TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

**Katariina Parkja**

# Tehisintellekti abiga kõne põhjal piltide genereerimine

**Bakalaureusetöö (9 EAP)**

Juhendaja:
Ardi Tampuu, PhD

Tartu 2025

## Tehisintellekti abiga kõne põhjal piltide genereerimine

**Lühikokkuvõte:**

Bakalureusetöö eesmärk oli arendada Delta õppehoones eksponeeritav rakendus, mis demonstreerib tehisintellektil põhinevaid kõnetuvastuse ja piltide genereerimise tehnoloogiaid ühtse töövoona. Töö tulemusena valmis lokaalselt töötav programm, mis võimaldab kasutajal genereerida pilte nii eesti- kui ingliskeelse kõne põhjal. Demoperioodil testiti lahenduse ingliskeelset versiooni reaalsete kasutajatega. Töö annab ülevaate katsetatud kõnetuvastuse, keeletuvastuse, masintõlke ja pildiloome mudelitest, keskendudes lahendustele, mis toimivad ilma internetiühenduseta. Kirjeldatakse kasutatud tehnoloogiate valikut ning analüüsitakse süsteemi töökindlust ja kasutajate tagasisidet.

**Võtmesõnad:** Tehisintellekt, kõnetuvastus, pildigeneratsioon, Whisper, Stable Diffusion

**CERCS:** P176 Tehisintellekt

## Image Generation Based on Spoken Input Using AI

**Abstract:**

The aim of this bachelor’s thesis was to develop an application to be exhibited in the Delta academic building, which demonstrates artificial intelligence-based speech recognition and image generation technologies as a unified workflow. As a result of the work, a locally operating program was created that allows users to generate images based on both Estonian and English speech. During the demo period, the English version of the solution was tested with real users. The thesis provides an overview of the speech recognition, language detection, machine translation, and image generation models that were tested, focusing on solutions that work without an internet connection. The thesis also describes the selection of the used technologies and analyzes the system's reliability and user feedback.

**Keywords:** artificial intelligence, automatic speech recognition, image generation, Whisper, Stable Diffusion

**CERCS:** P176 Artificial intelligence

# Sisukord

# 1. Sissejuhatus

Järgnev lõik põhineb Bengesi jt artiklil arengutest generatiivse tehisintellekti valdkonnas [1]. Generatiivsete vastandvõrkude (ingl *generative adversarial network* ehk GAN) kasutuselevõtt 2014. aastal tähistas uue generatiivse tehisintellekti (ingl *generative artificial intelligence* ehk GAI) loomise ajastut. Enne seda olid sügavõppe mudelid peamiselt kirjeldavad, loodud eesmärgiga selgitada andmemustreid ja teha olemasoleva info põhjal ennustusi. Erinevalt kirjeldavatest mudelitest on generatiivsete mudelite eesmärk luua uusi andmeelemete, mis sarnanevad treeningandmetes täheldatud mustritele. ChatGPT avaldamine 2022. aastal tõi kaasa GAI suure populaarsuse kasvu laiema avalikkuse seas. Lisaks suurtele keelemudelite on GAI valdkonnas muljetavaldavaid tulemusi saavutatud kõnetuvastuses ja piltide genereerimisel.

Selle bakalauseusetöö eesmärk on luua rakendus, mis demonstreerib kõnetuvastuse ja tekstipõhise pildiloome tehisintellekti tehnoloogiaid ühtse töövoona meelelahutusliku demoprojekti vormis. Lõplik lahendus annab kasutajale võimaluse tehisintellekti abil pilte genereerida nii eesti- kui ka ingliskeelse suulise kirjelduse põhjal. Demo eesmärk on pakkuda meelelahutust, tutvustada Delta külastajatele tehisintellekti tehnoloogiad, näidata, et selline tehnoloogia on kättesaadav ja kasutatav mitte ainult suurtele ettevõtetele, vaid bakalaureuse taseme töös.

Töö Tausta peatükis tutvustatakse lühidalt tehisintellekti põhiseid äratussõna tuvastamise, kõnetuvastuse ja pildiloome tehnoloogiaid. Meetodite peatükk annab ülevaate rakenduse töövoost, kasutatud riistvarast ning katsetatud lähenemistest kasutaja pöördumise tuvastamiseks, kõnetuvastuseks, masintõlkeks ja piltide genereerimiseks. Tulemuste peatükis kirjeldatakse rakenduses kasutatud tehnoloogiate valikut; lisaks antakse ülevaade kõnetuvastuse ja pildiloome moodulite tulemustest ning kasutajate tagasisidest. Viimases peatükis antakse üldine ülevaade töö tulemustest, analüüsitakse valminud programmi puuduseid ja kirjeldatakse selle edasiarendamise võimalusi. Lisades on tagasiside vormi küsimused ja link loodud programmi koodile.

Lõputöö kirjutamisel on kohati kasutatud ChatGPT 4o mudeli abi sõnastuse parandamiseks.

# 2. Taust

Kõne tuvastamine ja selle põhjal piltide genereerimise töövoog koosneb mitmest sammust – alates kasutaja pöördumise tuvastamisest kuni pildi genereerimiseni. Selles peatükis tutvustatakse lähemalt tehisintellektil põhinevaid tehnoloogiaid äratussõna tuvastamiseks, kõne transkribeerimiseks ja teksti alusel piltide genereerimiseks.

## 2.1 Äratussõna tuvastamine

Äratussõna (ingl *wake word, hot word* või *trigger word*) tuvastamine on üks keskseid komponente hääljuhtimisega süsteemides. See on protsess, mille ülesandeks on pidevalt jälgida helisisendit, et tuvastada kindel eelmääratud fraas [2]. Mõned tuntumad näited hääljuhitavates süsteemides kasutatavatest äratussõnadest on „OK Google“, „Hey Siri“ ja „Alexa“ [2, 3]. Fraasi tuvastamine annab süsteemile märku, millal alustada aktiivset häältuvastust.

Üks selline mudel, mida saab mitteärilisetel eesmärkidel ka tasuta kasutada on Picovoice'i Porccupine[1]. Porcupine'i kasutamise teeb arendaja jaoks eriti mugavaks see, et enda projektile kohandatud äratussõna on võimalik Picovoice Console'i keskkonnas luua vaid mõne sekundiga[1]. Teiste sarnaste mudelite puhul võib olla vajalik ise treeningandmete genereerimine (näiteks openWakeWord[2] puhul) või on vaja mudeli loojatega ühendust võtta (näiteks DaVoice.io[3]).

Äratussõna tuvastamisele järgneb tavaliselt häälkäskluste kuulamise samm ehk järgnevale kasutaja kõnele rakendatakse kõnetuvastust [2], kuid näiteks targa kodu lahendustes võib sellele järgneda ka mõni muu tegevus nagu näiteks tulede põlema panemine või muusika mängimine [4].

## 2.2 Kõnetuvastus

Automaatne kõnetuvastus (ingl *automatic speech recognition*) on tehisintellekti valdkond, mis tegeleb helisignaali põhjal kõne teisendamisega kirjalikuks tekstiks. Umbes alates 2010. aastast saavutavad sügavad närvivõrgud kõnetuvastuses paremaid tulemusi, kui varasemalt standardiks olnud statistilised mudelid [5]. Algselt kasutati sügavaid närvivõrke koos peidetud Markovi mudelitega, kuid üks olulisemaid arenguetappe oli otsast-lõpuni (ingl *end-to-end*)

[1] https://picovoice.ai/platform/porcupine/
[2] https://github.com/dscripka/openWakeWord?tab=readme-ov-file
[3] https://github.com/frymanofer/Python_WakeWordDetection

mudelite kasutuselevõtt [6]. Need mudelid võimaldasid otse heli tekstiks teisendamise, ilma et oleks vaja vahepealseid foneetilisi esitusi.

Tänapäeval domineerivad kõnetuvastuse valdkonnas transformerite arhitektuuri rakendavad lahendused [7] nagu OpenAI Whisper [8], Google'i Conformer [9]. Automaatse kõnetuvastuse teenust pakuvad ka näiteks Microsoft Azure Speech Service[4] ja Amazon Transcribe[5], kuid nende täpset arhitektuuri pole avalikustatud. Moodsad mudelid suudavad mitte ainult tuvastada ning kirja panna helisid, vaid sisaldavad piisavalt teadmist sihtkeelest, et eemaldada transkriptsioonist mitte-tähenduslikke häälitsusi ning lisada vajalikke kirjavahemärke [8, 10]. Neist arvatavasti enim kasutatav on OpenAI loodud Whisper, mida eristab teistest toodud näidetest see, et see on vabavaraline[6].

Whisperi loojad ütlevad loodud mudeli arhitektuuri ja tulemusi tutvustavas töös [8], et kõnetuvastuse süsteemi eesmärk peaks olema toimida usaldusväärselt erinevates kasutuskeskkondades, ilma, et oleks vajadust juhendatud peenhäälestamiseks. Selle saavutamiseks keskenduvad nad oma töös lisaks treeningandmete mahu suurendamisele ka ingliskeelse kõnetuvastuse mitmekeelseks ja mitmeülesandeliseks laiendamisele – 680 000 tunni jagu treeningandmeid sisaldab lisaks ingliskeelsetele andmetele 117 000 tundi audioandmeid 96 muus keeles ja 125 000 tundi tõlkeandmeid. Nad leiavad, et piisavalt suurte mudelite puhul ei kaasne mitmekeelse ja mitmeülesandelise treenimisega negatiivseid külgi nagu täpsuse langus ühe või teise ülesande lahendamisel.

Kuigi Whisper on treenitud rohkem kui 90 keele peal, on täpsused varieeruvad ja seega toetab Whisper OpenAI sõnul large mudeli tulemuste põhjal 57 keelt, mille seas on ka eesti keel[7]. Teisisõnu Whisper võib luua transkriptsioone rohkem kui 90 keelele, kuid kõik need ei lähe toetatud keelte alla. Whisperi kasutajate kogukond[8] on püüdnud ka välja selgitada, milliseid keeli base mudel transkribeerida suudab. Vastavasisulises blogipostituses kirjeldatakse, kuidas base mudeli toetatud keelte välja selgitamiseks küsitleti 98 keele jaoks kahte kuni kolme seda keelt emakeelena kõnelevat inimest ja paluti hinnata kolme video transkriptsiooni kvaliteeti skaalal 0-5. Eestikeelse kõne transkribeerimisele Whisper base mudeliga anti hindeks 0, mis

[4] https://learn.microsoft.com/en-us/azure/ai-services/speech-service/
[5] https://aws.amazon.com/transcribe/
[6] https://github.com/openai/whisper
[7] https://platform.openai.com/docs/guides/speech-to-text/#supported-languages
[8] https://blog.merjek.com/2024/05/14/openai-is-wrong-they-do-not-support-over-90-languages-with-their-whisper-module/

tähendas, et transkriptsioonid olid täiesti arusaamatud. Samas peaksid tulemused olema Whisperi suuremate mudelite puhul palju paremad [8], seega väärib see siiski katsetamist.

## 2.3 Piltide genereerimine

Piltide genereerimine tähistab siinkohal generatiivse tehisintellekti alamvaldkonda, kus tehisintellekti mudel õpib suurtest pildiandmekogudest mustreid ning on seejärel võimeline sünteesima uusi visuaale, mis meenutavad õppematerjalina kasutatud andmeid. Selles töös keskendutakse tekstilise sisendi põhjal piltide genereerimisele (ingl text-to-image generation ehk TTI).

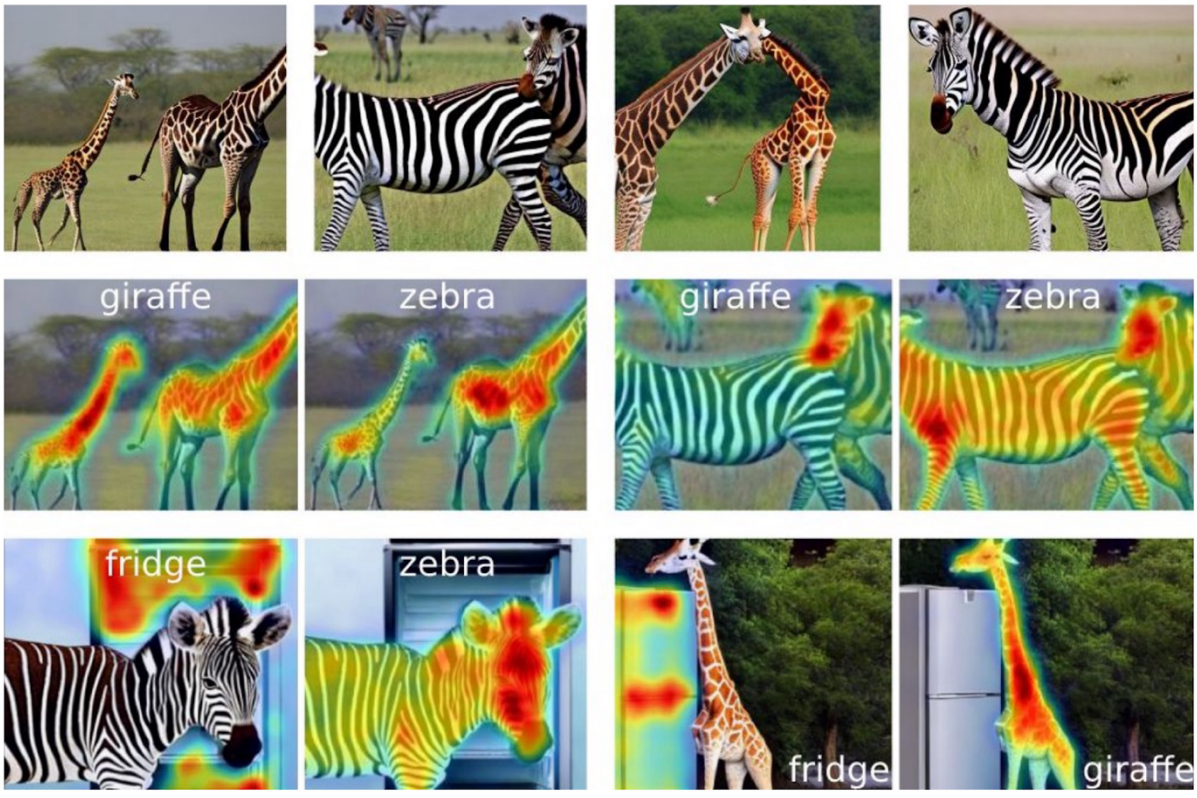
Esimesed edukamad lähenemised tekstist pildi genereerimisel põhinesid generatiivsetel vastandvõrkudel [11]. Nende tugevuseks on pildi genereerimise kiirus, kuid täpsuse ja detailirohkuse osas jäävad alla uuematele lähenemisele [12]. Teised kaks peamist lähenemist TTI valdkonnas on autoregressiivsel transformer (ingl *autoregressive Transformer*) arhitektuuril põhinevad mudelid ja difusioonimudelid (ingl *diffusion model*) [11]. Mõlemad võimaldavad väga realistlikke piltide genereerimist, kuid esimese puuduseks on, et sellised mudelid on tavaliselt väga suured ja see piirab kasutust lokaalsetel seadmetel [11]. Difusioonimudelid on üldiselt väiksemaid ja seega rohkem võimalusi lokaalselt kasutamiseks, samas võib nende treenimisprotsess olla ajamahukam [11].

Difusioonimudelite treenimine taandub kahele sammule: esiteks lisatakse pärisuunalise difusiooni sammus treeningandmetele Gaussi müra ja teine samm on selle müra samm-sammult eemaldamine (ingl *denoising*) [13]. Treenitud mudeliga saab müra eemaldamise protsessi juhtida tekstilise sisendi abil ja luua seeläbi uusi pilte [13]. Lisaks tavapärastele difusioonimudelitele, mis töötlevad pilte piksli-tasandil ja nõuavad seetõttu suurt arvutuslikku ressurssi, on loodud latentsed difusioonimudelid, mis töötlevad pilte eeltreenitud autokooderite abil latentses ruumis ja on seetõttu efektiivsemad [14]. Selles töös kasutatud Stable Diffusion on latentne difusioonimudel.

Lisaks tavalisele (positiivsele) sisendtekstile, mis kirjeldab, mida pilt kujutama peaks, saab Stable Diffusion mudelitele pildi genereerimisel kaasa anda ka negatiivse viiba (ingl *negative prompt*), mis laseb kasutajatel kirjeldada, mida loodud pilt kujutada ei tohiks [15]. Kui proovida kirjeldada positiivses viibas, et mingi element ei tohiks pildil kujutada On aga paslik

märkida, et negatiivse viiba lisamine on Stable Diffusion v2 mudelil märgatavalt suurema mõjuga, kui v1.5 puhul[9].

Tehisintellekti genereeritud pildid ei vasta aga alati täpselt sisendtekstile. TTI mudelitel on raskusi faktiliselt korrektsete piltide loomisega; vastavasisulises artiklis analüüsiti pildi hallutsinatsioonide (ingl *image hallucination*) esinemist Dalle-3 ja nelja Stable Diffusion mudeli näitel ja kuigi Dalle-3 mudelile anti üldiselt parem hinnang, esineb neid probleeme kõigi testitud mudelite piltidel [15]. Sellised pildi hallutsinatsioonid võivad muuhulgas tekkida näiteks molekulide või ajaloolise konteksti kujutamisel. Probleeme esineb ka elementide loendamisega: näiteks sisendtekstiga "viis õuna ja kümme sidrunit laua peal" on difusioonimudelite loodud pildil tihtipeale vale arv objekte [16]. On leitud ka, et kaashüponüümide esinemine sisendtekstis halvendab Stable Diffusioni pildiloome kvaliteeti [17]. Näiteks sisendtekstiga "sebra ja kaelkirjak" võivad loomad ühte sulanduda või ei kujutata mõlemat, samas "sebra ja külmkapp" või "kaelkirjak ja külmkapp" puhul on need kaks elementi genereeritud pildil suurema tõenäosusega selgelt eristatavad (vt Joonis 1).




Joonis 1. Read ülevalt alla: genereeritud pildid kaashüponüümidest "kaelkirjak ja sebra", soojakaardid (mis näitavad, millisel osal pildist valitud objekti on kujutatud) kahel esimesel pildil, soojakaardid sebra-külmkapp ja kaelkirjak-külmkapp paaridest genereeritud piltidel [17]

[9] https://stable-diffusion-art.com/how-to-use-negative-prompts/

# 3. Meetodid

Selles peatükis kirjeldatakse lahendusi, mida katsetati kasutaja pöördumise tuvastamiseks, kõne transkribeerimiseks nii eesti kui ka inglise keeles, masintõlkeks inglise keelest eesti keelde ning pildi genereerimiseks. Lisaks annab see peatükk ülevaate graafilise kasutajaliidese loomisest ja riistvarast, millele projekt tugines.

## 3.1 Töövoo ülevaade

Järgnevalt on loetletud tähtsaimad sammud rakenduse töövoos, et oleks paremini aru saada, millist rolli iga katsetatud tehnoloogia mängib.

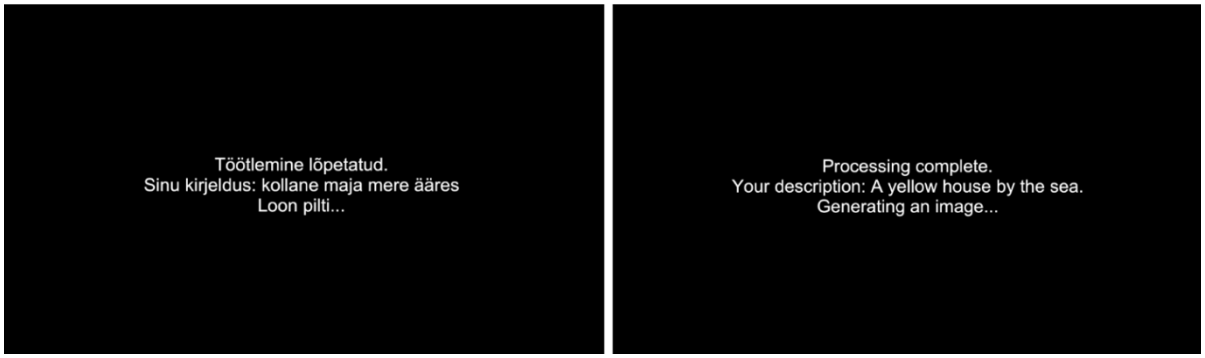
1. Ekraanil on kuvatud juhised rakenduse kasutamiseks (Joonis 2).


Joonis 2. Juhised ekraanil programmil käivitamisel.

2. Kasutaja alustab pöördumist, kasutajaliides annab märku, et alustati kuulamist. Algab heli salvestamine.
3. Kasutaja lõpetab pöördumise. Heli salvestamine lõpeb.
4. Tuvastatakse, kas kõneldi eesti või inglise keeles.
5. Kasutajale antakse märku, et kõne transkribeeritakse, ekraanil kuvatava teksti keele määrab tuvastatud keel (Joonis 3).
    a. Kõne transkribeeritakse, milline mudel kõnetuvastust teeb on määratud tuvastatud keele poolt.
    b. Transkriptsioonist eemaldatakse taustamüra märgendid.
    c. Kui kõne oli eestikeelne, tõlgitakse transkriptsioon inglise keelde.

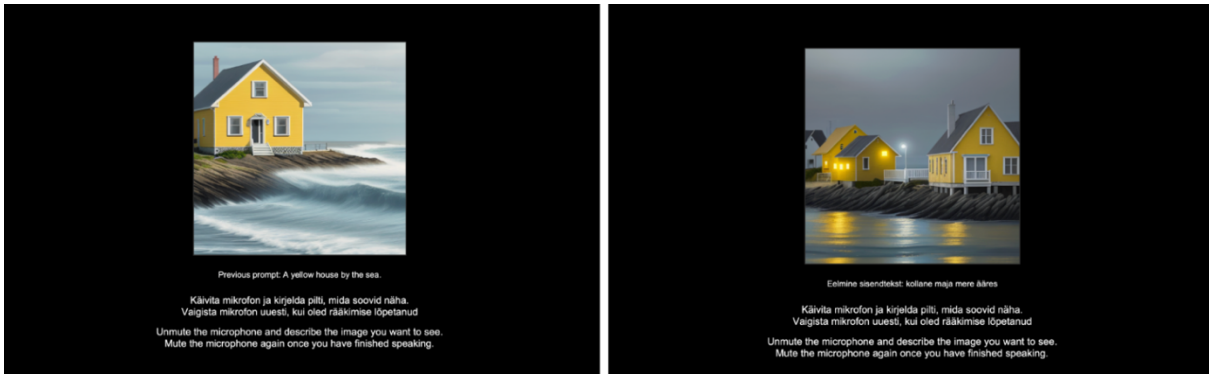

Joonis 3. Transkribeerimise teade tuvastatud keeles.

6. Kasutajale kuvatakse tuvastatud tekst originaalkeeles (Joonis 4).
    a. Alustatakse pildi genereerimisega, võttes tekstiviibana sisendiks ingliskeelse transkriptsiooni või tõlke.




Joonis 4. Transkriptsiooni kuvamine.

7. Kasutajale kuvatakse genereeritud pilt, tuvastatud tekst ja juhised uue pildi loomiseks (Joonis 5).




Joonis 5. Genereeritud pildi kuvamine.

Järgnevalt antakse ülevaade riistvarast, mida kasutati rakenduse arendamiseks ja testimiseks.

## 3.2 Riistvara ja nõuded

Rakenduse loomiseks kasutati peamiselt Apple M1 kiibiga sülearvutit (16GB RAM). Programmi töö demonstreerimiseks kasutatav riistvara valik oli määratud Tartu Ülikooli pakutavate võimaluste poolt. Kasutati mikrofoni Jabra Speak 510. Arvutina oli kasutusel Ubuntu operatsioonisüsteemiga arvuti, millel oli AMD Ryzen 9 3950X CPU ning Nvidia GeForce RTX 3080 graafikakaart, millel on 10GB VRAM-i. Võrdlemise madala mäluga graafikakaart piiras oluliselt võimalikke lahendusi kõne transkribeerimiseks ja pildi genereerimiseks. Et näidata, mis oleks võimalik võimsama riistvaraga, on tulemuste peatükis võrdluseks antud ka teatud võimekamate mudelite väljundid. Neid mudeleid sai antud riistvaral käivitada ühe kaupa või neid tuli käivitada Google Colab keskkonnas.

## 3.3 Kasutaja pöördumise tuvastamine

Kasutaja pöördumise tuvastamine on programmi töövoos oluline etapp, see annab teada, millal tuleb alustada heli salvestamist, et seda hiljem transkribeerida. Samaväärselt oluline on mõista, millal kasutaja pöördumine süsteemi poole lõpeb. See võib mürarikkas keskkonnas osutuda keeruliseks. Katsetati kahte lähenemist:

- Esimeses lähenemises katsetati äratussõna tuvastamiseks Picovoice'i Porcupine teeki. Porcupine'i eelisteks olid see, et sõna tuvastamine töötab lokaalselt ning et see võimaldab luua kohandatud äratussõnu ja ei vaja seejuures mudeli eraldi treenimist. Kuigi Porcupine'i ei toeta eesti keelt, oli see siiski katsetamiseks sobiv, kuna toetatud keelte hulgas on inglise ja saksa keel – keeled, mille sõnavaras leidub laensõnu ka eesti keeles. Lisaks jääb ka võimalus luua äratussõna, millel ei ole küll eesti keeles tähendust, kuid on siiski eesti keele kõnelejal mugav öelda. Porcupine'i lisamine Pythoni programmi osutus väga lihtsaks, ühtegi tehnilist probleemi seejuures ei tekkinud. Katsetuste käigus kasutati nii vaikimisi Porcupine'iga kaasas olevaid äratussõnu („Alexa", „Jarvis", „Picovoice", „Terminator") kui ka üht kohandatud äratussõna: saksa keeles fraas „Joon ist da", mille kõlapilt meenutab eestikeelset sõna „joonista". Eesmärk oli leida sõna või fraas, mida oleks eestikeelsele kasutaja jaoks mugav öelda ning mida Porcupine usaldusväärselt tuvastaks. Vaikimisi äratussõnu „Alexa" ja „Jarvis" jõuti katsetada kolme kasutaja peal ning vaikses keskkonnas toimis see piisavalt töökindlalt. Samas kohandatud äratussõna proovimine näitas, et kuigi ühe kasutaja peal toimis tuvastamine stabiilselt, võib teise kasutaja puhul olla selle töökindlus varieeruv. Lisaks ilmnes, et mürarikkama keskkonna puhul võib mudel

vaikse taustajutu seest ekslikult tuvastada äratussõna, mida tegelikult ei öeldud. Teine probleem äratussõna kasutamisel oli see ei aita tuvastada, millal kasutaja on rääkimise lõpetanud. Esialgu lahendati kõne lõpu tuvastamine helienergia analüüsi kaudu – kui järjestikuste heliakende energiatasemed olid kahe sekundi jooksul alla teatud lävendi, loeti see vaikuseks. Samas ei ole seda lävendit võimalik üheselt määrata keskkonna jaoks, kus taustamüra tase võib eri hetkedel väga palju erineda.

- Teises lähenemises kasutati kasutaja pöördumise alguse ja lõpu tuvastamiseks mikrofoni vaigistuse nuppu. Suhtluse alustamiseks peab selles lähenemises kasutaja Jabra mikrofoni vaigistamise nupu välja lülitama ning suhtluse lõpetamiseks uuesti mikrofoni vaigistama. Selle lahendusega sai töökindlalt määrata vaigistuse maha võtmise kasutaja pöördumise alguseks ning mikrofoni vaigistamise pöördumise lõpuks. Seega lahendas selline lähenemine lindistamise lõpu määramise probleemi mürarikkas keskkonnas ning eemaldas valepositiivsete probleemi.

Äratussõna tuvastamisega seotud probleeme arvesse võttes implementeeriti töökindlam lähenemine, mis eeldab kasutajapoolseid nupulevajutusi oma soovide väljendamiseks. Kasutaja jaoks intuitiivsem lahendus oleks *push-to-talk* nupuga mikrofoni kasutuselevõtt, mille puhul toimuks heli lindistamine ajal, mil nupp on alla vajutatud. Kui kasutaja pöördumine tuvastatud ja lindistatud, on järgmiseks sammuks pöördumise transkribeerimine.

## 3.4 Kõne transkribeerimine

Kõne transkribeerimine tekstiks on oluline, et suulise sisendi alusel luua pildi genereerimiseks vajalik tekst. Algne plaan oli selleks kasutada OpenAI Whisper mudelit, kuna see toetab nii eesti kui ka inglise keelt.

Arendus algas Apple Macbook M1 Pro sülearvutil, mille puhul ilmnes kiirelt, et ka Whisperi väiksemaid mudeleid ei olnud mälupiirangute tõttu võimalik edukalt jooksutada. Selle probleemi lahendamiseks kasutati whisper.cpp[10] Pyhtoni teeki[11], mis võimaldab Whisperi mudeleid jooksutada ka piiratud mäluga seadmetel, seal hulgas Apple Silicon arhitektuuril.

Katsetused näitasid, et kuigi Whisperi väiksemad mudelid saavad ingliskeelse kõne transkribeerimisega edukalt hakkama, ei suuda need eestikeelset kõne usaldusväärselt tuvastada – näiteks transkribeeris base mudel lause „Kõrb päikeseloojangu ajal“ kui

[10] https://github.com/ggml-org/whisper.cpp
[11] https://github.com/aarnphm/whispercpp

„Kirppäiksel ojanku jaal“. Whisperi large mudel suutis enamasti ka eesti keelt aktsepteeritaval tasemel tekstiks teisendada, kuid väiksemaid vigu siiski esines. Näiteks võis „Vihmasadu mägedes impressionistlikus stiilis“ transkriptsioon olla „Vihma sadumägedes impressionistlikus stiilis“.

Samas selgus, et large mudelit ei ole võimalik demoversiooniks kasutada mälu- ja ajapiirangute tõttu. Demoversiooniks oli kasutada 10GB VRAM-iga GPU ja large mudel üksi vajab umbes 10GB VRAM-i, seega ei jätaks see pilte loovale generatiivsele mudelile mäluruumi. Isegi Whisperi medium ja small mudelitel jäi demoarvutil mälust puudu ja seda ka vaid mõnesekundiliste helifailide puhul. Whisper.cpp lahendab küll mäluprobleemid, aga töötab seejuures tunduvalt aeglasemalt ja ei sobi seetõttu reaalajas toimiva demonstratsiooni jaoks.

Lisaks ilmnes pikema katsetamise jooksul, et ilma keelt ette määramata ei pruugi Whisperi transkriptsioon olla vastavuses sellega, mis keeles helifailis kõneldi. Vaadates näiteks teksti „Impressionistlikus stiilis maal tuletornist päikeseloojangu taustal.", mille transkriptsiooniks andis Whisper large mudel „Impressionistic style painting of a fire tower in the background of a sunny lake.“, on selge, et probleem ei tule sellest, et eesti keelne tekst kõlas nagu inglise keel, vaid Whisper tõlgib vahesammuna saadud transkriptsiooni. Probleemi vältimiseks on soovitatud lahendus mudelile keel helifailiga kaasa anda[12], seda varianti kasutades enam soovimatut tõlkimist ei esinenud. Samas algsete katsetuste jaoks kasutatud whisper.cpp Pythoni teek sellise parameetri lisamist ei võimalda.

Kuna Whisperi väiksemad mudelid ei olnud eesti keele jaoks piisavad ning „large“ mudeli kasutamine oli tehniliselt piiratud, katsetati alternatiivina Hugging Face’i kaudu kättesaadavat TalTechNLP Estonian Espnet2 ASR mudelit. See töötas edukalt MacOS-i peal, kuid programmi demo jaoks kasutatud Linux arvuti peale tõstes tekkisid probleemid, mida piisavalt kiirelt lahendada ei õnnestunud, ja seega läks esimeseks demonstratsiooniks kasutajate ette testimiseks vaid inglise keelt tuvastav versioon.

## 3.5 Masintõlge

Stable Diffusion mudelid on treenitud valdavalt ingliskeelsete kirjelduste põhjal[13] ning need annavad seetõttu ingliskeelsel sisendtekstil parimaid tulemusi. Seega oli vaja töövoosse lisada ka masintõlke samm, et tõlkida eestikeelsed transkriptsioonid inglise keelde.

[12] https://community.openai.com/t/whisper-is-translating-my-audios-for-some-reason/86468
[13] https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5

Masintõlkeks kasutati MarianMT[14] raamistiku mudelit „Helsinki-NLP/opus-mt-et-en“. Selle mudeli eelisteks on lihtne integreerimine Python-rakendusse (kuna see on osa Hugging Face'i Transformers teegist) ning võimalus kasutada seda lokaalselt, ilma internetiühenduseta. Tõlkekvaliteedi hindamiseks ei viidud läbi struktureeritud teste, kuid subjektiivsel hinnangul olid tõlked piisavalt täpsed, et edastada lause tähendus inglise keeles ning toetada pildi genereerimist.

## 3.6 Pildi genereerimine

Pildiloome mudeli valikul oli loomulikuks valikuks populaarne vabavaraline Stable Diffusion, alternatiivid nagu DALL-E[15] ja Midjourney[16] välistati, kuna need on kättesaadavad vaid läbi API või kommertsteenusena.

Uuemad Stability AI Stable Diffusion mudelid nagu 3.5 variatsioonid ja SDXL on kasutatud riistvaral jooksutamiseks liiga mahukad. Väiksemast Stable Diffusion mudelitest on populaarseim v1.5, selle põhjal on kasutajaskond loonud väga paljusid peenhäälestatud mudeleid; v2.x mudelite kasutajaskond on võrdlemisi väike[17].

Stable Diffusioni baasmudelid on üldotstarbelised (ei ole treenitud looma kindla stiiliga pilte) ja algne eesmärk oligi, et saaks luua ükskõik millise stiiliga pilte. Kuid katsetades ilmnesid suured probleemid inimeste ja loomade kujutamisel, eriti fotorealistliku stiili puhul – nt üleliigsed, puuduvad või ebaloomulikult asetsevad jäsemed, moonutatud näod jms (vt Joonis 6). Samas olid tulemused paremad näiteks impressionistliku stiili, õli- või akvarellmaali puhul (vt Joonis 7). Selliste stiilide puhul ei mõju moonutused ka niivõrd ebaloomulikult, kuna võivad paista osana valitud stiilist.

[14] https://huggingface.co/docs/transformers/en/model_doc/marian
[15] https://openai.com/index/dall-e-3/
[16] https://www.midjourney.com/home
[17] https://medium.com/@promptingpixels/comparing-stable-diffusion-models-2c1dc9919ab7

Joonis 6. Stable Diffusion v1.5 genereeritud pildid sisendtekstiga „a kid and a puppy“.


Joonis 7. Stable Diffusion v1.5 genereeritud pildid stiili määratlusega. Sisendtekst on kirjas pildi kohal.

Seetõttu järeldati, et parimate võimalike tulemuste saamiseks on vaja mingil kujul piirata genereeritava pildi stiili. Selleks kaaluti kahte viisi: sisendtekstile stiilimääratluste lisamine ning konkreetse stiili peal peenhäälestatud mudelite kasutamine. Erinevatel peenhäälestatud mudelitel saadud tulemused ning lõplik mudeli valik on kirjeldatud Tulemuste peatükis.

## 3.7 Graafiline kasutajaliides

Graafilist kasutajaliidest (GUI) on vaja, et kuvada kasutajale juhised rakenduse kasutamiseks, tuvastatud kõne transkriptsioon ning genereeritud pilt (ning sellega koos juhised uue pildi loomiseks). Kuna GUI loomiseks oli oluline vaid teksti ja pildi kuvamise võimalus, lähtuti tehnoloogia valikul selle kasutamise lihtsusest. ChatGPT soovitusel valiti selleks Pyhtoni teek Tkinter.

# 4. Tulemused

Vaid inglise keelt toetav versioon rakendusest oli testimiseks üleval kuupäevadel 28.03.2025-31.03.2025 Tartu Ülikooli Delta õppehoone vaba lava alal (Joonis 8). Demoperioodil kasutati demorakendust kokku 202 korda. Arvuti kõrval oli täpsem kasutusjuhend, ning QR-kood, mille kaudu said kasutajad anda rakendusele tagasisidet. Selles peatükis kirjeldatakse demoversiooni ning eraldi testitud keeletuvastuse, eestikeelse kõnetuvastuse ja eesti-inglise tõlke töövoo tulemusi. Lisaks antakse ülevaade demoperioodil kogutud tagasisidest.

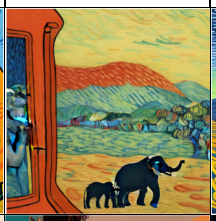
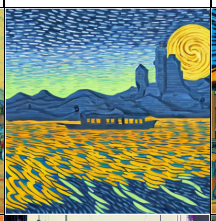
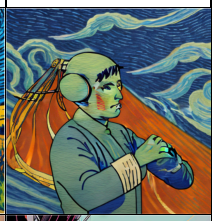
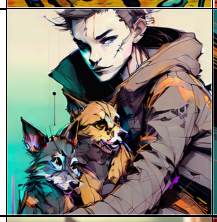
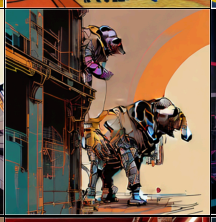
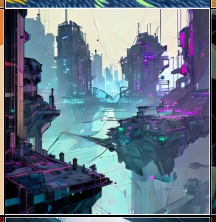
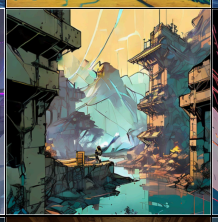
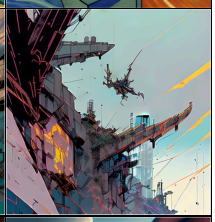
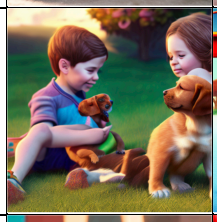
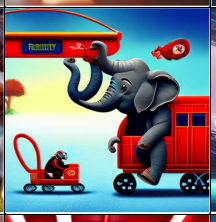
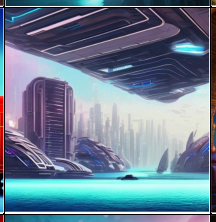
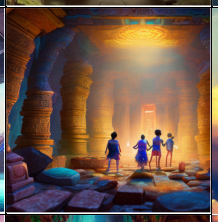
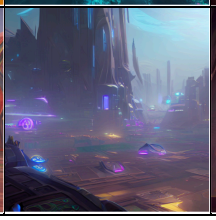
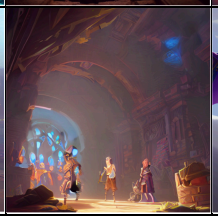

Joonis 8. Demo Delta vaba lava alal 28.03.2025.

## 4.1 Sobivaima pilte genereeriva mudeli valik

Tabelis on toodud mõned näited katsetatud Stable Diffusion v1.5 mudelil põhinevatest pildi genereerijate loodud piltidest (Tabel 1). Kaks esimest sisendteksti on autori loodud ning ülejäänud kolm on loodud ChatGPT o3 abiga. Mõne mudeli puhul peab sisendteksti algusesse või lõppu lisama kindla märgendi, et saada soovitud stiilis pilt (nt dallinmackay/Van-Gogh-

diffusion[18] puhul on vaja algusse lisada „lvngvncnt“), selguse huvides on tabelis toodud sisendtekstist see osa, mis oli kõikide mudelite puhul sama. Tajutud kvaliteedi alusel valiti projekti jaoks dreamlike-art/dreamlike-photoreal-2.0[19] mudel.

Tabel 1. Näiteid Stable Diffusion 1.5 mudelil põhinevate mudelite loodud piltidest.

| Sisendtekst / Mudel | „Three kids playing with two puppies” | „An elephant and a monkey riding a red train” | „A futuristic city floating above an ocean” | „A group of explorers discovering a hidden underground temple filled with treasure” | „A robotic dragon soaring through the clouds” |
|---|---|---|---|---|---|
| **dallinmackay/Van-Gogh-diffusion** |  |  |  |  |  |
| **Envvi/ Inkpunk-Diffusion[20]** |  |  |  |  |  |
| **lykon/ dreamshaper-8[21]** |  |  |  |  |  |
| **prompthero/ openjourney-v4[22]** |  |  |  |  |  |
| **dreamlike-art/dreamlike-photoreal-2.0** |  |  |  |  |  |
| **nitrosocke/ Arcane-Diffusion[23]** |  |  |  |  |  |

[18] https://huggingface.co/dallinmackay/Van-Gogh-diffusion
[19] https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0
[20] https://huggingface.co/Envvi/Inkpunk-Diffusion
[21] https://huggingface.co/Lykon/dreamshaper-8
[22] https://huggingface.co/prompthero/openjourney-v4
[23] https://huggingface.co/nitrosocke/Arcane-Diffusion

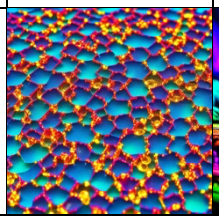
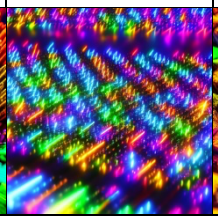
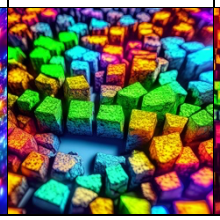
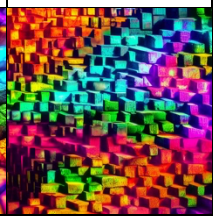
Selleks, et vältida ebasündsate piltide genereerimist, lisati negatiivne viip (ingl *negative prompt*) „nude, naked“. Mudeli näidispiltide[24] sisendtekstidele tuginedes lisati programmi pildi genereerimiseks kasutaja poolt antud sisendtekstile kvaliteetsemate tulemuste saamise eesmärgil järgnev: "highly detailed, cinematic lighting, vibrant colors, crisp edges".

## 4.2 Pildi genereerimise mooduli tulemused

Selles peatükis antakse ülevaade pildi genereerimise mooduli tulemustest, tuuakse näited demoperioodil õnnestunult genereeritud piltidest ja esinenud probleemidest. Ühe pildi genereerimiseks kulus demoarvutil umbes 10 sekundit.

Need korrad, kus ei olnud selget transkriptsiooni või see koosnes sõnadest, mis ei kirjeldanud mingit objekti, olid tulemuseks mitte midagi konkreetset kujutavad värvilised pildid (Tabel 2). Värvide rohkus tulenes tõenäoliselt sisendtekstile lisatud “vibrant colors” osast, lühikese või segase transkriptsiooni puhul on selliste täienduste mõju suurem.

Tabel 2. Näiteid demoperioodil genereeritud piltidest segaste transkriptsioonide korral.

| **Transkriptsioon** | 1 2 3 1 2 3 | Okay. | Nothing. | Come on. | as manyilns we mustfacing against the |
|---|---|---|---|---|---|
| **Genereeritud pilt** |  |  |  |  |  |

Selgete sisendtekstide puhul olid tulemused paremad, mõned näited on toodud Tabelis 3.

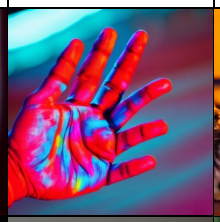
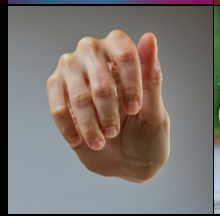
Tabel 3. Näited demoperioodil genereeritud piltidest, mis vastasid sisendtekstile.

| **Transkriptsioon** | 8 bit pixel art a nice cabin in the woods, smoke coming out of the chimney stone cabin small dirt path leading up to it pine woods | Elephant King is climbing on a mountain, probably rainbows. | A herd of cattle running on the streets of Manhattan. | Captain Jack Sparrow | Salmon in the river. |
|---|---|---|---|---|---|
| **Genereeritud pilt** |  |  |  |  |  |

[24] https://dreamlike.art/create

Samas on ka selliseid näited, kus genereeritud pilt ei vasta hästi sisendtekstile (vt Tabel 4). Siiski olid esinenud probleemid valdavalt ootuspärased: kaashüponüümide kujutamine (näiteks kassid ja koerad), elementide loendamine, täpse kellaaja kujutamine[25]. Sellised probleemid on Stable Diffusion 1.5 tasemega mudelite puhul tavapärased, kuid on uuemates, rohkem videomälu nõudvates mudelites mingil määral lahendatud. Stable Diffusion 3.5 Large saab paremini hakkama kolme kella ja inimkäe kujutamisega. Teised näidetena toodud sisendtekstid valmistavad raskusi ka suuremale mudelile, kuid sisendtekstiga „10 dogs riding 15 cats“ loodud pildil on erinevalt demoversioonile näha ka koera ja liikumist, seega vastab see pilt siiski paremini sisendtekstile.

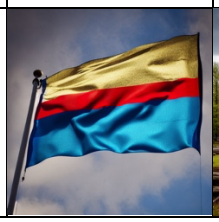
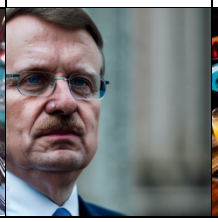
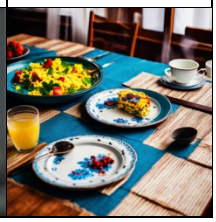
Tabel 4. Näiteid demoperioodil genereeritud piltidest, kus loodud pilt ei vastanud hästi sisendtekstile, võrdluseks on toodud Stable Diffusion 3.5 Large mudeli genereeritud pildid.

| **Transkriptsioon** | A clock showing 3pm. | clock without hands | 3 clocks all showing different times. | A human with a normal amount of fingers. | 10 dogs riding 15 cats. |
|---|---|---|---|---|---|
| **Demoperioodil genereeritud pilt** |  |  |  |  |  |
| **Stable Diffusion 3.5 Large loodud pilt** |  |  |  |  |  |

Probleemid pildi hallutsinatsioonidega tulid eriti selgelt välja näiteks Eesti lipu kujutamisel. Järgnevas tabelis on toodud näited pildiloome tulemustest Eestiga seonduvate transkriptsioonide korral (Tabel 5).

[25] https://generativeai.pub/in-the-ai-art-world-the-time-is-almost-always-10-10-cb38eed88aec

Tabel 5. Näited demoperioodil Eestiga seonduvate transkriptsioonide puhul genereeritud piltidest.

| **Transkriptsioon** | The Estonian flag. | a teenager holding an Estonian flag. | Estonian Christmas meal with sauerkraut, black sausages and potatoes. | Estonian President. | Traditional Estonian breakfast |
|---|---|---|---|---|---|
| **Genereeritud pilt** |  |  |  |  |  |

Kokkuvõttes töötas pildi genereerimise moodul ootuspäraselt. Esines teadaolevaid vigu nagu probleemid loendamisega ja kaashüponüümide kujutamisega ning pildi hallutsinatsioonid. Arusaadavalt olid piltide kvaliteet kehvem segaste transkriptsioonide korral. Selgete trankriptsioonide korral, mis ei palunud genereerida teadaolevalt mudelile raskusi pakkuvaid asju, olid tulemused piisavalt head.

## 4.3 Kõnetuvastuse mooduli tulemused

Demoperioodil kasutati kõnetuvastuseks Whisper base.en mudelit, kuna see oli parim, mis graafikakaardi videomällu mahtus. Tabelis 6 on välja toodud mõned näited demoperioodi kogutud helifailide transkriptsioonidest (Tabel 6). Esimeses tulbas on kirjas autori transkriptsioon salvestatud heliklipile, teises tulbas demo ajal tuvastatud tekst ehk transkriptsioon Whisper base.en poolt. Lisaks on antud transkriptsioon Whisper large-v3 mudeli poolt, et näidata, mis oleks olnud võimalik saavutada võimsama riistvara olemasolul. Baasmudeli puhul on suuremaks probleemiks vaikuse hetkel salvestatud heliklipid: mudel kipub taustamüra põhjal kõneldud teksti hallutsineerima. Tabeli kolmas tulp näitab, et Whisper large mudel saab sellise taustamüraga paremini hakkama ja suudab töökindlamalt eraldada helist kõne, mida on vaja transkribeerida. Homofoonia ehk sarnaselt kõlavad sõnad/fraasid on testitud näidete põhjal erineval määral probleemiks mõlemale (vt read 1 ja 2). Mõne vea põhjuseks võib olla ka vähene Eestiga seonduva treeningmaterjali hulk: kumbki mudel ei suutnud õigesti tuvastada Tartu linna nime, base mudel ei saanud aru fraasist „Estonian song festival“, kuigi inimkõrvale oli selles klipis kõne piisavalt selge.

Tabel 6. Valik demoperioodil valesti transkribeeritud pöördumisi. Võrdluseks Whisper large-v3 transkriptsioon samadel heliklippidel.

| **Kõneldud tekst** | **Whisper base.en transkriptsioon** | **Whisper large-v3 transkriptsioon** | **Märkused** |
|---|---|---|---|
| Salmon swimming in the river | Salomon swimming in the river. | Salmon in the river | |
| I am an Estonian 27 year old PhD student from Tartu. What do I look like? | I am an Estonian 27 year old PhD student from Dart to what do I look like? | I am an Estonian 27 year old PhD student from DART. What do I look like? | Homofoonia: Tartu ja dart to |
| Happy people | Happy Beeping Booth. | Happy people. | |
| Wedding cake | I'm going to squeeze out my wedding cake. | Wedding cake | Klipi alguses taustamüra |
| Tired people at school | I'm tired people at school. | Tired people at school | Klipi alguses taustamüra |
| Estonian song festival | Medd for two. | Estonian Song Festival | |
| | as manyilns we mustfacing against the | Thank you. | Klipp koosnes ainult mürast |

Eestikeelse kõne transkribeerimise jaoks parima mudeli valimiseks tõlgiti 20 demoperioodil loodud heliklippi eesti keelde ja loeti autori poolt sisse. Salvestatud heliklipid anti transkribeerimiseks Whisperi large-v3 ning TalTechNLP loodud espnet2_estonian, whisper-medium-et ja whisper-large-et mudelitele. Mõned näited saadud tulemustest on toodud järgnevas tabelis (Tabel 7). On näha, et Whisperi suurima mudeli täpsus jääb alla eesti keele peal peenhäälestatud mudelitele. TalTechNLP espnet2_estonian transkribeeris õigesti 18, whisper-medium-et 15, whisper-large-et 16 ja Whisperi large-v3 vaid 10 helifaili 20st. Mõlemad vead, mida espnet2_estonian testklippidel transkribeerimisel tegi olid seotud võõrnimedega: “Van Gogh” - “vanoo” ja “Caesari salat” - “see särisevad”. Saadud transkriptsioonide täpsuste põhjal valiti projekti jaoks TalTechNLP espnet2_estonian mudel.

Tabel 7. Valik eesti keeles sisseloetud lausetest ja nende transkriptsioonist erinevate kõnetuvastuse mudelitega. Eksimustega transkriptsioonid on välja toodud punase taustavärviga.

| **Kõneldud tekst** | **TalTechNLP/espnet2_estonian transkriptsioon** | **TalTechNLP/whisper-medium-et transkriptsioon** | **TalTechNLP/whisper-large-et transkriptsioon** | **Whisper large-v3 transkriptsioon** |
|---|---|---|---|---|
| Õnnelikud inimesed | õnnelikud inimesed | õnnelikud inimesed . | Õnnelikud inimesed . | Õnnedikud inimesed! |
| Lõhe jões ujumas | lõhe jões ujumas | lõhe jões ujumas . | lõhe jões ujumas . | Lõhe ju eesujumas! |
| Pulmatort | pulmatort | pulmatort . | ulmatort . | Pulmatort! |

| | | | | |
|---|---|---|---|---|
| Väsinu inimesed koolis | väsinud inimesed koolis | täsinud inimesed koolis . | väsinud inimesed koolis . | Päsinud inimest koolis. |
| Eesti laulupidu | Eesti laulupidu | Eesti laulupidu . | Eesti laulupidu . | Eesti laulupidu. |
| Ma olen kahekümne seitsme aastane Eesti doktorant Tartust. Milline ma välja näen? | Ma olen kahekümne seitsme aastane Eesti doktorant Tartust milline ma välja näen | ma olen kahekümne seitsme aastane Eesti doktorant Tartust , milline ma välja näen ? | ma olen kahekümne seitsme aastane Eesti doktorant Tartust . milline ma välja näen ? | Ma olen 27-aastane Eesti doktorin Tartust. Milline ma välja näen? |
| Suvaline spordiüritus Eestis | suvaline spordiüritus Eestis | suvaline spordiüritus Eestis . | suvaline spordiüritus Eestis . | Suvaline spordiüritus Eestis. |
| Caesari salat lõhe ja krutoonidega | see särisevad lõhe ja krutoonidega | see sõrmi salat lõhe ja krutoonidega . | see sari salat lõhe ja krutoonidega . | See sõri salat lõhe ja krutoonidega. |
| Pelikan sõidab jalgrattaga | pelikan sõidab jalgrattaga | Pelikan sõidab jalgrattaga . | Pelikan sõidab jalgrattaga . | Pelik sõidab jalgrattaga. |
| Me ronime Mount Everesti tippu | Me ronime Mount Everesti tippu | me ronime Mount Everest'i tippu . | me ronime Mount Everesti tippu . | Me roonime mounteväresti tippu. |

Eestikeelsete päringute tõlkimiseks inglise keelde prooviti esmalt Helsinki-NLP/opus-mt-et-en mudelit ning kuna seda oli lihtne kasutada ja see andis piisavalt häid tulemusi nii tõlke kvaliteedi kui ka kiiruse mõttes, otsustati, et ei ole vajadust proovida teisi mudeleid. Tabelis 8 on toodud näited tõlgetest Tabelis 7 näidatud transkriptsioonidele. Võib öelda, et kvaliteetse transkriptsiooni puhul on ka tõlge kvaliteetne. Mõne väiksema vea puhul transkriptsioonis võib tõlkimine muuta teksti isegi paremaks (vt Tabel 8 rida 2 „Õnnedikud“ – „Lucky“). Kuid üldjoontes võib öelda, et kvaliteetse tõlke jaoks on vaja kvaliteetset transkriptsiooni.

Tabel 8. Valik eestikeelsetest transkriptsioonidest ja Helsinki-NLP/opus-mt-et-en loodud ingliskeelsetest tõlgetes, valesti transkribeeritud sõnad on märgitud rasvases kirjas.

| **Transkriptsioon** | **Tõlge** |
|---|---|
| õnnelikud inimesed | Happy People |
| **Õnnedikud** inimesed! | Lucky people! |
| lõhe jões ujumas | salmon swimming in the river |
| Lõhe **ju eesujumas**! | It's a gap in the front. |
| väsinud inimesed koolis | tired people at school |
| **täsinud** inimesed koolis | perfected people at school |
| **Päsinud** inimest koolis. | A person in school. |
| Eesti laulupidu. | The Estonian Song Festival. |

| | |
|---|---|
| Ma olen kahekümne seitsme aastane Eesti doktorant Tartust milline ma välja näen | I am a twenty-seven-year-old Estonian doctoral student from Tartu what I look like |
| Ma olen 27-aastane Eesti **doktorin** Tartust. Milline ma välja näen? | I'm a 27-year-old Estonian doctorate from Tartu. What I look like? |
| suvaline spordiüritus Eestis | any sporting event in Estonia |
| Pelikan sõidab jalgrattaga | Pelican rides a bicycle |
| Me ronime Mount Everesti tippu | We're climbing Mount Everest to the top. |
| Me **roonime mounteväresti** tippu. | We're gonna climb the top of the mountain block. |

Keeletuvastuse mudeleid testiti 40 heliklipi peal, millest pooled sisaldasid eestikeelset ja pooled ingliskeelset kõne. Whisperi mudelid klassifitseerivad klippe 100 klassi (keelde), SpeechBrain lang-id-commonlanguage_ecapa ja lang-id-voxlingua107-ecapa vastavalt 45 ja 107 klassi. Seega tekkis ka olukordi, mil mudeli poolt valitud keel ei olnud ei eesti ega inglise keel. Samas on võimalik saada kätte tõenäosused iga klassi kohta ja võrreldes vaid tõenäosusi, kas heliklipp on eesti või inglise keeles ja klassifitseerida selle põhjal. Nii saab eemaldada olukorrad, kus tuvastatud keeleks on mõni kolmas keel, millega rakendus toime ei tule, ja tõsta keeletuvastuse täpsust. Tabelis on välja toodud, mitmel protsendil juhtudest tuvastas mudel õige keele ning mitmel protsendil juhtudest oli vaid eesti ja inglise keele tõenäosusi võrreldes õige keele tõenäosus suurem (Tabel 9). Kasutades kõiki klasse oli keeletuvastuse täpsus suurim Whisper large-v3 mudelil, kuid kitsendades võrdluse vaid eesti ja inglise keeltele, oli täpsus veidi parem SpeechBrain lang-id-voxlingua107-ecapa mudelil (85%). Seega valiti rakenduse jaoks viimane.

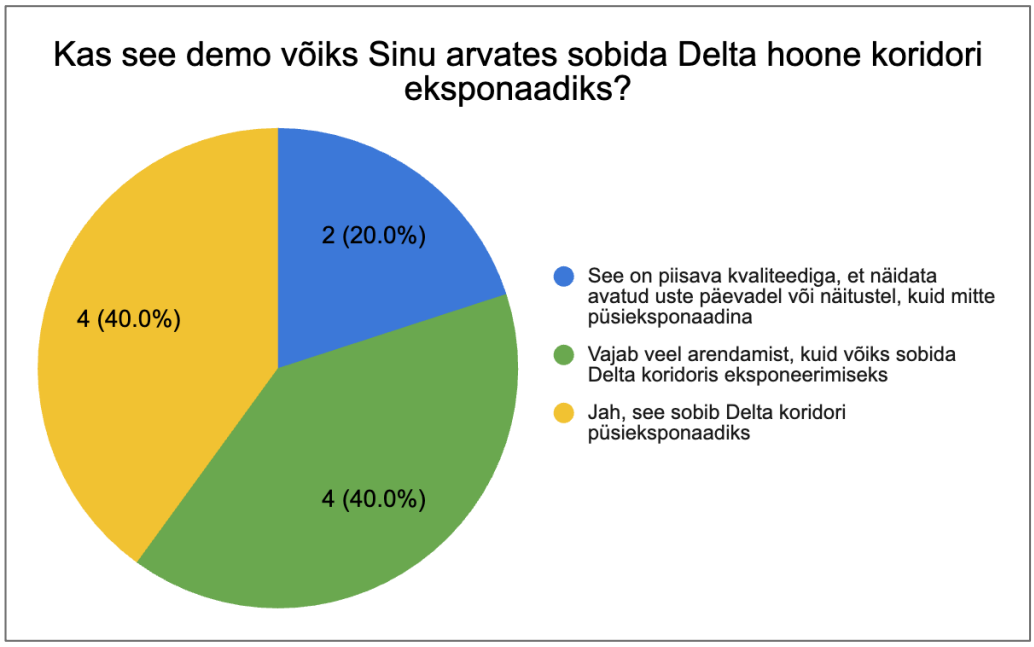
Tabel 9. Erinevate mudelite keeletuvastuse täpsused testides eesti- ja ingliskeelsetel heliklippidel.

| | Whisper large-v3 | SpeechBrain lang-id-commonlanguage_ecapa | SpeechBrain lang-id-voxlingua107-ecapa |
|---|---|---|---|
| Täpsus | 75% | 47,5% | 60% |
| Täpsus võrreldes eesti ja inglise keele tõenäosusi | 82,5% | 80% | 85% |

Demoperioodil oli kõnetuvastuse mooduli jaoks üks suurimaid probleeme müra esinemise heliklipi alguses; rohkem videomälu nõudev Whisper large-v3 saaks sellise müraga aga tunduvalt paremini hakkama. Eestikeelse kõne transkribeerimiseks valitud TalTechNLP/espnet2_estonian mudel sobib hästi demo kasutusjuhtudele omaste lühikeste heliklippide transkribeerimiseks.

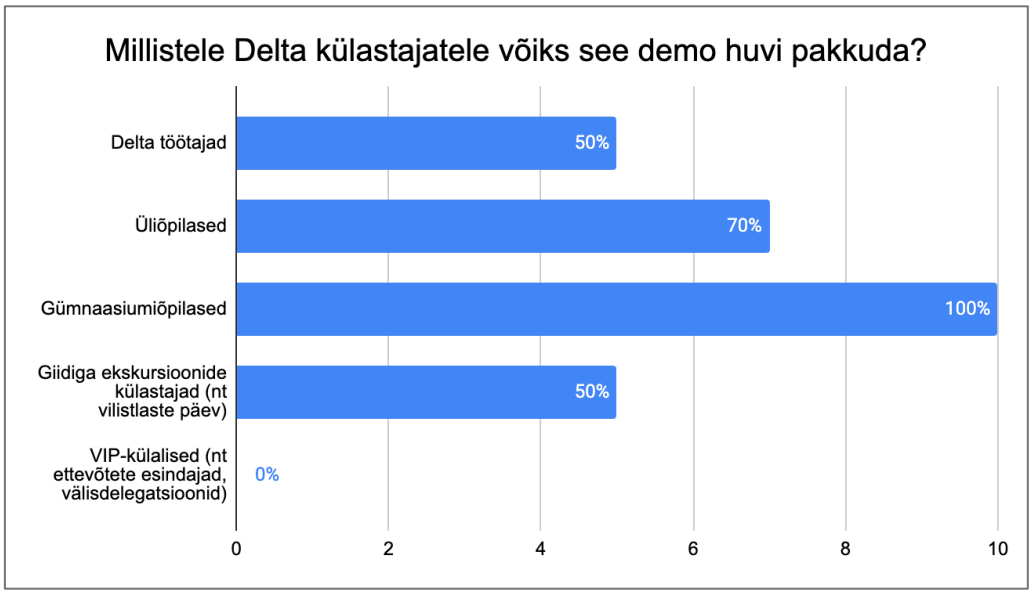
## 4.4 Demoperioodil kogutud tagasiside

Demoperioodi jooksul vastas tagasiside küsimustikule 10 inimest. Küsimustik on Lisa 1. Valminud programmi lähtekood. Esimene küsimus küsis kasutajatelt, kas demo võiks nende arvates sobida Delta õppehoone koridori eksponaadiks (vt Joonis 9). Kaks vastajat kümnest leidsid, et demoprojekt oli piisava kvaliteediga, et näidata seda avatud uste päevadel või näitustel, kuid see ei sobiks püsieksponaadiks. Ülejäänud kaheksa vastust jagunesid kaheks: pooled arvasid, et seda võiks eksponeerida Delta koridoris nagu olemasolevad demod, ning pooled arvasid, et see vajaks veel arendamist, kuid võiks siiski sobida püsieksponaadiks. Varianti, et projekt ei ole piisavalt hea või meelelahutuslik Delta hoones eksponeerimiseks, ei valinud ükski vastaja.



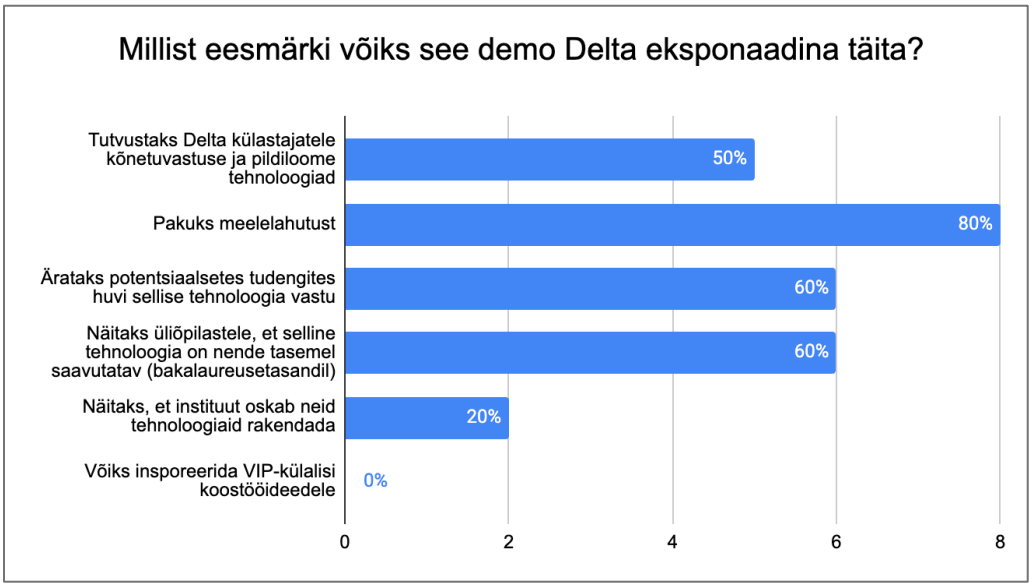

Joonis 9. Hinnangud demo sobivusele Delta koridoris eksponeerimiseks

Kasutajatelt küsiti ka, millistele Delta külastajatele see demo huvi võiks pakkuda (vt Joonis 10). Kõik vastajad nõustusid, et see oleks huvitav gümnaasiumiõpilaste jaoks. 70% vastajatest leidsid, et demo võiks olla huvi pakkuda ka üliõpilastele. Variante, et sihtrühmaks sobiksid ka Delta töötajad või giidiga ekskursioonide külastajad, valiti mõlemat viiel korral. Ükski vastaja ei arvanud, et demo võiks olla huvitav VIP-külalistele.



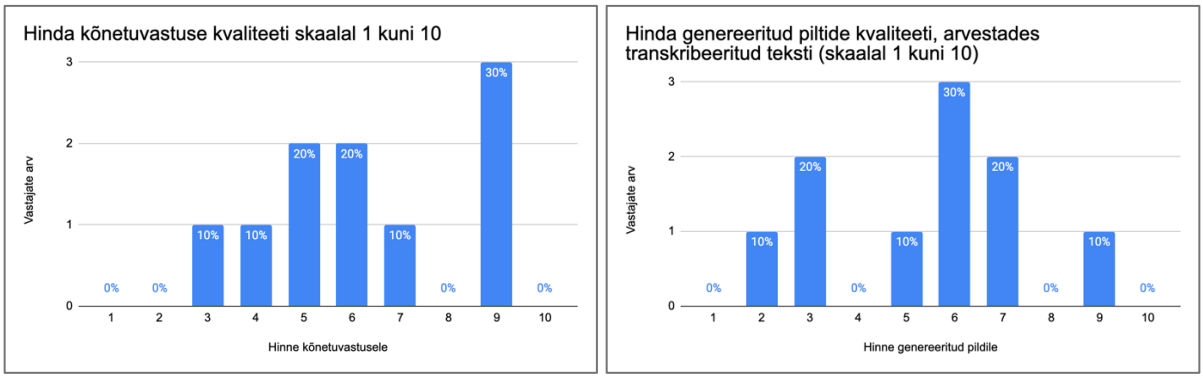

Joonis 10. Vastused küsimusele "Millistele Delta külastajatele võiks see demo huvi pakkuda?"

Vastajatel paluti hinnata, milliseid eesmärke selline demo täita võiks Delta püsieksponaadina (vt Joonis 11). Kaheksa vastajat arvasid, et see pakuks meelelahutust. Vastusevariante, et demo ärataks potentsiaalsetes tudengites huvi kasutatud tehnoloogiate vastu või ärataks huvi praegustes tudengites näidates, et selline projekt on saavutatav bakalaureuse tasemel, valiti mõlemat kuuel korral. Pooled vastajatest leidsid, et see aitaks tutvustada kõnetuvastuse ja piltide genereerimise tehnoloogiaid Delta külastajatele. Vastajate arvates ei aitaks see aga VIP-külastajaid koostööle inspireerida.




Joonis 11. Vastused küsimusele, millist eesmärki võiks demo täita.

Kasutajatel paluti hinnata kõnetuvastuse kvaliteeti skaalal ühest kümneni, kus 1 tähistas täiesti ebatäpset või segast ja 10 perfektset transkriptsiooni. Keskmiseks hindeks kõnetuvastusele kujunes 6,3 ning standardhälbeks oli 2,16. Samuti hindasid vastajad genereeritud piltide kvaliteeti arvestades transkribeeritud teksti skaalal ühest kümneni, kus üks tähendas, et genereeritud pilt oli halva kvaliteediga või kujutatu ei vastanud absoluutselt sisendtekstile, ning 10 tähendas, et pilt oli hea kvaliteediga ja vastasid täpselt transkribeeritud tekstile. Keskmine hinne genereeritud piltidele oli 5,4 standardhälbega 2.17. Joonis 12 näitab hinnangute jaotumist täpsemalt.




Joonis 12. Hinnangud kõnetuvastuse ja genereeritud piltide kvaliteedile.

Viimased kaks küsimust olid avatud vastusega küsimused ning neile vastamine polnud vastuse salvestamiseks kohustuslik. Tagasisidena prinditud juhistele soovitasid kaks vastajat need lühemaks muuta. Positiivsena toodi välja graafilise kasutajaliidese lihtsus. Edasiarenduste soovitused olid järgmised: QR-koodi lisamine genereeritud pildi alla

laadimiseks, kõnetuvastuse muutmine interaktiivsemaks kuvades tuvastatud teksti jooksvalt rääkimise ajal, kõnetuvastuse vigade parandamise võimaluse lisamine ning parema pildi genereerimise mudeli kasutamine, mis kuulab paremini sisendteksti ja suudab kujutada rohkemaid stiile.

# 5. Tulemuste arutelu

Selles peatükis antakse üldine ülevaade saadud tulemustest ning kirjeldatakse valminud programmi puuduseid ja edasiarendamise võimalusi.

## 5.1 Tulemuste üldine kirjeldus

Demoperioodil oli kasutusel pigem kehv kõnetuvastuse mudel Whisperi base.en, mille puhul mõjutasid märgatavalt transkriptsiooni kvaliteeti pausid kõnes, mille ajal oli kosta taustamüra. Eraldi testiti eesti keele töövoogu (keeletuvastus, eestikeelse kõne transkribeerimine ja tõlkimine inglise keelde) demoperioodil kogutud kasutusjuhtude põhjal ja tulemus oli rahuldav.

Genereeritud piltide kvaliteet on kõikuv. Lihtsamate sisendtekstide puhul see vastab ootustele. Samas on teada probleemid elementide loendamisega ja kaashüponüümide kujutamisega. Pildi hallutsinatsioonid ehk faktiliselt ebakorrektsed kujutised tulid esile näiteks Eestiga seonduva kujutamisel. Esines probleeme ka selgete nägude kujutamisega, kui sisend küsib tegevuse kujutamist, kuid portreepiltide puhul olid näod enamasti selged ja loomulike proportsioonidega.

## 5.2 Puudused ja edasiarendamise võimalused

Suurimad probleemid saaks lahendada suurema mälumahuga. Ingliskeelse kõne transkribeerimine annaks tunduvalt paremaid tulemusi kasutades Whisperi large mudelit ning piltide kvaliteeti ja vastavust sisendtekstile saaks tõsta võttes kasutusele mõne mälunõudlikuma mudeli, näiteks Stable Diffusion 3.5.

Piltide genereerimisel on olulisel kohal ka sisendtekst. Saaks parandada sisendteksti täiendust ja negatiivset viipa. Kuna sisendteksti täiendusi on keeruline lisada nii, et need sobiks ükskõik, millise kasutaja sisendiga, võib olla parem sisendteksti täiendamisel kasutada mõne suure keelemudeli abi.

Kasutajate tagasiside põhjal võiks üheks edasiarenduseks olla genereeritud pildi juurde QR-koodi lisamine, mille kaudu saaks kasutaja loodud pildi alla laadida. Lisaks saaks kasutajakogemuse mugavamaks muuta muutes kõne transkribeerimise protsessi nii, et ekraanil näidataks jooksvalt transkribeerimise tulemust juba rääkimise ajal. Nii näeks kasutaja kohe, kui transkriptsioonis tekib viga ja saaks probleemset sõna korrata, et see siiski õigel kujul pildiloome mudeli sisendiks jõuaks.

Praeguse lahendusega peab kasutaja demorakendusega suhtlemiseks tegema kaks nupulevajutust: kõne alguse ja lõpu märkimiseks. Rakendusega suhtlemise muudaks kasutaja jaoks intuitiivsemaks *push-to-talk* nupuga mikrofoni kasutusele võtmine. Sellise mikrofoniga peaks kasutaja lihtsalt pildisoovi kirjeldamise ajal nuppu all hoidma.

# 6. Kokkuvõte

Valmis lokaalselt töötav programm, mis võimaldab kasutajal genereerida pilte nii eesti- kui ingliskeelse kõne põhjal. Programmis kasutati eesti- ja ingliskeelse kõne transkribeerimiseks vastavalt TalTechNLP loodud espnet2_estonian mudelit ja OpenAI Whisperit, keeletuvastuseks SpeechBraini lang-id-voxlingua107-ecapa mudelit, tõlkimiseks Helsinki-NLP opus-mt-et-en mudelit ning pildiloomeks valiti dreamlike-art/dreamlike-photoreal-2.0 mudel. Demoperioodil 28.03.2025-31.03.2025 testiti programmi vaid inglise keelt toetavat versiooni reaalsete kasutajatega. Ülejäänud kõnemoodulist ehk keeletuvastust, eestikeelse kõne transkribeerimist ja masintõlget eesti keelest inglise keelde testiti eraldi.

Demoperioodil kasutati ingliskeelse kõne transkribeerimiseks Whisper base.en mudelit, millega tegi tihti vigu, kui kõnes esines pause, mille ajal oli kosta vaid taustamüra. Taustamürast tulenevate hallutsinatsioonide probleemi lahendaks suurel määral Whisperi large mudeli kasutamine, kuid selleks on vaja kasutusele võtta suurema videomäluga arvuti. Piltide genereerimisel esines ootuspäraseid probleeme: objektide loendamine, kaashüponüümide kujutamine ja faktilisiselt ebakorrektsed kujutised, mis tuli demoperioodil eriti selgelt välja Eesti lipu kujutamisel. Siiski lihtsamate ja selgete sisendtekstidega sai pildiloome mudel hästi hakkama.

Tagasiside põhjal võiks selline demoprojekt üldiselt inimestele meeldida ja pakkuda meelelahutust. Kõik tagasiside küsitlusele vastanutest arvasid, et taoline rakendus sobiks Delta õppehoones eksponeerimiseks, kas püsivalt või vähemalt üritustel nagu avatud uste päev või näitustel. Peamisteks sihtrühmadeks peeti gümnaasiumiõpilasi ja üliõpilasi, vähesemal määral arvati, et see võiks huvi pakkuda ka Delta töötajatele ja õppehoone ekskursioonidel osalejatele.

# Viited


[1] S. Bengesi, H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu and T. Oladunni, "Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers," *IEEE Access,* vol. 12, p. 69812–69837, 2024.

[2] Y. Wang, "Wake word detection and its applications," Johns Hopkins University, Baltimore, Maryland, 2021.

[3] A. H. Michaely, X. Zhang, G. Simko, C. Parada and P. Aleksic, "Keyword spotting for Google assistant using contextual speech recognition," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Okinawa, 2017.

[4] S. Venkatraman, A. Overmars and M. Thong, "Smart Home Automation—Use Cases of a Secure and Integrated Voice-Control System," *Systems,* vol. 9, no. 4, p. 77, 2021.

[5] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine,* vol. 29, no. 6, p. 82–97, 2012.

[6] V. Delić, Z. Perić, M. Sečujski, N. Jakovljević, J. Nikolić, D. Mišković, N. Simić, S. Suzić and T. Delić, "Speech Technology Progress Based on New Machine Learning Paradigm," *Computational Intelligence and Neuroscience,* pp. Article ID 4368036, 19 pages, 2019.

[7] H. Kheddar, M. Hemis and Y. Himeur, "Automatic Speech Recognition Using Advanced Deep Learning Approaches: A Survey," *Information Fusion,* vol. 109, p. 102422, 2024.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," San Francisco, 2022.

[9] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu and R. Pang, "Conformer: Convolution-Augmented Transformer for Speech Recognition," Ithaca, NY, 2020.

[10] N. V. Vitale, L. Schettino and F. Cutugno, "Rich Speech Signal: Exploring and Exploiting End-to-End Automatic Speech Recognizers' Ability to Model Hesitation Phenomena," in *25th Annual Conference of the International Speech Communication Association (INTERSPEECH 2024)*, 2024.

[11] F. Bie, Y. Yang, Z. Zhou, A. Ghanem, M. Zhang, Z. Yao, X. Wu, C. Holmes, P. Golnari, D. A. Clifton, Y. He, D. Tao and S. L. Song, "RenAIssance: A Survey Into AI Text-to-Image Generation in the Era of Large Model," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 47, no. 3, p. 2212–2231, 2025.

[12] P. Dhariwal and A. Q. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in *Advances in Neural Information Processing Systems*, 2021.

[13] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui and M.-H. Yang, "Diffusion Models: A Comprehensive Survey of Methods and Applications," *ACM Computing Surveys ,* vol. 56, no. 4, pp. 1-39, 2023.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis With Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[15] Y. Lim, H. Choi and H. Shim, "Evaluating Image Hallucination in Text-to-Image Generation with Question-Answering," *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 39, no. 25, p. 26290–26298, 2025.

[16] W. Kang, K. Galim, H. Il Koo and N. I. Cho, "Counting Guidance for High Fidelity Text-to-Image Synthesis," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[17] R. Tang, L. Liu, A. Pandey, Z. Jiang, G. Yang, K. Kumar, P. Stenetorp, J. Lin and F. Ture, "What the DAAM: Interpreting Stable Diffusion Using Cross Attention," arXiv preprint arXiv:2210.04885, 2022.

# Lisad

## Lisa 1. Valminud programmi lähtekood.

Loodud programmi kood on nähtav GitHubi repositooriumis: https://github.com/Katariina-Parkja/speakDraw

## Lisa 2. Tagasiside küsimustik.

**Feedback on BSc thesis project "SpeakDraw"**

**Küsimus 1** (ühe valikuvõimalusega valikvastusega küsimus)

Do you think this demo would be an interesting exhibit in the hallways of Delta?

- Yes, it could be in the hallway like other existing demos (magic mirror, ID card demo)
- It needs more work, but it could be in the hallway like other existing demos
- It is sufficient quality to be displayed during open-doors events and exhibitions, but not as a permanent demo
- This demo is not good or entertaining enough

**Küsimus 2** (mitme valikuvõimalusega valikvastusega küsimus)

Which guests of Delta do you think this demo would interest or be entertaining for?

- Employees of Delta
- Students
- High-school students
- Guided tour guests during events such as Alumni day
- VIP guests such as business representatives, foreign guests

**Küsimus 3** (mitme valikuvõimalusega valikvastusega küsimus)

If this demo was placed permanently in Delta, what goals would it serve in your opinion?

- Educate guests of Delta about speech recognition and image generation technologies (is it new to them?)
- Make the space in Delta more entertaining
- Attract interest in this technology in potential students by looking interesting
- Attract interest in students by showing it is achievable at their competence level (at Bsc level)
- Demonstrate that our institute is capable of applying such technologies (is this demonstration needed?)
- Generate collaboration ideas in VIP guests by demonstrating the technology

**Küsimus 4**

From a scale of 1 to 10, rate the quality of the speech recognition

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| It detected absolute gibberish, I did not say those things | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | It understood me perfectly |

**Küsimus 5**

From a scale of 1 to 10, rate the quality of the images generated, given the detected text

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| The generated images were abnormal or did not include the content mentioned in transcription | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | The images were looking good and followed the prompt faithfully |

**Küsimus 6** (avatud vastusega küsimus)

Give feedback on the current UI and printed "how to use" guidelines. What could make it better?

**Küsimus 7** (avatud vastusega küsimus)

How could this demo be developed further? Suggest what could make it mode educative or more entertaining.

# Litsents

**Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks**

Mina, Katariina Parkja,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Tehisintellekti abiga kõne põhjal piltide genereerimine“, mille juhendaja on Ardi Tampuu, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;

2. annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, 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.

Katariina Parkja

**18.05.2025**