Anja Virkkunen
DI, väitöskirjatutkija, Signaalinkäsittelyn ja akustiikan laitos, Aalto-yliopisto

Anssi Moisio
DI, väitöskirjatutkija, Signaalinkäsittelyn ja akustiikan laitos, Aalto-yliopisto

Tamás Grósz
PhD, tutkijatohtori, Signaalinkäsittelyn ja akustiikan laitos, Aalto-yliopisto

Mikko Kurimo
professori, Signaalinkäsittelyn ja akustiikan laitos, Aalto-yliopisto

MITÄ KONE KUULEE?
Puheen ja äänten tunnistus vanhoista kotimaisista elokuvista

Oletko joskus halunnut tutkia, kuinka paljon vanhoissa elokuvissa lauletaan tai soitetaan musiikkia? Tai sitä, miten paljon kiroilua eri aikakauden elokuvissa esiintyy? Ja sitten luovuttanut, koska tilastojen kerääminen käsin on liian työlästä ja aikaa vievää? Puheentunnistus ja ääntenhavainnointi ovat tekoälyn osa-alueita, jotka voisivat huomattavasti helpottaa tällaista elokuvahistorian tutkimusta. Kone siis etsisi ja tunnistaisi elokuvissa esiintyvät puheen ja äänet puolestasi.

Matkassa on kuitenkin yksi mutka. Koneoppimisalgoritmit toimivat parhaiten aineistolla, joka muistuttaa niiden opetukseen käytettyä aineistoa. Tekoälylle, joka on tottunut tunnistamaan 2000-luvulla nauhoitettua puhetta ja ääntä, vanhojen elokuvien puhe ja äänet ovat uutta ja outoa. Ajatellaan vaikkapa tässä artikkelissa esimerkkinä toimivia 1950-luvun fiktioelokuvia: niiden ääniraidat ovat huomattavan erilaisia nykypäivään verrattuna sekä sisällöllisesti että tekniseltä laadultaan. Sisällössä muutoksia on tapahtunut niin puheessa kuin ympärillä kuuluvissa äänissäkin. Esimerkiksi kielemme sanasto, puhetyylit ja murteiden käyttö ovat muuttuneet sitten 1950-luvun. Samoin esimerkiksi lankapuhelimet ovat vaihtuneet älypuhelimiin ja kirveet moottorisahoihin. Tekninen kehitys puolestaan on mahdollistanut monikanavaisen äänen ja korkeamman äänenlaadun.

Tässä artikkelissa käymme läpi, mitä puheentunnistus ja ääntenhavainnointi ovat ja millaisin koneoppimismenetelmin niitä mallinnetaan. Lisäksi käsittelemme viimeaikaisia kehitysaskeleita molempien osalta. Sen jälkeen esittelemme, kuinka puheentunnistuksen ja ääntenhavainnoinnin soveltaminen elokuvatutkimukseen onnistui Movie Making Finland eli MoMaF-projektissa. Projektissa tutkimme sitä, miten Suomen modernisoituminen näkyy 1950-luvun näytelmäelokuvissa mittaamalla esimerkiksi kaupunkeihin ja maaseutuun liittyvien avainsanojen ja äänten määriä puheentunnistus- ja ääntenhavainnointituloksissa.

MoMaF-projektissa saamamme tulokset osoittavat, että moderneilla tekoälymalleilla ääntenhavainnointi ja puheentunnistus elokuvista onnistuu kohtalaisesti. Mallit tekevät kuitenkin yhä virheitä edellä kuvatuista syistä johtuen, joten lopuksi pohdimme ratkaisuja virheiden vähentämiseksi.

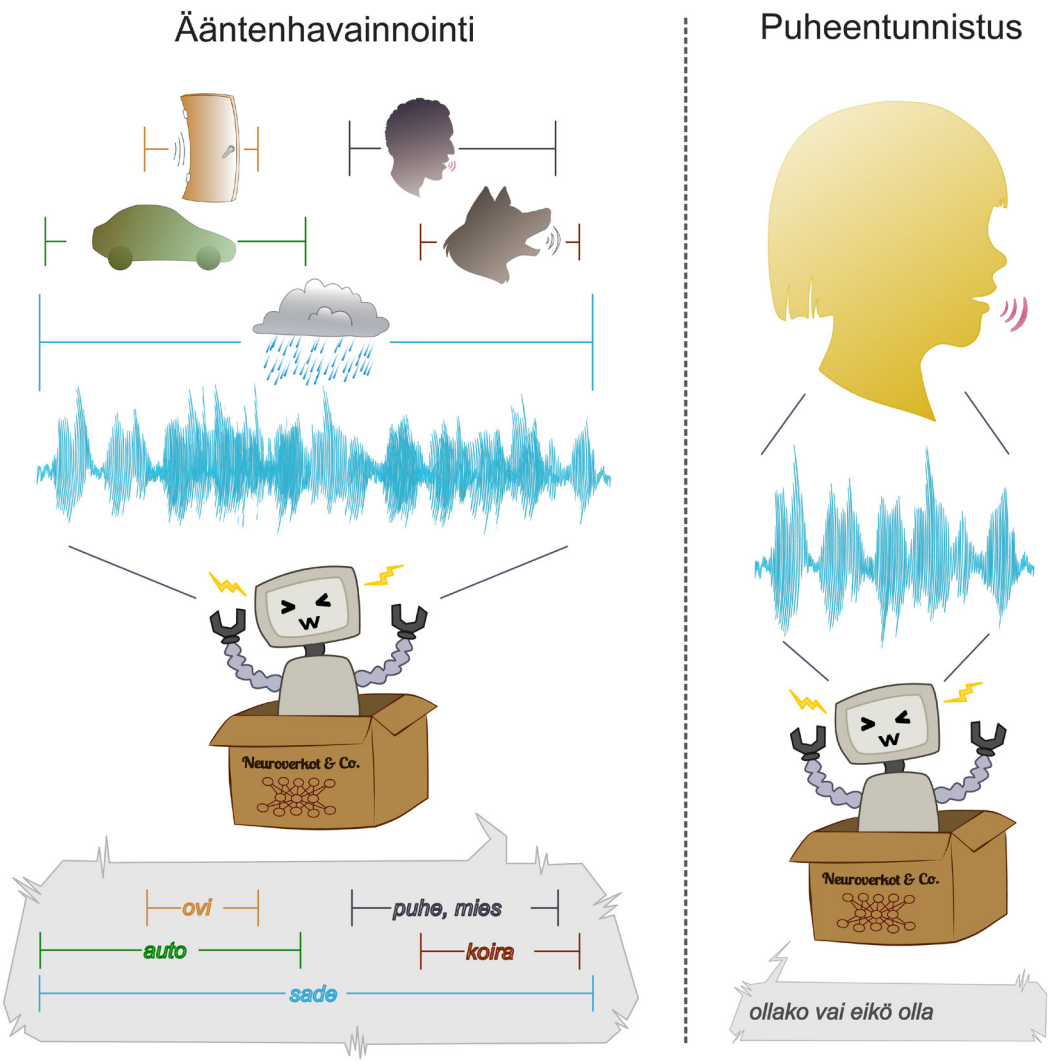
KATSAUKSET • Anja Virkkunen, Anssi Moisio, Tamás Grósz ja Mikko Kurimo: Mitä kone kuulee? Puheen ja äänten tunnistus vanhoista kotimaisista elokuvista, 57–69.

Mitä ovat ääntenhavainnointi ja puheentunnistus?

Ääntenhavainnoinnissa (*sound event detection*, *SED*) on tehtävänä havaita ääniraidassa esiintyvät äänet, kuten puheensorina tai rummunpärinä, sekä tunnistaa niiden ajoitus ja kesto. Lisähaastetta tulee siitä, että eri äänet voivat kuulua päällekkäin ja niiden kesto on erilainen, esimerkiksi sateenropina tai ukkosenjyrähdys. Puheentunnistuksessa (*automatic speech recognition*, *ASR*) tehtävänä on puolestaan muuttaa äänisignaalin sisältämä puhe sitä vastaavaksi tekstiksi, ilman välimerkkejä. Esimerkiksi suomen kielen puheentunnistimen tuloste sisältää pelkkiä pieniä kirjaimia ja välilyöntejä. Kuva 1 havainnollistaa tehtävien eri tavoitteet.

Ääntenhavainnoinnille ja automaattiselle puheentunnistukselle on yhteistä syötteenä toimiva äänisignaali, mutta muuten ne ovat koneoppimisongelmina hyvin erityyppisiä. Ensin mainitussa ääntenhavainnointimalli ennustaa kullekin tuntemalleen ääniluokalle, kuuluuko ääni sillä hetkellä signaalissa vai ei. Puheentunnistin taas tuottaa tekstiä eli kirjainten ja välilyöntien muodostamia merkkijonoja.

Eroista huolimatta molempia koneoppimisongelmia on mallinnettu samanlaisilla menetelmillä. Ensin vallalla olivat tilastollisiin malleihin lukeutuvat Gaussin



Kuva 1. Ääntenhavainnointi ja puheentunnistus ovat koneoppimisongelmina erilaisia.

mikstuurimallit (*Gaussian mixture models, GMM*) ja Markovin piilomallit (*hidden Markov models, HMM*), jotka sopivat hyvin puheentunnistuksen kaltaisen ongelman mallintamiseen. Puheentunnistuksessa näitä malleja käytetään yhdessä – GMM arvioi opetusaineistosta laskettujen tilastojen perusteella yksittäisten foneemien todennäköisyyksiä, kun taas HMM ennustaa saman aineiston pohjalta todennäköisiä foneemien yhdistelmiä tai jonoja. Esimerkiksi foneemijono ”/s/ /s/ /a/” on huomattavasti todennäköisempi kuin vaikkapa ”/g/ /s/ /g/”. Puheentunnistuksen vanavedessä näitä tekniikoita sovellettiin myös ääntenhavainnointiin, mutta äänten mahdollinen päällekkäisyys teki soveltamisesta vaikeaa (Mesaros et al. 2021). Puheentunnistuksessa ennustetaan kerrallaan vain yhtä foneemia, joten samaa ongelmaa ei ollut.

Ihmisen aivojen hermosolujen eli neuronien verkottuneesta rakenteesta inspiroituneet neuroverkkotekniikat ovat sittemmin vallanneet alaa joustavuutensa ja monipuolisuutensa avulla aluksi puheentunnistuksen puolella ja sen jälkeen myös ääntenhavainnoinnissa. Monipuolisuuden mahdollistaa se, että erilaisia neuroneja ja tapoja yhdistellä niitä on paljon – niin paljon, että mille tahansa matemaattiselle funktiolle on olemassa sitä approksimoiva neuroverkko.

Seuraavaksi käymme läpi sitä, mistä osista tyypilliset ääntenhavainnointi- ja puheentunnistusmallit koostuvat. Lisäksi sivuamme viimeaikaisia kehityskulkuja molempien alojen tutkimuksessa.

Ääntenhavainnointi

Ääntenhavainnoinnissa pyritään siis tunnistamaan, mitkä äänet ovat kuultavissa milläkin hetkellä. Vaikean tehtävästä tekee se, että mahdollisia äänilähteitä on lukemattomia ja ne voidaan luokitella eri tarkkuuksilla, vaikkapa eläimen äänenä tai hevosen hirnahduksena. Tästä ja yleisesti hyväksytyn ontologian puutteesta seuraa, että käytännön sovelluksissa käyttökohde ja saatavilla oleva data rajaavat havainnointimallin kykyä tunnistaa erilaisia ääniä (Mesaros et al. 2021). Dataa on kuitenkin pääsääntöisesti tarjolla vain vähän, mistä on neuroverkkojen vakiinnutettua asemansa ensisijaisena mallityyppinä tullut ongelma – neuroverkot kun tarvitsevat toimiakseen yleensä suuria määriä aineistoa. Ääntenhavainnoinnin tutkimus onkin alkanut yhä enemmän keskittyä menetelmiin, joilla uutta dataa voidaan kerätä tehokkaasti tai neuroverkkojen oppimista olemassa olevilla aineistoilla parantaa. Käydään ensin kuitenkin läpi hieman sitä, miten mallinnus neuroverkoilla käytännössä tapahtuu.

Itse ääntenhavainnointijärjestelmän voidaan nykyisellään ajatella koostuvan kolmesta osasta: esikäsittelystä, neuroverkkomallista ja jälkikäsittelystä. Esikäsittelyssä on kyse *piirreirrotuksesta* eli siitä, kuinka äänisignaalista saadaan poistettua ylimääräistä informaatioita ja samalla tiivistettyä sitä pienempään tilaan. Tähän on pitkään käytetty erilaisia signaalinkäsittelyn menetelmiä, joilla pyritään toisintamaan ihmiskorvan toimintaa. Ihmisen kuulo toimii tarkimmin 200–4 000 hertsin taajuuksilla, joten piirreirroituksessa pyritään vahvistamaan näitä taajuuksia ja vähentämään sitä korkeampien taajuuksien vaikutuksia. Tiivistämistä tarvitaan siksi, että koneoppimissovelluksissa käytettävät äänisignaalit on tallennettu 16 000 hertsin näytteenottotaajuudella, eli yhden sekunnin äänitiedosto tietokoneella muodostuu 16 000 peräkkäisestä luvusta. Laskentakapasiteettia säästyy, kun piirreirroituksella signaali tiivistetään vaikkapa kymmenesosaan alkuperäisestä. Olennaisen informaation poimiminen valmiiksi helpottaa myös koneoppimismallin tehtävää, koska muuten mallin pitäisi oppia tunnistamaan olennainen informaatio itse.

Havainnointimalli koostuu nykyisellä neuroverkkojen aikakaudella yhdestä neuroverkosta, jossa yhdistellään erilaisia neuroverkkokerroksia. Viimeisessä ker-

roksessa jokaiselle ennustettavalle luokalle on oma neuroninsa, joka antaa todennäköisyyden sille, onko luokka juuri kyseisellä hetkellä kuultavissa ääniraidalla. Jälkikäsittelyssä mallin tuottamista ennusteista siivotaan vielä pois epätodennäköisiä tapahtumia, joita mallin epätarkkuus aiheuttaa. Asiantuntija voi määritellä jokaiselle mallin tunnistamalle äänelle tunnuslukuja, kuten keskimääräisen keston ja esiintymistiheyden. Vaihtoehtoisesti nämä voidaan laskea myös suoraan opetusaineistosta. Jälkikäsittelyssä sitten tarkistetaan, poikkeavatko mallin havaitsemat äänet näistä tunnusluvuista. Esimerkiksi sade ei yhtäkkiä taukoa 400 ms:n ajaksi, joten jälkikäsittelyssä tauon erottamat sadeäänet yhdistetään yhdeksi pidemmäksi sadeääneksi.

Uuden aineiston keräämiseen manuaalisesti ihmisvoimin on kaksi lähestymistapaa. Asiantuntijatyönä teetetyt tarkat merkinnät ovat hitaita ja kalliita tuottaa, joten tämä lähestymistapa mahdollistaa vain joidenkin kymmenien tuntien kokoiset aineistot. Toinen vaihtoehto on hyödyntää joukkoistamista, mutta tällä tavoin tuotetun datan laatu vaihtelee paljon. Esimerkiksi suurimman tunnetun äänten-havainnointiaineiston AudioSetin (Gemmeke et al. 2017) on osoitettu sisältävän huomattavan määrän virheellisiä merkintöjä. AudioSet on Youtubesta kerättyjen 10 sekunnin videoklippien aineisto, jonka merkinnät on hankittu joukkoistamalla. Aineisto sisältää yli 500 eri ääniluokkaa ja noin 5 000 tuntia ääntä. Lisäksi AudioSetissä on niin sanotusti ”heikot” merkinnät (*weak labels*), eli kunkin äänitteen kohdalla kerrotaan vain, mitä ääniä se sisältää äänien tarkkojen alku- ja loppuaikojen sijaan. Neuroverkot pystyvät oppimaan kelvollisen mallin tällaisesta heikompilaatuisestakin datasta, mutta tarkkuuskriittisissä sovelluksissa on parempi käyttää tarkkoja merkintöjä. Uuden datan keräämisen lisäksi aineistoaan voi kasvattaa luomalla signaalinkäsittelyllä keinotekoista dataa jo olemassa olevasta aineistosta. Signaalinkäsittelyn tekniikat mahdollistavat muun muassa näytteiden keston nopeuttamisen ja hidastamisen, äänen taajuuden muuttamisen, näytteiden yhdistelyn ja taustakohinan lisäämisen. Näin mallia voidaan totuttaa tilanteisiin, joita alkuperäisessä aineistossa ei ole. Aineistossa voi esimerkiksi olla erilliset näytteet koiran haukusta ja sateesta mutta ei haukusta ulkona sateessa. Puuttuva näyte saadaan yhdistämällä sade- ja haukkunäytteet. Toinen esimerkki: aineistossa on näytteitä tavallisesta puheesta, mutta nopeuttamalla tai hidastamalla näitä näytteitä saadaan myös nopeaa ja hidasta puhetta.[1]

Toinen lähestymistapa datan rajallisuuden ongelmaan on tehostaa neuroverkkojen oppimista. Opinsiirrossa (*transfer learning*) suuri malli oppii ensin harjoitusongelman ja suuren aineiston avulla tunnistamaan syötteestä olennaisia piirteitä, minkä jälkeen malli opetetaan erikseen erikoistumaan pienemmän aineiston tehtävään (Cramer et al. 2019). Esimerkiksi Cramer et al. (2019) opettavat mallin ensin tunnistamaan, ovatko vai eivätkö yksittäinen videoruutu ja sekunnin pätkä ääntä peräisin samasta videoklipistä. Vasta sitten, kun malli on oppinut tämän tehtävän, vaihdetaan tehtäväksi äänten-havainnoinnin oppiminen. Opettaja–oppilas-menetelmissä (*student-teacher methods*) oppilasmallia opetetaan oikeiden merkintöjen sijaan opettajamallin tuottamilla ennusteilla (Lin et al. 2020). Ajatuksena on opettaa oppilasmallille epävarmuutta. Sen sijaan, että mallille kerrotaan ”tässä on hälytyssireenin ääni”, opettajamalli kertoo, että ”tässä on 88 % todennäköisyydellä hälytyssireenin ääni”. Tämä mahdollistaa merkitsemättömien aineistojen käytön oppilasmallin opetukseen, ja epävarmuuden opettaminen vähentää riskiä mallin ylisovitukselle eli sille, että malli toimisi vain opetusaineistollaan. Aktiivisessa oppimisessa (*active learning*) malli pyytää ihmistä merkitsemään mallin oppimista eniten auttavat näytteet, jolloin jo hyvin pienellä määrällä merkittyä aineistoa saadaan hyviä malleja (Shuyang et al. 2020).

1 Äänitteen nopeutta voi samalla tavalla säätää esimerkiksi podcast-sovelluksissa.

Puheentunnistus

Puheentunnistuksessa lähtökohtana on puhetta sisältävä äänite. Mikäli äänitteellä on pitkiä taukoja ilman puhetta, kuten elokuvan ääniraidalla tyypillisesti on, ne kannattaa ensin leikata pois äänitteestä. Puheäänen havainnointi (*voice activity detection*, *VAD*) tarkoittaa puhetta sisältävien kohtien erottamista äänitteestä. Tähän voidaan hyödyntää esimerkiksi edellä kuvattua tapaa poimia äänitteestä vain kohdat, jotka ääntenhavainnointijärjestelmä on luokitellut 'puhe'-luokkaan.

Kuten ääntenhavainnoinnissa edellä, myös puheäänen havainnoinnissa syntyvistä puheenpätkistä pyritään *piirreirroituksella* poistamaan epäolennainen tieto ennen puheentunnistusjärjestelmään syöttämistä. Puheentunnistusjärjestelmä ottaa parametrinä puheesta irrotetut piirteet, joiden perusteella se tulostaa transkriptiot. Perinteisesti puheentunnistusjärjestelmä koostuu kolmesta erikseen opetetusta mallista: 1) äännemallista, joka muuntaa piirrejonot äänteiksi, 2) ääntämissanakirjasta, joka kääntää äänteiden yhdistelmät sanoiksi, sekä 3) kielimallista, joka määrittää, kuinka todennäköisiä eri sanajonot (esimerkiksi lauseet) ovat itsessään.

Äännemalli on tyypillisesti Markovin piilomallin (HMM) ja neuroverkon (*deep neural network*, *DNN*) yhdistelmä eli niin sanottu HMM-DNN-malli, joka opetetaan käsin litteroidulla puheaineistolla. Hinton et al. (2012) tarjoaa tarkemman katsauksen tähän menetelmään. Malli saa syötteenä puheesta erotetut piirteet ja oppii muuntamaan ne äänteiden jonoksi.

Ääntämissanakirjan avulla äännemallin tunnistamat äänteet muutetaan kirjaimiksi ja sanoiksi. Ääntämissanakirjan toteuttaminen suomen kielelle on suoraviivaista verrattuna moniin muihin kieliin, kuten englantiin, koska suomessa yksi kirjain vastaa tavallisesti yhtä äännettä. Siksi suomenkielisen sanakirjan voi toteuttaa yksinkertaisesti muuttamalla jokaisen kirjaimen vastaavaksi äänteeksi. Esimerkiksi sana "äänne" kääntyy äänneketjuksi "/æ/ /æ/ /n/ /n/ /e/". Toisaalta kaikkien mahdollisten suomen sanamuotojen luetteleminen ääntämissanakirjaksi olisi hankalaa, koska sanamuotojen määrä on taivutusten ja yhdyssanojen vuoksi valtava. Siksi äännesanakirjaan listataan yleensä vain sanamuodot, jotka ovat kielimallin opetusaineistossa.

Kielimalli opetetaan laajoilla tekstiaineistoilla joko tilastollisesti tai neuroverkkoalgoritmein (Bengio et al. 2000). Opetukseen käytettävät tekstiaineistot voidaan kerätä esimerkiksi lukuisilta tekstiä sisältäviltä verkkosivustoilta. Tavallisin tilastollinen kielimallityyppi on n-gram-kielimalli, jossa "n-gram" tarkoittaa n:ää peräkkäistä sanaa. Esimerkiksi 3-gram-kielimalli muodostetaan laskemalla, kuinka usein mikäkin sana seuraa mitäkin kahta aiempaa sanaa, ja muuttamalla saadut tilastot todennäköisyyksiksi. Käytettyjen sanaketjujen pituus vaihtelee tyypillisesti kahdesta viiteen. Neuroverkkomenetelmissä puolestaan opetetaan neuroverkkoa ennustamaan seuraava sana aiempien sanojen perusteella. Kun neuroverkko on käynyt läpi miljoonia tai miljardeja sanoja tekstiä yrittäen aina ennustaa seuraavan sanan ja hioa arvauksia onnistumisten ja epäonnistumisten perusteella, se tuottaa luotettavia todennäköisyyksiä eri lauseille. Puheentunnistuksessa kielimallin tuottamat todennäköisyydet auttavat karsimaan äännemallin sanayhdistelmistä ne vaihtoehdot, joita kielessä ei yleensä esiinny. Esimerkiksi vaihtoehdoista a) "kisa naukuu" ja b) "kissa naukuu" kielimalli antaisi b:lle suuremman todennäköisyyden, koska se on nähnyt tämän sanayhdistelmän useammin opetustekstiaineistossa. Sanojen sijasta lauseiden todennäköisyydet voidaan yhtä hyvin laskea myös pilkkomalla sanat ensin pienempiin osiin, mikä onkin järkevää esimerkiksi suomessa erilaisten sanamuotojen suuren määrän vuoksi. Esimerkiksi sana "esimerkiksi" voidaan jakaa osiin "esi", "merki" ja "ksi". Koska näitä osia käytetään myös muissa sanoissa (esim. "*esi*kuva", "*merki*tä", ja "puhee*ksi*"), sanan osien kokonaismäärä on pienempi kuin kokonaisten sanojen määrä.

Tälle perinteiselle kolmen mallin järjestelmälle on erityisesti 2010-luvulla kehitetty korvaavia menetelmiä, joissa erilliset mallit korvataan yhdellä neuroverkolla. Verkko opetetaan tuottamaan tekstiä suoraan piirteistä eli äänitteestä erotetusta numeerisesta representaatiosta ilman välivaiheita (esim. Graves et al. 2006; Graves 2012; Chorowski et al. 2014; Chan et al. 2016). Tällainen yhtenäinen (*end-to-end*) malli on yksinkertaisempi eikä vaadi eri osien yhteensovittamista. Kääntöpuolena tämä lähestymistapa vaatii huomattavasti enemmän opetusaineistoa, koska mallin pitää oppia epäsuorasti mallintamaan äänteitä ja kielen rakenteita sen sijaan, että puheentunnistus pilkottaisiin osatehtäviksi, joiden opettaminen erikseen on helpompaa. Litteroitua puheaineistoa ei myöskään läheskään aina ole saatavilla tarpeeksi (eli mieluiten vähintään satoja tunteja) yhtenäismallin opettamiseksi.

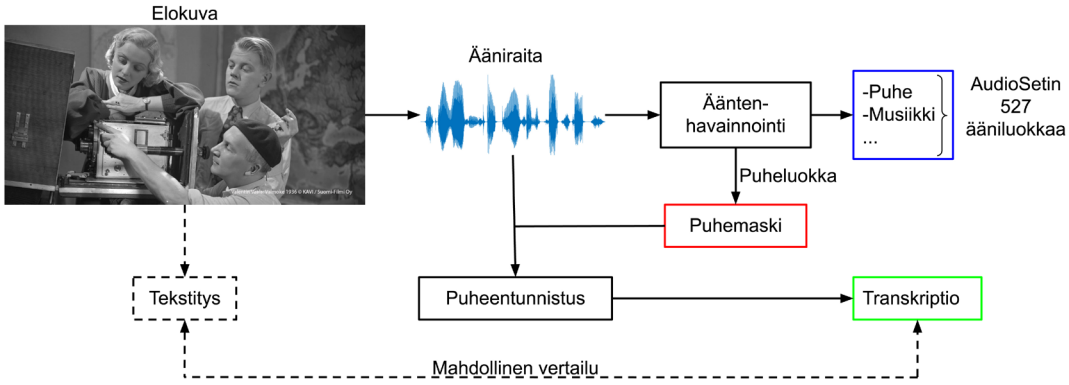
Viimeisimmät edistysaskeleet puheentunnistuksessa ovat perustuneet itseohjautuvaan oppimiseen (Schneider et al. 2019) eli samaan menetelmään, jota on käytetty neuroverkkokielimallien opettamiseen. Samaan tapaan kuin kielimalli käy läpi tekstiä, itseohjautuva äännemalli käy läpi puheääniteaineistoa yrittäen joka hetki ennustaa, miten äänite jatkuu. Tämän menetelmän etu on se, ettei puhetta tarvitse litteroida. Kun malli on oppinut taitavaksi tässä tehtävässä, se jatko-opetetaan tunnistamaan puhetta litteroidulla puheaineistolla. Jatko-opetuksessa malli hyödyntää aiemmassa itseohjautuvassa opetuksessa oppimiansa representaatioita puheesta. Menetelmän itseohjautuva vaihe vaatii vielä enemmän puheaineistoa kuin yhtenäiset mallit. Ja vaikka aineistoa ei tarvitsekaan litteroida käsin, niin yhden tällaisen mallin opettaminen voi vaatia kymmenien tuhansien eurojen verran laskentaresursseja. Itseohjautuvia malleja on kuitenkin julkisesti saatavilla, sillä jotkin yritykset, kuten Facebook (Baevski et al. 2020), ovat opettaneet ja julkaisseet omia mallejaan. Näitä malleja kuka tahansa voi jatko-opettaa omalla pienelläkin litteroidulla aineistollaan ja käyttää puheentunnistukseen.

Data ja käytetyt menetelmät

Kaikista 1950-luvulla ensi-iltaan tulleista 208 näytelmäelokuvasta meillä oli analyysiamme varten käytössä 140 elokuvaa. Osa elokuvista (22) karsiutui pois tutkimusluvan puutteen takia ja osan (46) jätimme itse pois, koska elokuvat sijoittuivat tekoajan sijaan selvästi menneisyyteen. Analysoimistamme elokuvista 44:lle oli olemassa myös tekstitykset, joita hyödynsimme mallien testauksessa. Tämän lisäksi poimimme 50:stä satunnaisesti valitusta elokuvasta muutaman minuutin pätkiä, joihin merkitsimme niissä kuultavat äänet ja puheen. Äänten merkinnässä käytimme samoja ääniluokkia kuin AudioSetissä. Yhteensä pätkiä oli noin sadan minuutin edestä. Tällä testiaineistolla pystyimme vertailemaan eri mallien toimivuutta 1950-luvun näytelmäelokuviin.

Kuva 2 havainnollistaa prosessin, jolla kaikki 140 elokuvaa käsiteltiin. Ensin jokaisen elokuvan ääniraita ajettiin ääntenhavainnointimallin läpi, minkä jälkeen puhetta sisältävät kohdat annettiin puheentunnistimelle. Ääntenhavainnointimalliksi valitsimme alustavien kokeiden jälkeen esiopetetun audioneuroverkon (*pre-trained audio neural network*, *PANN*), joka on rakennettu 14 konvoluutiokerroksesta[2] (Kong et al. 2020). Malli on opetettu AudioSet-korpuksen (Gemmeke et al. 2017) avulla tunnistamaan 527 eri ääniluokkaa. Valintaa puolsi tunnistettavien äänien monipuolisuus

2 Konvoluutiokerros on yksi niistä neuroverkkokerrostyypeistä, joista neuroverkkoja rakennetaan. Se pohjautuu nimensä mukaisesti matemaattiseen konvoluutio-operaatioon.



Kuva 2. Prosessi, jolla käsittelimme kaikki tutkimukseen valikoituneet elokuvat. Lähde: Grósz et al. (2022).

ja hyvät testitulokset useilla eri testiaineistoilla.[3] Lisäksi varmistimme vielä omalla testiaineistollamme, että malli selviytyy kohtuullisesti myös tutkimistamme elokuvista.

Puheentunnistimeksi valitsimme tutkimusryhmämme itse kehittämän HMM-DNN-äännemallin ja 4-gram-kielimallin. Kokeilimme myös yhtenäisiä ja itseohjautuvia puheentunnistusmalleja, mutta ne eivät tuottaneet aineistollamme yhtä hyviä tuloksia kuin HMM-DNN-malli. Käyttämämme äännemalli oli opetettu noin 1600 tunnilla Lahjoita puhetta -kampanjassa[4] kerättyä suomenkielistä puhetta (Moisio et al. 2022). Kielimallin opetukseen käytimme kolmea eri aineistoa: 1) edellä mainitun puheaineiston transkriptioita, 2) internetistä kerättyä tekstiaineistoa (Enarvi 2018) ja 3) OpenSubtitles-tekstitysaineistoa (Lison & Tiedemann 2016). Näistä viimeisin auttaa mukauttamaan kielimallin lähemmäksi elokuvissa käytettyä kieltä, vaikka tämän aineiston elokuvien ja TV-sarjojen tyylilaji ei vastaakaan 1950-luvun elokuvien aineistoa.

Äänten havainnoinnin ja puheentunnistuksen tarkkuudesta

Kuinka hyvin mallit sitten lopulta toimivat 1950-luvun elokuvilla? Kuvassa 3 on annettu esimerkit mallien onnistumisista ja epäonnistumisista. Niistä näkee, että tunnistus onnistuu vaihtelevasti. Vasemmalla puheentunnistus (punainen teksti) on onnistunut virheettömästi, vaikka kyse on laulusta. Oikealla puheentunnistus taas on epäonnistunut. Valkoinen teksti on alkuperäinen elokuvatekstitys ja vihreä teksti kertoo, mitä muita ääniä kone kuulee tässä kohdassa elokuvaa. Mutta kuinka paljon mallien tarkkuus elokuva-aineistolla poikkeaa modernimmista aineistoista? Entä miten onnistui urbaanien ja maaseudun sanojen ja äänien tunnistus elokuvista? Urbaaneihin ääniin ja sanoihin lukeutuivat ajoneuvot ja laitteet (esim. juna, auto ja puhelin), kun taas maaseudun puolelle kuuluivat kotieläimet ja veneet (esim. hevonen, sika, soutuvene ja kajakki).

3 Testiaineistoina oli mm. AudioSetin oma testiaineisto ja vuosien 2018 ja 2019 DCASE-kilpailujen testiaineistot (ks. https://dcase.community).

4 Lisätietoa kampanjasta: https://www.kielipankki.fi/lahjoita-puhetta/

Kuva 3. Kaksi esimerkkiä ääntenhavainnoinnin ja puheentunnistuksen tuloksista elokuvasta *Hei, rillumarei!* (1954).

Ääntenhavainnointi

Ensimmäiseen kysymykseen vastaamisen voimme aloittaa tarkastelemalla kuinka hyvin valitsemamme malli toimii moderneilla aineistolla. Testeissään Kong et al. (2020) osoittavat, että elokuvissa tyypillisille äänille, kuten puheelle ja musiikille, tunnistustarkkuus on 85 % tasolla. Muilla yleisillä äänillä keskimääräinen tarkkuus (*average precision*, *AP*) vaihtelee 20 % ja 60 % välillä muutamia poikkeuksia lukuun ottamatta. Tutkimuksemme kannalta tärkeiden luokkien lähempi tarkastelu osoittaa, että junat havaitaan keskimäärin 70 %:ssa, ajoneuvot 50 %:ssa, autot 40 %:ssa ja hevoset 50 %:ssa tapauksista.

Mallin toimivuutta 1950-luvun elokuvilla mittasimme kahdella tavalla: 1) vertaamalla puhehavaintojen kestoa ja ajoitusta tekstityksen vastaaviin aikoihin ja 2) testaamalla kokoamallamme pienellä testiaineistolla.

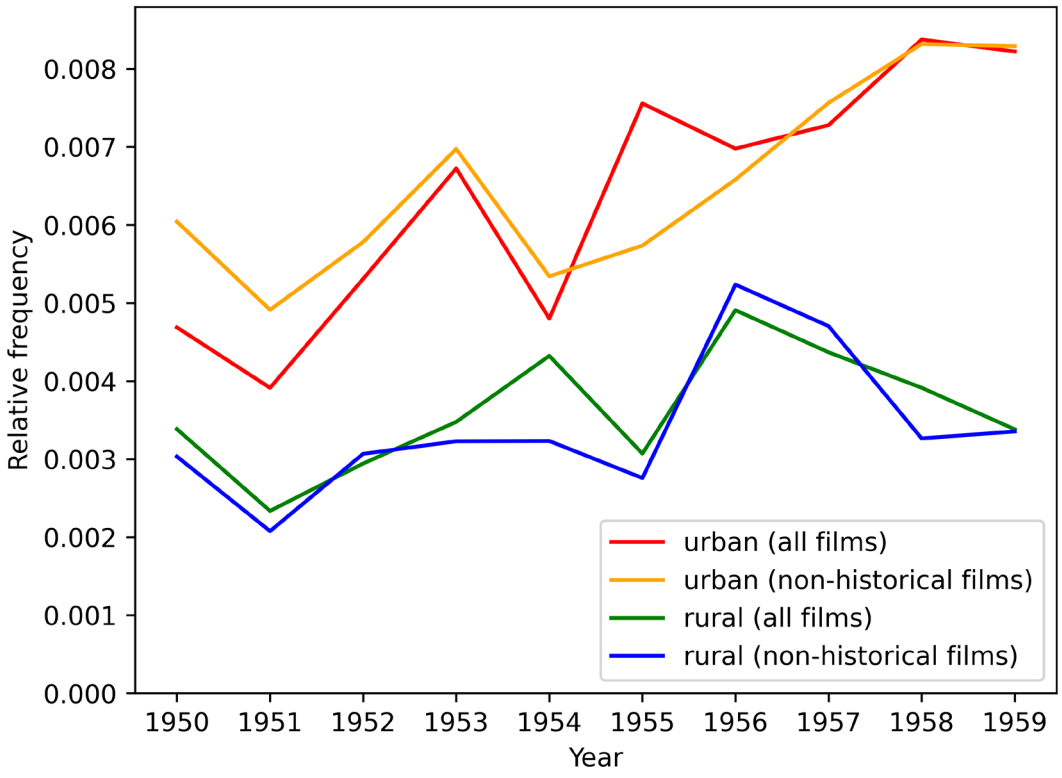
Tekstitysvertailussa lähtökohtana oli, että tekstitetyissä kohdissa on puhetta. Otimme tekstitystiedostoista aikaleimat ja vertasimme niitä ääntenhavannointimallin puhehavaintoihin. Nämä kaksi olivat päällekkäisiä keskimäärin 85 % ajasta. Yleisesti ottaen havainnointimalli löysi kohdat, joissa on tekstitystä, mutta malli arvioi puheen keston tekstitystä pidemmäksi. On kuitenkin tärkeää huomata, että tekstitysten ajoitukset eivät ole täysin tarkkoja. Tekstitykset saattavat esimerkiksi näkyä ruudulla varsinaista keskustelua pidempään. Vääriä positiivisia puhehavaintoja aiheutuu myös siitä, että havainnointimalli prosessoi yhden sekunnin mittaisia palasia, kun taas tekstitykset on ajoitettu millisekuntien tarkkuudella. Kaiken kaikkiaan 85 % päällekkäisyys on erinomainen puheenhavainnointitulos vanhojen elokuvien usein heikkolaatuiselle äänelle. Myöhemmässä kokeessa paransimme vielä tulosta soveltamalla opettaja–oppilas-menetelmää (Dinkel et al. 2021). Tämä ratkaisu nosti puhehavaintojen ja tekstityksen päällekkäisyyden 90 prosenttiin.

Ihmisen tekemiin merkintöihin vertailtaessa keskityimme äänihavaintojen osalta kolmeen yläkategoriaan: musiikkiin, maaseudun ääniin ja urbaaneihin ääniin. Musiikin, joka oli havaituista äänistä toiseksi yleisin ja olennainen osa elokuvia, malli havaitsi hyvin, sillä saanti (*recall*) oli täydet 100 % ja tarkkuus (*precision*) noin 85 %. Tämä tarkoittaa, että malli löysi kaikki näytteet, joissa oli musiikkia, ja tuotti vain muutaman ylimääräisen musiikkihavainnon.

Seuraavaksi tutkimme maaseudun ääniä, jollaisiksi laskimme tutkimuksessa ensisijaisesti maatilan eläinten äänet. Ihmisen merkitsemistä maatilan äänistä malli havaitsi laskujemme mukaan 62,5 % tapauksista. Osan luokista, kuten hevosen äänet, malli havaitsi täydellisellä 100 % tarkkuudella.

Urbaanien ja modernien äänten kohdalla vertailua vaikeutti datan vähäisyys. Suurinta osaa urbaaneista ääniluokista esiintyi vain yhdessä tai kahdessa näytteessä, minkä takia laskelmat tarkkuudesta olivat epäluotettavia. Vaihtelu tarkkuudessa oli suurta eri luokkien välillä. Keskimäärin tarkkuus oli noin 50 %, mutta monille luokille tarkkuus oli odottamaamme huonompaa. Esimerkiksi aseen laukaukset havaittiin noin 50 % tarkkuudella, mutta usein ne luokiteltiin väärin papattimatoiksi.[5] Junan äänien kohdalla tarkkuus oli vain 25 %, mikä oli huomattava pudotus AudioSet-aineiston alkuperäisestä 70 % tarkkuudesta. Vastaavasti veneiden äänien havainnointitarkkuus oli matala, sillä ne sekoittuivat monesti auton ääniin. Muihin yleisiin mallin tekemiin virheisiin lukeutuivat muun muassa kellokortti- ja kassakoneen äänten sekoittaminen toisiinsa ja ajoneuvo-yläkategorian käyttö tarkemman luokan, esimerkiksi moottoriveneen, junan tai auton tilalla.

Vertasimme maaseudun ja kaupungin äänien suhteellista esiintyvyyttä 1950-luvun eri vuosina julkaistuissa elokuvissa. Vertailun tuloksia havainnollistaa kuva 4. Kaupunkimaisten äänien määrässä on havaittavissa loiva kasvu verrattaessa vuosikymmenen alku- ja loppupuolta, kun taas maaseudun äänien määrä elokuvissa pysyy suurin piirtein samana läpi 1950-luvun. Erotimme myös historialliset elokuvat 1950-lukuun sijoittuvista elokuvista, mutta näiden välillä ei tässä vertailussa näkynyt juurikaan eroa.



Kuva 4. Havaittujen kaupunkiin ja maaseutuun viittaavien äänien suhteelliset määrät eri vuosien elokuvissa. Lähde: Grósz et al. (2022).

5 On myös mahdollista, että aseenlaukausääniä on tuotettu 1950-luvulla papattimattojen avulla, sillä tiedetään, että aitojen äänien sijasta on silloin tällöin käytetty vastaavia ääniefektejä.

Puheentunnistus

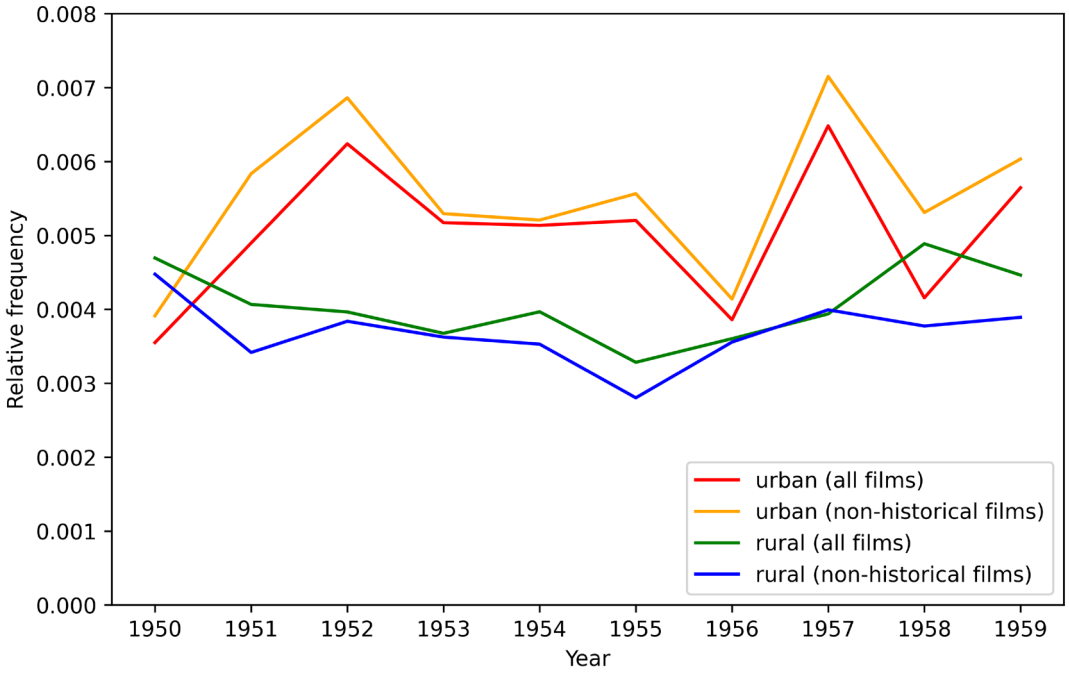
Puheentunnistukseen käyttämämme malli saavutti tuoreella Lahjoita puhetta -aineiston testijoukolla sanavirheprosentin (*word error rate*, *WER*) 23,8 % ja merkkivirheprosentin (*character error rate*, *CER*) 9,5 %. Nämä luvut lasketaan jakamalla virheiden määrä todellisessa transkriptiossa olevalla sanojen (WER) tai kirjainten (CER) määrällä. Elokuva-aineistolla puolestaan pystyimme laskemaan karkean arvion virheprosentista 44 elokuvan tekstityksistä ja tämän lisäksi tarkan arvion vain pienestä 100 minuutin testiaineistostamme.

Tekstitysvertailussa sanavirheprosentti (WER) oli noin 62,3 % ja merkkivirheprosentti (CER) 40,4 %, mikä on huomattavasti huonompi kuin Lahjoita Puhetta -testijoukon tulos. Merkittävä suureen eroon vaikuttava tekijä on se, että tekstitykset ovat harvoin tarkkoja transkriptioita: niissä tiivistetään puhe niin, että se mahtuu rajalliseen tilaan, poistetaan epäröinnit, korjaukset ja ei-sanalliset ilmaukset sekä joskus muunnetaan puhekielisiä, epävirallisia sanamuotoja muodollisempaan tyyliin. Tämän vuoksi nämä luvut ovat luultavasti paljonkin todellisia WER- ja CER-arvoja korkeammat. Hypoteesia tukee myös se, että 100 minuutin tarkasti litteroidulla testiaineistolla virheprosentit olivat odotetusti pienempiä: 52,9 % ja 27,5 % WER:n ja CER:n osalta. Lopuksi vertasimme vielä testiaineiston transkriptioita tekstityksiin saadaksemme käsityksen siitä, kuinka paljon ne erosivat toisistaan. WER oli tässä tapauksessa 42,2 % ja CER 22,6 %, mikä osoittaa tekstitysten poikkeavan selvästi transkriptiosta. Koska on epätodennäköistä, että puheentunnistustulos voisi päästä lähemmäs elokuvatekstitystä kuin ihmisen tekemä transkriptio, jota se on opetettu jäljittelemään, tämä ”tekstitysvirheiden” määrä antaa käytännössä alarajan sille virhemäärälle, mitä täydellisesti onnistunut puheentunnistus voisi tuottaa, kun sitä verrataan tekstityksiin.

Analyysiamme varten etsimme elokuvien puheentunnistustuloksista maaseutu- ja kaupunkiympäristöön viittaavia sanoja. Näiden avainsanaryhmien perustana olivat ääntenhavainnoinnin luokat, mutta muokkasimme ja laajensimme näitä sanalistoja sisällyttääksemme niihin sanoja, jotka todennäköisemmin esiintyisivät keskustelussa. Esimerkiksi ”junan torvi” on hyvä ääntenhavainnoinnin luokka, mutta ei todennäköisesti esiinny keskustelussa sellaisenaan.

Ääntenhavainnoinnin luokista valitsimme noin 20 sanaa kumpaankin ryhmään, mutta otoskoon kasvattamiseksi keräsimme lisää sanoja Word2Vec-mallin avulla (Mikolov et al. 2013). Word2Vec-mallit muuntavat sanat matemaattisiksi vektoreiksi niin, että samalla tavoin käytettävät sanat sijaitsevat lähellä toisiaan. Tätä ominaisuutta hyödyntäen poimimme kumpaankin luokkaan noin 150 sanaa, jotka muistuttivat merkitykseltään eniten kahta alkuperäistä listaa. Täydensimme alkuperäisiä listoja näillä sanoilla ja etsimme sitten kaikkia avainsanoja puheentunnistustuloksista. Kuvassa 5 esitetään kunkin vuoden elokuvissa havaittujen maaseutu- ja kaupunkiavainsanojen määrä jaettuna kunkin vuoden elokuvien sanojen kokonaismäärällä, jolloin saadaan suhteellinen frekvenssi. Kummassakaan luokassa ei ole havaittavissa selkeää nousevaa tai laskevaa suuntausta. Historialliset elokuvat kuitenkin sisälsivät odotetusti enemmän maaseutumaisia ja vähemmän kaupunkimaisia avainsanoja, mikä näkyy kaikkien elokuvien ja muiden kuin historiallisten elokuvien käyrien välisissä eroissa.

Lopuksi laskimme avainsanojen havaitsemisen tarkkuuden vertaamalla sitä tekstityksiin. Oletamme, että avainsanojen osalta tekstitykset ovat luotettavia transkriptioita, koska avainsanat ovat sisältösanoja, joita ei yleensä jätetä pois tekstiä tiivistettäessä. Lisäksi käytimme sanoista niiden kantoja, esimerkiksi ”hevonen” typistettiin muotoon ”hevo”, jotta kirjakielisten ja puhekielisten sanapäätteiden ero ei vaikuttaisi tuloksiin (vrt. ”hevosia” ja ”hevosii”). Yhdistettynä nämä kaksi



Kuva 5. Elokuvadialogien puheentunnistustranskriptioissa havaittujen kaupunki- ja maaseutuavainsanojen suhteelliset määrät eri vuosien elokuvissa. Lähde: Grósz et al. (2022).

avainsanaluetteloa sisälsivät 350 sanaa, jotka esiintyivät tekstityksissä yhteensä 1804 kertaa. Puheentunnistin löysi näistä esiintymistä 62 %, ja kaikista sen löytämistä esiintymistä oikeita oli 70 %.

Johtopäätökset

Artikkelissa teimme katsauksen puheen ja taustaäänten tunnistusmenetelmiin ja niiden kehitykseen ja kerroimme MoMaF-projektin tutkimustuloksista.[6] Tutkimuksessa kokeilimme, miten nykyaineistoilla opetetut tunnistusmallit selviävät 1950-luvun kotimaisten elokuvien ääniraidoista. Haasteina olivat sekä vanhojen äänitysten laatu että äänten ja varsinkin kielen erilaisuus nykyaineistoihin verrattuna. Tunnistustulosten mittaamisen teki hankalaksi ja epätarkaksi lisäksi se, ettei elokuvia ollut litteroitu eikä niissä esiintyviä ääniä kuvailtu.

Elokuva-aineiston äänitenhavainnoinnin ja puheentunnistuksen tulosten vertailu referenssituloksiin osoitti, että moderneilla aineistoilla opetetut tunnistusmallit tuntuivat toimivan kohtalaisesti aineiston ja tulosten mittaamisen haastavuudesta huolimatta. Yhteenvetona äänitenhavainnoinnin tuloksista voi todeta, että puhe, musiikki ja eläinten äänet havaittiin hyvin heikommasta äänen laadusta huolimatta, mutta ihmisen rakentamien teknisten koneiden ja laitteiden kohdalla havainnointitarkkuus oli heikompaa. Toisaalta analyysimme kannalta heikkoa tarkkuutta kompensoi jonkin verran se, että urbaanit äänet menivät keskenään sekaisin maaseudun ääniin sekoittumisen sijaan. Puheentunnistuksen tarkkuus oli selvästi heikompaa

6 Mallien tunnistustuloksia voi tarkastella myös itse osoitteessa http://momaf-data.utu.fi/momaf_bboxtool/annotations.html.

kuin monologeja ja vain vähän taustaääniä sisältävässä vertailuaineistossa, mutta se vaikutti silti riittävältä useimpien avainsanojen tunnistamiseksi.

Tarkempien tunnistustulosten saavuttamiseksi tarvittaisiin vanhoja elokuvia vastaavia äänten ja puheen opetusaineistoja. Mutta vaikka vanhoja elokuvia onkin tallella melko paljon, niitä tarvittaisiin hyvien mallien opettamiseen tuhansia. Lisäksi elokuvien sisältämän puheen litterointi ja äänien kuvailu opetusta ja testausta varten on hidasta ja raskasta käsityötä, mikä on käytännössä mahdotonta tässä mittakaavassa. Yksi jatkotutkimuksen aihe olisikin lisätä opetusdataa muokkaamalla suuria nykyaikaisia puhe- ja ääniaineistoja keinotekoisesti enemmän vanhoihin elokuviin sopivaksi. Merkityn kohdeaineiston puute hankaloittaa tätä samalla tavalla kuin valmiiden modernien mallien mukauttamista.

Lähteet

Baevski, Alexei; Zhou, Yuhao; Mohamed, Abdelrahman & Auli, Michael (2020) wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Teoksessa *Advances in Neural Information Processing Systems*. Curran Associates, 33, 12449–12460.

Bengio, Yoshua; Ducharme, Réjean & Vincent, Pascal (2000) A Neural Probabilistic Language Model. Teoksessa *Advances in Neural Information Processing Systems*. MIT Press, 13, 932–938.

Chan, William; Jaitly, Navdeep; Le, Quoc & Vinyals, Oriol (2016) Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. Teoksessa *2016 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP). 4960–4964.

Chorowski, Jan; Bahdanau, Dzmitry; Cho, Kyunghyun & Bengio, Yoshua (2014) End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results. *arXiv pre-print:1412.1602 [cs, stat]*. Saatavilla: <https://doi.org/10.48550/arXiv.1412.1602>.

Cramer, Jason; Wu, Ho-Hsiang; Salamon, Justin & Bello, Juan Pablo (2019) Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings. Teoksessa *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP). 3852–3856.

Dinkel, Heinrich; Wang, Shuai; Xu, Xuenan; Wu, Mengyue & Yu, Kai (2021) Voice Activity Detection in the Wild: A Data-Driven Approach Using Teacher-Student Training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29, 1542–1555.

Enarvi, Seppo (2018) *Modeling Conversational Finnish for Automatic Speech Recognition*. Väitöskirja, kieliteknologia, Sähkötekniikan korkeakoulu, Aalto-yliopisto. Saatavilla: <http://urn.fi/URN:ISBN:978-952-60-7908-0>.

Gemmeke, Jort F.; Ellis, Daniel P. W.; Freedman, Dylan; Jansen, Aren; Lawrence, Wade; Moore, R. Channing; Plakal, Manoj & Ritter, Marvin (2017) Audio Set: An ontology and human-labeled dataset for audio events. Teoksessa *2017 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP). 776–780.

Graves, Alex (2012) Sequence Transduction with Recurrent Neural Networks. *arXiv:1211.3711 [cs, stat]*. Saatavilla: <https://doi.org/10.48550/arXiv.1211.3711>.

Graves, Alex; Fernández, Santiago; Gomez, Faustino & Schmidhuber, Jürgen (2006) Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. Teoksessa *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. New York, NY, USAACM, 369–376.

Grósz, Tamás; Kallioniemi, Noora; Kiiskinen, Harri; Laine, Kimmo; Moisio, Anssi; Römpötti, Tommi; Virkkunen, Anja; Salmi, Hannu; Kurimo, Mikko & Laaksonen, Jorma (2022) Tracing Signs of Urbanity in the Finnish Fiction Film of the 1950s: Toward a Multimodal Analysis of Audiovisual Data. Teoksessa *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB)*. Uppsala, Ruotsi, 63–78.

Hinton, Geoffrey; Deng, Li; Yu, Dong; Dahl, George E.; Mohamed, Abdel-rahman; Jaitly, Navdeep; Senior, Andrew; Vanhoucke, Vincent; Nguyen, Patrick; Sainath, Tara N. & Kingsbury, Brian (2012) Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine* vol. 29:6, 82–97.

Kong, Qiuqiang; Cao, Yin; Iqbal, Turab; Wang, Yuxuan; Wang, Wenwu & Plumbley, Mark D. (2020) PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28, 2880–2894.

Lin, Liwei; Wang, Xiangdong; Liu, Hong & Qian, Yueliang (2020) Guided Learning for Weakly-Labeled Semi-Supervised Sound Event Detection. Teoksessa *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP). 626–630.

Lison, Pierre & Tiedemann, Jörg (2016) OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. Teoksessa *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia. European Language Resources Association (ELRA), 923–929.

Mesaros, Annamaria; Heittola, Toni; Virtanen, Tuomas & Plumbley, Mark D. (2021) Sound Event Detection: A tutorial. *IEEE Signal Processing Magazine* vol. 38:5, 67–83.

Mikolov, T; Chen, K; Corrado, G & Dean, J (2013) Efficient Estimation of Word Representations in Vector Space. Teoksessa *1st International Conference on Learning Representations*. Scottsdale, AZ, USA.

Moisio, Anssi; Porjazovski, Dejan; Rouhe, Aku; Getman, Yaroslav; Virkkunen, Anja; AlGhezi, Ragheb; Lennes, Mietta; Grósz, Tamás; Lindén, Krister & Kurimo, Mikko (2022) *Lahjoita puhetta: A Large-Scale Corpus of Spoken Finnish with Some Benchmarks. Language Resources and Evaluation*. Saatavilla: <https://doi.org/10.1007/s10579-022-09606-3>.

Schneider, Steffen; Baevski, Alexei; Collobert, Ronan & Auli, Michael (2019) wav2vec: Unsupervised Pre-training for Speech Recognition. *arXiv pre-print:1904.05862 [cs]*. Saatavilla: <https://doi.org/10.48550/arXiv.1904.05862>.

Shuyang, Zhao; Heittola, Toni & Virtanen, Tuomas (2020) Active Learning for Sound Event Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28, 2895–2905.