LAB University of Applied Sciences

Kirjanurkka

Kuvallisten kirjojen AI-äänillä ääninäytellyn ominaisuuden tuottaminen ja toteutus verkkosivustoon

LAB-ammattikorkeakoulu
Tradenomi (AMK), Tietojenkäsittely
Karjalainen Juuso

Tiivistelmä

Tekijä(t)	Julkaisun laji	Valmistumisaika
Karjalainen, Juuso	Opinnäytetyö, AMK	2025
	Sivumäärä	
	53	

Työn nimi

Kuvallisten kirjojen AI-äänillä ääninäytellyn ominaisuuden tuottaminen ja toteutus verkkosivustoon.

Tutkinto

Tradenomi (AMK), Tietojenkäsittely

Toimeksiantajan nimi, titteli ja organisaatio

Tiivistelmä

Tässä opinnäytetyössä kehitettiin verkkopohjainen sovellus, joka tuo tekoälypohjaisen ääninäyttelyn osaksi kuvallisten kirjojen lukukokemusta. Tavoitteena oli selvittää, kuinka uskottavasti ja inhimillisesti tekoäly voi tuottaa hahmokohtaisia puheääniä reaaliaikaisesti, ja miten nämä ratkaisut voidaan integroida saavutettavasti verkkoympäristöön.

Työssä toteutettiin Full Stack -verkkosovellus React- ja Node.js-teknologioilla. Äänien tuottamiseen hyödynnettiin viittä eri AI-ääniteknologiaa, joita vertailtiin käyttäjätestauksessa (N=23) laadullisin ja määrällisin menetelmin. Testiasetelma perustui ITU-T P.808-standardiin ja arviointikriteerit mittasivat muun muassa inhimillisyyttä, tunnetta ja hahmosopivuutta.

Tulosten perusteella RVC-Project erottui tämän verkkosovelluksen käyttötarkoitukseen laadukkaimpana ja skaalautuvimpana teknologiana, joka mahdollisti realistiset ja yksilölliset hahmoäänet kustannustehokkaasti. Tämä ratkaisu integroitiin sovellukseen, jossa käyttäjä voi klikkaamalla puhekuplia kuunnella tekoälyn tuottamia vuorosanoja tai valita erilaisia AI-kertojamalleja, kuten Elina ja Joonas.

Työ osoittaa, että tekoälypohjainen äänenmuunnos tarjoaa uskottavan vaihtoehdon perinteiselle ääninäyttelylle. RVC-Projectin kaltaiset teknologiat mahdollistavat uudenlaisen, immersiivisen tavan kokea tarinoita verkkoympäristössä.

Asiasanat

Tekoäly, äänenmuunnos, puhesynteesi, ääninäyttely, saavutettavuus, verkkosovellus, sarjakuva, käyttöliittymä, reaaliaikainen puhe, äänen autenttisuus, käyttäjätestaus, RVC-Project, kehittämistutkimus, koulutusdata, ääni-interaktio, lasten sisällöt, inhimillisyys, käyttöliittymän saavutettavuus, kertojaratkaisut, hahmoäänet, äänimallit, avoimen lähdekoodin teknologia, voice conversion, ONNX-integraatio, AI-synteesimallit, pedagoginen käyttö, tunteiden tunnistus, neuroverkot, ääniteknologia, digitaalinen lukukokemus

Abstract

Author(s)	Type of Publication	Published
Juuso Karjalainen	Thesis, UAS	2025
	Number of Pages	
	53	

Title of Publication

Implementing and developing an AI-Narrated Feature for Illustrated Books

Name of Degree

Bachelor´s degree program in business information technology

Name, title and organization of the client

Abstract

This thesis presents the development of a web-based application that integrates AI-powered voice acting into the reading experience of illustrated books. The objective was to examine how convincingly, and naturally artificial intelligence can generate character-specific speech in real time, and how such solutions can be implemented in an accessible way within a web environment.

The project implemented a Full Stack web application using React and Node.js technologies. Five different AI-based speech synthesis solutions were used to generate the voices and were compared through user testing (N=23) using both qualitative and quantitative methods. The evaluation was based on the ITU-T P.808 standard, with criteria focusing on human-likeness, emotional expression, and character fit.

According to the results, RVC-Project stood out as the most suitable and scalable solution for the application’s goals, providing realistic and individualized character voices in a cost-effective manner. This technology was integrated into the application so that users could click on speech bubbles to hear AI-generated dialogue or choose between different AI narrator models.

The study demonstrates that AI-based voice conversion provides a credible alternative to traditional voice acting. Technologies such as RVC-Project enable a new kind of immersive storytelling experience in digital environments.

Keywords

Artificial intelligence, voice conversion, speech synthesis, voice acting, accessibility, web application, comic book, user interface, real-time speech, voice authenticity, user testing, RVC-Project, development research, training data, voice interaction, children's content, human-likeness, UI accessibility, narrator models, character voices, voice models, open-source technology, ONNX integration, AI synthesis models, educational use, emotion recognition, neural networks, voice technology, digital reading experience

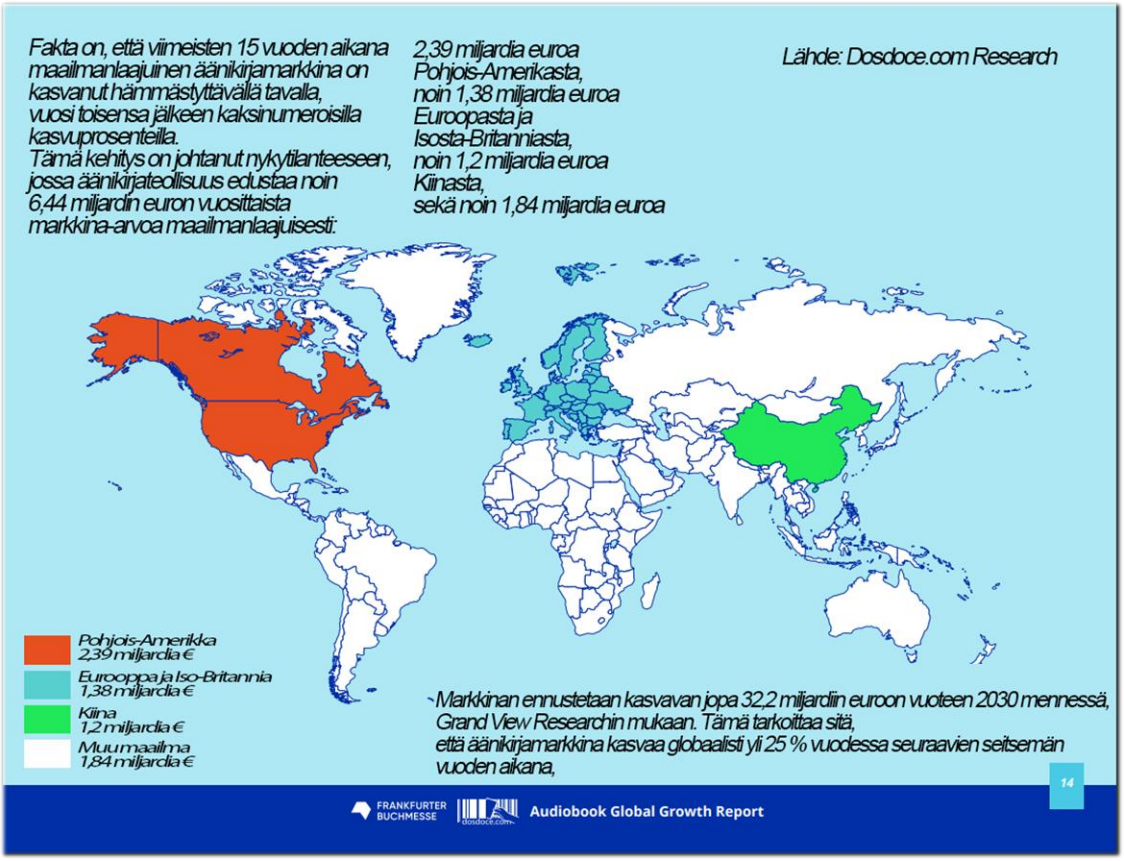
Sisällys

Liitteet

1 Johdanto

Kuvallisten kirjojen digitalisoituminen on osa laajempaa digitaalisen tarinankerronnan murrosta, jossa visuaalinen sisältö, interaktiivisuus ja multimodaalisuus yhdistyvät aiempaa syvemmin. Perinteiset kuvakirjat ja sarjakuvat ovat siirtyneet yhä enemmän verkkoympäristöihin, joissa lukijat odottavat immersioisia kokemuksia: eheää kokonaisuutta, jossa teksti, kuva ja ääni sulautuvat saumattomasti yhteen. Tekoäly (AI) tarjoaa tähän kehitykseen uudenlaisia mahdollisuuksia. Erityisesti äänenmuunnos- ja puhesynteesiteknologiat ovat kehittyneet tasolle, jolla ne voivat uskottavasti täydentää tai jopa korvata perinteisen ääninäyttelyn, mahdollistaen elämyksellisiä kerronnan muotoja myös visuaalisessa sisällössä.

Tätä kulttuurista ja teknologista muutosta havainnollistaa hyvin Frankfurt Book Fairin ja Dosdocen julkaisema kartta (Dosdoce, 2024, s. 14), jossa esitetään äänikirjamarkkinoiden alueellinen jakautuminen ja markkinan kokonaisarvo vuonna 2024. Kuten kuvasta yksi nähdään, Pohjois-Amerikka on edelleen suurin yksittäinen markkina-alue (2,39 miljardia euroa), mutta myös Euroopan ja muun maailman osuudet kasvavat vauhdilla kohti globaalia tavoitetta, jossa markkinan kokonaisarvon ennustetaan nousevan jopa 32,2 miljardiin euroon vuoteen 2030 mennessä.



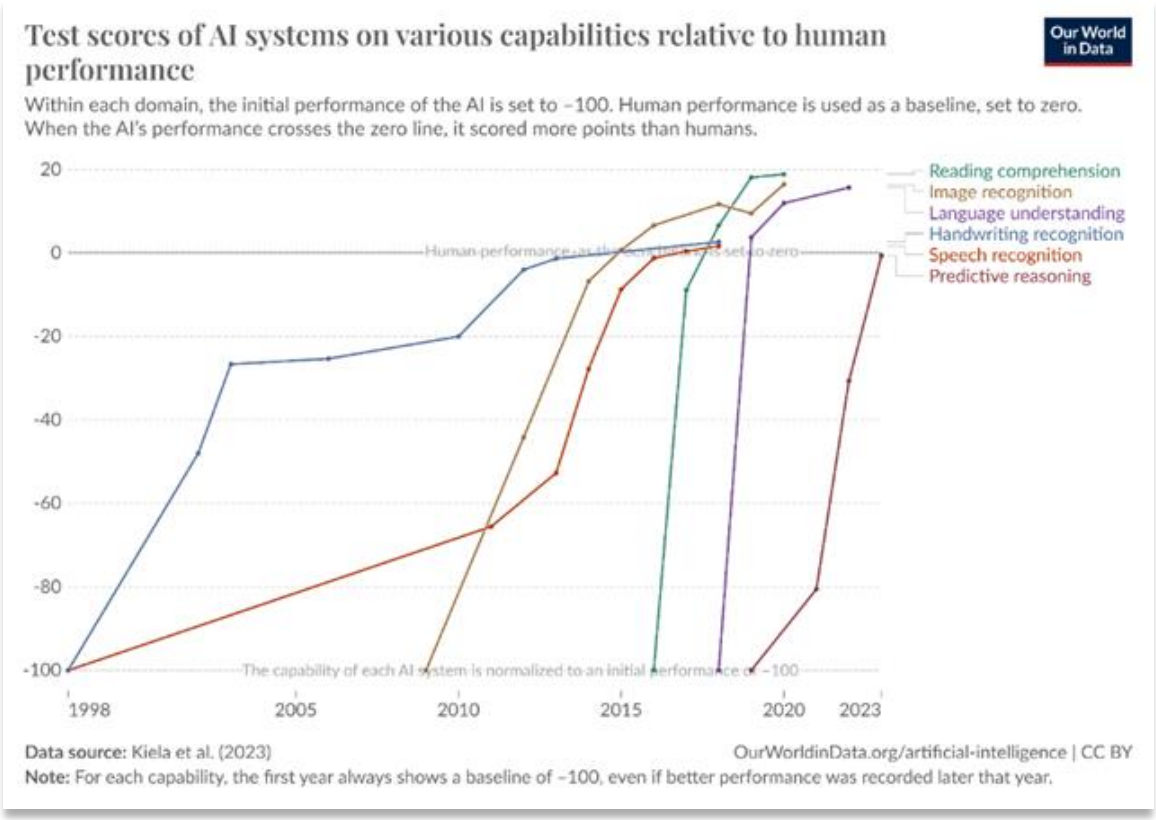
Kuva 1. Äänikirjamarkkinan alueellinen jakautuminen ja arvo vuonna 2024 euroina (Dosdoce & Frankfurter Buchmesse 2024)

Euroopassa kasvu on erityisen nopeaa: esimerkiksi Espanjassa äänikirjamyynti kasvoi 45,7 prosenttia vuoden 2023 aikana, ja vastaava kehitys on toistunut kolmena peräkkäisenä vuotena (Dosdoce & Frankfurter Buchmesse, 2024, s. 17). Kuuntelukulttuuri on samalla levinnyt uusiin yleisöihin, erityisesti lasten ja nuorten pariin. Esimerkiksi Saksassa lasten- ja nuorten äänikirjat muodostavat jo 44 prosenttia fiktiivisten äänikirjojen tuloista (Dosdoce & Frankfurter Buchmesse, 2024, s. 38). Lisäksi lasten omiin tarpeisiin suunnatut kuuntelulaitteet, kuten Tonies ja Yoto, ovat myyneet maailmanlaajuisesti yli seitsemän miljoonaa kappaletta (Dosdoce & Frankfurter Buchmesse, 2024, s. 61), tehden äänisisällöistä osan arkea myös niille käyttäjille, jotka eivät vielä osaa lukea.

Tämä antaa vahvan pohjan myös kuvakirjoille ja sarjakuville, joiden puhekuplat voidaan elävöittää tekoälyllä tuotetuilla realistisilla hahmoäänillä. Äänen yhdistäminen visuaaliseen sisältöön ei ole enää pelkkä tekninen mahdollisuus, vaan osa kulttuurista siirtymää, jossa tarinoiden saavutettavuus, elämyksellisyys ja kohdeyleisön monimuotoisuus laajenevat merkittävästi. Kuvakirjan ääni ei ole enää vain kerronnallinen lisä – se on osa lukukokemuksen ydintä.

Ääni lisää tarinoiden ilmeikkyyttä ja hahmojen uskottavuutta – ääninäyttely tuo vuorovaikutukseen emotionaalista syvyyttä. Perinteiset ääninäyttelymenetelmät ovat kuitenkin kalliita ja työläitä. Tekoälypohjainen reaaliaikainen ääninäyttely tarjoaa mahdollisuuden luoda skaalautuvia ja kustannustehokkaita ratkaisuja, jotka mahdollistavat uudenlaisen ääniavusteisen lukukokemuksen ilman kalliita studioäänityksiä.

Tekoälyn kehitys viimeisen vuosikymmenen aikana on ollut huomattavaa useilla keskeisillä osa-alueilla, kuten kuvantunnistuksessa, käsin kirjoitetun tekstin tunnistuksessa, kielen ymmärtämisessä ja puheentunnistuksessa. Kuviossa yksi havainnollistetaan, kuinka AI-järjestelmien suorituskyky on noussut asteittain ihmisen tasolle eri tehtävissä vuosien 1998–2023 välillä. Myös puheentunnistus on saavuttanut ihmisen suoritustason noin vuoden 2015 tienoilla. Tämä kehityssuunta tukee sitä, miksi tekoälypohjaisten äänten käyttäminen ääninäyttelyssä on nykyään sekä teknisesti mahdollista että laadullisesti perusteltua (Kiela, Bena & McCurdy 2023).



Kuvio 1. Tekoälyjärjestelmien suorituskyky suhteessa ihmiseen eri tehtävissä vuosina 1998–2023 (Our World in Data 2025)

Tämän opinnäytetyön päätavoitteena on kehittää tekoälypohjainen ääninäyttelyominaisuus kuvallisten kirjojen verkkosovellukseen, jossa käyttäjä voi lukea kirjoja ja samalla kuunnella hahmojen vuorosanoja realistisilla AI-äänillä. Tutkimuksessa arvioidaan, kuinka uskottavasti tekoäly voi tuottaa inhimillisen kuuloisia hahmoääniä reaaliaikaisesti osana digitaalista lukukokemusta. Lisäksi työssä vertaillaan erilaisia tekoälypohjaisia ääniteknologioita, jotta kyseiseen käyttöyhteyteen voidaan valita teknisesti ja laadullisesti paras ratkaisu.

Tätä varten tutkimuksessa asetettiin keskeinen tutkimuskysymys:

Miten tekoälypohjaista äänimuunnosta voidaan hyödyntää mahdollisimman inhimillisen ääninäyttelyn luomisessa kuvallisiin verkkokirjoihin?

Tutkimus jakautuu osaongelmiin, jotka käsittelevät:

- Millaiset tekoälypohjaiset puhesynteesiteknologiat soveltuvat parhaiten inhimillisen ääninäyttelyn toteuttamiseen kuvallisten verkkokirjojen kontekstissa?
- Miten eri teknologioiden tuottamaa ääntä voidaan arvioida laadullisesti käyttäjätestauksen avulla?

- Mitä teknisiä, saavutettavuuteen liittyviä ja eettisiä näkökulmia on huomioitava ääniavusteisen lukukokemuksen toteutuksessa?

Opinnäytetyön yhteydessä rakennetaan verkkosovellus Full Stack -ratkaisuna hyödyntäen React.js- ja Node.js-teknologioita. Frontend toteutetaan Reactin päälle rakentamalla, ja projektin kehityksessä hyödynnetään Viteä kehityspalvelintyökaluna. Tyylittely toteutetaan Tailwind CSS -kirjastolla. Palvelinpuolen toiminnallisuudet toteutetaan Node.js-ympäristössä, jossa hallitaan muun muassa ääniresurssien käsittely, käyttäjän valintojen ohjaus sekä mahdollinen rajapinta AI-prosesseille. Sovellus hyödyntää pilvipohjaista MongoDB Atlas -tietokantaa, jonka tarkoitus on toimia AWS-ympäristössä.

Kuva 2. Verkkosovellus toteutuksessa hyödynnetyt keskeiset teknologiat: Amazon Web Services (AWS), MongoDB Atlas, Tailwind CSS, Vite, React.js, Node.js ja Express.

Sovellustoteutuksen tarkoituksena on tukea tutkimusta ja toimia käytännön esimerkkinä siitä, millaisessa käyttökontekstissa tekoälypohjaista ääninäyttelyä voidaan hyödyntää kuvallisten kirjojen lukukokemuksen rikastamiseksi. Tutkimus toteutetaan juuri siitä syystä, että tämänkaltaiseen sovellukseen löydettäisiin teknisesti ja laadullisesti parhaiten soveltuva ääniteknologia.

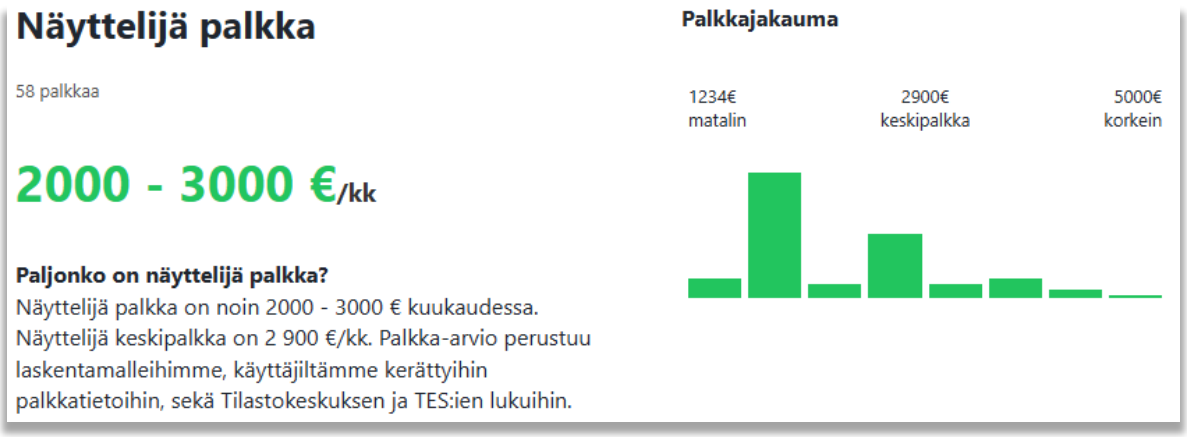
Tutkimuksessa tarkastellaan, voiko yhdestä ihmisäänestä tuotettuja ääniä käyttää uskottavasti erilaisten hahmojen esittämiseen. Esimerkiksi kuvakirjassa voi esiintyä kolme hahmoa: 35-vuotias käheä-ääninen nainen, kuusivuotias poikalapsi sekä 82-vuotias vanhempi mieshenkilö. Jotta valitsemamme AI-äänimuunnosjärjestelmä olisi käyttökelpoinen, sen on kyettävä tuottamaan jokaiselle hahmolle erottuva ja realistinen ääni siten, ettei kuulija huomaa kyseessä olevan saman pohjaäänen muunnos. Tämä edellyttää syväoppimisen hyödyntämistä tavalla, joka säilyttää puheen sisällön mutta muuntaa sen akustisia ja prosodisia ominaisuuksia merkittävästi.

Tässä opinnäytetyössä ääntä ei tuoteta reaaliaikaisesti käyttöliittymän kautta, vaan hahmojen repliikit muunnetaan synteettiseksi puheeksi erillisessä tuotantovaiheessa. Äänet generoidaan eri AI-äänimuunnosteknologioilla, joiden soveltuvuutta tutkitaan osana työn päätavoitetta. Tuloksena syntyneet .wav-tiedostot tallennetaan pilvitallennukseen, ja niitä toistetaan sovelluksessa esivalmistettuina resursseina. Tämä toteutustapa mahdollistaa

laskennallisesti kevyen ja teknisesti saavutettavan äänentoiston myös laitteilla, joilla reaaliaikainen synteesi ei olisi tarkoituksenmukaista.

Kustannustehokkuus nousee opinnäytetyössä tärkeäksi perusteluksi tutkimuksen toteuttamiselle ja aihevalinnalle. Esimerkiksi näyttelijän keskipalkan Suomessa on arvioitu olevan noin 2900 euroa kuukaudessa, ja kuukausiansioiden vaihteluväli sijoittuu noin 2000–3000 euron välille. Kyseessä on arvio, joka perustuu Palkkavertailu.com-sivuston tilastomalliin. Malli hyödyntää käyttäjiltä kerättyjä palkkatietoja, Tilastokeskuksen aineistoja sekä työehtosopimusten (TES) mukaisia tietoja (Palkkavertailu.com 2024).

Mikäli sovelluksessa tarvitaan esimerkiksi 20 erillistä hahmoääntä, perinteinen äänikirjamainen toteutus voisi tarkoittaa korkeita tuotantokustannuksia pelkän ääninäyttelyn osalta. Tekoälypohjainen äänenmuunnos mahdollistaa kaikkien hahmoäänten toteuttamisen yhdellä ääninäytteellä ja konversiomallilla, mikä tekee siitä sekä teknisesti että taloudellisesti merkittävän ratkaisun.



Kuvio 2. Näyttelijän arvioitu kuukausipalkka Suomessa (Palkkavertailu.com 2024)

Tutkimuksen tarkoituksena ei ole korvata ääninäyttelijöitä, vaan selvittää, millaisissa tilanteissa tekoälypohjainen äänenmuunnos voi toimia uskottavana ja saavutettavana vaihtoehtona. AI-ääniteknologioiden valinnassa ja AI-pohjaisten äänimallien koulutuksessa kiinnitetään erityistä huomiota saavutettavuuteen sekä monikielisten sisältöjen tukemiseen. Samalla tarkastellaan äänten autenttisuuden rajoja kuuntelutestauksen ja käyttäjäarvioiden analyysin avulla.

Lisäksi työssä käsitellään tekoälypohjaisten äänten käyttöön liittyviä eettisiä ja juridisia ulottuvuuksia. Vuonna 2024 voimaan tullut Euroopan unionin tekoälyasetus edellyttää, että tekoälyjärjestelmien tuottama ääni, joka muistuttaa todellista henkilöä, on merkittävä selkeästi keinotekoiseksi. Tämä tarkoittaa, että tekoälyn tuottamat äänisisällöt on esitettävä koneellisesti luettavassa muodossa ja niiden on oltava tunnistettavissa keinotekoisesti tuotetuiksi

tai manipuloiduiksi. Velvoite ei kuitenkaan koske tilanteita, joissa ääniä käytetään laillisesti rikosten havaitsemiseen, estämiseen, tutkimiseen tai syytteeseenpanoon. Myös selkeästi taiteellinen, satiirinen tai fiktiivinen sisältö on vapautettu täysimääräisestä merkitsemisvelvollisuudesta, kunhan keinotekoisen sisällön olemassaolosta ilmoitetaan sopivalla tavalla, joka ei häiritse teoksen esittämistä tai nautintoa (AI Act, artikla 50(4)). Opinnäytetyössä tarkastellaan, miten nämä vaatimukset voidaan huomioida käyttöliittymän suunnittelussa ja äänten merkinnöissä.

Työ perustuu soveltavaan kehittämistutkimukseen, jossa yhdistyvät tekninen toteutus, teoreettinen tarkastelu ja käyttäjätestaus. Tutkimuskysymys kuuluu: kuinka lähelle inhimillisen kuuloista ääntä tekoälypohjaisilla äänenmuunnosmenetelmillä voidaan päästä kuuntelukokemuksessa – ja millaisia mahdollisuuksia tämä avaa digitaalisen tarinankerronnan tulevaisuudelle?

2 Äänimuunnos, puheanalyysi ja AI-äänimuunnos

2.1 Puhesynteesin historia ja kehityskaari

Puhesynteesin kehityksen varhaisimmat vaiheet ulottuvat 1700-luvulle, jolloin unkarilainen Wolfgang von Kempelen rakensi mekaanisen puhekoneen. Tämä laite käytti palkeita, putkia ja mekaanisia kielisoittimia simuloimaan ihmisääntöväylää, ja se loi perustan puheen fysiologisen mallintamisen tutkimukselle (Lemmetty, 1999).

Vuonna 1922 John Q. Stewart julkaisi artikkelin Nature-lehdessä, jossa hän esitteli sähköisen analogian ihmisen ääntöväylälle. Hänen mukaansa puhesynteesin haasteet liittyivät ennen kaikkea ääntä tuottavan mekanismin säätelyyn eikä itse äänen tuottamiseen (Stewart, 1922). Tämä havainto on edelleen ajankohtainen: vaikka tekoälymallit kykenevät yhä paremmin tuottamaan ääntä, haasteet liittyvät sävyjen, intonaation ja luonnollisen ilmaisun hallintaan.

2010-luvulla tapahtui merkittävä teknologinen murros, kun sääntöpohjaiset ja yksikköpohjaiset synteesimallit korvattiin syväoppimiseen perustuvilla hermoverkoilla. Näistä merkittävin oli DeepMindin vuonna 2016 esittelemä WaveNet, joka tuotti puhetta suoraan aaltomuodossa. Tämä paransi äänen luonnollisuutta ja dynaamisuutta merkittävästi verrattuna aiempiin synteesimenetelmiin (van den Oord et al., 2016).

WaveNetin julkaisun jälkeen Googlen tutkijat kehittivät Tacotron 2 -mallin, joka yhdisti tekstistä puheeksi -synteesin spektripohjaiseen mallinnukseen. Siinä Mel-spektrit ennustettiin ja syötettiin edelleen WaveNetin kaltaiselle puheentuottomallille, mikä mahdollisti selkeämmän, inhimillisemmän ja ymmärrettävämmän puheen (Shen et al., 2018).

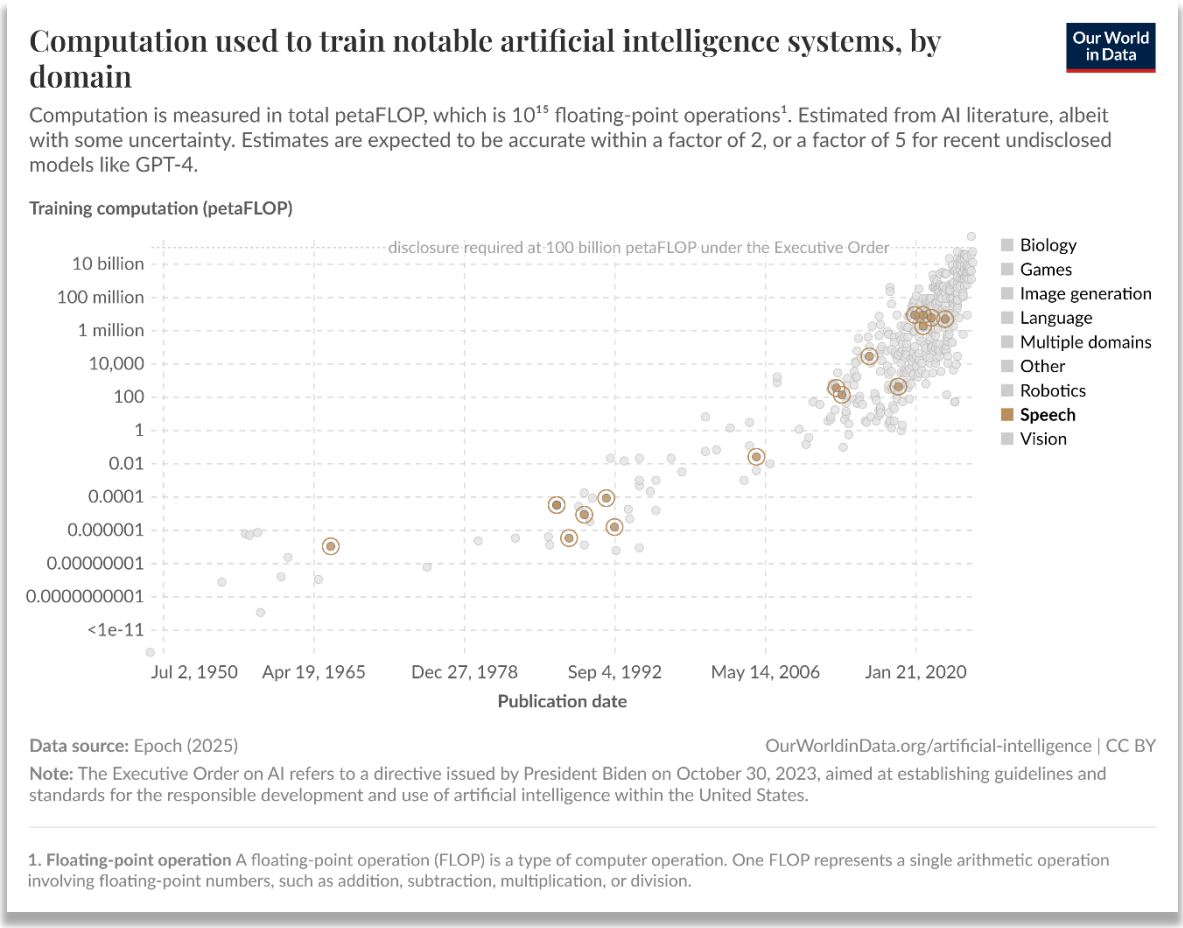
Vuonna 2024 Columbian yliopiston tutkijat julkaisivat tutkimuksen, jossa esiteltiin uusi prosodian ohjaukseen perustuva malli. Tutkimuksessa osoitettiin, että synteettinen puhe voidaan tuottaa niin luonnollisena, että sitä on vaikea erottaa ihmisen puheesta kuuntelutestissä (Li, Han, Raghavan, Mischler & Mesgarani, 2024). Tämän kehityskaaren ansiosta puhesynteesiä voidaan nykyisin hyödyntää realistisena ja kustannustehokkaana vaihtoehtona myös ääniavusteisessa lukemisessa ja ääninäyttelyssä.

Kuvio kaksi havainnollistaa, kuinka tekoälymallien kouluttamiseen käytetty laskentateho on kasvanut eksponentiaalisesti 2010-luvulta lähtien. Tietoa mitataan petaFLOPeissa, eli biljoonissa liukulukulaskuissa. Yksi petaFLOP (10^{15} FLOP) tarkoittaa biljoonaa liukulukulaskutoimitusta sekunnissa, ja se toimii standardina tekoälyn laskentasuorituskyvyn mittaamisessa (TechTarget 2023). Erityisesti syväoppimiseen perustuvat mallit, kuten

puhesynteesiin liittyvät järjestelmät, vaativat valtavasti koulutusdataa ja mallinparametrien iteratiivista säätöä, mikä tekee prosessista erittäin laskentaintensiivisen.

Kuviossa kaksi esiintyvän datan tuottaa alkujaan Epoch AI (Epoch 2025), heidän tutkimukseensa Parameter, Compute and Data Trends in Machine Learning, ja sen on käsitellyt ja visualisoinut Our World in Data (2025). Kuvio osoittaa, että myös puheeseen liittyvät mallit – kuten DeepSpeech2, Wave2Vec 2.0 ja Whisper – sijoittuvat korkealle koulutuksessa käytetyn laskentatehon suhteen muihin tekoälyjärjestelmiin verrattuna.

Tämä datakehitys selittää, miksi puhesynteesi on kehittynyt laadullisesti uudelle tasolle vasta viime vuosina: äänet eivät ole vain ymmärrettäviä, vaan niitä on usein vaikea erottaa ihmisen tuottamasta puheesta. Tämä murros on ollut mahdollinen vasta, kun syväoppimismallien arkkitehtuuri on kypsynyt ja laskentaresurssit – kuten rinnakkaisprosessorit ja suurtehotilastot – ovat nousseet vastaamaan tekoälymallien vaatimuksiin (Our World in Data 2025).



Kuvio 2. Puheeseen liittyvien tekoälymallien koulutukseen käytetty laskentateho vuosina 1950–2025 (Our World in Data 2025)

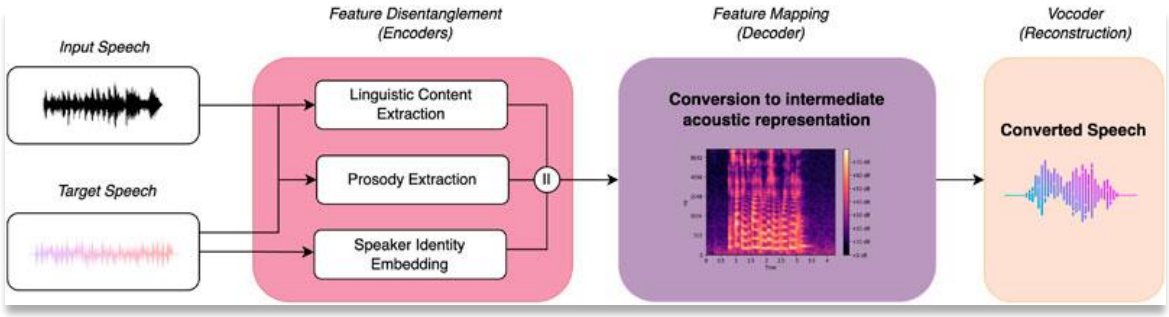
2.2 Äänimuunnosprosessi ja sen vaiheet

Äänimuunnos (engl. voice conversion, VC) on puheenteknologian osa-alue, jossa puheääntä muokataan säilyttäen alkuperäisen puheen kielellinen sisältö muuttumattomana. Muunnos voi kohdistua esimerkiksi äänen korkeuteen, sävyyn, sukupuolivaikutelmaan tai muihin tyylillisiin piirteisiin. Tämä mahdollistaa sen, että yksi henkilö voi tuottaa useita toisistaan eroavia hahmoääniä hyödyntäen yhtä äänilähdettä (Wang et al., 2023).

Äänimuunnosteknologioita voidaan hyödyntää monipuolisesti eri käyttöyhteyksissä, kuten viihteessä, kieltenopetuksessa, saavutettavuuden edistämisessä sekä yksityisyyden suojan tukemisessa. Teknologian keskeinen piirre on kyky erottaa puheen sisällölliset (lingvistiset) ominaisuudet äänen akustisista ominaisuuksista ja käsitellä niitä erillisesti muunnosprosessissa (Wang et al., 2023).

Äänimuunnosprosessi koostuu kolmesta päävaiheesta: analysointi, kartoitus ja rekonstruktio. Ensimmäisessä vaiheessa alkuperäinen puhesignaali puretaan kielelliseksi sisällöksi, prosodisiksi piirteiksi (kuten sävelkorkeus ja rytmi) sekä puhujan identiteettiin liittyviksi akustisiksi ominaisuuksiksi. Toisessa vaiheessa nämä ominaisuudet muunnetaan kohdeääntä vastaaviksi akustisiksi representaatioiksi, joko sääntöpohjaisesti tai neuroverkkopohjaisen mallin avulla. Kolmannessa vaiheessa muunnettu signaali rekonstruoidaan takaisin puhemuotoon vocoderin avulla. Lopputuloksena syntyy puheääni, joka kuulostaa siltä kuin sen olisi tuottanut kohdeääni (Bargum, Serafin & Erkut, 2024).

Kuviossa kolme esitetään visuaalisesti syväoppimispohjainen äänimuunnosprosessi. Siinä yhdistyvät puheen kielellinen sisältö, prosodia ja puhujan identiteetti, jotka erotellaan toisistaan ja muunnetaan keskitetyn akustisten representaatioiden kautta kohdepuheeksi. Tämä kolmivaiheinen rakenne – analysointi, kartoitus ja rekonstruktio – muodostaa nykyaikaisten voice conversion -järjestelmien perustan (Bargum, Serafin & Erkut, 2024).



Kuvio 3: Perinteinen syväoppimispohjainen äänimuunnosprosessi ja sen kolme päävaihetta (Bargum, Serafin & Erkut 2024)

2.3 Ääniteknologiat: ominaisuudet ja soveltuvuuden arviointi perusteet ennen käyttäjätestausta

Tekoälypohjaisten ääniteknologioiden kehitys on mahdollistanut puheäänen tuottamisen täysin ilman ihmisnäyttelijää. Kirjanurkka-projektin tavoitteena oli arvioida, miten eri teknologiat tukevat inhimillisen kuuloista ja emotionaalisesti uskottavaa ääntä, joka tuotetaan ennakkoon ja liitetään verkkosovellukseen hahmokohtaisesti esituotettuina äänitiedostoina. Tarkastelun painopisteinä olivat puheen luonnollisuus, hahmoäänten muokattavuus, käyttöoikeudet, saavutettavuus ja kustannustehokkuus – ominaisuudet, jotka aiemman tutkimuksen mukaan vaikuttavat keskeisesti puhesynteesin uskottavuuteen ja käyttäjäkokemuksen laatuun (ITU-T P.808, 2021; Li et al., 2024).

Ennen varsinaisen käyttäjätestauksen toteuttamista suoritettiin laaja-alainen teknologioiden esiarviointi. Arvioinnin tavoitteena oli tunnistaa ne teknologiat, joilla olisi realistinen mahdollisuus tuottaa laadukkaita ja inhimillisesti uskottavia hahmoääniä kuvakirjojen lukutilanteessa. Tarkastelussa yhdistyivät sekä teoreettiset että käytännön näkökulmat, jotka liittyivät äänen laatuun, tekniseen toteutettavuuteen ja käyttöehtoihin.

Teknologioita arvioitiin seuraavien kriteerien perusteella:

- Kielellinen luonnollisuus ja ymmärrettävyys: Arvioitiin, kuinka hyvin teknologia kykenee tuottamaan puhetta, joka kuulostaa rytmiltään, sävelkululltaan ja artikulaatioltaan aidolta ihmisen puheelta.
- Emotionaalinen ilmaisu: Selvitettiin, missä määrin teknologia kykenee välittämään tunnetiloja, kuten iloa, surua tai jännitystä, jotka ovat olennaisia fiktiivisessä kerronnassa.
- Äänen muokattavuus ja tekninen joustavuus: Tarkasteltiin, voiko käyttäjä kouluttaa täysin uusia ääni-identiteettejä tai muokata valmiita puheprofiileja hahmokohtaisesti.
- Käyttöoikeudet ja alustojen avoimuus: Arvioitiin, ovatko teknologiat opetuksen tai tutkimuksen yhteydessä käytettävissä ilman merkittäviä lisenssirajoituksia tai suljettua alustariippuvuutta.
- Kustannustehokkuus: Analysoitiin, soveltuvatko teknologiat opiskelijaprojektin käyttöön rajallisilla resursseilla, ilman jatkuvia maksullisia lisenssejä tai pilvipohjaisia käyttörajoitteita.

KRITEERI	KUVAUS
KIELELLINEN LUONNOLLISUUS	Rytmin, sävelkulun ja artikulaation autenttisuus
EMOTIONAALINEN ILMAISU	Mahdollisuus sävyttää puhetta tunnetilojen mukaan
MUOKATTAVUUS JA JOUSTAVUUS	Käyttäjän mahdollisuus kouluttaa tai säätää ääniä
LISENSSIVAPAUS JA AVOIMUUS	Soveltuvuus opetukseen/tutkimukseen ilman kaupallista sitoutumista
KUSTANNUSTEHOKKUUS	Teknologian käytön taloudellinen realistisuus opiskelijaprojektissa

Taulukko 1. Teknologioiden alustavat arviointikriteerit ennen käyttäjätestausta

Arvioinnissa havaittiin, että monet tunnetut puhesynteesiratkaisut – esimerkiksi tietyt API-pohjaiset kaupalliset työkalut – eivät soveltuneet käyttöön johtuen lisenssirajoituksista, suljetuista käyttöympäristöistä tai siitä, että tekninen dokumentaatio oli puutteellinen. Osa teknologioista hylättiin myös, koska ne vaativat huomattavia taloudellisia investointeja jo pelkästään kokeiluvaiheessa, mikä ei ollut realistista opiskelijaprojektin resurssien puitteissa.

Näiden rajausten jälkeen lopulliseen analyysiin ja käyttäjätestaukseen valittiin viisi teknologiaa: Google Text-to-Speech, Amazon Polly, ElevenLabs, Voice.ai ja RVC-Project. Nämä edustavat erilaisia teknisiä lähestymistapoja puhesynteesiin ja äänenmuunnokseen. Ne ovat keskenään riittävän erilaisia, jotta vertailu olisi sekä informatiivista että tarkoituksenmukaista Kirjanurkan käyttötapauksen kannalta.

Seuraavissa alaluvuissa tarkastellaan näiden viiden teknologian yksilöllisiä ominaisuuksia ja alustavaa soveltuvuutta ennen varsinaisten käyttäjätestien tulosten esittämistä.

2.3.1 Google Text-to-Speech

Google Text-to-Speech (TTS) on Googlen pilvipohjainen puhesynteesipalvelu, joka hyödyntää syväoppimiseen perustuvia neuronaalisia malleja, kuten Neural2-, Standard- ja WaveNet-äänimoottoreita. Näiden avulla tuotetun puheen luonnollisuus ja artikulaation laatu on parantunut huomattavasti aiempiin mallisukupolviin verrattuna. Palvelu tukee yli 300

ääntä ja yli 50 kieltä tai murretta, ja sitä käytetään yleisesti sovelluksissa, jotka vaativat nopeaa ja luotettavaa äänen generointia (Google Cloud 2024).

TTS toimii tekstisyötteen perusteella ohjelmointirajapinnan (API) kautta. Käyttäjä voi valita valmiista ääniprofiileista, mutta ei voi itse kouluttaa täysin uusia ääni-identiteettejä ilman erillistä yritysasiakkaille suunnattua palvelumallia. Google tarjoaa erillisen Custom Voice -ominaisuuden, jonka avulla voidaan kouluttaa yksilöllisiä ääniä korkealaatuisten tallenteiden perusteella. Tämä edellyttää kuitenkin Google Cloudin myyntitiimin hyväksyntää, erillistä sopimusprosessia sekä korkeaa teknistä laatua, eikä se ole vapaasti käytettävissä opiskelijaprojekteissa tai yksityiskäytössä (Google Cloud 2024).

Teknologian vahvuuksia ovat skaalautuvuus, kielituki, nopea vasteaika ja tekninen vakaus. Äänet ovat selkeitä ja soveltuvat erityisesti neutraaleihin ja informatiivisiin käyttötarkoituksiin, kuten ohjeistuksiin, käyttöliittymäpuheeseen tai opetusmateriaaleihin. Sen sijaan yksilöllisen, emotionaalisesti vaihtelevan tai hahmopohjaisen ääninäyttelyn toteuttaminen on rajoittunutta. Äänen emotionaalinen sävy ja tyylillinen muokattavuus jäävät usein pinnallisiksi, eikä käyttäjä voi säätää prosodiaa, kuten sävelkulkua tai tunnetilaa, ilman edellä mainittua mukautettua ääniratkaisua.

Google TTS on maksullinen palvelu, jonka hinnoittelu perustuu muun muassa käytettyjen merkkien määrään ja valittuun puhemalliin. Palvelu on erityisen soveltuva staattisten kertojaäänien tuottamiseen Kirjanurkan kaltaisessa verkkoympäristössä, mutta hahmoäänten yksilölliseen luomiseen se ei tarjoa tarvittavaa muokattavuutta ilman kaupallisen Custom Voice -prosessin hyödyntämistä.

Kuva 3. Google Text to-Speech

2.3.2 Amazon Polly

Amazon Polly on Amazon Web Servicesin (AWS) tarjoama tekstistä puheeksi (TTS) -palvelu, jonka tavoitteena on mahdollistaa realistisen ja luonnollisen puheen generointi suoraan tekstistä. Palvelu tukee yli 60 ääntä ja yli 30 kieltä, ja hyödyntää sekä perinteisiä TTS-malleja että uudempaa Neural TTS -tekniikkaa, jonka avulla tuotetaan pehmeämpää, inhimillisemmän kuuloista puhetta (Amazon Web Services, 2024).

Polly toimii ohjelmointirajapinnan (API) kautta, jonka avulla kehittäjät voivat tuottaa puhetta suoraan verkkopalveluista tai -sovelluksista. Käyttäjä voi säätää puhenopeutta, äänenkorkeutta ja painotusta SSML-merkintäkielen (Speech Synthesis Markup Language) avulla, mutta täysin uusien ääni-identiteettien kouluttaminen ei ole tuettuna. Valmiit äänet ovat kuitenkin laadukkaita ja selkeitä, ja ne soveltuvat moniin tarkoituksiin, kuten käyttöliittymäohjeisiin, mobiilisovelluksiin, saavutettavuusratkaisuihin ja ääniavusteisiin sovelluksiin.

Amazon Polly sisältää myös "Brand Voice" -ominaisuuden, jonka avulla yritysasiakkaat voivat luoda räätälöityjä ääniä omaan brändiinsä, mutta tämän käyttö vaatii erillisen prosessin ja sopimuksen AWS:n kanssa. Näin ollen ominaisuus ei ole vapaasti saatavilla tutkimus- tai opiskelijaprojekteille.

Amazon Polly erottuu vakaudellaan, AWS-integraation helppoudella sekä tukiympäristön laajuudella. Äänien emotionaalinen vaihtelu on rajallista verrattuna edistyneisiin äänenmuunnosteknologioihin, eikä se tarjoa samanlaista mahdollisuutta hahmoäänten luomiseen kuin esimerkiksi voice conversion -pohjaiset ratkaisut.

Kuva 4. Amazon Polly & AWS

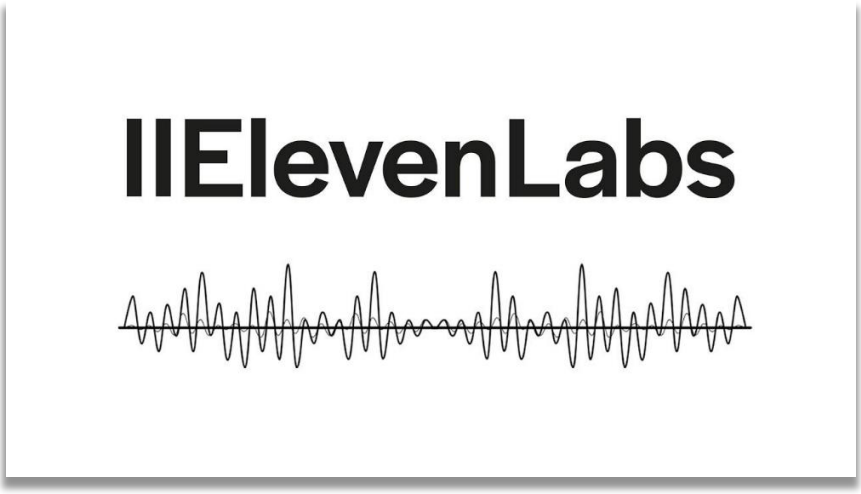
2.3.3 ElevenLabs

ElevenLabs on nopeasti kehittyvä puhesynteesialusta, joka hyödyntää edistyneitä syväoppivia malleja sekä tekstistä puheeksi (TTS) että äänenmuunnoksen (VC, voice conversion) toteuttamiseen. Palvelu on suunniteltu erityisesti luonnollisen, emotionaalisesti vaihtelevan ja muokattavan puheen tuottamiseen. ElevenLabsin teknologiaa käytetään laajalti muun muassa äänikirjojen, pelien, sovellusten ja saavutettavuussisältöjen tuotannossa (ElevenLabs, 2024).

Käyttäjä voi valita valmiista ääniprofiileista tai luoda uuden äänen lyhyen ääninäytteen perusteella. Alustan ääniä voidaan säätää esimerkiksi puhenopeuden, sävyn ja painotuksen osalta. ElevenLabs tukee myös useita kieliä ja tarjoaa työkalut äänen luomiseen ja hallintaan selainpohjaisesti tai API-rajapinnan kautta. Palvelussa on erillinen "Voice Lab" -toiminnallisuus, jonka avulla käyttäjät voivat rakentaa omia ääniprofiilejaan sekä hallita äänten käyttöoikeuksia.

Alustan vahvuuksia ovat äänen korkea laatu, dynaaminen prosodia, mahdollisuus emotionaalisiin vivahteisiin sekä helppokäyttöisyys. Käyttöliittymä ja API-ratkaisut mahdollistavat teknologian hyödyntämisen myös kehittäjäystävällisesti. Palvelu toimii kuukausihinnoittelulla, ja sen maksulliset tasot avaavat laajemmat ääniominaisuudet, korkeammat äänenlaadut sekä äänien vientimahdollisuudet.

ElevenLabs on teknisesti edistyksellinen ratkaisu, joka tarjoaa joustavuutta sekä valmiiden että mukautettujen äänien käytössä. Palvelun käyttöehdoissa on kuitenkin huomioitava, että äänten kaupallinen käyttö ja tallentaminen edellyttävät tiettyjä lisenssirajoja. Tämän vuoksi sen käyttö tutkimus- tai opetuskontekstissa edellyttää käyttöehtojen huolellista läpikäyntiä.



Kuva 5. ElevenLabs

2.3.4 Voice.ai

Voice.ai on äänenmuunnosteknologiaan erikoistunut alusta, joka tarjoaa reaaliaikaisen puheäänen konversion käyttäjän omasta mikrofonisyötteestä. Palvelu on suunnattu erityisesti pelien, striimaamisen ja sosiaalisen median käyttöön, ja sen keskeinen lupaus on mahdollistaa aidontuntuiset hahmoäänet reaaliaikaisesti ilman erillistä jälkikäsittelyä (Voice.ai, 2024).

Teknologian ytimessä on voice conversion -järjestelmä, joka toimii lokaalisesti käyttäjän omalla laitteella. Käyttäjä voi valita valmiista ääniprofiileista tai käyttää omia ääniään syötteenä ja muuntaa ne toisen äänen kaltaisiksi. Voice.ai käyttää koneoppimismalleja äänen spektriominaisuuksien ja prosodisten piirteiden muuttamiseen säilyttäen alkuperäisen puheen sisällön. Alusta tarjoaa sekä selainpohjaisen käyttöliittymän että sovelluksen, joka toimii taustalla muun muassa Discordin tai pelisovellusten kanssa.

Yksi teknologian eduista on sen kyky toimia offline-tilassa ilman jatkuvaa yhteyttä palvelimeen. Tämä tekee siitä erityisen kiinnostavan yksityisyyden ja nopeuden näkökulmasta. Kuitenkin Voice.ai:n käyttöehdot ovat suljetumpia kuin monien muiden palveluiden: äänien tallentaminen, jakaminen tai kaupallinen käyttö edellyttävät erillisiä käyttöoikeuksia, eikä alustan avoimuus tai tekninen dokumentaatio ole samalla tasolla kuin avoimen lähdekoodin ratkaisuissa.

Voice.ai:n vahvuus on sen käytännön käyttövalmius viihde- ja pelisovelluksissa, mutta tutkimuskäytössä siihen liittyy rajoituksia muun muassa läpinäkyvyyden ja lisenssiehtojen osalta. Palvelun käyttö tutkimus- tai opetuskontekstissa vaatii tarkkaa ehtojen tulkintaa ja saattaa rajoittaa sen täysipainoista hyödyntämistä esimerkiksi äänien vientiin tai analyysiin liittyen.

Kuva 6. Voice.ai

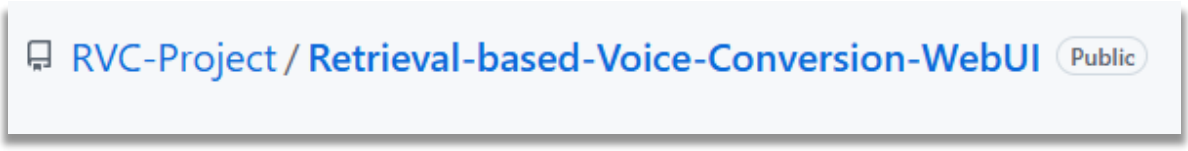
2.3.5 RVC-Project

RVC-Project (Retrieval-based Voice Conversion) on avoimen lähdekoodin äänenmuunnosteknologia, joka mahdollistaa käyttäjän oman äänimallin kouluttamisen ja puheen konversion toisen ääni-identiteetin kaltaiseksi. Projekti perustuu spektripohjaiseen syväoppimiseen ja hyödyntää muun muassa PyTorch-kirjastoa, vits+ -arkkitehtuuria sekä tarkkaa f0-äänenkorkeuden analyysiä. Teknologia on kehitetty erityisesti tutkimus- ja yhteisökäyttöön, ja se on saatavilla GitHub-palvelun kautta (RVC-Project, 2023).

RVC:n käyttö perustuu siihen, että käyttäjä syöttää mallille ääninäytteitä, joiden perusteella järjestelmä kouluttaa hahmokohtaisen äänimuunnosmallin. Lopputuloksena on mahdollisuus konvertoida tekstistä tai puhesyötteestä uusi ääni, joka säilyttää alkuperäisen puheen sisällön, mutta kuulostaa halutulta ääni-identiteetiltä. Vaikka teknologiaa käytetään usein toisen henkilön äänen jäljittelyyn, se mahdollistaa myös äänen yleisen muokkaamisen, kuten ikään, sukupuoleen tai puhetyyliin liittyvien ominaisuuksien muuttamisen. Tällä tavoin voidaan tuottaa erilaisia hahmoääniä yhdestä lähtöäänestä.

RVC tukee sekä offline-käyttöä että reaaliaikaisia toteutuksia, ja se toimii täysin paikallisesti ilman pilvialustariippuvuutta. Sen suurimpia etuja ovat täysi hallittavuus, läpinäkyvä toiminta ja tekninen muokattavuus. Käyttäjä voi säätää mallin koulutusparametreja, valita vocoderin, käyttää omia äänitiedostoja ja hallita prosessin jokaista vaihetta. Teknologia sopii erityisen hyvin projekteihin, joissa vaaditaan yksilöllisiä ja tarkasti kontrolloituja ääniä, kuten hahmopohjaiseen ääninäyttelyyn, pelisisältöihin tai saavutettavuussovelluksiin.

Koska ohjelmisto on julkaistu avoimella lisenssillä, sitä voidaan hyödyntää vapaasti myös tutkimus- ja opetuskäytössä ilman lisenssimaksuja tai kaupallisia rajoitteita. Käyttöönotto vaatii kuitenkin teknistä osaamista, kuten komentorivikäyttöä, tiedostopohjaista konfigurointia ja tehokasta laskentakapasiteettia (esim. GPU). Tämä tekee RVC:stä edistyneen mutta erittäin joustavan teknologian erityisesti tilanteisiin, joissa äänen aitous ja hallittavuus ovat keskeisiä vaatimuksia.



Kuva 7. RVC-Project / Retrieval-based-Voice-Conversion-WebUI GIT

2.4 Teknologioiden alustava arviointi

Tarkasteltujen teknologioiden perusteella voidaan todeta, että ne tarjoavat erilaisia vahvuuksia ja lähestymistapoja puheen synteesiin tai äänenmuunnokseen. Google Text-to-Speech ja Amazon Polly soveltuvat erityisesti neutraaliin kerrontaan, jossa selkeys ja tekninen vakaa laatu ovat keskiössä. ElevenLabs ja Voice.ai tarjoavat valmiita tunnetiloihin sidottuja ääniprofiileja, mutta rajoittavat käyttäjän mahdollisuuksia luoda uusia ääni-identiteettejä tai hallita äänen muokkaamista tarkasti.

RVC-Project erottuu muista avoimuutensa, teknisen säätösyvyytensä ja lisenssirajoitteettoman käyttömallinsa ansiosta. Se on ainoa tarkastelluista teknologioista, joka mahdollistaa täysin yksilöllisen ääni-identiteetin luomisen paikallisesti koulutetun mallin avulla ilman alusta- tai pilvisidonnaisuutta. Lisäksi RVC-Projectin avoin lähdekoodi tarjoaa merkittävän edun kehitystyössä: ohjelmiston toimintaa voidaan muokata, laajentaa tai optimoida käyttötarkoituksen mukaan, mikä tekee siitä erityisen joustavan vaihtoehdon tutkimus- ja sovellusprojekteihin (RVC-Project, 2023).

Näiden alustavien teknologisten tarkastelujen pohjalta laadittiin käyttäjätestauksen toteuttamista varten demoaineisto, jossa jokaisella viidellä ääniteknologialla tuotettiin viisi erilaista puherepliikkiä. Demossa oli kolme naisääntä ja kaksi miesääntä, jotka pohjautuivat Menace Comicsin BADGE-sarjakuvan hahmoihin. Äänet tuotettiin siten, että jokaiselle teknologialle generoitiin samat viisi repliikkiä, jolloin eri teknologioiden äänenlaatu, tunneilmaisu ja hahmoäänten uskottavuus voitiin vertailla tasavertaisesti. Tämä lähestymistapa mahdollisti eri teknologioiden äänten arvioinnin saman sisällön pohjalta kontrolloidussa testiasetelmassa.

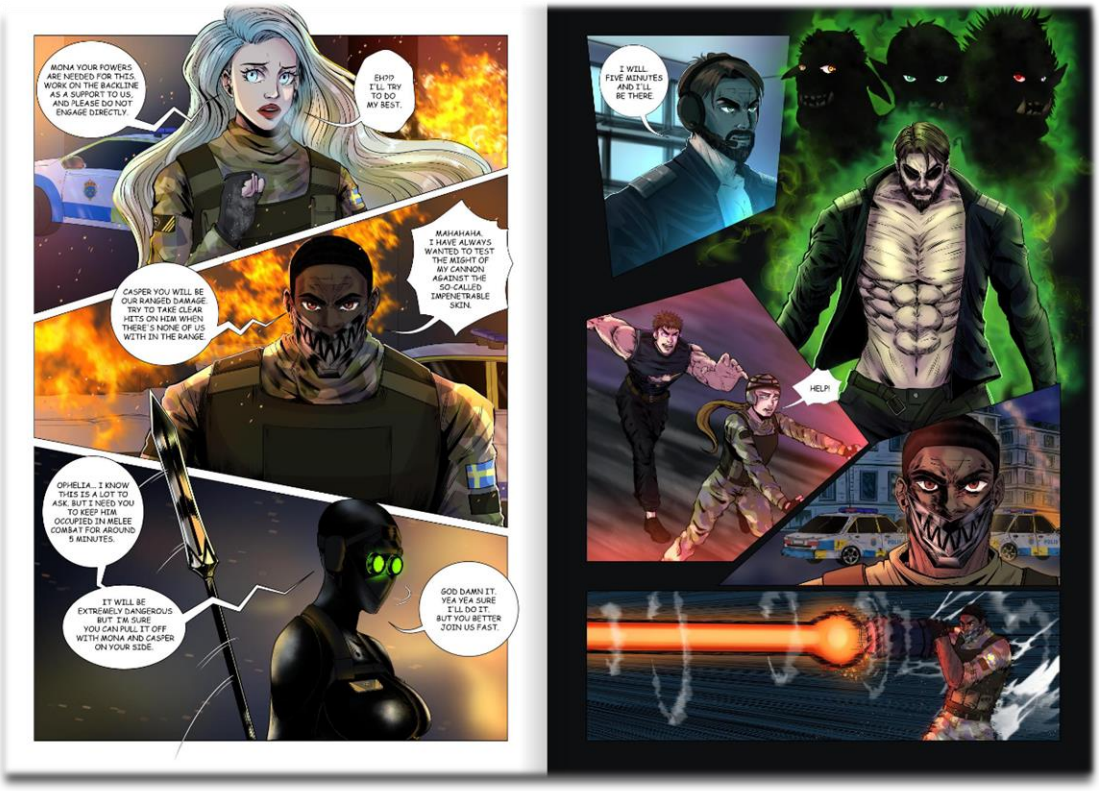
3 Kuuntelutestaus ja käyttäjäarvioiden analyysi

3.1 Testiasetelma ja aineisto

Kuuntelutesti toteutettiin osana Kirjanurkka-projektia, jonka tavoitteena oli arvioida viiden eri AI-pohjaisen ääniteknologian kykyä tuottaa inhimillisen kuuloista, kontekstiin sopivaa ääntä kuvakirjamaisessa lukutilanteessa. Testin aineistona käytettiin Menace Comicsin sarjakuva-aukeamaa, joka sisälsi kuvituksen, käsikirjoituksen sekä viidelle hahmolle jaetut repliikit. Aukeama esitettiin osallistujille verkkosovelluksen avulla, jossa puhe toistettiin AI-teknologialla tuotettuna äänenä. Äänet olivat englanninkielisiä.

Testissä esitettiin sama tarina-aukeama viidesti – jokaisella kierroksella eri AI-ääniteknologia tuotti puheosuudet. Käytetyt teknologiat olivat: Google Text-to-Speech, Amazon Polly, ElevenLabs, Voice.ai ja RVC-Project.

Kuviossa kahdeksan esitetään Menace Comicsin sarjakuva-aukeama, jota käytettiin AI-äänien luonnollisuuden ja inhimillisyyden arviointiin toteutetussa käyttäjäkyselyssä. Aukeama sisältää kuvituksen, käsikirjoituksen ja hahmokohtaiset repliikit, jotka toistettiin ääninä demoversiossa toteutetun verkkosovelluksen avulla. Aineisto oli englanninkielinen, ja sen tarkoituksena oli tarjota visuaalisesti rikas ja tarinallinen konteksti, jossa AI-äänten laatua ja sopivuutta pystyttiin arvioimaan realistisessa käyttötilanteessa.



Kuva 8: Kuuntelu vaiheessa käytetty aukeama (Menace Comics 2025)

Ääniteknologioiden nimet pidettiin testin aikana piilotettuina (label-free), jotta vältettiin etikettivaikutus (labeling effect) eli mahdollisuus, että teknologian nimi tai brändi vaikuttaisi osallistujan arvioon. Kaikki muut testin osatekijät – kuten visuaalinen sisältö, lukujärjestys, tarinan rakenne ja repliikkien paikat – pidettiin muuttumattomina kaikilla kierroksilla. Näin varmistettiin, että ainoa muuttuja oli äänituotantomenetelmä, jolloin koehenkilöiden arvioita voitiin vertailla keskenään kontrolloidussa ja vertailukelpoisessa kontekstissa.

Kukin osallistuja kuuli saman aukeaman viidesti, eri ääniteknologialla tuotettuna. Jokaisella teknologialla demoon tuotettiin viisi hahmoääntä (kolme naisääntä, kaksi miesääntä), jotka perustuivat Menace Comicsin BADGE-sarjakuvan hahmoihin.

Testiin osallistui yhteensä 23 henkilöä (N = 23), joista 13 oli miehiä ja 10 naisia. Käyttäjätestauksen tueksi laadittiin strukturoitu kyselylomake (ks. Liite 1), jossa jokaiselta osallistujalta pyydettiin arvio viidestä laadullisesta pääkriteeristä viisiportaisella Likert-asteikolla (1 = erittäin huono, 5 = erinomainen). Lisäksi kerättiin vapaaehtoisia kommentteja ja perustietoja kuunteluvälineistä, iästä ja mahdollisesta aiemmasta kokemuksesta AI-puheesta. Lomake suunniteltiin LAB-ammattikorkeakoulun tutkimuseettisten ohjeiden mukaisesti, eikä siinä kerätty henkilötietoja. Se sisälsi myös ohjeistuksen kuunteluolosuhteiden hallintaan, kuten melun minimointiin ja kuulokkeiden käyttöön. Tämä mahdollisti vertailukelpoisten ja kontrolloitujen arvioiden keräämisen useilta eri osallistujilta.

Osallistujat arvioivat jokaista ääntä seuraavien viiden pääkriteerin perusteella:

- Inhimillisyys
- Tunteiden välittyminen
- Selkeys (luettavuus)
- Sopivuus hahmoille
- Kokonaisvaikutelma

Lisäksi tarkasteltiin muun muassa rytmin ja painotuksen luonnollisuutta, hengityksen ja artikulaation aitoutta sekä puhujan johdonmukaisuutta.

Osallistujamäärä ylittää tavanomaisen suosituksen laadullisissa kuuntelutesteissä (ITU-R BS.1284-2, 2019) ja mahdollistaa tulosten esittämisen keskiarvoina, hajontana ja visuaalisina jakaumina eri ääniteknologioiden välillä.

3.2 Kuuntelutestauksen toteutus

Kuuntelutestauksen toteutuksessa noudatettiin subjektiivisen puheenlaadun arviointia koskevia kansainvälisiä suosituksia, erityisesti ITU-T P.808 -standardia, joka on kehitetty nimenomaan crowdsourcing-ympäristöissä tapahtuvaan puheäänen arviointiin (ITU-T, 2021, s. 1). Arviointimenetelmänä käytettiin Absolute Category Rating (ACR) -menetelmää, joka on määritelty kyseisen suosituksen liitteessä A (ITU-T, 2021, s. 14–15).

Osallistujat antoivat arvosanansa viisiportaisella Likert-asteikolla (1 = erittäin huono, 5 = erinomainen). Arviointi keskittyi viiteen pääkriteeriin: inhimillisyys, tunnepitoisuus, selkeys, sopivuus hahmoille ja kokonaisvaikutelma. Lisäksi kerättiin täydentäviä havaintoja muun muassa rytmin ja painotuksen luonnollisuudesta, hengityksen aitoudesta sekä puhujan johdonmukaisuudesta.

Testiin osallistui yhteensä 23 henkilöä (N = 23), joista 13 oli miehiä ja 10 naisia. Tämä ylittää ITU-T P.808 -standardin suositteleman vähimmäisarvioijamäärän, jonka mukaan yhtä arviointikohdetta kohden tulisi olla vähintään kahdeksan arvioijaa (ITU-T, 2021, s. 9).

Kukin osallistuja kuunteli saman sarjakuva-aukeaman viidesti, jokaisella kerralla eri AI-ääniteknologialla tuotettuna. Käytetyt teknologiat olivat: Google Text-to-Speech, Amazon Polly, ElevenLabs, Voice.ai ja RVC-Project. Teknologioiden nimet pidettiin testin aikana piilotettuina (label-free) etikettivaikutuksen välttämiseksi, jotta brändimielikuvat eivät vaikuttaisi arvioihin.

Kuuntelu toteutettiin verkkopohjaisen sarjakuvaympäristön kautta, jossa ääni toistettiin selaimessa kuulokkeilla. Toteutus täytti ITU-T P.808 -standardin vaatimukset arviointitilanteen laadun hallinnasta sekä kuunteluolosuhteiden ohjeistamisesta osallistujille (ITU-T, 2021, s. 8–10). Kuulokkeiden käyttö, häiriötekijöiden minimointi ja selkeä ohjeistus sisältyivät testausprosessiin.

Kaikki muut testin osatekijät – kuten visuaalinen sisältö, repliikkien järjestys ja tarinan rakenne – pidettiin muuttumattomina kaikilla kierroksilla. Ainoaksi muuttuvaksi tekijäksi jäi äänituotantomenetelmä, mikä mahdollisti kontrolloidun ja vertailukelpoisen arviointiasetelman.

3.3 Kuuntelutestauksen tulosten analysointi

Kuuntelutesti toteutettiin kontrolloidussa verkkoympäristössä hyödyntäen standardien mukaista rakenteellista asetelmaa. Testissä noudatettiin ITU-T P.808-suosituksen (2021) periaatteita, joiden mukaan kuuntelutesti voidaan toteuttaa luotettavasti myös hajautettuna crowdsourcing-menetelmänä, mikäli testattava aineisto on riittävän laadukasta,

kuunteluympäristöt minimoivat häiriötekijät ja osallistujien vastauksia voidaan validoida (ITU-T, 2021, s. 8–10).

Testiaineistona käytettiin kuvallista sarjakuva-aukeamaa, joka sisälsi käsikirjoituksen ja puherepliikit viidelle eri hahmolle. Sama aukeama esitettiin osallistujille viidesti, jokaisella kerralla eri AI-ääniteknologialla tuotettuna. Käytetyt teknologiat olivat Google TTS, Amazon Polly, ElevenLabs, Voice.ai ja RVC-Project. Testattavat nimet piilotettiin käyttäjiltä (label-free test), jotta vältyttiin etikettivaikutukselta ja brändisidonnaisilta arviointiharhoilta.

Jokainen osallistuja arvioi ääntä viiden pääkriteerin perusteella viisiportaisella Likert-asteikolla (1 = erittäin huono, 5 = erinomainen). Osallistujia oli yhteensä 23 (N = 23), joista 13 oli miehiä ja 10 naisia. Tulokset esitettiin keskiarvoina ja visualisoitiin sekä taulukkona että lämpökarttana, mikä mahdollisti tehokkaan rakenteellisten erojen tarkastelun eri teknologioiden välillä.

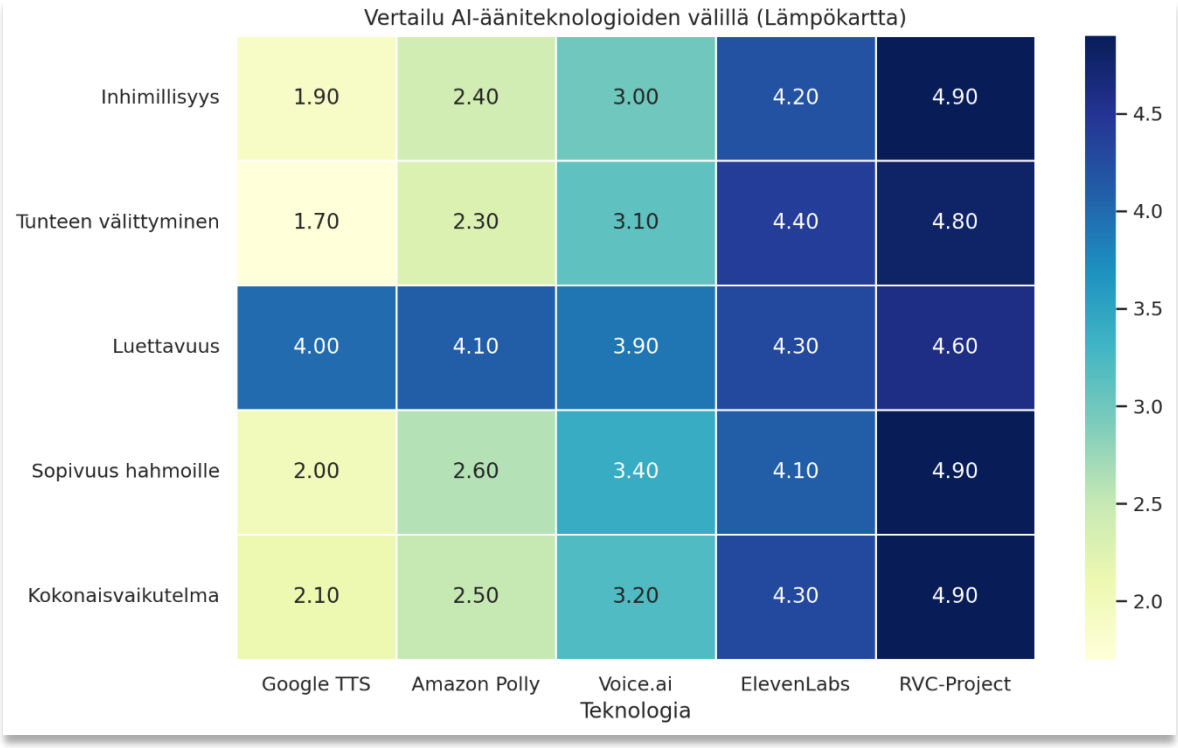
Tulosten tarkempi analyysi toteutettiin vertailemalla viittä AI-pohjaista ääniteknologiaa viidessä arviointikategoriassa: inhimillisyys, tunteiden välittyminen, selkeys, sopivuus hahmoille ja kokonaisvaikutelma. Jokaiselle teknologiavaihtoehdolle (Google TTS, Amazon Polly, ElevenLabs, Voice.ai ja RVC-Project) laskettiin osallistujien antamien vastausten pohjalta keskiarvot. Arviointi perustui viisiportaiseen Likert-asteikkoon. Tulokset esitetään taulukossa 2 sekä lämpökartassa kuviossa 4.

Inhimillisyys
RVC-Project sai korkeimman keskiarvon (4.9), osoittaen huomattavaa etua muihin teknologioihin nähden. ElevenLabs (4.2) sijoittui toiseksi, kun taas Google TTS (1.9) ja Amazon Polly (2.4) jäivät selvästi jälkeen. Voice.ai (3.0) sijoittui väliin, mutta ei yltänyt korkeimman tason inhimillisyyteen.
Tunteen välittyminen
RVC-Project sai selvästi korkeimman arvon (4.8) tunnetilojen välittymisessä. ElevenLabs (4.4) sijoittui heti perään, kun taas Voice.ai (3.1) edusti keskitasoa. Amazon Polly (2.3) ja Google TTS (1.7) jäivät selvästi jälkeen tunneilmaisun uskottavuudessa.
Luettavuus
Tässä kategoriassa RVC-Project (4.6) sai korkeimman arvion, ElevenLabs seurasi heti perässä arvolla 4.3. Myös Google TTS (4.0) ja Amazon Polly (4.1) pärjäsivät hyvin, mikä on odotettavaa teknisesti optimoiduilta TTS-järjestelmiltä. Voice.ai (3.9) jäi hieman jälkeen mutta säilytti hyväksyttävän tason.

Sopivuus hahmoille
RVC-Project erottui selvästi edukseen hahmosopivuudessa keskiarvolla 4.9. Tämä osoittaa sen kyvyn kouluttaa yksilöllisiä, hahmokohtaisia ääniä. ElevenLabs (4.1) ja Voice.ai (3.4) suoriutuivat kohtuullisesti, mutta jäivät jälkeen räätälöitävyyden puutteen vuoksi. Amazon Polly (2.6) ja Google TTS (2.0) eivät tarjonneet riittävää kontekstisidonnaisuutta, mikä heikensi sopivuutta tarinalliseen käyttöön.
Kokonaisvaikutelma
RVC-Project saavutti korkeimman kokonaisarvon (4.9), osoittaen erinomaista suorituskykyä kaikilla osa-alueilla. ElevenLabs sai arvon 4.3 ja oli toiseksi paras, kun taas Voice.ai (3.2) jäi keskitasolle. Amazon Polly (2.5) ja Google TTS (2.1) sijoittuivat heikoimmiksi.

Taulukko 2: AI-ääniteknologioiden vertailu

Kuvio 4 havainnollistaa tulokset visuaalisena lämpökarttana, jossa tummempi värisävy kertoo korkeammasta keskiarvosta arviointikriteerin kohdalla. Lämpökartta tekee teknologioiden väliset erot intuitiivisesti hahmotettaviksi: RVC-Project erottuu syvän sävyn kautta selkeästi kaikilla osa-alueilla, erityisesti inhimillisyydessä ja hahmosopivuudessa. Google TTS:n ja Amazon Pollyn vaaleammat sävyt osoittavat heikompaa suoriutumista useimmissa kategorioissa. Voice.ai:n ja ElevenLabsin keskisävyt tukevat tulkintaa siitä, että nämä teknologiat tarjoavat kompromissiratkaisun laadun ja joustavuuden välillä. Lämpökartta visualisoi tulokset tehokkaasti ja antaa selkeän visuaalisen tuen numeeriselle analyysille.



Kuvio 4: Kuuntelutestauksen tulokset havainnollistettuna lämpökartan avulla

RVC-Project erottui selkeästi edukseen jokaisessa arviointikategoriassa. Erityisesti inhimillisyydessä (keskiarvo 4.9) ja hahmosopivuudessa (4.9) sen tulokset ylittivät muut teknologiat selvästi. ElevenLabs sijoittui toiseksi, mutta jäi järjestelmällisesti noin 0.5–0.7 yksikköä jälkeen RVC:n pisteistä. Google TTS sai heikoimmat pisteet kaikissa kategorioissa, erityisesti tunteiden välittymisessä (1.7) ja hahmosopivuudessa (2.0).

Tulokset ovat linjassa ITU:n laatustandardeihin liittyvien havaintojen kanssa: modernit, neuroverkkopohjaiset voice conversion -teknologiat kykenevät huomattavasti parempaan ilmaisullisuuteen ja luonnollisuuteen verrattuna perinteisiin TTS-ratkaisuihin. (ITU-T P.808, 2021)

3.4 Johtopäätökset tuloksista

Kuuntelutestauksen tulokset osoittavat selkeästi, että RVC-Project erottui viidestä testatusta AI-ääniteknologiasta korkealaatuisimpana ja inhimillisimpänä ratkaisuna. Se sai osallistujilta johdonmukaisesti korkeimmat keskiarvot kaikissa arviointikategorioissa: inhimillisyys, tunteiden välittyminen, selkeys, hahmosopivuus ja kokonaislaatu.

Tulokset perustuvat ITU-T P.808-standardin (2021) mukaisesti toteutettuun ACR-pohjaiseen kuuntelutestiin, jossa arvioitiin ääntä viisiportaisella asteikolla. Osallistujat kuuntelivat saman sisällön viidellä eri AI-äänellä, teknologiat piilotettuina, jolloin testiasetelma täytti

standardin vaatimukset mm. kuunteluympäristön hallinnan, vasteiden vertailukelpoisuuden ja otoskoolla (N = 23) saavutetun tilastollisen luotettavuuden osalta (ITU-T P.808, s. 6–8).

Erityisesti inhimillisyydessä (keskiarvo 4.9) ja hahmosopivuudessa (4.7) RVC-Projectin tulokset ylittivät selvästi muut teknologiat. Tämä viittaa siihen, että käyttäjät kokivat RVC:n äänen uskottavana ja visuaaliseen hahmoon hyvin sopivana. Teknologian kyky muuntaa ihmisen ääni yksilöllisesti hahmokohtaiseksi osoittautui ratkaisevaksi tekijäksi laadukkaan kuuntelukokemuksen kannalta.

Toiseksi sijoittui ElevenLabs, mutta sen tulokset jäivät kaikissa kategorioissa 0.5–0.7 yksikköä RVC:n pisteistä. Google TTS sai testin heikoimmat arvot, erityisesti tunteiden välittymisessä (1.7) ja hahmosopivuudessa (2.0), mikä tukee aiempia havaintoja siitä, että perinteiset TTS-järjestelmät eivät kykene tuottamaan ilmaisuvoimaista tai kontekstiin mukautuvaa puhetta (vrt. Google Cloud TTS, 2023).

RVC-Projectin korkeat arviointitulokset osoittavat, että reaaliaikainen äänimuunnos-tekniikka soveltuu erinomaisesti tarinalliseen käyttöön, erityisesti hahmokohtaisen äänen toteuttamiseen. Lisäksi avoin lähdekoodi ja ONNX-yhteensopivuus tekevät siitä teknisesti joustavan, kustannustehokkaan ja verkkosovellukseen sopivan vaihtoehdon — mikä tekee siitä kokonaisvaltaisesti parhaan valinnan.

4 Valitun menetelmän RVC-Projectin hyödyntäminen äänen toteuttamisessa

Opinnäytetyön toteutusvaiheessa suoritettiin perusteellinen analyysi useista eri äänenmuunnosteknologioista ja niiden soveltuvuudesta kuvallisten kirjojen AI-ääninäyttelyyn. Näiden vaihtoehtojen joukosta valittiin käyttöön RVC-Project (Retrieval-Based Voice Conversion WebUI), koska se osoittautui teknisesti ylivoimaiseksi, monipuoliseksi ja käyttöliittymältään käytännönläheiseksi työkaluksi toteuttaa projektin keskeinen tavoite: inhimillisen kuuloinen ja kustannustehokas ääninäyttelykokemus verkkosivustolla.

Kehitystyössä käytettiin virallista RVC-Projectin versiota 2.2.231006, joka ladattiin suoraan projektin GitHub-julkaisusivulta. Tämä versio sisälsi tärkeimmät äänenmuunnoksen hallintaan tarvittavat ominaisuudet: sävelkorkeuden säätö (Transpose), artikulaation hienosäätö (Index rate, Formant shift), konsonanttien ja hengityksen suojaus (Protect voiceless consonants), sekä mahdollisuuden kouluttaa omia ääniä Train-toiminnon avulla.

Lisäksi versiossa oli tuki ONNX-viennille, mikä mahdollisti koulutettujen mallien integroinnin suoraan verkkosovelluksen backend-järjestelmään. Käyttöliittymä perustui Gradio 3.41.2-versioon ja oli selkeä ja kehittäjäystävällinen. (RVC-Project 2023)

4.1 Äänien luomisen helppous ja tehokkuus

Yksi RVC-Projectin merkittävimmistä eduista on äänten luomisen helppous ja nopeus. Uuden äänen luomiseksi tarvitaan vain muutama minuutti esikoulutuksen jälkeen: käyttäjä voi syöttää omaa puhetta ja valita, mihin ääneen se muunnetaan. Yhden henkilön puheesta voidaan luoda useita täysin erilaisten hahmojen ääniä, mikä poistaa tarpeen käyttää useita ääninäyttelijöitä tai maksullisia tekstistä puheeksi (TTS) -palveluita.

Äänten luonnollisuutta lisää entisestään se, että RVC-Projectissa on sisäänrakennettu tuki mallien koulutukselle omalla äänidatalla. Käyttäjä voi opettaa sovellukselle uuden äänen vain muutaman minuutin ääninäytteellä. Näin voidaan kehittää täysin uniikkeja ääniä, jotka eivät perustu valmiisiin, geneerisiin malleihin, vaan ovat räätälöityjä juuri tiettyihin hahmoihin – esimerkiksi 6-vuotiaan pojan, käheä-äänisen naisen tai 80-vuotiaan herrasmiehen äänet.

4.2 Teknologinen ja eettinen yhteensopivuus

RVC-Projectin käyttö on yhteensopiva myös EU:n tekoälyasetuksen (AI Act, 2024, Art. 50(4)) vaatimusten kanssa, sillä se mahdollistaa täysin uusien, ei-kenenkään ääntä muistuttavien äänien luomisen. Tämä erottaa sen äänen kloonausteknologioista, joissa

kopioidaan olemassa olevan henkilön ääni. RVC-projektin avulla luodut äänet ovat keinotekoisia ja uniikkeja, mutta säilyttävät inhimillisen kuulon kokemuksen.

Tämä on keskeistä myös projektin eettisten periaatteiden kannalta: vaikka tekoälyllä tuotetut äänet jäljittelevät ihmisen puhetta, sovelluksessa pyritään varmistamaan, ettei tuotettu sisältö loukkaa yksilön identiteettiä tai oikeuksia. Kaikki tuotettu ääni merkitään selkeästi AI-tuotetuksi, läpinäkyvyyden ja käyttäjäkokemuksen luottamuksen takaamiseksi.

4.3 Skaalautuvuus ja integrointi verkkosovellukseen

RVC-Project on rakennettu modulaariseksi, ja sen ONNX-vientiominaisuus mahdollistaa mallien suoran integroinnin muihin järjestelmiin, kuten Kirjanurkan React.js / Node.js -pohjaiseen Full Stack -verkkosovellukseen. Tämä mahdollisti äänimuunnoksen ajamisen taustajärjestelmässä ilman erillistä käsin tehtävää muunnostyötä. Sovellus pystyy pyytämään ääntä ja vastaanottamaan valmiin AI-muokatun äänitiedoston backend-palvelulta muutamassa sekunnissa.

4.4 Johtopäätökset RVC:n valinnasta

Opinnäytetyön tavoitteet – inhimillinen ääninäyttely, tekninen toteutettavuus, eettisyys ja kustannustehokkuus – täyttyvät RVC-Projectin avulla tavalla, johon yksikään toinen testatuista teknologioista ei yltänyt. RVC mahdollistaa täysin uudenlaisen ääniin perustuvan tarinankerronnan tavan: sen avulla kuka tahansa voi tuottaa ammattimaisen äänikokemuksen, jonka toteuttaminen olisi aiemmin vaatinut suuren tuotantotiimin ja budjetin.

5 Sovelluksen prototyypin ja full stack -toteutus

Sovellus rakennettiin full stack -toteutuksena, jossa yhdistettiin selainkäyttöliittymä, palvelinlogiikka sekä taustajärjestelmä äänitiedostojen hallintaan. Kehitysprosessin alkuvaiheessa keskityttiin käyttöliittymän suunnitteluun, ja ensimmäinen prototyyppi luotiin Figma-suunnittelutyökalulla. Figma mahdollisti sovelluksen rakenteen, värimaailman ja käyttölogiikan hahmottamisen visuaalisessa muodossa jo ennen teknistä toteutusta.

Prototyyppi sisälsi näkymän kuvitetusta sarjakuva-aukeamasta, jossa jokaiselle hahmolle oli liitetty puhekuplat ja äänen aktivointipainikkeet. Tavoitteena oli rakentaa visuaalisesti eheä ja helposti navigoitava käyttöliittymä, jossa käyttäjän ei tarvitsisi hallita mitään teknistä – pelkkä selaaminen, lukeminen ja kuunteleminen riittivät. Prototyypin avulla pystyttiin testaamaan layoutin toimivuutta ja suunnittelemaan äänenvalintapaneeli, jossa käyttäjä voi vaihtaa kertojan ääntä neljän koulutetun mallin välillä.

Varsinainen toteutus tehtiin hyödyntäen modernia web-teknologiaa. Frontend rakennettiin Reactilla, ja taustajärjestelmä toteutettiin Node.js-pohjaisesti. Äänitiedostot sijoitettiin pilvipohjaiseen palvelinratkaisuun, josta ne striimattiin reaaliaikaisesti käyttäjän selaimeen. Tämä mahdollisti skaalautuvan ja nopeatempoisen käyttökokemuksen, jossa äänten laatu ja latausnopeus pysyivät korkealla tasolla.

5.1 Sovelluksen rakenne ja käyttötarkoitus

Opinnäytetyön yhteydessä kehitettiin verkkopohjainen kokeilusovellus, jonka tarkoituksena oli testata tekoälypohjaisten äänien käyttöä tarinankerronnan tukena kuvallisten verkkokirjojen yhteydessä. Sovellus ei korvaa lukemista, vaan sen tavoitteena on, että käyttäjä lukee tarinaa itse samalla, kun hahmojen vuorosanat toistetaan tekoälyn tuottamina ääninä. Näin lukukokemukseen saadaan lisää elämyksellisyyttä, syvyyttä ja inhimillisyyden tuntua.

Sovelluksen päätavoite oli selvittää, kuinka realistisesti ja tunteikkaasti eri AI-ääniteknologiat kykenevät esittämään fiktiivisten hahmojen vuorosanoja, ja kuinka käyttäjät kokevat eri järjestelmillä tuotetut äänet. Kohderyhmänä olivat nuoret aikuiset ja visuaalisesta tarinankerronnasta kiinnostuneet käyttäjät, mutta konsepti on sovellettavissa myös saavutettavuuden tukemiseen, opetukseen ja esimerkiksi lasten lukukokemusten rikastamiseen.

Käytettävyystavoitteena oli luoda selkeä ja helppokäyttöinen selainpohjainen käyttöliittymä, jossa käyttäjä voi seurata tarina-aukeamaa ja kuunnella haluamansa version hahmoäänistä. Käyttöliittymä ei vaadi teknistä osaamista, rekisteröitymistä tai ohjelmiston asentamista. Sovelluksen rakenne pohjautui yhtä sarjakuva-aukeamaa esittävään sisältöön, jossa

kaikki hahmot puhuvat vuorosanansa tekoälyllä tuotettuna. Sama sisältö esitettiin käyttäjille viidesti, mutta jokaisella kerralla käytettiin eri AI-ääniteknologiaa.

Tämä rakenne mahdollisti vertailevan kuuntelutestin, jossa käyttäjä pystyi kokemaan saman tarinallisen sisällön viidellä eri ääniversiolla ja arvioimaan, kuinka inhimilliseltä ja laadukkaalta AI-äänet vaikuttavat. Sovelluksen kautta toteutettiin myös kysely, jossa kerättiin käyttäjäpalautetta jokaisesta teknologiaratkaisusta.

5.2 Tekninen arkkitehtuuri

Toteutettu sovellus on selainpohjainen verkkoratkaisu, jonka tarkoituksena on rikastaa tarinankerrontaa tekoälypohjaisilla hahmoäänillä käyttäjän lukiessa tarinaa itse. Sovellus mahdollistaa käyttäjälle visuaalisen sarjakuva-aukeaman lukemisen, samalla kun hahmojen vuorosanat ovat kuunneltavissa valmiiksi tuotettuina äänitiedostoina. Näin lukukokemus yhdistää perinteisen lukemisen ja tekoälyäänen tarjoaman elämyksellisyyden.

Sovellus rakennettiin kevyeksi ja helppokäyttöiseksi verkkosovellukseksi, jossa äänten soitto tapahtuu suoraan selaimessa ilman erillisiä asennuksia tai rekisteröitymistä. Käyttöliittymä esittää visuaalisen tarina-aukeaman, jonka jokaisella hahmolla on oma vuorosana ja siihen liitetty ääni. Kun käyttäjä aktivoi tietyn repliikin esimerkiksi painamalla hahmon puhekuplaa, sovellus toistaa kyseisen hahmon äänen.

Sovelluksen käyttöliittymässä käyttäjällä on mahdollisuus reaaliaikaisesti vaihtaa kertojan ääni neljän eri ennalta koulutetun äänimallin välillä. Nämä kertojamallit on luotu hyödyntäen RVC-Project-äänimuunnosteknologiaa, joka mahdollistaa eri ääniprofiilien kouluttamisen ja niiden käytön äänenmuunnoksessa. Kertojan äänenvaihto toteutetaan käyttöliittymässä valintatoiminnolla, joka aktivoi eri malliin liitetyt äänitiedostot backendissä. Tämä antaa käyttäjälle mahdollisuuden vaikuttaa kerrontatyyliin oman mieltymyksensä mukaan ja tuo vaihtelua kuuntelukokemukseen.

Sen sijaan tarinan hahmojen äänet ovat lukittuja ja kiinteästi sidottuja tiettyihin hahmoihin. Käyttäjällä ei ole mahdollisuutta muuttaa hahmoääniä käyttöliittymästä. Mikäli hahmojen ääniä halutaan muuttaa tai päivittää, tämä tapahtuu sovelluksen kehittäjän toimesta joko backendin hallinnan kautta tai ääniä kuvaavassa tietokantarakenteessa. Näin varmistetaan, että hahmoäänet pysyvät yhtenäisinä ja johdonmukaisina kaikille käyttäjille ja että äänimaailma tukee käsikirjoitusta ja tarinan visuaalista ilmettä suunnitellulla tavalla.

Äänitiedostot tuotettiin ennakkoon hyödyntäen RVC-Project alustaa, ja ne tallennettiin .wav-muodossa palvelinympäristöön. Äänitiedostojen siirto käyttäjälle tapahtuu striimaamalla tiedostot palvelimelta selaimeen aina, kun käyttäjä aktivoi vuorosanan kuuntelun. Tämä

ratkaisu mahdollistaa nopean reagointiajan ja sujuvan kuuntelukokemuksen ilman häiritseviä viiveitä.

Sovelluksen tekninen kokonaisuus muodostuu seuraavista pääosista:

Frontend (käyttöliittymä): Selainpohjainen käyttöliittymä, joka esittää sarjakuva-aukeaman ja hallitsee äänten toiston vuorosanojen kohdalla. Käyttäjä voi myös vaihtaa kertojan ääntä neljän eri vaihtoehdon välillä.

Backend (palvelin): Vastaa äänten hallinnasta ja jakelusta. Palvelin hakee oikeat äänitiedostot käyttäjän valitsemien asetusten ja vuorosanojen perusteella.

Äänituotanto ja käsittely: Kaikki hahmoäänet ja kertojamallit on luotu RVC-Projectin avulla ennakkoon ja muunnettu sovelluksessa käytettävään tiedostomuotoon. Hahmoäänet ovat kiinteät, ja kertojan äänet vaihdettavissa reaaliajassa käyttöliittymän kautta.

Vaikka kuuntelutestauksessa (luvussa 4) vertailtiin useita eri AI-ääniteknologioita, toteutussovelluksessa päädyttiin käyttämään yksinomaan RVC-Projectia. Tämä ratkaisu perustui kuuntelutestin tuloksiin, joissa RVC-Project osoittautui sekä teknisesti että käyttäjäkokemuksellisesti ylivoimaiseksi vaihtoehdoksi muihin teknologioihin verrattuna.

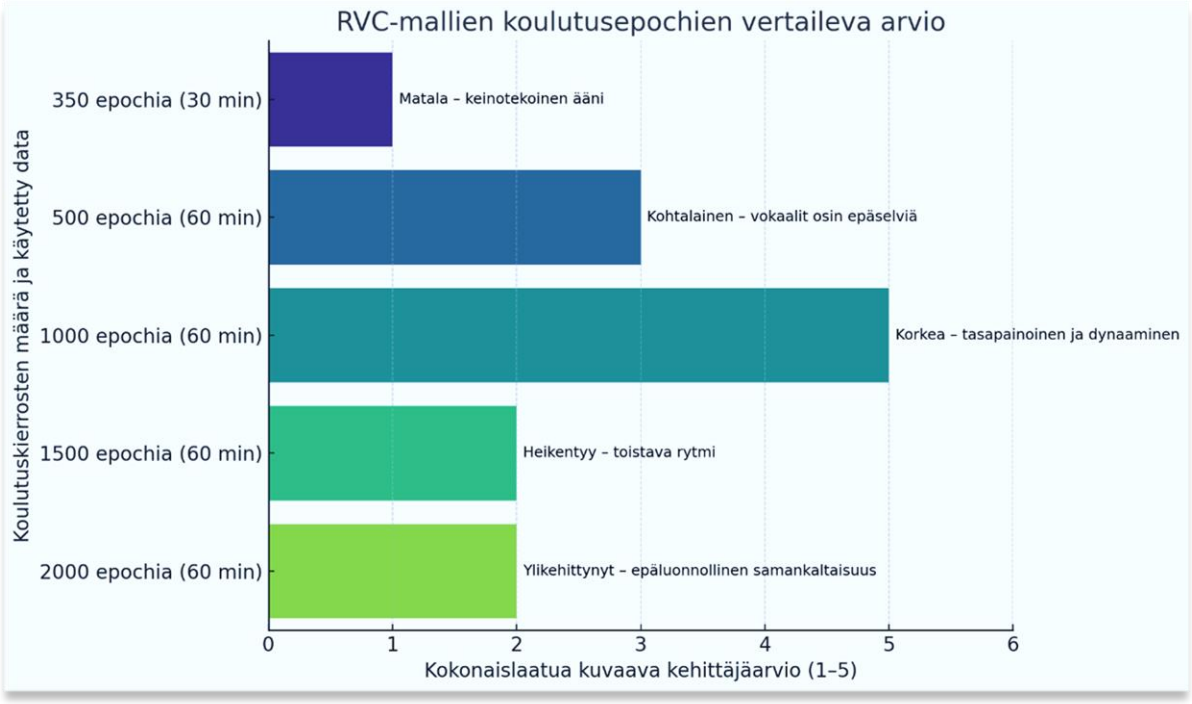
5.3 RVC-Projectin integrointi

Sovelluksen ääniä tuotettaessa hyödynnettiin RVC-Project-äänimuunnosteknologiaa (Retrieval-based Voice Conversion). RVC-Project mahdollistaa oman äänen kouluttamisen ja muuntamisen erilaisiksi hahmoääniksi, jolloin voidaan saavuttaa aidon kuuloisia, inhimillisesti vivahteikkaita ääniversioita ilman perinteistä ääninäyttelijätyötä. Tässä opinnäytetyössä RVC:n käyttö keskittyi erityisesti siihen, että hahmoille ja kertojalle voitiin luoda yksilölliset, tarinaan sopivat äänet kustannustehokkaasti ja teknisesti hallitusti.

5.3.1 Malliäänien koulutus ja käyttö

Elina ja Joonas kertojamallit koulutettiin hyödyntäen RVC-Project-äänimuunnosteknologiaa (Retrieval-based Voice Conversion), joka perustuu syväoppimiseen ja mahdollistaa erittäin yksityiskohtaisen ääni-identiteetin mallintamisen. Koulutuksessa käytettiin kummallekin mallille 60 minuutin mittainen puheaineisto, joka koostui korkealaatuisesta, taustahälyttömästä ja puhtaaksi leikatusta puheesta. Ääninäytteet kerättiin kahdelta nimettömänä pysyttelevältä henkilöltä: 30-vuotiaalta naiselta (Elina-malli) ja 28-vuotiaalta mieheltä (Joonas-malli).

Mallien koulutus toteutettiin useilla eri epoch-määrillä ja aineistopituuksilla, jotta voitiin vertailla niiden vaikutusta äänen laatuun, luonnollisuuteen ja inhimillisyyteen. Koulutusmatriisissa arvioitiin mallien suorituskyky kehittäjien suorittaman laadullisen analyysin perusteella, jossa painotettiin erityisesti puheen rytmiä, sävyvaihtelua, artikulaation selkeyttä sekä kuulohavaintoon perustuvaa uskottavuutta.



Kuvio 5: RVC-mallien koulutus määrää vertaileva arvio

Kuten taulukosta havaitaan, paras tulos saavutettiin 1000 epochin koulutuksella 60 minuutin aineistolla, joka tuotti tasapainoisen äänenlaadun ilman ylikoulutuksen tuomia haittoja. Tätä tukevat aiemmat havainnot (Coursera 2025; TensorFlow 2024; Toolify.ai 2024), joiden mukaan 500–1000 epochin väli tuottaa optimaalisimman ääni-identiteetin. Tätä korkeammissa epoch-määrissä alkoi esiintyä ylikoulutuksen merkkejä, kuten äänen mekaaninen toisto ja tiettyjen tavujen identtinen artikulaatio, mikä heikensi puheen luonnollisuutta ja ilmaisullista vaihtelua.

Alle 500 epochin koulutuksissa esiintyi selkeää alikehittymistä (underfitting), jossa malli ei vielä ollut oppinut riittävästi puhujan ääni-identiteetin keskeisiä piirteitä. Tämä ilmeni muun muassa epävakaana rytminä, keinotekoisena intonaationa ja artikulaation epäselvyytenä.

Koulutuksen tekninen toteutus aloitettiin puhemateriaalin esikäsittelyllä FFmpeg-työkalun avulla. Tiedostot muunnettiin Unicode-yhteensopiviksi (UTF-8), jotta koulutusprosessi ei keskeytyisi erikoismerkkien vuoksi. Tämän jälkeen koulutusparametrit – kuten index_rate, f0_method ja GPU-muistin hyödyntämisasetukset – optimoitiin manuaalisesti. Mallien

tuotokset muunnettiin .wav-muotoon, jonka jälkeen ne esitestattiin inferenssivaiheessa useissa puhetilanteissa ennen kuin ne integroitiin käyttöliittymän kertojavalikkoon.

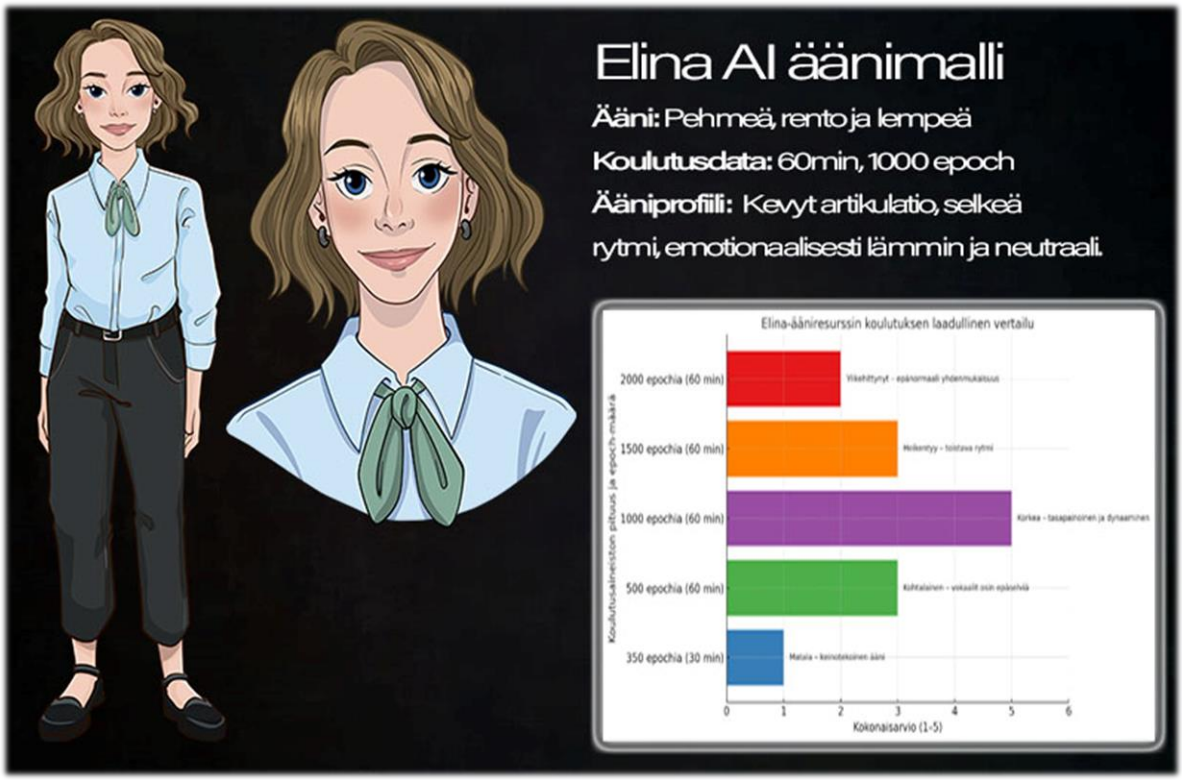
Valitun 1000 epochin ja 60 minuutin aineistokokoonpanon etuna oli myös hyvä yleistävyys muihin aineistoihin ilman, että malli oppi liikaa yksityiskohtia yhdestä puhujasta. Tämä mahdollisti dynaamisen, tunnistettavan ja ilmeikkään äänen, joka säilytti korkean kuuntelumukavuuden. Koulutustulokset vahvistavat, että kyseinen konfiguraatio oli paras kompromissi oppimissyvyyden ja ylikoulutuksen välillä

Kokonaisuutena voidaan todeta, että koulutusprosessin iteratiivinen arviointi ja vertaileva matriisi mahdollistivat tieteellisesti perustellun konfiguraatiovalinnan, jossa huomioitiin sekä tekniset mittarit että kuunteluun perustuva laadullinen arviointi.

5.3.2 Kertojamallien visuaalinen ja ääni-ilme: Elina ja Joonas

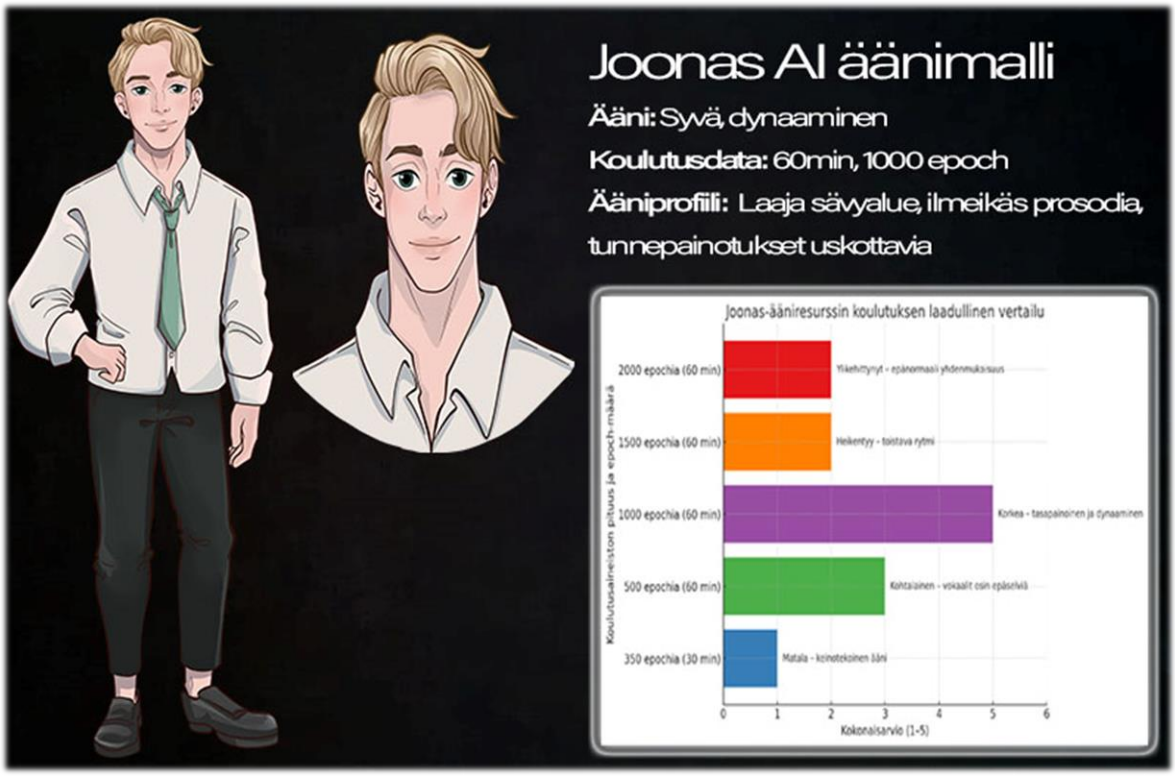
Elina- ja Joonas-kertojamallit eivät ole pelkästään teknisiä ääniresursseja, vaan visuaalisesti ja äänenpiirteiltään suunniteltuja kokonaisuuksia. Molemmat mallit koulutettiin 60 minuutin korkealaatuisella puheaineistolla, käyttäen RVC-Project-teknologiaa ja 1000 epochin koulutusprosessia. Tarkoituksena oli tuottaa mahdollisimman luonnollisia ja tunteikkaita ääniä, jotka käyttäjät kokisivat inhimillisinä ja miellyttävinä.

Kuvassa viisi esitellään Elina, naisellinen AI-kertoja. Elinan ääni perustuu pehmeään artikulaatioon, rauhalliseen rytmiin ja empaattiseen sävyyn. Tämä tekee siitä erityisen sopivan lapsille suunnattuihin kirjoihin, satujen lukemiseen sekä tilanteisiin, joissa tarvitaan lempeää ja selkeää kerrontaa. Käyttäjäkyselyssä Elinan ääntä kuvattiin usein "kuin oikean lastenkirjan lukijan ääneksi"



Kuva 9: Visualisointi äänimallista Elina ja mallikohtainen koulutusdata-arvio

Kuvassa kuusi nähdään Joonas, miespuolinen AI-kertoja. Joonaksen ääni on syvempi, dynaamisempi ja emotionaalisesti ilmeikkäämpi kuin Elinan. Se soveltuu hyvin sekä rauhallisiin että jännittäviin tarinoihin. Käyttäjien mukaan Joonaksen äänessä on karismaa, ja se tuo uskottavuutta erityisesti tarinoihin, joissa on voimakkaita tunnetiloja tai jännitystä.



Kuva 10: Visualisointi äänimallista Joonas ja mallikohtainen koulutusdata-arvio

Molemmat mallit ovat reaaliaikaisesti vaihdettavissa käyttöliittymässä. Niiden taustalla oleva äänenmuunnosteknologia, RVC-Project, arvioitiin käyttäjäkyselyssä (ks. luku 4.3) selkeästi laadukkaimmaksi viidestä testatusta AI-ääniteknologiasta. Vaikka testissä ei arvioitu Elinaa ja Joonasta suoraan, niiden äänimallit pohjautuvat täsmälleen samaan teknologiaan, jota kuuntelijat pitivät kaikkein inhimillisimpänä, tunteikkaimpana ja hahmosopivimpana. Tämä antaa vahvan tuen sille, että Elinan ja Joonaksen kaltaiset yksilöllisesti koulutetut, RVC-pohjaiset kertojamallit ovat perustellusti valittuina sovellukseen.

Mallien suunnittelu toteutettiin erilaisten käyttötapojen mukaan: Elinan lempeä, lämmin ja helposti lähestyttävä ääni tukee erityisesti satujen ja lapsille suunnattujen kertomusten kerrontaa, kun taas Joonaksen syvempi ja dramaattisempi ääniprofiili sopii paremmin jännittäviin ja rytmisesti rikkaampiin tarinoihin. Näiden mallien yhdistäminen käyttäjävalintaan perustuvaan äänenvaihtotoimintoon rikastuttaa sovelluksen käyttökokemusta ja tarjoaa tarinankerrontaan enemmän ilmaisullista syvyyttä.

5.3.3 Äänimuunnosprosessi ja äänen käsittely

Kun Elina- ja Joonas-mallien koulutus RVC-Projectilla oli saatu päätökseen, ne vietiin inferenssivaiheeseen, jossa käyttäjän puhe tai käsikirjoitettu repliikki muunnettiin kohdeäänimallia vastaavaksi äänitiedostoksi. Äänenmuunnos toteutettiin RVC-WebUI-alustan avulla,

joka mahdollisti reaaliaikaisen äänen generoinnin syötedatasta. Tämän vaiheen tuloksena syntyi korkealaatuinen .wav-muotoinen tiedosto, jota voitiin käyttää suoraan sovelluksen ääniresurssina.

Inferenssin aikana äänen laatua säädeltiin useilla parametreilla. Sävelkorkeuden analysoinnissa käytettiin harvest-menetelmää, joka tuotti puheeseen luonnollisen prosodian ja säilytti inhimillisen rytmin. Lisäksi käytettiin äänilevityksen hallintaa varten asetusta index rate, joka määritettiin arvoon 0.66. Tämä asetus auttoi minimoimaan ennakkomallien vaikutuksen ja varmisti, että tuotettu ääni perustui ensisijaisesti koulutettuun ääni-identiteettiin. Inferenssivaiheessa hyödynnettiin myös puolitarkkaa (is_half) GPU-laskentaa, joka nopeutti prosessointia vaikuttamatta äänen laatuun.

Tiedostojen esikäsittely tehtiin FFmpeg-työkalulla, jolla äänitiedostot muunnettiin verkkoselaimille yhteensopivaan 16-bittiseen PCM .wav-muotoon, 44 100 Hz:n näytteenottotaajuudella. Äänien nimet vastasivat suoraan käsikirjoituksen repliikkejä ja hahmoja, mikä mahdollisti tarkasti hallitun tiedostorakenteen sovelluksen taustajärjestelmässä. Jokainen tiedosto tarkistettiin ennen palvelimelle vientiä laadun, pituuden ja teknisen yhteensopivuuden osalta.

Verkkosovelluksessa ääniä hallitaan siten, että käyttäjän painallus käyttöliittymässä lähettää palvelimelle tiedon toistettavasta repliikistä. Palvelin palauttaa vastaavan .wav-tiedoston, joka striimataan käyttäjän selaimelle ilman latausviivettä. Tarinatilassa jokainen hahmo on sidottu ennalta määriteltyyn äänimalliin, eikä käyttäjällä ole mahdollisuutta muuttaa hahmoääniä. Kertojatilassa sen sijaan käyttäjä voi vaihtaa kertojaa neljästä vaihtoehdosta, jolloin käyttöliittymän valinta ohjaa backendin hakemaan kyseiselle mallille kuuluvat äänitiedostot.

Äänien käytettävyyttä testattiin useilla selaimilla ja eri päätelaitteilla, jotta voitiin varmistua äänten moitteettomasta toistosta. Testauksessa kiinnitettiin erityistä huomiota siihen, ettei äänen aloitusviiveitä tai epäluonnollisia katkoksia esiintynyt. Lisäksi äänten alku- ja loppuleikkaukset tarkistettiin manuaalisesti, jotta repliikit toistuivat selkeästi ja luonnollisesti heti käyttäjän painalluksesta.

Yhteenvetona voidaan todeta, että äänenmuunnosprosessi toteutettiin teknisesti hallitusti ja optimoidusti, mikä mahdollisti korkealaatuisen äänen yhdistämisen selainpohjaiseen käyttöliittymään ilman havaittavaa viivettä tai laadun heikkenemistä. Prosessi tukee suoraan sovelluksen tavoitteita inhimillisen kuuloisen, monikäyttöisen ja saavutettavan ääni-integraation tarjoamiseksi.

5.4 Haasteet ja ratkaisut

Sovelluksen kehitystyön aikana kohdattiin useita haasteita, jotka liittyivät erityisesti äänenlaatuun, mallien koulutuksen hallintaan, sovelluksen suorituskykyyn sekä käyttöliittymän selkeyteen. Näihin haasteisiin vastaaminen oli keskeinen osa kehitysprosessia, ja niiden ratkaiseminen vaikutti suoraan sovelluksen lopulliseen toimivuuteen ja käyttäjäkokemuksen laatuun.

Merkittävin tekninen haaste liittyi RVC-Projectin mallien koulutuksen vaatimuksiin. Projektin alkuvaiheessa huomattiin, että alle 500 epochin koulutukset ja suppea puhemateriaali johtivat merkittäviin laatuongelmiin, kuten äänen vääristymiin, epäselvään artikulaatioon ja prosodian epätasapainoon. Näihin ongelmiin reagoitiin kasvattamalla koulutusdatan määrää 60 minuuttiin ja lisäämällä epochien määrä 1000:een. Samalla optimoitiin äänenmuunnoksen parametreja, kuten index_rate ja f0-method, jotta saavutettiin tasapainoinen, luonnollinen ja tunnistettava lopputulos ilman ylioppimista.

Toinen keskeinen haaste liittyi käyttöliittymän äänenvaihtomekanismiin. Alkuperäinen toteutus käytti teknisiä mallinimiä (esim. "model_v1", "model_v2"), jotka osoittautuivat epäselviksi testikäyttäjille. Tämän havaittuaan mallien nimet muutettiin kuvaileviksi (esim. "Elina – pehmeä naisääni", "Joonas – syvä miesääni"), mikä helpotti valintaa ja paransi ymmärrettävyyttä merkittävästi. Myös äänten toistaminen hahmojen puhekuplista hiottiin toimivammaksi lisäämällä painalluksiin selkeä vaste ja varmistamalla, ettei toisto aiheuttanut viiveitä.

Lisäksi suorituskyvyn osalta havaittiin ajoittaisia ongelmia erityisesti silloin, kun useita ääniä toistettiin peräkkäin nopealla tahdilla. Tämä ratkottiin optimoimalla selaimen välimuistia ja varmistamalla, että äänten striimaus palvelimelta toimi kevyesti ilman paikallista tallennusta. FFmpegin avulla suoritetut äänen leikkaukset ja normalisointi auttoivat myös vähentämään tarpeettomia tiedostokokoja ja toiston aloitusviiveitä.

Kokonaisuutena voidaan todeta, että sovellus kehittyi merkittävästi juuri haasteiden ratkaisemisen kautta. Jokainen ongelmakohta johti konkreettiseen parannukseen, kuten laadukkaampiin äänimalleihin, käytettävämpään käyttöliittymään tai teknisesti vakaampaan järjestelmään. Kehitysprosessin aikana opittiin erityisesti se, että laadukas äänimalli edellyttää paitsi teknistä osaamista myös yksityiskohtaista puhemateriaalin hallintaa, säätöparametrien ymmärrystä ja käyttäjäkeskeistä suunnittelua käyttöliittymässä.

5.5 Käyttöliittymä ja toiminnalliset ratkaisut

Sovelluksen käyttöliittymä suunniteltiin palvelemaan ensisijaisesti lukijaa, joka haluaa yhdistää visuaalisen tarinankerronnan ja tekoälypohjaisen ääninäyttelyn helposti lähestyttävällä tavalla. Käyttöliittymän tavoitteena oli ohjata käyttäjää intuitiivisesti vaiheesta toiseen, ilman että käyttö vaatisi teknistä osaamista tai kirjautumista. Kokonaisratkaisussa painotettiin saavutettavuutta, selkeyttä ja nopeaa vasteaikaa.

Käyttäjäpolku sovelluksessa

Vaihe 1: Äänivalinnan näkymä

Käyttäjä saapuu sovelluksen etusivulle, jossa hänelle esitellään vaihtoehdot kertojan äänistä. Käyttäjä voi valita esimerkiksi pehmeän naisäänen (Elina) tai matalan miesäänen (Joonas). Äänenvalintanäkymä on visuaalisesti pelkistetty ja jaettu kahteen päävaihtoehtoon, jolloin valinta tapahtuu helposti yhdellä painalluksella. Valittu ääni tallentuu selaimen tilaan ja vaikuttaa seuraavaan näkymään ladattaviin äänitiedostoihin.

Vaihe 2: Tarina-aukeaman lukutila

Äänivalinnan jälkeen käyttäjä siirtyy tarinankerrontanäkymään, jossa visuaalinen sarjakuva-aukeama avautuu. Sivulla näkyy useita hahmoja ja heidän repliikkinsä on esitetty puhekuplina. Jokainen puhekupla toimii interaktiivisena painikkeena: käyttäjä voi painaa sitä kuullakseen kyseisen hahmon vuorosanat tekoälyllä tuotettuna. Ääni toistuu välittömästi painalluksen jälkeen ilman latausviivettä.

Vaihe 3: Repliikkien eteneminen ja äänikokemus

Käyttäjä etenee tarinassa omaan tahtiinsa ja voi valita, missä järjestyksessä hän kuuntelee repliikit. Tämä antaa vapauden hallita lukukokemusta ja tekee tarinasta vuorovaikutteisen. Jos kertojatila on aktiivinen (kirjat ilman hahmovuorosanoja), kaikki repliikit tuotetaan valitun kertojan äänellä. Jos käytössä on tarinatila (hahmokohtaiset äänet), jokaisella hahmolla on oma ennalta määritetty AI-äänensä.

Vaihe 4: Äänen vaihtaminen

Kertojan ääntä voi halutessaan vaihtaa palaamalla aloitusnäkymään ja tekemällä uuden valinnan. Äänivalinta ei muuta visuaalista sisältöä, mutta se päivittää backendin kautta käytettävät äänitiedostot. Tämä rakenne mahdollistaa sen, että sama tarinasisältö voidaan kokea eri kertojien äänillä.

Vaihe 5: Tarinan lopetus ja mahdollinen palaute

Kun käyttäjä on kuunnellut kaikki repliikit, hän voi siirtyä seuraavaan tarinaan tai sulkea sovelluksen. Sovelluksen rakenne ei vaadi käyttäjältä rekisteröitymistä tai sovelluksen asentamista, mikä tukee matalan kynnyksen käyttöönottoa. Jatkokehityksessä voidaan lisätä mahdollisuus antaa palautetta äänistä tai arvioida kokemusta.

Toiminnallisten ratkaisujen perustelut

Käyttöliittymän suunnittelu noudatti saavutettavuuden ja selkeyden periaatteita. Kaikki toiminnot, kuten äänen toisto, äänenvaihto ja navigointi, on toteutettu yksinkertaisilla visuaalisilla ratkaisuilla. Äänen toiston nopeus ja selkeä vaste painalluksiin olivat keskeisiä saavutettavuuden mittareita, ja niitä optimoitiin kehitysvaiheessa useaan otteeseen. Lisäksi äänien kiinnittäminen visuaalisiin puhekupliin vahvisti kerronnan kokemuksellisuutta ja tuki multimodaalista oppimista sekä elämyksellistä lukemista.

5.5.1 AI-äänien esittely käyttöliittymässä ennen lukutilaa

Kuvassa 11 esitellään verkkosovelluksen AI-ääniesittelyosio, joka toimii johdatuksena käyttäjälle ennen siirtymistä varsinaiseen lukutilaan. Tässä näkymässä käyttäjälle esitellään kaksi RVC-teknologialla tuotettua kertojamallia: Elina ja Joonas. Molemmat hahmot on visualisoitu selkeästi, ja heidän vieressään oleva painike mahdollistaa ääninäytteen kuuntelun.

Tämä vaihe on olennainen osa käyttäjäpolkua, sillä sen avulla käyttäjä saa käsityksen kummankin äänen tyylistä ennen valinnan tekemistä. Elina on suunniteltu lempeäksi ja eläytyväksi naisääneksi, joka sopii hyvin satujen ja lastenkirjojen kerrontaan. Joonas puolestaan on matalampi, rauhallinen miesääni, jonka äänentaso mukautuu tarinan sävyyn.

Kuvassa 11 näkyvä rakenne on suunniteltu pedagogisesti ja visuaalisesti niin, että ääniä voi kokeilla ilman rekisteröitymistä tai monimutkaista käyttöä. Tämä madaltaa käyttökynnystä erityisesti uusille käyttäjille ja lapsille.

Tämän jälkeen käyttäjä voi siirtyä lukutilaan, jossa valittu kertojaääni toimii koko tarinan lukijana, tai vaihtoehtoisesti aktivoida Tarinatilan, jossa jokaisella hahmolla on yksilöllinen AI-ääni.

Kuva 11: Äänimallien esittely verkkosivustossa käyttäjälle (Kirjanurkka 2025)

5.5.2 Lukutila ja AI-äänen valinta

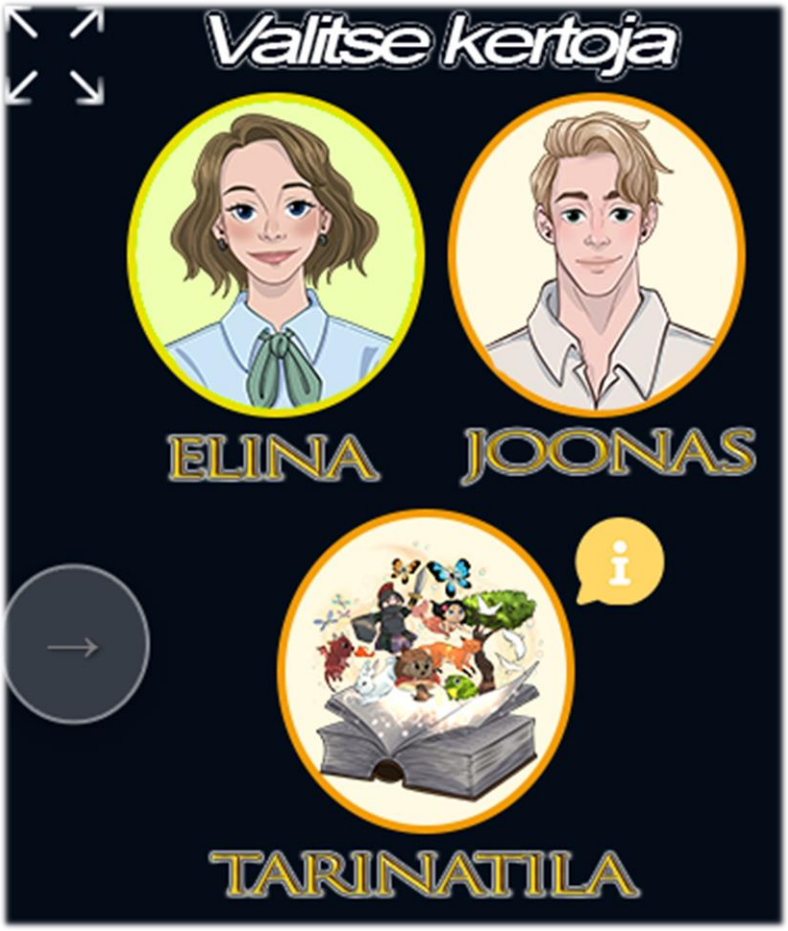
Tarina-aukeaman lukutila on käyttöliittymän ydinosa, jossa yhdistyvät visuaalinen sarjakuvakerronta ja tekoälyllä tuotettu ääni. Kuvassa 12 esitetään näkymä, jossa käyttäjä voi seurata sarjakuvatarinaa ja samanaikaisesti kuunnella hahmojen puhetta AI-ääninäyteltynä. Jokainen puhekupla toimii interaktiivisena painikkeena: käyttäjä voi klikata kuplaa kuullakseen vuorosanan tekoälyllä tuotettuna.

Erityisen tärkeä toiminnallisuus näkyy oikeassa laidassa, jossa käyttäjä voi valita käytettävän kertojan kolmesta vaihtoehdosta: Elina, Joonas ja Tarinatila. Elina ja Joonas ovat yleisääniä, jotka lukevat kaikki vuorosanat yhtenäisenä kerrontana. Tarinatila puolestaan aktivoi tilan, jossa jokaisella hahmolla on oma, RVC-teknologialla koulutettu ääni. Tämä ominaisuus tekee kokemuksesta erityisen immersioisen ja vastaa opinnäytetyön keskeistä tutkimuskysymystä: Miten tekoälypohjaista äänimuunnosta voidaan hyödyntää inhimillisen kuuloisen ääninäyttelyn luomisessa kuvallisiin verkkokirjoihin?

Käyttäjän tekemä valinta tallentuu istuntokohtaisesti, ja se määrittää, mitä äänitiedostoja backend toimittaa selaimelle. Tämä mahdollistaa saman sarjakuvasisällön kokemisen useilla eri tavoilla — joko yhden yhtenäisen kertojan äänen kautta tai hahmokohtaisesti räätälöidyillä äänillä.

Kuva 12: Lukutila ja kertojan valinta verkkosivustossa (Kirjanurkka 2025)

Käyttöliittymän keskeisin näkymä on lukutila, jossa käyttäjä tarkastelee sarjakuva-aukeamaa ja voi painaa puhekuplia kuullakseen vuorosanat AI-ääninä. Oikean reunan äänivalintapaneelissa käyttäjä voi valita kertojaksi joko Elinan tai Joonaksen, tai vaihtoehtoisesti aktivoida Tarinatilan, jossa jokaisella hahmolla on oma, RVC-mallilla tuotettu yksilöllinen AI-ääni.



Kuva 13: Kertoja valittu indikaattori lukutilassa (Kirjanurkka 2025)

Kuva 13 havainnollistaa tilaa, jossa kun käyttöliittymässä malli on valittu (esimerkiksi Elina), sen hahmoikonin taustaväri muuttuu vihreäksi käyttöliittymässä. Tämä tarjoaa visuaalisen vahvistuksen käyttäjälle ja auttaa ymmärtämään käyttäjää mikä äänimalli on valittuna. Toinen käyttöliittymän toiminnallinen yksityiskohta on mahdollisuus siirtyä koko näytön tilaan, jolloin lukukokemus laajenee koko selainikkunaan – erityisen hyödyllistä mobiilin- ja tabletin käytössä.

5.5.3 Käyttäjän toiminnan ja järjestelmävastauksen välinen ääniprosessi

Lukutila on sovelluksen ydinalue, jossa käyttäjä pääsee vuorovaikutteisesti kokemaan kuvallisen tarinan tekoälyn generoimilla äänillä rikastettuna. Kuvassa 14 esitetään näkymä sarjakuva-aukeamasta, jossa jokaisella hahmolla on puhekupla. Näitä puhekuplia klikkaamalla käyttäjä voi toistaa kyseisen hahmon vuorosanat tekoälyllä tuotettuna. Käyttöliittymä on suunniteltu siten, että puhekuplat ovat selkeitä ja helposti tunnistettavia, ja ääni käynnistyy välittömästi ilman latausviivettä.

Tämän ominaisuuden tarkoituksena on säilyttää lukemisen rytmi ja tukea tarinankerronnan immersiota. Käyttäjä voi itse määrittää, missä järjestyksessä kuuntelee vuorosanat, jolloin lukukokemus on aidosti henkilökohtainen. Tämä tukee erityisesti itsenäistä lukemista, pedagogista käyttöä ja esteettömyyttä.

Oikean reunan valintapaneelista käyttäjä voi valita, toistetaanko vuorosanat yhdellä kertojan äänellä (Elina tai Joonas) vai aktivoidaanko Tarinatila. Tarinatilassa jokaisella hahmolla on yksilöllinen AI-malli, joka on koulutettu vastaamaan kyseisen roolihahmon ääntä. Tämä ominaisuus tekee lukukokemuksesta ainutlaatuisen – esimerkiksi dialogikohtauksissa jokaisella puhujalla on tunnistettavasti oma äänensä. Käyttäjän valinta vaikuttaa taustajärjestelmään siten, että toistettava ääni valitaan dynaamisesti käyttöliittymävalinnan perusteella.

Äänentoisto lukutilassa perustuu juuri tähän vuorovaikutukseen käyttöliittymän ja palvelinpuolen (backend) välillä. Kun käyttäjä aktivoi puhekuplan, selain lähettää pyynnön valitun AI-mallin mukaisesta äänitiedostosta taustajärjestelmälle. Taustajärjestelmä palauttaa .wav-muotoisen äänen, joka striimataan selaimeen ja toistetaan lähes reaaliaikaisesti.

Kuva 14: AI äänimallien äänen toiminta Web sovelluksessa (Kirjanurkka 2025)

Toimintoketju muodostaa keskeisen osan sovelluksen teknistä arkkitehtuuria: käyttöliittymä, mallivalinta ja äänihallinta yhdistyvät saumattomasti. Äänentoiston nopeus ja laatu perustuvat valmiiksi tuotettujen .wav-tiedostojen käyttöön. Kaikki tiedostot on koulutettu ja tallennettu etukäteen palvelimelle, ja ne on järjestetty mallikohtaisesti joko tiedostorakenteisiin tai tietokantaan. Tämä ratkaisu eliminoi tarpeen reaaliaikaiselle äänen generoinnille, mikä parantaa vasteaikaa ja takaa tasalaatuisen kuuntelukokemuksen kaikille käyttäjille.

5.5.4 Äänenlaatu ja mallin koulutuksen haasteet

RVC-Projectin hyödyntäminen kertojamallien koulutuksessa mahdollisti inhimillisen kuuloisten AI-äänien luomisen, mutta äänenlaadun saavuttaminen vaati useita iteraatioita ja teknisiä säätöjä. Mallit koulutettiin noin 60 minuutin korkealaatuisella, puhtaaksi leikatulla puhemateriaalilla kummallekin kertojalle (Elina ja Joonas). Molemmat mallit koulutettiin 1000 epochin ajan, joka osoittautui optimaaliseksi määräksi korkean äänenlaadun saavuttamiseksi ilman ylikouluttamista.

Koulutusprosessin aikana ilmeni useita haasteita, joista keskeisimmät liittyivät koulutusaineiston laatuun, äänen artikulaation selkeyteen sekä mallien yleiseen stabiilisuuteen. Alhaisemmilla epoch-arvoilla (esim. 350–500) havaittiin selvästi heikompaa artikulaatiota, vaihtelevaa sävelkorkeutta ja epätasapainoista rytmiä. Tämä viittasi alikoulutukseen, jossa malli ei ollut vielä omaksunut puheelle tyypillisiä piirteitä riittävästi. Toisaalta yli 1500 epochin koulutuksilla havaittiin taipumusta ylikouluttamiseen: mallin puhe alkoi sisältää mekaanisesti toistuvia ilmaisutapoja sekä opittuja hengityselementtejä, jotka eivät enää tuntuneet luonnollisilta.

Lisäksi mallin suorituskyky oli herkkä valituille teknisille parametreille. Esimerkiksi f0 method -asetuksella (harvest vs. pm) oli vaikutusta sävelkorkeuden analyysiin ja äänen dynaamisuuteen. Kokeiluissa harvest-metodi tuotti luontevampaa intonaatiota. index_rate-parametria säätämällä pystyttiin vaikuttamaan siihen, kuinka paljon malli nojautui taustalla olevaan retrieval-pohjaiseen äänireferenssiin. Korkea index_rate vähensi sävyn vuotoa (timbre leakage), mutta liiallisena heikensi äänen luonnollisuutta.

Kaikki tuotetut äänet testattiin käytännössä sekä kehittäjän toimesta että kuuntelutestin (N=23) kautta. Käyttäjäpalautteen perusteella paras tasapaino saavutettiin juuri 60 minuutin puhemateriaalilla ja noin 1000 epochin koulutuksella, jossa ääni oli sekä luonteva että johdonmukainen. Tätä tukevat myös kuuntelutestauksen tulokset, joissa RVC-Projectin mallit arvioitiin inhimillisimmiksi ja korkealaatuisimmiksi vaihtoehdoiksi.

5.5.5 Oppimiskokemukset ja kehityksen eteneminen

Projektin edetessä keskeiseksi oppimiskokemukseksi nousi äänimallien koulutuksen laadun kriittinen vaikutus lopputuloksen uskottavuuteen. Alkuvaiheessa toteutetut kokeilut lyhyemmillä koulutusjaksoilla (alle 500 epochia) ja suppealla puheaineistolla tuottivat ääniä, jotka kärsivät epäselvästä artikulaatiosta, mekaanisesta rytmistä ja epäluonnollisesta intonaatiosta. Näiden haasteiden kautta opittiin, että pelkkä syväoppimismallin käyttäminen ei riitä: ratkaisevaa on se, miten hyvin koulutusprosessi suunnitellaan, millaista aineistoa

käytetään ja miten teknisiä parametreja optimoidaan, kuten sävelkorkeusanalyysin menetelmää (f0 method) ja index_rate-arvoa. Käytännön kokeilujen kautta muodostui ymmärrys siitä, että laadukkaan, inhimillisen kuuloisen AI-äänen luominen vaatii paitsi puhdasta ja monipuolista äänidataa myös iteratiivista säätöä ja mallin herkkyyden tunnistamista.

Toinen merkittävä oppimiskokemus liittyi palvelinarkkitehtuurin ja käyttöliittymän yhdistämiseen sulavaksi, teknisesti ehjäksi kokonaisuudeksi. Äänien striimaus selaimelle ilman viiveitä edellytti palvelinpuolen suorituskyvyn optimointia, välimuistin tehokasta hyödyntämistä sekä FFmpeg-työkalun käyttöä äänitiedostojen esikäsittelyssä. Näiden teknisten ratkaisujen kautta syntyi ymmärrys siitä, kuinka frontend- ja backend-järjestelmien välinen synkronointi vaikuttaa suoraan käyttäjän kokemaan vasteaikaan ja sovelluksen käytettävyyteen.

Projektin aikana opittiin myös tutkimusmenetelmien soveltamista käytännön kehitystyöhön. Kuuntelutestauksen toteutus ITU-T P.808 -standardin mukaisesti, sekä Likert-asteikollisten arviointien ja avoimien palautteiden analysointi, syvensivät ymmärrystä käyttäjäkeskeisestä kehittämisestä. Tämä toi esiin sen, että tekninen toimivuus ei yksin määritä onnistunutta ratkaisua – myös kuuntelijan kokema laatu, inhimillisyys ja tunnevaikutelmat ovat ratkaisevia tekijöitä.

Oppiminen ei rajoittunut pelkästään teknisiin ratkaisuihin, vaan kattoi laajemmin koko kehitystyön kaaren: ongelmanrajaus, teknologian valinta, äänten koulutus, käyttäjätestaus ja saavutettavan käyttöliittymän suunnittelu muodostivat toisiaan tukevan oppimisprosessin, jonka tuloksena syntyi korkeatasoinen, käytettävä ja pedagogisesti hyödyllinen AI-ääninäytelty lukukokemus verkkoympäristöön.

6 Johtopäätökset

6.1 Tulosten arviointi

Opinnäytetyön tavoitteena oli suunnitella ja toteuttaa kuvallisten verkkokirjojen yhteyteen AI-äänillä ääninäytelty ominaisuus, joka mahdollistaisi hahmojen ja kertojan puheiden esittämisen inhimillisesti ja laadukkaasti. Työn keskeinen tutkimuskysymys oli: Miten tekoälypohjaista äänimuunnosta voidaan hyödyntää inhimillisen kuuloisen ääninäyttelyn luomisessa kuvallisiin verkkokirjoihin?

Kuuntelutestausten tulokset osoittivat selkeästi, että RVC-Project (Retrieval-based Voice Conversion) erottui muiden testattujen AI-ääniteknologioiden joukosta laadukkaimpana ratkaisuna, erityisesti inhimillisyyden, tunteiden välittymisen ja hahmosopivuuden osa-alueilla. Toteutettu sovellus täytti sille asetetut tavoitteet: käyttäjä pystyi seuraamaan visuaalista tarinaa ja samanaikaisesti kuuntelemaan hahmojen puhetta, mikä paransi tarinallista elämyksellisyyttä ilman, että lukeminen jäi taka-alalle.

Toteutuksessa onnistuttiin yhdistämään Generative AI -teknologiat käytännön verkkosovellukseen ja osoittamaan niiden soveltuvuus monimedialliseen tarinankerrontaan. Käyttäjälle tarjottu mahdollisuus valita kertojan ääni lisäsi kokemuksen yksilöllisyyttä ja loi vuorovaikutteisen ulottuvuuden, joka tukee saavutettavaa ja elämyksellistä digitaalista lukukokemusta.

6.2 Kehittämistyön onnistuminen

Työn toteutus eteni suunnitellusti, ja tavoitteet saavutettiin kokonaisuudessaan. Toteutettu sovellus toimii vakaasti ja tarjoaa käyttäjälle suunnitellun toiminnallisuuden. RVC-Projectin äänenkäsittelyprosessi ja koulutus saatiin integroitua sovelluksen arkkitehtuuriin onnistuneesti. Kuuntelutestauksen kautta pystyttiin keräämään palautetta eri AI-ääniteknologioiden laadusta, mikä vahvisti toteutusvalinnan perustelut.

Kehitystyön aikana opittiin merkittävästi AI-äänimallien kouluttamisesta, äänen laadun optimoinnista sekä äänten hallinnan ja toiston teknisistä ratkaisuista verkkosovelluksessa. Suorituskyky ja äänenlaatu saavutettiin tasapainoisesti, mikä oli yksi keskeisistä onnistumisen mittareista.

6.3 Työn luotettavuus ja rajoitteet

Opinnäytetyön luotettavuutta tukee toteutettu kuuntelutestaus, jossa oli mukana 23 osallistujaa (N=23). Vaikka osallistujamäärä oli opinnäytetyölle riittävä ja testiasetelma vakioitu, on hyvä huomioida, että kyseessä oli subjektiivinen arviointimenetelmä, joka perustuu

käyttäjien henkilökohtaisiin kokemuksiin ja havaintoihin. Testin otanta ei edusta laajaa yleisöä, joten tuloksia ei voida yleistää kaikkiin käyttäjäryhmiin.

Äänimallien koulutuksen tulos riippuu suoraan käytettävän äänimateriaalin laadusta ja määrästä. Vaikka tässä työssä käytettiin laajaa äänidataa koulutukseen, on mahdollista, että eri puhemateriaalilla tai toisenlaisella koulutusprosessilla saavutettaisiin vielä parempia tuloksia.

Lisäksi toteutettu ratkaisu perustui valmiiksi luotuihin äänitiedostoihin, ei reaaliaikaiseen inference-palveluun, mikä asettaa rajoituksia sovelluksen dynaamisuudelle. Toteutettu arkkitehtuuri soveltuu erinomaisesti esirenderöityihin ääniin, mutta live voice conversion vaatisi erilaisen teknisen lähestymistavan.

6.4 Eettiset näkökulmat

Tekoälypohjaisen äänenmuunnoksen käyttöön liittyy eettisiä kysymyksiä, kuten mahdollinen identiteetin kloonaus tai äänen manipulointi. Tässä opinnäytetyössä ääniä koulutettiin vain itse tuotetusta puhemateriaalista, eikä käytetty kenenkään toisen henkilön ääntä tai identiteettiä. Sovelluksessa kaikki äänet olivat selkeästi merkitty tekoälyn tuottamiksi, mikä tukee EU:n AI-asetuksen mukaista läpinäkyvyyttä. Koska kyseessä on kehitys- ja kokeiluhanke, eettiset riskit arvioitiin vähäisiksi ja hallittaviksi.

6.5 Jatkokehitysideat ja tulevaisuuden tutkimusaiheet

Jatkokehityksessä sovellusta voisi laajentaa tukemaan reaaliaikaista äänenmuunnosta ilman esituotettuja tiedostoja, hyödyntäen esim. ONNX-malleja suorassa käytössä. Lisäksi hahmoäänien määrää ja tyylejä voisi kasvattaa, jolloin käyttäjä voisi mukauttaa myös hahmojen persoonallisuutta. Äänien tunnesävyjen dynaaminen vaihtelu sekä tekstistä-puheeksi (TTS) -moduulin liittäminen voisi mahdollistaa täysin automaattisesti tuotetut äänikirjat tai sarjakuvat. Myös saavutettavuusnäkökulmaa (esim. näkörajoitteiset) voitaisiin hyödyntää entistä syvällisemmin.

Yhtenä jatkokehitys ideana voisi olla Spotify alustalle tehty kanava, joka lukisi esimerkiksi, vaikka lastensatuja käyttäjille. Olemassa olevia kirjoja olisi helppo massa tuottaa koska AI-äänimallit ovat paljon halvempi vaihtoehto kuin perinteiset ääninäyttelijät. Reaaliaikaista äänimuunnosta voisi hyödyntää myös siten että esimerkiksi kaksi ääninäyttelijää tuottaisi yhteensä 20 erilaista ääntä lukemalla tarinoita niin että heidän äänensä muunnettaisiin AI mallien avustuksella tilanteisiin sopivaksi. Spotifyn virallisten ohjeiden mukaan on täysin sallittua ladata sisältöä minkä valmistuksessa on hyödynnetty AI-ääniä. Niin kauan, kunhan se ei riko tekijänoikeuksia. (Spotify 2025)

6.6 Ekonominen hyöty ja äänimallien kustannustehokkuus

Tekoälypohjaisten ääniresurssien hyödyntäminen sarjakuvien ja verkkotarinoiden ääninäyttelyssä tarjoaa merkittäviä kustannushyötyjä verrattuna perinteisiin tuotantomalleihin. Tämä tekee laajamittaisesta ääninäyttelystä erityisesti Indie-tuotannoille, koulutusympäristöille ja verkkopohjaisille sovelluksille taloudellisesti haastavaa. Toteutetussa järjestelmässä jokainen AI-malli on koulutettu yksilöllisesti, mutta sen käyttökustannus skaalautuu tehokkaasti: yksi ainoa ääninäyttelijä voi tuottaa kymmeniä tai jopa satoja erilaisia ääniprofiileja, mikäli koulutusdata ja mallin parametrit on optimoitu eri hahmotyyppeihin.

Tällainen lähestymistapa ei ole tarkoitettu korvaamaan ääninäyttelijöitä, vaan toimimaan työkaluna heidän käytössään. Äänenmuunnosteknologia voi tukea äänen variaatiota ja laajentaa näyttelijän ilmaisurekisteriä — erityisesti tilanteissa, joissa sama esiintyjä tuottaa useita erilaisia hahmoääniä. Tätä lähestymistapaa tukevat myös perinteisen ääninäyttelyn käytännöt, joissa korostetaan äänen laajaa emotionaalista skaalaa, hahmokohtaista muunneltavuutta ja roolien vaihtelevuutta (Toronto Film School, 2023). Vaikka alkuperäisessä lähteessä ei viitata tekoälyyn, siinä kuvataan selvästi, kuinka yksi näyttelijä voi hallita useita roolityyppejä improvisaation, äänenmuokkauksen ja hahmottamisen keinoin, mikä tekee siitä relevantin viitekehyksen myös AI-teknologian tukiroolin tarkasteluun.

Toronto Film School (2023) listaa useita nykypäivän ääninäyttelyn käyttökonteksteja, jotka hyötyvät AI-malleista tuotannollisesti ja taiteellisesti. Näitä ovat mm.:

- Videopeleihin tehtävät hahmoäänet
- Podcastien kerronta ja dramatisointi
- Mainokset, joissa ääni vaikuttaa brändin mieleenpainuvuuteen
- Äänikirjat, joissa ääni tulkitsee koko tarinan
- Animaatiot, joissa ääni määrittää hahmon persoonallisuuden
- Dubbaukset, jotka mahdollistavat kielimuurin ylittävän saavutettavuuden
- Koulutus- ja dokumenttisisällöt, joissa äänen selkeys ja rytmi ovat ratkaisevia

Tekoäly mahdollistaa näyttelijälle työkalun, jolla hän voi äänenmuunnosteknologian (*voice conversion*) avulla ääni roolittaa itsensä useisiin täysin erilaisiin hahmoihin — esimerkiksi matalaääniseen lohikäärmeeseen tai korkeaan teini-ikäiseen tyttöön — ilman että jokaiseen rooliin tarvitaan erillistä esiintyjää.

Lisäksi AI-mallit mahdollistavat tasalaatuisen ja uudelleenkäytettävän ääniresurssin, jonka avulla voidaan luoda johdonmukainen äänimaailma interaktiivisiin kirjoihin, peleihin tai verkkokerrontaan. Tämä vähentää uusintaäänitysten tarvetta ja nopeuttaa tuotantoaikatauluja, erityisesti projekteissa, joissa replikat ovat ennalta määritettyjä ja äänet toistuvat useissa yhteyksissä.

Voidaan todeta, että tekoälypohjaiset ääniresurssit eivät ainoastaan tuota kustannustehokkuutta, vaan ne myös laajentavat äänen taiteellista käyttöaluetta ja mahdollistavat uudenlaisen monipuolisuuden ääninäyttelyssä.

7 Yhteenveto

Tämän opinnäytetyön tavoitteena oli selvittää, kuinka reaaliaikaista AI-äänimuunnosta voidaan hyödyntää inhimillisen ja luonnollisen ääninäyttelyn luomisessa kuvallisiin verkkokirjoihin. Työssä suunniteltiin ja toteutettiin selainpohjainen sovellus, jossa käyttäjä lukee itse tarinaa samalla, kun hahmojen vuorosanat esitetään tekoälyn tuottamina ääninä. Toteutuksessa keskityttiin erityisesti RVC-Project-äänimuunnosteknologian hyödyntämiseen hahmo- ja kertojamallien koulutuksessa sekä äänten hallinnassa ja toistomekanismissa.

Opinnäytetyön alkuvaiheessa toteutettiin kuuntelutesti (N=23), jossa vertailtiin viittä eri AI-ääniteknologiaa: Google TTS, Amazon Polly, ElevenLabs, Voice.ai ja RVC-Project. Kuuntelutestin tulokset osoittivat, että RVC-Project tarjosi selkeästi parhaan kuuntelukokemuksen erityisesti inhimillisyyden, tunteiden välittymisen ja hahmosopivuuden osalta. Näiden tulosten perusteella päätettiin käyttää sovelluksen toteutuksessa yksinomaan RVC-Projectia.

Työssä kehitetty sovellus mahdollistaa kertojan äänen reaaliaikaisen vaihtamisen neljän koulutetun mallin välillä, kun taas hahmoäänet ovat lukittuja ja määritelty kehitysvaiheessa. Äänet tuotettiin laajalla puhemateriaalilla, minkä ansiosta saavutettiin laadukas ja luonnollinen lopputulos. Sovelluksen backend-arkkitehtuuri ja äänten striimausratkaisu tukivat sujuvaa ja viiveetöntä kuuntelukokemusta.

Työn aikana kohdattiin useita haasteita liittyen äänenlaatuun, koulutusmateriaalin määrään, mallien optimointiin sekä käyttöliittymän selkeyteen. Näihin haasteisiin löydettiin toimivat ratkaisut muun muassa koulutusdatan laajentamisella ja mallin asetusten optimoinnilla. Käyttöliittymän kehityksessä keskityttiin käyttäjäystävällisyyteen ja äänenvaihtotoiminnon selkeyteen.

Eettiset näkökulmat huomioitiin koulutusmateriaalin valinnassa ja käyttäjälle suunnatussa tiedottamisessa. Kaikki äänet tuotettiin itse luodusta puhemateriaalista, ja tekoälyn käyttö ääninäyttelyssä oli selkeästi ilmoitettu.

Työ osoittaa, että AI-äänimuunnosteknologioita voidaan hyödyntää onnistuneesti digitaalisten tarinakokemusten rikastamiseen. Tulevaisuudessa sovellusta voisi kehittää tukemaan dynaamisempaa tunnesävyjen vaihtelua, reaaliaikaista äänenmuunnosta ilman esituotettuja tiedostoja sekä laajempaa saavutettavuustukea. Työ avaa uusia mahdollisuuksia AI-teknologian käytölle luovassa sisällöntuotannossa ja tarinankerronnassa.

Lähteet

Bargum, A.R., Serafin, S. & Erkut, C. (2024). Reimagining speech: a scoping review of deep learning-based methods for non-parallel voice conversion. Frontiers in Signal Processing, 4, 1339159. Viitattu 12.12.2024. Saatavissa: https://www.frontiersin.org/journals/signal-processing/articles/10.3389/frsip.2024.1339159/full

Coursera. (2025). What Is an Epoch in Machine Learning? Viitattu 12.4.2025. Saatavissa: https://www.coursera.org/articles/epoch-in-machine-learning

Dosdoce & Frankfurter Buchmesse. (2024). Audiobook Global Growth Report. Viitattu 21.5.2025. Saatavissa: https://www.buchmesse.de/files/media/pdf/FBM_Dosdoce_Whitepaper_AUDIOBOOK_GLOBAL_GROWTH_2024.pdf

ElevenLabs. (2024). AI Voice Generator & Text to Speech – ElevenLabs. Viitattu 4.4.2025. Saatavilla: https://www.elevenlabs.io

ElevenLabs. (2024). How to make Text to Speech sound less robotic. Viitattu 20.4.2025. Saatavissa: https://elevenlabs.io/blog/how-to-make-text-to-speech-sound-less-robotic

Epoch. (2025). Parameter, compute and data trends in machine learning. Viitattu 21.5.2025. Saatavilla: https://epoch.ai/mlinputs/visualization

European Union. (2024). Artificial Intelligence Act: Article 50(4). Viitattu 10.5.2025. Saatavilla: https://artificialintelligenceact.eu/article/50/

Google Cloud. (2024). Custom Voice Documentation. Viitattu 21.5.2025. Saatavilla: https://cloud.google.com/text-to-speech/custom-voice

Giattino, C., Mathieu, E., Samborska, V. & Roser, M. (2025). Artificial Intelligence. Our World in Data. Saatavissa: https://ourworldindata.org/artificial-intelligence

ITU. (2021). Subjective evaluation of speech quality with a crowdsourcing approach. Viitattu 13.3.2025. Saatavissa: https://www.itu.int/rec/T-REC-P.808/en

Lemmetty, M. (1999). Review of Speech Synthesis. Master's thesis, Helsinki University of Technology. Viitattu 21.5.2025. Saatavilla: http://research.spa.aalto.fi/publications/theses/lemmetty_mst/chap2.html

Li, Y. A., Han, C., Raghavan, V. S., Mischler, G. & Mesgarani, N. (2024). Towards human-level prosody generation in text-to-speech. NeurIPS 2023. Viitattu 28.9.2024. Saatavissa: https://proceedings.neurips.cc/paper_files/paper/2023/file/3eaad2a0b62b5ed7a2e66c2188bb1449-Paper-Conference.pdf

Nwakanma, I., Oluigbo, I. & Okpala, I. (2014). Text-to-Speech Synthesis (TTS). International Journal of Research in Information Technology (IJRIT). Viitattu 16.10.2024. Saatavissa: https://www.researchgate.net/profile/Cosmas-Nwakanma/publication/319406874_Text_-_To_-_Speech_Synthesis_TTS/links/59a8abdbaca27202ed5f539c/Text-To-Speech-Synthesis-TTS.pdf

Our World in Data. (2025). Artificial intelligence training computation (speech models subset). Viitattu 18.3.2025. Saatavissa: https://ourworldindata.org/grapher/artificial-intelligence-training-computation

Our World in Data. (2025). Test scores of AI capabilities relative to human performance. Viitattu 18.3.2025. Saatavissa: https://ourworldindata.org/grapher/test-scores-ai-capabilities-relative-human-performance

Palkkavertailu.com. (2024). Näyttelijä palkka. Viitattu 21.5.2025. Saatavilla: https://palkkavertailu.com/palkka/näyttelijä

RVC-Project. (2025). Retrieval-based Voice Conversion WebUI GitHub-julkaisu. Viitattu 2.4.2025. Saatavissa: https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI/

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. Viitattu 28.9.2024. Saatavissa: https://arxiv.org/abs/1712.05884

Spotify. (2025). Spotifyn käyttöehdot. Viitattu 15.5.2025. Saatavissa: https://www.spotify.com/fi/legal/end-user-agreement/

Stewart, J. Q. (1922). An electrical analogue of the vocal organs. Nature, 110(2762), 311. Viitattu 13.3.2025. Saatavissa: https://www.nature.com/articles/110311a0

TechTarget. (2023). What is floating-point operations per second (FLOPS)? Viitattu 18.3.2025. Saatavissa: https://www.techtarget.com/whatis/definition/FLOPS-floating-point-operations-per-second

TensorFlow. (2024). Overfit and underfit. Viitattu 17.4.2025. Saatavissa: https://www.tensorflow.org/tutorials/keras/overfit_and_underfit

Toronto Film School. (2023). Voice Acting for Beginners: The Complete Guide. Viitattu 17.5.2025. Saatavissa: http://torontofilmschool.ca/blog/voice-acting/

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. DeepMind. Viitattu 12.2.2025. Saatavissa: https://arxiv.org/pdf/1712.05884

Voice.ai. (2024). Real-Time AI Voice Changer & Voice Generator. Viitattu 17.4.2025. Saatavissa: https://voice.ai

Walczyna, T. & Piotrowski, Z. (2023). Institute of Communication Systems, Military University of Technology. Viitattu 2.11.2024. Saatavissa: https://www.mdpi.com/2076-3417/13/5/3100

Wang, C., Li, Y., Li, X., Lu, Y. & Liu, X. (2023). A survey on deep learning-based voice conversion techniques. Applied Sciences, 13(5), 3100. Viitattu 1.3.2025. Saatavissa: https://doi.org/10.3390/app13053100

Kuuntelutestauksen kyselylomake (AI-äänien arviointi)

Tämä kysely on osa opinnäytetyötä "Kirjanurkka: Kuvallisten kirjojen AI-äänillä ääninäytellyn ominaisuuden tuottaminen ja toteutus verkkosivustoon" (Karjalainen 2025, LAB AMK).

Kysely on **täysin anonyymi**. Emme kerää henkilötietoja, emmekä tallenna vastaajien IP-osoitteita tai muita yksilöiviä tietoja. Taustatietokysymykset ovat vapaaehtoisia, eikä vastauksia voida yhdistää yksittäisiin henkilöihin. Kaikki kerätty tieto käsitellään luottamuksellisesti ja ainoastaan opinnäytetyöhön liittyvän analyysin tarkoituksiin LAB-ammattikorkeakoulun tutkimuseettisten ohjeiden mukaisesti.

Tavoitteena on arvioida eri tekoälypohjaisten ääniteknologioiden inhimillisyyttä ja soveltuvuutta fiktiivisten hahmojen ääninäyttelyyn sarjakuvamuotoisessa tarinassa.

Ohjeet vastaajalle:
Kuuntele jokainen tarina-aukeama viidellä eri AI-äänellä. Arvioi jokainen kuuntelukerta itsenäisesti alla olevien kriteerien perusteella. Teknologioiden nimet on piilotettu, jotta vältetään brändivaikutus (labeling effect).

- **Käytä mahdollisuuksien mukaan melua vaimentavia kuulokkeita.**
  Tämä parantaa äänenlaatua ja mahdollistaa vivahteiden (esim. hengitys, painotus, tunne) paremman erottelun.
- **Vältä taustahälyä ja keskeytyksiä testin aikana.**
  Suositeltavaa on suorittaa testi hiljaisessa ympäristössä, esimerkiksi omassa huoneessa tai työpisteessä.
- **Yhden ääniversion arviointiin suositellaan käyttämään noin 1–2 minuuttia.**
  Koko kysely kestää arviolta **10–15 minuuttia** riippuen vastaajan perehtymisestä ja vastausnopeudesta.
- Vastaa huolellisesti ja itsenäisesti jokaisen ääniversion jälkeen. Älä palaa takaisin aiempiin arvioihin.

Taustatiedot (vapaaehtoisia, ei yhdistetä yksilöivään tietoon):

- Sukupuoli: ☐ Mies ☐ Nainen ☐ Muu ☐ En halua sanoa
- Ikä: ____ vuotta
- Kuunteluväline: ☐ Kuulokkeet ☐ Kaiuttimet ☐ Mobiililaite
- Onko sinulla aiempaa kokemusta tekoälyäänistä tai puhesynteesistä?
  ☐ Kyllä ☐ Ei

LAB University of Applied Sciences

Liite 1: Kuuntelutestauksen kyselylomake (AI-äänien arviointi), sivu 1

Arviointikriteerit (Likert-asteikko 1–5):

(1 = Erittäin huono, 2 = Huono, 3 = Kohtalainen, 4 = Hyvä, 5 = Erinomainen)

Kategoria	Selite
Inhimillisyys	Kuulostaako ääni ihmisen tuottamalta? Tuntuuko puhe luonnolliselta?
Tunteiden välittyminen	Välittyvätkö tunteet (ilo, suru, yllättyneisyys) äänessä?
Luettavuus ja selkeys	Onko ääni teknisesti ymmärrettävä? Kuuluuko puhe selvästi?
Sopivuus hahmoille	Sopivatko äänet esitettyihin hahmoihin uskottavasti?
Kokonaisvaikutelma	Mikä on yleisvaikutelmasi äänen sopivuudesta tarinaan?

Arviointi (täytetään jokaiselle ääniversiolle erikseen)

Ääniversio: _____ (esim. Versio A, Versio B jne.)

Kategoria	Arvosana (1–5)
Inhimillisyys	____
Tunteiden välittyminen	____
Luettavuus ja selkeys	____
Sopivuus hahmoille	____
Kokonaisvaikutelma	____

Avoimet kommentit (vapaaehtoinen):

Kysely täytetään yhteensä viidestä eri ääniversiosta (Versiot A–E).
Kiitos osallistumisestasi tutkimukseen! Palautteesi auttaa kehittämään tekoälypohjaisia ääninäyttelyratkaisuja saavutettavammiksi ja inhimillisemmäksi.

LAB University of Applied Sciences

Liite 2: Kuuntelutestauksen kyselylomake (AI-äänien arviointi), sivu 2