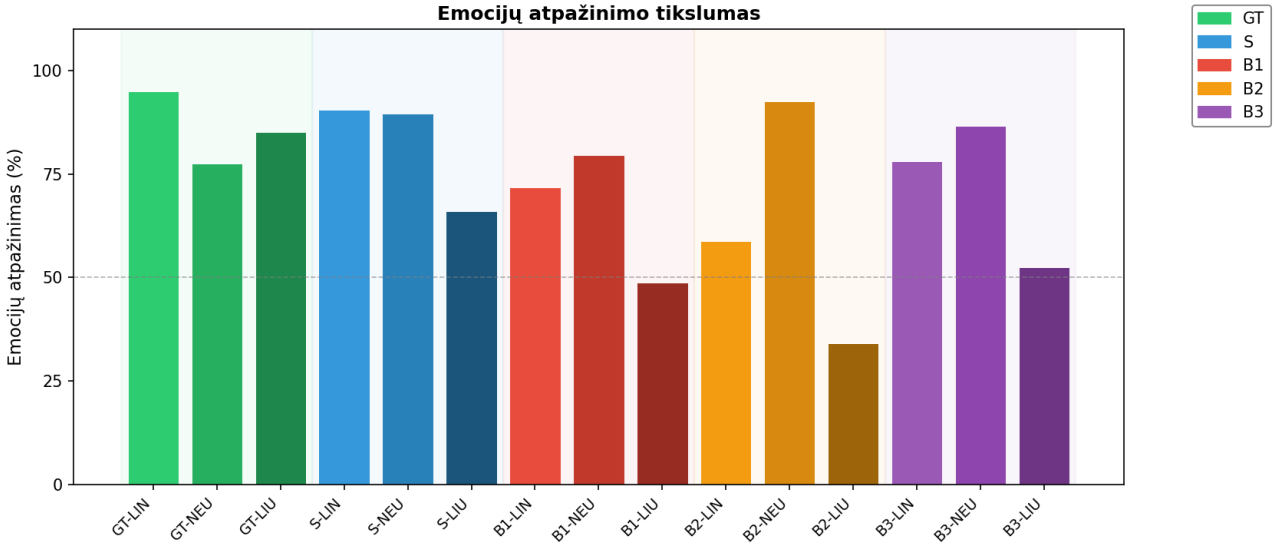
Turint visų įrašų įvertinimus, kiekvieno modelio MOS reikšmei apskaičiuotas vidurkis ir standartinis nuokrypis, bei įvertinta teisingai atpažintų emocijų procentinė dalis. Rezultatai pateikti 19 lentelėje, o grafiškai atvaizduoti 15 ir 16 pav. Lentelėje realūs įrašai žymimi santrumpa GT (angl. *ground truth*) – tai originalūs žmogaus balso įrašai, naudojami kaip atskaitos taškas sintetizuotų modelių kokybei įvertinti.

19 lentelė. *Subjektyvaus vertinimo rezultatai*

Modelis	**MOS**	**±Std**	**Emocijų atpažinimas (%)**
Realūs įrašai (GT)			
GT-LIN	4,74	0,54	95,0
GT-NEU	4,67	0,55	77,5
GT-LIU	4,64	0,58	85,2
Vidurkis	*4,69*		*85,9*
S modeliai			
S-LIN	3,59	1,09	90,6
S-NEU	3,70	1,03	89,6
S-LIU	3,33	1,02	66,1
Vidurkis	*3,54*		*82,1*
B1 modeliai			
B1-LIN	2,27	0,90	71,8
B1-NEU	2,67	1,12	79,7
B1-LIU	2,42	1,09	48,7
Vidurkis	*2,45*		*66,8*
B2 modeliai			
B2-LIN	1,42	0,57	58,8
B2-NEU	1,24	0,47	92,6
B2-LIU	1,20	0,42	34,2
Vidurkis	*1,29*		*62,2*
B3 modeliai			
B3-LIN	1,98	0,88	78,2
B3-NEU	2,18	0,94	86,7
B3-LIU	2,02	0,88	52,5
Vidurkis	*2,06*		*72,5*



15 pav. *Modelių natūralumo vertinimas (MOS)*



16 pav. *Modelių emocijų atpažinimo tikslumas*

Realūs įrašai (GT) pasiekė aukščiausius MOS įvertinimus – vidutiniškai 4,69 balo, kas patvirtina apklausos patikimumą ir dalyvių gebėjimą atskirti natūralią kalbą nuo sintetizuotos. S grupės modeliai pasiekė aukštesnius MOS įvertinimus nei visi B grupės modeliai. Geriausias S grupės modelis buvo S-NEU (MOS 3,70), geriausias B grupės – B1-NEU (MOS 2,67). Bendrai B1 modeliai surinko geriausius rezultatus iš visų B tipo modelių. Prasčiausius rezultatus gavo B2 tipo modeliai, kurie vidutiniškai surinko mažiausiai balų tiek natūralumo, tiek emocijų atpažinimo srityse.

Lyginant su atliktos objektyviosios analizės rezultatais, pastebimas sutapimas S modeliuose ir iš dalies B1 modeliuose – modeliai kurie akustiškai buvo artimesni tikslinėms emocijoms, taip pat gavo aukštesnius subjektyvius įvertinimus. B2 ir B3 modelių rezultatai buvo mažiau nuoseklūs – pavyzdžiui, B2-NEU pasiekė aukštą emocijų atpažinimą (92,6%), tačiau itin žemą MOS įvertinimą (1,24), kas rodo kad modelis generavo atpažįstamą, bet nenatūraliai skambančią kalbą.

Apibendrinus galima teigti, kad papildomo mokymo metodas yra efektyvus būdas, kuriant emocinės kalbos modelius su limituotu emocinės kalbos įrašų kiekiu. Vis dėlto siekiant kokybiško modelių apmokymo, galima išskirti kelis svarbius veiksnius:

1. Didelis kiekis geros kokybės duomenų baziniam modeliui – papildomai apmokytų modelių natūralumas stipriai priklauso nuo bazinio modelio kokybės.

2. Akustinis panašumas tarp bazinio modelio mokymo duomenų ir papildomo mokymo emocinių duomenų – darbe geriausius rezultatus parodė S ir B1 tipo modeliai, kurių bazinio modelio mokymo metu buvo atrinkti akustiškai panašūs į papildomo mokymo diktoriaus įrašus diktoriai.

3. Kokybiški papildomo mokymo duomenys – tyrime nustatyta, kad studijos kokybės emociniai įrašai leido pasiekti žymiai geresnius rezultatus nei buitinėmis sąlygomis įrašyti duomenys.

4.7. Papildomų duomenų kiekis ir skaičiavimo resursai

Vienas iš praktiškai svarbių papildomo mokymo metodo privalumų – žymiai mažesni skaičiavimo resursų reikalavimai lyginant su mokymu nuo nulio. S bazinio modelio mokymas su 28 105 įrašais truko apie 30 valandų. Kiekvienas papildomas S modelio mokymas su apie 2 400 emocinių įrašų truko apie 4 valandas – tai sudaro tik apie 13% bazinio modelio mokymo laiko. Taigi vieno emocinio modelio sukūrimui užteko apie 8,5% bazinio modelio mokymo duomenų kiekio ir apie 13% mokymo laiko. Kadangi šiame darbe buvo apmokyti trys papildomi S modeliai (linksma, neutrali, liūdna), bendras papildomo mokymo laikas sudarė apie 12 valandų – tai yra 40% bazinio modelio mokymo laiko. Tai rodo, kad turėdami vieną kokybišką bazinį modelį, galima efektyviai sukurti kelis emocinius modelius su žymiai mažesnėmis laiko, skaičiavimo resursų ir duomenų sąnaudomis. Lyginant su hipotetiniu scenarijumi, kai emociniai modeliai būtų mokomi nuo nulio su pilnu duomenų kiekiu, papildomo mokymo metodas leido sutaupyti apie 91,5% mokymo duomenų ir apie 87% skaičiavimo laiko kiekvienam emociniam modeliui.

Rezultatai ir išvados

Šiame darbe sukurta:

1. LESS duomenų rinkinys, kurį sudaro mokymo ir testavimo rinkiniai. Mokymo rinkinį sudaro 900 garso įrašų (300 sakinių $\times$ 3 emocijos), testavimo rinkinį – 60 garso įrašų (20 sakinių $\times$ 3 emocijos). Duomenų rinkinyje pateikiamos trys emocinės kategorijos: linksma, neutrali ir liūdna.

2. Keturi baziniai modeliai B1, B2, B3 ir S, gebantys sintezuoti neutralios emocijos lietuvišką kalbą, naudojant skirtingus duomenų rinkinius ir mokymo strategijas.

3. Dvylika papildomai apmokytų modelių, gauti papildomo mokymosi metodu iš bazinių modelių B1, B2, B3 ir S, naudojant LESS ir ELKG duomenų rinkinius. Šie modeliai geba sintezuoti suprantamą lietuvių kalbą su akustiškai pastebimomis emocinėmis savybėmis.

Darbo išvados:

1. Papildomo mokymo metodas yra tinkamas būdas kurti emocines kalbos sintezės sistemas su ribotu emocinių duomenų kiekiu – visi papildomai apmokyti modeliai (B1, B2, B3 iš dalies) perėmė tikslinės emocijos akustines charakteristikas ir sintezavo suprantamą lietuvių kalbą.

2. Bazinio modelio ir papildomo mokymo duomenų diktorių akustinis panašumas yra svarbus veiksnys sėkmingam emocinių savybių perėmimui – kuo akustiškai artimesni diktorių balsai, tuo geriau modelis perima tikslinės emocijos charakteristikas.

3. Objektyvi akustinė analizė ir subjektyvus MOS vertinimas parodė suderintus rezultatus – modeliai kurie akustiškai tiksliau perteikė tikslinės emocijos savybes, taip pat gavo aukštesnius subjektyvius natūralumo įvertinimus.

4. Papildomo mokymo metodas žymiai sumažina laiko, duomenų ir skaičiavimo resursų sąnaudas lyginant su mokymu nuo nulio – turėdami kokybišką bazinį modelį, emocinius modelius galima sukurti per trumpą laiką su nedideliu kiekiu emocinių duomenų.

Literatūra ir šaltiniai

[ABD+19] R. Ardila, M. Branson, K. Davis, M. Henretty ir kiti. „Common Voice: A Massively-Multilingual Speech Corpus“. Iš: *CoRR* abs/1912.06670 (2019). URL: http://arxiv.org/abs/1912.06670.

[BAD13] A. Balyan, S. S. Agrawal, A. Dev. „Speech synthesis: a review“. Iš: *International Journal of Engineering Research & Technology (IJERT)* 2.6 (2013), puslapiai 57–75.

[BTD24] H. Barakat, O. Turk, C. Demiroglu. „Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources“. Iš: *EURASIP Journal on Audio, Speech, and Music Processing* 2024.1 (2024), puslapis 11.

[Ekm92] P. Ekman. „An argument for basic emotions“. Iš: *Cognition & emotion* 6.3-4 (1992), puslapiai 169–200.

[ESS+15] F. Eyben, K. Scherer, B. Schuller, J. Sundberg ir kiti. „The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing“. Iš: *IEEE Transactions on Affective Computing* 7 (2015), puslapiai 1–1. https://doi.org/10.1109/TAFFC.2015.2457417.

[IM24] M. Ikeda, K. Markov. „FastSpeech2 Based Japanese Emotional Speech Synthesis“. Iš: *2024 IEEE 12th International Conference on Intelligent Systems (IS)*. IEEE. 2024, puslapiai 1–5.

[JL03] P. Juslin, P. Laukka. „Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code?“ Iš: *Psychological Bulletin* 129 (2003), puslapiai 770–814. https://doi.org/10.1037/0033-2909.129.5.770.

[Kel95] E. Keller. *Fundamentals of speech synthesis and speech recognition: basic concepts, state-of-the-art and future challenges*. John Wiley ir Sons Ltd., 1995.

[KKS21] J. Kim, J. Kong, J. Son. „Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech“. Iš: *International conference on machine learning*. PMLR. 2021, puslapiai 5530–5540.

[LTK+18] S. Laurinčiukaitė, L. Telksnys, P. Kasparaitis, R. Kliukienė, V. Paukštytė. „Lithuanian Speech Corpus Liepa for Development of Human-Computer Interfaces Working in Voice Recognition and Synthesis Mode“. Iš: *Informatica* 29.3 (2018), puslapiai 487–498. ISSN: 0868-4952. https://doi.org/10.15388/Informatica.2018.177.

[MYD21] Z. Mu, X. Yang, Y. Dong. „Review of end-to-end speech synthesis technology based on deep learning“. Iš: *CoRR* abs/2104.09995 (2021). URL: https://arxiv.org/abs/2104.09995.

[MKC22] S. Moon, S. Kim, Y.-H. Choi. „Mist-tacotron: End-to-end emotional speech synthesis using mel-spectrogram image style transfer“. Iš: *IEEE Access* 10 (2022), puslapiai 25455–25463.

[NHW+19] Y. Ning, S. He, Z. Wu, C. Xing, L.-J. Zhang. „A review of deep learning based speech synthesis“. Iš: *Applied Sciences* 9.19 (2019), puslapis 4050.

[Plu80] R. Plutchik. „A general psychoevolutionary theory of emotion“. Iš: *Theories of emotion*. Elsevier, 1980, puslapiai 3–33.

[PM17] S. PS, G. Mahalakshmi. „Emotion models: a review“. Iš: *International Journal of Control Theory and Applications* 10.8 (2017), puslapiai 651–657.

[PRP05] J. Posner, J. A. Russell, B. S. Peterson. „The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology“. Iš: *Development and psychopathology* 17.3 (2005), puslapiai 715–734.

[Rad22] A. Radzevičius. „Lithuanian Speech Synthesis Using Neural Networks“. Magistro darbas. 2022.

[Sch05] K. R. Scherer. „What are emotions? And how can they be measured?“ Iš: *Social science information* 44.4 (2005), puslapiai 695–729.

[SPW+18] J. Shen, R. Pang, R. J. Weiss, M. Schuster ir kiti. *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*. 2018. URL: https://arxiv.org/abs/1712.05884.

[Sto15] Ł. Stolarski. „Pitch Patterns in Vocal Expression of'Happiness' and'Sadness' in the Reading Aloud of Prose on the Basis of Selected Audiobooks“. Iš: *Research in Language* 13.2 (2015), puslapiai 140–161.

[Tha23] Q. H. N. Thanh X. Le An T. Le. „Emotional Vietnamese Speech Synthesis Using Style-Transfer Learning“. Iš: *Computer Systems Science and Engineering* 44.2 (2023), puslapiai 1263–1278. https://doi.org/10.32604/csse.2023.026234. URL: http://www.techscience.com/csse/v44n2/48270.

[THD19] N. Tits, K. E. Haddad, T. Dutoit. „Exploring Transfer Learning for Low Resource Emotional TTS“. Iš: *CoRR* abs/1901.04276 (2019). URL: http://arxiv.org/abs/1901.04276.

[TQS+21] X. Tan, T. Qin, F. Soong, T.-Y. Liu. „A survey on neural speech synthesis“. Iš: *arXiv preprint arXiv:2106.15561* (2021).

[Vil26] Vilniaus universiteto Duomenų mokslo ir skaitmeninių technologijų institutas. *Emocinės lietuvių kalbos garsynas*. Licencija: CC BY 4.0. 2026.

[VSR+24a] P. S. Varadhan, A. Sankar, G. Raju, M. M. Khapra. „Rasa: Building expressive speech synthesis systems for indian languages in low-resource settings“. Iš: *arXiv preprint arXiv:2407.14056* (2024).

[VSR+24b] P. S. Varadhan, A. Sankar, G. Raju, M. M. Khapra. *Rasa: Building Expressive Speech Synthesis Systems for Indian Languages in Low-resource Settings*. 2024. URL: https://arxiv.org/abs/2407.14056.

[WLL+19] P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, L. Dai. „End-to-end emotional speech synthesis using style tokens and semi-supervised training“. Iš: *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2019, puslapiai 623–627.

[ZY23] W. Zhao, Z. Yang. „An emotion speech synthesis method based on vits“. Iš: *Applied Sciences* 13.4 (2023), puslapis 2225.

Santrumpos

VITS – Variational Inference Text-to-Speech
MOS - Mean Opinion Score
TTS – Text To Speech
LESS – Lithuanian Emotional Speech Set
CVL – Common Voice Lithuanian
F0 – Pagrindinis tonas
CMOS - Comparison Mean Opinion Score
DMOS - Differential Mean Opinion Score
RMSE - Root Mean Squared Error
MCD - Mel-Cepstral Distortion
GPE - Gross Pitch Error
VDE - Voicing Decision Error
CER - Character Error Rate
FFE - f0 Frame Error
WER - Word Error Rate
VAE - Variational autoencoder