TARTU ÜLIKOOL
Loodus- ja täppisteaduste valdkond
Arvutiteaduse instituut
Andmeteaduse õppekava

Kevin Lehtsalu

Kindlustusettevõtte kõnede automaatne transkriptsioon ja sentimendi analüüs

Magistritöö (15 EAP)

Juhendaja(d): Kairit Sirts, PhD

Tartu 2025

Kindlustusettevõtte kõnede automaatne transkriptsioon ja sentimendi analüüs

Lühikokkuvõte:

Automaatne kõnede transkriptsioon ja analüüs on loomuliku keele töötlemise valdkonnas kiiresti arenev suund, mis võimaldab ettevõtetel struktureerimata audioteabest väärtuslikku teavet saada. Käesolev magistritöö uurib, kuidas selliseid lahendusi oleks võimalik rakendada kindlustusvaldkonnas, kus salvestatud kliendikõned sisaldavad olulist teavet klientide vajaduste, teenuste kvaliteedi ja sisemiste tööprotsesside kohta. Kuigi automaatse töötluse vahendid on olemas, ei ole neid konkreetses organisatsioonis seni süstemaatiliselt rakendatud - kõnede sisu ja kvaliteeti on seni hinnatud käsitsi.

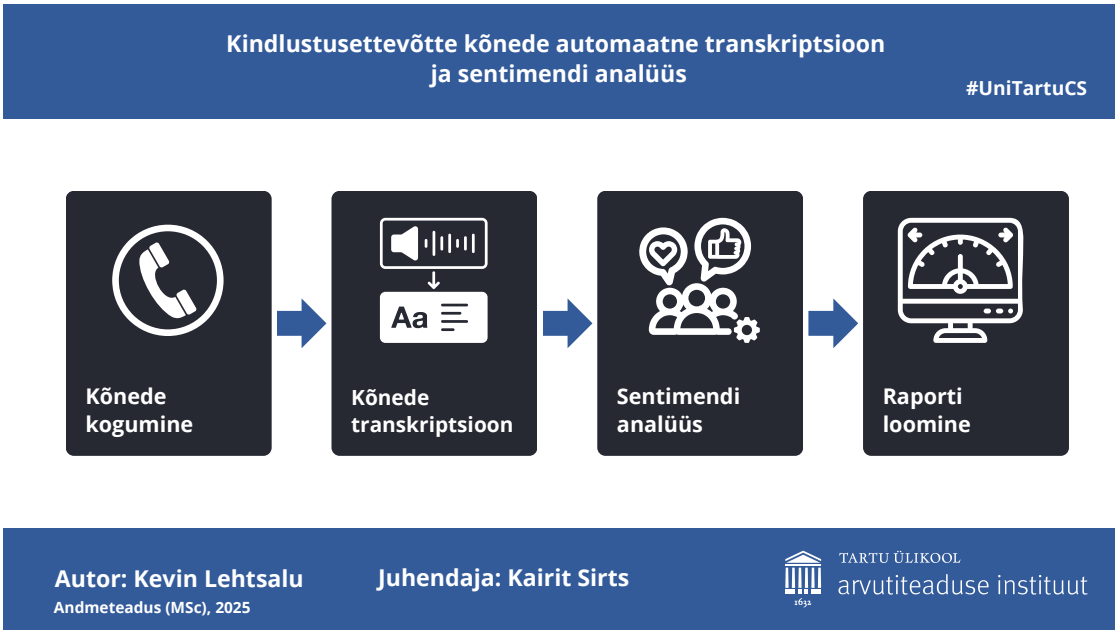
Käesolev magistritöö jaguneb kahte põhiossa. Töö praktiline osa keskendub esmalt erinevate automaatsete transkriptsioonimudelite (Whisper, Kaldi ja Wav2Vec 2.0) katsetamisele, et välja selgitada, milline neist sobib kõige paremini eesti keeles toimunud kindlustuskõnede töötlemiseks. Selles osas hinnatakse transkriptsioonide täpsust ja mudelite tehnilist rakendatavust, arvestades väikekeeltega seotud eripärasid nagu morfoloogiline keerukus ja piiratud treeningandmed.

Töö teises osas analüüsitakse transkribeeritud tekstide põhjal kõnede sentimenti. Selleks võrreldakse leksikonipõhiseid ja masinõppel põhinevaid meetodeid, eesmärgiga uurida, milliseid hinnanguid need annavad kindlustusvaldkonna kõnedele ning kuidas erinevate mudelite tulemused omavahel erinevad. Võrdlustulemused pakuvad esialgset ülevaadet meetodite käitumisest, kuid täpsema valideerimise jaoks on tulevikus vaja täiendavaid märgendatud andmeid. Selline analüüs on oluline loomaks alust süsteemile, mis võiks tulevikus aidata hinnata kliendi emotsionaalset hoiakut või rahulolu.

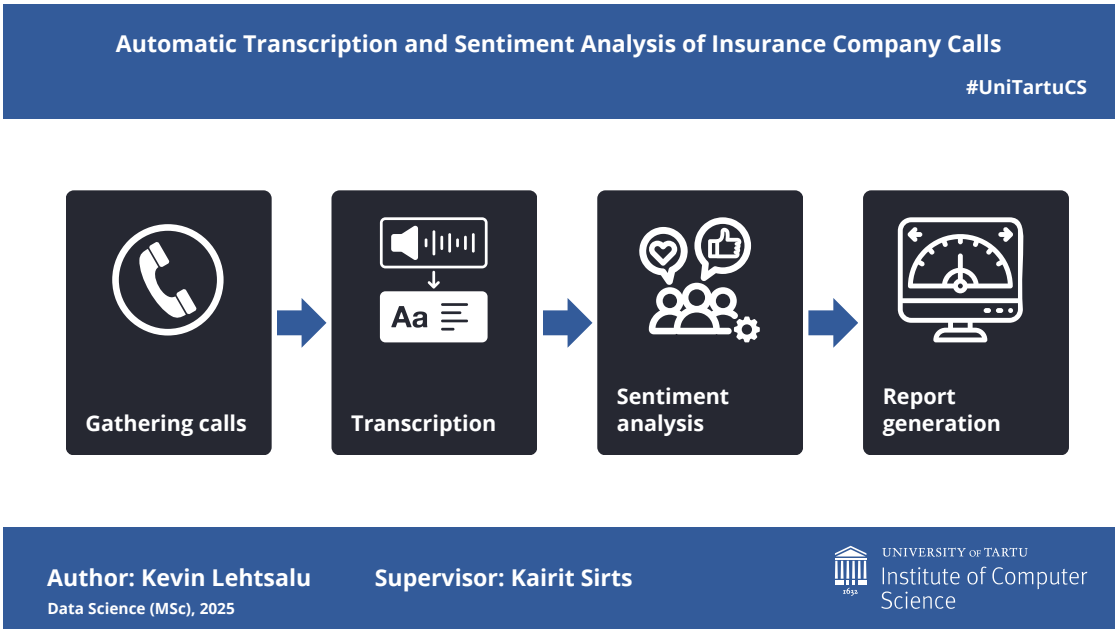
Töö tulemuste põhjal on võimalik anda soovitusi sobivaima transkriptsioonimudeli valikuks ning hinnata, millistel tingimustel võiks sentimentide automaatne tuvastamine pakkuda lisaväärtust. Edasiste sammudena saab ettevõte kaaluda automaatanalüüsil põhineva prototüübi loomist, mis toetaks salvestiste sisupõhist töötlemist ning võimaldaks tõhustada nii klienditeeninduse kvaliteedikontrolli kui ka andmehaldust.
Visuaalne abstrakt on kujutatud Joonisel 1

Võtmesõnad:
kõnede transkriptsioon, loomuliku keele töötlemine (NLP), sentimendianalüüs

CERCS koodid: P176 - Tehisintellekt, P170 – Arvutiteadus



Joonis 1. Visuaalne kokkuvõte magistritööst.



Joonis 2. Visual abstract.

Automatic Transcription and Sentiment Analysis of Insurance Company Calls

Abstract:
Automatic call transcription and analysis is a rapidly developing field within natural language processing, enabling organizations to extract valuable information from unstructured audio data. This master's thesis explores how such solutions could be applied in the insurance domain, where recorded customer calls contain important insights into client needs, service quality, and internal processes. Although tools for automatic processing exist, they have not been systematically implemented in the organization under study - the content of calls has so far been assessed manually.

In the first part of the thesis, various automatic transcription models (Whisper, Kaldi, and Wav2Vec 2.0) are tested to determine which performs best for processing insurance related calls in Estonian. The models are evaluated in terms of transcription accuracy and technical applicability, taking into account the specific challenges of low resource languages, such as morphological complexity and limited training data.

The second part focuses on sentiment analysis based on the transcribed texts. Both lexicon based and machine learning based methods are compared to assess their ability to detect customers emotional stance or satisfaction. Such information is valuable for improving customer experience and gathering meaningful feedback.

Based on the results, the thesis provides recommendations for selecting the most suitable transcription model and assesses under which conditions automatic sentiment detection may offer added value. As a next step, the organization could consider developing a prototype based on automatic analysis to support content-based processing of call recordings and improve both service quality monitoring and data management.
Visual abstract is depicted on Joonis 2

Keywords: call transcription, natural language processing (NLP), sentiment analysis

CERCS codes: P176 - Artificial intelligence, P170 – Computer science

Sisukord

Sissejuhatus **7**

1 Teoreetilised lähtekohad **9**
1.1 Kõnede transkriptsioon 9
1.1.1 Akustilised tunnused 10
1.2 Transkriptsioonimudelid 12
1.2.1 Whisper 12
1.2.2 Kaldi 12
1.2.3 Wav2Vec ja Wav2Vec 2.0 13
1.2.4 Võrdlus: Whisper, Kaldi ja Wav2Vec 14
1.3 Mudelite arhitektuurilised erinevused 15
1.4 Sentimendianalüüs 17
1.5 Andmekaitse ja privaatsuse kaalutlused 18
1.6 Balti keelte eripärad ja mudelite sobivus 19

2 Varasemad uuringud ja võrdlev analüüs **20**
2.1 Tööstuslikud lahendused Balti riikides 20
2.2 Semantilise analüüsi kvaliteet ASR-i baasil 21
2.3 Kokkuvõte 21

3 Metoodika **22**
3.1 Andmestik 22
3.2 Transkriptsioonimudelid 23
3.2.1 Whisper-põhised mudelid 24
3.2.2 Wav2Vec 2.0-põhine mudel 25
3.2.3 Kaldi-põhine transkriptsioon (TalTech EST-ASR-Pipeline) . . . 25
3.3 Mudelite võrdlusmetoodika 26
3.4 Sentimendi tööriistad 27
3.5 Rakenduskeskkond ja töövoog 29

4 Eksperimentaalsed tulemused **31**
4.1 Transkriptsioonimudelite täpsus 31
4.2 Töötlemiskiirus 33
4.3 Sentimendianalüüs 36
4.4 Kokkuvõte 38

5 Arutelu **39**
5.1 Tulemuste kriitiline hindamine 39
5.2 Eesti keele jaoks sobivaimad mudelid 39

5.3 Esinenud väljakutsed ja arendusvõimalused 39
5.4 Vastus uurimisküsimustele 40
5.5 Jätkutöö ja rakendusvõimalused 41

Viidatud kirjandus **44**

Lisad **45**
I. Glossary 45
II. Litsents 48

Sissejuhatus

Loomuliku keele töötlemine (ingl. *natural language processing*, NLP) on viimastel aastatel kiiresti arenenud, pakkudes uusi võimalusi struktureerimata andmete analüüsimiseks ja automatiseerimiseks (Hinton et al., 2012; Radford et al., 2022). Üheks oluliseks suunaks selles valdkonnas on automaatne kõnede transkriptsioon ja analüüs, mis võimaldab ettevõtetel ammutada väärtuslikku teavet helisalvestistest. Kindlustusettevõtetes sisaldavad kliendi- ja ettevõttesisesed kõned teavet klientide vajaduste, teenuste kvaliteedi ning tööprotsesside kohta, muutes need oluliseks andmeallikaks kliendisuhtluse ja teenuse parendamisel. Samas toob nende salvestiste haldamine kaasa mitmeid väljakutseid. Eelkõige vajadust hinnata salvestiste asjakohasust ning järgida andmekaitse ja isikuandmete töötlemise nõudeid (Nautsch et al., 2019; Weitzenboeck et al., 2022).

Automaatne kõnetuvastus ja sellele järgnev tekstianalüüs toimivad tänapäeval üsna täpselt sellistes keeltes nagu inglise või hispaania, mille kohta on olemas suured ja kvaliteetsed treeningandmestikud. Sellistele keeltele on arendatud mitmeid töökindlaid lahendusi, mis võimaldavad nii transkribeerimist kui ka sentimentide ja vestluse eesmärgi tuvastamist reaalajas.

Väiksemate keelte puhul, nagu eesti, läti ja leedu, on olukord keerulisem. Kuigi paljud tänapäevased mitmekeelsed mudelid toetavad tehniliselt ka neid keeli (Parikh et al., 2023), on transkriptsioonide täpsus sageli madalam. Selle põhjuseks on ühelt poolt piiratud treeningandmete kättesaadavus ja teiselt poolt keelte morfoloogiline keerukus ning häälduslikud variatsioonid, mis esitavad tehnoloogilistele lahendustele täiendavaid nõudmisi (Fedorchenko and Alumäe, 2025; Olev and Alumäe, 2022).

Samuti on transkribeeritud kõnede sisuline analüüs väikekeeltes piiratum ning seda on vähem uuritud. Näiteks sentimentide või eesmärgi (nt kas tegemist on müügi- või kahjukäsitluskõnega) automaatne tuvastamine ei ole sageli piisavalt täpne, et seda saaks kasutada automatiseeritud protsessides. Samas on just selline analüüs oluline kliendikogemuse parandamisel ja sisupõhise andmehalduspoliitika kujundamisel, sealhulgas näiteks mittevajalike salvestiste kustutamisel vastavalt andmekaitse põhimõtetele (European Parliamentary Research Service, 2020).

Käesoleva magistritöö eesmärk on katsetada erinevaid automaatseid kõnetuvastusmudeleid (Whisper, Kaldi ja Wav2Vec 2.0), et selgitada välja, milline neist sobib kõige paremini eestikeelsete kindlustusvaldkonna kõnesalvestiste transkribeerimiseks. Lisaks uuritakse, kuivõrd on võimalik transkribeeritud tekstide põhjal hinnata kõnede sentimenti, kasutades selleks nii leksikonipõhiseid kui ka masinõppel põhinevaid sentimendianalüüsi lähenemisi. Töö keskendub esmajoones eesti keelele, kuid metoodika on üles ehitatud viisil, mis võimaldab sarnast analüüsi tulevikus laiendada ka läti ja leedu keelele.

Töö keskendub järgmistele uurimisküsimustele:

- Milline automaatne transkriptsioonimudel sobib kõige paremini eestikeelsete kindlustuskõnede transkribeerimiseks?

- Milliseid hinnanguid annavad erinevad sentimendianalüüsi mudelid transkribeeritud kindlustusvaldkonna kõnedele ja kuivõrd erinevad need hinnangud üksteisest kontekstis, kus valideeritud andmestik puudub?

Küsimustele vastamiseks võrreldi erinevaid automaatseid kõnetuvastusmudeleid (Whisper, Wav2Vec2 ja Kaldi), kasutades eestikeelseid kindlustuskõnesid. Transkriptsioonide täpsust hinnati võrreldes neid käsitsi koostatud tekstidega. Seejärel tehti sentimentide analüüs, et näha, kas kõne oli pigem positiivne või negatiivne.

Kuigi töö algses plaanis oli hõlmatud ka kõnede eesmärgi tuvastamine, selgus uurimistöö käigus, et see vajab eraldi märgendatud andmestikku ja valideerimisraamistikku. Seetõttu keskendub käesolev töö kahele peamisele uurimisküsimusele: transkriptsioonimudelite võrdlusele ja sentimendianalüüsi võrdlevale uuringule. Eesmärgituvastuse võimalused jäävad tulevaste uuringute valdkonda.

Käesoleva töö koostamisel on vormistuslikul eesmärgil kasutatud toetavat tööriista ChatGPT 4o (OpenAI), mida on rakendatud teksti toimetamisel ja vormistamisel.

1 Teoreetilised lähtekohad

Käesolevas peatükis antakse ülevaade teoreetilistest ja tehnoloogilistest lähtekohtadest, mis on olulised automaatse kõnede transkriptsiooni ja sellele järgneva sentimendianalüüsi mõistmiseks kindlustusvaldkonna kontekstis. Alustuseks käsitletakse automaatse kõnetuvastuse põhimõtteid ja mudelite arengut, sealhulgas traditsioonilisi ning kaasaegseid lahendusi nagu Kaldi, Whisper ja Wav2Vec 2.0. Seejärel tutvustatakse peamisi arhitektuurilisi erinevusi nende mudelite vahel ning käsitletakse sentimendianalüüsi teoreetilisi aluseid, sealhulgas leksikonipõhiseid ja süvaõppel põhinevaid lähenemisi. Lõpuks vaadeldakse andmekaitse ja väikekeeltega seotud eripärasid, mis mõjutavad tehniliste lahenduste valikut ja rakendatavust Eesti kindlustussektoris.

1.1 Kõnede transkriptsioon

Automaatne kõnede transkriptsioon on viimastel aastatel teinud märkimisväärseid edusamme tänu süvaõppe meetodite arengule. Varasemad lahendused põhinesid valdavalt statistilistel mudelitel, millest levinumad olid peidetud Markovi mudelid (ingl. *Hidden Markov Models*, HMM) ja Gaussi segu mudelid (ingl. *Gaussian Mixture Models*, GMM).

HMM-i näol on tegemist järjestikuste olekute ahelaga, kus iga olek vastab ligikaudu mingile keelelisele ühikule, näiteks foneemile. Süsteem ei jälgi otseselt, millises olekus ta parasjagu on (olek on "peidetud"), kuid igast olekust pärinevad akustilised tunnused on jälgitavad. Seega võimaldab HMM modelleerida kõne ajas muutuvat struktuuri, võttes arvesse nii olekute üleminekuid kui ka kestuseid. (Rabiner, 1989)

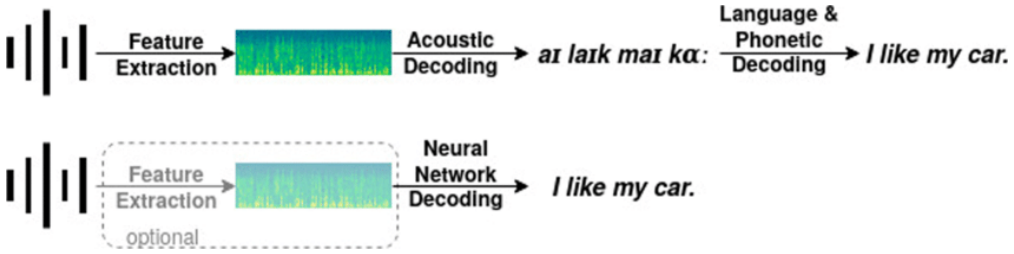
GMM-i kasutatakse omakorda iga oleku puhul selleks, et kirjeldada akustiliste tunnuste jaotust. Kuna üheainsa Gaussi jaotusega ei saa realistlikult modelleerida kõne varieeruvust (nt erinevad hääldused, aktsendid, taustamüra), kasutatakse Gaussi jaotuste kombinatsioone. GMM võimaldab hinnata, kui tõenäoline on konkreetne helilõik antud foneemi jaoks, võttes arvesse selle akustilist mitmekesisust.

HMM ja GMM olid enam kui kaks aastakümmet automaatsete kõnetuvastussüsteemide aluseks. Need mudelid võimaldasid akustilise signaali põhjal tuvastada kõige tõenäolisema foneemide või sõnade jada. Nende tugevuseks oli modulaarne ülesehitus - HMM modelleeris kõne ajastruktuuri, GMM kirjeldas akustilisi tunnuseid, hääldussõnastik sidus foneemid sõnadega ning eraldi keelemudel aitas eelistada grammatilisi sõnejärjestusi. Selline arhitektuur võimaldas süsteeme hästi kohandada uutele keelte ja domeenidele, kuid oli keeruline ehitada ning iga mooduli täiendamine või asendamine nõudis eraldi ekspertiisi (Rabiner, 1989).

Süvaõppe levikuga 2010. aastate alguses hakati HMM-GMM süsteeme järk-järgult asendama süvanärvivõrgul (ingl. *deep neural network*, DNN) põhinevate süsteemidega. Esialgu asendati ainult GMM komponendid sügavate närvivõrkudega, säilitades ülejäänud töövoo (ingl. *pipeline*). Edasi arendati ka täielikult läbivteenuse arhitektuurid (ingl.

end-to-end, E2E), kus kogu transkriptsioon toimub üheainsa närvivõrgu sees - alates helisignaali sisendist kuni lõpliku tekstini (Hinton et al., 2012).

E2E lähenemine muudab kõnetuvastussüsteemi lihtsamaks, kuna kaob vajadus eraldi käsitsi koostatud komponentide, nagu foneemitabelite ja hääldussõnastike järele. Mudel õpib otseselt helisignaali ja selle tekstilise vaste seoseid, ilma et oleks vaja eelnevalt määratleda keele reegleid või foneemistruktuuri. See muudab süsteemi ülesehituse ühtlasemaks ning võimaldab sageli saavutada parema täpsuse eeldusel, et treeningandmeid on piisavalt. Samas tähendab selline lähenemine ka väiksemat kontrolli mudeli sisemiste tööetappide üle, mis võib vähendada selgitatavust ning raskendada kohandamist konkreetsete valdkondade vajadustele (Georgescu et al., 2021).



Joonis 3. Kõnetuvastussüsteemide areng: traditsiooniline pipeline-lähenemine (üleval) ja end-to-end lähenemine (all). Kohandatud allikast: (Georgescu et al., 2021).

1.1.1 Akustilised tunnused

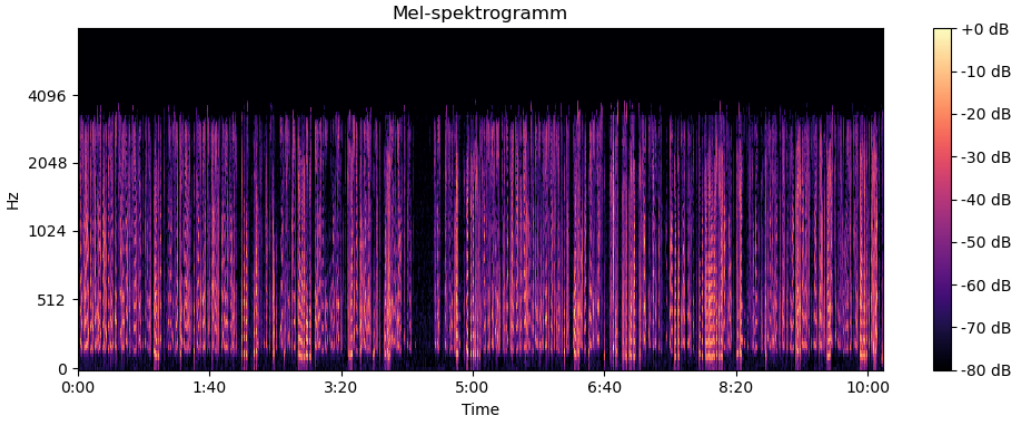
Kõnetuvastussüsteemide esimene samm on helisignaali teisendamine tunnusteks (ingl. *features*), mis säilitavad olulisima info kõne akustilise struktuuri kohta. Levinumad lähenemised sisaldavad:

Mel-spektrogramm (ingl. *mel spectrogram*) on sagedusanalüüsi graafiline kujutus, kus helisignaali võimsust mõõdetakse erinevates sagedusribades, mis on jaotatud *mel*-skaalal - inimkõrva tajule vastavalt logaritmilise tihedusega skaalal. Selline teisendus aitab närvivõrkudel paremini tabada kõnele iseloomulikke jooni ning on laialdaselt kasutusel süvaõppel põhinevates kõnetuvastussüsteemides (sh end-to-end mudelites). (Stevens et al., 1937)

Joonisel 4 on toodud näide 10-minutilise eestikeelse kõnesalvestise mel-spektrogrammist. Kujutis on genereeritud Pythonis, kasutades `librosa`, `matplotlib` ja `numpy` teeke.

[1] Kasutatud helifail on pärit kindlustusvaldkonna tegelikust müügikõnest.

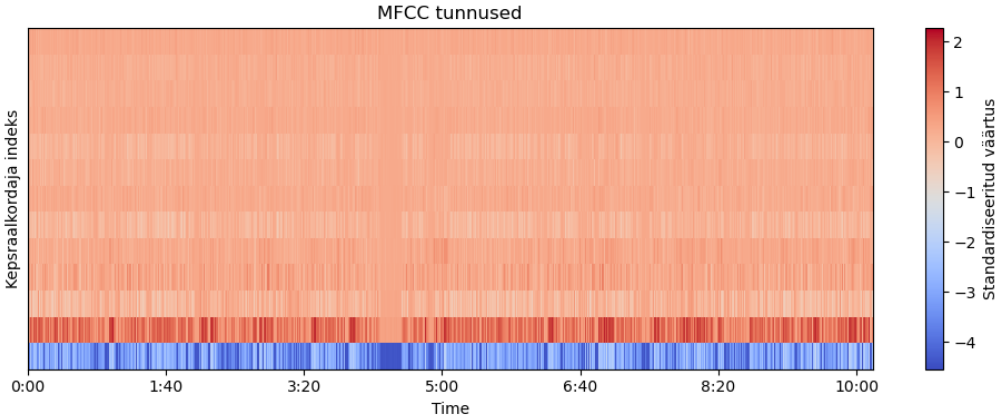
[1] Spektrogrammi genereerimise koodi aluseks oli ChatGPT 4o (OpenAI, 2025) poolt loodud näidis, mida autor kohandas töö spetsiifikale sobivaks.



Joonis 4. Mel-spektrogramm 10-minutilisest eestikeelsest kõnesalvestisest. Vertikaaltelg näitab sagedust (Hz), horisontaaltelg aega (min:s). Värv tähistab signaali tugevust detsibellides (dB). Kõne pausid ja aktiivsed lõigud on selgelt eristatavad.

MFCC Teine levinud akustiline tunnus on mel-sageduslikud kepstraalkordajad (ingl. *mel-frequency cepstral coefficients*, MFCC), mida kasutatakse sageli traditsioonilistes GMM/HMM-põhistes süsteemides. MFCC-d annavad ülevaate sellest, kuidas heli ajas muutub, arvestades nii kõla kui ka häälduse seotud omadusi. Tänu sellele aitavad need paremini eristada erinevaid kõneelemente. Nende muutumist ajas illustreerib joonis 5 (Davis and Mermelstein, 1980)

[2] Kasutatud helifail on pärit kindlustusvaldkonna tegelikust müügikõnest.



Joonis 5. MFCC (mel-sageduslikud kepstraalkordajad) 10-minutilisest eestikeelsest kõnesalvestisest. Joonisel on kujutatud 13 kepstraalkordaja väärtused ajas.

[2]MFCC genereerimise koodi aluseks oli ChatGPT 4o (OpenAI, 2025) poolt loodud näidis, mida autor kohandas töö spetsiifikale sobivaks.

Mel-spektrogramm ja mel-sageduslikud kepstraalkordajad (MFCC) on levinud viisid helisignaali kujutamiseks. Mel-spektrogramm näitab, kui tugevad on erinevad sagedused ajas. MFCC-d arvutatakse Mel-spektrogrammi põhjal ning annab lühema ja lihtsama arvulise esitluse, mida on mugavam kasutada masinõppes. Kui mel-spektrogramm sobib hästi visuaalseks analüüsiks, siis MFCC sobib paremini arvutile sisendiks andmiseks.

1.2 Transkriptsioonimudelid

Selles peatükis antakse ülevaade automaatse kõnetuvastuse lahendustest, mida magistritöö raames katsetati. Eesmärgiks on võrrelda erinevate transkriptsioonimudelite — Whisperi, Wav2Vec 2.0 ja Kaldi — sobivust eestikeelsete kindlustuskõnede töötlemiseks. Iga mudeli puhul käsitletakse selle arhitektuurilisi aluseid, treeningmetoodikat ning praktilisi kasutusvõimalusi väikese ressursiga keelte kontekstis. Võrdluse keskmes on mudelite rakendatavus ja potentsiaal täpsete ning usaldusväärsete transkriptsioonide loomiseks.

1.2.1 Whisper

Whisper on OpenAI poolt 2022. aastal välja töötatud E2E kõnetuvastussüsteem, mis pakub mitmeid eeliseid võrreldes varasemate lahendustega (Radford et al., 2022). Kõnetuvastussüsteemi Whisper treenimisel on kasutatud 680 000 tundi mitmekeelset andmestikku, mis sisaldab transkriptsioone 98 erinevas keeles, sealhulgas limiteeritud määral ka eesti keeles.

Whisperi arhitektuur põhineb kodeerija-dekodeerija tüüpi transformer-mudelil (ingl. *encoder-decoder transformer model*), mis võimaldab efektiivselt töödelda pikki helisalvestisi. Mudel jagab helifailid 30-sekundilisteks lõikudeks, teisendab need mel-spektrogrammideks ning kasutab seejärel kodeerijat (*encoder*), et heli esitada kompaktsel kujul, ja dekodeerijat (*decoder*), et genereerida vastav tekstiväljund.

Eesti keele jaoks on loodud mitu Whisper-põhist mudelit, mis on peenhäälestatud (ingl. *fine-tuned*) eestikeelsel audiomaterjalil. Need mudelid pakuvad potentsiaalselt paremat täpsust kui üldine mitmekeelne Whisper mudel (Fedorchenko and Alumäe, 2025).

1.2.2 Kaldi

Kaldi on avatud lähtekoodiga kõnetuvastuse tööriistakomplekt, mis on laialt kasutusel nii akadeemilises maailmas kui ka tööstuses (Povey et al., 2011). Erinevalt Whisperist järgib Kaldi traditsioonilisemat töövoopõhist lähenemist, kus akustiline modelleerimine, hääldussõnastik ja keelemudelid on eraldi komponendid.

Kaldi kasutab tavaliselt sügava närvivõrgu ja peidetud Markovi mudeli hübriidarhitektuuri (ingl. *deep neural network - hidden Markov model*, DNN-HMM), kus närvivõrgud

on kombineeritud statistiliste mudelitega. See võimaldab saavutada kõrget täpsust, eriti kui on saadaval keelespetsiifilised ressursid nagu foneemide kataloog ja häälduss õnastik.

Eesti keele jaoks on Kaldi baasil välja töötatud mitmeid süsteeme, sealhulgas TalTechis loodud eesti keele kõnetuvastussüsteemid. Need süsteemid on kohandatud arvestama eesti keele fonoloogiliste ja morfoloogiliste eripäradega, mis on olulised täpse transkriptsiooni saavutamiseks (Alumäe, 2021).

Kaldi eeliseks on selle kohandatavus ja läbipaistvus, mis võimaldab süsteemi täiustada vastavalt konkreetse valdkonna vajadustele. Samas on Kaldi süsteemide seadistamine ja treenimine tehniliselt keerukam võrreldes E2E lähenemistega (nt Whisper).

1.2.3 Wav2Vec ja Wav2Vec 2.0

Wav2Vec on Facebook AI Research (FAIR) poolt välja töötatud kõnetuvastustehnoloogia, mille esimene versioon avaldati 2019. aastal ning edasiarendus Wav2Vec 2.0 2020. aastal (Baevski et al., 2020). Tegemist on läbivteenus arhitektuuriga, mis võimaldab õppida helisignaali akustilisi omadusi otse töötlemata helivoo põhjal, kasutades isejuhitud õppemeetodit (ingl. *self-supervised learning*).

Esialgne Wav2Vec mudel (Wav2Vec 1) töötles helisignaali konvolutsiooniliste närvivõrkude (ingl. *convolutional neural networks*, CNN) abil, mille eesmärk oli teisendada heli lühemaks ja tähenduslikuks vektorkujuks ehk varjatud representatsiooniks. Kuigi mudel suutis heli struktuurist mõningaid mustreid õppida, ei olnud selle väljund piisavalt informatiivne, et otse tekstiks teisendada. Selle asemel tuli lisada eraldi klassifikaator ja keelemudel, mis vajasid suurel hulgal käsitsi märgendatud andmeid. Seetõttu jäi Wav2Vec 1 täpsus ja paindlikkus piiratuks, eriti väikeste keelte ja keerukama kõne korral. (Schneider et al., 2019)

Wav2Vec 2.0 lahendas eelkäija piirangud, lisades konvolutsioonikihi järel transformerarhitektuuri, mis võimaldab mudelil paremini mõista heli sisulist ülesehitust ja arvestada kaugemate seostega kõne struktuuris. Mudelis kasutatakse ainult kodeerijat (ingl. *encoder-only transformer*), mis teisendab helisignaali järjestikuseks vektorite jadaks. Sellele esitlusele järgneb ennustav kiht, mis määrab igale ajahetkele sobiva sümboli. Kuna kõnes ei ole täpseid sõnapiire, kasutatakse tekstiväljundi loomiseks spetsiaalset algoritmi *Connectionist Temporal Classification* (CTC), mis võimaldab ennustada järjestusi muutliku pikkusega sisendi ja väljundi korral. (Baevski et al., 2020)

CTC-algoritm võimaldab mudelil arvata, millal üks sõna või häälik lõpeb ja järgmine algab, ilma et neid oleks vaja eelnevalt ajaliselt täpselt märgistada. Selle tulemusel saab mudel toota järjestatud tekstiväljundi ka siis, kui kõne ei ole täpselt segmenteeritud.

Lisaks muudeti ka õppemeetodit: esmalt lastakse mudelil kuulata suurt hulka märgendamata helisalvestisi ja õppida nende põhjal üldisi mustreid. Alles seejärel õpetatakse mudelit väikese hulga transkribeeritud kõnede abil heli ja teksti vahelisi seoseid looma. Selline lähenemine võimaldab saavutada häid tulemusi isegi siis, kui käsitsi märgendatud andmeid on vähe, mis on väikeste keelte puhul sageli vältimatu.

Wav2Vec 2.0 eeliseks on kõrge täpsus isegi piiratud suurusega märgendatud andmestiku korral, mistõttu sobib see hästi väiksemate keelte jaoks nagu eesti keel. Selle arhitektuur on suutnud saavutada tulemusi, mis on võrreldavad või paremad kui keele spetsiifilised süsteemid - eriti juhul, kui kasutada sobivat peenhäälestust. Nagu näitab tabel 1, on mudel suuteline saavutama madala sõnavigade määra (ingl. *word error rate*, WER) isegi juhul, kui märgendatud andmeid on äärmiselt vähe (Baevski et al., 2020).

Tabel 1. Wav2Vec 2.0 täpsus erineva märgendatud andmehulgaga (WER, %).
Mudel: LARGE + Transformer LM + LV-60k (53k tundi märgendamata kõnet)

Märgendatud andmete hulk	**WER (clean)**	**WER (other)**
10 minutit	4,8%	8,2%
1 tund	2,9%	5,8%
10 tundi	2,4%	4,9%
100 tundi	2,0%	4,0%

Allikas: (Baevski et al., 2020), tabel 1 - Wav2Vec 2.0 (LARGE, LV-60k)

Eesti keele jaoks on välja töötatud ka mitmeid Wav2Vec 2.0 baasil edasitreenitud mudeleid eelkõige TalTechNLP labori poolt. Need mudelid on peenhäälestatud eestikeelsetel andmetel ning suudavad pakkuda paremat transkriptsioonitäpsust võrreldes mitmekeelsete üldmudelitega.(TalTechNLP, 2022)

1.2.4 Võrdlus: Whisper, Kaldi ja Wav2Vec

Whisper, Kaldi ja Wav2Vec esindavad kolme erinevat lähenemist kõnetuvastuse probleemile, millest igaühel on oma tugevused ja piirangud. Need tehnoloogiad peegeldavad kõnetuvastussüsteemide arengut erinevatel ajastutel, pakkudes alternatiive vastavalt keeleressursside kättesaadavusele ja rakenduse vajadusele.

Kaldi (Povey et al., 2011) järgib traditsioonilist moodulpõhist arhitektuuri, kus akustiline mudel, hääldussõnastik ja keelemudel on eraldi komponendid. Selle eeliseks on kõrge kohandatavus ja komponentide läbipaistev juhtimine, mis võimaldab täpsustada tulemusi konkreetsele keelele, kuid see eeldab tehnilisi oskusi ning keelespetsiifiliste ressursside olemasolu.

Wav2Vec 2.0 (Baevski et al., 2020) esindab sammu isejuhitud õppe suunas, kus mudel õpib esmalt suurel hulgal märgendamata helisalvestistel helikujutisi ning alles seejärel toimub peenhäälestus väiksemal hulgal märgendatud andmetel. See lähenemine võimaldab saavutada häid tulemusi ka piiratud andmestike korral, mistõttu sobib eriti väikese ressursiga keeltele nagu eesti keel.

Whisper (Radford et al., 2022) on täielikult läbivteenus arhitektuuriga kõnetuvastussüsteem, mis põhineb kodeerija-dekodeerija transformer mudelil. Mudel on treenitud 680 000 tunni erineva keele ja kvaliteediga heliandmetel, mis võimaldab tal käsitleda mitmekeelseid sisendeid ilma eraldi keelemudeliteta ja teostada automaatset keeletuvastust. See muudab Whisperi sobivaks kiireks kasutuselevõtuks erinevates keeleruumides.

Järgnevalt vaatleme lähemalt nende mudelite arhitektuurilisi erinevusi ja töötamispõhimõtteid, mis aitavad mõista nende sobivust eestikeelsete kindlustuskõnede transkribeerimiseks.

1.3 Mudelite arhitektuurilised erinevused

Kuigi kõik kolm mudelit lahendavad sama probleemi – kõne teisendamist tekstiks – kasutavad nad selleks fundamentaalselt erinevaid lähenemisi.

Arhitektuurilised põhimõtted. Kaldi süsteem koosneb eraldiseisvatest moodulitest, kus igas töötlusetapis rakendatakse spetsialiseeritud komponenti. Akustiline mudel (DNN-HMM hübriid) tuvastab foneeme, hääldussõnastik seob foneemid sõnadega ja keelemudel aitab valida kontekstis sobivat sõnajärjestust. See modulaarne ülesehitus võimaldab iga komponenti eraldi optimeerida, kuid eeldab tervikliku töötlusahela seadistamist ja hooldust. (Povey et al., 2011)

Whisper kasutab kodeerija-dekodeerija transformer arhitektuuri, kus kodeerija teisendab helisignaali varjatud esituseks ning dekodeerija genereerib selle põhjal järk-järgult tekstiväljundi. Transformerid võimaldavad tähelepanumehhanismi abil (ingl. *attention mechanism*) arvestada kogu helilõigu konteksti korraga, mis parandab pikaajaliste sõltuvuste tuvastamist kõnes. (Radford et al., 2022)

Wav2Vec 2.0 rakendab ainult kodeerija-põhist arhitektuuri, kus heliandmetest leitakse varjatud esitus, millest transkriptsioon luuakse CTC algoritmi abil. CTC võimaldab töödelda helisignaali ilma spetsiifilise joondamiseta, leides igale heliüksusele vastava sümboli või tühisümboli. See lähenemine on eriti tõhus, kui täpne joondamine pole teada või on keeruline määrata. (Baevski et al., 2020)

Treenimismeetodid. Kaldi süsteemide treenimine toimub etapiviisiliselt, kus iga komponenti treenitakse eraldi andmestikul. Akustilise mudeli jaoks vajatakse kõnesegmente koos nende transkriptsioonidega, keelemudeli jaoks aga suurt tekstikorpust. See nõuab märkimisväärset käsitööd ja keelespetsiifiliste ressursside ettevalmistamist. (Povey et al., 2011)

Whisper on treenitud juhendatud õppe meetodil, kus mudel õpib otseselt helisalvestise ja transkriptsiooni vahelisi seoseid. Suure treeningandmete mahu tõttu (680 000 tundi) suudab mudel üldistada erinevat laadi kõnet ja keeli, sealhulgas mürarikast või aktsendiga kõnet. (Radford et al., 2022)

Wav2Vec 2.0 kasutab esimeses faasis isejuhitud eelõpet, kus mudel ennustab varjatud akustilisi tunnuseid ilma teksti transkriptsioonita. Alles teises faasis toimub peenhäälestus väiksemal hulgal märgendatud andmetel. Selline kaheastmeline lähenemine võimaldab mudelil õppida helistruktuure ka kontekstides, kus märgendatud andmeid on vähe. (Baevski et al., 2020)

Kontekstuaalsed võimed ja mitmekeelsus. Whisper suudab töödelda kuni 30-sekundilisi helilõike tervikuna ning võtab arvesse laiemat lingvistilist konteksti. Mudel on disainitud algusest peale mitmekeelsena, toetades 98 keelt, sealhulgas eesti keelt. See võimaldab mudelil sooritada null-eeltreeningu tõlkimist, kus sisendkeel tuvastatakse automaatselt. (Radford et al., 2022)

Wav2Vec 2.0 töötleb heli järjestikuste raamidena, mis võimaldab tuvastada üksikuid foneeme ja sõnu, kuid pikema konteksti haldamise eest vastutab eraldi keelemudel. Wav2Vec 2.0 mudelid on sageli keelespetsiifilised või piiratud keeltevalikuga ning vajavad selget sisendkeele määratlemist. (Baevski et al., 2020)

Kaldi puhul on konteksti arvestamine jagatud erinevate komponentide vahel – akustiline mudel tegeleb lühikeste helisegmentidega, samas kui keelemudel arvestab pikema kontekstiga. Iga keele jaoks on vaja eraldi ressursse ja mudeleid, mis muudab mitmekeelsuse toetamise keerulisemaks. (Povey et al., 2011)

Tabel 2. Kõnetuvastusmudelite võrdlus: Whisper, Kaldi ja Wav2Vec 2.0

Omadus	**Whisper**	**Kaldi**	**Wav2Vec 2.0**
Arhitektuur	Läbivteenus (kodeerija-dekodeerija transformer)	Moodulpõhine arhitektuur, DNN-HMM hübriid	Läbivteenus, isejuhitud eelõpe + peenhäälestus
Keeleressursside vajadus	Madal (null-eeltreeningu võimekus)	Kõrge (foneemid, sõnastik, keelemudel)	Keskmine (märgendamata andmed + väike annoteeritud korpus)
Täpsus väikestes keeltes	Hea, kuid jääb alla spetsialiseeritud mudelitele	Väga hea, kui ressurssid olemas	Väga hea, kui teostatud sobiv peenhäälestus
Kohandatavus konkreetsele domeenile	Piiratud, mudel on must kast	Väga hea, iga komponenti saab eraldi kohandada	Hea, võimalik domeenipõhiselt peenhäälestada
Seadistamise keerukus	Madal, valmislahendus	Kõrge, tehniline töö nõutav	Keskmine, valmis mudelid kättesaadavad
Töötlemiskiirus	Väikeste mudelite puhul kiire, suured aeglased	Keskmine, sõltub konfiguratsioonist	Keskmine, sõltub mudeli suurusest
Avatus ja paindlikkus	Avatud lähtekood, piiratud muudetavus	Täielik kontroll, skriptid ja komponendid kohandatavad	Avatud lähtekood, kuid mudelistruktuur on keeruline
Sobivus mitmekeelseks kasutuseks	Väga hea, 98 keele toetus	Halb, iga keel vajab eraldi ressursse	Keskmine, sõltub eeltreeningu mitmekeelsusest

Tabelis 2 on kokkuvõtlikult võrreldud kolme mudeli peamised omadused ja võimekused erinevate kriteeriumite alusel. Ilmneb, et Whisper sobib paremini mitmekeelseks ja kiireks kasutuselevõtuks, Kaldi süvitsi kohandatud ja kontrollitud lahendusteks ning Wav2Vec 2.0 projektideks, kus on piiratud kogus märgendatud andmeid, kuid soovitakse head üldistust väikestes keeltes nagu eesti keel. (Baevski et al., 2020; Radford et al., 2022; Povey et al., 2011)

1.4 Sentimendianalüüs

Kõnede transkriptsioonide põhjal teostatav sentimendianalüüs on olulisel kohal kindlustussektori kliendisuhtluse parendamisel ning võib pakkuda taustainfot andmete asjako-

hasuse hindamisel. Sentimendianalüüsi abil on võimalik hinnata kliendi emotsionaalset hoiakut ning potentsiaalseid rahulolu või rahulolematuse allikaid. Kuigi sellised hinnangud võivad olla heaks lähtekohaks, on nende täpseks rakendamiseks vajalik põhjalik valideerimine antud valdkonna kontekstis.

Traditsioonilised sentimendianalüüsi meetodid baseeruvad sõnastikupõhistel lähenemistel, kus igale sõnale on määratud positiivne või negatiivne väärtus (Raees and Fazilat, 2024). Kaasaegsed lähenemised kasutavad masinõppe mudeleid, eelkõige sügavaid närvivõrke, mis suudavad arvesse võtta konteksti ning tuvastada keerukamaid keelelisi mustreid (Guo, 2024).

Vähese ressursiga keelte (nt eesti, läti ja leedu) puhul on sentimendianalüüs eriti keerukas, kuna puuduvad kvaliteetsed lause tasemel märgendatud andmestikud ning olemasolevad tööriistad ei pruugi arvestada nende keelte morfoloogilise ja kultuurilise mitmekesisusega. Sellistes oludes on osutunud tõhusaks lähenemiseks leksikonipõhine eelõpe, kus ei kasutata üldse lause tasemel treeningandmeid. Selle asemel treenitakse mudelit üksiksõnade põhjal, kasutades mitmekeelseid sentimentileksikone, kus igale sõnale on määratud hinnanguline meeleolu väärtus (nt positiivne või negatiivne valents). Mudel õpib seoste kaudu üldistama ja suudab hiljem hinnata ka tervete lausete sentimenti isegi keeltes, mille kohta puuduvad lause tasemel näited (Koto et al., 2024).

Tulevikus võiks automaatselt transkribeeritud kõnede analüüs hõlmata ka eesmärgi (ingl. *intent*) tuvastamist, mis võimaldaks klassifitseerida kõnesid nende funktsiooni alusel, näiteks müügikõned, kahjukäsitlus või teeninduspäringud. Käesolevas töös on fookuses siiski erinevate sentimendianalüüsi mudelite võrdlemine ning nende käitumismustrite analüüs, mis paneb aluse edasistele uuringutele.

Kõnede eesmärgi automaatne tuvastamine on seda olulisem, et kindlustusettevõtted salvestavad igapäevaselt suurel hulgal kõnesid, millest vaid osa on strateegiliselt oluline säilitada. Automaatne klassifikatsioon aitab vähendada manuaalse läbivaatuse vajadust ning toetab isikuandmete kaitse põhimõtete rakendamist, võimaldades kustutada ebaolulised salvestised vastavalt andmekaitse regulatsioonidele.

1.5 Andmekaitse ja privaatsuse kaalutlused

Kindlustussektori kõnesalvestiste analüüsimisel on andmekaitse ja privaatsuse tagamine äärmiselt oluline. Euroopa Liidu isikuandmete kaitse üldmääruse (ingl. *General Data Protection Regulation*, GDPR) kohaselt kvalifitseeruvad helisalvestised, mille kaudu saab tuvastada konkreetset isikut, isikuandmeteks ning nende töötlemine peab vastama rangetele õiguslikele nõuetele (European Parliamentary Research Service, 2020).

Selliste salvestiste säilitamine ja töötlemine peab toimuma selgelt määratletud eesmärkidel ning piiratud ajaperioodil. Seetõttu on oluline välja töötada usaldusväärsed meetodid, mis võimaldavad hinnata kõnede asjakohasust ja säilitamise vajadust, minimeerides samal ajal isikuandmete liigset kasutamist. (Nautsch et al., 2019) Eriti tundlikuks

muudab olukorra väikekeelte kontekst, kus automaatsete filtrite ja kustutamismehhanismide täpsus võib otseselt mõjutada andmete töötlemise õiguspärasust.

Kaasaegsed kõnetuvastuse ja loomuliku keele töötlemise vahendid võivad aidata luua privaatsust säilitavaid lahendusi - näiteks tuvastada tundliku sisuga lõike, hinnata kõne sisu väärtust või vältida ebavajalike salvestiste säilitamist. Samas tuleb tagada, et kasutatavad mudelid oleksid piisavalt täpsed, läbipaistvad ja vastaksid andmekaitsepõhimõtetele (Weitzenboeck et al., 2022).

Need lahendused on eriti olulised eestikeelses kontekstis, kus valmis kommertslahendused on piiratud ja ettevõtted peavad sageli looma kohandatud süsteeme, mis vastavad kohalikele vajadustele. Seetõttu on käesolev töö oluline samm tõhusama ja õiguspärasema andmehalduse suunas.

1.6 Balti keelte eripärad ja mudelite sobivus

Eesti, läti ja leedu keel kuuluvad küll geograafiliselt Balti piirkonda, kuid on lingvistiliselt märkimisväärselt erinevad: eesti keel kuulub soome-ugri keelkonda, samas kui läti ja leedu on indoeuroopa keeled. Selline keeleline mitmekesisus esitab väljakutseid ühtsete kõnetuvastus- ja analüüsilahenduste väljatöötamisel.(Olev and Alumäe, 2024; Rehm and Way, 2023)

Balti keelte peamised eripärad, mis mõjutavad kõnetuvastussüsteemide täpsust ja kohandatavust, on järgmised:

1. Morfoloogiline keerukus - eesti keeles esineb 14 käändevormi ja keeruline sõnamoodustussüsteem, mis suurendab oluliselt võimalike sõnavormide arvu võrreldes analüütilisemate keeltega, näiteks inglise keelega.

2. Fonoloogilised eripärad - eesti keele spetsiifilised häälikud nagu *õ*, *ä*, *ö* ja *ü* nõuavad akustilistes mudelites erikäsitlust ja täiendavat kohandamist. (Asu and Teras, 2009)

3. Ressursside piiratus - võrreldes suurte maailmakeeltega on eestikeelsete kõnekorpuste ja annoteeritud andmestike maht piiratud, mis raskendab suuremahuliste mudelite treenimist. (Rehm and Way, 2023)

4. Dialektide ja aktsendi mõju - eriti vanema põlvkonna kõnelejate seas esinevad piirkondlikud erinevused ja tugevad aktsendid, mis võivad vähendada mudelite üldistustugevust.

Whisperi ja Kaldi mudelite sobivus Balti keeltele sõltub nende võimest kohanduda nimetatud keeleliste eripäradega. Whisperi mitmekeelne arhitektuur ja nullnäiteline *zero-shot* võimekus pakuvad potentsiaalset eelist väiksema ressursiga keelte puhul, kuid spetsiifiliste keeleliste konstruktsioonide käsitlemisel võib mudeli täpsus jääda alla keelespetsiifiliselt kohandatud lahendustele. Kaldi põhine lähenemine võimaldab

üksikasjalikumat keele- ja akustikamudelite kohandamist, kuid nõuab oluliselt rohkem keeleressursse ja tehnilist ekspertiisi.

Kindlustussektori kõnede transkribeerimisel on oluline hinnata, milline mudel suudab parimal moel käsitleda nii eesti keele üldiseid eripärasid kui ka vastava valdkonna spetsiifilist sõnavara. Transkriptsioonimudelite võrdlev analüüs aitab määratleda optimaalse lahenduse, mis pakub tasakaalu täpsuse, töökiiruse ja kohandatavuse vahel.

2 Varasemad uuringud ja võrdlev analüüs

Käesolevas jaotises keskendutakse Balti regiooni väikekeeltele (eesti, läti ja leedu) suunatud praktilistele kõnetuvastuse ja kõneanalüüsi lahendustele, mida rakendatakse erinevates tööstuslikes ja avalikes kontekstides. Varasemates peatükkides käsitletud teaduslike ja avatud lähtekoodiga lahenduste (nt Kaldi, Whisper, Wav2Vec2) taustal analüüsitakse siinkohal, kuidas neid või sarnaseid tehnoloogiaid on integreeritud ettevõtete ja riigiasutuste tööprotsessidesse.

Eraldi tähelepanu pööratakse sellistele tegijatele nagu FeelingStream, Tilde, TalTech (tekstiks.ee) ja Oxus.AI, kes on välja töötanud funktsionaalsed süsteemid kliendisuhtluse, kohtumenetluste, parlamendiarutelude ja teeninduskõnede automaatseks töötlemiseks.

Peatüki lõpuosas käsitletakse eraldi küsimust, kuivõrd piisav on automaatse transkriptsiooni täpsus semantilise analüüsi (nt meeleolu või kavatsuse tuvastamine) läbiviimiseks ning millised on peamised kitsaskohad praktikas.

2.1 Tööstuslikud lahendused Balti riikides

FeelingStream

FeelingStream on Eesti ettevõte, mis pakub reaalajas kõneanalüüsi teenuseid mitmes keeles, sh eesti, vene, soome ja inglise keeles. Süsteem kasutab kohandatud Kaldi-põhist kõnetuvastusmootorit ning sellele järgnevat loomuliku keele töötlust (Feelingstream, 2021), kus rakendatakse näiteks BERT-põhiseid mudeleid. Transkriptsioonide alusel määratakse sentiment, vestluse tüüp ning tuvastatakse võimalikud müügivihjed. Lahenduse tugevuseks on reaalajas töötlemine (RTF $\approx$ 0,2) ning kõrge praktiline kasutatavus. Samas võib WER ulatuda ligikaudu 20% tasemele, mistõttu tekst vajab kriitilistes rakendustes endiselt järeltöötlust (Feelingstream, 2024).

Tilde ja Salme projekt

Läti ettevõte Tilde pakub mitmekeelseid ASR-teenuseid ning on arendanud domeenipõhiseid mudeleid, sealhulgas Eesti kohtusüsteemi jaoks loodud Salme projekt. Salme süsteem treeniti enam kui 800 tunni eestikeelsete helide ja 800 miljoni sõna tekstikorpuse põhjal. Tulemuseks saavutati WER vahemikus 8-15%, sõltuvalt kõne kvaliteedist ja

kõnelejate arvust salvestisel. Süsteem võimaldab reaalaja- ja järeltranskriptsiooni ning automaatset subtiitrite ajastust. Salme on kasutusel kõikides Eesti kohtutes ning arendati koostöös CGI Eestiga (Tilde, 2023).

TalTech Tekstiks.ee

Tallinna Tehnikaülikooli loodud teenus `tekstiks.ee` põhineb Wav2Vec 2.0 suurmudelil, mis on peenhäälestatud umbes 760 tunni eestikeelse kõne peal. Teenuse eelversioon tugines Kaldi-põhisele süsteemile, kuid uue arhitektuuriga saavutati keskmiselt 27% parem täpsus. Akadeemilise testi põhjal saavutati järgmised tulemused: WER 6̃,9% ringhäälingu salvestistel, 9,6% konverentsidel ning 16,6% reaalsetel kasutajate üleslaaditud failidel (Olev and Alumäe, 2022). Suure mudeli tõttu on süsteem oluliselt aeglasem kui Kaldi, mistõttu vajab GPU tuge, et vältida koormusest tingitud ooteaegu.

Oxus.AI ja kliendispetsiifiline ASR

Leedus tegutsev Oxus.AI on keskendunud klienditoe ja müügikõnede automaatsele analüüsile väikekeeltes. Ettevõte pakub lahendusi, kus iga kliendi jaoks treenitakse kohandatud kõnetuvastusmudel, tõenäoliselt suurte eeltreenitud arhitektuuride (nt Whisper või Wav2Vec2) baasil. Süsteem suudab automaatselt hinnata kõnekvaliteeti, tuvastada vestluste põhiteemad ja anda soovitusi meeskonna töö parendamiseks. Ettevõtte enda väitel saavutatakse reaalsetes kasutusoludes üle 94% täpsus (Oxus.AI, 2021).

2.2 Semantilise analüüsi kvaliteet ASR-i baasil

Uuringud ja praktika näitavad, et sentimentide ja kõne eesmärgi tuvastamine on võimalik ka osaliselt ebatäpsetest transkriptsioonidest. Näiteks FeelingStream kasutab reaalajas hinnangut kliendi emotsioonile juba 80% täpsuse juures. Sentimentimudelid (nt BERT, RoBERTa) on teatud määral vigade suhtes robustsed, kui võtmesõnad jäävad alles. Kriitiline mõju esineb vaid siis, kui tuvastamata jäävad võtmeverbid nagu „tühistan“, „lahkun“ vms.

Vestluse eesmärgi tuvastamine sõltub samuti võtmemõistetest. Vigade mõju leevendamiseks kasutatakse praktikas kas ebaselgete juhtude suunamist käsitsi kontrolli või otsustuskünnise nihutamist.

2.3 Kokkuvõte

Balti väikekeelte jaoks on arendatud mitmeid efektiivseid ASR-lahendusi, mis saavutavad WER 10-20% vahemikus ning võimaldavad edasist semantilist analüüsi. End-to-end süsteemid nagu Wav2Vec2 ja Whisper pakuvad head täpsust, kuid vajavad rohkem ressursse. Kaldi on jätkuvalt sobiv valik, kui eesmärgiks on üles ehitada täielikult kontrollitav

ja kohandatav kõnetuvastustorustik. See on eelistatud olukordades, kus on vaja täpset kontrolli akustilise ja keelemudeli koostamise üle või kus nõutakse lokaalselt jooksvat süsteemi, mis ei sõltu välistest API-dest. Kuigi Kaldi ei ole lihtsasti kasutatav ja nõuab märkimisväärset tehnilist ettevalmistust, on see endiselt väärtuslik platvorm uurimis- ja arenduskeskkondades, kus soovitakse süsteemi sisemist töötlust täielikult mõista ja vajadusel ise optimeerida.

Tööstuslikud lahendused, nagu FeelingStream ja Tilde, on edukalt kombineerinud kohalikku keeletehnoloogiat ja spetsialiseeritud NLP-d, pakkudes reaalajas väärtust ärirakendustes. Teaduslikud võrdlusuuringud toetavad väidet, et semantilise analüüsi edukus sõltub suuresti transkriptsiooni täpsusest, kuid tänapäevased mudelid võimaldavad saavutada piisava kvaliteedi ka väikekeeltes.

3 Metoodika

Selles peatükis kirjeldatakse kasutatud andmestikku, rakendatud transkriptsioonimudeleid, tekstianalüüsi töötlustehnikaid, mudelite võrdlemise metoodikat ning eksperimendi läbiviimise tehnilist keskkonda ja töövoogu.

3.1 Andmestik

Käesolevas töös kasutatav andmestik koosneb kindlustussektori kliendisuhtluses toimunud kõnesalvestistest, mis hõlmavad nii müügi- kui ka kahjukäsitlusega seotud kõnesid. Kõnede pikkus varieerub, jäädes keskmiselt 5 kuni 12 minuti vahele.

Kõikide helifailide transkriptsiooni aluseks olevad korrektsed tekstid (ingl. *ground truth*) on koostatud käsitsi, järgides võimalikult täpset kõneleja öeldu fikseerimise põhimõtet. See võimaldab objektiivselt võrrelda automaatsete transkriptsioonimudelite väljundeid tegelike tekstidega.

Salvestised olid valdavalt stereoformaadis, mis võimaldas eristada kahte kõnelejat erinevates kanalites. Vajadusel teostati kanalite eraldamine (ingl. left/right split), mis võimaldas igat kõnelejat töödelda ja transkribeerida eraldi. Seejärel koondati tulemused kronoloogiliselt ühtseks dialoogiks.

Enne transkriptsioonimudelite rakendamist viidi läbi mitmesugused helitöötlused, mille eesmärk oli parandada mudelite sisendandmete kvaliteeti ja suurendada transkriptsioonide täpsust. Katsetati kolme erinevat lähenemist, millest igaüks keskendus mõnele konkreetsele aspektile helisignaali parendamisel.

Esimeses töötlusvariandis rakendati helitugevuse normaliseerimist ja tempo vähendamist, mille käigus ühtlustati kõne valjus kogu salvestise ulatuses ning aeglustati selle tempot ligikaudu 5% võrra. Selline lähenemine on kasulik eriti olukordades, kus kõne sisaldab kiiret diktsiooni või suuri valjususe kõikumisi. Aeglasem tempo võimaldab

transkriptsioonimudelil paremini eristada sõnapiire ning potentsiaalselt vähendada akustiliste elementide kattumist. Tehnilises teostuses kasutati Pydub'i abil helitugevuse ühtlustamiseks `effects.normalize()` funktsiooni ning helifaile töödeldi edasi Librosa raamistikus, kus rakendati tempot säilitavat ajamoonutust (ingl. *time stretching*) eraldi mõlemale stereokanalile.

Teises töötlusvariandis keskenduti taustamüra vähendamisele stereoformaadis, et säilitada kahe kõneleja akustiline eraldatus. Kindlustusvaldkonna kõnesalvestised on üldjuhul salvestatud kahekohaliselt, kus klient ja teenindaja kõnelevad eraldi helikanalites. Selline lahendus võimaldab töödelda mõlemat kõnelejat iseseisvalt, mis omakorda suurendab müra eemaldamise täpsust. Müra vähendati iga kanali puhul eraldi, kasutades `noisereduce` teeki, mis rakendab spektraalanalüüsil põhinevat kohanduvat mürasummutust. See lähenemine aitab säilitada soodsa signaali ja müra suhte ning minimeerib oluliste kõneelementide kadumise riski.

Kolmandas töötluses viidi läbi stereofaili teisendamine monoks koos järgneva töötlusega, mille käigus ühendati vasak- ja parempoolne helikanal üheks monoheliks. Seejärel rakendati kogu helile nii müravähendust kui ka helitugevuse normaliseerimist. See variant sobib paremini mudelitele, mis eeldavad ühekanalilisi sisendeid (nt Whisper või Wav2Vec2 baasil mudelid, mis on treenitud monoformaadis helil), võimaldades vältida stereoanalüüsi lisakompleksust. Kuna mono kanaliks teisendamisel võivad kanalite summamisel tekkida artefaktid või amplituudikõikumised, viidi helitase uuesti tasakaalu ning vajadusel rakendati müravähendus kogu signaalile tervikuna.

Kõik helitöötlusprotsessid viidi läbi eraldiseisvate skriptidena Pythonis, kasutades kombineeritult teeke nagu `pydub`, `librosa`, `soundfile` ja `noisereduce`. Eelprotsessimise eesmärk ei olnud ainult kuuldava kvaliteedi parandamine, vaid ka testida, kas sellised parendused suurendavad transkriptsioonimudelite suutlikkust käsitleda keerulisema akustilise struktuuriga salvestisi - näiteks kõnesid, kus esineb vaikseid kõnelejaid, kõnet katkestavat müra või kiiret kõnetempot. Katsete tulemusel selgus siiski, et kuigi helitöötlused parandasid kuuldelist kvaliteeti, põhjustasid need mõningatel juhtudel transkriptsiooni täpsuse langust, eriti akustiliste tunnuste moonutamise tõttu. Seetõttu viidi lõplikud transkriptsioonid läbi töötlemata, kuid formaadilt ühtlustatud monohelidega (16 kHz, 16-bit WAV), mis osutusid mudelite jaoks usaldusväärsemaks sisendiks.

3.2 Transkriptsioonimudelid

Transkriptsiooniks kasutati mitmeid erinevaid eeltreenitud mudeleid, mis jagunevad kahte suuremasse kategooriasse: end-to-end põhised mudelid (Whisper ja Wav2Vec 2.0) ning traditsiooniline töövoo põhine Kaldi süsteem. Kasutatud mudelid olid kas spetsiaalselt eesti keele jaoks kohandatud või mitmekeelsed eeltreenitud versioonid.

Selline lähenemine võimaldab võrrelda erinevate arhitektuuride sobivust eesti, läti ja leedu keele transkribeerimiseks, mis kuuluvad väikese keeletehnoloogilise toega keelte hulka. Sarnast metoodikat rakendas ka Parikh jt (2023), kes võrdlesid Kaldi ja Wav2Vec

2.0 jõudlust nii hea kui ka napilt varustatud keelte kontekstis, leides, et end-to-end mudelid suudavad saavutada parema täpsuse, kuid vajavad reaalajas kasutamiseks GPU tuge (Parikh et al., 2023).

3.2.1 Whisper-põhised mudelid

Whisper on OpenAI poolt välja töötatud end-to-end kõnetuvastusmudel, mille arhitektuur põhineb encoder-decoder transformeril. Töö käigus kasutati järgmisi Whisperi baasil loodud mudeleid:

Peenhäälestatud eestikeelsed mudelid

- **TalTechNLP/whisper-large-v3-turbo-et-subs** - Whisper V3 Turbo versioon, mis on spetsiaalselt peenhäälestatud eestikeelsete subtiitrite põhjal. Kohandatud TalTechi poolt kasutamiseks väikekeelelises kontekstis. (Fedorchenko and Alumäe, 2025)
- **Whisper-Large-V3-ET-Subs** - eksperimentaalne Whisper V3 mudel, mille peenhäälestus põhineb eestikeelsetel subtiitritel. Sobib domeenispetsiifilisteks rakendusteks.(Fedorchenko and Alumäe, 2025)
- **TalTechNLP/whisper-large-et** - varasem suur Whisper-mudel, mis on kohandatud eestikeelsete kõnekorpuste põhjal ning sobib üldotstarbeliseks eestikeelseks transkribeerimiseks. (TalTechNLP, 2023a)
- **TalTechNLP/whisper-medium-et** - keskmise suurusega Whisperi variant, peenhäälestatud eestikeelsetel andmetel. Vähem ressursinõudlik kui large-mudelid, kuid säilitab arvestatava täpsuse. (TalTechNLP, 2023b)

Üldised (mitte peenhäälestatud) Whisperi mudelid

- **OpenAI/whisper-large-v3** - OpenAI mitmekeelne suur mudel, mis suudab nulleeltreeningu (zero-shot) režiimis käsitleda üle 90 keele. Ei ole kohandatud eesti keele spetsiifikale. (Radford et al., 2022)
- **OpenAI/whisper-large-v3-turbo** - optimeeritud versioon eelmisest, mille fookus on inference'i kiirusel ja skaleeritavusel. Sobib mitmekeelseks rakenduseks. (Radford et al., 2022)
- **OpenAI/whisper-large-v3-turbo** koos `language="et"` seadistusega - sama mudel, kuid fikseeritud eestikeelseks töörežiimiks, et vältida automaatse keele tuvastuse eksimusi.
- **Faster-Whisper-Large-V3** - kiirendatud ja ressursisäästlik versioon Whisper V3 mudelist, sobib hästi CPU-põhiseks töötlemiseks. Kasutab väiksema täpsusega arvutustüüpe (nt `int8`). (SYSTRAN, 2023)

- **OpenAI/whisper-tiny** - Whisperi väikseim ametlik mudel, loodud kiiruse ja minimaalse ressursitarbimise jaoks. Sobib võrdluseks või sisendiks süsteemidesse, kus täpsus ei ole peamine eesmärk. (Radford et al., 2022)

Kõik Whisperi mudelid töötlesid helifaile 30-sekundiliste lõikudena (*chunk size* = 30 s), mis vastab mudelite optimaalsele tööpõhimõttele. TalTechNLP mudelid olid täiendavalt kohandatud eesti keelele subtiitrite ja kõnekorpuste põhjal.

3.2.2 Wav2Vec 2.0-põhine mudel

Lisaks Whisperile kasutati transkriptsiooniks ka Wav2Vec 2.0 baasil loodud mudeleid:

- **TalTechNLP/xls-r-300m-et** - Wav2Vec 2.0 XLS-R arhitektuuril põhinev mudel, millel on ligikaudu 300 miljonit parameetrit ja mis on treenitud spetsiaalselt eesti keele kõnekorpustel. (TalTechNLP, 2022)
- **Facebook/wav2vec2-base-960h** - originaalne ingliskeelne Wav2Vec 2.0 mudel, mis on treenitud 960 tunni ulatuses LibriSpeech andmestikul. Kasutatud võrdlusmudelina eesti keele mittespetsiifilise transkriptsiooni hindamiseks. (Baevski et al., 2020)

Wav2Vec 2.0 mudelid ühendavad end-to-end õppe ja isejuhitud eelõppe tugevused, võimaldades saavutada kõrget täpsust ka piiratud annoteeritud andmestiku korral.

3.2.3 Kaldi-põhine transkriptsioon (TalTech EST-ASR-Pipeline)

Lisaks end-to-end (läbivteenus, E2E) mudelitele kasutati töös ka traditsioonilist Kaldi baasil loodud kõnetuvastussüsteemi. Transkriptsioon viidi läbi TalTechi NLP-labori poolt arendatud `est-asr-pipeline` projekti abil (Alumäe, 2021; Olev and Alumäe, 2022).

Süsteemi kasutamiseks rakendati Docker-konteinerit `alumae/est-asr-pipeline:0.1.4`, mis sisaldas kogu vajalikku Kaldi ökosüsteemi koos keelemudelite ja tugiteekidega.

Kuna konteiner ei pakkunud otse kasutusvalmis transkribeerimise skripte, klooniti vastav töövoog GitHubi repositooriumist:

```
https://github.com/taltechnlp/est-asr-pipeline
```

ning transkriptsioon käivitati Nextflow töövoohalduri abil, kasutades faili `transcribe.nf`.

Transkribeerimiseks kasutatud helifailid olid standardiseeritud 16 kHz proovivõtusageduse, monoformaadi ja WAV-faililaiendiga. Töövoog koosnes mitmest järjestikusest etapist: esmalt teostati helifailide ettevalmistus ning keele tuvastus, millele järgnes kõne segmentimine. Seejärel ekstraheeriti akustilised tunnused - eelkõige melsageduse kepstraalkordajad - ning dekodeerimine toimus Kaldi hübriidarhitektuuri abil,

mis ühendab peidetud Markovi mudeli ja sügava närvivõrgu. Pärast esmast dekodeerimist viidi tulemused läbi täiendava hindamise rekurrentsel närvivõrgul (recurrent neural network, RNN) põhineva keelemudeli abil. Lõpetuseks muudeti dekodeeritud tekst loetavamaks, lisades automaatselt punktuatsiooni ja viimistledes väljundi tekstitasandil.

Kogu Kaldi-põhine tööprotsess toimus lokaalses tööjaamas, kus kasutati Nextflow versiooni 22.10.0. Lähenemise tugevuseks on süsteemi suur kohandatavus ja läbipaistvus, mis võimaldab seda vajadusel täiendada vastavalt domeenispetsiifilistele nõuetele. Samas on see tehniliselt nõudlikum kui kaasaegsed E2E-mudelid, nagu Whisper või Wav2Vec 2.0.

3.3 Mudelite võrdlusmetoodika

Transkriptsioonimudelite võrdlemisel keskenduti peamiselt täpsusele ja töötlemise kiirusele. Selleks kasutati järgmisi standardseid mõõdikuid:

- **Word Error Rate (WER)** - transkriptsiooni täpsuse mõõdik, mis arvestab kolme tüüpi vigu: *asendused* (S, substitutions), *kustutamised* (D, deletions) ja *lisamised* (I, insertions). WER arvutati võrrandiga:

$$\text{WER} = \frac{S + D + I}{N} \times 100\%,$$

  kus N on viitesõnade koguarv.

- **Realtime Factor (RTF)** - töötlemise kiirus, mis näitab, kui palju aega kulus heli töötlemiseks võrreldes selle tegeliku kestusega. Väärtus RTF < 1 viitab sellele, et mudel töötas kiiremini kui reaalajas.

Transkriptsiooni täpsuse hindamiseks kasutati sõnavigade määra ja selle komponente: asendused (S), kustutamised (D) ja lisamised (I). Arvutused viidi läbi Pythonis, kasutades `jiwer` teeki, mis võimaldab võrrelda mudelite transkriptsioone käsitsi koostatud korrektsete tekstidega (ingl. *ground truth*).

Veatüüpide täpsustamine Levensteini kauguse abil

Lisaks traditsioonilisele sõnavigade määrale (WER), mis arvestab kõiki asendusi, lisamisi ja kustutamisi võrdselt, viidi läbi ka täiendanalüüs, mis hindab konkreetsete asenduste suurust tähemärgipõhiselt Levensteini kauguse abil (*Levenshtein distance*). (Levenshtein, 1966)

Iga substitution-tüüpi operatsiooni (asendus) puhul arvutati kahe sõna vaheline Levensteini kaugus (tähemärkide põhjal) ning asendused jaotati kahte klassi:

- **Väikesed asendused (minor substitutions)** - Levensteini kaugus ≤ 2 tähemärki;

- **Suured asendused (major substitutions)** - Levensteini kaugus > 2 tähemärki.

Lävendiks valiti kaks tähemärki, kuna selline erinevus viitab enamasti foneetilistele tajuvigadele, morfoloogilistele kõikumistele või diktsiooniprobleemidele (nt "kindlustus" $\rightarrow$ "kinldustus"), mitte aga sisulisele tähendusmuutusele. Suuremad erinevused, nagu "kindlustus" $\rightarrow$ "kahju", viitavad tõenäoliselt valele sõna tuvastusele ning võivad mõjutada hilisemat semantilist analüüsi, sealhulgas sentimendi või eesmärgi määramist.

Analüüsi tulemused salvestati CSV-vormingus ning arvutati iga mudeli kohta nii väikeste kui suurte asenduste koguarv ja osakaal. See võimaldas võrrelda mitte ainult mudelite üldist täpsust, vaid ka eksimuste tõsidust ning hinnata, millised mudelid sobivad paremini edasiseks semantilise sisuga töötluseks.

Lisaks viidi läbi joonduspõhine võrdlus (ingl. *forced alignment*), mille eesmärk oli analüüsida, kui täpselt kattuvad transkribeeritud üksused algse helisalvestise segmentidega nii ajaliselt kui ka sõnalises järjestuses.

Joonduste kasutamine võimaldab tuvastada mustrilisi vigu, mis esinevad automaatses transkriptsioonis. Näiteks aitab see välja selgitada, kas transkribeeritud tekst kipub ajaliselt hilinema või kiirustama võrreldes algse heliga, kas mõni sõna on valesti tuvastatud ebatäpse häälduse tõttu või kas esineb regulaarselt sõnalisi lisamisi või kustutamisi kindlates kõnefaasides.

Kõik mudelid testiti identse sisendandmestiku ja töövoo tingimustes, et tagada õiglane ja võrreldav hindamine. Võrdlused viidi läbi samade transkriptsiooniskriptide ning eel- ja järelprotsessidega.

3.4 Sentimendi tööriistad

Kõnede transkriptsioonide emotsionaalse hoiaku (sentimendi) hindamiseks kasutatakse kahte lähenemist: lihtsamat leksikonipõhist meetodit ning keerukamat masinõppepõhist mudelit. Mõlema eesmärgiks on määrata, kas vestlus on emotsionaalselt positiivne, negatiivne või neutraalne.

Leksikonipõhine lähenemine. Esimene lähenemine põhineb Eesti Keele Instituudi valentsisõnastikul, mis sisaldab üle 38 000 eestikeelse sõna koos vastava sentimendiskooriga. Tekstide töötlemisel rakendatakse eeltöötlust, mille käigus eemaldatakse numbrid ja kirjavahemärgid ning sõnad viiakse kas tüvivormi (lemmatiseeritakse) või jäetakse algkujule sõltuvalt tööriistade saadavusest. Lemmatiseerimiseks kasutatakse `EstNLTK` teeki, mille olemasolul saab leksikoni sobitada ka morfoloogiliselt muundunud sõnadele.

Iga kõneanalüüsi käigus arvutatakse esmalt kogu transkribeeritud tekstile vastav üldskoor, mida tõlgendatakse edasi, võttes arvesse ka keelelisi ja semantilisi tegureid. Sentimendihinnangu täpsustamiseks arvestatakse muu hulgas keeleliste intensiivistajate,

nagu väga ja eriti, esinemist, mis võivad tugevdada lause emotsionaalset värvingut. Samuti analüüsitakse võimalike eituste - näiteks ei, mitte või pole - esinemist, mis võivad muuta sentimenti vastupidiseks. Lõplik arvutuslik hinnang skaleeritakse ja normaliseeritakse vahemikku [-1, 1], võimaldades kvantitatiivselt võrrelda erinevaid kõnesid ning jälgida ajas muutusi sentimentides.

Tulemuseks saadakse iga vestluse kohta üldine sentimendiskoor koos positiivsete ja negatiivsete sõnade osakaaluga. Lisaks tuuakse automaatselt esile kõige emotsionaalsemad laused - nii positiivsed kui ka negatiivsed - mis aitavad mõista kõne tonaalsuse äärmusi. Visualiseerimise eesmärgil genereeritakse ülevaatlikud graafikud, sealhulgas tulpdiagrammid ja ajas muutuvat sentimenti kujutavad joonised, mis toetavad analüüsi tõlgendamist nii kvantitatiivselt kui ka kvalitatiivselt.

Masinõppepõhine lähenemine. Teise lähenemisena kasutatakse süvaõppe mudelit, mis tugineb eeltreenitud keelemudelile RoBERTa. Selle meetodi eesmärk on klassifitseerida kõne transkriptsioonide emotsionaalne hoiak kogu teksti tähendusvälja põhjal, mitte ainult üksiksõnade sentimentide kaupa. Lähenemine sobib olukordades, kus väljendatud hoiak tuleneb sõnade koosmõjust või kontekstist (nt varjatud iroonia, mitmetähenduslikkus).

Käesolevas töös katsetati mitut eelnevalt treenitud sentimendianalüüsi mudelit:

- `tartuNLP/EstBERT128_sentiment` - eesti keelele kohandatud mudel, mis põhineb BERTi arhitektuuril ning on treenitud eestikeelsel sentimentandmestikul;
- `cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual` - mitmekeelne RoBERTa-põhine mudel, mis on treenitud suurandmestikel ja sobib sentimentide klassifitseerimiseks erinevates keeltes, sh eesti keeles;
- `nlptown/bert-base-multilingual-uncased-sentiment` - mitmekeelne BERT-põhine mudel, mille väljundiks on sentiment viiepalliskaalal; sobib üldiseks emotsioonihinnanguks;
- `ProsusAI/finbert` - ingliskeelne mudel, mis on peamiselt mõeldud finantstekstide sentimentide klassifitseerimiseks, kuid katsetati ka üldises kõnekontekstis.

Et käsitleda pikemaid vestlusi, mis ületavad mudeli sisendmahu piiranguid (nt 128 tokenit), jaotatakse transkribeeritud tekst väiksemateks osalõikudeks, kasutades libiseva akna meetodit (ingl. *strided windowing*). Iga lõiku analüüsitakse eraldi Hugging Face'i `transformers` teegi `TextClassificationPipeline` abil ning lõplik sentiment arvutatakse lõikude skooride keskmistamise teel. Tulemuseks saadakse mitmekihiline hinnang, mis sisaldab mudeli ennustatud üldhoiakut (positiivne, negatiivne või neutraalne), emotsionaalse hinnangu intensiivsust (positiivse ja negatiivse tõenäosuse vahe), täpset

tõenäosusjaotust kõigi kolme klassi vahel ning töödeldud lõikude arvu koos sentimentide sisemise variatsiooniga.

Kõik tulemused salvestatakse struktureeritud kujul: iga faili kohta luuakse JSON-vormingus detailne kirjeldus ning koondtabel CSV-kujul, mis võimaldab andmeid edasi analüüsida ja visualiseerida. Selline lähenemine pakub olulist eelist tänu oma keelelisele tundlikkusele ja kontekstitunnetusele, võimaldades usaldusväärsemalt tuvastada kliendisuhtluses esinevaid emotsionaalseid mustreid ka keerukamate väljendusviiside puhul.

Leksikonipõhise ja masinõppepõhise lähenemise kombineerimise eesmärk on võrrelda, milline meetod sobib paremini kindlustussektori kliendikõnede sentimentide üldiseks hindamiseks. Võrdluse käigus analüüsitakse esmalt, kui järjekindlalt ja tõlgendatavalt kumbki meetod suudab määrata kõne emotsionaalse hoiaku kogu transkribeeritud vestluse ulatuses. Eriti oluline on hinnata, kuidas süsteemid tuvastavad tugeva emotsionaalse alatooniga kõnesid, näiteks rahulolu, pettumust või frustratsiooni. Lisaks uuritakse, kuivõrd tõhusalt aitavad need lähenemised eristada informatsioonivaeseid ja ebaolulisi salvestisi, mille automaatne eemaldamine võimaldaks vähendada andmemahu haldust ning toetada isikuandmete töötlemise piirangute rakendamist.

3.5 Rakenduskeskkond ja töövoog

Kõik katsed viidi läbi lokaalses tööjaamas, mis võimaldas tagada protsesside täieliku kontrolli ja kordusanalüüsi võimalikkuse. Süsteemi konfiguratsioon oli järgmine:

- Protsessor: AMD Ryzen 5 5500U kuue tuumaga keskprotsessor (12 lõime), mis võimaldas paralleelset töötlemist transkriptsioonide segmentimisel;
- Muutmälu: 40 GB, mis osutus piisavaks nii Whisper kui Wav2Vec2 mudelite inference'i jooksutamiseks ilma GPU-ta;
- Salvestus: SSD-andmekandja, mis tagas kiire lugemis-/kirjutamiskiiruse helifailide ja tulemuste töötlemisel.

Kogu töötlusprotsess viidi läbi Pythonis kirjutatud skriptide abil. Valitud programmeerimiskeel ja tööriistad võimaldasid efektiivselt kombineerida nii automaatse kõnetuvastuse, loomuliku keele töötluse kui ka statistilise analüüsi.

Peamised kasutatud teegid ja töövahendid olid järgmised:

- `whisper`, `transformers`, `torchaudio`, `fairseq` - kõnetuvastuseks ning Whisperi ja Wav2Vec2 mudelite inference'i läbiviimiseks;
- `jiwer` - transkriptsioonide täpsuse hindamiseks, arvutades sõnavigade määra (WER) ning selle komponente: asendused (S), lisamised (I), kustutamised (D);

- `pandas`, `scikit-learn`, `matplotlib` - eksperimentide tulemuste talletamiseks, töötlemiseks ja visualiseerimiseks;
- `tqdm`, `os`, `time` - failide töötlemise tsüklite haldamiseks ja ajamõõtmisteks;
- `pydub` - helifailide formaadi konversiooniks ning eeltöötluseks, sh mono/stereo teisendused ja sagedusnormaliseerimine;
- `WhisperProcessor`, `WhisperForConditionalGeneration`, `Wav2Vec2Processor`, `Wav2Vec2ForCTC` - Hugging Face'i platvormi kaudu mudelite laadimiseks ning inference'i sooritamiseks.

Eksperimentaalne töövoog koosnes neljast selgelt eristuvast etapist:

1. **Helifailide ettevalmistus.** Kõik algsed MP3-formaadis helifailid teisendati 16 kHz mono WAV-formaadiks. Vajadusel lõigati failid segmentideks, et tagada kokkusobivus mudelite eeldatavate sisendite pikkustega. Samuti katsetati erinevaid müravähenduse ja normaliseerimise tehnikaid aga lõppanalüüsi kaasati vaid töötlemata, kuid formaadilt ühtlustatud helifailid.
2. **Automaatne transkriptsioon.** Failid transkribeeriti järjestikku kõigi valitud mudelitega. Transkriptsioonid salvestati tekstifailidena (UTF-8, plain text) ning igale väljundile lisati teabesilt, mis viitas kasutatud mudelile ja töötlemise parameetritele. Protsessi jooksul mõõdeti ka kestust, mille alusel arvutati hiljem reaalaegne tegur (RTF).
3. **Sentimendi- ja eesmärgianalüüs.** Transkribeeritud tekstid suunati sentimentide analüüsi plokki. Kasutati kahte lähenemist: leksikonipõhist meetodit (Eesti Keele Instituudi valentsisõnastik) ning masinõppel põhinevat lahendust (nt RoBERTa ja EstBERTi mudelid). Igale failile arvutati sentimentide jaotused ja domineeriv hoiak.
4. **Tulemuste analüüs ja koondamine.** Transkriptsioonide ja sentimentide tulemusi võrreldi käsitsi koostatud tõetekstidega (ingl. *ground truth*). Hindamiseks arvutati nii WER kui ka selle alamkomponendid ning teostati statistiline ja visuaalne analüüs. Tulemused koondati CSV-vormingus tabelitesse ning visualiseeriti jooniste ja boxplot-graafikutena.

Selline struktureeritud töövoog võimaldas läbipaistvat ja reprodutseeritavat eksperimenteerimist, tagades võrreldavuse kõigi mudelite vahel nii täpsuse, kiiruse kui ka semantilise väljundianalüüsi osas.

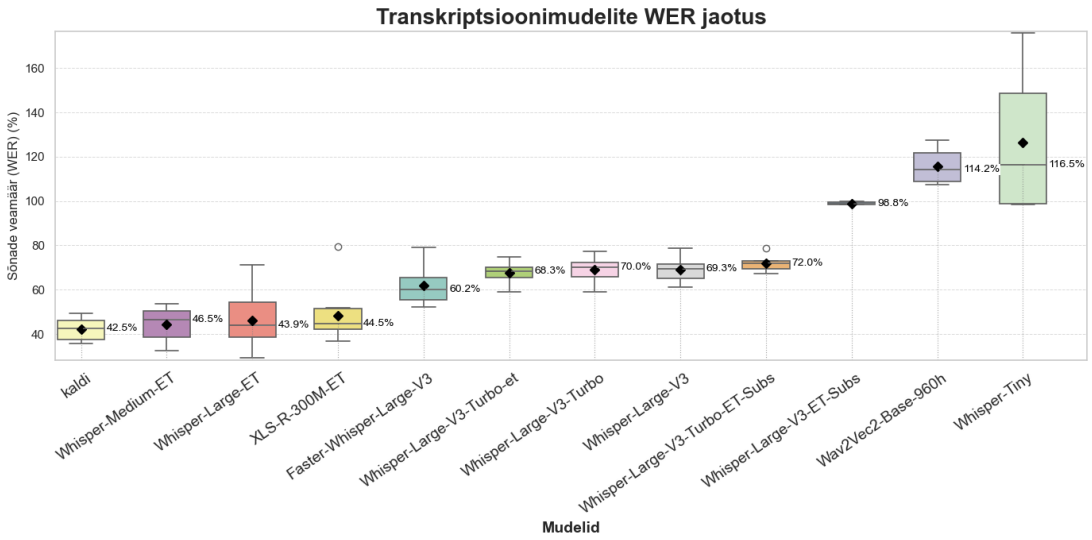
4 Eksperimentaalsed tulemused

Selles jaotises esitatakse automaatsete transkriptsioonimudelite võrdlus nii täpsuse kui töötlemiskiiruse alusel ning antakse esialgne hinnang transkribeeritud kindlustusettevõtte kõnesalvestiste põhjal teostatud sentimendi- ja eesmärgianalüüsi tulemustele.

4.1 Transkriptsioonimudelite täpsus

Transkriptsiooni kvaliteeti võrreldi sõnavea määra (WER) põhjal, mille arvutamisel võeti arvesse kolme tüüpi vigu: asendused (S), lisamised (I) ja kustutamised (D). Kõik mudelid said sisendiks samad helifailid, mis olid ühtlustatud 16 kHz monovormingusse. Eksperimentides kasutati nii peenhäälestatud (kohandatud) kui ka mitmekeelseid üldmudeleid, mille töökeel määrati osaliselt käsitsi (nt `language="et"`) ja osaliselt jäeti automaatse tuvastuse hooleks.

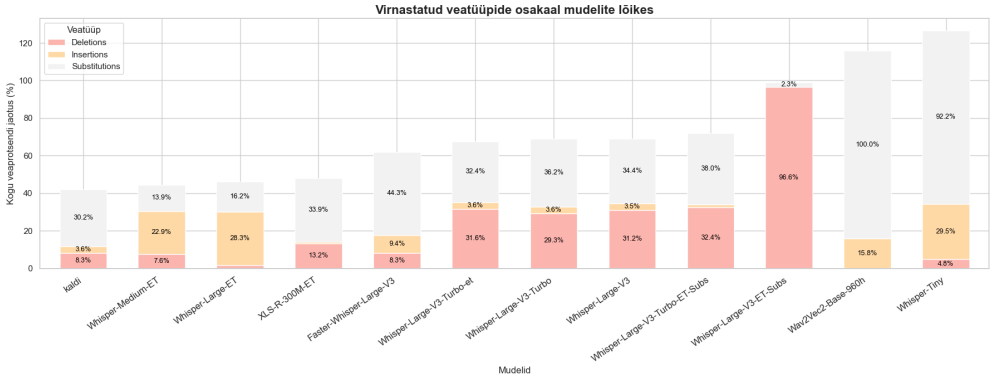
Joonisel 6 on kujutatud kõigi testitud mudelite WER-i jaotus. On näha, et parimad tulemused saavutati spetsiaalselt eesti keele jaoks treenitud mudelitega, samas kui üldmudelitel esines suurt hajuvust ja olulisi täpsuse kaotusi.



Joonis 6. Sõnavigade määra (WER) jaotus mudelite lõikes (boxplot).

Lisaks üldisele veamäärale hinnati, milliseid veatüüpe mudelid enim tekitasid. Joonisel 7 on toodud kumulatiivne jaotus. Eriti tihti esines mitmekeelsetel mudelitel lisamisi ja asendusi, mis viitab valedele sõnatuvastustele või liigsele täiendamisele.

Kuigi WER koos asenduste, lisamiste ja kustutamistega annab hea ülevaate üldisest täpsusest, ei kajasta see alati seda, kui sisuliselt olulised need vead on. Seetõttu hinnati täiendavalt transkriptsioonide täpsust Levensteini kauguse alusel, mis mõõdab sõnade



Joonis 7. WER-i komponentide kumulatiivne jaotus mudelite lõikes (kihiline tulpdiagramm).

erinevuse ulatust tähemärkide lõikes. Analüüsi eesmärk oli eristada väiksemad vead - näiteks kirjavead või morfoloogilised kõikumised - sisu moonutavatest asendustest.

Joonisel 8 on kujutatud iga mudeli keskmine sõnavigade määr (WER) võrreldуna väikeste vigade osakaaluga. Väikesteks loeti sellised erinevused, kus mudeli poolt tuvastatud sõna erines referentsist kuni kahe tähemärgi ulatuses.

Graafikult nähtub, et isegi sarnase WER-i korral võivad vigade sisulised omadused oluliselt erineda. Mõnel mudelil koosneb enamik vigu väikestest tähemärgilistest erinevustest (nt *kasko* → *kaso*), samas kui teistel esinevad sagedamini tähenduslikumad moonutused. Näiteks XLS-R-300M-ET mudel paistab silma kõrge väikeste vigade osakaalu poolest, mis viitab sellele, et tema transkriptsioonid jäävad semantiliselt tihti algteksti lähedale isegi siis, kui need pole sõnasõnaliselt täpsed. Selline erinevus on eriti oluline edasise semantilise analüüsi (nt sentimendi või eesmärgi määramise) usaldusväärsuse seisukohalt.

Kokkuvõtlik võrdlus on esitatud tabelis 3, kus on näidatud iga mudeli keskmist WER-i ja töötlemiskiirust (RTF).

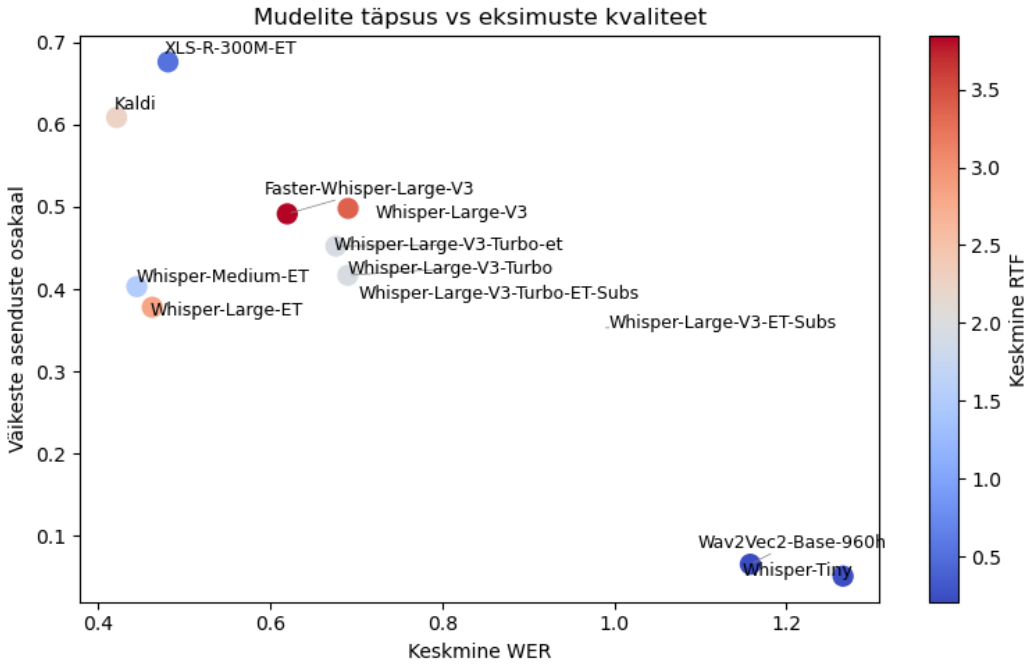
Parim täpsus saavutati Whisper-Large-ET mudeliga (WER 39,48 %), mis on OpenAI Whisperi suur mudel, peenhäälestatud eesti keelele TalTechi andmestiku alusel. Võrdluseks: sama mudeli mitmekeelne versioon ilma kohandamiseta - Whisper-Large-V3 - andis tulemuseks WER 69,03 %, mis näitab peenhäälestuse tähtsust.

Teise koha saavutas XLS-R-300M-ET, mis kuulub Wav2Vec2 mudeliperre ning on samuti eestikeelne versioon. Tema täpsus oli 41,12 %, kuid RTF kõigest 0,55, mis teeb sellest efektiivse valiku praktiliseks kasutuseks.

Kaldi DNN–HMM süsteem jäi täpsuselt kolmandaks (46,93 %), kuid selle tugevuseks on häälduss õnastiku kasutamine ja tulemuste reprodutseeritavus foneemitasandil.

Kõige kehvemini esinesid:

- Whisper-Tiny - väga kiire, ent ebatäpne (WER 126,58 %);



Joonis 8. Mudelite keskmine WER võrrelduna väikeste Levensteini erinevuste ($\leq$ 2 tähemärki) osakaaluga.

- Wav2Vec2-Base-960h - inglise keelele mõeldud mudel, mille WER oli 115,80 %;
- Whisper-Large-V3 ja Turbo - kuigi ressursimahukad, ei näidanud nad eesti keeles paremust võrreldes varem peenhäälestatud versioonidega.

Eksperimentides võrreldi ka seda, kas mudelitele anti keeleparameeter käsitsi (‘language="et„) või lasti mudelil keel automaatselt tuvastada. Tulemused viitasid sellele, et eesti keele määramine parandas mõningal määral Whisperi mudelite täpsust.

Kokkuvõttes näitab analüüs, et transkriptsioonimudeli valik sõltub otseselt kasutuskontekstist: maksimaalse täpsuse saavutamiseks tasub kindlasti eelistada peenhäälestatud mudeleid, ressursside piirangute korral on parim valik XLS-R-300M-ET.

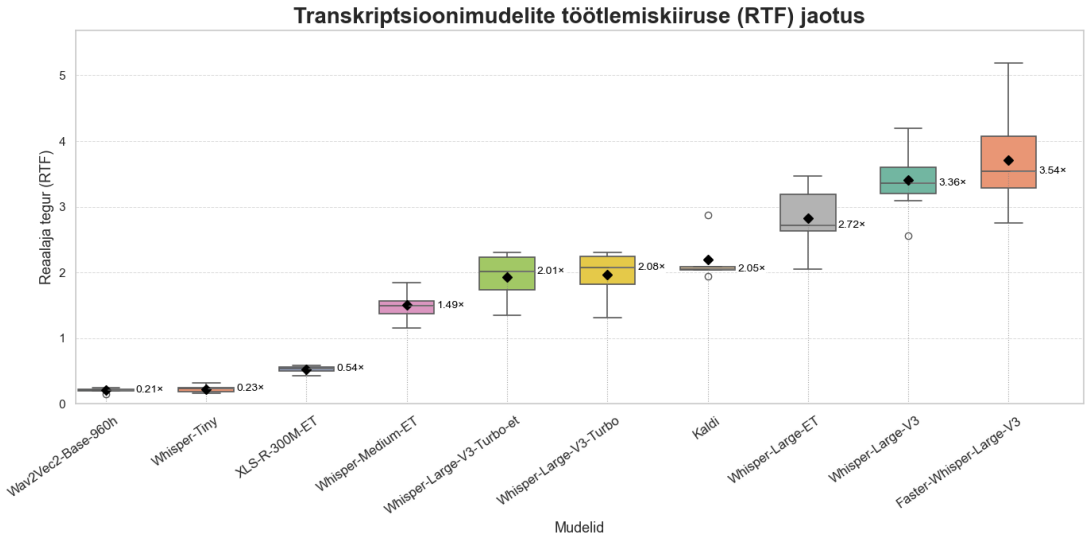
4.2 Töötlemiskiirus

Lisaks täpsusele hinnati mudelite efektiivsust ajakulu alusel. Transkriptsioonikiirust mõõdeti reaalaegse tegurina (RTF), mis näitab, kui palju aega kulub heli töötlemiseks võrreldes selle tegeliku kestusega. Näiteks RTF = 2 tähendab, et mudel vajab kaks korda rohkem aega kui heli pikkus, samas kui RTF < 1 viitab reaalajast kiiremale töötlemisele.

Tabel 3. Transkriptsioonimudelite keskmine WER ja RTF kõnesalvestiste peal.

Mudel	**WER (%)**	**RTF**
Whisper-Large-ET	39,48	3,12
XLS-R-300M-ET	41,12	0,55
Whisper-Medium-ET	44,47	1,63
Kaldi DNN–HMM	46,93	2,34
Faster-Whisper-Large-V3	59,08	4,14
Whisper-Large-V3-Turbo-et	67,62	2,08
Whisper-Large-V3	69,03	3,62
Wav2Vec2-Base-960h	115,80	0,22
Whisper-Tiny	126,58	0,22

Joonisel 9 on kujutatud mudelite RTF jaotus. Täpsus ja kiirus on sageli kompromissis: kiiremad mudelid on reeglina vähem täpsed ning täpsemad mudelid vajavad rohkem töötlemisaega.

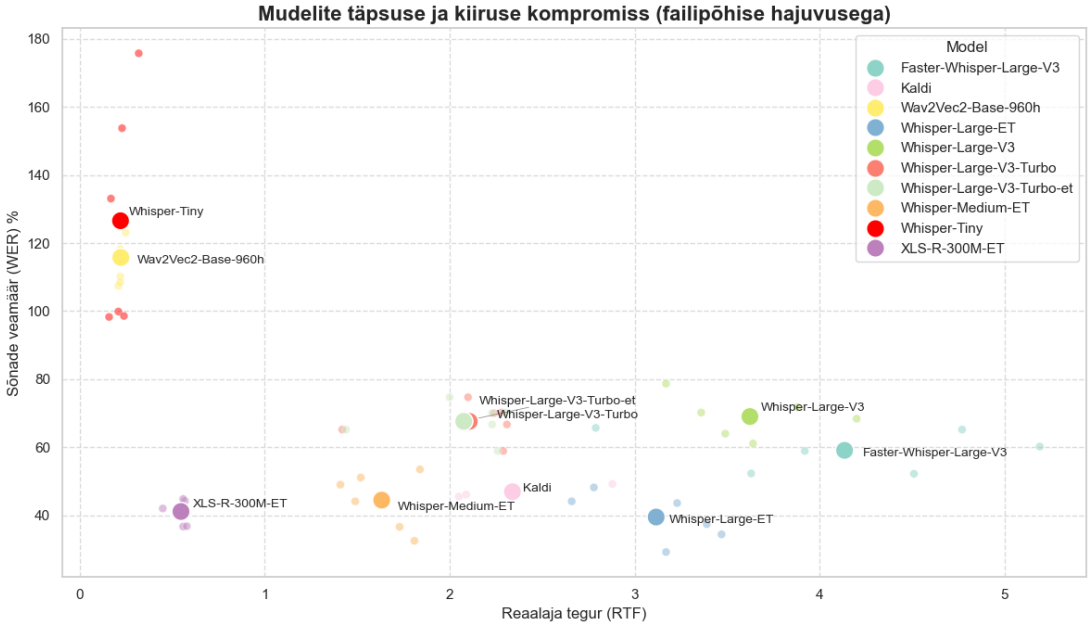


Joonis 9. RTF-i jaotus mudelite lõikes (boxplot).

Lisaks esitati täpsuse ja kiiruse omavahelist seost hajuvusjoonisena joonisel 10, kus mudelid paiknevad WER-RTF telgedel. See võimaldab hinnata mudelite sobivust konkreetsetesse kasutusstsenaariumitesse.

Olulisemad tähelepanekud:

- XLS-R-300M-ET paistis silma kui parim kompromissmudel, saavutades madala



Joonis 10. Mudelite paigutus täpsus-kiirus teljestikus (WER vs RTF).

WER (41,12 %) ja väga kiire töötlemise (RTF 0,55). Selle RTF väärtus jäi stabiilselt alla 1, võimaldades reaalajast kiiremat töötlemist ka CPU-põhises keskkonnas.

- Whisper-Tiny ja Wav2Vec2-Base-960h olid kõige kiiremad mudelid (RTF 0,22), kuid nende täpsus oli oluliselt alla vastuvõetava piiri (WER > 100 %).
- Whisper-Large-ET, kuigi kõige täpsem, oli ka üks aeglasemaid (RTF 3,12), mistõttu see ei sobi ressursipiiratud olukordadesse ilma GPU-ta.
- Faster-Whisper-Large-V3 ei suutnud oma nimes lubatud kiirust saavutada - selle RTF oli lausa 4,14, mis viitab sellele, et kiirendus ei realiseerunud praktikas CPU süsteemis.

Kõige tasakaalustatumaks valikuks osutus XLS-R-300M-ET, mis tagas mõõduka täpsuse ja väga hea töökiiruse. Kui eesmärgiks on maksimaalne täpsus ja ressursid ei ole piiratud, on soovitatav kasutada Whisper-Large-ET mudelit. Kui aga kiirus ja skaleeritavus on olulisemad, siis sobib paremini XLS-R mudel või vajadusel väiksem Whisper (nt Medium).

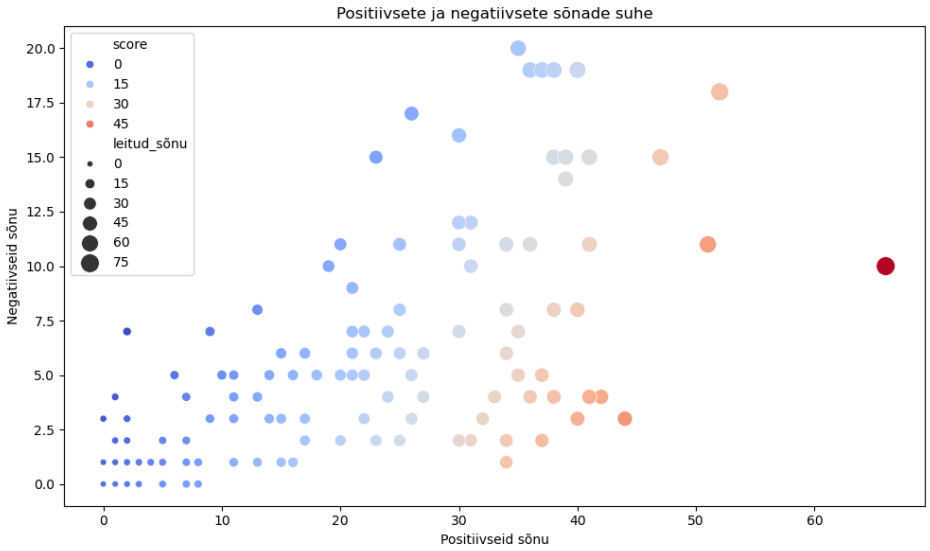
Sõltuvalt kasutusstsenaariumist - olgu selleks klienditeeninduse reaalajalisus, suurte arhiivide töötlemine või kvaliteedinõuded - tuleb langetada teadlik otsus täpsuse ja kiiruse vahel.

4.3 Sentimendianalüüs

Kõnede emotsionaalse hoiaku hindamiseks rakendati kahte erinevat lähenemist:

- **Leksikonipõhine lähenemine**, mis tugineb Eesti Keele Instituudi valentsisõnastikule ning sobib hästi morfoloogiliselt lihtsustatud tekstile. Igale sõnale on määratud positiivne või negatiivne skoor, mille põhjal arvutati kogu transkribeeritud vestluse sentiment.
- **Masinõppel põhinev lähenemine**, kus kasutati eeltreenitud sügavaid närvivõrke, nagu RoBERTa ja BERT-variantide arendused, sh `EstBERT`, `Multilingual-BERT`, `XLM-Roberta-Multi` ning finantsvaldkonnale kohandatud `FinBERT`.

Leksikonipõhise sentimendianalüüsi täpsemaks visualiseerimiseks on joonisel 11 kujutatud positiivsete ja negatiivsete sõnade omavaheline suhe iga transkribeeritud faili kohta. Punktide suurus tähistab leksikonist tuvastatud sõnade koguarvu ning värv esindab koguskoori - st positiivsete ja negatiivsete väärtuste vahet.



Joonis 11. Positiivsete ja negatiivsete sõnade suhe leksikonipõhise analüüsi puhul. Punktide suurus vastab leitud sõnade hulgale, värv tähistab üldskoori (pos - neg).

Enamus kõnesid sisaldas kuni 20 positiivset ja 10 negatiivset sõna, mis viitab sentimentide vähesele esinemissagedusele kõnesalvestistes üldiselt. Samas võib näha, et failid, kus leksikon kattus suure hulga sõnadega (nt üle 60), kaldusid andma tugevalt positiivse üldskoori, mis võib viidata spetsiifilistele kõnedele (nt müügivestlused või teeninduse tagasiside).

Punktide tihedus nulli ümber näitab, et paljud vestlused sisaldasid tasakaalus positiivseid ja negatiivseid väljendeid või jäid sentimenti osas määramatuks. See viitab leksikonipõhise lähenemise konservatiivsele loomusele - eriti dialoogilises kõnes, kus emotsionaalne toon ei pruugi väljenduda sõnatasandil selgelt.

Graafik toetab ka varasemat järeldust, et leksikonil põhinev meetod on küll tõlgendatav ja struktureeritud, kuid selle tundlikkus sentimentide tabamisel jääb piiratud, eriti keerukamate kõnekäändude või kaudsete hoiakute puhul.

Masinõppepõhise lähenemise puhul katsetati mitmeid eeltreenitud mudeleid, sealhulgas eestikeelne `EstBERT`, mitmekeelsed mudelid `XLM-Roberta-Multi` ja `Multilingual-BERT` ning ingliskeelsele finantstekstile kohandatud `FinBERT`.

Joonisel 12 on kujutatud iga mudeli sentimentide skoorid kõigi failide lõikes. Hinnanguks kasutati positiivse ja negatiivse skoori vahet, kus positiivne väärtus viitab valdavalt heatahtlikule ja negatiivne väärtus kriitilisele suhtumisele.

Joonis 12. Sentimendiskoorid failide lõikes eri mudelite puhul (pos - neg)

Tulemustest selgus, et mudelite hinnangutes esineb märkimisväärne varieeruvus, mis on oluline tähelepanek ilma valideeritud võrdlusandmestikuta kontekstis. Näiteks FinBERT andis valdavalt tugevalt positiivseid hinnanguid, kusjuures skoorid koondusid kitsasse vahemikku. EstBERT oli seevastu tugevalt negatiivne ning samuti madala hajuvusega. XLM-Roberta-Multi oli ainus mudel, mille hinnangud jaotusid laiemalt ja hõlmasid kogu spektrit alates negatiivsest kuni tugevalt positiivseni, mis võib viidata selle paremale kohanemisvõimele, kuigi ilma valideeritud andmestikuta ei saa me teha järeldusi mudeli täpsuse kohta. Multilingual-BERT jäi üldiselt neutraalseks ning näitas mõõdukat hajuvust.

Mudelite tulemuste nii suur erinevus on tähelepanuväärne ja näitab, et isegi kui transkriptsiooni täpsus on kõrge, võivad sellele järgnevad analüüsivahendid anda äärmiselt erinevaid tulemusi sõltuvalt nende tehnilisest ülesehitusest. Seetõttu on tulevikus vaja põhjalikku valideerimist spetsiifiliselt kindlustusvaldkonna eestikeelsete kõnede kontekstis, kus sentimentide tõlgendamine sõltub tugevalt kasutatava mudeli eelhäälestusest ja treeningandmete sobivusest. Praegune võrdlus annab ülevaate mudelite üldistest käitumismustritest, kuid ei võimalda veel teha järeldusi nende tegeliku täpsuse või usaldusväärsuse kohta.

4.4 Kokkuvõte

Eksperimentide tulemused kinnitasid, et väikekeelte automaatne töötlemine nõuab tasakaalu täpsuse, ressursikasutuse ja usaldusväärsuse vahel. Transkriptsioonimudelite võrdluses osutus kõige täpsemaks eestikeelsele andmestikule peenhäälestatud `Whisper-Large-ET` mudel, mille keskmine sõnavigade määr (WER) jäi 39,48% juurde. Samas oli selle mudeli töökiirus madal (RTF 3,12), mistõttu võib selle praktiline rakendamine olla piiratud ilma spetsiaalse riistvaratoeta.

Kõige parema täpsuse ja töökiiruse suhte pakkus `XLS-R-300M-ET`, mis kuulus *Wav2Vec 2.0* mudeliperre. Selle mudeli WER oli vaid 1,64 protsendipunkti kõrgem kui parimal tulemuse andnud mudelil, kuid töötlemiskiirus oli oluliselt parem (RTF 0,55). Seega sobib see mudel paremini olukordadesse, kus on vaja analüüsida suuremat helimahtu kiirelt ja stabiilselt.

Sentimendianalüüsi puhul ilmnes märgatav erinevus leksikonipõhise ja masinõppepõhise lähenemise vahel. Leksikonil põhinev analüüs osutus stabiilseks ja hästi tõlgendatavaks, eriti neutraalse sisuga vestluste puhul. Selle tugevuseks oli läbipaistvus: igale sõnale vastas konkreetne emotsionaalne kaal, mistõttu oli võimalik hinnata, millest tulemused koosnesid. Siiski jäi see lähenemine sageli hätta varjatud või kontekstist sõltuvate emotsioonide tuvastamisel, kuna paljud transkribeeritud failid sisaldasid vähe sentimentileksikonis esinevaid sõnu.

Masinõppemudelid (sh `FinBERT`, `XLM-Roberta-Multi`, `EstBERT`) pakkusid märksa varieeruvamaid hinnanguid ning suutsid tuvastada ka peenemaid hoiakuid. Samas ilmnes nende puhul mudelipõhine kallutatus: näiteks `FinBERT` kaldus tugevalt positiivsete hinnangute poole, samas kui `EstBERT` andis valdavalt negatiivseid tulemusi. See viitab vajadusele kontrollida, kuivõrd sobib iga mudel konkreetse keele ja domeeni jaoks. Lisaks raskendab nende tulemuste tõlgendamist asjaolu, et otsust loovate sisemiste kaalude ja tegurite kohta puudub selge nähtavus.

Kokkuvõttes sobib leksikonipõhine lähenemine esmasesse sõelumiseks ja neutraalsete kõnede hindamiseks, samas kui masinõppemudelid on paremad emotsionaalselt laetud juhtumite tuvastamisel. Kõige usaldusväärsema tulemuse võiks anda kahe lähenemise kombineeritud kasutus: esmalt leksikonipõhine analüüs kõnede üldiseks hindamiseks, seejärel suunamine süvaanalüüsi ainult neile kõnedele, mis viitavad emotsionaalsele või

riskantsele sisule. Selline hübriidne lähenemine tasakaalustab skaleeritavuse, tõlgendatavuse ja tundlikkuse nõudeid kindlustussektori kontekstis.

5 Arutelu

Selles jaotises analüüsitakse kriitiliselt katsete tulemusi, hinnates, mis töötas hästi, millised mudelid sobisid väikekeele (eelkõige eesti keele) konteksti, millised raskused ilmnesid ning kuivõrd töö alguses püstitatud hüpoteesid kinnitust leidsid.

5.1 Tulemuste kriitiline hindamine

Kõnetuvastuse vallas pakkus parimat sõnatasandi täpsust `Whisper-Large-ET`, mille peenhäälestus oli suunatud just eestikeelsetele andmetele. Samas oli selle mudeli tööaeg pikk (RTF > 3), mis muudab selle sobivamaks uurimuslikuks kasutuseks kui suurmahuliseks rakendamiseks. Parima täpsuse ja kiiruse kompromissi pakkus `XLS-R-300M-ET`, mille sõnavigade määr jäi alla 42% ja tööaeg oli kordades väiksem. See mudel sobib hästi praktilisteks lahendusteks, kus on oluline töökindlus ilma GPU-toeta.

Whisperi uuemad mitmekeelsed versioonid (nt `Large-V3`) ei toonud täiendavat täpsuse kasvu ning mitmel juhul esines probleeme automaatse keele tuvastusega. Samuti tõlgendasid väiksemad mudelid, nt `Whisper-Tiny`, kõnet ebaühtlaselt, viies WER-i üle 100%.

Sentimendianalüüsis näitasid masinõppepõhised mudelid suuremat tundlikkust tugeva emotsionaalse kallakuga kõnede tuvastamisel. Eriti `FinBERT` ja `XLM-Roberta-Multi` andsid selgeid ja stabiilseid hinnanguid, samas kui `EstBERT` kaldus tugevalt negatiivse hoiaku poole. Leksikonipõhine lähenemine, kuigi tõlgendatav ja morfoloogiliselt tundlik, jäi sageli neutraalseks ja suutis sentimenti tabada vaid siis, kui emotsionaalselt laetud sõnad kattusid sõnastiku kirjetega.

5.2 Eesti keele jaoks sobivaimad mudelid

Peenhäälestatud mudelid, mis olid otseselt treenitud eestikeelsete korpuste peal (nt `Whisper-Large-ET`, `XLS-R-300M-ET`), andsid oluliselt paremaid tulemusi võrreldes üldiste mitmekeelsete mudelitega. Nende täpsus oli parem ja transkribeeritud tekstid struktureeritumad ning stabiilsemad. Samuti ilmnes, et domeenikohane peenhäälestamine parandab mudelite käitumist - näiteks `FinBERT` sobis hästi müügisuunitlusega kõnede tuvastamiseks, kuigi selle üldistus väikekeele kontekstis võib jääda piiratuks.

5.3 Esinenud väljakutsed ja arendusvõimalused

Katsete läbiviimisel ilmnesid mitmed tehnilised ja metodoloogilised väljakutsed:

- **Kaldi** süsteemi ülesseadmine oli ajamahukas ning sõltus tugevalt Docker- ja Nextflow-põhisest töövoost, mis ei sobi hästi mitte-tehnilisele kasutajale.
- **Whisper** suuremad mudelid olid CPU-põhises seadmes väga aeglased, muutes suurte kõnepartiide töötlemise ebapraktiliseks.
- **Sentimendianalüüsi mudelite tulemused** osutusid vastuoluliseks, eriti neutraalse ja negatiivse hoiaku eristamisel, mis ilmnes ka mudelite vahelise varieeruvuse kaudu.
- **Leksikonipõhise katvuse** piiratus mõjutas sentimentide üldskoore - paljud failid ei sisaldanud piisavalt leksikonis leiduvaid sõnu, mis viis nullilähedaste hinnanguteni.

5.4 Vastus uurimisküsimustele

Töö käigus püstitatud uurimisküsimused said järgmised vastused:

- **Milline automaatne transkriptsioonimudel sobib kõige paremini eestikeelsete kindlustuskõnede transkribeerimiseks?**
  Transkriptsioonitäpsus oli parim mudelitel, mis olid otseselt peenhäälestatud eesti keelele. Eriti silmapaistev oli Whisper-Large-ET mudel (WER 39,48%), millele järgnes XLS-R-300M-ET (41,12%). See kinnitab, et väikekeele puhul on keeleline spetsialiseeritus kriitilise tähtsusega. Parima kiiruse ja täpsuse suhtega oli XLS-R-300M-ET, mis sobib hästi praktiliseks kasutamiseks.
- **Milliseid hinnanguid annavad erinevad sentimendianalüüsi mudelid transkribeeritud kindlustusvaldkonna kõnedele ja kuivõrd erinevad need hinnangud üksteisest kontekstis, kus valideeritud andmestik puudub?**
  Sentimendianalüüsi mudelite väljundid erinesid märkimisväärselt – FinBERT andis valdavalt positiivseid hinnanguid, EstBERT negatiivseid, XLM-Roberta-Multi kasutas kogu sentimendiskaalat ning Multilingual-BERT jäi valdavalt neutraalseks. See viitab, et eri mudelitel on tugevad eelhäälestuslikud kalduvused, mis mõjutavad nende hinnanguid. Leksikonipõhine lähenemine andis stabiilsemaid, kuid sageli neutraalsemaid tulemusi.

  Oluline on märkida, et ilma valideeritud võrdlusandmestikuta ei saa me veel teha järeldusi, milline mudel on kõige täpsem. XLM-Roberta-Multi laiem väljundite jaotus võib viidata suuremale kohanemisvõimele, kuid selle tegelikku täpsust tuleks tulevikus hinnata spetsiaalselt märgendatud andmetel. Tulevases töös on vajalik luua kindlustusvaldkonna kõnedele spetsiifiline hindamisraamistik, et määrata, milline sentimendianalüüsi lähenemisviis on täpseim ja usaldusväärseim.

5.5 Jätkutöö ja rakendusvõimalused

Edasised arendussuunad hõlmavad mitut tehnilist, metoodilist ja organisatsioonilist aspekti:

- **Shiny-põhise juhtpaneeli (dashboard) loomine**, mis võimaldaks kasutajatel visualiseerida WER-i, sentimentide jaotust, kõnede pikkust, eesmärki ja potentsiaalset relevantsust. Selline rakendus toetaks sisekommunikatsiooni, andmehaldust ja tagasisideprotsesse.
- **Multimodaalsete suurte keelemudelite (LLM) kasutuselevõtt**, sh natiivne transkribeerimine ja eesmärgipõhine klassifitseerimine (nt kas kõne on seotud müügi, teeninduse või kahjukäsitlusega). Kuigi käesolevas töös ei olnud võimalik LLM-e rakendada arvutusvõimekuse ja andmekaitseliste piirangute tõttu, pakuvad need tulevikus potentsiaali eelkõige keerulisemate vestluste semantilisel mõistmisel.
- **Eesmärgi tuvastamise süsteemi arendamine tulevikus**, mis käesolevas töös ei olnud fookuses. See eeldab kindlustusettevõtte siseselt selget ja ühtset definitsiooni erinevate kõneliikide kohta (nt müügikõned, kahjukäsitluskõned või oluline ja ebaoluline). Selleks on vaja luua struktureeritud klassifikatsioonisüsteem ning koguda märgendatud andmeid, mille alusel saaks mudeli treenida ja valideerida. Sentimendianalüüsi tulemused võiksid pakkuda täiendavat sisendit ka sellisele eesmärgituvastusele.
- **Andmekaitse- ja kustutamisloogikate arendamine**, et hinnata mitte ainult kõne sisu, vaid ka selle säilitamise vajadust GDPR-i põhimõtetest lähtudes. Tulevikulahendus peaks olema võimeline tuvastama ebaolulised või aegunud salvestised ning integreeruma ettevõtte siseste arhiveerimis- ja kustutuspoliitikatega.
- **Eesti keele mudelite täiendav peenhäälestus**, sh Wav2Vec2 ja RoBERTa-põhiste mudelite treenimine ettevõtte oma andmestiku põhjal. Selle abil oleks võimalik suurendada mudelite domeenikohasust ning vähendada üleanalüüsimise või valepositiivsete tulemuste osakaalu.

Kokkuvõttes loob käesolev töö aluse praktilisele süsteemile, mis suudab automatiseeritult hinnata kõnede olulisust, eesmärki ja emotsionaalset hoiakut ning toetada andmepõhiseid otsuseid kindlustussektoris. Tulevikutöö keskendub süsteemi skaleeritavusele, selgitatavusele, vastavustele reaalsete ärivajadustega ning regulatiivsetele ootustele.

Viidatud kirjandus

Alumäe, T. e. a. (2021). Estonian asr pipeline. https://github.com/taltechnlp/est-asr-pipeline. Accessed: 01.05.2025.

Asu, E. L. and Teras, P. (2009). Illustrations of the ipa: Estonian. *Journal of the International Phonetic Association*, 39(3):367–372.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

European Parliamentary Research Service (2020). Artificial intelligence: How does it work, why does it matter, and what can we do about it? European Parliament, PE 641.530.

Fedorchenko, A. and Alumäe, T. (2025). Optimizing estonian TV subtitles with semi-supervised learning and LLMs. In *Proceedings of the 25th Nordic Conference on Computational Linguistics (NoDaLiDa)*.

Feelingstream (2021). Feelingstream & medicum – reach 2021 technical slidedeck. https://www.reach-incubator.eu/wp-content/uploads/2021/05/Feelingstream-Medicum-REACH-2021-technical-slidedeck.pdf. Külastatud: 09.05.2025.

Feelingstream (2024). Speaker app – real-time speech to text. https://speaker.feelingstream.com. Külastatud: 09.05.2025.

Georgescu, A., Pappalardo, A., Cucu, H., and Blott, M. (2021). Performance vs. hardware requirements in state-of-the-art automatic speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021.

Guo, X. (2024). Sentiment analysis based on roberta for amazon review: An empirical study on decision making.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Koto, F., Beck, T., Talat, Z., Gurevych, I., and Baldwin, T. (2024). Zero-shot sentiment analysis in low-resource languages using a multilingual sentiment lexicon.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Nautsch, A., Jasserand, C., Kindt, E., Todisco, M., Trancoso, I., and Evans, N. (2019). The gdpr & speech data: Reflections of legal and technology communities, first steps towards a common understanding. In *Interspeech 2019*, interspeech_2019. ISCA.

Olev, A. and Alumäe, T. (2022). Estonian speech recognition and transcription editing service. *Baltic Journal of Modern Computing*, 10(3):409–421.

Olev, A. and Alumäe, T. (2024). Open source platform for estonian speech transcription. *Language Resources and Evaluation*. Published online: 2024-10-16.

Oxus.AI (2021). Oxus.ai raises €800k to bring advanced speech analytics solution for sales teams to 7,000 languages. `https://www.oxus.ai/post/manage-your-blog-from-your-live-site`. Külastatud: 09.05.2025.

Parikh, A., ten Bosch, L., van den Heuvel, H., and Tejedor-Garcia, C. (2023). Comparing modular and end-to-end approaches in ASR for well-resourced and low-resourced languages. In Abbas, M. and Freihat, A. A., editors, *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 266–273, Online. Association for Computational Linguistics.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision.

Raees, M. and Fazilat, S. (2024). Lexicon-based sentiment analysis on text polarities with evaluation of classification models.

Rehm, G. and Way, A. (2023). *European Language Equality: Introduction*, pages 1–10. Springer International Publishing, Cham.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. pages 3465–3469.

Stevens, S.Š., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190.

SYSTRAN (2023). Faster whisper large v3. https://huggingface.co/Systran/faster-whisper-large-v3. Accessed: 01.05.2025.

TalTechNLP (2022). xls-r-300m-et. https://huggingface.co/TalTechNLP/xls-r-300m-et. Accessed: 01.05.2025.

TalTechNLP (2023a). whisper-large-et. https://huggingface.co/TalTechNLP/whisper-large-et. Accessed: 01.05.2025.

TalTechNLP (2023b). whisper-medium-et. https://huggingface.co/TalTechNLP/whisper-medium-et. Accessed: 01.05.2025.

Tilde (2023). Estonian courts shift to automated transcription with salme. https://tilde.ai/asr-solution-for-estonian-courts. Külastatud: 09.05.2025.

Weitzenboeck, E. M., Lison, P., Cyndecka, M., and Langford, M. (2022). The gdpr and unstructured data: is anonymization possible? *International Data Privacy Law*, 12(3):184–206.

Lisad

I. Glossary

Andmekaitse üldmäärus (ingl. *General Data Protection Regulation*, GDPR) - Euroopa Liidu määrus, mis sätestab isikuandmete töötlemise reeglid ja privaatsuse tagamise nõuded.

Automaatne kõnetuvastus (ingl. *Automatic Speech Recognition*, ASR) - tehnoloogia, mis võimaldab arvutil tuvastada ja transkribeerida inimkõnet tekstiks.

BERT (ingl. *Bidirectional Encoder Representations from Transformers*) - Google'i arendatud eeltreenitud keeleesitusmudel, mis põhineb transformer-arhitektuuril ning suudab mõista konteksti nii vasakult kui paremalt. Seda kasutatakse laialdaselt loomuliku keele töötluses ülesannetes nagu teksti klassifitseerimine, märgendamine ja küsimustele vastamine.

CTC algoritm (ingl. *Connectionist Temporal Classification*) - algoritm, mis võimaldab treenida närvivõrke järjestusülesannetes ilma täpse joonduseta sisendi ja väljundi vahel.

Encoder-decoder transformer mudel - süvaõppemudel, kus kodeerija (encoder) teisendab sisendi vaheesituseks ja dekodeerija (decoder) genereerib vastuse; kasutatakse kõnetuvastuses ja masintõlkes.

Fine-tuning (eestikeelne: *peenhäälestamine*) - eelnevalt treenitud mudeli kohandamine konkreetse andmestiku jaoks.

Forced alignment (eestikeelne: *sundjoondus*) - meetod, millega ajaliselt joondatakse helisignaal ja selle transkriptsioon.

Gaussi segu mudel (ingl. *Gaussian Mixture Model*, GMM) - statistiline mudel, mida kasutatakse akustiliste tunnuste tõenäosusjaotuste modelleerimiseks kõnetuvastuses.

Keelemudel (ingl. *language model*) - statistiline või tehisnärvivõrgul põhinev mudel, mis hindab sõnade järjestuse tõenäosust ja aitab kõnetuvastussüsteemidel valida kontekstis kõige sobivamaid sõnu.

Latendiesitus (ingl. *latent representation*) - tihendatud ja informatiivne andmete esitusvorm, mida kasutatakse näiteks helisignaali sisuliseks kokkuvõtteks.

Levenshteini kaugus (ingl. *Levenshtein distance*) - meetrika, mis mõõdab kahe sõne vahelise erinevuse suurust tähtede või sümbolite lisamiseks, eemaldamiseks või asendamiseks vajalike operatsioonide minimaalse arvu kaudu.

Loomuliku keele töötlemine (ingl. *natural language processing*, NLP) - arvutiteaduse haru, mis tegeleb inimkeele analüüsi ja mõistmisega.

Läbivteenus arhitektuur (ingl. *end-to-end architecture*, E2E) - mudel, mis töötleb andmeid otse sisendist väljundini ilma eraldiseisvate vahemoduliteta.

Mel-spektrogramm (ingl. *mel spectrogram*) - helisignaali sageduslik kujutus, mille teljestik põhineb inimese kuulmistajul (mel-skaalal).

MFCC (ingl. *Mel-Frequency Cepstral Coefficients*) - akustilised tunnused, mida kasutatakse laialdaselt kõnetuvastuses ja mis põhinevad helisignaali mel-sageduse spektri analüüsil.

Peidetud Markovi mudel (ingl. *Hidden Markov Model*, HMM) - tõenäosuslik mudel järjestikuste sündmuste modelleerimiseks, kus osa olekuid ei ole otseselt jälgitavad.

Peidetud olek (ingl. *hidden state*) - olek mudelis, mis ei ole otseselt vaadeldav, kuid mõjutab väljundeid; nt HMM puhul kõneolekud.

Reaalaja tegur (ingl. *realtime factor*, RTF) - töötlemise kiiruse mõõdik, mis näitab, kui kiiresti mudel suudab heli töödelda võrreldes selle pikkusega.

RoBERTa (ingl. *Robustly Optimized BERT Pretraining Approach*) - täiustatud versioon BERT mudelist, mis on optimeeritud parema jõudluse saavutamiseks erinevates keeleülesannetes.

Sentimendianalüüs (ingl. *sentiment analysis*) - loomuliku keele töötlemise tehnika, mis tuvastab ja klassifitseerib emotsionaalset tooni tekstis, näiteks positiivne, negatiivne või neutraalne.

Süvaõpe (ingl. *deep learning*) - masinõppe haru, mis kasutab mitmekihilisi närvivõrke keeruliste mustrite õppimiseks andmetes.

Tähelepanumehhanism (ingl. *attention mechanism*) - transformer-mudelite oluline komponent, mis võimaldab mudelil keskenduda sisendi erinevatele osadele väljundit genereerides.

Töövoog (ingl. *pipeline*) - protsesside järjestus, kus andmeid töödeldakse samm-sammult kindlas järjekorras.

WER (ingl. *Word Error Rate*) - sõnavea määr, mis mõõdab transkriptsiooni täpsust, arvutades asenduste, lisamiste ja kustutamiste summa ja jagades selle kogu sõnade arvuga.

Zero-shot transkriptsioon - võime transkribeerida keelt, millel mudel ei ole otseselt treenitud (nt Whisperi puhul väikekeeled nagu eesti keel).

II. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Kevin Lehtsalu**,
*(*autori nimi)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

   Kindlustusettevõtte kõnede automaatne transkriptsioon ja sentimendi analüüs,
   *(*lõputöö pealkiri)

   mille juhendaja(d) on Kairit Sirts,
   *(*juhendaja nimi)

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Kevin Lehtsalu
15.05.2025