Kauno technologijos universitetas

Elektros ir elektronikos fakultetas

Garsyno LIEPA atpažinimo su Kaldi paketu sistemos sukūrimas ir tyrimas

Baigiamasis magistro projektas

Gytis Baltrušaitis

Projekto autorius

Doc. dr. Kastytis Ratkevičius

Vadovas

Kaunas, 2020

Kauno technologijos universitetas

Elektros ir elektronikos fakultetas

Garsyno LIEPA atpažinimo su Kaldi paketu sistemos sukūrimas ir tyrimas

Baigiamasis magistro projektas

Valdymo technologijos (6211EX014)

Gytis Baltrušaitis

Projekto autorius

Doc. dr. Kastytis Ratkevičius

Vadovas

Doc. dr. Tomas Tekorius

Recenzentas

Kaunas, 2020

Kauno technologijos universitetas

Elektros ir elektronikos fakultetas

Gytis Baltrušaitis

Garsyno LIEPA atpažinimo su Kaldi paketu sistemos sukūrimas ir tyrimas

Akademinio sąžiningumo deklaracija

Patvirtinu, kad mano, Gyčio Baltrušaičio, baigiamasis projektas tema „Garsyno LIEPA atpažinimo su Kaldi paketu sistemos sukūrimas ir tyrimas“ yra parašytas visiškai savarankiškai ir visi pateikti duomenys ar tyrimų rezultatai yra teisingi ir gauti sąžiningai. Šiame darbe nei viena dalis nėra plagijuota nuo jokių spausdintinių ar internetinių šaltinių, visos kitų šaltinių tiesioginės ir netiesioginės citatos nurodytos literatūros nuorodose. Įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs.

Aš suprantu, kad išaiškėjus nesąžiningumo faktui, man bus taikomos nuobaudos, remiantis Kauno technologijos universitete galiojančia tvarka.

_____	_____
(vardą ir pavardę įrašyti ranka)	(parašas)

Baltrušaitis, Gytis. Garsyno LIEPA atpažinimo su Kaldi paketu sistemos sukūrimas ir tyrimas. Magistro baigiamasis projektas / vadovas doc. dr. Kastytis Ratkevičius; Kauno technologijos universitetas, Elektros ir elektronikos fakultetas.

Studijų kryptis ir sritis: Elektronikos inžinerija (Inžinerijos mokslai).

Reikšminiai žodžiai: LIEPA, ASR, kaldi, TDNN.

Kaunas, 2020. 67 p.

Santrauka

Šio projekto tikslas yra išanalizuoti „Kaldi“ priemonių rinkinio perspektyvas automatinio balso atpažinimo kontekste. Pirmas garsyno paketas, su kuriuo buvo pradėta dirbti, pavadintas Medicininiai terminai, kurį sudarė 30 diktorių. Ištyrus šį garsyną visais galimais metodais, buvo gauti idealūs rezultatai, apie 99,99 proc., todėl buvo nuspręsta užtriukšminti medicininių terminų žodyną 5dB baltu triukšmu. Užtriukšminus įrašus bei pradėjus tirti, gauti apie 100 proc. tikslumo rezultatai, todėl buvo nuspręsta pereiti prie platesnio, atviro kodo garsyno LIEPA. Šiame darbe išskiriami ir įvertinami metodai, kurie yra skirti diktoriaus atpažinimui. Automatinio balso atpažinimo procesą apima keli etapai: garsintojo balso apdorojimas, požymių išskyrimas bei diktoriaus apmokymas ir patikrinimas.

Pagrindinis šio baigiamojo magistro darbo tikslas – ištirti Kaldi paketo funkcionalumą panaudojant skirtingus metodus kalbos atpažinimui. Darbo objektas ir metodai – Kaldi programiniu paketu apmokomas automatinis kalbos atpažintuvas, naudojant bendro naudojimo žodyną LIEPA, panaudojant lietuviškus 350 diktorių. Žodynas susideda iš 11346 žodžių, jo duomenų bazė susideda iš 310000 žodžių. Naudojamas kompiuteris su Ubuntu operacine sistema. Tyrimai atliekami panaudojant monofininį, trifoninį, LDA+MLLT, LDA+MLLT+SAT, SGMM2 ir TDNN pnorm. bei TDNN tanh metodus.

Kaldi paketo paruošimo failai, wav.scp, text.txt, utt2spk.txt, corpus.txt, spk2gender. Šie paketai pateikti, kaldi tyrimo struktūros paruošimui bei testavimui.

Kaldi iliustraciniai garso dokumentai: spk2gender.txt, wav.scp, text.txt, utt2spk.txt, corpus.txt buvo pateikti paklausimui, kurie reikalingi karkaso paruošimui ir testavimui, kaip ir pirminis run.sh įrašas, vaizduojantis išnaudotus pajėgumus.
Pradėjus analizuoti LIEPA garsyną, šis buvo išskaidytas į 3 skirtingas dalis: LIEPA_ZOD, LIEPA_SEK, LIEPA_SAK. Patikrinus kiekvieną prieš tai minėtą dalį, pastebima, kad tiriant sekas, žodžius ar sakinius, tikslesni atpažinimo rezultatai, gauti naudojant žodžių klaidų dažnį, lyginant su sakinių klaidų dažniu.

Baltrušaitis, Gytis. Creation and investigation of LIEPA speech corpus recognition system using Kaldi package. Master's Final Degree Project / supervisor doc. Dr. Kastytis Ratkevičius; Faculty of Electrical and electronics engineering, Kaunas University of Technology.

Study field and area (study field group): Electronics Engineering (Engineering Sciences)

Keywords: LIEPA, ASR, kaldi, TDNN

Kaunas, 2020. 67p.

Summary

The aim of this project is to clarify the perspectives of the Kaldi toolkit in the context of automatic voice recognition. The first sound package to be worked on was called Medical Terms, which consisted of 30 narrators. After examining this sound system with all possible methods, ideal results were obtained by ~ 99.99%, so it was decided to noise the dictionary of medical terms with 5dB white noise. After making noise and starting to research, the results were close to 100%, so it was decided to move to a wider, open source sound system, LIEPA. In this work, methods for recognizing a narrator are singled out and evaluated. The process of automatic voice recognition involves several steps, including the processing of the loudspeaker's voice, the extraction of features, and the training and verification of the narrator.

The main goal of this master's thesis is to investigate the functionality of the Kaldi package using different methods for language recognition. Object and methods of the work - Kaldi software package teaches automatic speech recognition, using the common dictionary LIEPA, using Lithuanian 350 narrators. The dictionary consists of 11346 words, its database consists of 310000 words. A computer with an Ubuntu operating system is used. The studies are performed using monophonic, triphonic, LDA + MLLT, LDA + MLLT + SAT, SGMM2 and TDNN pnorm. methods

Kaldi package preparation files, wav.scp, text.txt, utt2spk.txt, corpus.txt, spk2gender. These packages are provided for the preparation and testing of the study structure.

Kaldi illustrative audio documents: spk2gender.txt, wav.scp, text.txt, utt2spk.txt, corpus.txt were provided on request, which are required for the preparation and testing of the framework, as well as the initial run.sh recording depicting the capacity utilized.

After starting to analyze the LIEPA sound system, this one was divided into 3 different parts: LIEPA_ZOD, LIEPA_SEK, LIEPA_SAK. Examining each of the above sections, we can observe that by examining both sequences, words, or sentences, we can observe that the more accurate recognition results obtained using the word error rate compared to the sentence error rate.

Turinys

Santrumpų ir terminų sąrašas

Santrumpos:

ASR – automatinis balso atpažinimas;

DARPA – gynybos pažangiųjų mokslinių tyrimų projektų agentūra;

HMM – paslėptasis Markovo modelis;

WER – žodžių klaidų tikimybė;

SER – sakinių klaidų tikimybė;

HTK – paslėptasis Markovo modelio įrankių rinkinys;

LDA – tiesinė diskriminantinė analizė (angl. Linear Discriminant Analysis);

MLLT – maksimali linijinės transformacijos tikimybė (angl. Maximum Likelihood Linear Transform);

SAT – adaptyvus mokymasis (angl. speaker-adaptely training);

fMLLR – bruožas-intervalas Maksimali linijinės regresijos tikimybė (angl. feature-space Maximum Likelihood Linear Regression);

SGMM2 – Gauso mišinių modeliai (angl. Subspace Gaussian Mixture Models);

TDNN – laiko delsos neuroniniai tinklai (angl. Time-Delay Neural Network);

MFCC – Melo dažnio cepstralinis koeficientas;

LIEPA_SEK – Garsyno LIEPA, sekų tyrimas;

LIEPA_ZOD – Garsyno LIEPA, žodžių tyrimas;

DNN – gilieji neuroniniai tinklai;

DBN – Deep Belief Network;

AM – akustinis modelis;

LM – kalbos modelis;

RLS – rekursyvus kvadrato modelis;

VTS – Vektor Taylor‘o serija;

STE – trumpo laiko energija;

ZCR – nulinio kirtimosi dažnis;

FTE – rėmo pagrindo Teagerio energija;

EEF – energijos entropijos būsena;

PLP – percepcinis tiesinis numatymas;

CMS – koeficientinio Cepstralinio vidurkio skirtumas;

LCA – pagrindinio komponentinė analizė;

MLE – maksimalus panašumo įvertinimas;

MAP – maksimalus Pasteriori;

MMIE – maksimali abipusė informacijos vertinimas;

MCE – minimali klasifikacijos klaida;

VQ – vektorių kvantisacija;

EM – tikimybės maksimalus algoritmas;

CART – klasifikavimo ir regresijos medis;

Terminai:

Diftonai – garsas, suformuotas sujungiant du balsius viename skiemenyje, kuriame garsas prasideda kaip viena balsė.

Skriptas – programa, sudaryta iš interpretavimui skirtų komandų. Skriptas laikomas pirminiu tekstu, kurį prireikus vykdo interpretatorius. Skriptų būna įvairių rūšių. Skriptas yra iš esmės tas pats, kas ir makroprograma skirta valdyti kompiuteriui.

White noise – lietuviškai baltasis triukšmas yra atsitiktinis signalas, kurio energetinis spektras yra pastovus visiems dažniams. Pavadinimas kilęs iš panašumo į baltą šviesą.

Open-Source – kompiuterinė programa, kuri platinama pagal atvirojo kodo licenciją. Tokia programinė įranga dažniausiai pasižymi šiomis savybėmis:

- nemokama;
- laisvai prieinami programos kodas;
- galima laisvai platinti ir modifikuoti, nekeičiant licenzijos

Ubuntu – tai programinės įrangos paketas skirtas Linux operacinėi sistemai.

Įvadas

Automatinis kalbos atpažinimas (ASR) yra dirbtinio intelekto metodas, kurio pagrindinis tikslas – leisti balsu susikalbėti kompiuteriui ir žmogui. ASR pagrindinis tikslas yra žmogaus kalba paversti į kompiuterinį tekstą, tačiau atsiranda kita problema – žmonių kalbos sudėtingumas, t.y. fonetika. Kalbėdami žmonės naudoja ne tik ausis, taip pat yra įvertinama pašnekovo kūno kalba, emocijos, bendravimo aplinka. Naudodami ASR mes gauname tik kalbos signalą. Šiame darbe bus vertini kiti būdai ir strategijos norint išgauti geriausią signalo rezultatą.

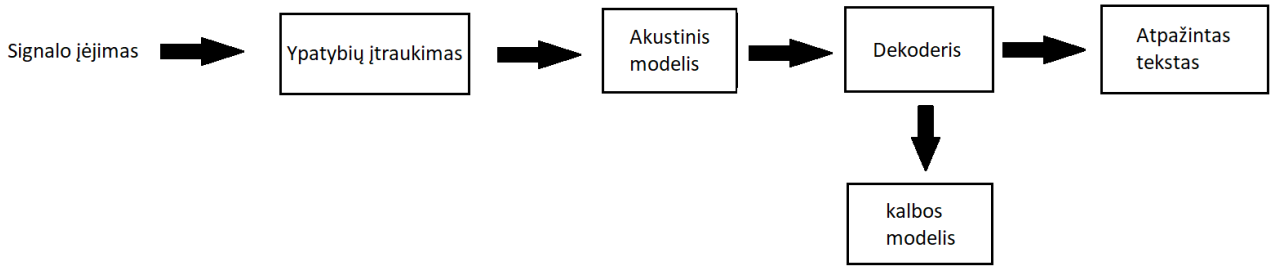
Pagrindinis baigiamojo magistro darbo tikslas – ištirti Kaldi paketo funkcionalumą panaudojant skirtingus metodus kalbos atpažinimui. Darbo objektas ir metodai – Kaldi programiniu paketu apmokomas automatinis kalbos atpažintuvas, naudojant bendro naudojimo žodyną LIEPA, panaudojant lietuviškus 350 diktorių. Žodynas susideda iš 11346 žodžių, jo duomenų bazė susideda iš 310000 žodžių. Naudojamas kompiuteris su Ubuntu operacine sistema. Tyrimai atliekami panaudojant monofininį, trifoninį, LDA+MLLT, LDA+MLLT+SAT, SGMM2 ir TDNN pnorm. metodus.

1. Literatūros analizė

1.1. Automatinis kalbos atpažintuvas

Automatinė kalbos atpažinimo technologija išgyveno kelis plėtros etapus, kad pasiektų dabartinę stadiją. Per pastaruosius šimtą metų mokslininkai visame pasaulyje buvo įtraukti į mašinos sukūrimą, kuri galėtų tapti žmogaus kopija. Mašina – kuri galėtų klausyti, šnekėti bei atsakyti bet kokia žinoma kalba (B.H. Juang and L.R. Rabiner, 2005).

Tipiškos ASR sistemos „blokinė“ schema pateikta 1 paveiksle. Ją sudaro keturi moduliai - funkcijų ypatybių įtraukimo modelis, akustinis modelis (AM), dekodavimo modelis ir kalbos modelis (LM) (Akella Amarendra Babu, Yellasiri Ramadevi and Akepogu Ananda Rao, 2014).



1 pav. Blokinė diagrama ASR sistemos

Įvesties signalo forma yra paverčiama į ypatybių parametrų rinkinių vektorius. Funkcijų išskyrimui naudojami Mel dažnio Cepstral koeficientai (MFCC), jos pirmosios eilės delta MFCC ir antrosios eilės delta MFCC (Khaled Abdalgadar and Andrew Skabar, 2012) (L. Rabiner, B. Juang and B Yegnanarayana, 2010) (Morgan, 2012). Įvairūs metodai, naudojami parametrų vaizdavimui pagerinti, pateikti 1 lentelėje.

1 lentelė. Metodai, skirti parametrų vaizdavimui pagerinti

Algoritmas	Tikslas
RLS VTS	Triukšmo slopinimas
STE ZCR FTE EEF	Galutinio taško atradimas bei kalbos segmentavimas
MFCC PLP CMS	Ypatybių įtraukimui
RASTA filtravimas	Triukšmingiems diktoriams
LCA LDA	Ypatybių transformacija

Akustinis modelis (AM) konvertuoja kalbos parametrinius vektorius į atitinkamas fonemų sekas. Akustiniam modeliavimui naudojami paslėpti Markovo modeliai (Issam Bazzi and James Glass, 2002) (Xinguang Li, Jiahua Chen, Zhenjiang Li, 2013). Įvairūs mokymosi ir dekodavimo būdai pateikti 2 lentelėje.

2 lentelė. Apmokymo ir dešifravimo technikos

MLE MAP	Prižiūrimas mokymo modelis
MMIE MCE	Diskriminantinis modelis
Atgalinis plitimas	MLP apmokymas
VQ K reikšmės algoritmas EM	Neprižiūrimasis modelis
CART	DTW
Dinaminis programavimas	HMM dekodavimas

Dešifratorius paverčia fonemų sekas į žodžius ir naudoja kalbos modelį (LM) semantiniam patvirtinimui (Huang, Acero and Hon, 2001) (Tao Tao, Su-Youn Yoon, Andrew Fister, Richard Sproat and ChengXiang Zhai, 2006).

1.1.1. Automatinio kalbos atpažinimo iššūkiai

Automatinio atpažinimo sistemos tvirtumas yra sistemos gebėjimas sėkmingai susidoroti su skirtingais kalbos signalo kintamumo aspektais. Kalbos atpažinimo sistemos tikslumą lemia keletas gerai žinomų veiksnių. Labiausiai pastebimi veiksniai: kalbėtojų kintamumas, tarimo kintamumas, regiono kintamumas, kalbos greičio kintamumas, konteksto kintamumas, kanalo kintamumas ir aplinkos kintamumas. Projektuojant kalbos atpažinimo sistemas, reikia atsižvelgti į šiuos iššūkius lemiančius veiksnius ir sukurti veiksmingus modelius, kurie užtikrintų gerą atpažinimo tikslumą, nepriklausomai nuo šių kintamumų (Forsberg, 2003). Aukštesniame lygmenyje kalbėjimo atpažinimo sistemos projektavimui reikia prieinamų algoritmų procesų automatiniam žodžių leksikų generavimui, automatiniam kalbos modelių generavimui naujoms užduotims, automatiniams kalbos segmentų algoritmams, optimalaus posakio patikrinimo-atmetimo algoritmams, pasiekiamiems ar pranokstamiems žmogaus sugebėjimams bei ASR užduotims.

1.1.2. Automatinės kalbos atpažinimo tikimybių teorija

Pagrindinis ASR sistemos tikslas yra iš pateiktos akustinės įvesties *O* rasti labiausiai tikėtiną diskretinę simbolių seką iš visų galiojančių sekų *L* (D. Jurafsky and J. H. Martin, 2009). Kaip paminėta aukščiau, įvestis traktuojama kaip atskira diskretinė įžvalga, tokia kaip:

$$O = o_1, o_2, o_3, \dots, o_t \quad (1)$$

Panašiai atpažįstama simbolių seka apibrėžiama taip:

$$W = w_1, w_2, w_3, \dots, w_n \quad (2)$$

Pagrindinis ASR sistemos tikslas:

$$W = argmax\, P(W|O) \quad for\, W \in L \quad (3)$$

Ši lygtis reiškia, kad tam tikrai *W* sekai ir akustinei įvesties sekai *O* reikia nustatyti *P (W | O)* tikimybę. Šiai lygčiai gauti galima pritaikyti Bayes'o teoremą:

$$P(W|O) = \frac{P(O|W)P(W)}{P(O)} \quad (4)$$

Kiekius dešinėje lygties pusėje lengviau apskaičiuoti nei *P (W | O)*. *P (W)* yra apibrėžiama kaip ankstesnė pačios sekos tikimybė. Tai apskaičiuojama naudojant išankstines žinias apie sekos *W* įvykius. Kadangi *P (O)* yra tas pats kiekvienam kandidato *W* sakiniui, todėl 4 lygtį galima supaprastinti taip:

$$W = argmax\frac{P(O|W)P(W)}{P(O)} = argmaxP(O|W)P(W) \qquad for\ W\ \in\ L \quad (5)$$

Tikimybė *P (O | W)*, kuri yra akustinio įėjimo *O* tikimybė, atsižvelgiant į seką *W*, yra apibrėžta kaip stebėjimo tikimybė, kuri gali būti vadinama akustiniu rezultatu. Šį kiekį galima nustatyti paslėpto Markovo modeliu.

1.1.3. Technologijos raida

1889 m. Alexander Graham Bell ir Charles Sumner Tainter buvo pirmieji, kurie išrado garso įrašymo mašiną, kuri galėjo reaguoti į balso dažnį. Jie panaudojo besisukantį vaško cilindrą, kuris turėjo griovelius, kuriais eidavo speciali adata, atkartojanti tam tikrus virpesius (Roe, J. G. Wilpon and D. B. Roe, 1992). 1930 m. Homer Dudley pirmasis inžinierius suprojektavęs elektroninės garso sistemos analizatorių, kuris vėliau buvo naudojamas antrojo pasaulinio karo metu siųsti užkoduotas balso žinutes. Šios mašinos išradimas tapo didelis žingsnis balso atpažinimo srityje po kurio daugelis mokslininkų bei tyrinėtojų pasistūmėjo kurti daugiau įvairių įrenginių, skirtų balso atpažinimui bei šifravimui.

1952 m. trys Bell laboratorijos mokslininkai – Balashek, Davis ir Biddulph sukūrė nesudėtingą balso atpažinimo sistemą „Audrey“, kuri buvo skirta vieno diktoriaus skaitmeniniam kalbos atpažinimui. Po ketverių metų Olsonas ir Belaras suprojektavo sistemą, kuri galėtų atskirti dešimt skiemenų, įterptų tarp dešimt atskirų žodžių vienam kalbėtojui vienu metu (B.H. Juang and L.R. Rabiner, 2005).

1959 m. keletas naujų ASR sistemų buvo pristatyta pasaulyje. Vienas iš pavyzdžių būtų MIT Lincoln laboratorijoje sukurta balsių atpažinimo sistema, kuri sugebėjo atpažinti dešimt balsių, ištartų to paties diktoriaus (B.H. Juang and L.R. Rabiner, 2005).

Septintajame dešimtmetyje buvo pristatyta daug įrangos bei prototipų skirtų ASR sistemoms. Viena iš labiausiai žinomų pavyzdžių – Suzuki ir Naka balsių atpažinimo sistema, Sakai ir Doshita fonemų atpažinimo sistema bei skaitmeninė atpažinimo sistema, sukurta NEC laboratorijose (History of Speech Recognition, 2015).

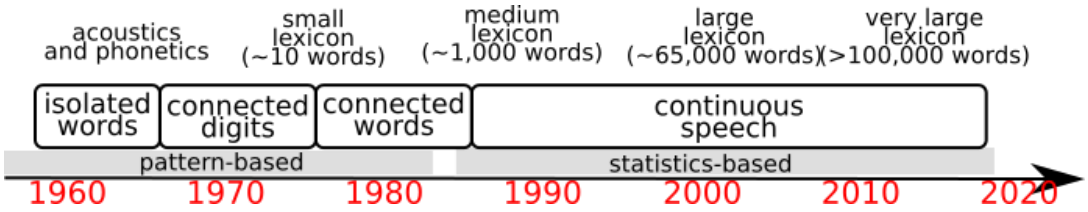
Didžiausias proveržis balso atpažinimo srityje buvo pastebėtas aštuntame dešimtmetyje. Vienas didžiausių atradimų tuo periodu buvo pristatytas statistinis Paslėptojo Markovo modelis (angl. Hidden Markov Model (HMM)) septinto dešimtmečio gale, aštunto pradžioje. Taip pat atlikti tyrimai Japonijoje bei Rusijoje atskleidė technologiją, kurioje buvo naudojamas atskiro žodyno atpažinimas arba izoliuotas tarimo atpažinimas – tai labai dažnas būdas sukurti itin paprastą ASR sistemą.

Tuo tarpu devintajame dešimtmetyje, Automatinė Kalbos Atpažinimo sistema susidūrė su daugybe patobulinimų, ypač modelio atpažinimo struktūroje, naudojant mažiausio klaidų klasifikavimo lygio

metodus. Ši pažanga lėmė naujų svarbių metodų atradimus, tokius kaip diskriminantinis apmokymas ir Kernel pagrindu teigta technika (B.H. Juang and L.R. Rabiner, 2005). Be to, gynybos pažangiųjų mokslinių tyrimų projektų agentūra (DARPA) tęsė savo darbą naudodamiesi dideliais leksikos rinkiniais bei automatiniu kalbos atpažinimo metodikos tobulinimu. Jie pradėjo vartoti žodžių klaidų tikimybes (WER) bei sakinių klaidų tikimybes (SER). Šie du metodai pradėti naudoti kaip ASR technologijos sistemos veikimo metrika.

Tuo metu buvo sukurta ir pristatyta daug programinės įrangos priemonių, pavyzdžiui, Paslėptojo Markovo modelio įrankių rinkinys (angl. Hidden Markov Model Tool Kit), kuris buvo sukurtas Kembridžo universitete ir, manoma, kad jis tapo svarbiausiu elementu bei įrankiu Automatinėje Kalbos Atpažinimo sistemoje (B.H. Juang and L.R. Rabiner, 2005).

Laikotarpis nuo 2000 m. iki 2010 m. tapo dideliu proveržiu Automatinės Kalbos Atpažinimo sistemose, dėl modernių programinės įrangos bei interneto atnaujinimų. Šiandien ASR sistemos laikomos viena iš svarbiausių technologijų, kurios yra naudojamos kasdien. Šias sistemas, galime sutikti: telefonuose, automobiliuose, karinėje pramonėje, robotikoje, sveikatos priežiūros įrankiuose bei kosmose. Automatinio Kalbos Atpažinimo raida per paskutinį šimtmetį pateikta 2 paveiksle (B.H. Juang and L.R. Rabiner, 2005).



2 pav. Automatinio kalbos atpažinimo raida

1.1.4. Kalbos atpažinimo taikymas

Visai neseniai, eksponentiškai didėjant didžiųjų duomenų ir skaičiavimo galiai, ASR technologija perėjo į stadiją, kai sudėtingesnės programos tampa realybe. Paieška balsu ir sąveika su mobiliaisiais įrenginiais (pvz.: „Siri“, „iPhone“, „Bing“ paieška balsu „WinPhone“ ir „Google“ dabar „Andriod“), balso valdymas namų pramogų sistemose (pvz.: „Kinect on xBox“) ir įvairus į kalbą orientuotas informacijos apdorojimas. Paraiškos, kurių pagrindą sudaro ASR rezultatų perdirbimas vėlesnėje dalyje (J. Li, L. Deng, R. H.Umbach and Yifan Gong, 2015). Kai kurios iš šių tipinių programų apima garso įrašymo sistemas, balso vartotojo sąsajas, rinkimą balsu, skambučių nukreipimą, buitinių prietaisų valdymą, paiešką balsu, paprastą duomenų įvedimą, laisvų rankų įrenginių programas bei neįgaliųjų mokymosi sistemą.

1.2. Lietuvių kalba automatinio atpažinimo srityje

Automatinio atpažinimo technologija Lietuvoje nėra naujiena, tačiau, jos taikymo bei tyrimų yra labai nedidelis kiekis. Tą įtakojo tai, kad lietuvių kalba yra gana sudėtinga. Sudėtinga gramatika, fonetika, kurioje įvertiname plačią garsų įtaką mūsų kalboje. Balsių, priebalsių, pusbalsių, dvibalsių, skiemenų įtaka mūsų žodynui. Įvertinant tai, kad lietuviškai kalba labai nedidelė pasaulio žmonių dalis, todėl sunku surinkti visus reikiamus parametrus kalbos modeliavimui bei tyrimui.

Išanalizuoti turimus duomenis ir sprendinius nėra lengva bei paprasta, kadangi automatinio atpažinimo sistemos Lietuvoje kinta kasmet, ne tik gaunamais duomenimis – diktorių skaičiumi, įrašų skaičiumi, žodynais, tačiau neretai keičiasi ir tobulėja metodikos bei įrankiai kalbos atpažinimo tikslams – vienos sistemos labiau tinkamos pavienių žodžių atpažinimui, kitos sakinių, o dar kitos pavieniams tekstams analizuoti (D. Sipavičius and R. Maskeliūnas, 2016).

1.2.1. Lietuviškas garsynas LIEPA

Įprasta valdymo sistema gauna komandą kompiuteriu, pele ar kitos įrangos sugeneruotu elektriniu signalu, kuris reaguoja į vaizdą, garsą ar judesį elektrinio signalo forma. Tačiau balsu valdoma sistema gauna užduotį ir pateikia atsakymą kalbos signalo ar balso forma. Tokios sistemos naudotojai, duodantys komandas balsu, gali būti skirtingo amžiaus, įvairių emocinių ar fizinių būsenų, ir juos gali supti skirtinga akustinė aplinka, kuri gali daryti įtaką atpažinimo rezultatui. Tokios sistemos sąlyga – galimybė fiksuoti, atpažinti bei išanalizuoti gautus sprendinius dėl kalbos subtilybių bei sugebėti teisingai ištarti atsakymą balsu. „Speech corpus“ – vienas iš įrankių, kuris suteikia galimybę suprojektuoti ir sukurti įrankius, tokius kaip balso atpažinimas ar kalbą į tekstą verčiančius sintezatorius sistemos valdomoms balsu. Sistemoms valdomoms balsu išauga dideli reikalavimai: anotuotų kalbų įrašų skaičius dideliais kiekiais, diktorių įvairovė, tarimų žodyno prieinamumas, klaidų minimalizavimo bei aukštos kokybės priemonės tyrimams atlikti. Duomenų bazės tobulinimas dažnai nesudėtingas, tačiau daug laiko užimantis procesas, kadangi į jo tobulinimą įeina daug laiko užimančio rankinio darbo. Atsiradusi tam tikra vystymosi proceso dalis galėtų būti naujų įrankių bei metodu išradimas norint eliminuoti rankinį žmogaus darbą. Nepilnai ištirtų kalbų tyrinėtojai visada ieško sprendimų, kaip įveikti kalbos duomenų trūkumą. Norėdami ištirti prastai išnagrinėtą temą, mokslininkai pasirenka jiems naudingai atrodančius mokymo pavyzdžius (Axelrod, A., Resnik, P., He, X. and Ostendorf, M., 2015), kurių pagalba gaunamas patikimas tarimų žodynas ir taip yra išnaudojami šaltiniai, ištirti glaudžiai susijusiai kalbai (Takahashi, N.,Naghibi, T., Pfister, B., 2016) .Vis dėlto, kalbos tyrimas yra neatsiejamas ir lemiamas įrankis kalbos vystymosi progrese tarp kompiuterio ir žmogaus veikiančio balso atpažinimu ar sintezės rėžimu (Samson, J.S., Besacier, L., Lecouteux, B. and Tan, T., 2014).

1.3. Ubuntu sistemos paruošimas

Kompiuteryje esant įdiegtai Ubuntu Linux operacinei sistemai, Kaldi programinio bloko įrašymui reikalingi papildomi paketai operacinės sistemos optimizavimui:

1. *zlib* – skirtas turimų duomenų optimizavimui;
2. *git* – paskirstytos peržiūros įvedimo į sistemą modelis;
3. *wget* – duomenų perdavimas HTTP/ HTTPS/ FTP protokolais;
4. *libtool* – skirtas kurti dinamines bei statines bibliotekas;
5. *svn* – įvedimo į sistemą blokas, skirtas Kaldi paketo parsisiuntimui bei įrašymui;
6. *automake* – skirta *Makefile* katalogams sukurti;
7. *autoconf* – bendro naudojimo programinės įrangos rinkinys, skirtas operacinės sistemos valdymui.

Visi šie bei kiti paketai, yra vykdomi per Ubuntu programinės įrangos sistemos terminalą, kuris naudoja Unix tipo komandas. Šiems visiems reikalingiems paketamas yra naudojama komanda – *sudo apt-get.*

Įsidiegus visus reikiamus programos paketus Kaldi vartotojo aplanke, galima rasti atvirojo naudojimo kodą – *check_dependencies.sh*, skirtą patikrinti ar tikrai visi reikalingi operacinės sistemos plėtiniai yra tinkamai įdiegti bei suteikia išsamią instrukciją, nesamiems paketams gauti. Tai check_dependencies.sh kodo fragmento ištrauka.

```
if ! dpkg -l | grep -E 'libatlas3gf|libatlas3-base' >/dev/null; then
  echo "You should probably do: "
  echo " sudo apt-get install libatlas3-base"
  printed=true
```

3 pav. Check_dependencies.sh kodo struktūra

Kaldi įrangą galime įrašyti į kompiuterį kai jame yra įdiegta aukščiau išvardinti paketai.

1.3.1. Kaldi paketo įdiegimas

Kaldi įranga yra parsiunčiama iš viešos GitHub platformos, panaudojus git komandą, Ubuntu sistemos terminale:

```
git clone https://github.com/kaldi-asr/kaldi.git kaldi --origin upstream
cd kaldi
```

4 pav. Komanda skirta Kaldi paketo įdiegimui

Atsisiuntus Kaldi paketą, jo aplanke randame instrukcijas, skirtas programinio paketo įrašymui. Jo įrašymas išskirstytas dalimis:

- paketų skirtų Kaldi įrangos patikrinimui;
- Kaldi paketo vykdymas.

Įsitikinus, kad paketai, sudarantys Kaldi sistemą, yra sėkmingai įrašyti, pereiname prie kompiuterio paruošimo:

```
./configure –shared
make depend -j 8
make -j 8
```

5 pav. Apdorojimo proceso kodas, skirtas konfigūruoti kompiuterį

Šiuo atveju -j 8 reiškia, kad apdorojimo procesas bus vykdomas lygiagrečiai, t.y. aštuonios užduotys bus atliekamos vienu metu. Prieš pasirenkant procesų skaičių, turime atsižvelgti į kompiuterio resursus ir įvertinti ar tiriama sistema nebus per sudėtinga seniems kompiuteriams.

1.3.2. Garsyno paruošimas projektui

Kaldi paketui norint atpažinti kokiems diktoriams ar kokioms transkripcijoms yra priskirti garso įrašai, reikia sukurti penkis tekstinius failus.

Pirmas failas – *spk2gender*. Šis failas skirtas atskirti diktorius pagal jų indentifikavimą bei lyti. Informacinio failo apie diktoriaus lytį pateikiama 4 paveiksle.

```
LP004 m
LP005 f
LP006 f
LP007 f
LP008 f
LP009 f
LP010 f
LP011 m
LP012 f
LP013 f
```

6 pav. Spk2gender kodo dalis

Diktoriaus identifikavimo failą sudaro diktoriaus šifruotas kodas: LP001, LP002, LP003 ir panašiai, taip pat prie diktoriaus žyminti m ir f raidės, parodo asmens lytį: f – moteris, m – vyras.
Antras failas – utt2spk, šio failo paskirtis, tai garso įrašų sąsaja su diktoriais. Failo struktūra – tai diktoriaus identifikatorius bei garso įrašo identifikatorius. 7 paveiksle pateikiama utt2spk failo informacinė struktūra.

```
Z002Mg_001_01 LP002
Z002Mg_001_02 LP002
Z002Mg_001_04 LP002
Z002Mg_001_06 LP002
Z002Mg_001_08 LP002
Z002Mg_001_10 LP002
Z002Mg_001_12 LP002
```

7 pav. Utt2spk kodo dalis

Garso failą identifikuoja bei jį sudaro jo šifras ir tariamas žodis. Garso įrašas atkartojamas nuo 00 iki 19.
Trečias failas – text, šio failo struktūrą sudaro diktoriaus identifikacija bei įrašo transkripcija, kad būtų aiškiau, 8 paveiksle pateikiamas pavyzdys su įrašų sąsaja ir tekstiniu formatu.

```
Z006Mg_020_09 tyliau
Z006Mg_020_10 garsiau
Z006Mg_020_11 padidink
Z006Mg_020_12 pilnas vaizdas
```

8 pav. Text failo fragmentas

Ketvirtas failas – wav.scp, tai turbūt vienas iš svarbiausių failų Kaldi paketo atkūrime – tai garso įrašų bei realių garso įrašų kompiuterio atmintyje subendrinimas. Šio failo struktūra susideda iš garso įrašo identifikavimo bei garsinio failo pavadinimo su tikslia katalogo vieta. 9 paveiksle parodyta, kaip atrodo failas, skirtas sąsajai su realiu garso įrašu.

```
Z002Mg_001_02 /home/LIEPA/D002/Z001/Z002Mg_001_02.wav
Z002Mg_001_04 /home/LIEPA/D002/Z001/Z002Mg_001_04.wav
Z002Mg_001_06 /home/LIEPA/D002/Z001/Z002Mg_001_06.wav
Z002Mg_001_08 /home/LIEPA/D002/Z001/Z002Mg_001_08.wav
Z002Mg_001_10 /home/LIEPA/D002/Z001/Z002Mg_001_10.wav
Z002Mg_001_12 /home/LIEPA/D002/Z001/Z002Mg_001_12.wav
```

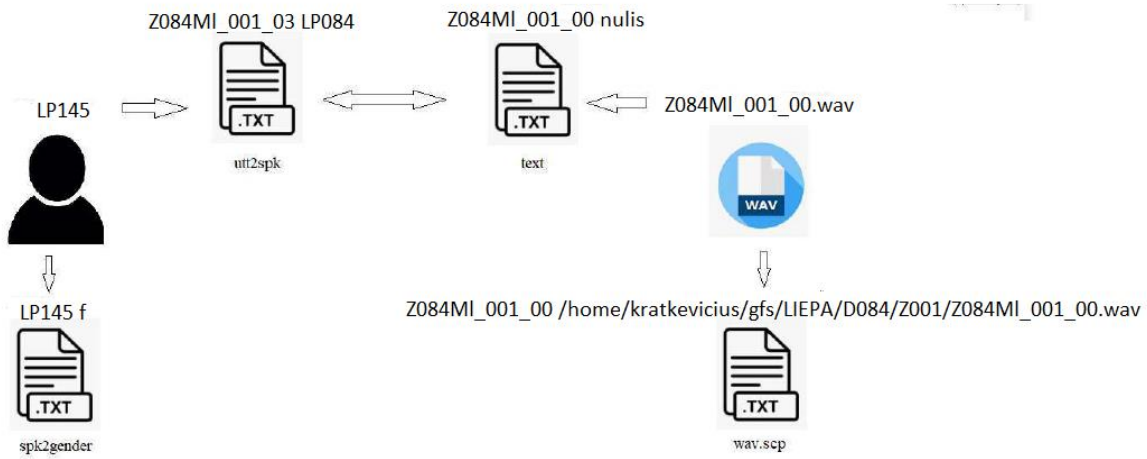
9 pav. Wav.scp kodo fragmento dalis

Penktas failas – corpus.txt, šiame faile paruošiamas failas su visų apmokymui ir testavimui skirtų garso failų transkripcijų sąrašu. Failą sudaro 27700 eilučių, kur kiekviena eilutė identifikuoja atskirą įrašą su atskiru žodžiu, pavyzdį matome 10 paveiksle.

```
pagalba
kà daryti
sustok
pirmyn
atgal
þemyn
aukðtyn
á kairæ
á deðinæ
toliau
atverk árađus
rodyk
paþymëk failà
pasirinkti
```

10 pav. Corpus.txt fragmentas

Aukščiau išvardintus failus reikia sukurti atskirai testavimo ir apmokymo dalims. Failų turinys priklauso nuo to, kokie diktoriai yra paskirti apmokymui ir testavimui. Testavimui visada yra naudojamas mažas kiekis garsyno diktorių. Taip pat yra kuriami atskiri failai, kurie yra talpinami kataloge LIEPA/data/test. Apmokymui skirti failai yra talpinami kataloge data/train. Bendras visų failų, išvardintų aukščiau, sąsajos pavaizduotos 11 paveiksle.



11 pav. Garsyno tarpusavio sąsajos ryšys

1.3.3. Fonemų paruošimas

Fonemų paruošimą sudaro: *non silence, silence phones, lexicon* bei *optional silence* tekstiniai failai.

Pirmas failas – lexicon.txt: Ištirti fonetines transkripcijas yra naudojamos grafemos, rašto vienetai, kurie yra lygūs kalbos vienetams, t.y. fonemoms, neįskaitant dvibalsių ie ir uo, kurie sudaro atskiras grafemas. Jos pateiktos 12 paveiksle.

```
adreso aA d' r' e s oo
adresus aA d' r' e s u s
adresyno a d' r' e s' ii n oo
adresynà a d' r' e s' ii n aa
adresà aA d' r' e s aa
adresø a d' r' e s uU
adverk a d' v' e r k
afaanasijus a f a a n a s' i j' u s
afanasijus a f a n aA s' i j' u s
afasijus a f a s' i j' u s
afganistane a f g a n' i s t a n' E
afri a f' r' i
afrika aA f' r' i k a
afrikinio a f' r' i k' i n' oo
afrikinis a f' r' i k' i n' i s
afriko a f' r' i k oo
afrikoe a f' r' i k oo e
afrikoj aA f' r' i k oo j
afrikoje aA f' r' i k oo j' e
afrikos aA f' r' i k oo s
afrikà aA f' r' i k aa
```

12 pav. Failo lexicon.txt iškarpa

Nonsilence_phone.txt. Fonemų sąrašas, kuriuose nėra tylos fonemų pateiktas 13 paveiksle.

```
l'
i
s
v'
Ie
a
d
t'
r'
iI
k'
e
t
u
I
p'
n'
S'
s'
ii
```

13 pav. Tariamų fonemų sąrašas

Silence_phones.txt. Tyliųjų fonemų sąrašas pateiktas 14 paveiksle.

```
sil
spn
_pauze
_ikvepimas
_iskvepimas
_cepsejimas
_nurijimas
_tyla
```

14 pav. Tyliųjų fonemų sąrašas

1.4. Garsyno LIEPA analizė

Garsyno LIEPA analizė ir atpažinimo tyrimai su Kaldi paketu.
Garsyną LIEPA galima išskaidyti į 3 dalis:

- LIEPA_ZOD: komandos, sudarytos iš vieno arba kelių žodžių ir įrašytos į atskirus failus;
- LIEPA_SAK: sakiniai;
- LIEPA_SEK: komandų sekos, atskirtos pauzėmis.

Garsyną sudaro garso įrašai, kurių diskretizavimo dažnis 22 kHz, įrašų anotacijos ANSI ir UNICODE formatuose bei įrašų foneminės anotacijos su fonemų trukmėmis. Garso įrašai perkoduoti į 16 kHz dažnį, kad tiktų tyrimams su Kaldi paketu.

1.4.1. Garsyno dalies LIEPA_ZOD tyrimas

Šią garsyno dalį sudaro 10 atskirų rinkinių, žymimų Z000, Z001, Z020, Z021, Z022, Z023, Z024, Z060, Z061 ir Z062. Kol kas liko nepanaudoti D556 diktoriaus įrašai, kadangi jo rinkiniai pažymėti kitaip: ZS001, ZS020 ir ZS060. Garsyno dalies LIEPA_ZOD charakteristikos pateiktos 3 lentelėje.

3 lentelė. Garsyno dalies LIEPA_ZOD charakteristikos

Rinkinys	Komandų skaičius	Komandų tipas
Z000	55	Balsiai, priebalsiai, dvibalsiai.
Z001	31	Skaitmenys: nulis, vienas, pirmas, pirma ir t.t.
Z020	30	Dažnai vartojamos komandos: paleisk, sustok, pirmyn, atgal ir t.t.
Z021	66	Kompiuterio valdymo komandos: paleisk skaičiuoklę ir pan.
Z022	58	Pašto programos valdymo komandos: persiųsti, atverti ir pan.
Z023	46	Teksto redagavimo, grotuvo valdymo komandos: paleisk, grok ir pan.
Z024	47	Rečiau vartojamos kompiuterio valdymo komandos: vykdyk, išsijunk ir t.t.
Z060	26	Naršyklės valdymo komandos:
Z061	66	Biologijos terminai: ląstelė, baltymas ir pan.
Z062	281	Įvairūs žodžiai: antis, arklys, arbata ir pan.

Atpažinimo tyrimuose nenaudotas rinkinys Z000, nes jį sudaro labai trumpi balsių, priebalsių ir dvibalsių ištarimai. Likusiuose rinkiniuose yra 651 komanda. Testavimui atrinkta 37 diktorių (33 moterys ir 4 vyrai) visų 9 rinkinių įrašai. Apmokymui panaudoti 145 diktorių (108 moterys ir 37 vyrai) garso įrašai. Keturių diktorių (D601, D603, D604, D605) įrašai panaudoti ir testavimui, ir apmokymui, bet testavimui panaudoti Z021, Z022, Z023, Z024 garso įrašų rinkiniai, o apmokymui – Z001, Z020 ir Z060 garso įrašų rinkiniai. Testavimo diktorių skaičius sudaro 20,3 proc. visų diktorių.
Testavimui panaudoti 5448 garso įrašai, o apmokymui – 22246 garso įrašai, testavimo garso įrašai sudaro 19,7 proc. visų įrašų.
Tyrimuose naudotas artimas SAMPA_LT fonemų rinkinys, viso 91 fonema. Žodžių transkripcijų failą *lexicon.txt* sudaro 13076 transkripcijos, t.y., visų garsyne LIEPA sutinkamų žodžių

transkripcijos, įskaitant atskirus balsius, priebalsius, dvibalsius bei kai kurių žodžių klaidingų ištarimų arba užrašymų transkripcijas, pvz.:

```
potvinius p oo t' v' i n' u s
potvynius p Oo t' v' ii n' u s
poverpoin p oo v' e r p oo j n
poverpoint p oo v' e r p oo j n t
poverpoit p oo v' e r p oo j t
powerpoint p oo v' e r p oo j n t
```

Matome klaidingą užrašymą *potvinius* bei keletą skirtingų to paties žodžio ištarimų ir transkripcijų.

1.4.2. Garsyno dalies LIEPA_SAK tyrimas

Kol kas tyrimams naudotas tik pirmos eilės kalbos modelis, t.y., tekstinis failas *corpus.txt*, kuriame surašytos tekstinės garsyną sudarančių sakinių transkripcijos. Garsyne LIEPA naudojami tokie pauzių žymėjimai: _pauze, _tyla, _ikvepimas, _iskvepimas, _cepsejimas, _nurijimas. Šie žymėjimai surašyti į failą *silence_phones.txt*:

```
sil
spn
_pauze
_ikvepimas
_iskvepimas
_cepsejimas
_nurijimas
_tyla
```

Tuo tarpu iš failų *txt* ir *corpus.txt* šie žymėjimai buvo pašalinti juos pakeičiant tuščiais tarpais. Šią garsyno dalį sudaro 45 rinkiniai, žymimi S001, S002, ... ,S045. Rinkinių statistiką pateikti būtų sudėtinga, kadangi kai kurie diktoriai įrašė nepilnus rinkinius, pvz.: diktorius D002 iš antro rinkinio įrašė tik 19 sakinių, o diktorius D003 – 41 sakinį. Rinkiniuose sakinių skaičius dažniausiai svyruoja nuo 30 iki 40, didžiausias sakinių skaičius yra 35-tame rinkinyje (51 skirtingas sakinys), o mažiausias – 19-tame rinkinyje (24 skirtingi sakiniai). Kol kas liko nepanaudoti D556 diktoriaus įrašai, kadangi jo rinkiniai pažymėti kitaip: SS013 ir SS014.

Testavimui atrinkti nedalyvavusių apmokyme 55 diktorių (41 moteris ir 14 vyrų) visų 45 rinkinių įrašai (4642 sakiniai). Apmokymui panaudoti 237 diktorių (160 moterų ir 77 vyrai) garso įrašai (17518 sakiniai). Testavimo garso įrašai sudaro 20,9 proc. visų įrašų. Garsyno dalies LIEPA_SEK charakteristikos pateiktos 4 lentelėje.

4 lentelė. Garsyno dalies LIEPA_SAK charakteristikos

Rinkiniai	Testavimo rinkinių skaičius	Apmokymo rinkinių skaičius	Testavimo diktorių skaičius	Apmokymo diktorių skaičius
1,20,21,22,23,24,25,26,27,28,29,30,31,32, 33,34,35,36,37,38,39,40,41,60,61,65,66,67 ,68	1130	2708	44 (37 f,7 m)	138 (77 f,61 m)

1.4.3. Garsyno dalies LIEPA_SEK tyrimas

Komandų sekos, tai komandos, atskirtos jau minėtais pažymėjimais (_pauze, _tyla ir t.t.) ir surašytos į atskirus failus. Viso yra 31 sekų rinkinių tipas, trijuose rinkiniuose (Z000, Z063, Z064) įrašytos balsių, priebalsių, dvibalsių sekos kartu su žodžių, prasidedančių kuria nors raide, pavyzdžiais (sekos Z063, Z064). Šios trys sekos nebuvo nagrinėjamos. Sekų Z001, Z020, Z021, Z022, Z023, Z024, Z060, Z061, Z062 turiniai sutampa su garsyno dalies LIEPA_ZOD turiniu, o sekos nuo Z025 iki Z041 ir sekos nuo Z065 iki Z068 skiriasi savo turiniu nuo LIEPA_ZOD dalies turinio.

Sekų rinkiniuose esančių sekų skaičius svyruoja nuo 1 iki 14, o vyrauja 2, 5 ir 6 sekos viename rinkinyje. Pauzių žymėjimai pašalinti iš tekstinių failų taip pat kaip garsyno dalies LIEPA_SAK tyrimuose. Visuose tyrimuose panaudotas tas pats žodžių transkripcijų failas *lexicon.txt.*

1.5. Kalbos rinkinys LIEPA: kalbos struktūra, raidos aprašymas

1.5.1. Kalbos rinkinio reikalavimai

Reikalavimai rinkiniui dažniausiai atkeliauja iš jau pilnai ištirto rinkinio pavyzdžio, tačiau reikia įvertinti ir tai, kad skirtingi rinkiniai bei kalba reikalauja skirtingų metodologijos bei kalbos sintezės įvertinimų. Balso atpažinimas, kaip technologija, labai retais atvejais naudoja kalbų rinkinius tiesiogiai, užuot tai, skirtingi metodai keičia rinkinį į statistinį modelį, kaip užslėptąjį Markovo modelį ar kaip neuroninius tinklus. Kadangi, kalbėjimo sintezė dažniausiai naudojama tiesiogiai, todėl įvairūs metodai yra naudojami mažoms dalims sujungti, kad būtų suprantama dikcija. Pagrindinis reikalavimas rinkiniams yra aukštos kokybės duomenų prieinamumas. Pagrindiniai skirtumai yra šie (Laurinčiūkaitė S., Telsknys L., Kasparaitis P., Kliūkienė R. and Paukštytė V., 2018):

- reikalavimai fonetikai: kalbos duomenys turi apimti visas lietuvių kalbos fonemas, t.y. jie turi atspindėti visus balsius, priebalsius, frikatyvinius garsus, sklandžius priebalsius, švarius dvibalsius bei mišrius dvibalsius. Platus fonetinis aprėpimas kalbos rinkinyje suteikia galimybę platesniam akustiniam modeliui bei tyrimui;

- kalbos duomenų kiekis, iš esmės, statistiniai metodai, kuriuos naudoja kalbos atpažinimas, lemia, kad reikia naudoti kuo daugiau duomenų kurie gali atskirti ir sugrupuoti panašius kalbos požymius. Kalbos duomenys paprastai naudojami netiesiogiai, apdorojant kalbos signalą, pvz.: kiekybinis nustatymas, filtravimas, funkcijų išskyrimas ir įvairių metodų, kuriuose naudojami paslėptojo Markovo modeliai, neuroniniai tinklai, kalbos modeliavimo ar adaptacijos metodai. Kalbos sintezei reikia mažiau kalbos duomenų ir jos gali būti naudojamos tiesiogiai;

- kalbos duomenų kokybė – dėl to, kad kalbos sintezė naudoja kalbos duomenis tiesiogiai, duomenys turi būti aukščiausios kokybės. Kalbos atpažinimas, naudojant įvairius išankstinio apdorojimo būdus ir metodus padeda įveikti ar valdyti triukšmą ir kitus artefaktus bei užfiksuoti kalbos duomenų ypatybes;

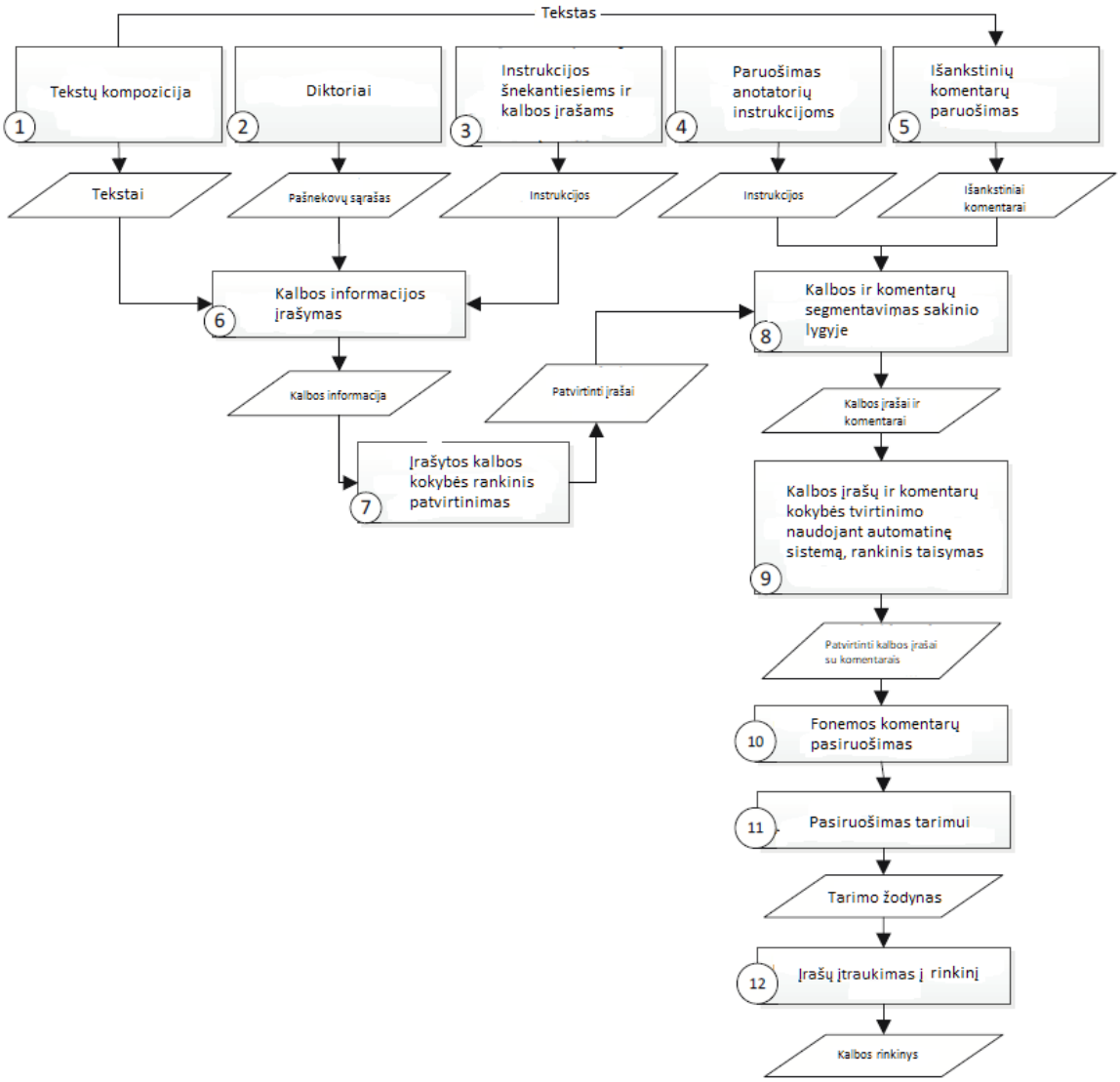
- diktorių skaičius – nuo kalbėtojų nepriklausomų kalbų atpažinimo sistemoms reikia surinkti kuo daugiau diktorių duomenų. Pagrindinis kriterijus renkantis diktorius – jiems tai turi būti

gimtoji kalba bei sugebėtų išraiškingai tarti. Kalbėtojai iš skirtingų amžiaus grupių, lyčių turi būti parinkti proporcingai. Kalbos sintezė vienai programai naudoja vieno ar kelių diktorių duomenis;

- triukšmas ar kiti įsikišimai į rinkinį – kuriant rinkinį visada reikia įvertinti žmogaus nevaldomus reiškinius, tokius kaip kosulys, juokas, čiaudulys ir kt. Iš to daroma prielaida, kad mažo lygio triukšmai neturi įtakos balso atpažinimo proceso rezultatams. Kalbos duomenys be reikšmingo triukšmo, fonetinio žodžių iškraipymo privalo būti naudojami. Kalbos sintezė atmeta galimybę naudoti duomenis, kuriuos įtakoja žmogus.

1.5.2. Rinkinio LIEPA vystymasis

Rinkinio LIEPA vystymasis panašus į esamų rinkinių vystymąsi be jokių neįprastų skirtumų ir į konkretų Kazlauskienės bei Raškinio modelį. Šio konkretaus ir pateikto proceso skirtumas yra rankų darbo mastas. Procese pavaizduoti pagrindiniai etapai paeiliui. Visos fazės yra susipynusios, skirtingi etapai vyko vienu metu. Unikalūs tam tikro proceso komponentai yra pažymėti 15 paveiksle pažymėta numeriu 9, 10 ir 11. Komponentai numeriais 9 ir 10 gali būti taikomi bet kuriai kalbai, nes jie apima nustatytą tikrinimo taisyklių rinkinį ir primityvios kalbos atpažinimo sistemos konstrukciją. Komponentas numeriu 11 yra būdingas tik kalbai ir yra jautrus foneminio atkūrimo pasikeitimui. Rinkinio liepa vystymasis pavaizduotas 15 paveiksle.



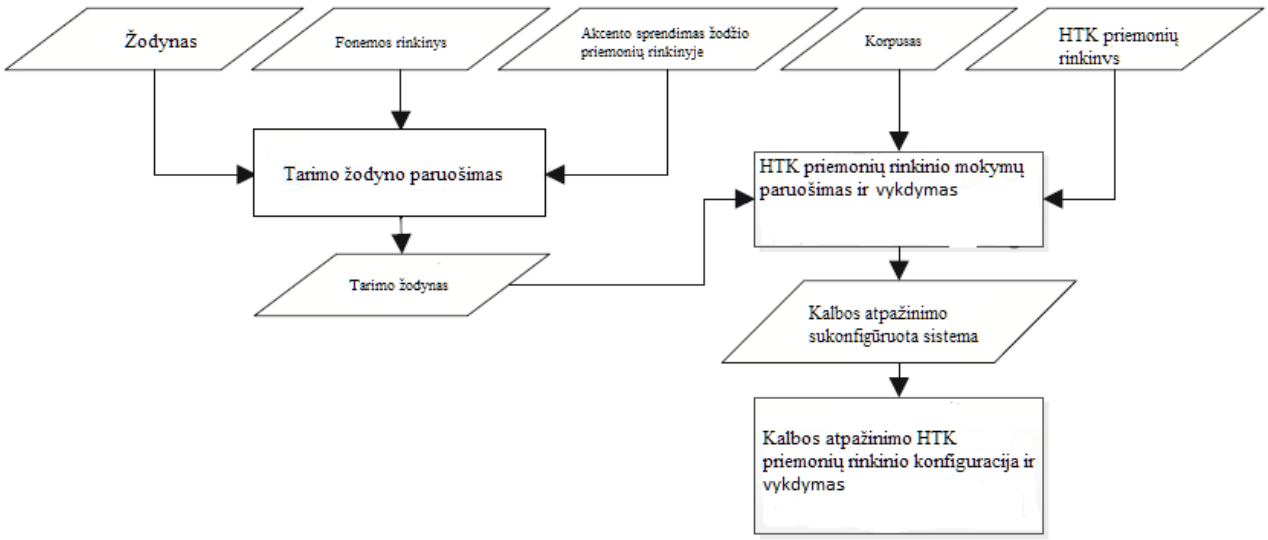
15 pav. Rinkinio LIEPA vystymosi struktūra

Pirmasis blokas susijęs su tekstų kompozicija. Kalbininkų indėlis yra svarbi sąlyga norint pasiekti reikiamą tekstą, kuris visiškai atspindėtų tam tikros kalbos foneminės erdvės unikalumą. Pateikti labai retas, tam tikros kalbos fonemas ir diftonus tampa iššūkiu ir dar sunkiau susitvarkyti su kitų kalbų fonetikos įsiliejimu į mūsų kalbą.

Antrasis blokas yra diktorių pasirinkimas. Kalbančiųjų amžius ir lytis turėtų būti pasiskirstę vienodai. Dėl papildomų kriterijų, tokių kaip etninė išskirtis, rinkinys tampa daugiafunkcinis. Šeštojo blokas – kalbos duomenų įrašymas. Septintasis blokas yra pirmasis patikrinimo etapas, kuris apima rankinį anotacijos atitikimo kalbos signalui patikrinimą. Sugeneruotas grįžtamasis ryšys apie šį bloką yra rekomendacinis, kuris leidžia pagerinti kalbos duomenų įrašymą.

Kalbos duomenų segmentavimas ir kalbos duomenų anotavimas sakinio lygiu apima aštunto bloko veiklą. Anotatorius turi išklausyti ir taisyti preliminarų komentarą, pritaikydamas jį prie garso įrašo. Antrasis devintojo bloko patikros etapas apima automatinį patikrinimą. Šis tikrinimo etapas parodytas 16 paveiksle ir susideda iš primityvios kalbos atpažinimo sistemos sukūrimo. Automatiniai įrankių rinkiniai padeda sekti kalbos duomenų ir jų anotacijų atitikimą, leidžia ištirti

failų formatus ir failų pavadinimų struktūrą, komentarų turinį ir kt. (Kazlauskiene, A., Raškinis, G., 2013).



16 pav. Antrasis automatinės patikros etapas

Dešimtajame bloke siūloma technika, aprašyta Laurinčiukaitės ir kt. (2009). Ji susideda iš kalbos duomenų pritaikymo kalbos atpažinimo procese, siekiant sukurti foneminio lygio anotacijas. Kalbos duomenų perskirstymas yra žinomas metodas, kurį naudoja daugelis tyrėjų. Jis imasi kalbos atpažinimo sistemos mokymo duomenų ir atlieka iteracinį procesą, norėdamas rasti fonemų laiko ribas kalbos duomenyse, kurios geriausiai atitinka išmoktus fonemų modelius. Šios technologijos taikymas kuriant kalbos rinkinį padeda automatizuoti anotacijos procesą. Vienuoliktas blokas padeda sudaryti tarimo žodyną: žodžių-fonemų transkripcijų forma. Pasirinktas fonemų rinkinys greičiausiai turės įtakos kalbos atpažinimo rezultatams, ir dėl šios priežasties galėtų padėti preliminarūs fonemų rinkinio tyrimai. Vienas iš galimų sprendimų yra naudoti jau sukurtą įrankių rinkinį – žodžių-fonemų. Perrašymui – konvertavimo įrankius, kad būtų sukurti įvairius fonemų rinkinius (Laurinčiukaite, S., Filipovič, M. and Telksnys, L., 2009).

1.6. Įrankiai, skirti ištirti LIEPA rinkinį

1.6.1. Tiesinė diskriminantinė analizė ir maksimali linijinės transformacijos tikimybė (LDA+MLLT)

Linijinė diskriminuojančioji analizė (LDA) yra Fišerio tiesinio diferenciatoriaus apibendrinimas – metodas, naudojamas statistikoje, modelio atpažinime ir mašininiame mokyme, siekiant surasti linijinį bruožų derinį, apibūdinantį arba atskiriantį dvi ar daugiau objektų ar įvykių klases. Gautas derinys gali būti naudojamas kaip linijinis klasifikatorius arba, paprastai, norint sumažinti matmenis prieš vėliau klasifikuojant (Fisher, 1936). LDA yra glaudžiai susijusi su dispersijos analize ir regresine analize, kurios bando vieną priklausomą kintamąjį išreikšti kaip linijinį kitų savybių ar matavimų derinį (McLachlan, 2004). Dispersinė analizė naudoja kategorinius nepriklausomus kintamuosius ir ištisinį priklausomą kintamąjį, tuo tarpu diskriminuojanti analizė turi nuolatinius nepriklausomus kintamuosius ir kategoriškai priklausomą kintamąjį (Wetcher-Hendricks, 2011). Logistinė regresija ir probitinė regresija yra panašesnė į LDA nei dispersine analizė, nes jie taip pat paaiškina kategorinį kintamąjį ištisinių nepriklausomų kintamųjų reikšmėms. Šie metodai yra

priimtinesni tais atvejais, kai nėra pagrįsta, kad nepriklausomi kintamieji paprastai yra paskirstomi, o tai yra pagrindinė LDA metodo prielaida.

Apdorojant signalus, elementų erdvės maksimalios tikimybės tiesinė regresija (fMLLR) yra visuotinė funkcijų transformacija, paprastai taikoma diktoriaus adaptacijos būdu, kai fMLLR akustines ypatybes paverčia diktoriaus pritaikytomis funkcijomis, padaugindamas operaciją su transformacijos matrica. Kai kurioje literatūroje fMLLR taip pat žinomas kaip ribotosios maksimalios tikimybės tiesinė regresija (MLLR) (Gales, 1998).

Maksimalių linijinės transformacijų tikimybių pranašumas:

- adaptacijos procesas gali būti atliekamas išankstinio apdorojimo etape ir yra nepriklausomas nuo ASR mokymo ir dekodavimo proceso;
- šio tipo pritaikytą funkciją galima pritaikyti giliuosiuose neuroniniuose tinkluose (DNN), kad būtų pakeista tradiciškai naudojama spektro gramą kalbėjimo atpažinimo modeliuose;
- maksimalių linijinės transformacijos tikimybių diktorių adaptacijos procesas lemia reikšmingą ASR modelių našumo padidėjimą, todėl lenkia kitas transformacijas ar tokias savybes kaip MFCC (Mel dažnio cestraliniai koeficientai) ir FBANKs (filtrų) koeficientai;
- maksimalios linijinės transformacijos tikimybės savybes galima efektyviai realizuoti naudojant tokius kalbų priemonių rinkinius kaip Kaldi.

Maksimalių linijinės transformacijų tikimybių trūkumas:

- duomenų apie adaptaciją kiekis yra ribotas, transformacijos matricos paprastai lengvai viršija duotus duomenis.

1.6.2. Adaptyvus mokymasis (SAT)

Akustinio modeliavimo metu diktorių adaptyvusis mokymas (SAT) buvo tradicinis Gauso mišinių modelių (GMM) metodas. Akustiniai modeliai, treniruojami naudojant SAT, tampa nepriklausomi nuo mokymo diktorių ir geriau apibendrina nematomus bandomuosius diktorius. Adaptyvaus mokymo idėja perkeliama į giliuosius neuroninius tinklus (DNN) ir siūloma sistema, leidžianti atlikti neuroninių tinklų objektų erdvėje adaptyvaus mokymo metodą. Naudodami „i“ vektorius kaip diktorių reprezentacijas, mūsų sistema išveda adaptacinį neuroninį tinklą, kad gautų diktoriaus normalizuotas savybes. Adaptuojami diktorių modeliai gaunami patikslinus DNN tokioje funkcijų erdvėje. Ši sistema gali būti taikoma įvairių tipų ypatybėms ir tinklo struktūroms, sukuriant labai bendrą SAT sprendimą (Y. Miao, H. Zhang and F. Metze, 2015).

1.6.3. Gauso mišinių modeliai (SGMM2)

Gauso mišinio modelis yra tikimybinis modelis, kuris daro prielaidą, kad visi duomenų taškai yra generuojami iš baigtinio skaičiaus – Gauso paskirstymo su nežinomais parametrais mišinio. Galima manyti, kad mišinių modeliai yra apibendrinantys k-reikšmių klasteriai, siekiant įtraukti informaciją apie duomenų kovariacinę struktūrą ir latentinių Gausų centrus.

Gauso mišinys yra funkcija, susidedanti iš kelių Gauso funkcijų, kurių kiekvienas žymimas k ∈ {1,…, K}, kur K yra mūsų duomenų rinkinio grupių skaičius. Kiekvieną Gauso k mišinyje sudaro šie parametrai:

- vidurkis μ, kuris nusako jo centrą;
- kovariancija Σ, apibrėžianti jo plotį. Tai prilygtų elipsoido matmenims daugialypiame scenarijuje;
- maišymo tikimybė π, apibrėžianti, kokia didelė ar maža Gauso funkcija bus (Bishop, 2006).

Automatinis kalbos atpažinimas labiausiai paplitęs kaip generatyvinio mokymosi metodas grindžiamas Gaussian-Mixture modeliu pagrįstais paslėptais Markovo dėsniais. Įprastinės kalbos atpažinimo sistemos naudoja Gauso mišinio modelį (GMM), pagrįstą paslėptais Markovo modeliais (HMM), kad pavaizduotų nuoseklią kalbos signalų struktūrą. HMM naudojami kalbai atpažinti, nes į kalbos signalą galima žiūrėti kaip į stacionarųjį fragmentą arba trumpai nejudantį. Per trumpą laiką kalbą galima palyginti kaip nusistovėjusį procesą. Kalba gali būti laikoma Markovo modeliu daugeliui stochastinių tikslų. Paprastai kiekviena HMM būsena naudoja Gausso mišinį garso bangos spektriniam vaizdavimui modeliuoti. GMM-HMM yra parametras $\alpha = (A, B, \pi)$, kur π yra būsenos pirmenybes vektorius, tikimybes: $A = a_{ij}$ yra būsenos perėjimo tikimybių matrica: $B = \{(b_1, \dots, b_n)\}$ ir yra aibė, kurioje b_j žymi Gauso mišinio būsenos j modelį. Būsena paprastai siejama su kalbėtojo fonemų subsegmentu (R. Lawrence and B.H. Juang, 1993) (M. A. Anusuya and S. K. Katti, 2009) (H. Sakoe and S. Chiba, 1978) (Baker, 1975) (Bilmes, 2006).

1.6.4. Laiko delsos neuroniniai tinklai (TDNN) bei gilieji neuroniniai tinklai

Laiko delsos neuroniniai tinklai (TDNN) – daugiasluoksnė dirbtinio neuroninio tinklo architektūra, kurios tikslas – 1) klasifikuoti modelius su poslinkio-invariancija ir 2) modelio kontekstą kiekviename tinklo sluoksnyje. Besikeičianti invariantinė klasifikacija reiškia, kad klasifikatoriui prieš klasifikuojant nereikia aiškių segmentų. Laikinajam modeliui (pavyzdžiui, kalbai) klasifikuoti – TDNN vengia nustatyti garsų pradžios ir pabaigos taškus prieš juos klasifikuodamas. Kontekstiniam modeliavimui TDNN kiekvienas nervinis mazgas iš kiekvieno sluoksnio gauna įvestį ne tik iš aktyvacijos, ypatybių, esančių žemiau esančiame sluoksnyje, bet ir iš vieneto išvesties modelio bei jo konteksto. Laiko signalams kiekvienas įtaisas, kaip įvestį gauna aktyvavimo schemas per tam tikrus vienetus iš žemiau esančių. Taikant dvimatę klasifikaciją (vaizdus, laiko ir dažnio modelius), TDNN gali būti treniruojamas esant poslinkio variacijai koordinačių erdvėje ir išvengia tikslaus segmentų pasiskirstymo erdvėje (Alexander Waibel, Tashiyuki Hanazawa, Geoffrey Hinton, Kiyohito Shikano and Kevin J. Lang, 1989).

Gilus mokymasis, kartais vadinamas reprezentaciniu mokymu arba neprižiūrimu savybių mokymu, yra nauja mašininio mokymosi sritis. Giluminis mokymasis tampa pagrindine kalbos atpažinimo technologija ir sėkmingai pakeitė Gauso kalbų atpažinimo ir ypatybių kodavimo mišinius. Pirmąjį tipą sudaro generatyvios giliosios architektūros, skirtos apibūdinti aukštesniųjų laipsnių duomenų koreliacines savybes arba bendrus statistinius matomų duomenų paskirstymus ir su jais susijusias klases. Naudojant Bajeso taisyklę, šios rūšies architektūra gali būti paversta diskriminantine. Tokio tipo pavyzdžiai yra įvairių formų gilieji automatiniai kodavimo įrenginiai, gilioji „Boltzmann“ mašina, suminių gaminių tinklai, originali „Giliųjų įsitikinimų tinklas“ (angl. Deep Belief Network)

forma ir jos išplėtimas į aukščiausio lygio „Boltzmann“ mašiną apatiniame sluoksnyje (D.Yu and L. Deng, 2015).

Antrojo tipo gilios architektūros yra diskriminantinė, jų paskirtis yra suteikti diskriminacinę galią klasifikuojant modelius ir tai padaryti apibūdinant klasių etikečių pasiskirstymą tolesnėje dalyje, atsižvelgiant į matomus duomenis. Pavyzdžiai yra giliai struktūruota CRF struktūra, tandemo-MLP architektūra, giliai iškilas (angl. Deep convex) ar sudedantis tinklas (angl. stacking network), jo tensorinė versija bei aptikimo pagrįsta ASR architektūra (L. Deng and X. Li, 2013).

Trečiojo tipo arba hibridinių giliųjų mokymų architektūrų tikslas yra taip pat diskriminacija, tačiau tam padeda generatyviosios architektūros rezultatai. Generacinis komponentas dažniausiai naudojamas siekiant padėti diskriminuoti kaip galutinį hibridinės architektūros tikslą (G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T., 2012)

2. Projekto tyrimo metodika

2.1. Medicininių terminų garsyno tyrimas

Pirmame baigiamojo magistro darbo ruošimo etape tirtas medicininių terminų garsynas. Lietuviškų ligų diagnozių nuo alkoholio toksinis poveikis iki tiesiosios žarnos įplyšos garso įrašų rinkinį sudaro dvylikos diktorių balso įrašai: 5 moterys ir 7 vyrai. Kiekvieno failo pavadinimas yra sudarytas iš lyties indikatoriaus (pirmoji raidė), pirmų trijų vardo raidžių, pirmų trijų pavardės raidžių, skaičiaus, kurį diktorius ištaria bei dviejų skaitmenų iteracijos – pvz.: FEM02716. Kiekviena liga ištariama 20 kartų. 12 diktorių, lietuviškų skaičių pavadinimų garso įrašų rinkinys sudarytas iš 7200 skirtingų balso įrašų, diskretizacijos dažnis – 16 kHz, viena atskaita koduojama 16 bitų. Apmokomiesiems duomenims priskiriami 10 diktorių garso įrašai, testavimo duomenims – 2 diktorių garso įrašai.

Tyrimams panaudoti 30 ir 100 diktorių garsynai. Kiekvienam metodui atliktas penkis kartus kryžminis patikrinimas taip, kad kiekvienas diktorius pakliūtų į testavimo ir apmokymo dalis. Toliau rezultatai bus pateikiami kiekvienam metodui atskirai, ir galiausiai padarytas palyginimas, kuris automatinio kalbos atpažintuvo akustinio modeliavimo metodas suteikia tiksliausią kalbos atpažinimą.

Tyrimų pradžioje buvo naudojamas 30 diktorių garsynas. Apmokymui skirti 10 diktorių garso įrašai talpinami *kaldi/egs/Gytis/audio/train* direktorijoje, o testavimui skirti 2 diktorių garso įrašai *kaldi/egs/Gytis/audio/test* direktorijoje. Tokiu principu buvo sukurti 10 atskirų garsynų aplankai, pakeičiant testavimo ir apmokymo duomenis taip, kad visi diktoriai pakliūtų į abi sritis. Rezultatams gauti *linux* terminale paleidžiamas *run.sh* skriptas. Kiekvieno bandymo rezultatai išvedami tame pačiame terminalo lange tokiu formatu, kaip pavaizduota 3 lentelėje:

3 lentelė. 12 diktorių garsyno atpažinimo rezultatai. Monofoninis metodas.

Metodai/ aplankai	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Monofoninis	0	0	0	0	0	0	0	0	0	0
Trifoninis	0	0	0	0	0	0	0	0	0	0
LDA+MLTT	0	0	0	0	0	0	0	0	0	0
LDA+MLTT+SAT	0	0	0	0	0	0	0	0	0	0

Kadangi matome, kad rezultatai su visais metodais gaunami idealūs, buvo nuspręsta, aptriukšminti garsyno failo įrašus su 5 dB baltu triukšmu (angl. white noise), kas sistemai suteiktų sunkumą susitvarkyti su failų apdorojimu. Tačiau pastebėta, kad atlikus tyrimą su užtriukšminimu, gauti rezultatai buvo artimi 0 proc. paklaidos tikimybei. Kas lėmė tai, kad šis medicininių terminų žodynas yra per trumpas, turintis nedaug duomenų su kuriais galima atlikti tyrimą. Todėl buvo nutarta pereiti prie daug sudėtingesnio, dar ne tiek daug ištirto LIEPA garsyno, kuris yra atviro kodo (angl. open-source).

2.2. Duomenų paruošimas

Duomenų paruošimas yra pirmas svarbus žingsnis, kurį reikia padaryti prieš kuriant automatinio kalbos atpažinimo sistemą. Šiuo žingsniu siekiama tinkamai suskaidyti ir paruošti duomenis prieš juos naudojant. Duomenis paruošiame taip:

- nurodome duomenis, kurie bus naudojami kuriant kalbos atpažinimo sistemą. Darbe naudojamas LIEPA žodynas. Šį korpusą sudaro dvi dalys: moterys ir vyrai. Ši ASR sistema bus pagrįsta pateiktais duomenimis. Žodyną sudaro 376 diktoriai, iš kurių 116 mokinių, 260 studentų ir šį skaičių sudarė 248 moterys ir 128 vyrai;

- LIEPA garsynas yra padalinamas į dvi dalis: apmokymo rinkinį ir bandomąjį rinkinį tiek vyrams, tiek moterims. Apmokymo rinkinį sudarė 30 proc. diktorių, klasifikatoriaus apmokymui likę – 70 proc. diktorių. Jei garsynas nėra padalintas, vartotojas, prieš pradėdamas dirbti su garsynu turi jį padalyti į šias dvi dalis;

- aplanke „Kaldi/egs“ sukuriame aplanką pasirinktu pavadinimu, pavyzdžiui „LIEPA“. Šis aplankas parodys naują ASR sistemą. Šiame aplanke sukuriamas kitas aplankas ir pavadinamas „s5“, kuris atspindės naujausią versiją;

- kitas žingsnis išskaidyti sukurtą LIEPA aplanką į dvi dalis. Aplanke „kaldi/LIEPA/s5“ sukuriamas aplankas pavadinimu „LIEPA_audio“. „Kaldi/LIEPA/s5/LIEPA_audio“. Viduje sukuriami du duomenų aplankai: „test“ ir „train“. Nukopijuojami visi duomenys, esantys „../LIEPA/WAV/test“, į „kaldi/LIEPA/s5/LIEPA_audio/test“. Taip pat nukopijuojami visi garso duomenys, esantys „../LIEPA/data/WAV/train“ į „kaldi/ LIEPA/ s5/LIEPA_audio /train“. Testo aplanke bus 113 diktorių, o apmokymo aplanke – 263 diktorių;

- grįžus į katalogą „kaldi/egs/LIEPA“ sukuriamas aplankas pavadinimu „data“; šiame aplanke sukuriami du antriniai aplankai „test“ ir „train“. Testo pakatalogis yra susijęs su testo duomenų rinkiniu, o apmokymo pakatalogis – su apmokymo duomenų rinkiniu. Kiekviename iš šių aplankų sukuriami tie patys failai (šie failai turi tuos pačius pavadinimus, bet yra susiję su skirtingais duomenų rinkiniais).

2.3. Akustinio modelio apmokymas

Sukūrus LIEPA katalogą ir paruošus akustinius bei kalbos duomenis, galima apmokyti akustinį modelį ir pritaikyti garsą su šiuo akustiniu modeliu. Šiame darbe buvo taikomi skirtingi mokymo metodai, siekiant gauti pagrįstų rezultatų ir patobulintų modelių. Kiekvieną mokymo procesą lygiuoja procesas, nes dauguma išankstinių apmokymo metodų priklauso nuo ankstesnių treniruotų akustinių modelių suderinimo verčių. Kitaip tariant, garso suderinimas su atskaitos nuorašu su naujausiu akustiniu modeliu leidžia išankstinio mokymo algoritmams naudoti šias pradines reikšmes, kad būtų patobulinti modelio parametrai.

Didžioji dalis apmokymų vyksta naudojant specialius programinius kodus, kuriuos pateikia Kaldi paketo rinkiniai. Norint pradėti apmokymo procesą be šių programinių kodų, reikia apibrėžti keletą būtinų argumentų. Šie argumentai apima:

- duomenų apmokymą data/train kataloge;

- kalbos duomenų katalogas data/lang, kuris aprėpia katalogą, kuriame yra visi kalbos modelio failai;

- šaltinio katalogas – parodo ankstesnio apmokyto modelio katalogą, exp/previous-model;

- paskirties katalogas – jame yra dabartinio modelio exp/currentmodel rezultatas;

- derinimo procesui reikėjo to paties argumentų apibrėžimo, išskyrus paskutinius du veiksmus, kur šaltinio katalogas žymi exp/current_model, o paskirties katalogas reiškia exp/current-model-ali“.

Darbe taikomi mokymo metodai išsamiai paaiškinami žemiau:

- **monofoninio modelio mokymas**: tai yra pirmasis mokymas, naudojamas monofono modeliui treniruoti. Šis „Monofono“ modelis nenaudoja jokios kontekstinės informacijos iš ankstesnio ar būsimo kalbos garso. Šis modelis nuo pat pradžių treniruojamas naudojant MFCC, delta ir pagreičio (delta + delta) funkcijas, kurios vėliau bus naudojamos kaip pradinis trifono modelių blokas. Mokyti monofoninius kalbos garsus galima naudojant programinį kodą „train_mono.sh“, kaip aprašyta 17 paveiksle:

```
#steps/train_mono.sh [options] <training-data-dir> <lang-dir> <exp-dir>"
echo
echo "===== MONO TRAINING ====="
echo

 steps/train_mono.sh --nj $nj --cmd "$train_cmd" data/train data/lang
exp/mono  || exit 1
```

17 pav. Programinio kodo train_mono.sh ištrauka

Kaip aukščiau parodyta 15 paveiksle, šaltinio katalogo apibrėžimo nėra, ir tai yra todėl, kad monofoninis apmokymas yra pirmasis testavimo žingsnis, kuris nėra priklausomas nuo bet kurio kito modelio.
Palyginimo procesą galima atlikti taikant šį kodą, matomą 18 paveiksle.

```
echo
 echo "===== MONO ALIGNMENT ====="
 echo

 steps/align_si.sh --nj $nj --cmd "$train_cmd" data/train data/lang exp/mono
exp/mono_ali || exit 1
```

18 pav. Palyginimo proceso kodas monofoniniam procesui

- **trifoninio modelio mokymas**: trifoninis modelis pateikia fonemos variantą dviejų kitų (kairės ir dešinės) fonemų kontekste. Trifono modelį galima išmokyti naudojant kodą: train_deltas. Šio mokymo metodo pradinis katalogas žymi monofoninio suderinimo katalogą exp/mono_aligment. Be to, norint pasirinkti trifoninio mokymo modelius, reikia nustatyti daugiau argumentų, HMM narių skaičių sprendimų medyje build_tree.sh ir Gausų skaičius lygtyje. Trifoninio modelio apmokymą galima atlikti, kaip pavaizduota 19 paveiksle.

```
echo
echo "=========TRI1 (first triphone pass) TRAINING ========"
echo

steps/train_deltas.sh --cmd "$train_cmd" 300 3000 data/train data/lang
exp/mono_ali exp/tri1 || exit 1
```

19 pav. Trifoninio modelio apmokymo kodas

Kode galime pamatyti dvi nežinomųjų reikšmes, tai būtų 300 ir 3000. Skaičius 300 reprezentuoja HMM būsenų skaičių, o skaičius 3000 žymi Gauso skaičių. Trifoninį modelį galime suderinti taip, kaip parodyta 20 paveiksle.

```
echo
echo "==============align for the tri1==============="
echo

steps/align_si.sh --nj $nj --cmd "$train_cmd" \
 --use-graphs true data/train data/lang exp/tri1 exp/tri1_ali
```

20 pav. Trifoninio modelio suderinimo kodas

- **LDA-MLLT**: LDA reiškia linijinę diskriminacijos analizę, o MLLT – maksimalią linijinės transformacijos tikimybę. LDA-MLLT apmokymo metodas yra geresnis daugumos operacijų metu lyginant su monofoniniu bei trifoninio mokymo metodais. Linijinė diskriminacijos analizė LDA taikoma suskaidytoms MFCC funkcijoms su kairiuoju ir dešiniuoju kanalais. Tiek LDA, tiek MLLT apdoroja objekto transformaciją dviem etapais. Pirmiausia, LDA naudoja funkcijų komponentus ir sumažina visų duomenų ypatybes iki 40, kad būtų sukurta HMM būsena. Antra, MLLT gauna sumažintą duomenų paketą iš LDA ir taiko tiesinę paprastąją transformaciją, kad kiekvienam iš diktorių būtų suteikta reikšminga transformacija (Plátek, 2014). Šį mokymo metodą galima atlikti naudojant kodą: train_lda-mllt.sh, parodytą 21 paveiksle.

```
echo
echo "=========train and decode tri2b [LDA+MLLT]====="
echo

steps/train_lda_mllt.sh --cmd "$train_cmd"\
 --splice-opts "--left-context=3 --right-context=3" \
300 3000 data/train data/lang exp/tri1_ali exp/tri2b
```

21 pav. LDA-MLLT apmokymo modelio kodas

Aukščiau pateiktame kode galima matyti, kad apmokymo kodas yra padalintas į dvi dalis, tai kairįjį bei dešinįjį, tačiau LDA-MLLT suderinimas gali būti atliekamas kaip parodyta 22 paveiksle.

```
steps/align_si.sh --nj $nj --cmd "$train_cmd" --use-graphs true \
  data/train data/lang exp/tri2b exp/tri2b_ali
```

22 pav. Kitas LDA-MLLT suderinimo varianto kodas

- **LDA-MLLT-SAT modelis**: SAT reiškia diktorių adaptyvųjį mokymą, kuris taikomas paslėpto Markovo modelio principu (HHM), kuris perima Gauso mišinių modelius (GMM), kalbos atpažinimo priemonėms. Taip pat SAT yra apmokymo metodika, kuria normalizuojamas diktorius ir triukšmas pritaikant kiekvienam diktoriui tam tikrą duomenų transformaciją, kad būtų sukurtos didelio našumo ir atpažinimo tikslumo kalbos atpažinimo sistemos (O. Tsubasa, M. Shigeki and X. Lu, 2014). Kaldi paketas pateikia kodą pavadinimu train_sat.sh, kurio pagalba galima apmokyti SAT modelius. Šio metodo vykdymas pavaizduotas 23 paveiksle.

```
echo
echo "============Do LDA+MLLT+SAT, and decode=========="
echo

steps/train_sat.sh 300 3000 data/train data/lang exp/tri2b_ali exp/tri3b
utils/mkgraph.sh data/lang exp/tri3b exp/tri3b/graph
steps/decode_fmllr.sh --config conf/decode.config --nj $nj --cmd
"$decode_cmd" \
  exp/tri3b/graph data/test exp/tri3b/decode
```

23 pav. LDA-MLLT-SAT apmokymo modelio kodas Kaldi pakete

Pasibaigus SAT modelio apmokymui, akustinis modelis bus mokomas pagal normalizuotas diktoriaus funkcijas, o ne pagal originalias jo savybes. Labai svarbu pašalinti diktoriaus tapatybę iš funkcijos prieš naudojant ją derinimo procese. Pašalinimo procesą galima atlikti įvertinant diktoriaus tapatybę, naudojant objekto erdvės maksimalios tikimybės tiesinės regresijos (fMLLR) matricos atvirkštinę funkciją ir pašalinti ją padauginus iš atvirkštinės matricos su ypatybės vektoriumi. Šis kodas 24 paveiksle paaiškina, kaip suderinti SAT trifono modelį su fMLLR.

```
echo
echo "===== Align all data with LDA+MLLT+SAT system (tri3b)=== = "
echo

steps/align_fmllr.sh --nj $nj --cmd "$train_cmd" --use-graphs true \
  data/train data/lang exp/tri3b exp/tri3b_ali
```

24 pav. SAT modelio suderinimo kodas

- **SGMM2 modelis**: SGMM taip pat turi GMM kiekvienoje nuo konteksto priklausančioje būsenoje, tačiau užuot tiesiogiai nurodę parametrus, kiekvienoje būsenoje nurodome vektorių vj ∈ RS kartu su globalinės erdvės žemėlapiu iš šios S matmenų vektorinės erdvės į pxj parametrus. Šio modelio apmokymo kodą galima pamatyti 25 paveiksle.

```
echo
echo “=============Do SGMM2, and decode==============”
echo

steps/train_sgmm2.sh for f in $data/feats.scp $lang/G.fst $lang/L_disambig.fst
$lang/phones/disambig.int \
$srcdir/final.mdl $srcdir/tree $olddir/lat.1.gz; do
[ ! -f $f ] && echo "$0: no such file $f" && exit 1;
“$decode_cmd”\
  Usage: steps/decode_sgmm2_fromlats.sh [options] <data-dir> <lang-dir> <old-decode-dir>
<decode-dir>"
```

25 pav. SGMM2 apmokymo kodas

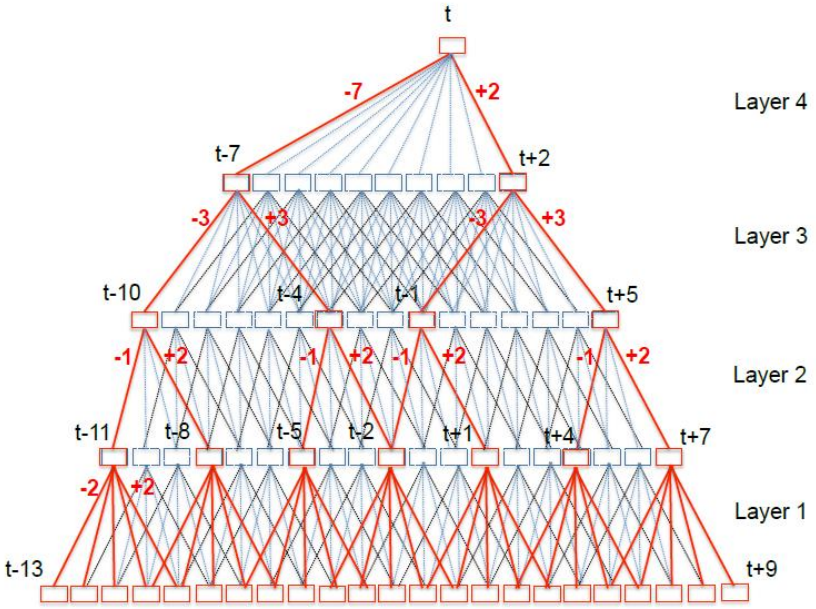
Teigiama, kad p(j|t) ≡ j(t) bus pateiktas naudojant tam tikrą standartinį pirmyn-atgal arba „Viterbi“ algoritmą. Mūsų lygtyje j(t) yra nulis arba dalimi pagrįsta Viterbi lygiavimu, kuris per keletą pirmųjų apmokymų pakartojimų gaunamas naudojant tikimybes iš pradinės GMM sistemos, o vėliau gaunamas naudojant patį SGMM (D. Poveya, L. Burgetb, M. Agarwalc, P. Akyazid and F. Kaie, 2010). Galime pamatyti derinimo kodą 26 paveiksle.

```
echo
echo “============Align all data with SGMM2 system===============”
echo

echo “$0: converting alignments”
$cmd JOB=1:$nj $dir/log/convert_ali.JOB.log\
convert-ali $alidir/final.mdl  $dir/0.mdl $dir/tree “ark:gunzip –c $alidir/ali.JOB.gz|”\
“ark:|gzip –c >$dir/ali.JOB.gz” || exit 1
```

26 pav. SGMM2 derinimo kodas

- **TDNN modelis**: Apdorojant platesnį laiko kontekstą, standartiniame DNN, pradiniame sluoksnyje sužinoma afinistinė transformacija visam laikiniam kontekstui. Tačiau TDNN architektūroje pradinės transformacijos išmokstamos siaurose situacijose, o gilesni sluoksniai apdoroja paslėptus aktyvinimus iš platesnės laiko juostos. Taigi aukštesnieji sluoksniai turi galimybę išmokti platesnių laiko santykių. Kiekvienas TDNN sluoksnis veikia skirtinga laiko skiriamąja geba, kuri padidėja pereinant į aukštesnius tinklo sluoksnius. TDNN architektūros transformacijos yra susietos per tam tikrus žingsnius ir dėl šios priežasties jos laikomos konvoliucinių neuroninių tinklų pirmtaku. Atliekant atgalinį dauginimą dėl susiejimo, apatiniai tinklo sluoksniai atnaujinami gradientu, sukauptu per visus įvesties laiko konteksto laiko žingsnius. Taigi, apatiniai tinklo sluoksniai yra priversti mokytis nekintamų vertimų transformacijų (Waibel, 1989). Skaičiavimo naudojant TDNN su atranka ar be jos pavaizduota 27 paveiksle.



27 pav. Skaičiavimas naudojant TDNN su atranka (raudona) ir be atrankos (mėlyna + raudona) (V. Peddinti, D. Povey and S. Khudanpur, 2015)

Taip pat TDNN apmokymo kodą galime pamatyti 28 paveiksle, o derinimo kodas pateiktas 29 paveiksle.

```
echo "                TDNN Hybrid Training & Decoding                "
echo ================================================================================

# DNN hybrid system training parameters
dnn_mem_reqs="--mem 1G"
dnn_extra_opts="--num_epochs 20 --num-epochs-extra 10 --add-layers-period 1 --shrink-interval 3"

steps/nnet2/train_tanh.sh --mix-up 5000 --initial-learning-rate 0.015 \
  --final-learning-rate 0.002 --num-hidden-layers 2 \
  --num-jobs-nnet "$train_nj" --cmd "$train_cmd" "${dnn_train_extra_opts[@]}" \
 data/train data/lang exp/tri3_ali exp/tri4_nnet

echo "                DNN Hybrid Decoding                "
echo
[ ! -d exp/tri4_nnet/decode ] && mkdir -p exp/tri4_nnet/decode
steps/nnet2/decode.sh --cmd "$decode_cmd" --nj "$decode_nj" "${decode_extra_opts[@]}" \
 --transform-dir exp/tri3/decode exp/tri3/graph data/test \
 exp/tri4_nnet/decode | tee exp/tri4_nnet/decode/decode.log
```

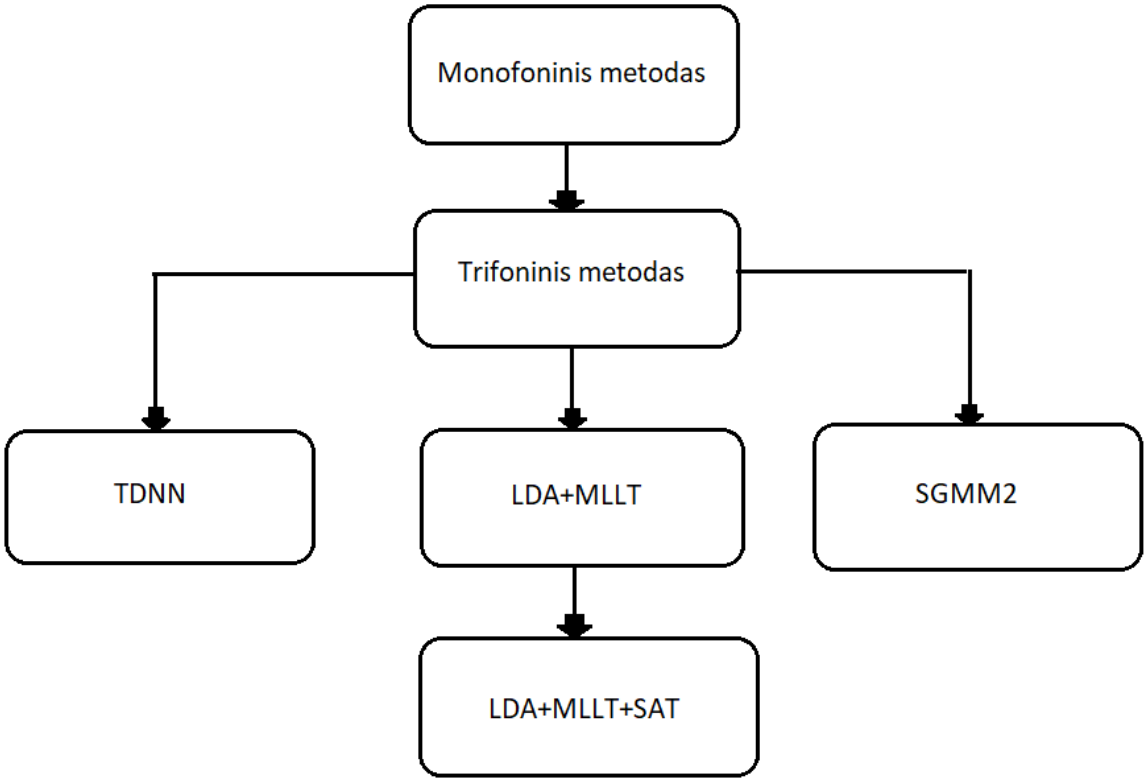
28 pav. TDNN apmokymo ir dekodavimo kodas

```
echo
echo "========Align all data with TDNN system=========="
echo

steps/nnet2/decode.sh\
--nj $(wc -1<data/$decode_set/spk2utt) –cmd "$decode_cmd" $iter_opts \
--online-ivector-dir exp/nnet2/ivectors_${decode_set} \
$graph_dir data/${decode_set}_hires $dir/decode_${decode_set}${decode_iter:+_$decode_iter}
||exit 1;
```

29 pav. TDNN derinimo kodas

Visų šių metodų tarpusavio ryšiai pateikti 30 paveiksle. Iš jo galima pamatyti visų metodų apmokymo ryšį.



30 pav. Apmokymo metodu tarpusavio ryšys

3. Rezultatai

Šio skyriaus pirmoje dalyje yra pateikiami garsyno LIEPA_ZOD atpažinimo tikslumo tyrimo rezultatai, antroje dalyje pateikiami LIEPA_SEK atpažinimo tikslumo rezultatai.

3.1. Garsyno LIEPA izoliuotų žodžių rezultatai

Garsyno LIEPA_ZOD atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatorių medžių šakų skaičių naudojant 5 atpažinimo metodus (monofoninį, trifoninį, LDA, SAT, SGMM2) bei nuo paslėptųjų sluoksnių bei neuronų skaičiaus juose naudojant TDNN metodo dvi modifikacijas.
Pereinant nuo vieno atpažinimo metodo prie kito paliekamos prieš tai naudotame metode surastos parametrų, duodančių mažiausią atpažinimo klaidą, reikšmės.

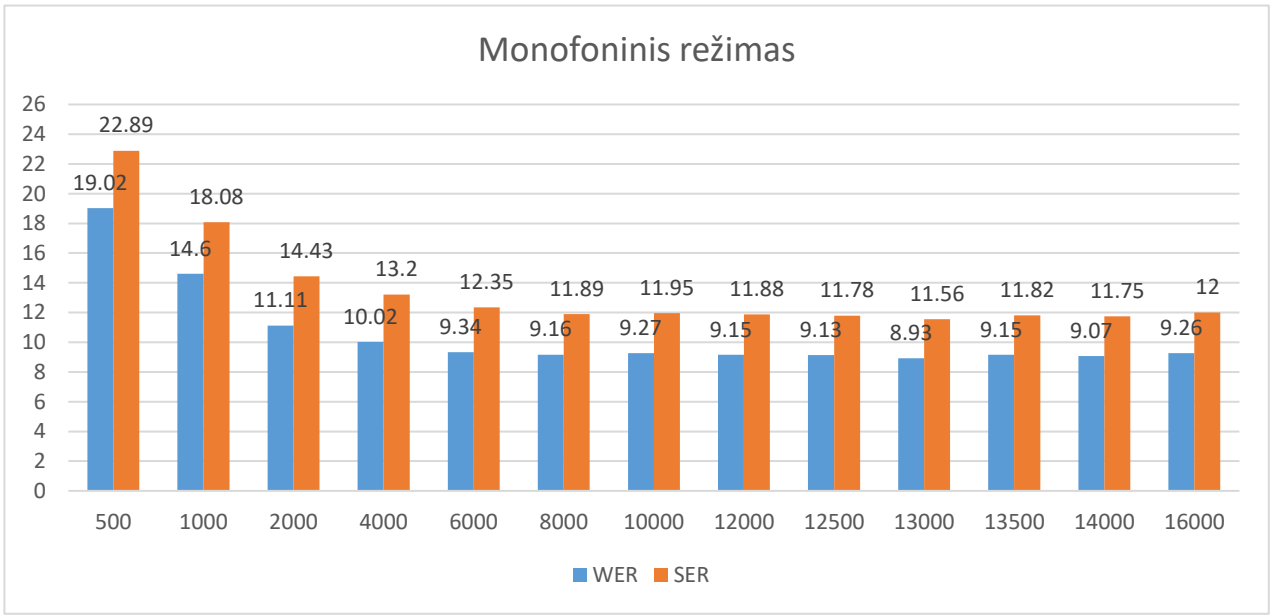
3.1.1. Izoliuotų žodžių atpažinimo tyrimas

- LIEPA_ZOD garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių skaičiaus monofoniniame režime:

Tyrimas atliktas keičiant parametrą *totgauss* nuo 500 iki 16000. Rezultatai – 3 lentelėje bei 19 paveiksle.

3 lentelė LIEPA_ZOD garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių skaičiaus monofoniniame režime tyrimo rezultatai

	Gauso mišinių skaičius												
	500	1000	2000	4000	6000	8000	10000	12000	12500	13000	13500	14000	16000
WER	19,02	14,60	11,11	10,02	9,34	9,16	9,27	9,15	9,13	8,93	9,15	9,07	9,26
SER	22,89	18,08	14,43	13,20	12,35	11,89	11,95	11,88	11,78	11,56	11,82	11,75	12,00



31 pav. Gauso mišinių skaičiaus ir žodžio klaidos dažnio bei sakinių klaidos dažnio rezultatas, monofoniniame rėžime, žodžių garsyne

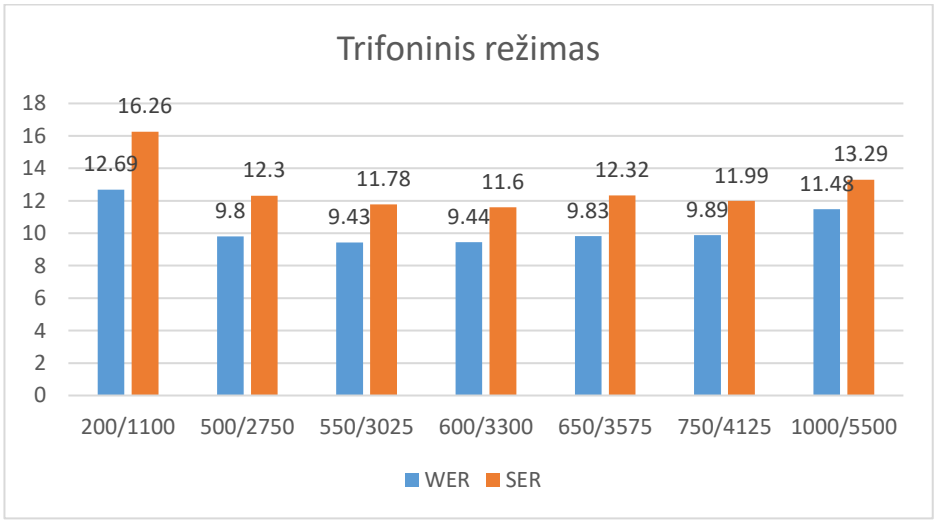
Matome, kad geriausi rezultatai, su mažiausia klaidos tikimybe gauti esant totgauss koeficientui lygiam 13000. Gauti rezultatai: WER (angl. Word error rate) – 8,93% ir SER (angl. Sentence error rate) – 11.56%.

- LIEPA_ZOD garsyno atpažinimo tikslumo rezultatai, gauti keičiant Gauso mišinių bei klasifikatorių medžių šakų skaičius trifoniniame režime:

Tyrimas atlikti keičiant numatytąsias parametrų reikšmes:
numLeavesTri1=2000
numGaussTri1=11000
Pradžioje buvo keičiamas *numLeavesTri1* ir ieškoma mažiausios atpažinimo klaidos perskaičiuojant parametrą *numGaussTri1* santykiu 1:5,5. Rezultatai – 4 lentelėje bei 20 paveiksle.

4 lentelė. LIEPA_ZOD garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičiaus trifoniniame režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius						
	200/1100	500/2750	550/3025	600/3300	650/3575	750/4125	1000/5500
WER	12,69	9,80	9,43	9,44	9,83	9,89	11,48
SER	16,26	12,30	11,78	11,60	12,32	11,99	13,29



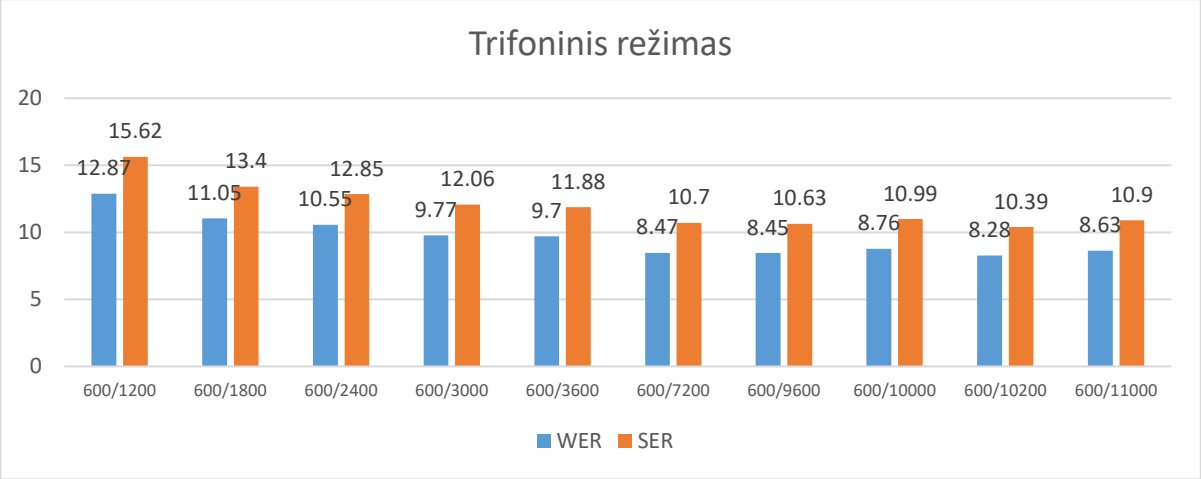
32 pav. Žodžių garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičiaus trifoniniame režime

Galime pastebėti rezultatų lentelėje, kad keičiant parametrus numLeavesTri1, geriausi rezultatai gaunami esant: numLeavesTri1 nustatymui 600/3300, kuriuo atžvilgiu, gauname, kad rezultatai WER – 9,44% ir SER – 11,60%.

Kitu atveju buvo keičiamas parametras *numGaussTri1* išlaikant pastovų *numLeavesTri1* bet keičiant santykį 1:5,5. Rezultatai matomi 5 lentelėje bei 21 paveiksle.

5 lentelė. LIEPA_ZOD garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus trifoniniame režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius									
	600/1200	600/1800	600/2400	600/3000	600/3600	600/7200	600/9600	600/10000	600/10200	600/11000
WER	12,87	11,05	10,55	9,77	9,70	8,47	8,45	8,76	8,28	8,63
SER	15,62	13,40	12,85	12,06	11,88	10,70	10,63	10,99	10,39	10,90



33 pav. Žodžių garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus trifoniniame režime

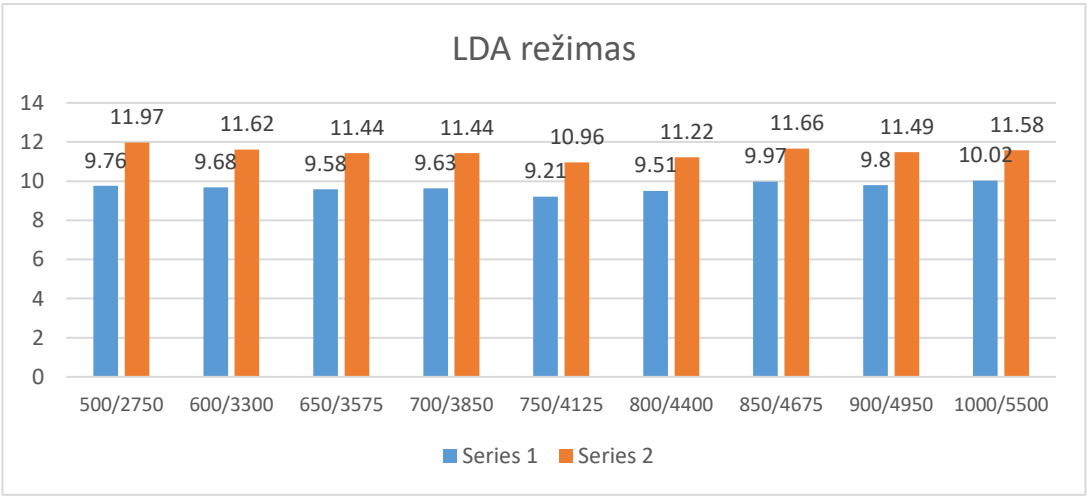
Geriausias rezultatas gautas, kai *numLeavesTri1=600, numGaussTri1=10200.* Gauname, kad rezultatas WER – 8,28% ir SER – 10,39%.

- LIEPA_ZOD garsyno atpažinimo tikslumo rezultatai, gauti keičiant Gauso mišinių bei klasifikatorių medžių šakų skaičius LDA režime:

Tyrimas atlikti keičiant numatytąsias parametrų reikšmes:
numLeavesMLLT=2000
numGaussMLLT=11000
Pradžioje buvo keičiamas *numLeavesMLLT* ir ieškoma mažiausios atpažinimo klaidos perskaičiuojant parametrą *numGaussMLLT* santykiu 1:5,5. Rezultatai – 6 lentelėje bei 22 paveiksle.

6 lentelė. LIEPA_ZOD garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičiaus LDA režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius								
	500/2750	600/3300	650/3575	700/3850	750/4125	800/4400	850/4675	900/4950	1000/5500
WER	9,76	9,68	9,58	9,63	9,21	9,51	9,97	9,80	10,02
SER	11,97	11,62	11,44	11,44	10,96	11,22	11,66	11,49	11,58

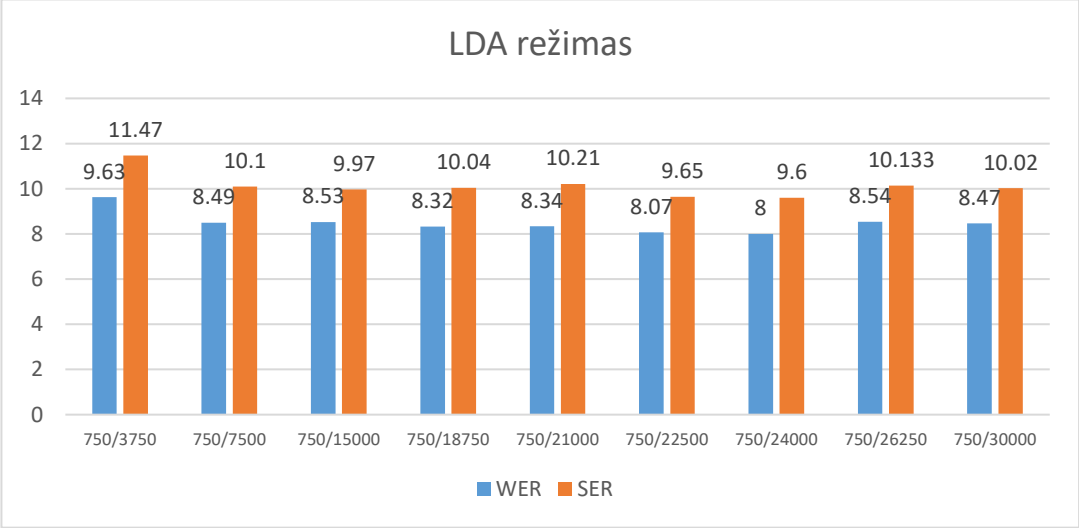


34 pav. Žodžių garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičiaus LDA režime

Po to buvo keičiamas parametras *numGaussMLLT* išlaikant pastovų *numLeavesMLLT* bet keičiant santykį 1:5,5. Rezultatai – 7 lentelėje bei 23 paveiksle.

7 lentelė. LIEPA_ZOD garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus LDA režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius								
	750/3750	750/7500	750/15000	750/18750	750/21000	750/22500	750/24000	750/26250	750/30000
WER	9,63	8,49	8,53	8,32	8,34	8,07	8,00	8,54	8,47
SER	11,47	10,10	9,97	10,04	10,21	9,65	9,60	10,13	10,02



35 pav. Žodžių garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus LDA režime

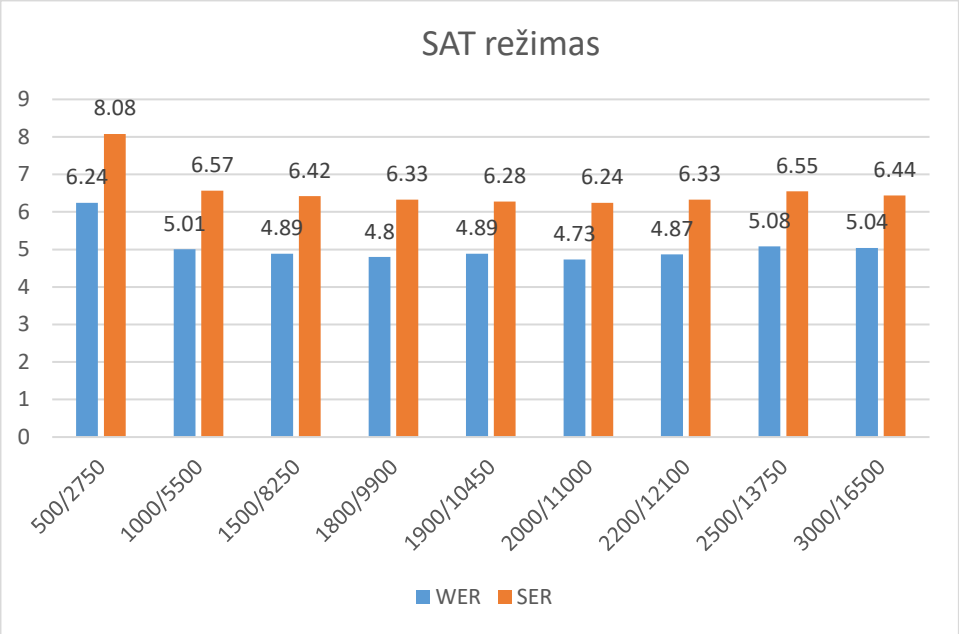
Geriausias rezultatas gautas, kai *numLeavesMLLT=750, numGaussMLLT=24000.*

- LIEPA_ZOD garsyno atpažinimo tikslumo rezultatai, gauti keičiant Gauso mišinių bei klasifikatorių medžių šakų skaičius SAT režime:

Tyrimas atlikti keičiant numatytąsias parametrų reikšmes:
numLeavesSAT=2000
numGaussSAT=11000
Pradžioje buvo keičiamas *numLeavesSAT* ir ieškoma mažiausios atpažinimo klaidos perskaičiuojant parametrą *numGaussSAT* santykiu 1:5,5. Rezultatai – 8 lentelėje bei 24 paveiksle.

8 lentelė. LIEPA_ZOD garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičiaus SAT režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius								
	500/2750	1000/5500	1500/8250	1800/9900	1900/10450	2000/11000	2200/12100	2500/13750	3000/16500
WER	6,24	5,01	4,89	4,80	4,89	4,73	4,87	5,08	5,04
SER	8,08	6,57	6,42	6,33	6,28	6,24	6,33	6,55	6,44

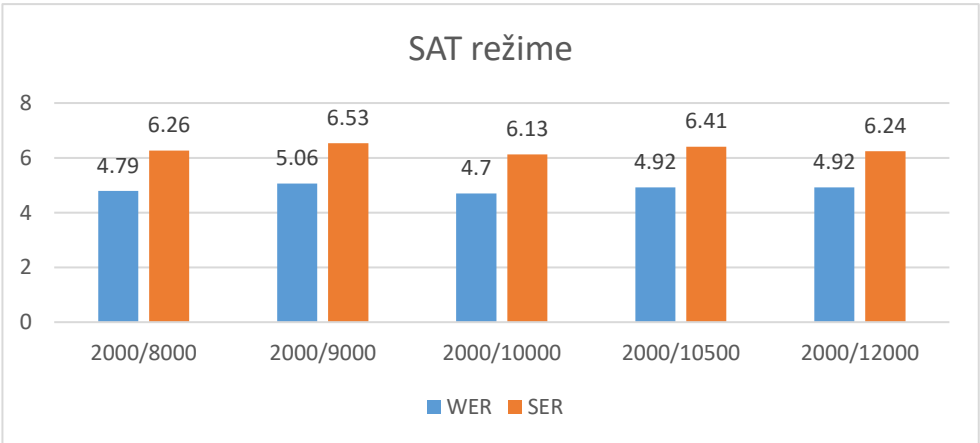


36 pav. Žodžių garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičiaus SAT režime

Po to buvo keičiamas parametras *numGaussSAT* išlaikant pastovų *numLeavesSAT* bet keičiant santykį 1:5,5. Rezultatai – 9 lentelėje bei 25 paveiksle.

9 lentelė. LIEPA_ZOD garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus SAT režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius				
	2000/8000	2000/9000	2000/10000	2000/10500	2000/12000
WER	4,79	5,06	4,70	4,92	4,92
SER	6,26	6,53	6,13	6,41	6,24



37 pav. Žodžių garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus SAT režime

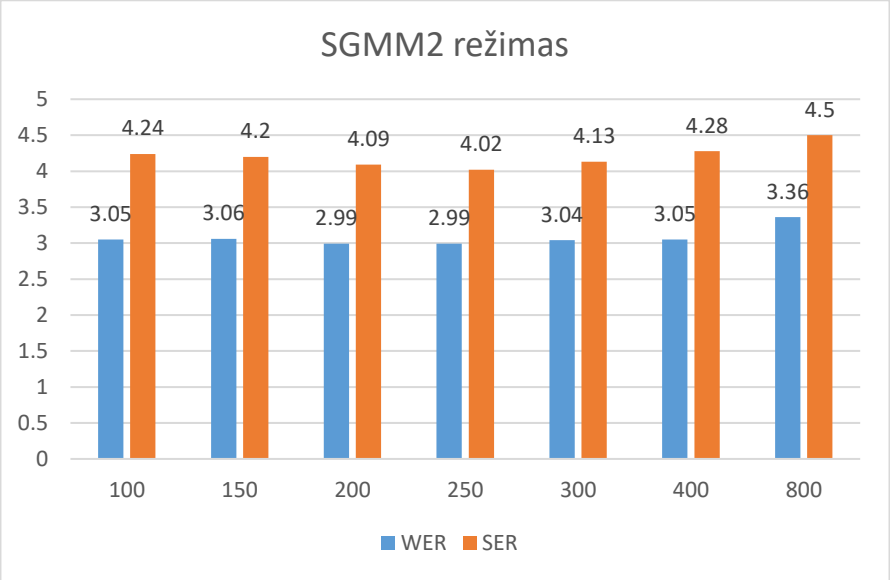
Geriausias rezultatas gautas, kai *numLeavesSAT=2000, numGaussSAT=10000.*

- LIEPA_ZOD garsyno atpažinimo tikslumo rezultatai, gauti keičiant Gauso mišinių bei klasifikatorių medžių šakų skaičius SGMM2 režime:

Tyrimas atlikti keičiant numatytąsias parametrų reikšmes:
numGaussUBM=400
numLeavesSGMM=7000
numGaussSGMM=9000
Pradžioje buvo keičiamas *numGaussUBM* ir ieškoma mažiausios atpažinimo klaidos kitus parametrus išlaikant tuos pačius. Rezultatai – 10 lentelėje bei 26 paveiksle.

10 lentelė. LIEPA_ZOD garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių UBM skaičiaus SGMM2 režime tyrimo rezultatai

Klaida	Gauso mišinių UBM skaičius (*numLeavesSGMM=7000, numGaussSGMM=9000)*						
	100	150	200	250	300	400	800
WER	3,05	3,06	2,99	2,99	3,04	3,05	3,36
SER	4,24	4,20	4,09	4,02	4,13	4,28	4,50

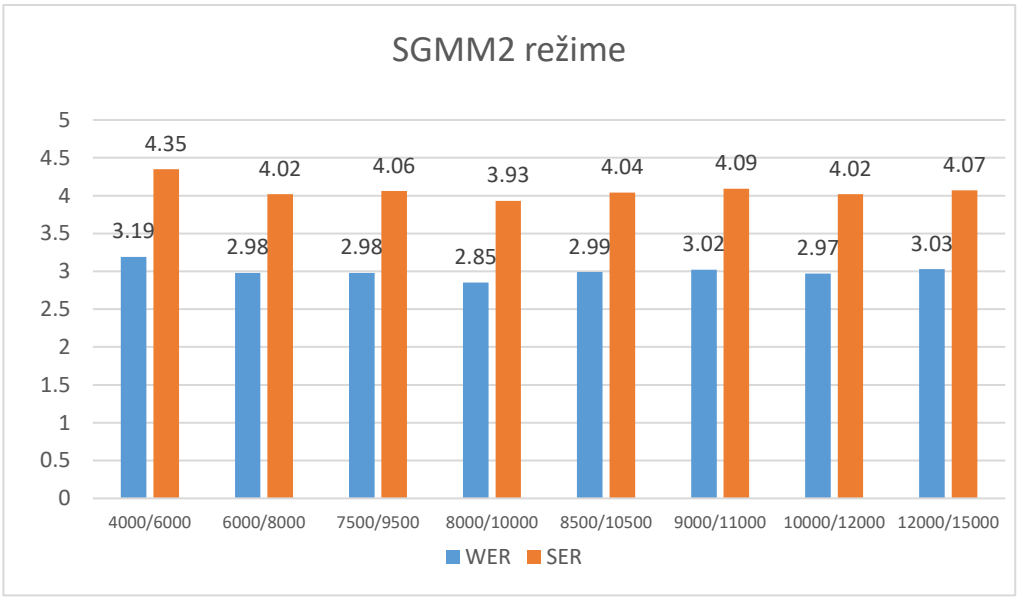


38 pav. Žodžių garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių UBM skaičiaus SGMM2 režime

Po to buvo keičiamas *numLeavesSAT* ir ieškoma mažiausios atpažinimo klaidos išlaikant *numGaussSAT* 2000 didesniu už *numLeavesSAT*. Rezultatai – 11 lentelėje bei 27 paveiksle.

11 lentelė. LIEPA_ZOD garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių SGMM bei klasifikatoriaus medžių šakų skaičiaus SGMM2 režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/Gauso mišinių SGMM skaičius (*numGaussUBM=250)*							
	4000/6000	6000/8000	7500/9500	8000/10000	8500/10500	9000/11000	10000/12000	12000/15000
WER	3,19	2,98	2,98	2,85	2,99	3,02	2,97	3,03
SER	4,35	4,02	4,06	3,93	4,04	4,09	4,02	4,07

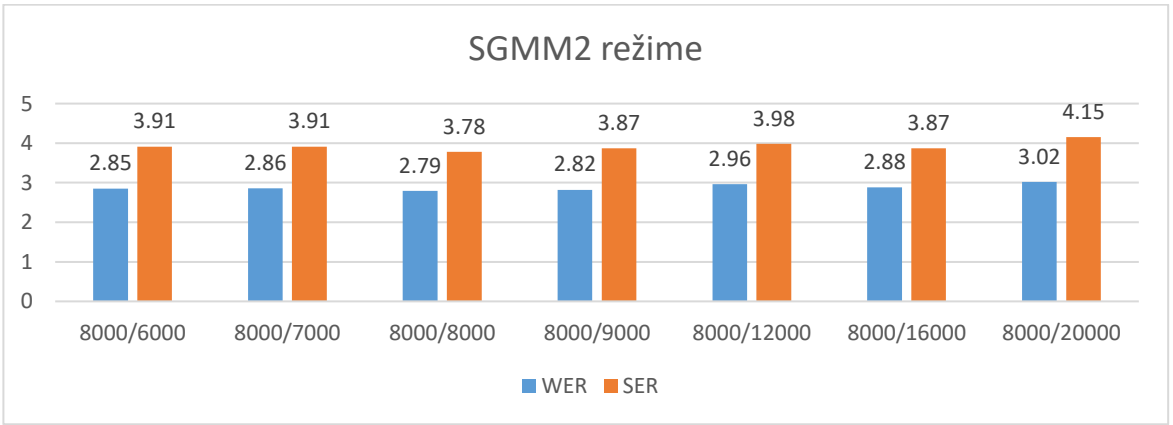


39 pav. Žodžių garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių SGMM bei klasifikatoriaus medžių šakų skaičiaus SGMM2 režime

Po to buvo keičiamas parametras *numGaussSAT* išlaikant pastovų *numLeavesSAT*. Rezultatai – 12 lentelėje bei 28 paveiksle.

12 lentelė. LIEPA_ZOD garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių SGMM skaičiaus SGMM2 režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių SGMM skaičius (*numGaussUBM=250)*						
	8000/6000	8000/7000	8000/8000	8000/9000	8000/12000	8000/16000	8000/20000
WER	2,85	2,86	2,79	2,82	2,96	2,88	3,02
SER	3,91	3,91	3,78	3,87	3,98	3,87	4,15



40 pav. Žodžių garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių SGMM skaičiaus SGMM2 režime

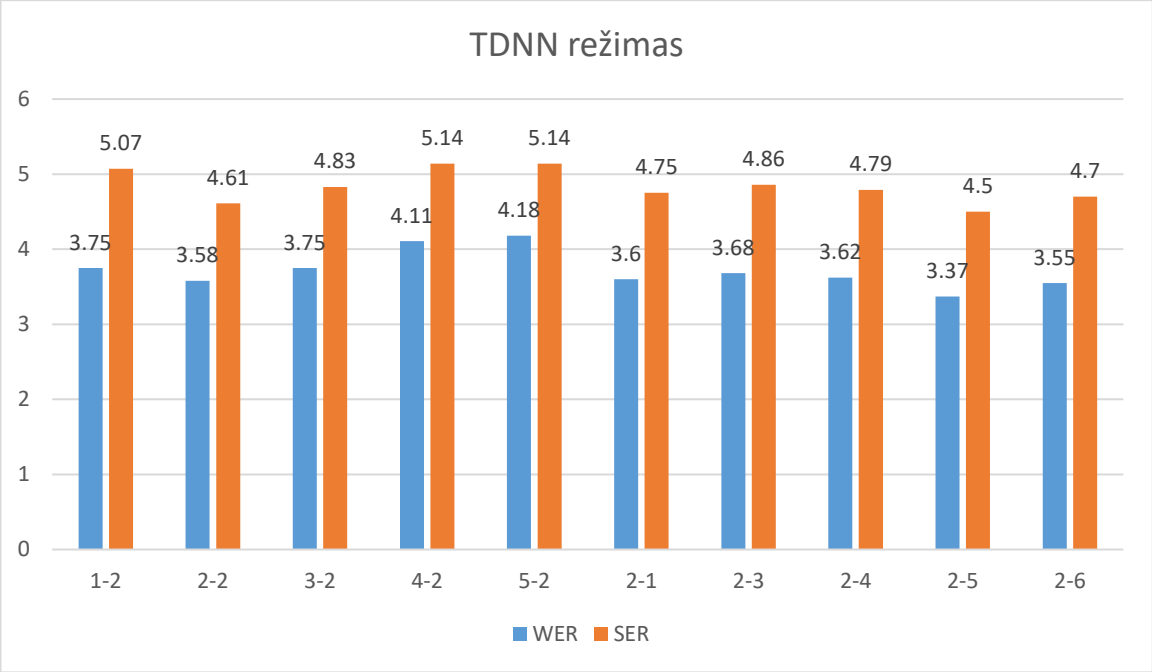
Geriausias rezultatas gautas, kai *numLeavesSAT=8000, numGaussSAT=8000.*

- LIEPA_ZOD garsyno atpažinimo tikslumo rezultatai, gauti keičiant paslėptųjų sluoksnių bei neuronų skaičių juose naudojant TDNN metodo dvi modifikacijas:

TDNN *pnorm* metodo atveju keičiamas paslėptųjų sluoksnių skaičius nuo 1 iki 6, parametras p nuo 1 iki 7 ir *pnorm_input_dim* bei *pnorm_output_dim*. Numatytosios šių parametrų reikšmės: 2 sluoksniai, p=2, *pnorm_input_dim=2000, pnorm_output_dim=200.* Tyrime naudota 20 epochų, realizuojamų kaip 60 iteracijų. Esminiai rezultatai – 13 ir 14 lentelėse taip pat 29 ir 30 paveiksluose.

13 lentelė. LIEPA_ZOD garsyno atpažinimo tikslumo priklausomybės nuo sluoksnių skaičiaus ir p parametro naudojant TDNN metodo *pnorm* modifikaciją tyrimo rezultatai

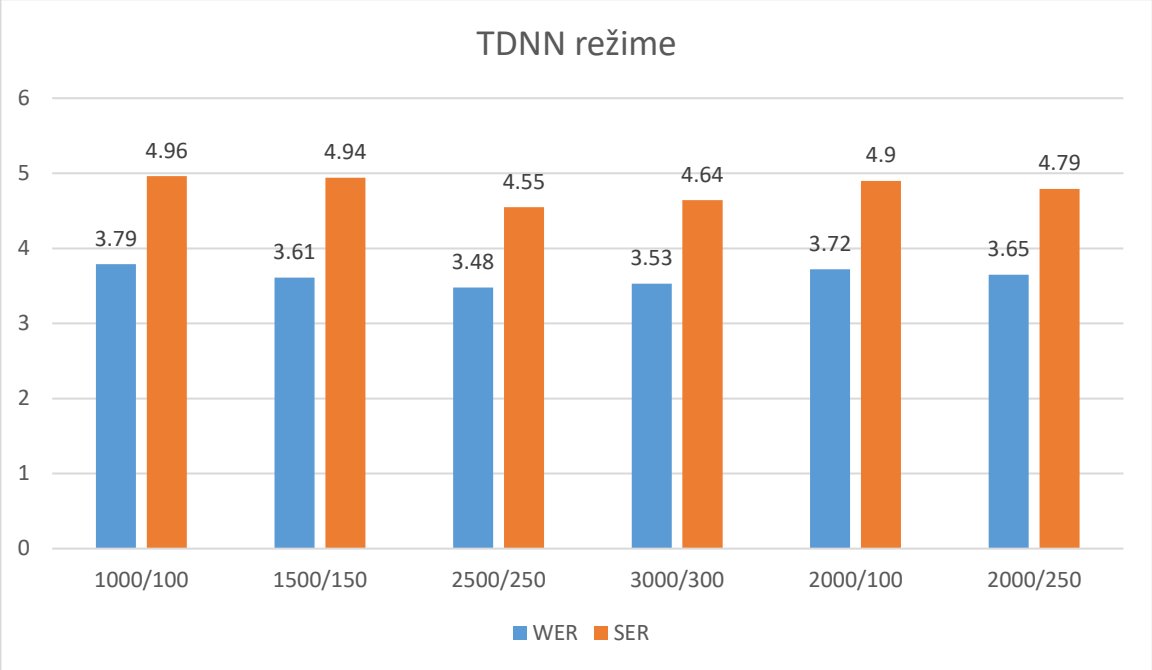
Klaida	Sluoksnių skaičius-p parametras (*input_dim=2000,output_dim=200)*									
	1-2	2-2	3-2	4-2	5-2	2-1	2-3	2-4	2-5	2-6
WER	3,75	3,58	3,75	4,11	4,18	3,60	3,68	3,62	3,37	3,55
SER	5,07	4,61	4,83	5,14	5,14	4,75	4,86	4,79	4,50	4,70



41 pav. Žodžių garsyno atpažinimo tikslumo priklausomybės nuo sluoksnių skaičiaus ir p parametro naudojant TDNN metodo *pnorm* modifikaciją

14 lentelė. LIEPA_ZOD garsyno atpažinimo tikslumo priklausomybės nuo parametrų *pnorm_input_dim* bei *pnorm_output_dim* naudojant TDNN metodo *pnorm* modifikaciją tyrimo rezultatai

Klaida	*input_dim/output_dim (2 sluoksniai, p=5)*					
	1000/100	1500/150	2500/250	3000/300	2000/100	2000/250
WER	3,79	3,61	3,48	3,53	3,72	3,65
SER	4,96	4,94	4,55	4,64	4,90	4,79

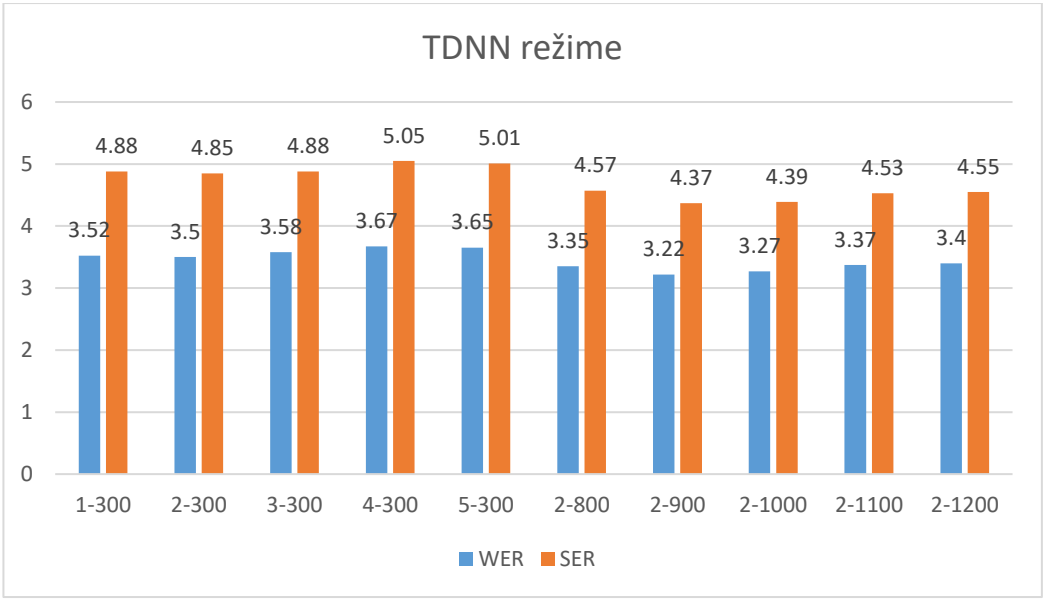


42 pav. Žodžių garsyno atpažinimo tikslumo priklausomybės nuo parametrų *pnorm_input_dim* bei *pnorm_output_dim* naudojant TDNN metodo *pnorm* modifikaciją

TDNN *tanh* metodo atveju keičiamas paslėptųjų sluoksnių skaičius nuo 1 iki 5 bei *hidden_layer_dim*. Numatytosios šių parametrų reikšmės: 2 sluoksniai, *hidden_layer_dim=300*. Tyrime naudota 20 epochų, realizuojamų kaip 20 iteracijų. Rezultatai – 15 lentelėje ir 31 paveiksle.

15 lentelė. LIEPA_ZOD garsyno atpažinimo tikslumo priklausomybės nuo sluoksnių skaičiaus ir sluoksnio dydžio naudojant TDNN metodo *tanh* modifikaciją tyrimo rezultatai

Klaida	Sluoksnių skaičius-parametras *hidden_layer_dim*									
	1-300	2-300	3-300	4-300	5-300	2-800	2-900	2-1000	2-1100	2-1200
WER	3,52	3,50	3,58	3,67	3,65	3,35	3,22	3,27	3,37	3,40
SER	4,88	4,85	4,88	5,05	5,01	4,57	4,37	4,39	4,53	4,55



43 pav. Žodžių garsyno atpažinimo tikslumo priklausomybės nuo sluoksnių skaičiaus ir sluoksnio dydžio naudojant TDNN metodo *tanh* modifikaciją

3.2. Garsyno LIEPA sekų atpažinimo rezultatai

3.2.1. Sekų atpažinimo tyrimas

LIEPA_SEK garsyno atpažinimo tikslumo priklausomybės

Patikrinsime LIEPA_SEK garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatorių medžių šakų skaičių naudojant 5 atpažinimo metodus (monofoninį, trifoninį, LDA, SAT, SGMM2) bei nuo paslėptųjų sluoksnių bei neuronų skaičiaus juose naudojant TDNN metodo dvi modifikacijas.

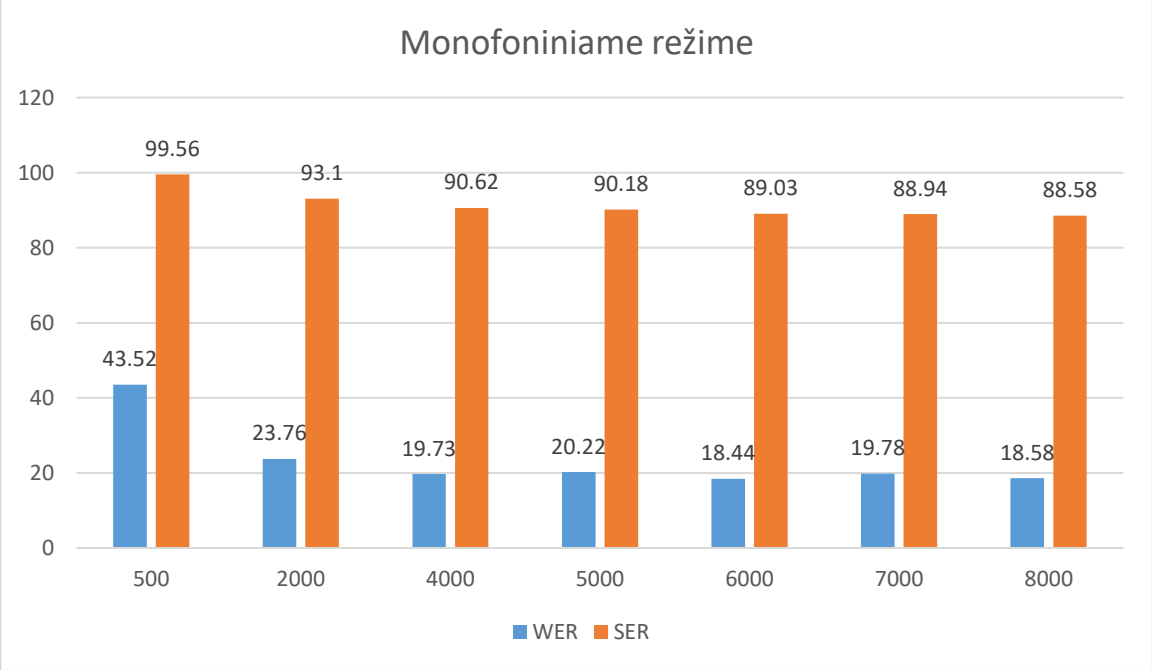
Pereinant nuo vieno atpažinimo metodo prie kito paliekamos prieš tai naudotame metode surastos parametrų, duodančių mažiausią atpažinimo klaidą, reikšmės.

- LIEPA_SEK garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių skaičiaus monofoniniame režime:

Tyrimas atliktas keičiant parametrą *totgauss* nuo 500 iki 16000. Rezultatai – 16 lentelėje ir 32 paveiksle.

16 lentelė. LIEPA_SEK garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių skaičiaus monofoniniame režime tyrimo rezultatai

Klaida	Gauso mišinių skaičius						
	500	2000	4000	5000	6000	7000	8000
WER	43,52	23,76	19,37	20,22	18,44	19,87	18,58
SER	99,56	93,10	90,62	90,18	89,03	88,94	88,58



44 pav. Sekų garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių skaičiaus monofoniniame režime

- LIEPA_SEK garsyno atpažinimo tikslumo rezultatai, gauti keičiant Gauso mišinių bei klasifikatorių medžių šakų skaičius trifoniniame režime:

Tyrimas atlikti keičiant numatytąsias parametrų reikšmes:

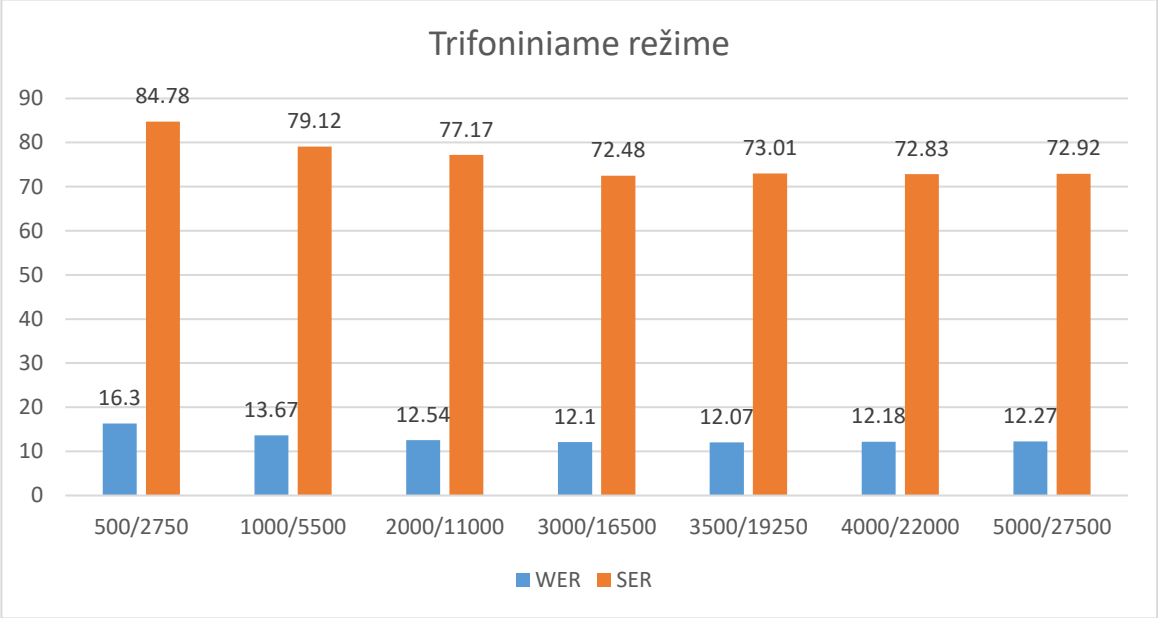
numLeavesTri1=2000

numGaussTri1=11000

Pradžioje buvo keičiamas *numLeavesTri1* ir ieškoma mažiausios atpažinimo klaidos perskaičiuojant parametrą *numGaussTri1* santykiu 1:5,5. Rezultatai – 17 lentelėje bei 33 paveiksle.

17 lentelė. LIEPA_SEK garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičiaus trifoniniame režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius						
	500/2750	1000/5500	2000/11000	3000/16500	3500/19250	4000/22000	5000/27500
WER	16,30	13,67	12,54	12,10	12,07	12,18	12,27
SER	84,78	79,12	77,17	72,48	73,01	72,83	72,92

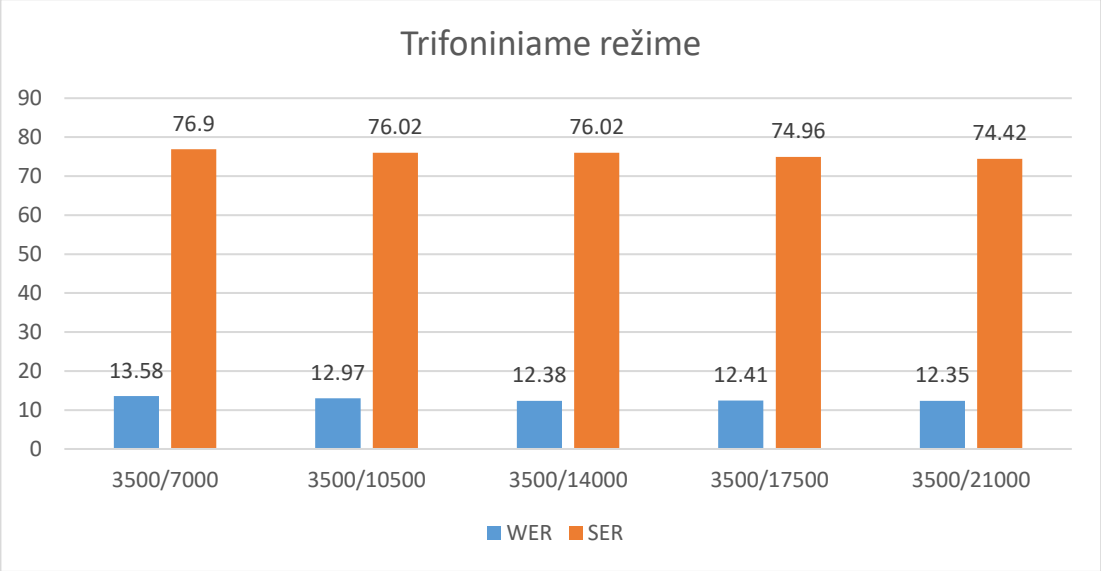


45 pav. Sekų garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičiaus trifoniniame režime

Po to buvo keičiamas parametras *numGaussTri1* išlaikant pastovų *numLeavesTri1* bet keičiant santykį 1:5,5. Rezultatai – 18 lentelėje bei 34 paveiksle.

18 lentelė. LIEPA_SEK garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus trifoniniame režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius				
	3500/7000	35000/10500	3500/14000	3500/17500	3500/21000
WER	13,58	12,97	12,38	12,41	12,35
SER	76,90	76,02	76,02	74,96	74,42



46 pav. Sekų garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus trifoniniame režime

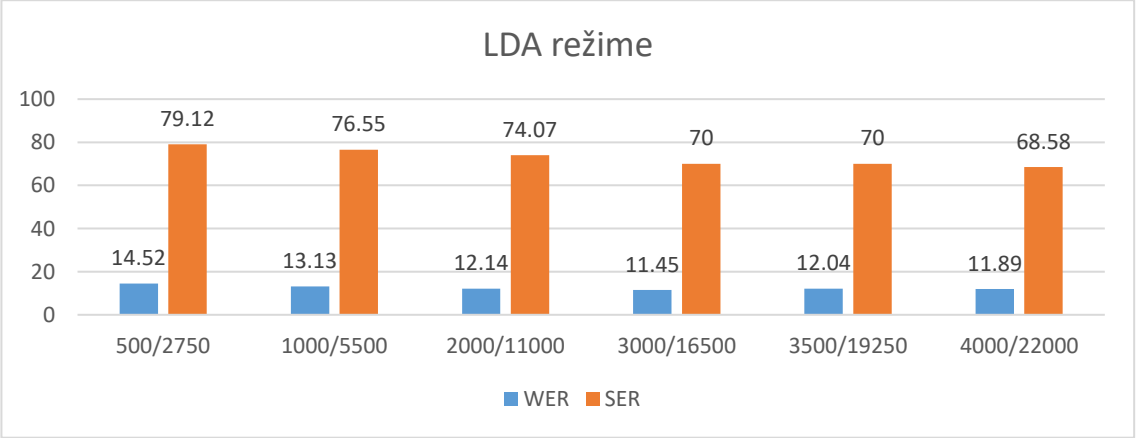
Geriausias rezultatas gautas, kai *numLeavesTri1=3500, numGaussTri1=19250.*

- LIEPA_SEK garsyno atpažinimo tikslumo rezultatai, gauti keičiant Gauso mišinių bei klasifikatorių medžių šakų skaičius LDA režime:

Tyrimas atlikti keičiant numatytąsias parametrų reikšmes:
numLeavesMLLT=2000
numGaussMLLT=11000
Pradžioje buvo keičiamas *numLeavesMLLT* ir ieškoma mažiausios atpažinimo klaidos perskaičiuojant parametrą *numGaussMLLT* santykiu 1:5,5. Rezultatai – 19 lentelėje ir 35 paveiksle.

19 lentelė. LIEPA_SEK garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičius LDA režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius					
	500/2750	1000/5500	2000/11000	3000/16500	3500/19250	4000/22000
WER	14,52	13,13	12,14	11,45	12,04	11,89
SER	79,12	76,55	74,07	70,00	70,00	68,58

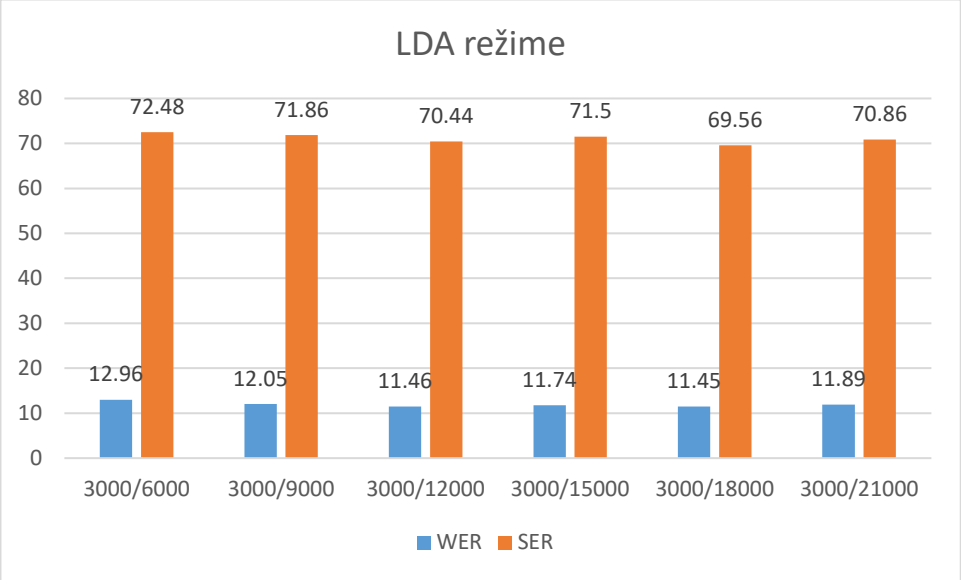


47 pav. Sekų garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičius LDA režime

Po to buvo keičiamas parametras *numGaussMLLT* išlaikant pastovų *numLeavesMLLT* bet keičiant santykį 1:5,5. Rezultatai – 20 lentelėje ir 36 paveiksle.

20 lentelė. LIEPA_SEK garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus LDA režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius					
	3000/6000	3000/9000	3000/12000	3000/15000	3000/18000	3000/21000
WER	12,96	12,05	11,46	11,74	11,45	11,89
SER	72,48	71,86	70,44	71,50	69,56	70,86



48 pav. Sekų garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus LDA režime

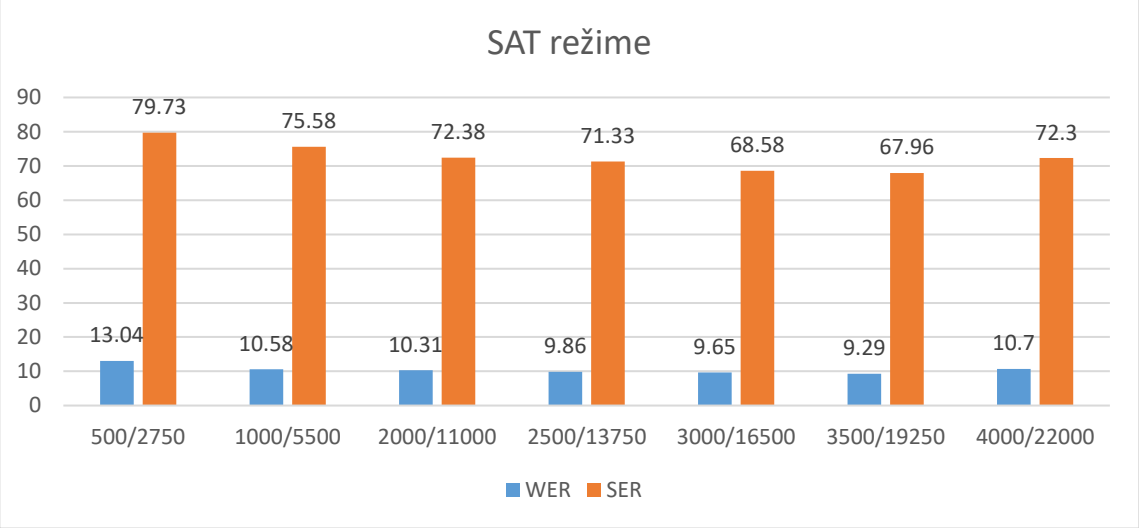
Geriausias rezultatas gautas, kai *numLeavesMLLT=3000, numGaussMLLT=16500.*

- LIEPA_SEK garsyno atpažinimo tikslumo rezultatai, gauti keičiant Gauso mišinių bei klasifikatorių medžių šakų skaičius SAT režime:

Tyrimas atlikti keičiant numatytąsias parametrų reikšmes:
numLeavesSAT=2000
numGaussSAT=11000
Pradžioje buvo keičiamas *numLeavesSAT* ir ieškoma mažiausios atpažinimo klaidos perskaičiuojant parametrą *numGaussSAT* santykiu 1:5,5. Rezultatai – 21 lentelėje bei 37 paveiksle.

21 lentelė. LIEPA_SEK garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičiaus SAT režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius						
	500/2750	1000/5500	2000/11000	2500/13750	3000/16500	3500/19250	4000/22000
WER	13,04	10,58	10,31	9,86	9,65	9,29	10,70
SER	79,73	75,58	72,38	71,33	68,58	67,96	72,30

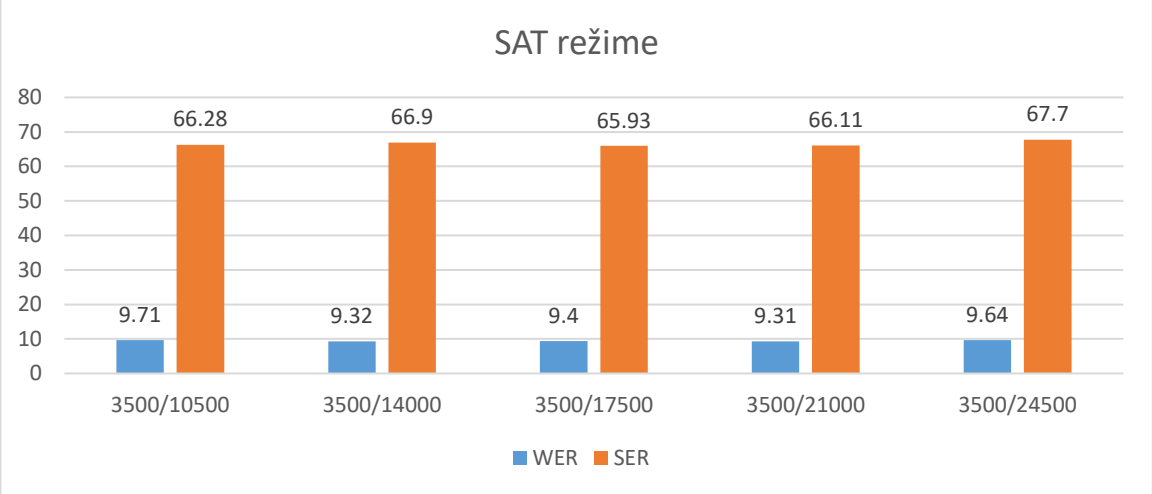


49 pav. Sekų garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičiaus SAT režime

Po to buvo keičiamas parametras *numGaussSAT* išlaikant pastovų *numLeavesSAT* bet keičiant santykį 1:5,5. Rezultatai – 22 lentelėje bei 38 paveiksle.

22 lentelė. LIEPA_SEK garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus SAT režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius				
	3500/10500	3500/14000	3500/17500	3500/21000	3500/24500
WER	9,71	9,32	9,40	9,31	9,64
SER	66,28	66,90	65,93	66,11	67,70



50 pav. Sekų garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus SAT režime

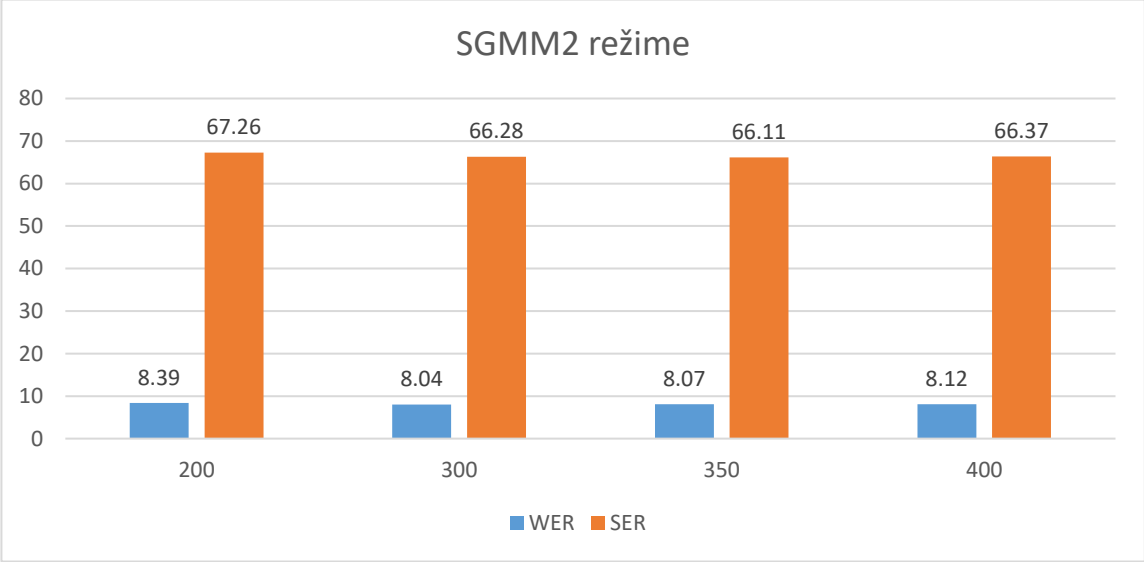
Geriausias rezultatas gautas, kai *numLeavesSAT=3500, numGaussSAT=19250.*

- LIEPA_SEK garsyno atpažinimo tikslumo rezultatai, gauti keičiant Gauso mišinių bei klasifikatorių medžių šakų skaičius SGMM2 režime:

Tyrimas atlikti keičiant numatytąsias parametrų reikšmes:
numGaussUBM=400
numLeavesSGMM=7000
numGaussSGMM=9000
Pradžioje buvo keičiamas *numGaussUBM* ir ieškoma mažiausios atpažinimo klaidos kitus parametrus išlaikant tuos pačius. Rezultatai – 23 lentelėje ir 39 paveiksle.

23 lentelė. LIEPA_SEK garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių UBM skaičiaus SGMM2 režime tyrimo rezultatai

Klaida	Gauso mišinių UBM skaičius (*numLeavesSGMM=7000, numGaussSGMM=9000)*			
	200	300	350	400
WER	8,39	8,04	8,07	8,12
SER	67,26	66,28	66,11	66,37

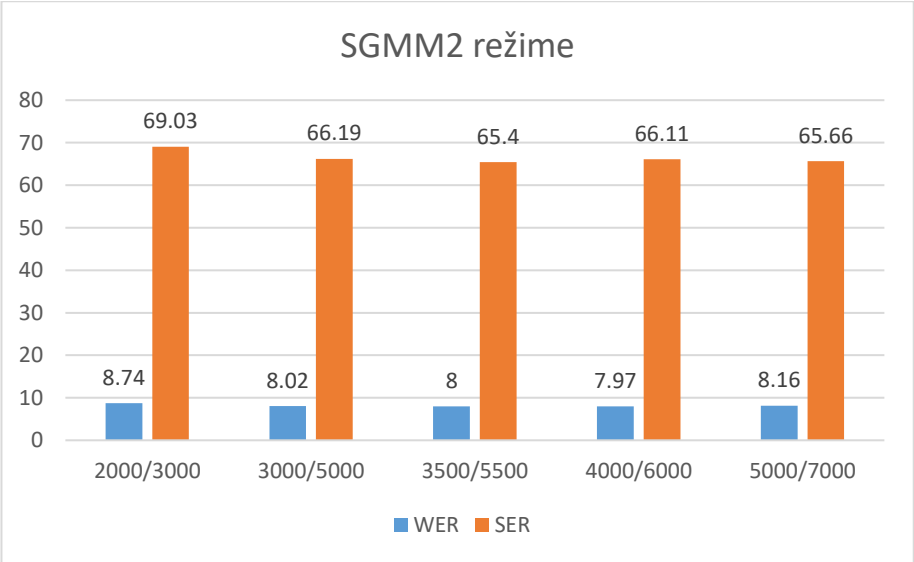


51 pav. Sekų garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių UBM skaičiaus SGMM2 režime

Po to buvo keičiamas *numLeavesSAT* ir ieškoma mažiausios atpažinimo klaidos išlaikant *numGaussSAT* 2000 didesniu už *numLeavesSAT*. Rezultatai – 24 lentelėje bei 40 paveiksle.

24 lentelė. LIEPA_SEK garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių SGMM bei klasifikatoriaus medžių šakų skaičiaus SGMM2 režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/Gauso mišinių SGMM skaičius (*numGaussUBM=300)*				
	2000/3000	3000/5000	3500/5500	4000/6000	5000/7000
WER	8,74	8,02	8,00	7,97	8,16
SER	69,03	66,19	65,40	66,11	65,66

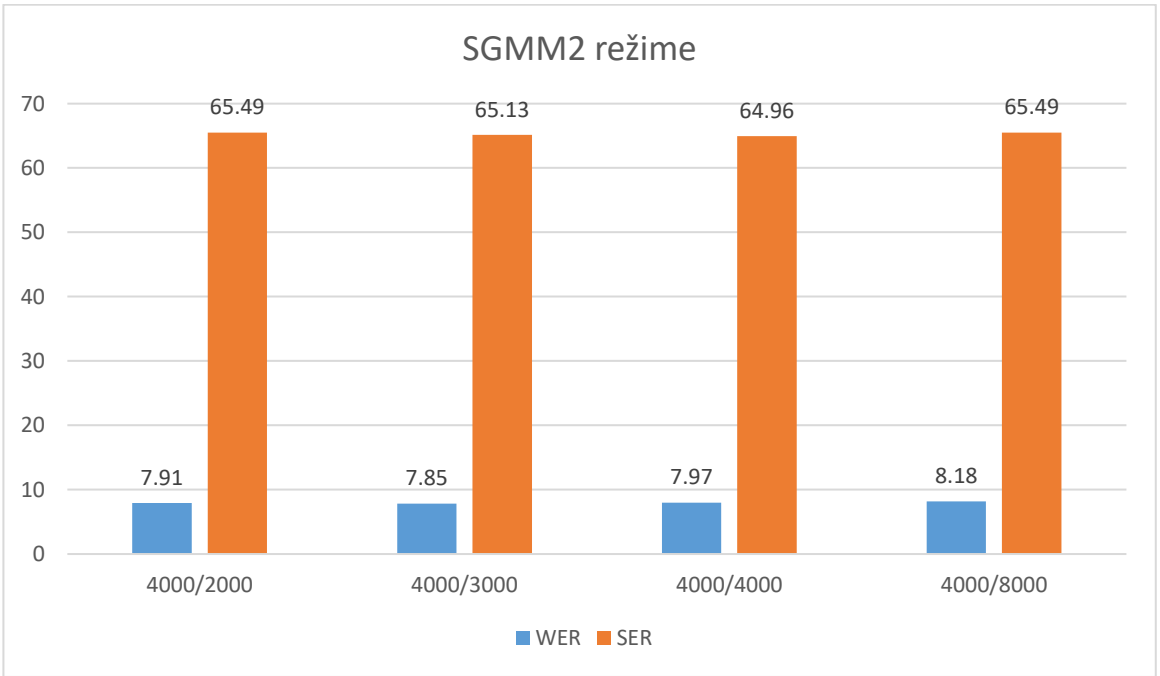


52 pav. Sekų garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių SGMM bei klasifikatoriaus medžių šakų skaičiaus SGMM2 režime

Po to buvo keičiamas parametras *numGaussSAT* išlaikant pastovų *numLeavesSAT*. Rezultatai – 25 lentelėje ir 41 paveiksle.

25 lentelė. LIEPA_SEK garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių SGMM skaičiaus SGMM2 režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių SGMM skaičius (*numGaussUBM=300)*			
	4000/2000	4000/3000	4000/4000	4000/8000
WER	7,91	7,85	7,97	8,18
SER	65,49	65,13	64,96	65,49



53 pav. Sekų garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių SGMM skaičiaus SGMM2 režime

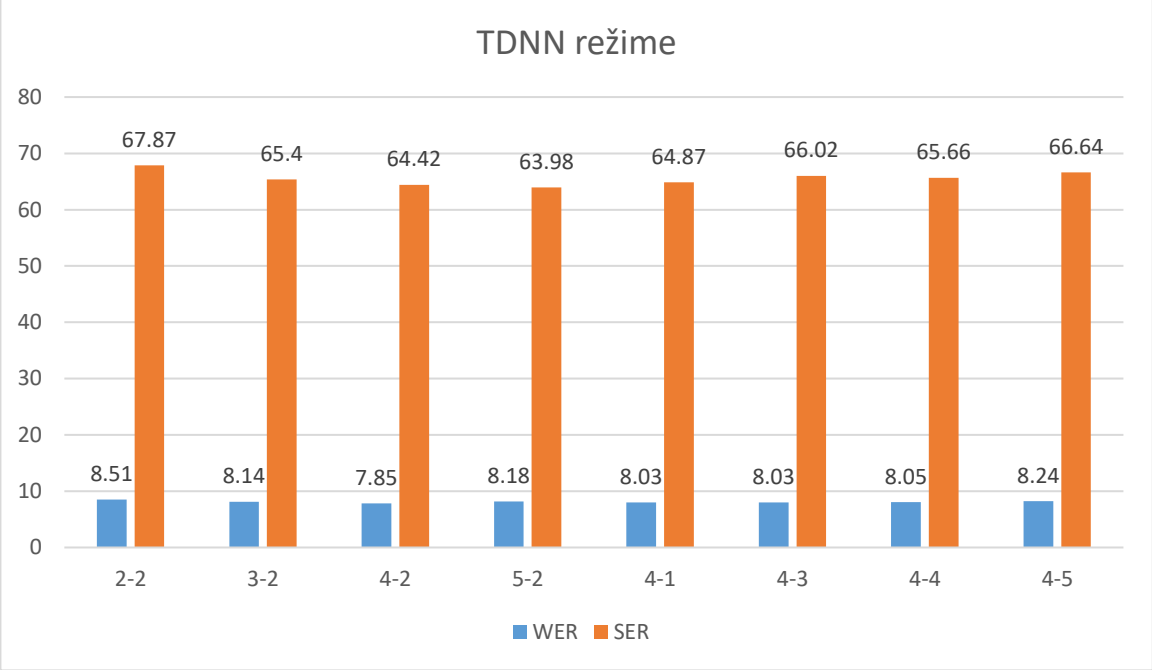
Geriausias rezultatas gautas, kai *numLeavesSAT=4000, numGaussSAT=3000.*

- LIEPA_SEK garsyno atpažinimo tikslumo rezultatai, gauti keičiant paslėptųjų sluoksnių bei neuronų skaičių juose naudojant TDNN metodo dvi modifikacijas:

TDNN *pnorm* metodo atveju keičiamas paslėptųjų sluoksnių skaičius nuo 2 iki 5, parametras p nuo 1 iki 5 ir *pnorm_input_dim* bei *pnorm_output_dim*. Numatytosios šių parametrų reikšmės: 2 sluoksniai, p=2, *pnorm_input_dim=2000, pnorm_output_dim=200.* Tyrime naudota 10 epochų, realizuojamų kaip 140 iteracijų. Esminiai rezultatai – 26 ir 27 lentelėse taip pat 42 ir 43 paveiksle.

26 lentelė. LIEPA_SEK garsyno atpažinimo tikslumo priklausomybės nuo sluoksnių skaičiaus ir p parametro naudojant TDNN metodo *pnorm* modifikaciją tyrimo rezultatai

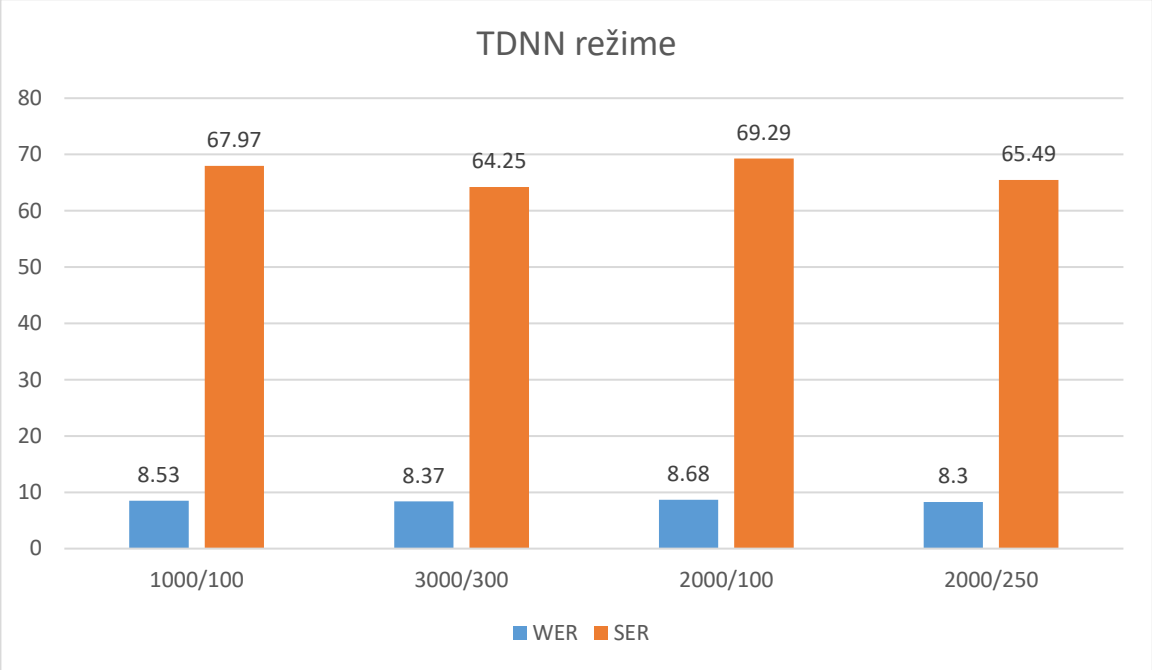
Klaida	Sluoksnių skaičius-p parametras (*input_dim=2000,output_dim=200)*							
	2-2	3-2	4-2	5-2	4-1	4-3	4-4	4-5
WER	8,51	8,14	7,85	8,18	8,03	8,03	8,05	8,24
SER	67,86	65,40	64,42	63,98	64,87	66,02	65,66	66,64



54 pav. Sekų garsyno atpažinimo tikslumo priklausomybės nuo sluoksnių skaičiaus ir p parametro naudojant TDNN metodo *pnorm* modifikaciją

27 lentelė. LIEPA_SEK garsyno atpažinimo tikslumo priklausomybės nuo parametrų *pnorm_input_dim* bei *pnorm_output_dim* naudojant TDNN metodo *pnorm* modifikaciją tyrimo rezultatai

Klaida	*input_dim/output_dim (4 sluoksniai, p=2)*			
	1000/100	3000/300	2000/100	2000/250
WER	8,53	8,37	8,68	8,30
SER	67,96	64,25	69,29	65,49

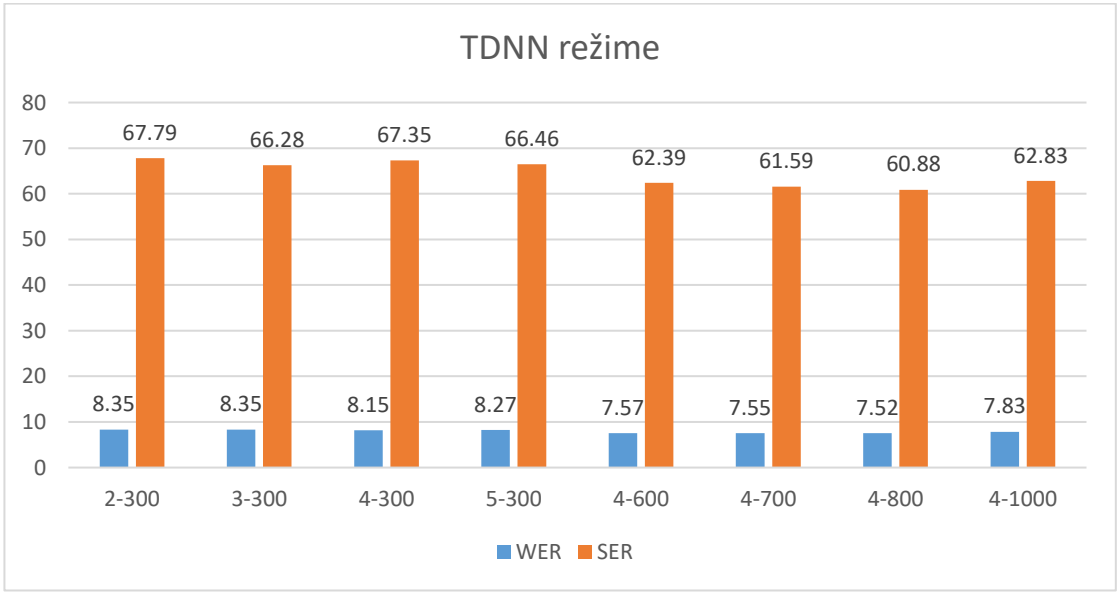


55 pav. Sekų garsyno atpažinimo tikslumo priklausomybės nuo parametrų *pnorm_input_dim* bei *pnorm_output_dim* naudojant TDNN metodo *pnorm* modifikaciją

TDNN *tanh* metodo atveju keičiamas paslėptųjų sluoksnių skaičius nuo 2 iki 4 bei *hidden_layer_dim*. Numatytosios šių parametrų reikšmės: 2 sluoksniai, *hidden_layer_dim=300*. Tyrime naudota 20 epochų, realizuojamų kaip 120 iteracijų. Rezultatai – 28 lentelėje ir 44 paveiksle.

28 lentelė. LIEPA_SEK garsyno atpažinimo tikslumo priklausomybės nuo sluoksnių skaičiaus ir sluoksnio dydžio naudojant TDNN metodo *tanh* modifikaciją tyrimo rezultatai

Klaida	Sluoksnių skaičius-parametras *hidden_layer_dim*							
	2-300	3-300	4-300	5-300	4-600	4-700	4-800	4-1000
WER	8,35	8,35	8,15	8,27	7,57	7,55	7,52	7,83
SER	67,79	66,28	67,35	66,46	62,39	61,59	60,88	62,83



56 pav. Sekų garsyno atpažinimo tikslumo priklausomybės nuo sluoksnių skaičiaus ir sluoksnio dydžio naudojant TDNN metodo *tanh* modifikaciją

3.3. Garsyno LIEPA sakinių atpažinimo rezultatai

3.3.1. Sakinių atpažinimo tyrimas

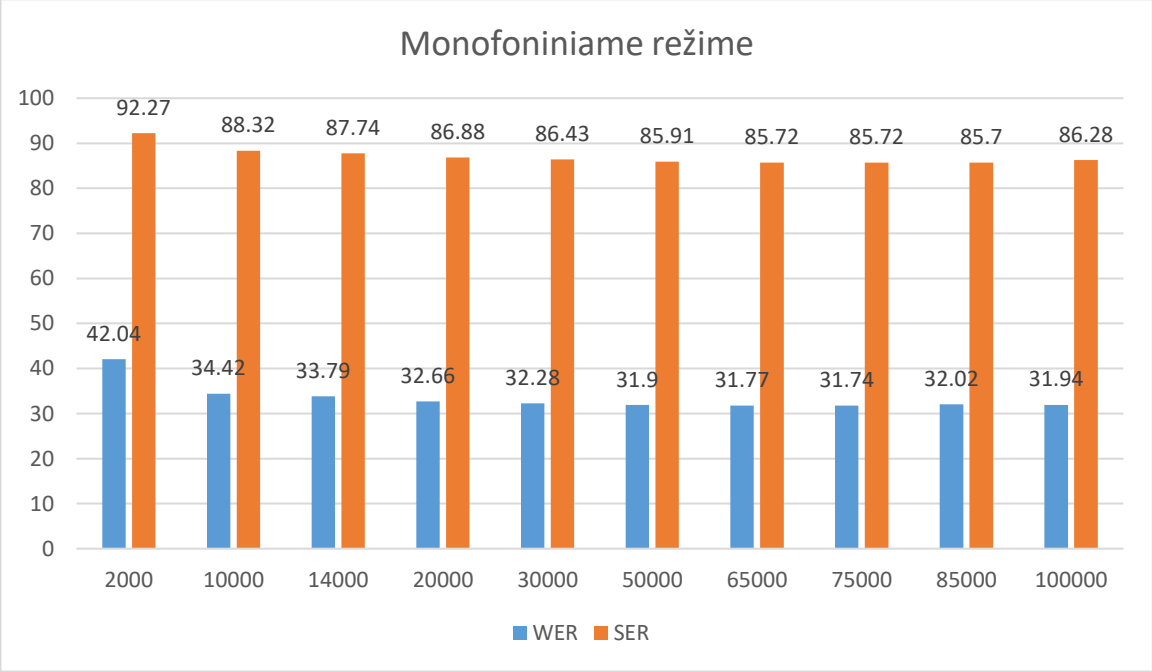
LIEPA_SAK garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatorių medžių šakų skaičių naudojant 5 atpažinimo metodus (monofoninį, trifoninį, LDA, SAT, SGMM2) bei nuo paslėptųjų sluoksnių bei neuronų skaičiaus juose naudojant TDNN metodo dvi modifikacijas.
Pereinant nuo vieno atpažinimo metodo prie kito paliekamos prieš tai naudotame metode surastos parametrų, duodančių mažiausią atpažinimo klaidą, reikšmės.

- LIEPA_SAK garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių skaičiaus monofoniniame režime:

Tyrimas atliktas keičiant parametrą *totgauss* nuo 2000 iki 100000. Rezultatai – 29 lentelėje bei 45 paveiksle.

29 lentelė. LIEPA_SAK garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių skaičiaus monofoniniame režime tyrimo rezultatai

Klaida	Gauso mišinių skaičius									
	2000	10000	14000	20000	30000	50000	65000	75000	85000	100000
WER	42,04	34,42	33,79	32,66	32,28	31,91	31,77	31,74	32,02	31,94
SER	92,27	88,32	87,74	86,88	86,43	85,91	85,72	85,72	85,70	86,28



57 pav. Sakinių garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių skaičiaus monofoniniame režime

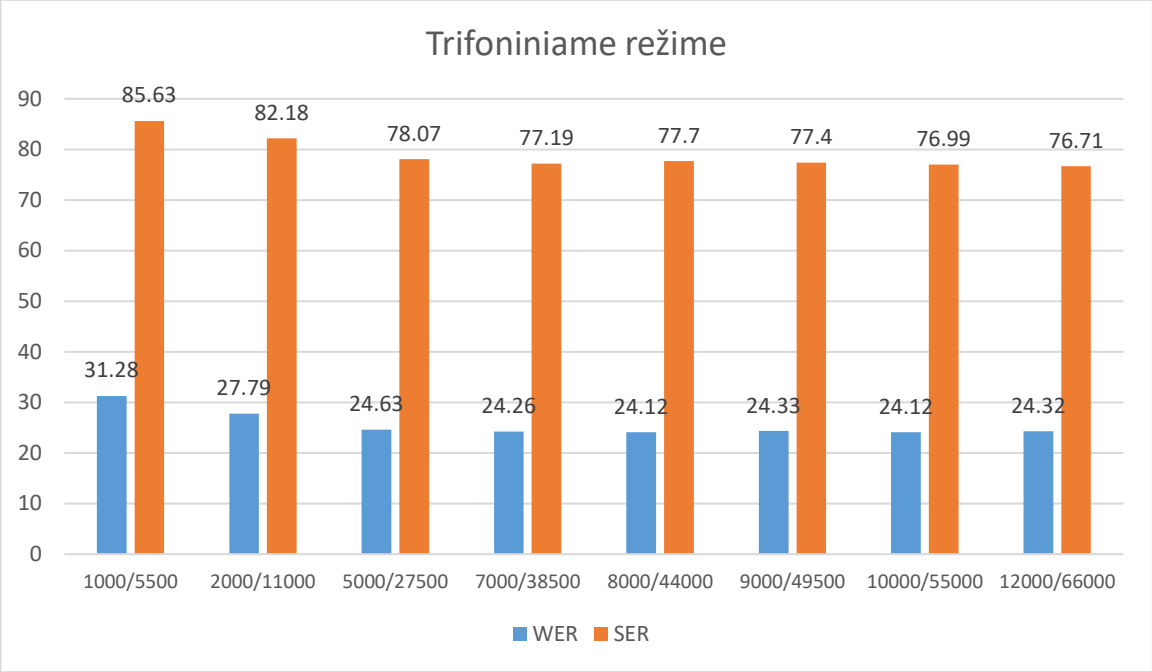
- LIEPA_SAK garsyno atpažinimo tikslumo rezultatai, gauti keičiant Gauso mišinių bei klasifikatorių medžių šakų skaičius trifoniniame režime:

Tyrimas atlikti keičiant numatytąsias parametrų reikšmes:
numLeavesTri1=2000
numGaussTri1=11000

Pradžioje buvo keičiamas *numLeavesTri1* ir ieškoma mažiausios atpažinimo klaidos perskaičiuojant parametrą *numGaussTri1* santykiu 1:5,5. Rezultatai – 30 lentelėje ir 46 paveiksle.

30 lentelė. LIEPA_SAK garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičiaus trifoniniame režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius							
	1000/5500	2000/11000	5000/27500	7000/38500	8000/44000	9000/49500	10000/55000	12000/66000
WER	31,28	27,79	24,63	24,26	24,12	24,33	24,12	24,32
SER	85,63	82,18	78,07	77,19	77,70	77,40	76,99	76,71

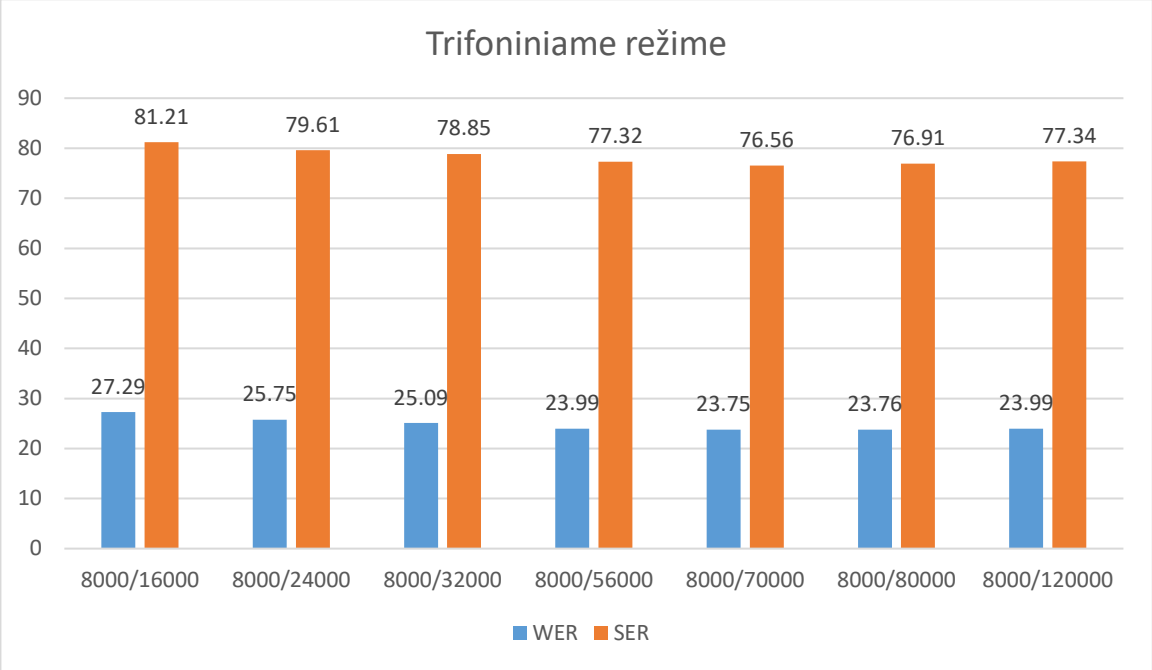


58 pav. Sakinių garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičiaus trifoniniame režime

Po to buvo keičiamas parametras *numGaussTri1* išlaikant pastovų *numLeavesTri1* bet keičiant santykį 1:5,5. Rezultatai – 31 lentelėje ir 47 paveiksle.

31 lentelė. LIEPA_SAK garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus trifoniniame režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius						
	8000/16000	8000/24000	8000/32000	8000/56000	8000/70000	8000/80000	8000/120000
WER	27,29	25,78	25,09	23,99	23,75	23,76	23,99
SER	81,21	79,61	78,85	77,32	76,56	76,91	77,34



59 pav. Sakinių garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus trifoniniame režime

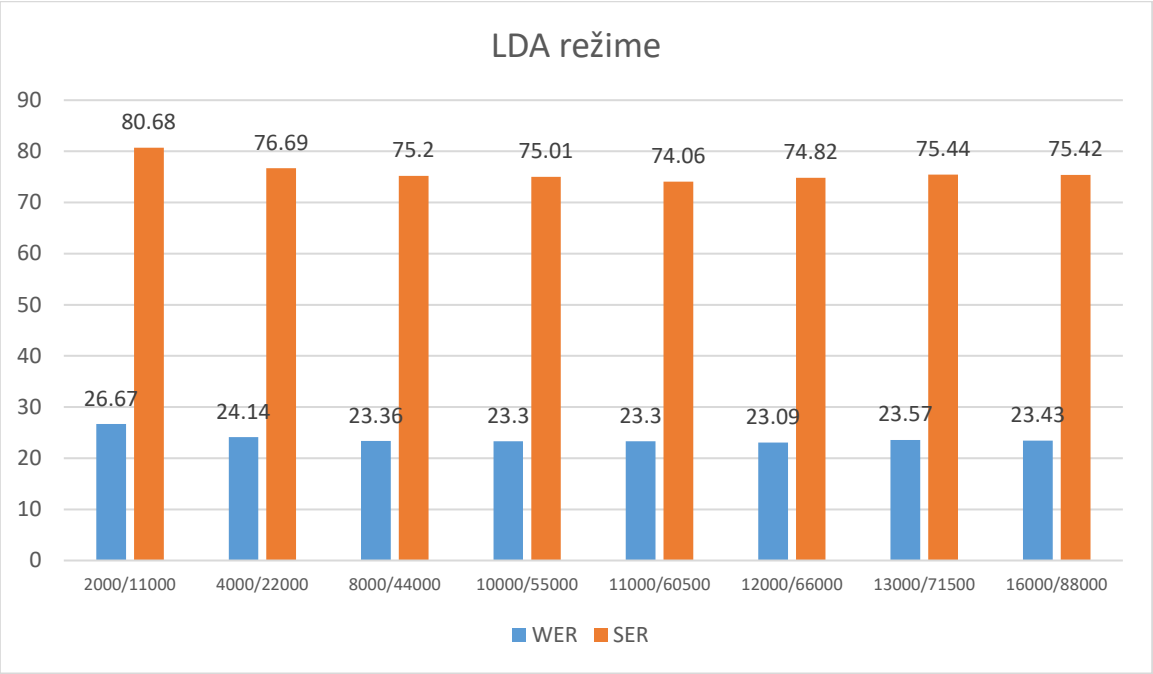
Geriausias rezultatas gautas, kai *numLeavesTri1=8000, numGaussTri1=70000.*

- LIEPA_SAK garsyno atpažinimo tikslumo rezultatai, gauti keičiant Gauso mišinių bei klasifikatorių medžių šakų skaičius LDA režime:

Tyrimas atlikti keičiant numatytąsias parametrų reikšmes:
numLeavesMLLT=2000
numGaussMLLT=11000
Pradžioje buvo keičiamas *numLeavesMLLT* ir ieškoma mažiausios atpažinimo klaidos perskaičiuojant parametrą *numGaussMLLT* santykiu 1:5,5. Rezultatai – 32 lentelėje bei 48 paveiksle.

32 lentelė. LIEPA_SAK garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičiaus LDA režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius							
	2000/11000	4000/22000	8000/44000	10000/55000	11000/60500	12000/66000	13000/71500	16000/88000
WER	26,67	24,14	23,36	23,31	23,30	23,09	23,57	23,43
SER	80,68	76,69	75,20	75,01	74,06	74,82	75,44	75,42

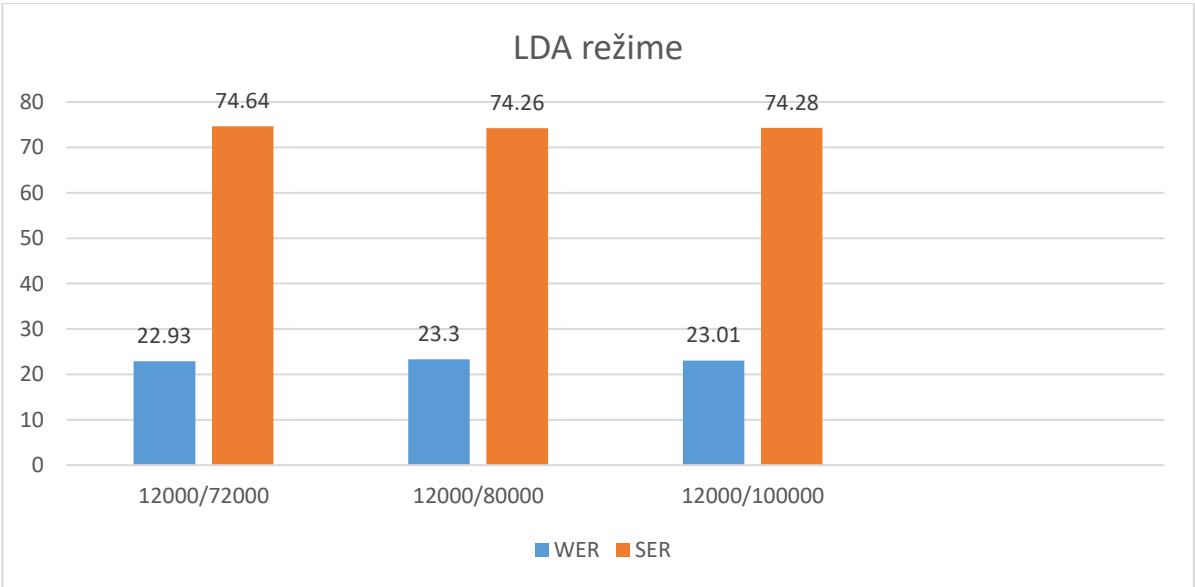


60 pav. Sakinių garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičiaus LDA režime

Po to buvo keičiamas parametras *numGaussMLLT* išlaikant pastovų *numLeavesMLLT* bet keičiant santykį 1:5,5. Rezultatai – 33 lentelėje ir 49 paveiksle.

33 lentelė. LIEPA_SAK garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus LDA režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius		
	12000/72000	12000/80000	12000/100000
WER	22,93	23,30	23,01
SER	74,64	74,26	74,28



61 pav. Sakinių garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus LDA režime

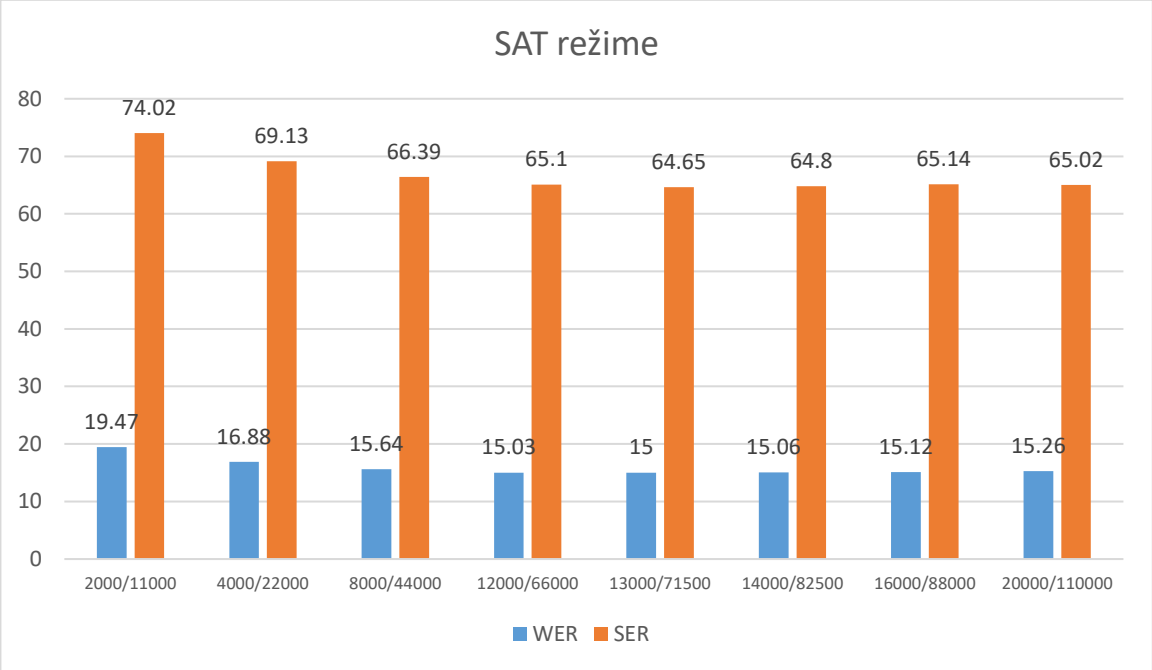
Geriausias rezultatas gautas, kai *numLeavesMLLT=12000, numGaussMLLT=72000.*

- LIEPA_SAK garsyno atpažinimo tikslumo rezultatai, gauti keičiant Gauso mišinių bei klasifikatorių medžių šakų skaičius SAT režime:

Tyrimas atlikti keičiant numatytąsias parametrų reikšmes:
numLeavesSAT=2000
numGaussSAT=11000
Pradžioje buvo keičiamas *numLeavesSAT* ir ieškoma mažiausios atpažinimo klaidos perskaičiuojant parametrą *numGaussSAT* santykiu 1:5,5. Rezultatai – 34 lentelėje ir 50 paveiksle.

34 lentelė. LIEPA_SAK garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičiaus SAT režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius							
	2000/11000	4000/22000	8000/44000	12000/66000	13000/71500	14000/82500	16000/88000	20000/110000
WER	19,47	16,88	15,64	15,03	15,00	15,06	15,12	15,26
SER	74,02	69,13	66,39	65,10	64,65	64,80	65,14	65,02

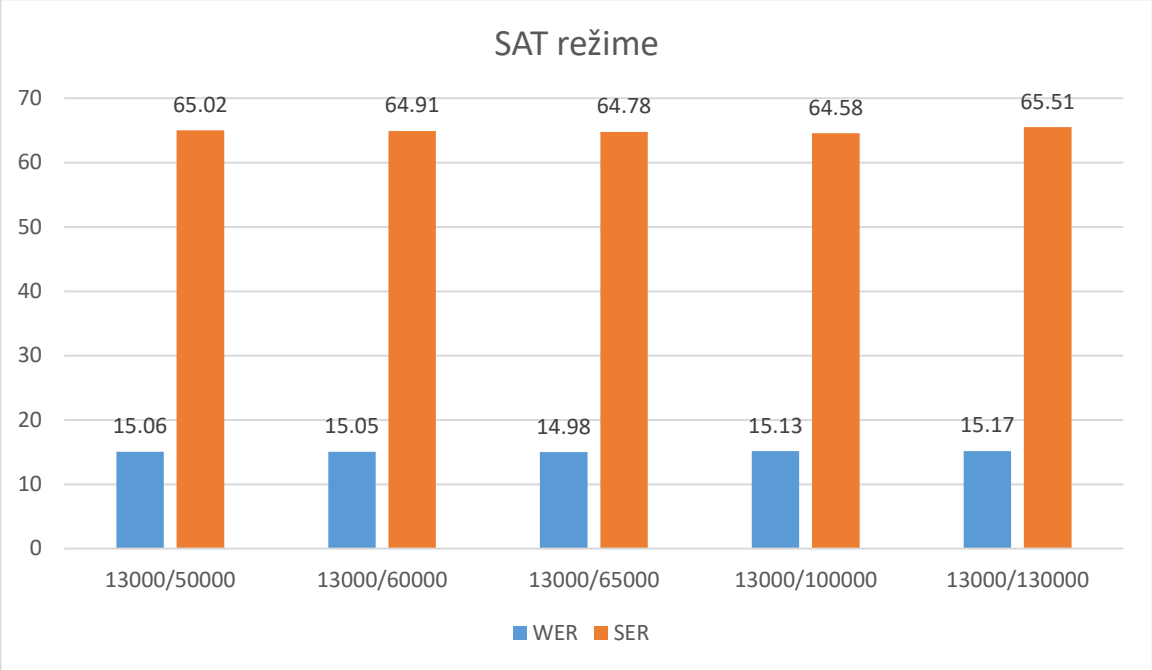


62 pav. Sakinių garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių bei klasifikatoriaus medžių šakų skaičiaus SAT režime

Po to buvo keičiamas parametras *numGaussSAT* išlaikant pastovų *numLeavesSAT* bet keičiant santykį 1:5,5. Rezultatai –35 lentelėje bei 51 paveiksle.

35 lentelė. LIEPA_SAK garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus SAT režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių skaičius				
	13000/50000	13000/60000	13000/65000	13000/100000	13000/130000
WER	15,06	15,05	14,98	15,13	15,17
SER	65,02	64,91	64,78	64,58	65,51



63 pav. Sakinių garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių skaičiaus SAT režime

Geriausias rezultatas gautas, kai *numLeavesSAT=13000, numGaussSAT=65000.*

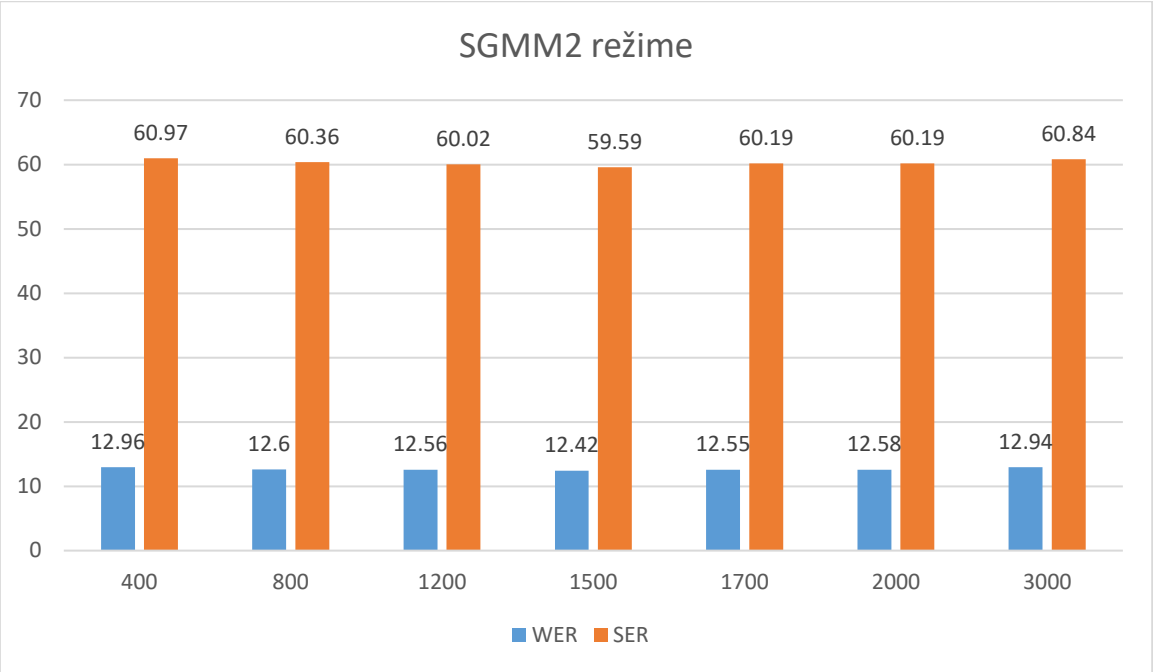
- LIEPA_SAK garsyno atpažinimo tikslumo rezultatai, gauti keičiant Gauso mišinių bei klasifikatorių medžių šakų skaičius SGMM2 režime:

Tyrimas atlikti keičiant numatytąsias parametrų reikšmes:
numGaussUBM=400
numLeavesSGMM=7000
numGaussSGMM=9000

Pradžioje buvo keičiamas *numGaussUBM* ir ieškoma mažiausios atpažinimo klaidos kitus parametrus išlaikant tuos pačius. Rezultatai – 36 lentelėje bei 52 paveiksle.

36 lentelė. LIEPA_SAK garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių UBM skaičiaus SGMM2 režime tyrimo rezultatai

Klaida	Gauso mišinių UBM skaičius (*numLeavesSGMM=7000, numGaussSGMM=9000)*						
	400	800	1200	1500	1700	2000	3000
WER	12,96	12,60	12,56	12,42	12,55	12,58	12,94
SER	60,97	60,36	60,02	59,59	60,19	60,19	60,84

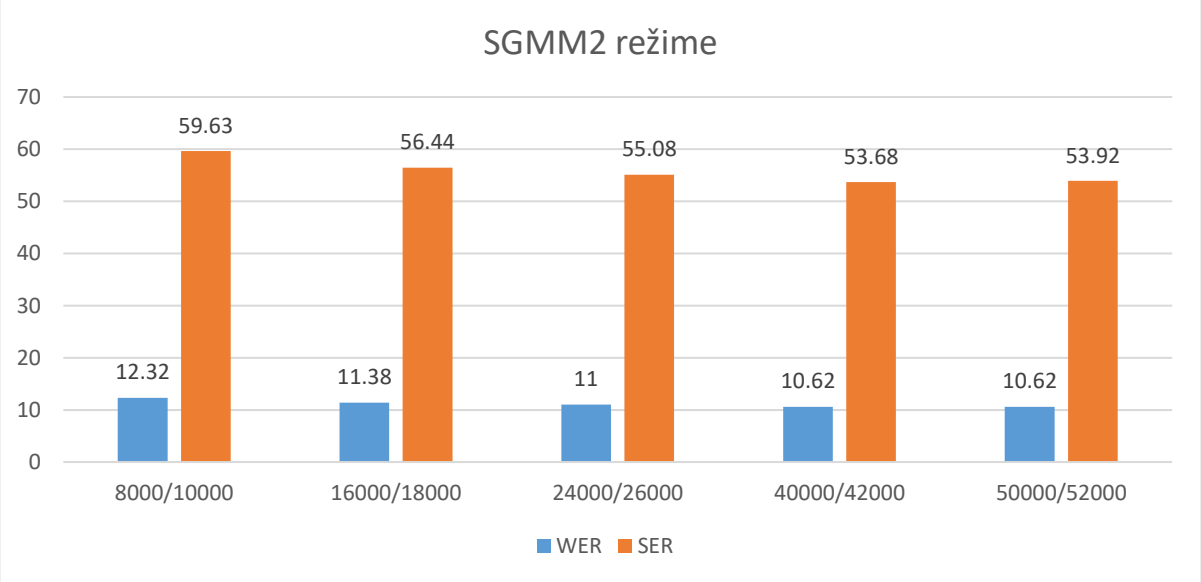


64 pav. Sakinių garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių UBM skaičiaus SGMM2 režime

Po to buvo keičiamas *numLeavesSAT* ir ieškoma mažiausios atpažinimo klaidos išlaikant *numGaussSAT* 2000 didesniu už *numLeavesSAT*. Rezultatai – 37 lentelėje ir 53 paveiksle.

37 lentelė. LIEPA_SAK garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių SGMM bei klasifikatoriaus medžių šakų skaičiaus SGMM2 režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/Gauso mišinių SGMM skaičius (*numGaussUBM=1500)*				
	8000/10000	16000/18000	24000/26000	40000/42000	50000/52000
WER	12,32	11,38	11,00	10,62	10,62
SER	59,63	56,44	55,08	53,68	53,92

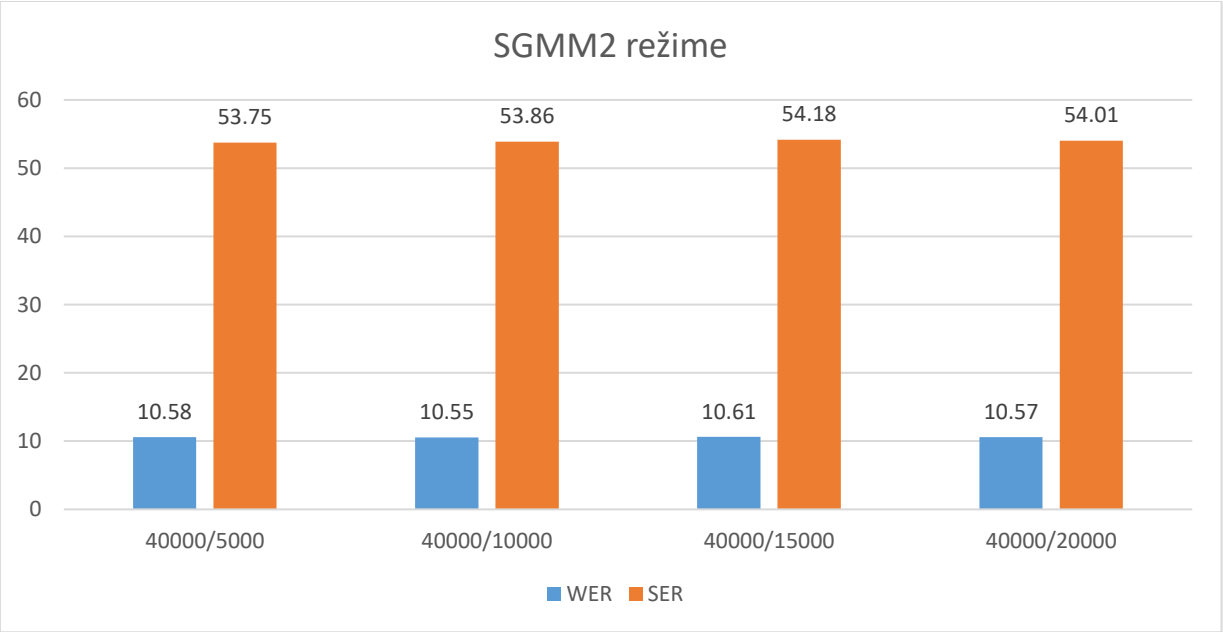


65 pav. Sakinių garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių SGMM bei klasifikatoriaus medžių šakų skaičiaus SGMM2 režime

Po to buvo keičiamas parametras *numGaussSAT* išlaikant pastovų *numLeavesSAT*. Rezultatai – 38 lentelėje bei 54 paveiksle.

38 lentelė. LIEPA_SAK garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių SGMM skaičiaus SGMM2 režime tyrimo rezultatai

Klaida	Medžių šakų skaičius/ Gauso mišinių SGMM skaičius (*numGaussUBM=1500)*			
	40000/5000	40000/10000	40000/15000	40000/20000
WER	10,58	10,55	10,61	10,57
SER	53,75	53,86	54,18	54,01



66 pav. Sakinių garsyno atpažinimo tikslumo priklausomybės nuo santykio tarp klasifikatoriaus medžių šakų bei Gauso mišinių SGMM skaičiaus SGMM2 režime

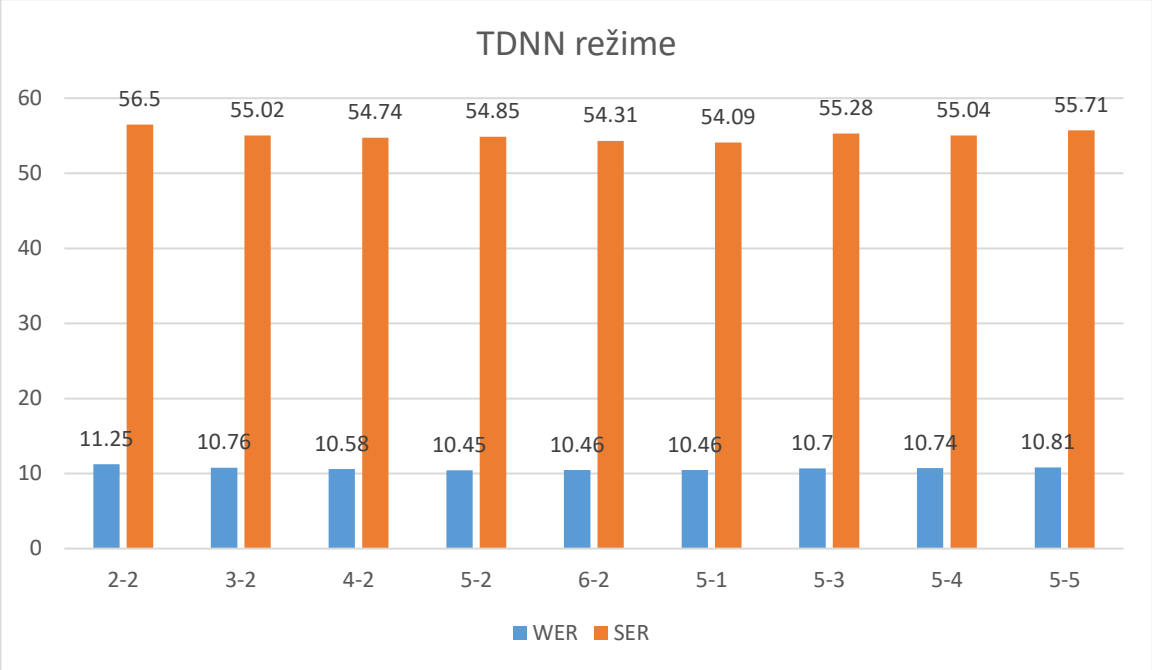
Geriausias rezultatas gautas, kai *numLeavesSAT=40000, numGaussSAT=10000.*

- LIEPA_SAK garsyno atpažinimo tikslumo rezultatai, gauti keičiant paslėptųjų sluoksnių bei neuronų skaičių juose naudojant TDNN metodo dvi modifikacijas:

TDNN *pnorm* metodo atveju keičiamas paslėptųjų sluoksnių skaičius nuo 2 iki 6, parametras p nuo 1 iki 5 ir *pnorm_input_dim* bei *pnorm_output_dim*. Numatytosios šių parametrų reikšmės: 2 sluoksniai, p=2, *pnorm_input_dim=2000, pnorm_output_dim=200.* Tyrime naudota 10 epochų, realizuojamų kaip 100 iteracijų. Esminiai rezultatai – 39 ir 40 lentelėse taip pat 55 ir 56 paveiksle.

39 lentelė. LIEPA_SAK garsyno atpažinimo tikslumo priklausomybės nuo sluoksnių skaičiaus ir p parametro naudojant TDNN metodo *pnorm* modifikaciją tyrimo rezultatai

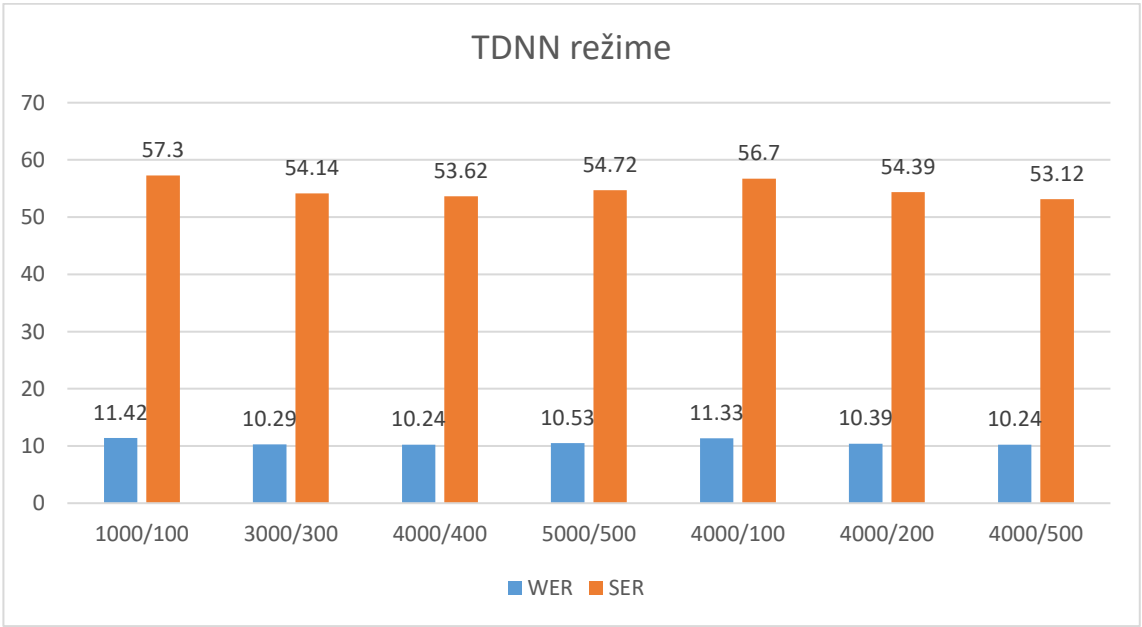
Klaida	Sluoksnių skaičius-p parametras (*input_dim=2000,output_dim=200)*								
	2-2	3-2	4-2	5-2	6-2	5-1	5-3	5-4	5-5
WER	11,25	10,76	10,58	10,45	10,46	10,46	10,70	10,74	10,81
SER	56,57	55,02	54,74	54,85	54,31	54,09	55,28	55,04	55,71



67 pav. Sakinių garsyno atpažinimo tikslumo priklausomybės nuo sluoksnių skaičiaus ir p parametro naudojant TDNN metodo *pnorm* modifikaciją

40 lentelė. LIEPA_SAK garsyno atpažinimo tikslumo priklausomybės nuo parametrų *pnorm_input_dim* bei *pnorm_output_dim* naudojant TDNN metodo *pnorm* modifikaciją tyrimo rezultatai

Klaida	*input_dim/output_dim (5 sluoksniai, p=2)*						
	1000/100	3000/300	4000/400	5000/500	4000/100	4000/200	4000/500
WER	11,42	10,29	10,24	10,53	11,33	10,39	10,24
SER	57,30	54,14	53,62	54,72	56,70	54,39	53,12

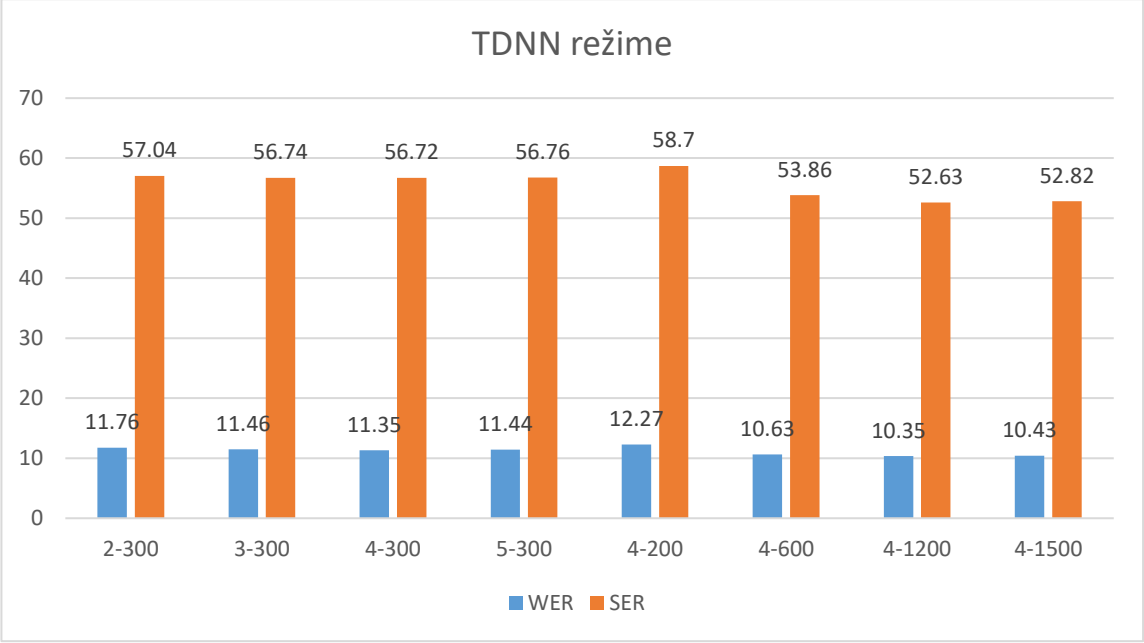


68 pav. Sakinių garsyno atpažinimo tikslumo priklausomybės nuo parametrų *pnorm_input_dim* bei *pnorm_output_dim* naudojant TDNN metodo *pnorm* modifikaciją

TDNN *tanh* metodo atveju keičiamas paslėptųjų sluoksnių skaičius nuo 2 iki 5 bei *hidden_layer_dim*. Numatytosios šių parametrų reikšmės: 2 sluoksniai, *hidden_layer_dim=300*. Tyrime naudota 20 epochų, realizuojamų kaip 60 iteracijų. Rezultatai – 41 lentelėje ir 57 paveiksle.

41 lentelė. LIEPA_SAK garsyno atpažinimo tikslumo priklausomybės nuo sluoksnių skaičiaus ir sluoksnio dydžio naudojant TDNN metodo *tanh* modifikaciją tyrimo rezultatai

Klaida	Sluoksnių skaičius-parametras *hidden_layer_dim*							
	2-300	3-300	4-300	5-300	4-200	4-600	4-1200	4-1500
WER	11,76	11,46	11,35	11,44	12,27	10,63	10,35	10,43
SER	57,04	56,74	56,72	56,76	58,70	53,86	52,63	52,82



69 pav. Sakinių garsyno atpažinimo tikslumo priklausomybės nuo sluoksnių skaičiaus ir sluoksnio dydžio naudojant TDNN metodo *tanh* modifikaciją

Išvados ir Rezultatai

1. Ištirtas atvirojo kodo lietuviškas garsynas LIEPA.
2. Geriausias rezultatas izoliuotų žodžių atpažinimo tyrime buvo gautas naudojant žodžių garsyno atpažinimo tikslumo priklausomybės nuo Gauso mišinių SGMM bei klasifikatoriaus medžių šakų skaičiaus SGMM2 režime. Žodžių klaidų dažnis – 2,78 proc. tiriamajame garsyne.
3. Išanalizavus sekų tyrimą, geriausi rezultatai žodžių klaidų dažnyje gauti naudojant TDNN pnorm 7,52 proc., tačiau sakinių klaidų dažnis gautas 68,22 proc. Tai geriausias rezultatas abiem metodams. Jis pasiektas naudojantis laiko delsos neuroninius tinklu.
4. Sakinių garsyno atpažinimo tikslumas priklauso nuo parametrų pnorm_input_dim ir pnorm_output_dim naudojant TDNN metodo pnorm modifikaciją. Ištyrus garsyną gauta, kad žodžių klaidų dažnis – 10,24 proc. Analizuojant sakinių tyrimą taip pat mažiausias sakinių klaidos tikimybės dažnis gautas nuo tikslumo priklausomybės bei parametrų pnorm_input_dim ir pnorm_output_dim naudojant TDNN metodo pnorm modifikaciją.

Literatūros sąrašas

1. Akella Amarendra Babu, Yellasiri Ramadevi and Akepogu Ananda Rao. (2014). Unsupervised Adaptation of ASR Systems Using hybrid HMM/ VQ model. *Proceedings of the International MultiConference of Engineers and Computer Scientists 2014 Vol I*.
2. Alexander Waibel, Tashiyuki Hanazawa, Geoffrey Hinton, Kiyohito Shikano and Kevin J. Lang. (1989). Phoneme Recognition Using Time-Delay Neural Networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 328-339.
3. Axelrod, A., Resnik, P., He, X. and Ostendorf, M. (2015). Data selection with fewer words. *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 58-65.
4. B.H. Juang and L.R. Rabiner. (2005). Automatic speech recognition–a brief history of the technology development. 1-24.
5. Baker, J. K. (1975). The Dragon System-An Overview. *IEEE Trans. on Acoustics Speech Signal Processing*, 24-29.
6. Bilmes, J. (2006). What HMMs can do. *IEICE Trans. Inf. Syst*, 869-891.
7. Bishop, C. M. (2006). Pattern Recognition and Machine Learning.
8. D. Jurafsky and J. H. Martin. (2009). Speech and Language Processing - An Introduction to Natural. *Prentice Hall*.
9. D. Poveya, L. Burgetb, M. Agarwalc, P. Akyazid and F. Kaie. (2010). The subspace Gaussian mixture model – a structured. 1-54.
10. D. Sipavičius and R. Maskeliūnas. (2016). 'Google' Lithuanian speech recognition efficiency evaluation research. *Inf. Softw. Technol. Proc. 22nd Int. Conf*, 602-616.
11. D.Yu and L. Deng. (2015). Automatic Speech Recognition - A Deep Learning Approach. *Springer-Verlag*.
12. Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 179-188.
13. Forsberg, M. (2003). Why Is Speech Recognition Difficult? *Chalmers University of Technology, Citeseer*.
14. G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Magazine*, 82-97.
15. Gales, M. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language*, 75-98.
16. H. Sakoe and S. Chiba. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 43-49.
17. *History of Speech Recognition*. (2015). Nuskaityta iš http://www.dragon-medical-transcription.com/.
18. Huang, Acero and Hon. (2001). Spoken Language Processing Guide to Algorithms and System Development.
19. Issam Bazzi and James Glass. (2002). A MULTI-CLASS APPROACH FOR MODELLING OUT-OF-VOCABULARYWORDS. *Proceedings of the 7th International Conference on Spoken Language Processing*, 1613-1616.
20. J. Li, L. Deng, R. H.Umbach and Yifan Gong. (2015). Robust Automatic Speech Recognition: A Bridge to. *Academic Press*.

21. Kazlauskiene, A., Raškinis, G. (2013). Principles of development of the intonational annotated spoken corpus. *Žmogus ir žodis: didaktine lingvistika*, 101-110.
22. Khaled Abdalgadar and Andrew Skabar. (2012). Unsupervised similaritybased word sense disambiguation using context vectors and sentential word importance. *ACM Transactions on Speech and Language Processing*.
23. L. Deng and X. Li. (2013). Machine learning Paradigms for Speech Recognition. *IEEE Transactions on Audio, Speech, and Language processing*, 1060-1089.
24. L. Rabiner, B. Juang and B Yegnanarayana. (2010). Fundamentals of Speech Recognition. *Prentice Hall, Englewood Cliff*.
25. Laurinčiūkaitė S., Telsknys L., Kasparaitis P., Kliūkienė R. and Paukštytė V. (2018). Lithuanian Speech Corpus Liepa for Development of Human-Computer Interfaces Working in Voice Recognition and Synthesis Mode. 487-496.
26. Laurinčiukaite, S., Filipovič, M. and Telksnys, L. (2009). Lithuanian continuous speech corpus LRN 1: an improvement. 203-207.
27. M. A. Anusuya and S. K. Katti. (2009). Speech Recognition by Machine:A Review. *International Journal of Computer Science and Information Security*.
28. McLachlan, G. J. (2004). Discriminant Analysis and Statistical Pattern Recognition. *Wiley Interscience*.
29. Morgan, N. (2012). Deep and Wide: Multiple Layers in Automatic Speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*.
30. O. Tsubasa, M. Shigeki and X. Lu. (2014). Speaker Adaptive Training using Deep Neural Networks. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
31. Plátek, O. (2014). Automatic speech recognition using Kaldi.
32. R. Lawrence and B.H. Juang. (1993). Fundamentals of Speech Recognition. *Prentice-Hall*.
33. Roe, J. G. Wilpon and D. B. Roe. (1992). AT&T Telephone Network Applications of Speech Recognition. *COST232 Workshop*.
34. Samson, J.S., Besacier, L., Lecouteux, B. and Tan, T. (2014). Using closely-related language to build an ASR for a very under-resourced language. *Proceedings of Co-Ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA)*, 1-5.
35. Takahashi, N.,Naghibi, T., Pfister, B. (2016). Automatic pronunciation generation by utilizing a semi-supervised deep neural networks. *Proceedings of the 17th Interspeech 2016*.
36. Tao Tao, Su-Youn Yoon, Andrew Fister, Richard Sproat and ChengXiang Zhai. (2006). "Unsupervised Named Entity Transliteration Using Temporal and Phonetic Correlation. *EMNLP 2006*, 250-257.
37. V. Peddinti, D. Povey and S. Khudanpur. (2015). A time delay neural network architecture for efficient modeling of long. 1-5.
38. Waibel, A. (1989). Modular construction of time-delay neural networks. *Neural computation*, 39-46.
39. Wetcher-Hendricks, D. (2011). Analyzing Quantitative Data: An Introduction for Social Researchers. 288.

40. Xinguang Li, Jiahua Chen, Zhenjiang Li. (2013). English Sentence Recognition Based on HMM and Clustering. *American Journal of Computational Mathematics*, 37-42.
41. Y. Miao, H. Zhang and F. Metze. (2015). Speaker Adaptive Training of Deep Neural Network Acoustic Models Using I-Vectors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 1938-1949.