TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Kunnar Kukk 204103IABM

# KÕNELDAVA KEELE TUVASTAMINE AKTSENDIGA KÕNEST

Magistritöö

Juhendaja: Tanel Alumäe
PhD

Tallinn 2023

# Autori deklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Kunnar Kukk

13.04.2023

# Annotatsioon

## Kõneldava keele tuvastamine aktsendiga kõnest

Lõputöö eesmärk on leida viise, kuidas parandada keelteüleselt kõneldava keele identifitseerimist, kui kõneleja räägib keelt võõrkeelse aktsendiga.

Kõnetuvastusrakendustes on kõneldava keele täpne tuvastamine kriitiline esimene samm edasiseks kõnetöötluseks. Võrreldes emakeelse kõnega põhjustab aktsendiga kõnest kõneldava keele tuvastamine seni teadaolevalt kolm korda enam vigu, mistõttu on oluline vähendada kõneldava keele tuvastamise mudelite veamäära.

Töö esmane eesmärk on analüüsida seniste akustikal põhinevate tippmudelite õigsust aktsendiga kõne puhul, analüüsides kõneldava keele identifitseerimise mudeleid aktsenti sisaldavates korpustes ning võrreldes tulemusi sama keelt emakeelena kõnelevate andmestike peal.

Töö teiseks eesmärgiks on uurida, kuidas aktsendiga kõne puhul parandada kõneldava keele tuvastamise täpsust. Mudelid treenitakse VoxLingua107 andmestiku peal ning tulemusi võrreldakse 6 erineva aktsendiga kõne ning emakeelse kõnekorpuse peal.

Töö näitab, et kõnetuvastushüpoteeside lisamine ning akustilise Wav2vec 2.0 arhitektuuriga tippmudeli koos-kasutamine parandab kõneledava keele tuvastust aktsendiga kõnest, vähendades õigsuse suhtelist veamäära 35-63%, andmata samal ajal järele õigsuse määras ka emakeelse kõne puhul ning lisamata täiendavaid andmeid.

Töö põhineb autori tehtud uurimistööl ning kahasse juhendajaga avaldatud samateemalisel artiklil kõnetehnoloogiakonverentsil Interspeech 2022.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 31 leheküljel, 7 peatükki, 9 joonist ja 3 tabelit.

# Abstract

## Spoken Language Identification from Accented Speech

The goal of the thesis is to find ways to improve current state-of-the-art of spoken language identification when a speaker speaks a language with foreign accent.

In speech recognition applications, accurate identification of spoken language is a critical first step for further speech processing. Compared to native language speech, the identification of the language spoken by speakers with an accent is known to cause three times as many errors compared to the native speakers, thus it is important to reduce the error rate of the spoken language identification models.

The primary objective of the work is to analyse the accuracy of the current acoustic state-of-the-art models for accented speech by analysing identification of the spoken language both in the accent-containing corporas and comparing the results on the same language native speech data sets.

The second aim of the work is to find ways how to improve accuracy of spoken language identification from accented speech. The models are trained using the multi-lingual VoxLingua107 dataset, and the results are compared on 6 different corporas of accented and native speech.

The experiments show that adding text hypotheses and using a Wav2vec2 architecture acoustic state-of-the-art model the accuracy of spoken language identification from accented speech, reduces the relative error rate from 35 to 63%, while not decreasing the accuracy rate of native speech or without adding any additional data.

The work is based on research carried out by the author and an article published on the same topic with supervisor. The article was published at the Interspeech 2022 speech technology conference.

The thesis is written in Estonian and contains 31 pages of text, 7 chapters, 9 figures and 3 tables.

# Lühendite ja terminite loetelu

| | |
|---|---|
| ASR | Kõnetuvastus, *Automatic Speech Recognition* |
| CMU | Carnegie Mellon University |
| CNN | Konvolutsiooniline närvivõrk, *Convolutional Neural Network* |
| CSLU | *Center for Spoken Language Understanding* |
| EFAC | Aktsendikorpus, *Estonian Foreign Accent Corpus* |
| GPU | Graafikaprotsessor, *Graphical Processing Unit* |
| HPC | Kõrgarvutusklaster, *High Performance Cluster* |
| LDA | Lineaardiskriminantanalüüs, *Linear Discriminant Analysis* |
| LID | Kõneldava keele tuvastus, *Language Identification* |
| LSTM | Pikk lühiajaline mälu, *Long Short-term Memory* |
| NB | Naive Bayes'i meetod |
| PLDA | Tõenäosuslik lineaardiskriminantanalüüs, *Probablistic Linear Discriminant Analysis* |
| SOTA | Tippmudel, *State-of-the-art* |
| TDNN | Aeg-viivitusega närvivõrk, *Time-delay Neural Network* |
| TTÜ | Tallinna Tehnikaülikool |

# Sisukord

# Joonistе loetelu

# Tabelite loetelu

# 1. Sissejuhatus

## 1.1 Probleem

Viimase kümnendi jooksul on inimeste igapäevaelu loomulikuks osaks saanud hääljuhitavad assistendid, häälliidesed ning kõnetuvastusrakendundused. Kõnetehnoloogia-alane teadus- ja arendustöö on aidanud luua teenuseid, mis otsivad podcastidest sisu, tuvastavad klientide probleeme, aitavad häirekeskuses määrata abivajaja asukohta ning aitavad juhtida automaatselt ressursse. Kõnetuvastustehnoloogia aitab iseseisvalt keeli õppida ning koduseadmeid juhtida, pakkuda teleriekraanile subtiitreid abistamaks kurte või stenografistina transkribeerida parlamendisaadiku sõnavõttu.

Ühiskonna arengu tulemusel ja riikidevaheliste piiride avatuse tulemusel on hõlbustunud töö- ja õpiränne. Euroopa Liidus 2016. aastal läbiviidud uurimuse kohaselt räägib 35.2% inimestest vähemalt ühte võõrkeelt [1]. Teisalt vapustused nagu Venemaa kallaletung Ukrainale 2022. aasta veebruaris on toonud kaasa suure sõjapõgenike laine Euroopas, kes satuvad täiesti uude keelekeskonda, mis toob väljakutseid sihtriigi teenustele ja sõjapõgenikele endile. Et hakkama saada uues keelekeskkonnas õpilase, töötaja või sõjapõgenikuna peavad inimesed suutma rääkida asukohariigi keeles või vähemasti esimest võõrkeelt piisaval tasemel, et tarbida teenuseid ning asuda sihtriigis tööle või õppima. Mitte-emakeelena keelt rääkivad inimesed on väljakutseks ka kõnetuvastusrakendustele.

Kõneldava keele tuvastust (LID) kasutatakse eeltöötlusetapina kõnetuvastusrakendustes nagu masintõlge, inimese ja masina vahelistes kommunikatsioonisüsteemides ja mitmekeelse kõne transkriptsioonisüsteemides. LID-i kasutatakse tavaliselt ka kõnede automaatses suunamises, suunates kõne antud keelt kõnelevale klienditeenidnajale [2].

Kõnetuvastusrakenduste arhitektuuris on kõneldava keele võimalikult täpne tuvastamine esimene vajalik samm selleks, et kõnetuvastuse tulemus oleks räägitavas keeles võimalikult kvaliteetne. Akustilisi tunnuseid kasutavad ise-juhendatud *(self-supervised)* mitmel keelel treenitud tippmudelid suudavad selle töö kirjutamise ajal väga väikese veamääraga tuvastada emakeelt rääkivate inimeste kõneldavat keelt, kuid mudelite õigsus aktsendiga kõnelejate puhul on oluliselt madalam ning mudelid kipuvad akustilisi mustreid õppima tuvastama nii, et võõrkeelne kõne klassifitseeritakse aktsendi tugevusest sõltuvalt rääkija emakeelseks kõneks. Antud probleem on ka näide sellest, kuidas masinõpe võib tehnoloogilise kallutatusega selle loojale tahtmatult asetada kehvemasse olukorda keelelise vähemuse

ja seeläbi vähendada võimalike teenuste kättesaadavust võrreldes emakeelerääkijatega. Tehnoloogilist kallutatust võivad esile kutsuda tasakaalustamata treeningandmestikud või ka näiteks mudeli arhitektuur.

Kõneldava keele tuvastamine on kõnetehnoloogias levinud klassifitseerimisülesanne, mille kohta on ilmunud mitmeid uurimusi ja teadustöid. Aktsendiga kõnest räägitava keele tuvastamist oli kuni 2022. aastani vähe uuritud. On küll uurimusi, mis käsitlevad kõneldava aktsendi tuvastamist [3] või aktsendi tugevuse uuringuid [4], mis tingimata ei tähenda, et oleks loodud mudel, mis tuvastab kõneldavat keelt. Olemasolevad vähesed uurimused näitavad [5, 6, 7], et kõnetuvastussüsteemid, mis saavad hästi hakkama emakeelsest kõnega, on vähemedukad mitte-emakeelse kõnega. Antud tulemus ei ole üllatav, sest kõnetuvastussüsteemid on tavaliselt treenitud emakeelt kõnelevate inimeste kõneandmetega. Uurijad [8] näitavad fonotaktiliste mudelitega, et aktsendiga kõne põhjustab keskmiselt kolm korda enam kõneldava keele tuvastamisvigu kui emakeelne kõne.

Kõnetuvastushüpoteeside lisamine kõneldava keele tuvastamisele ei ole täiesti uus idee. Varasemad tööd [9, 10] on näidanud, et akustiliste mudelite kombineerimine kõnetuvastushüpoteesidega parandab kõneldava keele tuvastamist märgatavalt ja vähendab 24-70% võrra suhtelist veamäära *(relative reduction error)* võrreldes baasmääraks võetud akustilise mudeli tulemusega. Käesolev töö kombineeribki lisaks n-grammidel baseeruvaid Naïve Bayes'i tekstiklassifikatsioonimudeleid akustilistel tunnustel põhinevate mudelitega, näidates seejärel, et kõneldava keele tuvastuse õigsus aktsendiga kõne korral paraneb märgatavalt.

## 1.2 Eesmärk

Käesolev töö soovib pakkuda täiendava vaate räägitava keele tuvastamisele aktsendiga kõne puhul.

Töö esmane eesmärk on analüüsida seniste akustikal põhinevate tippmudelite (SOTA) õigsust aktsendiga kõne puhul analüüsides kõneldava keele identifitseerimise mudeleid aktsenti sisaldavates korpustes ning võrreldes tulemusi sama keelt emakeelena kõnelevate andmestike peal. Töö näitab, et LID-süsteemid, mis annavad suurepäraseid tulemusi emakeelsest kõnest räägitava keele tuvastamisel, võivad võõrkeelse kõne ja regionaalsete aktsentide korral märgatavalt halveneda.

Töö teiseks eesmärgiks on ühe või mitme leksikonivaba kõnetuvastushüpoteesi lisamine lisaparameetrina kõneldava keele tuvastamise mudelitele, näidates sellega, et antud võte kahandab aktsendiga kõne puhul veamäära 35-63% halvendamata sealjuures emakeelsest

kõnest keele tuvastust.

## 1.3 Ülevaade tööst

Töö põhineb suures osas töö autori samateemalisel teadusartikli jaoks tehtud uurimistöö tulemustel [11], autori ja juhendaja samateemaline artikkel avaldati 2022. aasta septembris rahvusvahelisel kõnetehnoloogia konverentsil Interspeech 2022.

Käesolev, töö esimene peatükk, annab ülevaate probleemistikust, tööle püstitatud eesmärgist ning teeb ülevaate tööst.

Töö teine peatükk annab ülevaate seotud kirjandusest, sealhulgas Wav2Vec 2.0 olemusest ning senistest temaatilistest uuringutest.

Töö kolmas peatükk "Kasutatud andmestikud"kirjeldab ja analüüsib töös kasutatud 6 andmestikku, nende omadusi ja mahtu.

Töö neljas peatükk kirjeldab töös püstitatud hüpoteese ja metoodikat.

Töö viies peatükk annab üksikasjalikult ülevaate töö käigus tehtud eksperimentidest ja kirjeldab töös kasutatud mudeleid.

Töö kuues peatükk analüüsib eksperimentide käigus saadud tulemusi.

Viimases peatükis 7, võetakse kokku töö eesmärk ning olulisemad tulemused ja diskussioon.

Täiendavalt on esitatud töö lisad.

# 2. Seotud kirjanduse ülevaade

Kõneldava keele identifitseerimine liigitub klassifitseerimisülesandeks. Esialgsel ülevaatel on aktsendiga kõnest räägitava keele tuvastamist kõneldavast keelest väga vähe uuritud, enam leidub emakeelsest kõnest kõneldava keele tuvastamisega tegelevaid artikleid [12, 13, 14]. Samuti leidub aktsendituvastamisega seotud töid, ent need ei tegele tingimata kõneldava keele identifitseerimisega, vaid keskenduvad aktsendi identifitseerimisele ja klassifitseerimisele. Järgnevalt annab autor ülevaate seotud töödest, mis käsitlevad kõneldava keele tuvastamist, töös kasutatava ise-juhendatud õppe ning akustiliste ja kõnetuvastushüpoteeside mudelite probleemistikku.

## 2.1 Aktsendi olemus

Kaasajal ei ole aktsendile ühest definitsiooni ning selle tekkpõhjused on üheselt selgitamata, tegemist on keeruka ja mitmetahulise nähtusega, mille tekkepõhjused on kõneleja keele omandamise perioodis. Aktsendi olemust põhjalikult käsitlenud artiklis [15] selgitatakse aktsenti kui kõrvalekallet tüüpilisest keeleliset häädusest, mille tunnusteks on 1) kõneleja ei räägi oma emakeeles 2) kõne erineb kuuldavalt emakeelsest kõnest ning 3) aktsent ei ole patoloogiline ning 4) hälbed on kuuldeliselt tajutavad. Autorid selgitavad aktsendi tekkepõhjuseid imikueas omandatud emakeelsed häädus- ja tajumallidega, mille väljaarenemisel tekib lastel fonoloogiline kurtus emakeeles mitte-esinevate diskreetsete vokaalide ja konsonantide ehk kõnesegmentide suhtes. Teise hüpoteesi tasemel põhjusena märgivad autorid, et aktsendi teke on seotud võõrkeele kõneorganite erineva kasutusega võrreldes emakeelsega. Antud hüpotees pole aga piisavat tõestust leidnud.

Aktsendi määratlemisel on oluline ka täpsustada, kuivõrd aktsendi all käsitletakse regionaalseid erinevusi või võõrkeelset aktsenti, sest näiteks erinevad tugevalt briti inglise keel ning Ameerikas räägitav nn Texase aktsent samuti on häälduselt erinevad Eesti-siseselt saarte murrak ja Põhja-Eesti eestikeelne kõne.

Meister [15] määratleb aktsenti eelkõige võõrkeelse kõne puhul esinevaks ning ka antud töös kasutatavad andmestikud ja töö sisu ei keskendu kohalikele erinevustele vaid võõrkeelsele aktsendile.

## 2.2 Kõneldava keele tuvastamine emakeelest ja aktsendiga kõnest

Aktsendiga kõne põhjustab tuvastamisel enam vigu kui emakeelne kõne. Artiklis [8] leitakse fonotaktilisi mudeleid kasutades, et aktsent põhjustab ligi kolm korda enam kõneldava keele tuvastamise vigu võrreldes emakeelse kõnega, veamäär suurenes keskmiselt 10% 28%-le maksimaalselt 8-sekundilise telefonikõne jooksul. Artiklis ilmneb samuti, et ka keelte vahel on oluline erinevus. Kui inglisekeelse emakeele ja aktsendiga kõnest keele tuvastamisel on 1.4-kordne veamäära vahe, on prantsuse keelel see lausa 10-kordne. Autorid toovad välja, et kõneldava keele tuvastamine sõltub ka kõne pikkusest. Mida lühem on kõne, seda suurem on veamäär. Samuti lisavad autorid, et aktsendiga kõnest räägitava keele tuvastamine vajaks suuremaid andmebaase rohkemate aktsentidega ja pikema kõnega, mida artikli kirjutamise hetkel polnud veel loodud.

Kõnealune artikkel on aga enam kui 20 aastat vana ning vahepealseid uurimusi kõneldava keele tuvastamisel aktsendiga kõnest pole autorile teadaolevalt ilmunud, ehkki keeletuvastuse mudelid on vahepeal jõudsalt edasi arenenud. Samuti on vahepeal edasi arenenud kõnekorpused ning leidub mahukaid korpuseid, mis sisaldavad nii aktsendiga kui emakeelset kõnet.

Lisaks kõneldava keele tuvastamisele näitavad mitmed senised kõnetuvastuse-teemalised uuringud [5, 6, 7], et kõnetuvastuses tekib rohkem vigu aktsendiga kõne tuvastamisel võrreldes emakeelse kõnega. See ei ole üllatav avastus, sest enamasti on kõnetuvastusmudelid treenitud emakeelsete korpuste peal. Tulenevalt vajadusest aktsendiga tegeleda, leidub töid, mis käsitlevad aktsendist tingitud kallutatuse vähendamist kõnetuvastuses. Selle üheks näiteks on hiljuti [16] leidnud magistritöös, et aktsendist tekitatud kallutatuse vähendamiseks on mõttekas kasutada keelteüleselt augmenteeritud andmeid ning kõne kiiruse muutmist *(speed perturbation)* transformeritel põhinevatel kõnetuvastusmudelitel. Samuti on magistritöö leidnud ühe olulise järeldusena, et aktsendist tingitud kallutatus kõnetuvastuses on mudelitest sõltumatu.

Erinevalt aktsendiga kõnest kõneldava keele identifitseerimisest, mis on veel lahendamata probleem, võib väita, et emakeelest kõne tuvastamisel on saavutatud tippmudelitega tase, millega võiks lugeda antud probleemi teaduslikult peaaegu lahendatuks. 2021. aastal toimunud Oriental Language Challenge [12] võistlusel, kus võistlejatel tuli tuvastada 17 keelt, saavutasid 1. ja 2. koha pälvinud tippmudelid veamäära alla 1%.

Emakeelest kõneldava keele tuvastamisest on see-eest ilmunud rohkelt artikleid, nt [12, 17, 18], mis keskenduvad akustilistele mudelitele, mille tänane *state-of-the-art* on Wav2vec 2.0 arhitektuur.

## 2.3 Ise-juhendatud õpe mitmekeelsetel kõnekorpustel

Viimastel aastatel on kõneldava keele tuvastus teinud läbi kiire arengu eelkõige tänu isejuhendatud mudelitele (näiteks XLS-R [19]), mis on treenitud väga suurtel mitmekeelsetel andmestikel. Samuti on avalikuks kasutamiseks tulnud mitmeid mahukaid kõnekorpuseid nagu Mozilla CommonVoice [20] ja VoxLingua107 [21]. Antud korpustel treenitud mudelid on võimelised saavutama väga kõrge õigsuse *(accuracy)* eelkõige emakeelse kõne tuvastamisel, kuna treeningmaterjaliks kasutatud korpused sisaldavad peamiselt emakeelset kõnet.

Ise-juhendatud õppe kasutamine mitmekeelsetel mudelitel võimaldab efektiivselt kasutada üht mudelit mitme keele tuvastamisel, vältides sellega mitme emakeelse mudeli kooskasutamist [19]. Wav2vec 2.0 arhitektuuril treenitud mudelid on inspiratsiooni saanud loomuliku keele töötluses kasutatavatest ise-juhendatud mudelitest. Autorid näitavad ka, et ise-juhendatud õppel treenitud mitmekeelseid mudeleid on võimalik pärast ületreenimist ja peenhäälestamist kasutada mitme erineva ülesande lahendamiseks (nt kõnetuvastus, kõneldava keele identifitseerimine, kõneleja identifitseerimine, kõne tõlge). Samuti toovad autorid välja, et mitmekeelsed isejuhendatud õppel mudelid on võimelised edukalt hakkama saama keeltega, mil on vähe treeningandmeid.

## 2.4 Akustilistel tunnustel ja kõnetuvastushüpoteesidel mudelid

Akustikal ja kõnetuvastushüpoteesidel kasutatav ühestatud mudeli idee ei ole täiesti uus. Artiklis [9] pakuvad autorid välja idee kõnetuvastuses lisada akustilisele mudelile tekstipõhiseid parameetreid, vähendades LID-ülesande lahendamisel süvaneurovõrke kasutades klassifikatsiooni suhtelist veamäära 21.8% võrreldes baastasemega. Tekstipõhiste parameetrite lisamiseks teisendatakse kõnetuvastusel saadud transkriptsioonide tekstid vektor-kujule, sisaldades sõnade esinemissagedusi tekstis ning need vektorid antakse lisaparameetrina süvanärvivõrgule sõltuvalt mudeli arhitektuurist, kas liidetuna või eraldi akustiliste tunnustega. Tekstihüpoteeside kasutamine võimaldab kõneldava keele tuvastamise mudelil lisaks akustilistele tunnustele arvestada ka keele ülesehitusest tulenevaid iseärasusi, mis väljenduvad transkriptsioonis.

Artiklis [10] võrdlevad uurijad kolme LID mudelit mitmekeelsel korpusel: (a) akustilistel tunnustel pikk lühiajalise-mälu (LSTM) (b) tekstihüpoteesidel põhinevat LSTM ning (c) akustilistel tunnustel ning tekstihüpoteesidel ühestatud LSTM mudelit, milles on iga lausungi kohta antud $n$ sõnavektorit ja $k$ keelt, kus iga keele $k$ kohta on vähemalt üks tõene sõnavektor ning ülejäänud $n - k$ on nullvektorid. Lihtsustuse huvides kasutab

artikkel keelepaare, seega $k = 2$. Ehk, kui mudel c teeb lausungi kõnetuvastuse näiteks inglise ning hispaania keeles ning sisendiks antav lausung on hispaaniakeelne, tehakse mõlemas keeles kõnetuvastus ning söödetakse mudelis (c) nii inglisekeelse ASR mudeli kui ka hispaaniakeelse mudeli kõnetuvastustulemus lisaparameetritena koos akustiliste tunnustega. Autorite hinnangul käitub niiviisi etteantud tekst tihendatud audiotunnustena parandades LID-tulemusi.

Artiklis [10] esitatud tulemustel on akustiliste ja keeletunnuste kombineerimisel mudelis (c) võimalik vähendada suhtelist veamäära koguni 24-70% võrreldes akustilise mudeliga (a). Samuti parandavad tulemusi ka üksnes keeletunnustel mudel (b), kuid jäädes alla siiski mudelile (c). Artikli nõrkuseks on mitmekeelsete ja tuntud korpuste kasutamise asemel ettevõttesisese andmestiku kasutamine, mille tulemusi ei ole võimalik üksüheselt taasesitada, kuid artiklis toodud mudeli (c) ideed on võimalik praktikas rakendada ning tuntud korpustel tulemusi võrrelda.

# 3. Kasutatud andmestikud

Selles peatükis on kirjeldatud töös kasutatud 6 andmestikku. Järgnevalt on selgitatud andmestike valiku kriteeriumeid ning antud ülevaade iga andmestiku puhul selle peamistest karakteristikutest. Töö teostamiseks on olnud vajalikud eraldi kriteeriumid treeningandmestikule ning tulemuste testimiseks kasutatud andmestikele.

Treeninguks kasutatavate andmete puhul on oluline andmete kvaliteet, võimalikult väike kallutatus, mis ennetab loodava mudeli puhul moonutusi ning treeninguks kasutatavate andmete maht. Näitena kallutatusest võib tuua olukorra, kus ühe keele üle või ala-esindatus andmestikus mõjutab treenitud mudeli täpsust antud keele tuvastuses ja võib põhjustada situatsiooni, kus mudel on kallutatud eelistama üle-esindatud keelt. Kõnetuvastuses akustilised mudelid, mis on treenitud meeshäältel tuvastavad suurema veamääraga naiskõneleja teksti ja vastupidi. Aktsendiga kõnest keele tuvastamise puhul, kui mudel ei ole õppinud aktsendiga kõne tunnuseid, on mudeli klassifitseerimisvõime kehvem.

Mitmekeelsete suurkorpuste nt Mozilla Common Voice ja VoxLingua107 kasutamine on andnud häid tulemusi nii kõneldava keele [12] kui kõnetuvastuses [20]. Mitmekeelsete korpuste kasutamine võimaldab tuvastada rohkem keeli ning tõhusamalt tuvastada üksteisele foneetiliselt lähedasi keeli samuti loob võimaluse kasutada mudelis tunnuste ülekannet ühelt keelelt teisele, et seeläbi tuvastada haruldasemaid keeli, mille kohta on vähe treeningmaterjali.

Treeningkorpuse valikul osutus oluliseimaks kriteeriumiks mitmekeelsus ning võimalikult tasakaalustatud andmestik spontaanse kõnega, mis annab võimaluse treenida universaalseid mudeleid kõneldava keele tuvastuseks ja neid võrrelda. Järgnevalt on esitatud kriteeriumid, mille järgi valiti treeningandmestik.

1. Andmestik on mitmekeelne, sisaldades vähemalt 20 keelt
2. Andmestik on lausepikkustelt balansseeritud
3. Andmestik on balansseeritud nii, et erinevate keelte treeningmaterjal pikkus on võimalikult ühtlane
4. Andmestik on suur (enam kui 1000h treeningmaterjali)
5. Andmestik sisaldab spontaanset kõnet
6. Andmestik on teadustööks vabalt kättesaadav
7. Andmestik on tasuta

8. Andmestik on sildistatud

Tööga alustamise hetkeks augustis 2021 jäi esmasesse valikusse kaks suurandmestikku: Mozilla Common Voice [20] ning TalTechis loodud VoxLingua107 [21].

Ehkki Common Voice sisaldab töö kirjutamise hetkeks rohkem treeningmaterjali tundides, siis valiku VoxLingua107 kasuks osutus keelte arv (107 vs 87) ning balansseeritum hulk treeningmaterjali iga räägitava keele kohta, mis vähendab tõenäosust moonutusteks ning on täpsem. Teisalt võimaldades tuvastada rohkem keeli, olles seega universaalsem.

VoxLingua107 on kokku pandud Youtube's ilmunud videote töötluse tulemusel ja sisaldab spontaanset kõnet, jagunedes omakorda eeldefineeritud treening- ja testandmestikuks, tekitades sellega parema võrreldavuse ka hiljem ilmuvate teadustööde vahel, mis kasutavad sama andmestikku.

Töö tulemuste võrdlemiseks, oli vajalik leida andmestikud, mille puhul on võimalik võrrelda nii aktsendiga kõnel kõneldava keele tuvastuse täpsust kui ka aktsendita kõne puhul treenitud mudelite täpsust. Andmestike valiku tegemisel seati töös test-andmestikele kriteeriumid:

1. Andmestik peab sisaldama aktsendiga kõnet.
2. Aktsendiga kõnet sisaldavale korpusele peab olema võimalik kõrvutada emakeelse kõnega võimalikult sarnast korpust.
3. Andmestik peaks olema vabalt kättesaadav teadustööks
4. Andmestik peaks olema eelistatult tasuta kättesaadav või hinnaga alla 500 euro
5. Kui on tegu mitmekeelse andmestikuga, peab olema olemas info iga lausungi keele kohta.
6. Andmestikus peaks olema võimalikult erinevaid kõnelejaid
7. Andmestike hulgas peaks olema esindatud nii spontaanse kui ka dikteeritud kõnega korpuseid
8. Andmestikus ei tohi olla liigset taustamüra.
9. Andmestikul peaks olema kättesaadav kuulatav näidis.
10. Andmestikus olevad lausungid ei tohiks olla liiga pikad (mitte pikemad kui 30 sekundit).
11. Andmestik peaks olema piisavalt mahukas.
12. Andmestik peaks sisaldama erinevaid kõnelejaid.

Esmasesse valikusse valis autor välja 21 korpust, millest omakorda valis välja 6 andmestikku, millega edasi töötada. Nii lõppvalikusse jäänud CSLU Foreign Accented English kui

ka L2 Arcticule on võimalik leida vastavad emakeelsed andmestikud (CSLU 22 Languages English Subset ja CMU Arctic). TTÜs loodud Aktsendikorpus sisaldab nii aktsendiga kui ka aktsendita kõnet ning andmestikus on esitatud ka rääkija aktsendi tase.

Aktsendiga kõne andmestike välistavaks kriteeriumiks osutus peamiselt võrreldava aktsendita kõnega andmestiku kättesaadavus, mis andnuks võimaluse võrrelda, kas aktsendiga kõne keeletuvastuse tulemuse paranemisel mõjutab see ka aktsendita kõne tulemust. Lõppvalikus üheks välistavaks teguriks sai ka asjaolu, et leidus andmestikke, mille puhul ei olnud võimalik leida allalaadimisvõimalust või oli näidisfail vigane.

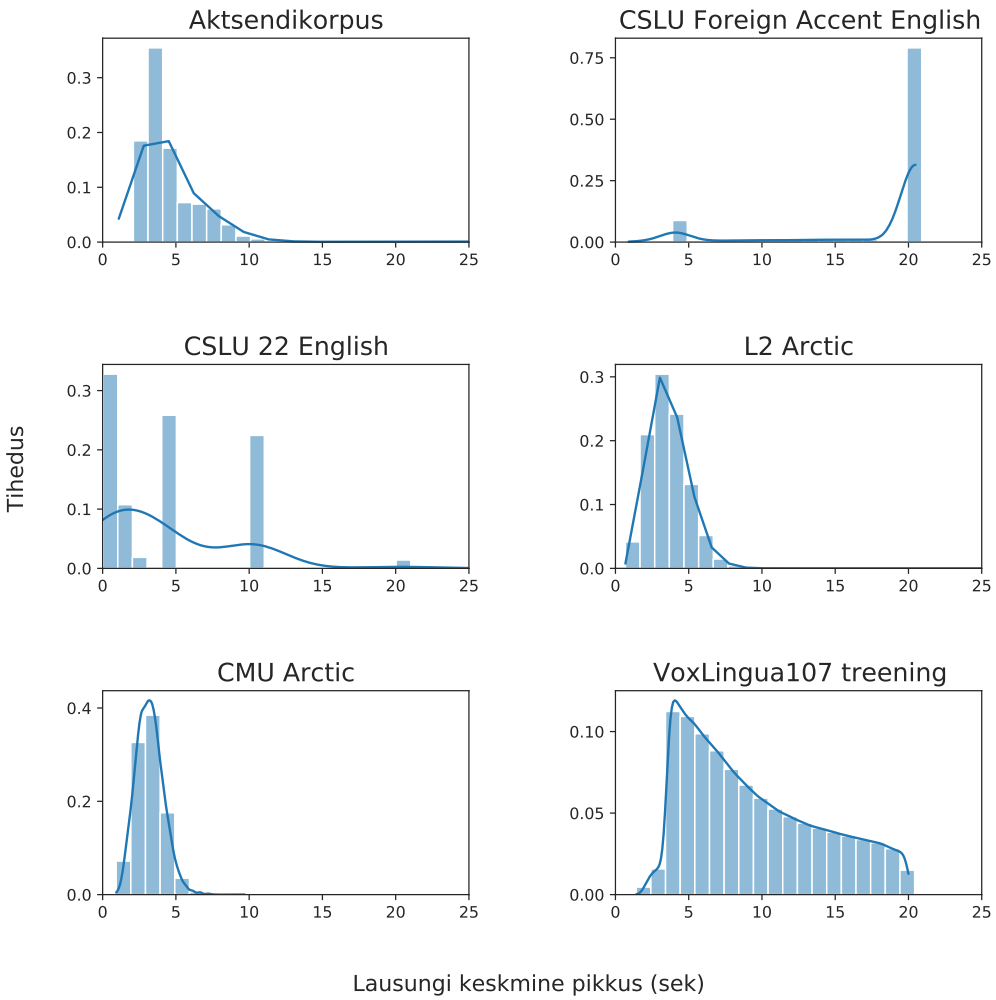
Lausungite pikkus mõjutab mudelite efektiivsust kõneldava keele tuvastamisel. Joonisel 1 on esitatud 6 korpuse lausete pikkuse tihedus. Sellest nähtub, et nii CSLU 22 English korpusel esineb alla 1-sekundilisi lauseid proportsionaalselt võrreldes teiste korpustega. Kõige ühtlasema pikkusega lausungid on VoxLingua107 andmestikus ning samuti L2 Arctic korpuses.

Kokkuvõtvalt on 6 andmestiku iseloomustus esitatatud tabelis 1.

Tabel 1. *Treening- ja testandmestike iseloomustavad näitajad.*

| **Andmestik** | **Keel** | **Aktsent** | **Disk. sage-dus (kHz)** | **Tüüp** | **Lausungeid** | **Lausungi kesk-mine pikkus (sek)** |
|---|---|---|---|---|---|---|
| Aktsendikorpus | ET | Jah/Ei | 44.1 | Spon/dikt | 32649 | 5.9 |
| CSLU Foreign Accented English | EN | Jah | 8 | Spon. | 4925 | 17.9 |
| CSLU 22 Languages (English) | EN | Ei | 8 | Spon/dikt | 2206 | 6.4 |
| CMU Arctic | EN | Ei | 16 | Dikt. | 14471 | 3.2 |
| L2 Arctic | EN | Jah | 44.1 | Dikt. | 25758 | 3.7 |
| VoxLingua107 train | 107 | Ei | 16 | Spon. | 2.54M | 9.4 |
| VoxLingua107 dev | 33 | Ei | 16 | Spon. | 1608 | 10.0 |

Järgmistes alapeatükkides annan detailsema ülevaate igast kasutatavast korpusest.




Joonis 1. *Andmestiku lausete keskmise pikkuse tihedus.*

## 3.1 Aktsendikorpus

Aktsendikorpus (versioon 1) koosneb 185 mitte emakeelt kõneleva inimese (L2) ja 20st eesti keelt emakeelena kõneleja lausungitest. Aktsendikorpus sisaldab igalt kõnelejalt keskmiselt 25-30 minutit kõnet [22]. Kõne on salvestatud stuudios, kasutades 16-bitist 44,1 kHz stereoformaati. Andmekogum koosneb 32649 lausungist, kokku 53,2 tundi (48,8 tundi võõrkeelset ja 3,4 tundi emakeelset kõnet), keskmine lausung on 5,9 sekundit.

Aktsendikorpus sisaldab näiteid spontaansetest ja dikteeritud kõnest (136 foneetiliselt rikkalikku lauset ja kaks lühikest teksti). Kõnekorpus sisaldab neutraalseid lauseid (130), milles on esindatud kõik eesti vokaalid ja konsonandid kolmes erivältelises sõnastruktuuris (kahesilbilistes sõnades), sagedasemad diftongid ja konsonantühendid, küsilaused (8) ja

kaks pikemat tekstilõiku (5–6 lauset). Lisaks spontaanset kõnet: kolme pildi kirjeldust, oma perekonna ja päritolu kirjeldust ning vastuseid eesti keele õpingute ja kasutamise kohta. [15] Korpuses on esindatud 18 eesti keelest erineva emakeele kõnelejad.

Andmekogum sisaldab ka katsealuste endi hinnangut eesti keele oskuse tasemest neljal tasemel emakeelena: a) algtase, b) keskmine, c) hea, d) väga hea.

## 3.2 CSLU Foreign Accented English

CSLU Foreign Accented English (CSLU FAE) [23] versioon 1.2 koosneb 22 inglise keelest erineva emakeelega keelejuhtide inglise keelsest kõnest.

Kõne on salvestatud telefoni teel, kasutades 16-bitist 8kHz mono kanalit. Korpus sisaldab kokku 4925 lausungit (24h) spontaanset telefonikõnet, teavet kõnelejate keelelise tausta kohta ja hinnanguid lausungites kõneldava aktsendi kohta. Kõnelejatel paluti rääkida endast 20 sekundit inglise keeles, keskmine lausungi pikkus korpuses on 17,9 sekundit.

## 3.3 CSLU 22 Languages English

CSLU 22 Languages English (CSLU 22) [24] alamkorpus sisaldab 3,9-tundi inglise keelt emakeelena rääkivate keelejuhtide kõnet. Korpuses on kõne, mis on salvestatud 16-bitise 8kHz mono kanaliga. Lausungid sisaldavad nii dikteeritud kõnet kui ka spontaanset telefonikõnet. Korpuses on kokku 2206 lausungiit, lausungite keskmine pikkus on 6,4 sekundit. Nagu on näha jooniselt 1, on suur osa lausungitest lühemad kui üks sekund.

## 3.4 CMU Arctic

CMU Arctic korpus [25] koosneb 12,9 tunnisest dikteeritud kõnest 18 inglise keelt emakeelena rääkivalt kõnelejalt, kellel on erinevad Ameerika aktsendid, kuid esineb ka Kanada, Šoti ja India dialektide rääkijaid. Korpus sisaldab 16-bitist 16 kHz mono kanalis salvestatud kõnet. Lausungite koguarv on 14471, lausungi keskmine pikkus on 3,2 sekundit.

## 3.5 L2 Arctic

L2 Arctic korpus koosneb 26,4 tunnist dikteeritud inglise keelsest kõnest 24-lt mitte-emakeelena kõnelevalt inimeselt, kes on 18 erineva keelelise taustaga. Korpus sisaldab keskmiselt 67,7 minutit kõnet kõneleja kohta. Kõne on salvestatud 16-bitises 44,1 kHz mono kanalis. Lausungeid on kokku 25758, keskmine lausungi pikkus on 3,7 sekundit.

## 3.6 VoxLingua107

VoxLingua107 [21] (Vox107) andmete puhul on eristatud töös treening- ja andmestiku loojate poolt eeldefineeritud testandmestik, mida ka selles töös on kasutatud.

Vox107 treeningkorpus sisaldab 6628 tundi emakeelset spontaanset kõnet, mis on eraldatud automaatselt kraabitud Youtube'i videotest.

Andmete kogumise protsess koosnes järgmistest sammudest. Esiteks genereeriti Vikipeediast juhuslikud trigrammi otsingufraasid otsitavas keeles. Otsingufraase kasutati YouTube'i videote otsimiseks, mille pealkiri ja/või kirjeldus vastas otsingufraasile. Seejärel kasutati tekstipõhist keeletuvastust videote filtreerimiseks, mille pealkiri ja kirjeldus ei ole tõenäoliselt antud keeles. Kõneaktiivsuse tuvastamiseks ja segmentide eraldamiseks rakendati kõneleja diariseerimist kõnet sisaldavatest videotest. Pikad kõnelõigud eraldati lausungitaolisteks kuni 20-sekundilisteks alasegmentideks. Järgnevalt kasutati andmepõhist järelfiltreerimist, et eemaldada andmestikust segmendid, mis tõenäoliselt ei olnud antud keeles. Selle tulemusel tõusis rahvahanke *(crowd-source)* meetodil õigesti märgendatud segmentide hulk 98%-ni.

VoxLingua107-s on kõneandmed 107 erinevas keeles, lausungite arv on 2,54 miljonit. Keskmine lausung on 9,4 sekundit. Keskmine andmemaht ühe keele kohta on 62 tundi varieerudes 3st - 155 tunnini.

VoxLingua107 testandmestik koosneb 4,5 tunnist spontaansest kõnest, mis sarnaselt treeningandmetega on kraabitud ja eraldatud Youtube'i videotest. Testandmestikus on esindatud 33 erinevat keelt, kokku on andmestikus 1608 lausungit. Testandmestiku sildid on verifitseeritud vähemalt kahe vastavat keelt hästi oskava vabatahtliku poolt.

# 4. Metoodika

Järgnev peatükk tutvustab töös kasutatavat metoodikat, andes ülevaate hüpoteesidest ja töö käigust.

Selleks, et parandada kõneldava keele tuvastamise täpsust, püstitasin töö alguses hüpoteesi: töö teostamise hetkel avalikult kättesaadavad maailma parimad mudelid tuvastavad aktsendiga kõnest räägitavat keelt oluliselt väiksema õigsusega kui emakeelest räägitavat keelt. Kuna aktsendiga kõnest räägitava keele tuvastamise kohta teadustöid on napilt, tuli esimesena kindlaks teha, millise õigsusega tänased SOTA mudelid kõneldavat keelt tuvastavad, et seejärel otsida võimalusi parandamiseks. Kui esialgne hüpotees leiab töö käigus kinnitust, otsib autor võimalusi, et saadud tulemusi parandada.

## 4.1 Andmete ettevalmistamine

Andmete ettevalmistamisel ehitas autor andmetorud *(data pipeline)* andmete treenimiseks ja mudelite hindamiseks. Kõik teisendamata andmed teisendati 16 kHz mono-kanal wav-formaati ning saadud andmeid kasutati hindamiseks. VoxLingua107 treenimiseks kasutati treeningskripti ettevalmistamisel juba varasemalt TTÜs loodud andmetoru.

## 4.2 Õigsus

Õigsus (ingl k *accuracy*) on masinõppes laialt kasutatav mõõdik, mis on esitatud valemi abil:

$$\tilde{O}igsus = \frac{TP + TN}{TP + TN + FP + FN}$$

$TP - true\ positives, õigesti\ tõeseks\ määratletud$
$TN - true\ negatives, õigesti\ negatiivseks\ määratletud$
$FP - false\ positives, vääralt\ tõeseks\ määratletud$
$FN - false\ negatives, vääralt\ negatiivseks\ määratletud$

Eesti keeles kasutatakse inglisekeelse $accuracy$ tavakeele tõlkena mõistet $täpsus$, et aga mitte segi ajada masinõppes laialt kasutusel olevaid mõisteid $precision$ ja $accuracy$, kasu-

tatakse siinkohal viimase puhul mõistet $\tilde{o}igsus$, mida pakub ka masinõppesõnastik [26]. Õigsuse valemist saab järeldada, mida rohkem on klassifitseeritud hulga elemente tõeselt, seda parem on mudel. Õigsuse pöördtehe on veamäär *(error rate)*, seega kõrgem õigsus tähendab väiksemat veamäära ja vastupidi. Klassifikatsiooni ülesande puhul kasutatakse vea määratlemisel mõistet klassifitseerimisviga. Õigsust kasutatakse töös peamise mudelite võrdlemise andmetel: mida kõrgem on mudeli õigsus ja järelikult väiksem klassifitseerimisviga, seda paremini suudab mudel tuvastada kõneldavat keelt.

## 4.3 Seniste eeltreenitud tippmudelite hindamine

Oluline on küsimus, milline õigsuse vähenemine on piisavalt oluline? Töö kirjutamise hetkel kättesaadavatest allikast [8] ilmneb, et fonotaktiline mudel tekitab aktsendiga kõnest keele tuvastamine ligikaudu kolm korda rohkem vigu kui emakeelsest kõnest. Kuna aga tegemist on enam kui 20 aastat vana artikliga, siis värskete tulemuste jaoks tuli kõigepealt hinnata state-of-the-art mudeleid. Emakeelest kõneldava keele tuvastamine on parimatel juhtudel võimalik 99.5% õigsusega [12] ja seega võib öelda, et tegemist on teaduslikult peaaegu lahendatud probleemiga, kuna selle õigsus on ligilähedane 100-le. Aktsendiga kõnet sisaldavate korpuste puhul oli autori esialgne hinnang, et see tulemus võib samade mudelite puhul jääda vahemikku 40-55%. Selleks aga, et hüpoteesi kehtivust kontrollida klassifitseeriti SOTA mudelid ning iga korpuse puhul arvutati õigsus.

## 4.4 Korpuste võrreldavus

Aktsendita ja aktsendiga kõne juures on võrreldavaid tegureid, mis sõltuvad korpuses olevast lause pikkusest, kõneleja karakteristikutest (mees, naine, laps, vanur) ja kõne tüübist (dikteeritud, spontaanne) samuti salvestise kvaliteedist ning müra olemasolust. Töö seisukohalt oluline, et tulemuste paranemise vaikimisi eeldus on, et paranemine ei tohi tulla aktsendita kõne arvelt või täiendavate andmete lisamise tulemusel. Eksperimendiks kasutatud mudelite puhul peaks mudel säilitama oma samaväärse õigsuse tuvastada aktsendita kõnet samal ajal tuvastades aktsendiga kõnet kõrgema õigsusega.

Võrreldavuse huvides peavad aktsendita ja aktsendiga kõne korpused olema võrreldavad, st korpused peaksid sisaldama nii emakeeles kui ka mitte-emakeeles kõnelejate kõnet. Aktsendikorpuse juures on viimane nõue täidetud, kuna see sisaldab emakeeles kui ka mitte-emakeelseid rääkijaid samuti on omavahel võrreldavad CSLU FAE ja CSLU 22 English korpus ning ka CMU Arctic ja L2 Arctic. Siiski tasub märkida, et CSLU FAE ja CSLU 22 korpused sisaldavad küll mõlemad telefonikõnet, kuid CSLU FAE korpuse suur osa lauseid on enam kui 20 sekundit pikad, samal ajal CSLU 22 English korpuse laused

on lühemad, sealjuures vähem kui 1 sekund. Lühikesed laused see-eest on väljakutseks kõneldava keele tuvastamise mudelitele, mis selgitab ka hiljem mudelite tulemusi tabelis 2

## 4.5 Eeltreenitud mudelite kasutamine

Pärast esialgsete tippmudelite hindamist, otsis autor võimalusi aktsendita kõnest keele tuvastamise õigsuse parandamiseks. Võrdlusbaasiks peenhäälestati ja treeniti XLS-R 300M VoxLingua107 andmestikul ning seejärel prooviti mudeleid täiendada, kasutades tekstihüpoteesidel genereeritud tõenäosusvektoreid lisasisendiks Multinomiaalse Naive Bayes'i mudelile kui ka katsetati LDA/PLDA ning ühestatud mudelitega.

Saadud mudeleid hinnati kõigil kuuel kirjeldatud andmestikel võrreldes kaalutud keskmist õigsust võrdlevalt ning hinnati, milline mudel osutus parimaks. Kogu treeningprotsess toimus TTÜ teadusarvutusklastris, kasutades vajadusel paralleelselt kuni kahte NVIDIA A100 GPUd.

# 5. Eksperimendid

Järgnev peatükk annab ülevaate kasutatavatest raamistikest, mudelite hetkeolukorrast ning seejärel tutvustab töös kasutatud eeltreenitud mudeleid, mida autor arendas edasi. Töös kasutatakse kõneldava keele tuvastamiseks eeltreenitud mudeleid, mida treeniti üle VoxLingua107 treeningandmestikuga.

## 5.1 Treeningkeskkond ja kasutatavad raamistikud

### 5.1.1 Speechbrain

SpeechBrain on PyTorchil põhinev avatud lähtekoodiga kõnetehnoloogia raamistik, mis sisaldab nii treeningskripte, mudeleid ning mudelite ennustusskripte [27].

SpeechBrain võimaldab lahendada erinevaid kõnetöötluses tekkivaid ülesandeid, artikli [27] väitel on neid 7 ent ajaga on ülesandeid lisandunud. SpeechBraini piiranguks on kõnesünteesi *(text-to-speech)* tüüpi ülesannete komponentide puudumine ja raamistik ei ole fokusseeritud reaalaja kõnetuvastusele, kuid siiski sisaldab raamistik audio klassifitseerimiseks kõiki vajalikke komponente.

SpeechBrain on PyTorchist pärineva GPU toega. Kõnetuvastuses, kõneleja-vaheldumises ning -identifitseerimises on SpeechBrain näidanud võrreldavaid tulemusi jõudluselt seni-ajani kasutusel olnud SOTA raamistikega, olles efektiivne ka kõneldava keele tuvastuses.

SpeechBrainis kasutatud treenimiseks on vajalik luua treeningskript, mis kasutab retsepti *(recipe)*, kus on defineeritud treeningskriptis kasutatavad funktsioonid ning mudeli arhitektuur. Treeningskript sisaldab andmelaadimist, mudeli treeningut ning hindamiskomponente. Ennustussskripti loomine sõltub sellest, kas mudel kasutab standardseid SpeechBraini sisse-ehitatud ennustusskripti või on vajalik luua unikaalne ennustusskript.

### 5.1.2 HuggingFace

HuggingFace'i repositooriumit on meediaväljaannetes kutsutud masinõppe Githubiks [28], sisaldades nii andmestikke kui ka eeltreenitud mudeleid, mida on võimalik uurijatel, andmeteadlastel või lihtsalt huvilistel paari koodireaga kasutada. Huggingface on tihedalt seotud ka Speechbrainiga, paljud kõnetehnoloogia mudelite treeningskriptid on ehitatud

Speechbraini raamistikul ning kasutajatel on lihtne neid üle treenida või juba valmis mudelit kasutades ennustada.

HuggingFace'i repositoorium sisaldab nii mudeleid kui ka andmestikke [29]. Ka töö käigus loodud VoxLingua107 andmestikul treenitud wav2vec XLS-R mudel on lisatud repositooriumisse.[1]

### 5.1.3 Kaldi

Kaldi on vabavaraline kõnetehnoloogiaraamistik, mis on kirjutatud C++'s. Kaldi on lõplikel keelekonverteritel *(tranducer)* lineaaralgebra toega ning kõnetuvastuseks vajalikke retsepte pakkuv kõnetuvastussüsteem, mis avaldati 2011. aastal [30]. Selles töös on kasutatud Kaldi sisend-väljundmoodulit eksperimentide jaoks andmete sisselugemisel.

### 5.1.4 PyTorch

PyTorch on avatud lähtekoodil masinõppe raamistik süvaõppe mudelite loomiseks. PyTorch kasutab arvutustel ja andmevoogudes tensoreid ja tensor-operatsioone, on GPU toega ja pakub arendajale kõrge abstraktsioonitasemega arendusraamistikku [31].

PyTorchi uue mudeli defineerimisel tuleb mudel defineerida klassi või klassidena, pärides baasklassi nn.Module omadused ning lisades mudelisse soovi korral erinevaid kihte ja kaofunktsioon, mis on samuti määratletud klassidena. Raamistikus on eeldefineeritud mudeli treenimiseks vajalikud funktsioonid, mida arendaja saab vastavalt vajadusele muuta.

Raamistik on kõrge jõudlusega võrreldes konkureerivate raamistikega. Selle mäluhaldus on üles ehitatud aktiivses kasutuses olevate tensorite loendamisele - kui loendur on null, siis aktiivne mäluviit vabastatakse, andmevood on asünkroonsed ning lisaks kasutab raamistik paralleelprotsesse, mida on GPU-l võimalik rakendada.

## 5.2 Hinnang senistele tippmudelitele

Selleks, et mõista, kui sügav on probleem kõneldava keele tuvastamiseks aktsendiga kõnest, oli vajalik anda hinnang olemasolevatele parimatele mudelitele. Töö teostamise hetkel osutus valikusse 2 mudelit, mis on näidanud häid tulemusi ja mille põhjal hinnang kujundati: TTÜs VoxLingua107 põhjal loodud ECAPA TDNN arhitektuuril mudelile[2],

[1] https://huggingface.co/TalTechNLP/voxlingua107-xls-r-300m-wav2vec
[2] https://huggingface.co/TalTechNLP/voxlingua107-epaca-tdnn-ce

mille veamäär *(error rate)* oli VoxLingua107 testandmestikuga 6.7% ning Speechbraini 45 keelel treenitud langid-commonlanguage-ecapa, mille õigsus on autorite väitel 85%[3]. Antud mudelite täpsused siiski ei ole omavahel võrreldavad, sest nende väljundklasside arv on erinev.

Läbi viidud eksperimendi tulemusel tabelist 2 nähtub, et aktsendiga kõne puhul on veamäär kõrgem aktsendita kõnest pea kõikidel juhtudel, olles suurim Aktsendikorpusel, parimal juhul veamäär puudub (L2 Arctic). Aktsendikorpuse (EFAC) puhul on tulemused kõige ühesemalt erinevad.

Eksperimendi tulemusel võib järeldada, et probleem aktsendiga kõnest kõneldava keele tuvastamiseks on olemas ja töö algne hüpotees õigsuse vähenemisest leidis kuuest võrdlusest nelja puhul kinnitust - aktsendiga kõne korral keelt tuvastavad mudelid annavad nõrgemaid tulemusi võrreldes emakeele rääkijatega. Veamäär Aktsendikorpusel on halvimal juhul koguni 71.8% võrreldes emakeele veamääraga 8.4%, ületades ka artikli [8] hinnangut veamäärale enam kui kahekordselt.

Tabel 2. *Akustiliste SOTA mudelite kõneldava keele identifitseerimise õigsus eesti ja inglise keeles.*

| | **Inglise keel** | | | | **Eesti keel** | |
|---|---|---|---|---|---|---|
| **Mudel** | **CMU Arctic Emakeel** | **L2 Arctic Võõrkeel** | **CSLU FAE Võõrkeel** | **CSLU 22 en Emakeel** | **EFAC Emakeel** | **EFAC Võõrkeel** |
| Vox107 epaca-tdnn | 78.7 | 47.7 | 59.5 | 48.6 | 91.6 | **28.2** |
| Speechbrain commonlanguage-ecapa | 80.9 | **81.0** | 46.4 | 46.7 | 93.6 | 46.5 |

## 5.3 Mudelite arhitektuur

### 5.3.1 Multinomiaalne Naive Bayes kõnetuvastushüpoteesidel

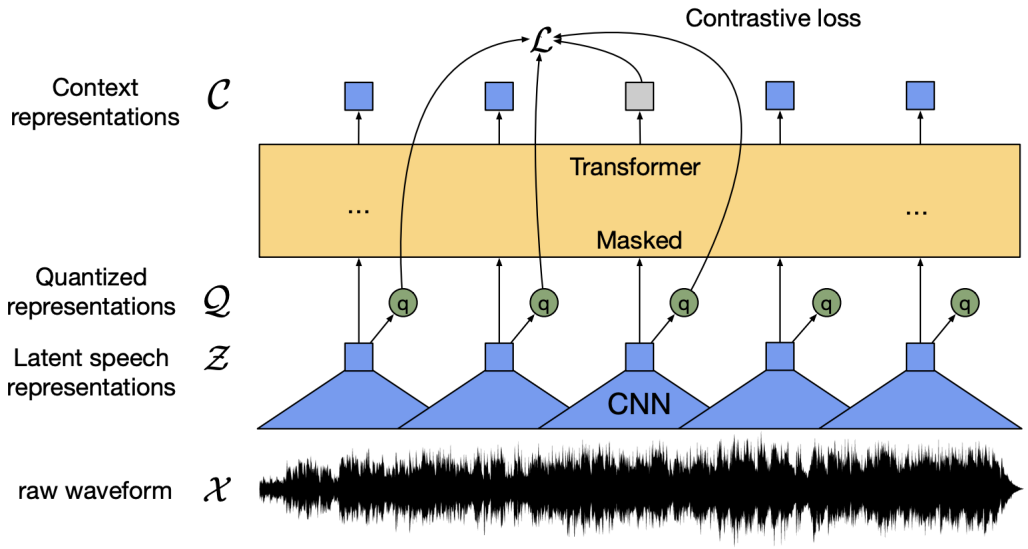
Multinomiaalne Naïve Bayes'i mudel (NB) on tõenäosuslik juhendatud õppe mudel, mis autorite [32] järgi lausungi $d$ kuulumist klassi $c$ järgmise valemi järgi:

$$P(c|d) \; \alpha \; P(c) \prod_{i \leq k \leq n_d} P(t_k|c),$$

[3] `https://huggingface.co/speechbrain/lang-id-commonlanguage_ecapa`

kus $P(t_k|c)$ on tingimuslik termi $t_k$ esinemistõenäosus lausungis, mis kuulub klassi $c$. $P(c)$ interpreteeritakse kui mõõdikut, kui palju tõendeid $c$ pakub, et $c$ on antud juhul õige klass. $P(c)$ on eelnev tõenäosus, et lausung kuulub klassi $c$. Kui lausungi termid ei paku selgeid tõendeid ühe või teise klassi eelistamiseks, valitakse selline klass, millel on kõrgem aprioorne tõenäosus.
$\langle_{1,2}, \ldots, _{n_d}\rangle$ on sõned lausungis $d$, mis on osa sõnastikust, mida kasutatakse klassifitseerimiseks ja $n_d$ on selliste sõnede arv lausungis $d$[32].

### 5.3.2 Wav2vec 2.0 arhitektuur




Joonis 2. Wav2vec2 arhitektuur[33].

Wav2vec 2.0 on isejuhendatud mudel, mille arhitektuur koosneb kolmest kihist: konvolutsioonilisest, transformer ning kvantifitseerimiskihist [33].

Wav2vec 2.0 arhitektuuri esimene kiht koosneb mitmekihilisest konvolutsioonilisest tunnuste enkoodrist $f : X \mapsto Z$, mis kodeerib audiosisendi latentseteks kõnetunnusteks $z_1, ...z_T$ (512-mõõtmelisteks vektoriteks), iga ajasammu $T$ (20ms) kohta. Erinevalt wav2vec esimesest generatsioonist arvestab konvolutsiooniliste tunnuste enkooder ka tulevikku [33].

Eeltreenimisel kõnetunnused seejärel liiguvad edasi kahte haruprotsessi. Esiteks liiguvad kõnetunnused kvantifitseerimiskihti, kus kontekstiesitus kvantifitseeritakse $Z \mapsto Q$. Kvantifitseerimisel seotakse konvolutsioonilisest tunnuste enkoodri väljund $Z$ klastri tsentri vektoresitusega $q$, klasterdamne toimub lähima naabri meetodil nii, et klastri tsentrite vaheline kaugus oleks minimaalne ning iga klastri väljundiks on klastri tsentri vektoresitus $q$ [33]. Kvantifitseerimisosa treenitakse koos ülejäänud mudeli osadega.

Kvantifitseerimisel genereeritakse kõigepalt $G$ gruppi ning igas grupis $V$ elementi (autorid on andnud väärtuseks $G = 2$ ja $V = 320$), moodustades elementidest maatriksi (koodiraamatud) $e \in \mathbb{R}^{d/G}$, kus $\mathbb{R}$ on ridade arv maatriksis, $d/G$ tähistab maatriksi dimensiooni $d$ jagatist gruppide koguarvuga $G$. Kvantifitseeritud tsentri vektoresitused $q$ luuakse seejärel juhuslikult maatriksist (koodiraamatust) $e$ valides juhusliku rea ning liites saadud vektorid $e_1, ..., e_G$. Seejärel rakendadatakse vektorile lineaarteisendust $\mathbb{R}^d \rightarrow \mathbb{R}^f$, ning saades niiviisi klastri tsentrile vastavuses oleva kvantifitseeritud vektori $q \in \mathbb{R}^f$.

Samal ajal teises haruprotsessis liiguvad algsel kujul kõnetunnuste vektorid ka transformer-kihti $g : Z \mapsto C$, mis koosneb mitmest transformer-enkooder kihist, mille väljund on vektoritena kontekstuaalne esitus $c_1, ...c_T$ üle kogu kõnetunnuste järjestuse. Transformer-kihis juhuslikult valitud kõnetunnused maskeeritakse, teisisõnu asendatakse juhuvektoritega, mida hiljem hakatakse taastama. Maskeerimise käigus valitakse kõnetunnuste 20ms 6.5% raamidest ning valitud ja 9 järgneva raami vektorid asendatakse. Selle tulemuseks on, et 15-sekundilise klipi korral keskmiselt 49% audiofaili raamidest on ülekattega maskeeritud. Transformer-kihi ennustuse tulemuseks peab olema vektor, mis on vastavuses kvantifitseerimise tulemusel saadud vektoriga, mis vastab klastri tsentrit esindavale vektorile $q$ [33].

Wav2vec 2.0 kaofunktsioon tagasilevil *(back-propagation)* avaldub kujul:

$$L = L_M + \alpha L_D$$

kus $L_M$ on kontrastiivne kaofunktsioon:

$$L_M = -\log \frac{\exp\left(sim(c_t, q_t)/\kappa\right)}{\Sigma_{\tilde{q} \sim Q_T} \exp\left(sim(c_t, \tilde{q})\right)/\kappa)}$$

kus $sim$ on koosinussarnasus kontekstuaalse esituse $c_T$ ja kvantifitseeritud esituse vektori $q_T$ vahel ning $\kappa$ on mittenegatiivne temperatuur [33].

Kontrastiivse kaofunktsiooni eesmärk on identifitseerida tõesed kvantifitseeritud (ja maskeeritud) tsentrid $q$ ajahetkel $T$ segajatest $\tilde{q}_T$. Kaofunktsioonis arvutatakse mõlemal juhul koosinussarnasuse $sim$ transformer-kihis saadud kontekstuaalse esituse $c_T$ ning kvantifitseeritud esituse $q$ vahel, jagades läbi konstantse mitte-negatiivse temperatuuriga $\kappa$, mille eesmärk on vähendada kontekstuaalse esituse ning kvantifitseeritud esituse juhuslikkust treeningu edenedes. Valemi lugejas on tõeste kvantifitseeritud sarnasused ning nimetajas nn segajate (maskeeritute) sarnasused, võimendades nii tõeseid ja vähendades

segajate olulisust. Viimaseks rakendatakse koosinussarnasusele $softmax$ funktsiooni ning $Gumble'i\ softmax$ funktsiooni, andes tulemuseks tõenäosusjaotuse, et kontekstuaalse esituse vektor $c$ ajahetkel $T$ vastab kvantifitseeritud esitusele $q$ ajahetkel $T$ [33].

Selleks, et mudeli entroopia oleks maksimaalne ehk mudelis valitakse võimalikult erinevaid (positiivseid ja negatiivseid) koodiraamatu $g$ elemente $v$ võimalikult ühtlase sagedusega, rakendadakse lisaks $L_M$ ka mitmekesisuse kaofunktsiooni (diversity loss) $L_D$, mis avaldub:

$$L_D = \frac{1}{GV} \sum_{g=1}^{G} -H(\overline{p}_g) = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \overline{p}_{g,v} \log \overline{p}_{g,v}$$

kus $\overline{p}_{g,v}$ on tõenäosusjaotus koodiraamatu $g$ sissekande $v$ esinemiseks valikul.

Mudeli peenhäälestamisel LID klassfikatsiooniülesende lahendamiseks lisatakse wav2vec kihile $attentive\ pooling$ kiht, $fully\ connected$ kiht koos $ReLU$ ja ploki normaliseerimisega *(batch normalization)* ning viimaseks keelte väljundiga klassifitseerija [33].

Wav2vec 2.0 arhitektuuril treenitud mudel töötab seega põhimõttel, õppides nii audiokui ka keeletunnuste kohta, maskeerides või "peites ära"transformerkihis juhusliku tüki enkooder-kihi tunnuseid, proovib seda tükki hiljem taastada.

### 5.3.3 XLSR-53 mudel

XLSR-53 baasmudel [34] on Wav2vec2 arhitektuuril treenitud 53 keelel mudel, mis kasutab CommonVoice, LibriSpeech ja BABEL treeningandmestikke 56 tuhande tunni ulatuses. Mudel sisaldab 300 miljonit parameetrit.

Selles töös kasutatakse inglisekeelsete tekstihüpoteeside genereerimiseks peenhäälestatud XLSR-53 mudelit, mis on ületreenitud XLSR-53 baasmudel CommonVoice 6.1 inglise keelsel andmestikul.[4]

### 5.3.4 XLS-R 300M mudel

XLS-R 300M on samuti Wav2vec 2.0 arhitektuuril treenitud mudel 300 miljoni parameetriga[19]. XLS-R on eeltreenitud ligi 500 tuhande tunni kõnematerjaliga 128 erinevas keeles, sisaldades lisaks XLS-R 53 kasutatud andmetele ka VoxLingua107 ning VoxPopuli

[4] `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english`

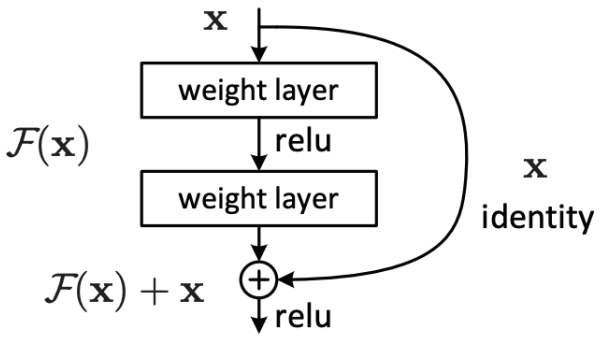
(VP-400K) korpusi. Mudelit on võimalik pärast peenhäälestamist kasutada nii kõnetuvastuseks, kõne tõlkeks, kõnelejatuvastuseks kui ka kõneldava keele identifitseerimiseks. XLS-R 300M on näitab võrdlustulemustes kõneldava keele identifitseerimises kui ka teistes ülesannetes paremaid tulemusi kui XLSR-53.

Antud töös kasutatakse kahel viisil. Esiteks, treeniti XLS-R 300M mudel akustilise mudeli võrdlusbaasiks kõneldava keele identifitseerimisel. Selleks loodi Speechbraini retsept ja lähtekood ning treeniti XLS-R 300M VoxLingua107 andmestikul[5][6].

Lisaks kasutati 800 tunnil eestikeelsel (peamiselt rahvusringhäälingu) materjalil peenhäälestatud XLS-R 300M mudelit tekstihüpoteeside genereerimiseks[7] sisendiks NB mudelile.

### 5.3.5 Jääkühikutega närvivõrgud

Jääkühikutega närvivõrk (*residual neural network*, lühidalt ResNet) on konvolutsiooniline süvaõppe närvivõrk, mille põhiomadusteks on sisendi liitmine väljundi tulemusele ning väljundi arvutamisel kihtide vahelejätmine *(skip)* nii, et vältida paljude konvolutsiooniliste kihtide lisamisega tekkiva kaduva gradiendi probleemi *(vanishing gradient problem)*, kus tagasilevil arvutatav gradiendi väärtus langeb treenimisel kiiresti nulli lähedale ning seejärel ei uuendata järgmisel iteratsioonil arvutatavaid mudeli kaalusid.



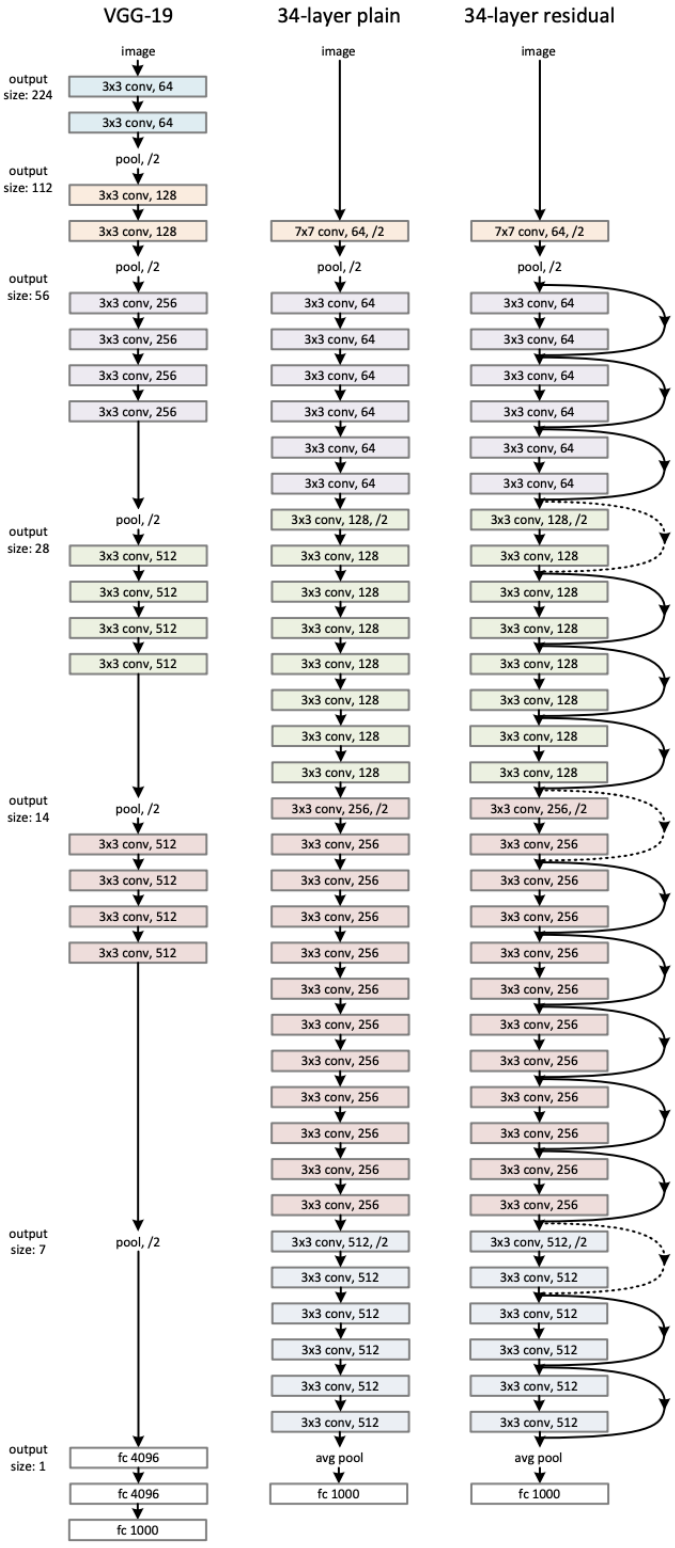

Joonis 3. *ResNet mudeli arhitektuur[35].*

ResNeti ühikuid defineeritakse:

$$y = F(x, \{W_i\}) + x.$$

kus $x$ ja $y$ on konvolutsiooniliste kihtide sisend- ja väljundvektorid ning funktsioon

[5] https://huggingface.co/TalTechNLP/voxlingua107-xls-r-300m-wav2vec
[6] https://github.com/kunnark/speechbrain/tree/vox107-xls-r-300m/recipes/VoxLingua107
[7] https://huggingface.co/TalTechNLP/xls-r-300m-et




Joonis 4. *ResNet34 mudeli arhitektuuri näide võrrelduna tavalise 34-kihilise konvolutsioonilise mudeli arhitektuuriga[35].*

$F(x, \{W_i\})$ esindab õpitavat jäägi-kaart *(identity mapping)*[35].

Töös kasutatav ResNet34 mudel on 34-kihiline mudel (joonis 4), kus algsed konvolutsioonilised blokid jääk ühikutega on asendatud tihenda-ja-tähelepanu moodulitega. Ajutiseks ahenduskihiks *(pooling layer)* kasutatakse mitmepealist tähelepanu-mehhanismi [21].

### 5.3.6 Konvolutsiooniline närvivõrk kõnetuvastushüpoteesidel

Enne, kui konvolutsioonilised närvivõrgud (*convolutional neural network* - CNN) leidsid efektiivset rakendust kõnetuvastuses, kasutati neid edukalt objektituvastuses [36]. Konvolutsioonilised närvivõrgud kasutavad *konvolutsiooni-* ning *ahenduskihte* [36]. Selleks, et tuvastada kõnest mustreid, on vajalik konvolutsioonilisse närvivõrku sööta ette tunnustekaart*(feature maps)*.

CNN arhitektuur koosneb kihtidepaaridest: konvolutsioonilisest kihist ja ahenduskihist.

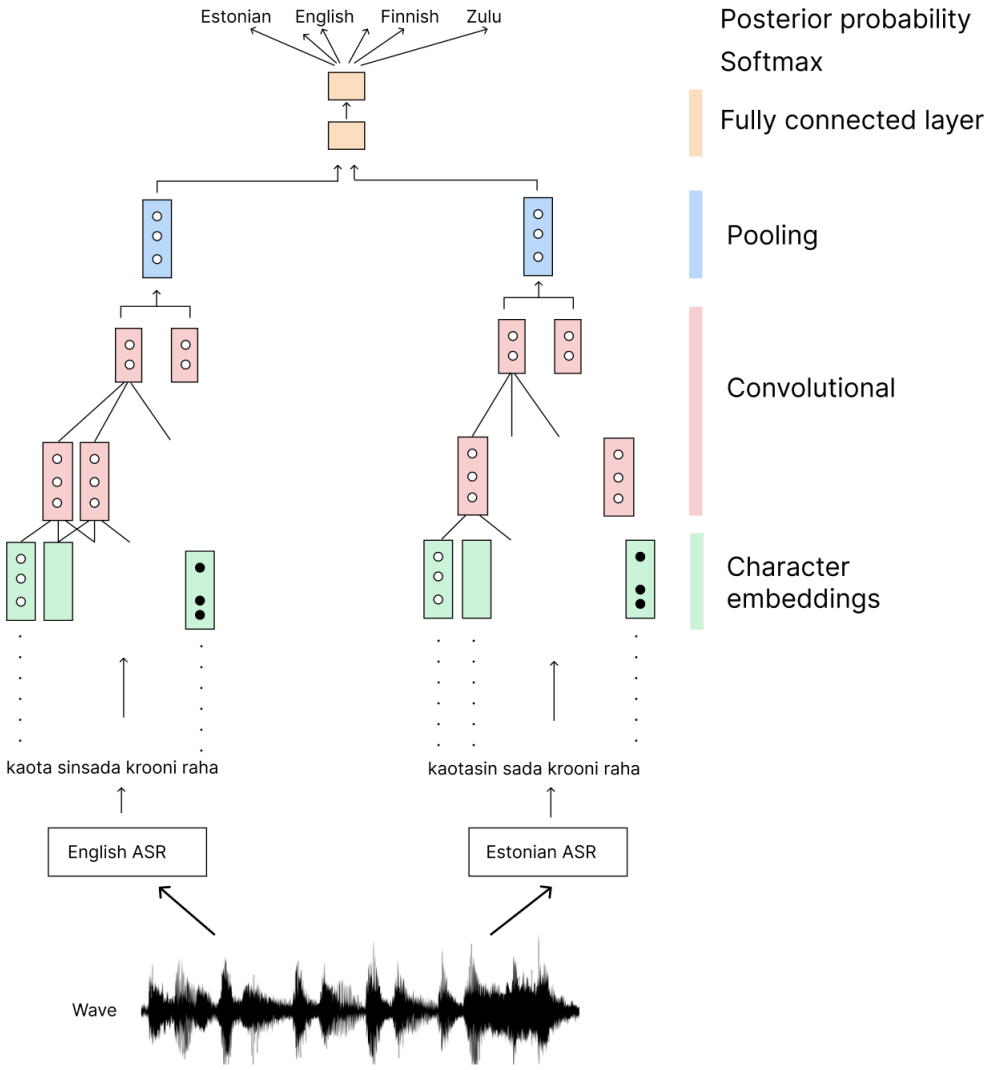
Konvolutsioniline kiht koosneb sisend ja väljund tunnustekaartidest, mida seob kaaludemaatriks. Konvolutsioonilist kihti võib vaadelda ka kui filtrit, mis määrab sagedusvahemike arvu igale sisendtunnustekaartile [36].

Konvolutsiooniline kiht leiab sisendandmetest spetsiifilisi mustreid, iga konvolutsioonioperatsiooni väljundiks on omakorda tunnuskaart. Konvolutsioonilisele kihile järgneb ahenduskiht, millega vähendatakse tunnuskaartide dimensionaalsust*(downsampling)*, teisiti võib seda mõista kui konvolutsioonilise kihi tunnuste üldisemale kujule viimist [36]. Tunnuskaartide suurus (resolutsioon) kahaneb ülemistes kihtides väiksemaks, kui tunnustele rakendatakse järjest enam konvolutsiooni- ja ahendamisoperatsioone [36].

Tavaliselt lisatakse lõpliku CNN-kihi peale üks või mitu täielikult ühendatud varjatud kihti, et ühendada tunnused kõigis sagedusribades enne väljundkihti ette söötmist [36].

Alternatiivina NB-põhisele tekstiklassifikatsiooni mudelile viisin läbi katsetuse konvolutsioonilise närvivõrguga (ConvNet).

Kavandatud ConvNet koosneb mitmest paralleelsest konvolutsioonilisest sisendharust, mille iga protsess töötleb ASR-i transkriptsiooni, mis omakorda on loodud konkreetse ASR-mudeli abil. Konvolutsiooniliste sisendharude väljundid ahendatakse lausungitel, kasutades max-poolingut, seejärel ühendatakse ja töödeldakse edasi, kasutades kahte täielikult ühendatud kihti *(fully connected layers)*. Mudel on treenitud, kasutades ristentroopia *(cross-entropy)* kaofunktsiooni.




Joonis 5. *Konvolutsioonilise närvivõrgu arhitektuur kõneldava keele identifitseerimisel.*

Konvolutsioonilised harud vastendavad kõigepealt tähemärgid õpitud 20-mõõtmelise vektoresitusega *(embeddings)* ja seejärel rakendavad vektoresitusele rea konvolutsioonilisi kihte.

Katsetes on kasutatud viit 1D-konvolutsioonilist kihti, mille tuumasuurus *(kernel size)* oli vastavalt (3, 1, 3, 1) ja kanalite arvuks määrati 512. Sarnaselt akustilistele LID mudelitele ei rakendata ConvNeti mudelit otseselt ennustamiseks, vaid kasutatakse 512-dimensioonilise vektoresituse eraldamiseks (esimese täielikult ühendatud kihi väljundist, mis tuleb pärast ahendamist). Lõpuks töödeldakse vektoresitusi LDA/PLDA mudeli abil.

### 5.3.7 Lineaarne diskriminantanalüüs ja tõenäosuslik lineaarne diskriminantanalüüs

Lineaarne diskriminantanalüüs (*Linear Discriminant Analysis* - LDA) on meetod uuritavate tunnuste dimensionaalsuse vähendamiseks, mis sobib klassifikatsiooniprobleemide lahendamiseks. LDA tuvastab alamruumi, milles eri klasside vahelised andmed on kõige rohkem hajutatud, võrreldes klassisisese levikuga igas klassis [37]. LDA puhul eristatakse klasse võrdse kaaluga, mis teeb keeruliselt lahendatavaks olukorra, kus mudel näeb uue klassi esindajat, mida senises treeningandmestikus ei ole ning mille tunnuste puhul mõned tunnused on olulisemad ja teised vähemolulisemad. Samuti on LDA meetodiga keeruline lahendada väiksema dimensionaalsusega tunnuste puhul probleemi, millised on olulisemad ja millised vähemolulisemad tunnused. Artiklis [37] on sellise olukorra tekkimisel, on seni kasutatud peakomponentanalüüsi (PCA) ning seejärel olulisemad tunnused viidud LDA alamruumi.

Selleks, et anda tunnustele klasside vahelistes erisustes kaalud, kasutatakse tõenäosuslikku lähenemist ehk tõenäosuslikku lineaardiskriminantanalüüsi (PLDA). PLDA annab suurema kaalu tunnustele, mis eristavad kahte klassi enam ja väiksema kaalu vastupidselt tunnustele, mis eraldavad kahte klassi vähem. PLDA erineb LDA-st, kuna parim võimalik dimensioonide arv valitakse välja automaatselt, andes vähem kaalu vähem eristavatele tunnustele. PLDA peamine eelis peitub võimaluses ennustusi teha klassidele, mis ei ole esindatud treeningandmestikus [37]. PLDA eeliseks LDA ees on veel ka see, et tõenäosusjaotusi õpitakse mitte üksnes klasside siseselt vaid arvestades ka klassi tsentreid [37].

## 5.4 Treeningute üksikasjad

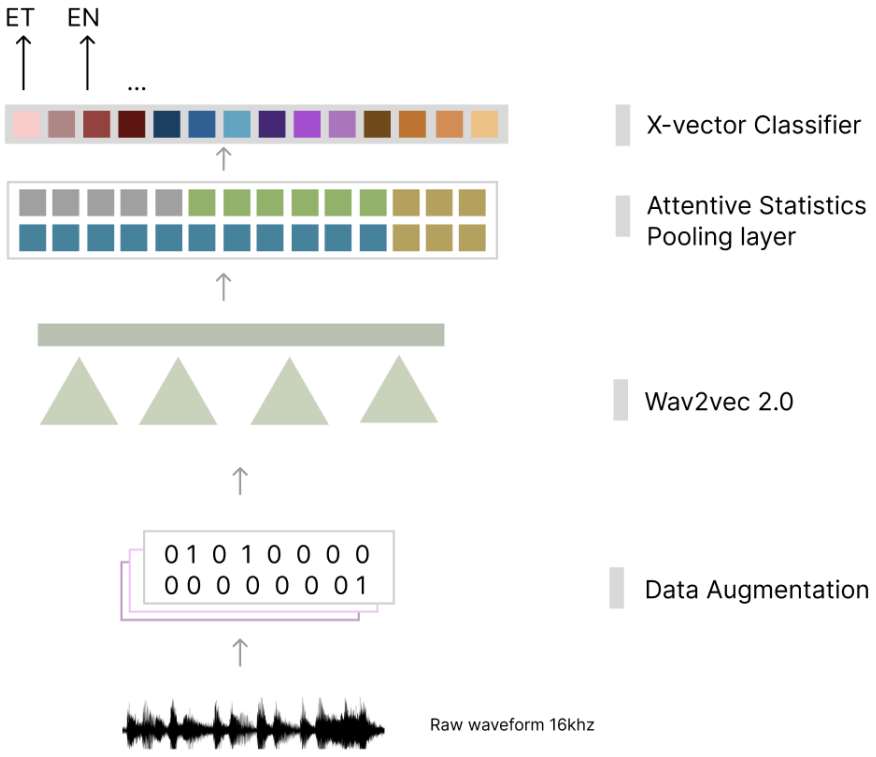
### 5.4.1 Multinomiaalne Naive Bayes

Siinses töös kasutatav NB mudel ennustab lausungi kõneldavat keelt, võttes sisendiks kõnetuvastuse transkriptsiooni sõnasiseseid 4-grammid. Sõnade puhul on välja jäetud sõnade äärised $(paddings)$ ja asendatud tühikutega. Mudel võtab arvesse kõiki n-gramme, mis esinevad treeningandmetes, et eemaldada nulle, funktsiooni silumiseks (nullvektorite asendamiseks) kasutatakse Laplace'i silumist parameetriga 0.95. Autor katsetas ka erinevate n-grammi pikkustega (2-5), parima tulemuse saavutas 4-grammidega.

ASR transkriptsioonide genereerimiseks kasutatakse kahte mudelit: inglise- ja eestikeelset. Mõlemad mudelid on peenhäälestatud mitmekeelsete Wav2vec 2.0 mudelitega, kasutades CTC *(Connectionist Temporal Classification)* kaofunktsiooni. Inglise keele mudel on

peenhäälestatud XLSR-53K mudelit kasutades, mis kasutab Mozilla CommonVoice inglise keele andmeid. Eestikeelne mudel on peenhäälestatud, kasutades XLS-R-300M mudelit, mis kasutab umbes 800 tundi mitmekesist eestikeelset kõnet (peamiselt eetris olev kõne). Kumbki mudel ei kasuta dekodeerimise ajal välist keelemudelit ja mõlemad kasutavad tähepõhiseid sõnastikke. Sellel on kaks eelist: esiteks on dekodeerimine GPU abil väga kiire, mistõttu on võimalik dekodeerida kogu VoxLingua107 6628 tundi andmeid, et genereerida treeningandmeid. Teiseks ei ole leksikonivaba kõnetuvastus-süsteemi väljund piiratud sõnavaraga, mille tulemuseks on väga väljendusrikkad ASR-transkriptsioonid muude keelte kui kõnetuvastuse sihtkeele jaoks.

### 5.4.2 XLS-R 300M

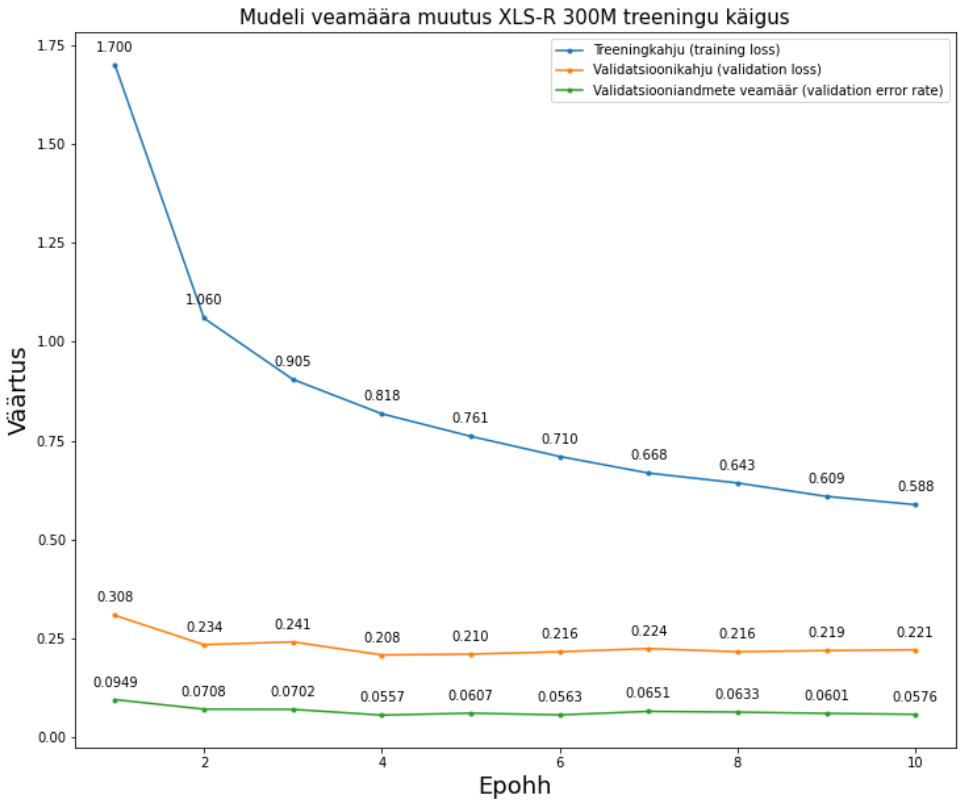



Joonis 6. *Wav2vec 2.0 kõneldava keele klassifitseerija arhitektuur.*

Mudeli treenimiseks lõi autor Speechbrainil põhineva treeningskripti[8] ja retsepti[9], mis on

[8] https://github.com/kunnark/speechbrain/blob/vox107-xls-r-300m/recipes/VoxLingua107/lang_id/train_wav2vec.py
[9] https://github.com/kunnark/speechbrain/blob/vox107-xls-r-300m/recipes/VoxLingua107/lang_id/hparams/train_wav2vec.yaml

saadaval Githubi repositooriumis. Mudel treeniti VoxLingua107 andmestikul 10 epohhi. Treeningu kestuvseks oli 60h TTÜ kõrgarvutusklastris, kasutades 2x NVIDIA A100 GPUd.




Joonis 7. *XLS-R 300M mudeli veamäära muutus treeningu jooksul.*

Treeningul kasutatav Wav2vec2 õpisamm *(learning rate)* on 100x väiksem võrreldes mudeli algse õpisammuga (0.001 vs 0.00001). Joonisel 7 on näha mudeli validatsiooniandmestiku veamäära muutust ning kahjufunktsioonide muutust treeningu jooksul. Mudeli treenimist ei peatatud ennatlikult või näidiku peale, treeniti 10 epohhi.

# 6. Tulemused ja järeldused

Järgnevalt on esitatud mudelite treeningu tulemused testandmetel ning nende analüüs.

Tabel 3. *Erinevatel mudelitel kõneldava keele identifitseerimine üle kõigi eesti- ja inglisekeelsete testandmestike.*

| | | Inglise keel | | | | Eesti keel | | Mitmekeelne |
|---|---|---|---|---|---|---|---|---|
| | | CMU Arctic | L2 Arctic | CSLU FAE | CSLU 22 en | EFAC | EFAC | V107 dev |
| ID | Mudel | Emakeel | Aktsent | Aktsent | Emakeel | Emakeel | Aktsent | Emakeel |
| A | ResNet | 77.4 | 60.5 | 67.1 | 57.7 | 93.3 | 43.9 | 91.9 |
| B | XLS-R 300M | 87.6 | 74.6 | 79.5 | 71.9 | **99.6** | 51.8 | **95.3** |
| D | NB en ASR 4-grammidel | 83.8 | 79.8 | 84.7 | 57.1 | 20.2 | 21.0 | 54.6 |
| E | NB et ASR 4-grammidel | 81.5 | 74.6 | 45.7 | 34.3 | 71.0 | 65.3 | 48.7 |
| F | Liidetud D, E | 91.9 | 88.2 | 81.7 | 58.7 | 68.2 | 62.8 | 58.5 |
| G | LDA+PLDA F-i log probs | 90.1 | 86.0 | 83.4 | 53.8 | 69.2 | 63.7 | 75.3 |
| H | ConvNet et+en ASR väljundil | 85.7 | 81.5 | 86.9 | 46.8 | 56.2 | 50.7 | 71.4 |
| I | Liidetud B, G | **95.5** | **90.5** | **88.2** | **72.6** | 99.5 | **69.5** | **95.3** |

Tabelis 3 on kirjeldatud treeningute tulemused. Tulemuste analüüsil on näha, et akustiline isejuhendatud Wav2vec2 arhitektuuril tippmudel XLS-R 300M (B) saab hästi hakkama ühekeelsetel korpustel emakeelse kõnest keele tuvastamisega (EFAC 99.6%), kuid jääb hätta aktsendiga kõnega (EFAC 51.8%). Mudel ületab kõikidel korpustel ResNeti põhist mudelit ning on väiksema veamääraga VoxLingua107 testandmestikul kui senine tippmudel (95.3% vs 94.6%). Võrreldes eesti ja inglise keelt, on XLS-R 300M tunduvalt edukam inglise keelse aktsendiga korpuste tuvastamisel, kuid emakeelse puhul jääb õigsus Aktsendikorpusel tuvastatud õigsusele alla.

Esmapilgul on see üllatav, et inglisekeelse CMU Arctic andmestiku õigsus on palju madalam kui õigsus Aktsendikorpuse emakeelsel osal. Lähemal uurimisel selgub, et CMU Arcticu õigsus erinevate rääkijate lõikes on väga erinev alates 35%-st hindi-pärase aktsendiga rääkijast puhul kuni 100%-ni rääkijani, kel ei ole märgitud häälduseripärasid. See näitab, et akustilisi tunnuseid kasutavad LID mudelid ei jää hätta mitte ainult võõrkeelse kõnega, vaid ka kohalike piirkondlike eripäradega.

Multinomiaalse Naive Bayes'i mudelite D ja E puhul (mis kasutavad kõnetuvastuseks peenhäälestatud XLS-R 53k ja XLS-R 300M mudeleid kõnetuvastuseks ja tekstihüpoteeside loomiseks) on näha märgatav paranemine ühekeelsete testandmestike korral, nii aktsendiga

kui ka emakeelse kõne puhul. Oodatult jäävad ükskeelsed mudelit hätta VoxLingua107 test-andmestikuga.

Mudelid D-H põhinevad kõik ASR-i transkriptsioonidel. Võrreldes mudeleid D (Naive Bayesi mudel, mis kasutab 4-gramme ingliskeelse mudeli abil loodud transkriptsioonidest) ja E (sama, aga eestikeelseid näidistranskriptsioone kasutades) on näha, et sihtkeele kõnetuvastuse tekstihüpoteeside olemasolu on kõneldava keele tuvastuse jaoks abiks rohkem kui teised ASR-i tekstihüpoteesid: nt Eesti ASR-i väljundil treenitud mudeli õigsus (mudel D) saavutab 71% õigsuse eesti keeles, langedes ingliskeelsete ASR-i tekstihüpoteeside kasutamisel vaid 20%-ni.

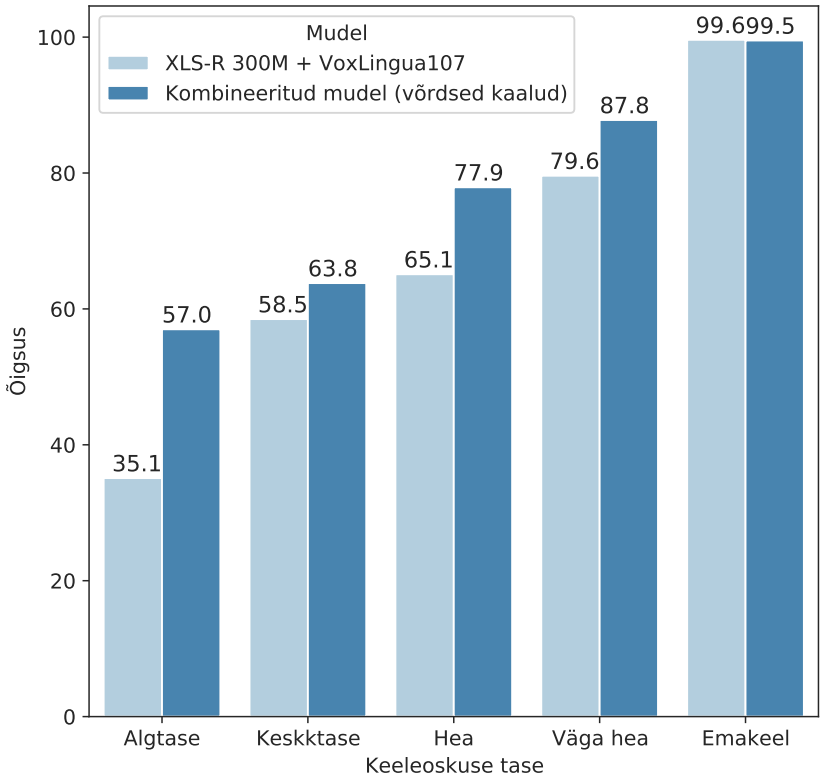
Mudelis F on liidetud mudelid D ja E, mis parandab tulemusi aktsendiga kõne korral ja samuti ka emakeelsest kõnest aktsendi tuvatamisega inglise keelel, samuti on tulemused paremad mitmekeelse mudeli korral.

Üksikute NB-mudelite liitmisel lineaarse interpolatsiooni abil suurendab LID-i jõudlust kõigi testandmestike jaoks. Mudeli G puhul, mis kasutab klassifitseerimisel LDA/PLDA posterior-tõenäosusi sisendina mudelist F, paraneb aktsendiga kõnetuvastus veelgi nii eesti kui inglise keele puhul. Samuti paraneb tugevalt ka mitmekeelse VoxLingua107 testandmestiku tulemus võrreldes mudeliga F. Üllataval kombel ei suuda ASR-i transkriptsioonidel treenitud ConvNet ületada NB mudeleid.

Kõige paremaid tulemusi pea kõikidel andmestikel näitab mudel I, mis kasutab omakorda mudeli G ja akustilise XLS-R 300M mudeli B liitmist. Mudel I kasutab kahe mudeli liitmisel võrdseid kaalusid. Mudeli hästi toimimise jaoks on oluline mitte optimeerida liidetud kaalusid emakeelsel kõnel, kuna akustilised esitused annavad kõrgema õigsuse emakeelsele kõnele ja liidetud mudel võib seetõttu taanduda üksnes akustiliseks mudeliks. See mudel tuvastab paremini nii aktsendiga ühekeelseid korpuseid kui ka emakeelseid ühekeelseid korpuseid ning näitab samaväärset tulemust mitmekeelsel VoxLingua107 tulemusel.

CSLU 22 EN andmestiku suhteliselt madalamaid tulemusi emakeelsel kõnetuvastusel võib selgitada sellega, et tegemist on telefonikõnedega, kus on palju alla sekundi pikkuseid lausungeid, mis mõjutab kõneldava keele tuvastuse kvaliteeti.

Joonis 8 võrdleb akustilise mudeli (B) õigsust liitmudeli (I) õigsusega Aktsendikorpuse andmetel, kasutades kõnelejate grupeerimiseks katsealuste hinnangut enda keeleoskusele. Joonis kinnitab, et keeleoskuse taseme ja LID-õigsuse vahel on tugev pöördkorrelatsioon. ASR-i tekstihüpoteeside kasutamine lisaparameetrina parandab LID tulemusi kõikides

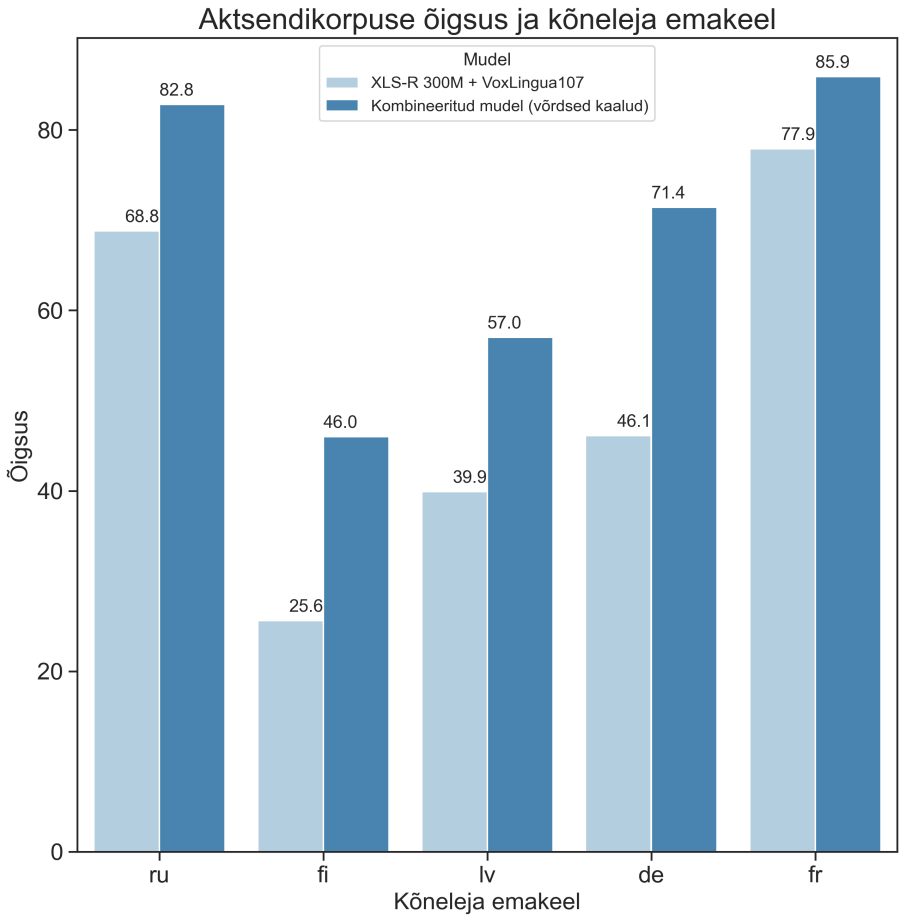



Joonis 8. *Keeleoskuse tase ning kõneldava keele tuvastuse õigsus Aktsendikorpuse 5-l enim esineval keelel.*

keeleoskuse tasemetes. Siiski on emakeele ja aktsendiga kõne LID õigsuse vahel endiselt märgatav lõhe, isegi väga heal tasemel kõnelejate ja emakeele kõnelejate vahel.

Joonis 9 näitab, et kõneleja emakeelest sõltuvalt on õigsuse määrades erinevused keelte vahel, kui tuvastatakse kõneldavat keelt. Erinevus on suurem soome keele puhul ja väiksem vene ning prantsue keele puhul. Saksa ja soome keelte puhul on näha, et kombineeritud LDA/PLDA mudeli kasutamine on andnud ka kõige suurema paranemise võrreldes akustilise mudeliga.

Täiendavalt vaatasin Aktsendikorpusel, mil määral mudelid ennustavad kõneldavaks keeleks kõneleja emakeelt. Soome keele puhul ilmnes, et XLS-R 300M mudel ennustab 73% soome keele juhtudest soome keeleks eesti keele asemel, kombineeritud lähenemisel on see protsent 53%. Järgnes läti keel (39% vs 19%), teiste keelte puhul oli see näitaja madalam. Sellest võib järeldada, et kuna soome keele puhul on tegu eesti keelele lähedase keelega, siis akustiline mudel ei suuda hästi eristada lähedasi keeli aktsendi puhul, tekstihüpoteeside lisamisel probleem leeveneb, kuid võib öelda, et lähedaste keelte puhul on aktsendist




Joonis 9. *Kõneleja emakeel ja kõneldava keele tuvastamise õigsus Aktsendikorpusel 5-l korpuses enam esineval keelel.*

kõneldava keele tuvastamine raskendatud ja vajaks edasiuurimist.

Kõige väiksem paranemine tekstihüpoteeside lisamisel on toimunud prantsuse keele puhul. Vahed kõikide keelte puhul on siiski märgatavad, millest võib järeldada, et tekstihüpoteeside lisamine aitab tugevalt kaasa kõneldava keele tuvastamisele.

Tabelist 3 järeldub seega, et parimaid tulemusi annab kombineeritud tekstihüpoteesidel 4-grammide kasutamine, LDA/PLDA dimensioonide vähendamisega klassifitseerimisel liidetuna senise isejuhendatud tippmudeli XLS-R 300M-ga. Aktsendiga kõnest keele tuvastus on siiski võrreldes emakeelse kõnega oluliselt madalam (Aktsendikorpusel 99.5% vs 69.5%).

# 7. Kokkuvõte

Käesoleva töö eesmärk oli analüüsida seniste akustikal põhinevate tippmudelite õigsust kõneldava keele tuvastamise aktsendiga kõne puhul ning seejärel pakkuda uusi lähenemisi aktsendiga kõnest kõneldava keele tuvastamiseks. Töö põhineb autori samateemalisel uurimistööl ja teadusartiklil kahasse juhendajaga, mis avaldati septembris 2022 kõnetehnoloogia konverentsil Interspeech 2022.

Aktsendiga kõnest kõneldava keele tuvastamise teemal on seniajani autori hinnangul väga vähe uurimusi ilmunud. Ainus teemakohane artikkel on enam kui 20 aastat vana ning adresseerib küll probleemi, kuid tänaseks on toonased meetodid aegunud kui ka uued korpused arenenud. Küll aga on teaduskirjanduses palju uuritud aktsendi tuvastamist.

Töö näitas esmalt, et akustikal põhinevad tippmudelid suudavad kõrge õigsusega kõneldavat keelt identifitseerida aktsendita kõne puhul, ent näitavad oluliselt nõrgemaid tulemusi aktsendiga kõne puhul.

Teiseks, töö pakub välja keelteülese lähenemise aktsendiga kõnest kõneldava keele tuvastamiseks. Autor leidis mudelite võrdlemisel, et parimaks lähenemiseks kõneldava keele tuvastamiseks tuleks kombineerida isejuhendatud õppel Wav2vec 2.0 arhitektuuril akustilist XLS-R 300M tippmudelit LDA/PLDA klassifitseerijat Multinomiaalse Naive Bayesi tekstiklassifitseerimismudelidega, mis kasutab sisendina ja lisatunnustena kõnetuvastushüpoteese mitmest erinevat keelt tuvastavast kõnetuvastajast.

Töö tulemusel vähenes suhteline veamäär kõneldava keele tuvastamisel 35-63% võrra. Antud lähenemine ei ole aga kõneldava keele tuvastamisel täiesti uus idee, kuid on olnud seni olnud rakendamata aktsendiga kõne puhul. Loodud ja testitud mudelite puhul ei tulene aktsendiga kõnest keele tuvastamine emakeele arvelt ega tulemuse parandamiseks pole lisatud täiendavaid andmeid. Mudelite treenimiseks on kasutatud TTÜs valminud mitmekeelset VoxLingua107 andmestikku ning valideerimiseks omavahel võrreldavaid eesti ja inglisekeelseid testandmestikke koos VoxLingua107 testandmestikuga. Töö tulemus on uudne ning on rakendatav kõnetuvastusrakendustes esimese sammuna kõneldava keele tuvastamisel.

Edasiseks uurimistööks tuleks esmalt senisest põhjalikumalt uurida keeltevahelisi erinevusi aktsendiga kõnest kõneldava keele tuvastamisel ning otsida võimalusi, kuidas parandada

omavahel lähedaste keelte puhul (kui kõneleja emakeel on kõneldavale keelele lähedane keel) kõneldava keele tuvastamist.

# Kasutatud kirjanduse loetelu


[1] *Statistics explained*. Accessed: 2022-11-04. Aprill 2019. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php/Foreign_language_skills_statistics.

[2] Enes Furkan Çiğdem, Ali Haznedaroğlu ja Levent M. Arslan. „Spoken Language Identification Using Call Center Data“. Teoses: *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*. 2021, lk. 1–4. DOI: 10.1109/ASYU52992.2021.9599040.

[3] Pierre Berjon, Avishek Nag ja Soumyabrata Dev. *Analysis of French Phonetic Idiosyncrasies for Accent Recognition*. 2021. DOI: 10.48550/ARXIV.2110.09179. URL: https://arxiv.org/abs/2110.09179.

[4] Ming Tu *et al. Investigating the role of L1 in automatic pronunciation evaluation of L2 speech*. 2018. DOI: 10.48550/ARXIV.1807.01738. URL: https://arxiv.org/abs/1807.01738.

[5] Siyuan Feng *et al.* „Quantifying bias in automatic speech recognition“. *ArXiv preprint* abs/2103.15122 (2021). URL: https://arxiv.org/abs/2103.15122.

[6] Yunhan Wu *et al.* „See what I'm saying? Comparing intelligent personal assistant use for native and non-native language speakers“. Teoses: *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. 2020, lk. 1–9.

[7] Abhijeet Awasthi *et al.* „Error-driven Fixed-Budget ASR Personalization for Accented Speakers“. Teoses: *ICASSP*. 2021, lk. 7033–7037.

[8] R Wanneroy *et al. Acoustic-phonetic modeling of non-native speech for language identification*. Tehniline raport. Multi-Lingual Interoperability in Speech Technology, 2000.

[9] Shengye Wang *et al.* „Signal combination for language identification“. *ArXiv preprint* abs/1910.09687 (2019). URL: https://arxiv.org/abs/1910.09687.

[10] Chander Chandak *et al.* „Streaming language identification using combination of acoustic representations and ASR hypotheses“. *ArXiv preprint* abs/2006.00703 (2020). URL: https://arxiv.org/abs/2006.00703.

[11] Kunnar Kukk ja Tanel Alumäe. *Improving Language Identification of Accented Speech*. 2022. DOI: 10.48550/ARXIV.2203.16972. URL: https://arxiv.org/abs/2203.16972.

[12] Tanel Alumäe ja Kunnar Kukk. „Pretraining Approaches for Spoken Language Recognition: TalTech Submission to the OLR 2021 Challenge“. Teoses: *Speaker Odyssey*. 2022.

[13] Roman Bedyakin ja Nikolay Mikhaylovskiy. *Low-Resource Spoken Language Identification Using Self-Attentive Pooling and Deep 1D Time-Channel Separable Convolutions*. 2021. DOI: 10.48550/ARXIV.2106.00052. URL: https://arxiv.org/abs/2106.00052.

[14] Christian Bartz *et al. Language Identification Using Deep Convolutional Recurrent Neural Networks*. 2017. DOI: 10.48550/ARXIV.1708.04811. URL: https://arxiv.org/abs/1708.04811.

[15] L Meister ja E Meister. „Aktsendikorpus ja võõrkeele aktsendi uurimine“. *Keel ja Kirjandus* 55.8-9 (2012), lk. 696–714.

[16] Yuanyuan Zhang. „Mitigating bias against non-native accents“. Magistritöö. Delft University of Technology, 2022. URL: http://resolver.tudelft.nl/uuid:bc989a6e-60b5-4cff-bb7f-999c616afc7c.

[17] Zhiyun Fan *et al. Exploring wav2vec 2.0 on speaker verification and language identification*. 2020. DOI: 10.48550/ARXIV.2012.06185. URL: https://arxiv.org/abs/2012.06185.

[18] Andros Tjandra *et al. Improved Language Identification Through Cross-Lingual Self-Supervised Learning*. 2021. DOI: 10.48550/ARXIV.2107.04082. URL: https://arxiv.org/abs/2107.04082.

[19] Arun Babu *et al.* „XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale“. *ArXiv preprint* abs/2111.09296 (2021). URL: https://arxiv.org/abs/2111.09296.

[20] Rosana Ardila *et al.* „Common Voice: A Massively-Multilingual Speech Corpus“. English. Teoses: *LREC*. Marseille, France, 2020, lk. 4218–4222. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.520.

[21] Jörgen Valk ja Tanel Alumäe. „VoxLingua107: a Dataset for Spoken Language Recognition“. Teoses: *Proc. IEEE SLT Workshop*. 2021.

[22] *Estonian Foreign Accent Corpus*. https://doi.org/10.15155/9-00-0000-0000-0000-0002BL. Accessed: 2022-03-21.

[23] Terri Lander. *CSLU: Foreign Accent English Release 1.2 (LDC2007S08)*. Linguistic Data Consortium, 2007.

[24] Terri Lander. *CSLU: 22 Languages Corpus (LDC2005S26)*. Linguistic Data Consortium, 2005.

[25] John Kominek ja Alan W Black. *CMU Arctic Databases for Speech Synthesis*. Tehniline raport. Language Technologies Institute School of Computer Science Carnegie Mellon University, 2003.

[26] Eesti andmeteaduse kommuun. *Masinõppesõnastik*. http://datasci.ee/masinoppe-sonastik/. Accessed: 2023-01-02. 2018.

[27] Mirco Ravanelli *et al. SpeechBrain: A General-Purpose Speech Toolkit*. 2021. DOI: 10.48550/ARXIV.2106.04624. URL: https://arxiv.org/abs/2106.04624.

[28] *Hugging Face wants to be GitHub of machine learning: a look at its history, key members, and major achievements*. Accessed: 2022-10-10. Mai 2022. URL: https://www.ai.nl/artificial-intelligence/hugging-face-wants-to-be-github-of-machine-learning-a-look-at-its-history-key-members-and-major-achievements/.

[29] Thomas Wolf *et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing*. 2019. DOI: 10.48550/ARXIV.1910.03771. URL: https://arxiv.org/abs/1910.03771.

[30] Daniel Povey *et al.* „The Kaldi Speech Recognition Toolkit". Teoses: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Catalog No.: CFP11SRW-USB. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society, detsember 2011.

[31] Adam Paszke *et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019. DOI: 10.48550/ARXIV.1912.01703. URL: https://arxiv.org/abs/1912.01703.

[32] Prabhakar Raghavan Christopher D. Manning ja Hinrich Schütze. „Introduction to Information Retrieval". Teoses: Accessed: 2022-09-28. Cambridge University Press, 2008. Ptk Naive Bayes text classification. URL: https://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html.

[33] Alexei Baevski *et al. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. 2020. DOI: 10.48550/ARXIV.2006.11477. URL: https://arxiv.org/abs/2006.11477.

[34] Alexis Conneau *et al.* „Unsupervised Cross-Lingual Representation Learning for Speech Recognition". Teoses: *Interspeech*. 2021. DOI: 10.21437/Interspeech.2021-329. URL: https://doi.org/10.21437/Interspeech.2021-329.

[35] Kaiming He *et al. Deep Residual Learning for Image Recognition*. 2015. DOI: 10.48550/ARXIV.1512.03385. URL: https://arxiv.org/abs/1512.03385.

[36] Ossama Abdel-Hamid *et al.* „Convolutional Neural Networks for Speech Recognition“. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.10 (2014), lk. 1533–1545. DOI: 10.1109/TASLP.2014.2339736.

[37] Sergey Ioffe. „Probabilistic Linear Discriminant Analysis“. Teoses: *Computer Vision – ECCV 2006*. Toim. Aleš Leonardis, Horst Bischof ja Axel Pinz. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, lk. 531–542. ISBN: 978-3-540-33839-0.

# Lisad

## Lisa 1 – Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks[1]

Mina, Kunnar Kukk

1. Annan Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Kõneldava keele tuvastamine aktsendiga kõnest“, mille juhendaja on Tanel Alumäe.
   1.1. reprodutseerimiseks lõputöö säilitamise ja elektroonse avaldamise eesmärgil, sh Tallinna Tehnikaülikooli raamatukogu digikogusse lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
   1.2. üldsusele kättesaadavaks tegemiseks Tallinna Tehnikaülikooli veebikeskkonna kaudu, sealhulgas Tallinna Tehnikaülikooli raamatukogu digikogu kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. Olen teadlik, et käesoleva lihtlitsentsi punktis 1 nimetatud õigused jäävad alles ka autorile.
3. Kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest ning muudest õigusaktidest tulenevaid õigusi.

13.04.2023

[1] Lihtlitsents ei kehti juurdepääsupiirangu kehtivuse ajal vastavalt üliõpilase taotlusele lõputööle juurdepääsupiirangu kehtestamiseks, mis on allkirjastatud teaduskonna dekaani poolt, välja arvatud ülikooli õigus lõputööd reprodutseerida üksnes säilitamise eesmärgil. Kui lõputöö on loonud kaks või enam isikut oma ühise loomingulise tegevusega ning lõputöö kaas- või ühisautor(id) ei ole andnud lõputööd kaitsvale üliõpilasele kindlaksmääratud tähtajaks nõusolekut lõputöö reprodutseerimiseks ja avalikustamiseks vastavalt lihtlitsentsi punktidele 1.1. ja 1.2, siis lihtlitsents nimetatud tähtaja jooksul ei kehti.