Mika Ihamäki

OHJELMISTO- JA LAITTEISTOPOHJAISEN PUHEENKÄSITTELYRATKAISUN FAR FIELD -SUORITUSKYVYN VERTAILU

OHJELMISTO- JA LAITTEISTOPOHJAISEN PUHEENKÄSITTELYRATKAISUN FAR FIELD -SUORITUSKYVYN VERTAILU

Mika Ihamäki
Opinnäytetyö
Syksy 2020
Tietotekniikan tutkinto-ohjelma
Oulun ammattikorkeakoulu

TIIVISTELMÄ

Oulun ammattikorkeakoulu
Tietotekniikan tutkinto-ohjelma, Laite- ja tuotesuunnittelu

Tekijä: Mika Ihamäki
Opinnäytetyön nimi: Ohjelmisto- ja laitteistopohjaisen puheenkäsittelyratkaisun far field -suorituskyvyn vertailu
Työn ohjaaja: Jaakko Kaski
Työn valmistumislukukausi ja -vuosi: Syksy 2020 Sivumäärä: 28

Opinnäytetyössä tutkittiin laitteisto- ja ohjelmistopohjaisten puheenkäsittelyratkaisujen far field -suorituskykyä puheohjauskäyttöliittymän tehtävissä ja vertailtiin näitä keskenään. Referenssijärjestelmä, johon suorituskykytuloksia peilattiin, oli puheohjauskäyttöliittymä ilman puheenkäsittelyratkaisua.

Suorituskykytutkimuksen tueksi selvitettiin kirjallisuudesta periaatteita tyypillisistä puheenkäsittelyratkaisun sisältävistä algoritmeista, kuten GSC-keilanmuodostus-, jälkisuodatus sekä jälkikaiunnanpoistoalgoritmeista. Lisäksi käytiin läpi puheohjauskäyttöliittymän muut osat, joita tarvitaan puheohjauskäyttöliittymän toimintaan: mikrofonit, audiorajapinnat, ASR (Automatic Speech Recognition) ja NLU (Natural Language Understanding).

Tuloksista saatiin ymmärrys laitteisto- ja ohjelmistopohjaisen ratkaisun kyvykkyydestä kahden mikrofonin puheohjauskäyttöliittymissä. Ohjelmistopohjainen ratkaisu osoittautui epävakaaksi kahdella mikrofonilla, koska FRR-prosentti (False Rejection Rate) vaihteli 6 %:ista 67 %:iin. Ratkaisuksi osoitettiin, miten mikrofonimatriisin vaihtaminen lineaarisesta pyöreään ja tällöin myös mikrofonien lukumäärän nosto kahdesta kolmeen käänsi ohjelmistopohjaisen ratkaisun epävakaasta vakaaksi puheenkäsittelyratkaisuksi. Tällöin FRR-prosentti vaihteli 1 %:sta 8 %:iin. Jatkokehitykseksi esitettiin erityyppisen keilanmuodostusalgoritmin soveltamista kahden mikrofonin puheenohjauskäyttöliittymäjärjestelmässä.

Asiasanat:

Automaattinen puheentunnistus
Keilanmuodostus
Puheenkäsittely
Akustiikka
Luonnollisen kielen ymmärrys

ABSTRACT

Oulu University of Applied Sciences
Degree Programme in Information Technology, Device and Product Design

---

Author: Mika Ihamäki
Title of thesis: Comparison of hardware- and software-based speech enhancement solutions' far field performance
Supervisor: Jaakko Kaski
Term and year when the thesis was submitted: Fall 2020 Number of pages: 28

During this thesis comparison of far field performance was done between hardware- and software-based speech enhancement solutions' in voice UI-tasks. The results were compared to results of a reference solution without speech enhancement algorithms.

To support the study typically used speech enhancement algorithms were researched, e.g. GSC-beamforming, postfiltering and dereverberation. In addition, other components of voice UI architecture that are essential to voice UI's were reviewed: microphones, audio interfaces, ASR (Automatic Speech Recognition) and NLU (Natural Language Understanding).

Results showed hardware- and software-based solutions' far field performance in two microphone voice UI systems. Software-based solution turned out to be unstable with two microphones showed by FRR % that ranged from 6 % to 67 %. Solution to the problem was proposed and tested. The effect of changing the microphone array from linear to circular and therefore increasing the number of microphones used to 3 gave better results: FRR % ranged from 1 % to 8 %. Proposition was made to apply a different type of beamforming technique on a 2-microphone voice UI system.

Keywords:

Automatic Speech Recognition
Beamforming
Speech Enhancement
Acoustics
Natural Language Understanding

SISÄLLYS

1 JOHDANTO

Ääniohjaus on vakiintunut käyttöliittymäksi älykkäissä laitteissa. Puheentunnistuksen suosio lisääntyy entistä enemmän älykkäissä laitteissa, kuten älykaiuttimissa ja televisioissa. Creoirin tuotekehitysprojekteissa on usein olennaisena ominaisuutena ääniohjauskäyttöliittymä.

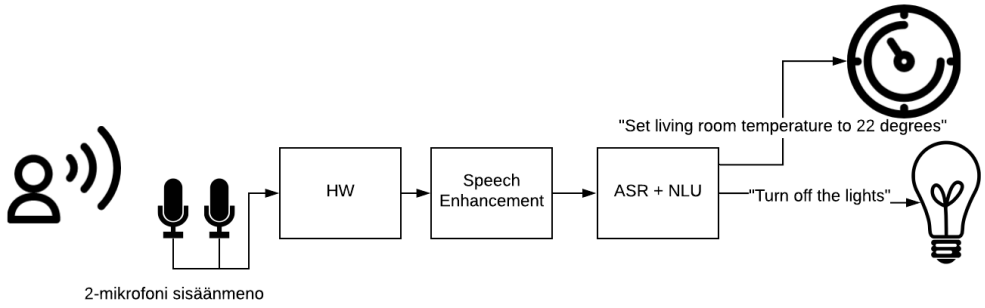
Tässä opinnäytetyössä tutustutaan ohjelmistopohjaisiin puheenkäsittelyratkaisuihin. Puheenkäsittelyalgoritmeilla parannetaan ääniohjauskäyttöliittymätuotteiden suorituskykyä akustisia häiriöitä sisältävissä olosuhteissa. Creoir on aikaisemmin käyttänyt lähinnä laitteistopohjaisia puheenkäsittelyratkaisuja. Tässä opinnäytetyössä tutkitaan mahdollisuuksia siirtyä laitteistopohjaisista ratkaisuista joissain tilanteissa ohjelmistopohjaisiin ratkaisuihin. Näiden käytöstä tuotekehityksessä on syntynyt rajoitteita, sillä puheenkäsittely tapahtuu laitteiston komponenteissa. Opinnäytetyö tulee tukemaan ohjelmistopohjaisten puheenkäsittelyratkaisujen suorituskyvyn ymmärtämistä sen tutkimustyyppisen lähestymistavan vuoksi.

Tutkimuksessa tullaan tekemään testejä neljässä erilaisessa akustisessa skenaariossa kahdella eri konfiguraatiolla: laitteistopohjaisella puheenkäsittelyratkaisulla sekä ohjelmistopohjaisella puheenkäsittelyratkaisulla. Testitulokset vertaillaan keskenään ja tehdään johtopäätökset siitä, minkälainen vaikutus molemmilla ratkaisuilla on ääniohjauskäyttöliittymän tehtävien suorittamiseen.

Lisäksi tavoitteena yrityksellä on tämän työn kautta kehittää ymmärrystä niistä teknologioista, joita ohjelmistopohjaisessa puheenkäsittelyratkaisussa käytetään. Tähän tavoitteeseen pääseminen avustaa tuotekehitysprosessin alkuvaiheessa teknologiavalintojen tekemisessä ja mahdollisten puheenkäsittelyratkaisujen parametrien säätämisessä käyttötarkoituksen tarpeen mukaan.

2 PUHEOHJAUSKÄYTTÖLIITTYMÄN OSAT JA NIIDEN TEHTÄVÄT

Creoir Oy tekee tuotekehitysprojekteissaan tuotteita, joiden pääasiallisena kyvykkyytenä on ääniohjauskäyttöliittymä. Tämä siis tarkoittaa sitä, että tuote pystyy tunnistamaan ihmisen puhetta (ASR, Automatic Speech Recognition) ohjelmiston ja laitteiston avulla sekä ymmärtämään puheen merkityksen ja tällöin tekemään toimintoja ymmärretyn puheen merkityksen tai aikomuksen perusteella (NLU, Natural Language Understanding). Näitä toimintoja voi olla esimerkiksi kodin valaistuksen ohjaus tai huoneen lämpötilan säätäminen. Kuvasta 1 nähdään kokonainen ääniohjauskäyttöliittymä eri kokonaisuuksineen.



KUVA 1. Korkean tason arkkitehtuurikuva puheohjauskäyttöliittymäjärjestelmästä.

Olosuhteista, joihin ääniohjauskäyttöliittymäjärjestelmät suunnitellaan sekä joita tämän opinnäytetyön testattavat ääniohjauskäyttöliittymäjärjestelmät käsittelevät, käytetään termiä ”far field”. Määritelmänä far field käsittää puheen kaappausta mikrofonilla vähintään 1 m:n päästä, mutta esimerkiksi laboratoriossa testattavissa testeissä käytetään 2,75 m:n etäisyyttä puhesignaalin lähteen ja testattavan laitteen välillä. Suorituskykytestien olosuhteista ja järjestelyistä kerrotaan lisää luvussa 3.

Huoneen olosuhteissa melun lähteitä on monenlaisia, joiden seasta järjestelmän on kyettävä kaappaamaan tarkoitettu puhesignaali. Nämä melunlähteet ovat taajuuksien tilastollisilta vaihtelevuuksiltaan staattisia tai dynaamisia. Staattisen melun lähteitä ovat esimerkiksi tuulettimien tai ilmanvaihtokoneiden hurina, joiden taajuusvaihtelevuus on kohtalaisen pieni. Dynaamisen melun lähteitä voivat olla olohuoneen kotiteatterin kaiutin, josta kuuluu musiikkia, tai sitäkin dynaamisempi melunlähde, puhuva ihminen, joita voi olla jopa useampi kappale paikan päällä. Tämän tyyppisiä melunlähteitä sisältävä akustinen skenaario on puheentunnistusjärjestelmälle erittäin haastava, mutta

myös hyvin realistinen nykymaailmassa. Tämä haastavuus nähtiin opinnäytetyön suorituskykytesteissä, joka näkyy myös lukujen 4.1, 4.2 ja 4.4 tuloksissa.

Näiden melulähteiden vaimennukseen ja häiriönsuodatukseen tarvitaankin laitteistolla tai ohjelmistolla suoritettavat puheenkäsittelyalgoritmit, joiden suorituskykyä tässä opinnäytetyössä vertaillaan. Näiden algoritmien tehtävänä on analysoida mikrofonien kaappaama signaali ja suorittaa melulähteiden vaimennus sekä häiriönsuodatus. Luvussa 3.3 syvennytään puheenkäsittelyalgoritmeihin.

Viimeiseksi luvuissa 3.4 ja 3.5 käydään läpi, kuinka itse puheentunnistus, puhesignaalin muuntaminen tekstiksi, tehdään ja kuinka voidaan ymmärtää tunnistettu puhe. Tällä ymmärtämisellä tarkoitetaan sitä, että analysoidaan mitä sanottiin ja selvitetään käyttäjän aikomus.

2.1 Laitteisto: prosessori, väylät, puheen näytteistys sekä dataformaatti

Audiojärjestelmän laitteistoon tyypillisesti kuuluu piirilevy, johon sisältyy prosessori, väylät sekä mikrofonit, joiden avulla puhe voidaan kaapata ja tallentaa digitaalisesti muistiin. Nykyään itse analogisen puhesignaalin muunnos digitaaliseen muotoon voidaan tehdä jo mikrofoneilla itsessään. Tämä vapauttaa analogia-digitaalimuuntimen (Analog-to-Digital Converter) piirilevyltä. Tunnetuin mikrofonityyppi on nimeltään MEMS-mikrofoni (Micro-Electromechanical-Systems), jonka ulostulona saadaan tämän opinnäytetyön kontekstissa digitaalista, PDM-muotoista (Pulse Density Modulation) tai I^2S-muotoista (Inter-IC-Sound) signaalia lähetettäväksi väyliin, jotka toimivat rajapintoina prosessorin ja mikrofonin välillä. Tämän jälkeen signaali muunnetaan PCM-muotoiseksi prosessorilla jatkokäsiteltäväksi.

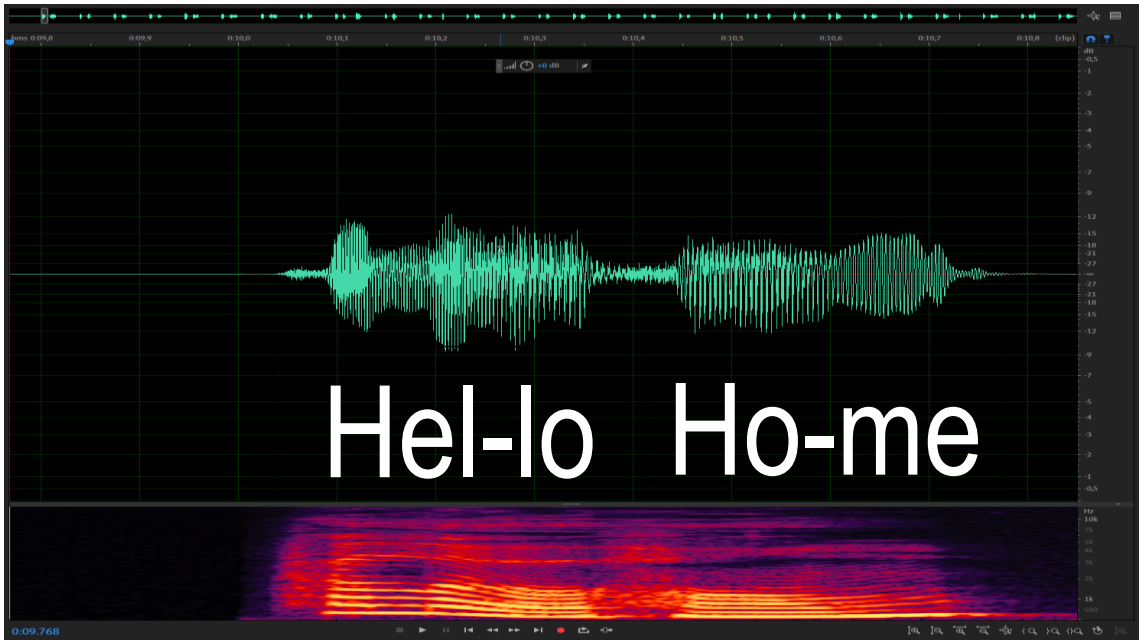
Puheääni näytteistetään PCM-muotoiseksi dataksi tietyllä näytteistystaajuudella. Tyypillinen näytteistystaajuus, jota käytetään, on 16 kHz, sillä puheentunnistusmoottori haluaa puhedatan näytteistettävän 16 kHz:n näytteistystaajuudella. 16 kHz riittää, sillä Nyquistin teoreeman mukaan 8 kHz:n signaalin uudelleenluontiin valitaan näytteistystaajuudeksi 16 kHz. Lisäksi bittileveydeltään puheen näytteet muunnetaan merkittyyn 16-bittiseen Little Endian -järjestettyyn formaattiin (S16_LE). Tämä bittileveys riittää, sillä 16-bittisillä luvuilla voidaan esittää tyypillisen MEMS-mikrofonin dynaaminen alue.

2.2 Signaalianalyysi

Ihmisen tuottaman puhe on signaalina monikaistainen signaali, jonka analyysiin tarvitaan taajuusanalyysia. Tätä analyysia voidaan käyttää sekä signaalinkäsittelyn ongelmaratkaisuun, kun käytössä on puhesignaalista tehty nauhoite, että puheenkäsittelyalgoritmien tekemien suodatuksien syöttönä. Tyypillisesti signaalinkäsittelyssä käytetään Fourier-muunnosta, jonka avulla voidaan hahmottaa ja visualisoida signaalien taajuudet. Tämäntyyppinen tapa analysoida signaalia on kuitenkin epätäydellinen. Syy tähän on se, että puhesignaalissa taajuudet muuttuvat hyvinkin pienessä ajassa huomattavasti. Tällöin tarvitaan tapa analysoida signaalia aika-taajuustasossa. (Virtanen – Vincent − Gannot 2018, luku 2.)

Yksi tavoista suorittaa aika-taajuus-analyysia, jonka avulla nähdään signaalin taajuus tietyllä ajan hetkellä, on lyhyen ajan Fourier-muunnos (STFT, Short-Time-Fourier-Transform). Tällä metodilla voidaan analysoida määritetyn aikaikkunan ajalta signaalin taajuuskäyttäytymistä.

Metodin visuaalisena havainnollistamisena käytetään spektrogrammia. Spektrogrammi on kolmiulotteinen malli, joka sisältää tietoa signaalin taajuudesta, ajasta sekä amplitudista. Signaalin taajuus nähdään y-akselilla, aika x-akselilla ja amplitudi logaritmisella asteikolla, tyypillisesti tietyn värin eri sävyjen spektrinä. Kuvasta 2 nähdään herätyssanan ”Hello Home” puheääninauhoitus.



KUVA 2. Puheääninauhoitus aikatasossa ja sen alapuolella spektrogrammi, joka kuvaa aika-taajuustasossa saman puheääninauhoituksen, jossa puhuttiin kirjaimilla ”Hello Home”.

2.3 Puheenkäsittelyalgoritmit

Opinnäytetyössä tehtävien suorituskykytestien tuloksissa voi välillä esiintyä vaihteluita. Näiden vaihtelujen syy-seuraus-analyysin vuoksi on hyödyllistä ymmärtää keinoja, joilla voidaan edistää mikrofonien kaappaaman puheen ymmärtävyyttä, kuten puheenkäsittelyalgoritmeilla.

Puheenkäsittelyalgoritmeilla pyritään edistämään signaali-kohinasuhdetta mikrofonien kaappaaman puheen sekä puheeseen kuulumattomien häiriöiden ja melujen välillä (SNR, Signal-To-Noise-Ratio), jotta puheentunnistusmoottori pystyy suoriutumaan tunnistustehtävästään mahdollisimman hyvin. Puheääneen voi helposti sekoittua eri melulähteiden sisältöä, kuten tuulettimen kohinaa tai toisen ihmisen puhetta.

Lisäksi ääniohjauskäyttöliittymäjärjestelmät suunnitellaan reaalimaailman ympäristöihin, jotka sisältävät huoneita, joissa on esteitä. Esteillä tarkoitetaan esimerkiksi seiniä, huonekaluja yms. objekteja, joista puheääni voi heijastua. Lisäksi äänen heijastuessa ympäri huonetta syntyy jälkikaikuja. Nämä ilmiöt heikentävät puheentunnistusjärjestelmän tunnistamisen suorituskykyä, jonka vuoksi tarvitaan algoritmeja vähentämään tai poistamaan näiden ilmiöiden vaikutuksia. Näitä algoritmeja ovat keilanmuodostus-, jälkisuodatus-, sokea lähde-erottelu- sekä jälkikaiunnanpoistoalgoritmit.

2.3.1 Keilanmuodostusalgoritmi

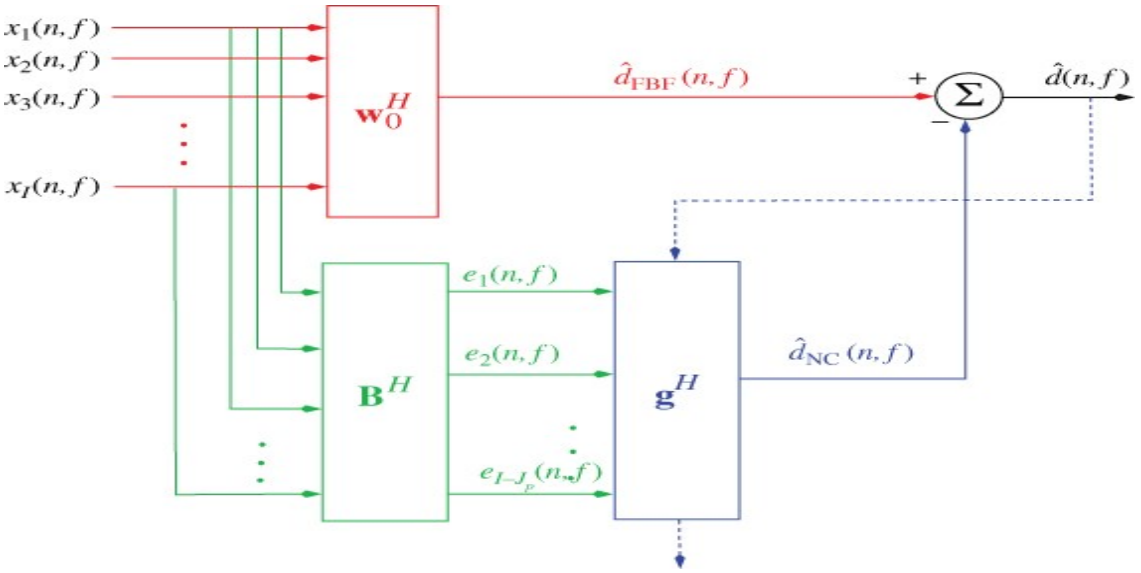
Keilanmuodostusalgoritmi on tapa tehdä äänen kulkuaikaan perustuvaa suodatusta mikrofoneilla, joiden etäisyys toisistaan on vakio. Keilanmuodostuksella on perinteisesti tarkoitettu tietystä suunnasta saapuvan signaalin keskittymän muodostamista. Nykyään kuitenkin termin määritelmä on laajentunut monen sisääntulon ja yksittäisen ulostulon järjestelmäksi, jossa yhdistetään tietty määrä eri suodatettuja mikrofonisignaaleja. (Markovich-Golan – Kellermann − Gannot 2018, luku 10.)

Keilanmuodostusalgoritmeja on suunniteltu useita eri tyyppejä erilaisiin konteksteihin, niin laajakaistaisiin kuin kapeakaistaisiin konteksteihin. Tämän opinnäytetyön kontekstista tiedetään, että ihmisen puhe sekä sen vallitsevien olosuhteiden häiriölähteiden signaalit ovat laajakaistaisia.

Siispä tämän opinnäytetyön kontekstissa käydään läpi laajakaistaisten signaalien käsittelyyn ja akustisesti dynaamisiin olosuhteisiin suunniteltu keilanmuodostusalgoritmi.

Yksi esimerkki algoritmeista on GSC-keilanmuodostusalgoritmi (GSC, Generalized Sidelobe Canceller). Tämä algoritmi on kehitetty LCMV-algoritmin (Linear Constraint Minimum Variance) pohjalta jakamalla sen toteutus kahteen osaan. Itse LCMV-algoritmin periaatteena on säilyttää tietyn yksittäisen äänilähteen äänenpainetasoa sekä samanaikaisesti vaimentaa ulkoisten äänilähteiden äänenpainetasoja. LCMV-keilanmuodostusalgoritmin rajoituksena on sen kyvyttömyys sopeutua pikaisesti melulähteen tuottaman signaalin muutoksiin, sillä sen suodattimen kertoimet pysyvät vakioina. GSC-keilanmuodostusalgoritmin toteutuksen alemmassa signaalipolussa suodattimen kertoimet taas kykenevät sopeutumaan pikaisiin muutoksiin melulähteen tuottaman signaalin tilastoissa. (GSC Beamformer. 2020.)

Kuvassa 3 on esitetty GSC-algoritmi vaiheittain. Syöttösignaalina tuodaan äänisignaalit aika-taajuusmuodossa, joihin liitetään aikaeropainotukset riippuen siitä, millä ajan hetkellä ne ovat saavuttaneet mikrofonit ja tällöin algoritmi saa tiedon signaalien saapumiskulmista.



KUVA 3. GSC-keilanmuodostusalgoritmin prosessoinnin eri vaiheet (Virtanen ym. 2018, luku 10.5).

Syöttösignaalit syötetään kahta polkua pitkin kohti vakioiden painotusten keilanmuodostusosiota $\mathbf{w}_0^H$ (punaisella merkitty), joka pyrkii säilyttämään pääasiallisia puhesignaaleja ($x_1(n,f)$, $x_2(n,f)$, $x_3(n,f)$, $x_I(n,f)$ jne.) sekä kohti estomatriisia, joka poistaa ne signaalit, joita ylempi polku säilyttää.

Näin estomatriisi ($\mathbf{B}^H$) pystyy tekemään arviot häiriöistä ja melusta, jotka sekoittuvat puheen äänilähteeseen. Häiriönpoisto-osio $\mathbf{g}^H$ käyttää näitä arvioita tehdäkseen arvion ylemmän $\mathbf{w}_0{}^H$ luoman keilanmuodostuksen jäljelle jäävästä häiriöstä ja melusta. Lopuksi lasketaan osioiden $\mathbf{w}_0{}^H$- ja $\mathbf{g}^H$-ulostulojen erotus ja tuodaan tämä tulos takaisinkytkentänä $\mathbf{g}^H$-osiolle. Takaisinkytkennän avulla voidaan sopeuttaa suodattimen kertoimet. (Markovich-Golan – Kellermann – Gannot 2018, luku 10.5.)

2.3.2 Jälkisuodatusalgoritmi

Jälkisuodatusalgoritmi toimii keilanmuodostusalgoritmin kanssa yhdessä niin, että se liitetään keilanmuodostusalgoritmin ulostuloon. Periaatteellisesti jälkisuodatusalgoritmi jatkaa keilanmuodostusalgoritmin häiriö- ja melusignaalien vaimennusta. Jälkisuodatusalgoritmin suodatus toteutetaan kuitenkin spektrisellä suodatuksella, joka on yksittäisen kanavan suodatusmetodi, toisin kuin keilanmuodostusalgoritmi, joka toteutetaan vähintään kahdella kanavalla. Lisäksi jälkisuodatusalgoritmi kykenee sopeutumaan nopeisiin muutoksiin häiriösignaaleissa, joka on tyypillinen ilmiö laitteen olosuhteissa. (Markovich-Golan – Kellermann – Gannot 2018, luku 10.6.)

2.3.3 Sokea lähde-erottelualgoritmi

Sokea lähde-erottelualgoritmi pyrkii erottelemaan lähteitä toisistaan tilanteessa, jossa ei tiedetä akustisen skenaarion melulähteiden ja puheäänilähteen ominaisuuksista mitään. Tunnettu akustinen skenaario, joka kuvaa tätä tilannetta on ns. *cocktail party problem*. Ainoa tieto, mitä lähteistä tiedetään, on niiden sekoitus ja tästä sekoituksesta olisi eroteltava lähteet, jotta voidaan maksimoida puheen selkeys ja suodattaa tai vaimentaa muita puheeseen kuulumattomia melulähteitä. (Hyvärinen – Karhunen – Oja 2001, luku 7.1.)

Tunnettu tapa suorittaa sokea lähde-erottelu on käyttää yksittäisen komponentin analyysia (ICA, Independent Component Analyysi), jonka avulla voidaan tehdä tilastollinen arvio käyttäen mikrofonien kaappaaman puheen signaalia, olettamalla, että jokainen alkuperäisistä lähteistä on tilastollisesti itsenäinen. (Hyvärinen ym. 2001, luku 7.1)

2.3.4 Jälkikaiunnanpoistoalgoritmi

Jälkikaiunnan ilmiö syntyy, kun puhe heijastuu heijastavista pinnoista. Ilmiönä jälkikaiunta jakautuu suoraan mikrofonille saapuvaan ääneen sekä kaikuun, jonka kesto ja voimakkuus on riippuvainen tilasta. Tilassa on heijastavia pintoja, jotka kasvattavat kaiun voimakkuutta ja kestoa. Ääntä imevät pinnat taas vähentävät kaiun voimakkuutta ja kestoa.

Puheentunnistuksen tapauksessa jälkikaiunnan ilmiö on haitallinen varsinkin voimakkaasti kaikuvissa huoneissa, joissa jälkikaiunnan vaikutuksesta puheen yksittäinen sana tai lause voi sekoittua seuraavaan sanottuun sanaan tai lauseeseen. Tällöin puheesta tulee epäselvää ja hyvin mahdollisesti puheentunnistusjärjestelmä ei kykene tunnistamaan puhetta vaan hylkää tunnistuksen. Tästä syystä puheentunnistusjärjestelmien testaukseenkin asetetaan vaatimus RT60-jälkikaiunta-ajalle (Reverberation Time) tiettyjen raja-arvojen väliltä. RT60-jälkikaiunta-aika kuvaa sitä aikaa, kuinka kauan kestää, että äänisignaalin äänenpainetaso on laskenut 60 dB:iä. (Habets – Naylor ym. 2018, luku 15.1.)

2.4 Automaattinen puheentunnistus (ASR)

Automaattinen puheentunnistus on vertailtavien laitteiden puheohjauskäyttöliittymien toiminnallisuuksien ytimenä. Sen tehtävänä on laskea tilastollinen arvio siitä, mikä sekvenssi sanoja mikrofonien kaappaamassa puheäänessä todennäköisimmin sanottiin. Tämän laskennan komponentteja ovat puheen ominaisuuksien erottelu, puheentunnistaja, akustinen malli, leksikaalinen sanasto sekä kielellinen malli. Akustinen malli tarkoittaa puheen ominaisuuksien todennäköisyyttä ja todennäköisten tilojen sekvenssiä foneemisekvenssissä. Leksikaalisessa sanastossa kuvataan foneemisekvenssit sanasekvenssien sisällä. Tämä sanasto on määritelty ääntämissanaston avulla. Kielellinen malli kuvaa sanojen sekvenssin ennakkotodennäköisyyttä, joka luodaan opettamalla neuroverkko laajalla tekstikorpuksella. (Watanabe – Virtanen − Kolossa 2018, luku 17.2.1.)

Tämän opinnäytetyön kontekstissa tarkastellaan sanastopohjaista puheentunnistusjärjestelmää. Sanastolla voidaan asettaa puheentunnistusjärjestelmälle sanat, jotka voidaan hyväksyä puheeksi. Näin ollen sanaston ulkopuolella jäävät sanat hylätään puhetta tunnistettaessa.

Kyseinen puheentunnistusjärjestelmä, joka on käytössä molemmissa testattavissa laitteissa, tekee kaiken tunnistukseen tarvittavan laskennan itse laitteen prosessorilla. Näin ollen voidaan kutsua puheentunnistusjärjestelmää adjektiivilla paikallinen tai offline. Tyypillisesti tämä tehdään pilvipohjaisissa web-palveluiden virtuaalikoneissa, joihin mikrofonien kaappaaman puheen data lähetetään. Esimerkiksi Amazon Alexa- virtuaaliavustaja toimii näin.

2.5 Luonnollisen kielen ymmärtäminen (NLU)

Luonnollisen kielen ymmärtäminen- osion (NLU, Natural Language Understanding) avulla voidaan ymmärtää tunnistetun puheen tarkoitus. Tämä komponentti ääniohjauskäyttöliittymäjärjestelmässä vastaanottaa puheentunnistusjärjestelmältä syöttönä tekstimuotoisia sanoja, joista muodostetaan ymmärrys siitä, mitä käyttäjä haluaisi mahdollisesti tehdä. Puheentunnistukseen kykenevissä laitteissa on tyypillistä nimenomaan tällainen kotiapuri-toiminnallisuus, joka ottaa pyyntöjä tai käskyjä vastaan käyttäjältä ja pyrkii toteuttamaan asiakkaan tarpeen. Näin ollen syntyy puheella toimiva käyttöliittymä ihmisen ja koneen välillä. Tämän opinnäytetyön testattavissa järjestelmissä NLU on toteutettu kotiautomaation ääniohjauskäyttöliittymänä. Ääniohjauskäyttöliittymällä voidaan ohjata valoja, säätää lämpötilaa tai sulkea ja avata ovia.

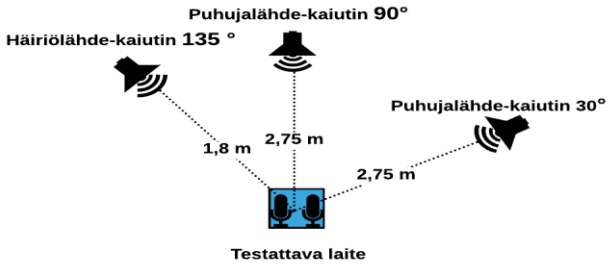
3 SUORITUSKYKYTESTIEN VALMISTELU, OLOSUHTEET JA METODIT

3.1 Testien järjestely sekä olosuhteet

Testauksen järjestely sekä olosuhteet noudattivat Alexa Voice Servicen akustiikkasertifioinnin spesifikaation asettamia vaatimuksia, joiden avulla voidaan itsenäisesti suorittaa akustinen testaus. Huoneelle on asetettu vaatimuksia, joiden on täytyttävä, jotta testaus on luotettava. Näitä olosuhteiden ominaisuuksia ovat huoneen mitat, testilaitteisto, huoneen pinnat sekä akustisen signaalin vaimentumisaika.

Huoneelle asetetut vaatimukset ovat: huoneen sisäiset mitat on oltava 4,4 m x 5,9 m 2,4 m, pohjakohinan äänenpainetaso saa olla maksimissaan 35 dBSPL (A) (SPL, Sound Pressure Level, A = a-painotus) ja akustisen signaalin jälkikaiunta-ajan täytyy olla suurempi kuin 0,2 s ja pienempi kuin 0,7 s 125 Hz:n ja 8 kHz:n välisillä taajuuksilla. Jotta jälkikaiunta-ajat voidaan saavuttaa, huone kalustetaan akustisella vaimennuksella. Viimeiseksi huoneen seinien ja testijärjestelmän kaiuttimien välillä sekä seinien ja testattavan laitteen välillä on oltava 0,5 m:n etäisyys. (Alexa Acoustic Testing Guide. 2020.)

Testijärjestelmän laitteisto koostuu kolmesta kaiuttimesta, joista kaksi kaiutinta on sijoitettu 2,75 m:n päähän testattavan laitteen mikrofoneista sekä yhdestä kaiuttimesta, joka on sijoitettu 1,8 m:n päähän testattavan laitteen mikrofoneista, Windows-työasemasta, ulkoisesta äänikortista sekä itse testattavasta laitteesta. Nämä kaiuttimet on sijoitettu lintuperspektiivistä katsottuna kuvan 4 mukaisesti: kaksi puhenauhoitetta toistavaa kaiutinta 30°:n sekä 90°:n kulmissa testattavasta laitteesta ja häiriönauhoitetta toistava kaiutin 135°:n kulmassa testattavasta laitteesta.



KUVA 4. Testattavan laitteen sekä kaiuttimien sijoitus geometrisesti testiympäristöön.

3.2 Testijärjestelmän kalibrointi

Testijärjestelmän kaiuttimien äänenpainetasot kalibroitiin tavoiteltuihin arvoihin, jotta voitiin varmistua halutusta signaali-kohina-suhteesta (SNR, Signal-To-Noise-Ratio). Tällöin tiedettiin myös tarkasti mitä ovat puheäänilähteen ja meluäänilähteen äänenpainetasot desibeleissä.

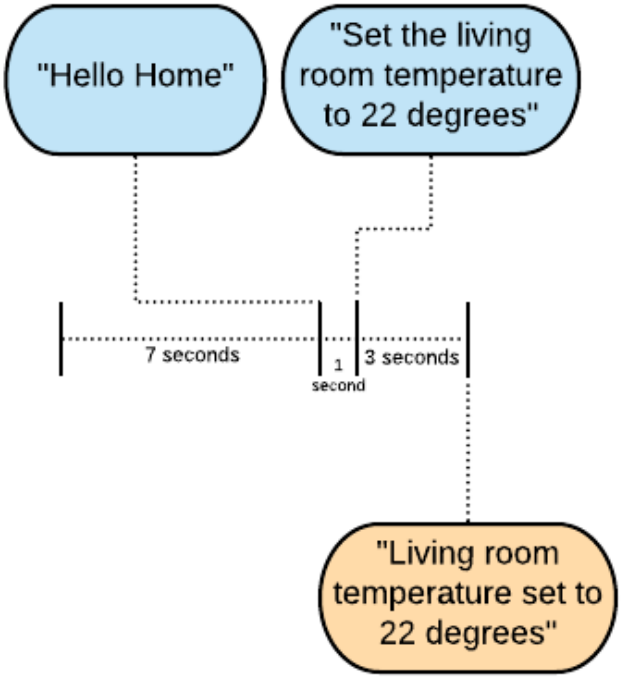
Kalibrointia varten tarvittiin äänenpainemittari. Äänenpaineentasot mitattiin äänenpainetaso-mittarilla (SPL Meter, Sound Pressure Level Meter). Lisäksi sen asetuksista asetettiin C-painotus, joka ottaa huomioon enemmän alhaisten taajuuksien energiaa kuin A-painotus. Tietokoneen äänenvoimakkuus asetus on useimmiten käytetty säädin, mutta Creoirin laboratorion monikanavaisella ulkoisella äänikortilla sekä Adobe Auditionin multitrack-toiminnon desibelisäädintä hyväksikäyttäen, voitiin asettaa äänenvoimakkuus erikseen jokaiselle ääniraidalle. Tällöin äänenvoimakkuusasetus asetettiin 100 %:iin ja tämä asetus pysyi vakiona. Adobe Auditionin asetus on ensisijaisesti +0 dB: ja tätä säätämällä ensiksi haettiin kokeilemalla tavoiteltu puheen äänenpainetaso.

Tavoiteltu puheen äänenpainetaso saavutettiin asettamalla puheääninauhoitteen sisältävän kanavan äänenvoimakkuustaso −27 dB:iin. Tämä kuvastaa akustista skenaariota ”Silence”. Seuraavaksi etsittiin sopiva äänenvoimakkuusasetus eri melunauhoituksien kanaville, joka pätee sekä ajan kuluessa taajuussisällön muuttuvuudeltaan staattiseen melunahoitukseen, kuten ”pinkkiin häiriöön”, joka kuvastaa testitapausta ”External Noise”, että musiikkimelunauhoitukseen, joka kuvastaa testitapausta ”External Music”. Viimeiseksi etsittiin sopiva äänenvoimakkuusasetus huomattavasti ajan kuluessa taajuussisällön muuttuvuudeltaan vielä dynaamisemmalle melunauhoitukselle, joka kuvastaa testitapausta ”External Babble Noise”. Tämän melunahoituksen kanavalle asetettiin äänenvoimakkuusasetukseksi −32 dB.

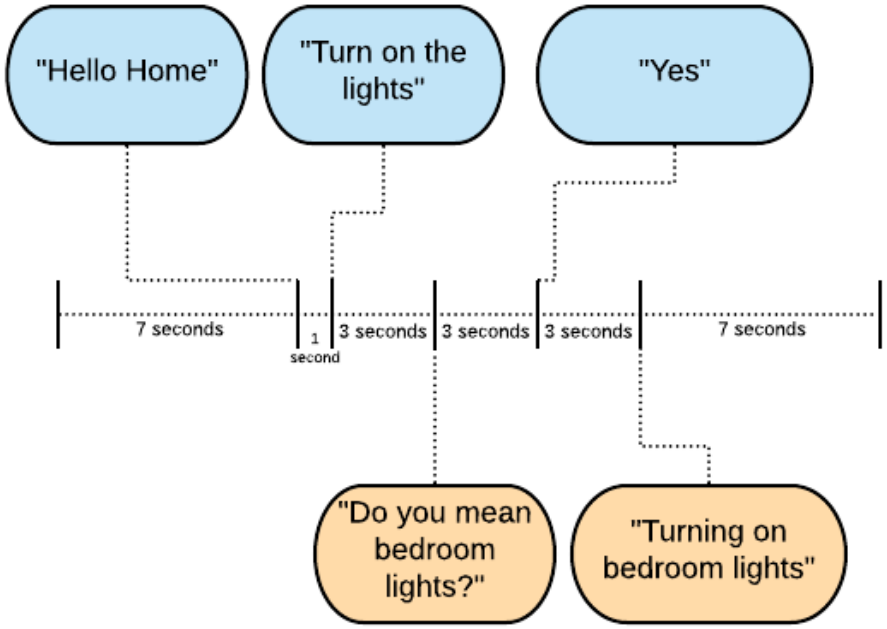
3.3 Testiraidan luonti

Testausta varten tilattiin puheääniraitoja kymmenen puhujan puhumana, viiden miehen ja viiden naisen, jotka olivat Amerikan Englannin natiivipuhujia, sillä puheentunnistusjärjestelmän kielellinen malli oli US-EN-tyyppiä (American English). Nämä 10 puhujan nauhoitteet sisälsivät 28 kertaa herätyssanan ”Hello Home” ja 32 lausetta, kuten ”Turn off the lights”. 28 lauseista sanottiin puheraidassa vaihtelevien aikavälien päästä herätyssanan sanomisesta (kuva 5). Loput 4 lausetta sanottiin vakioidun viiveen päästä vastauksena puheentunnistusjärjestelmän vastaukselle (kuva 6).

Tavoitteena oli luoda dialogi henkilön ja puheentunnistusjärjestelmän kanssa. Tämän tavoitteen saavuttamiseksi lisättiin 3 sekunnin tauko puheentunnistusjärjestelmän vastauksen tarkennukselle (kuva 6) sekä jokaisen henkilön puhuman sekvenssin jälkeen lisättiin 7 sekunnin tauko. Tällöin puheentunnistusjärjestelmälle annettiin tarpeeksi aikaa kertoa vastauksensa, ennen kuin puhuminen alkoi uudestaan eikä puheentunnistusjärjestelmän sanoma vastaus kuulunut päällekkäin puhujan seuraavan lauseen kanssa. Alla olevat kuvat havainnollistavat yhtä herätyssanan jälkeistä lauseen dialogia (kuva 5) ja herätyssanan, lauseen sekä käyttäjän lisäyksen dialogia (kuva 6).



KUVA 5. Dialogi käyttäjän ja puheentunnistusjärjestelmän välillä (käyttäjä sinisellä, kotiapuri oranssilla).



KUVA 6. Dialogi käyttäjän ja puheentunnistusjärjestelmän välillä käyttäjän lisäyksellä (käyttäjä sinisellä, kotiapuri oranssilla).

Puheraidan rinnalle tuotiin kappaleessa 3.2 esiteltyjen melunauhoitusten raidat kuvaamaan kolmea testitapausta: ”External Noise”, ”External Music” ja ”External Babble Noise”. Jokaiselle kolmelle testitapaukselle luotiin oma raita, sillä tapauksilla oli vaihtelevia äänenvoimakkuusasetuksia. Jokainen kolmesta meluraidasta synkronoitiin alkamaan samasta hetkestä kuin puheraitakin. Tekemällä näin saatiin vakioitua testit niin, että muuttuvia tekijöitä ei ole testin asettelusta peräisin: jokainen puhuttu sana kuuluu jokaisella testikierroksella samalla tietyllä ajan hetkellä tietyssä kohdassa häiriöraitaa. Tällöin, jos tuloksissa on muuttuvia tekijöitä, tekijöiden syy voidaan johtaa muuhun kuin vaihtelevuuksiin testien järjestelyissä testikertojen välillä.

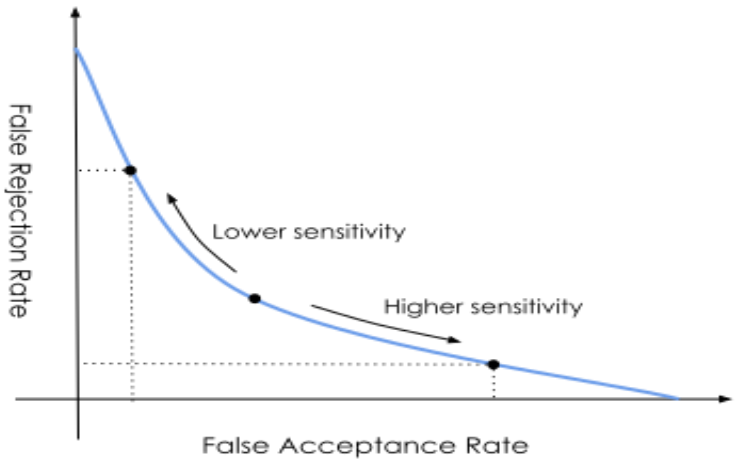
3.4 Suorituskyvyn vertailuun tehtävät testit

Suorituskyvyn vertailuun käytetään Amazonin Alexa Servicen akustisen testauksen ohjeen määrittämiä metriikoita, FRR (False Rejection Rate) ja FAR (False Acceptance Rate). FRR kuvaa prosenttilukua, joka määrää sitä osuutta, kuinka monta kertaa puheentunnistusjärjestelmä ei kyennyt tunnistamaan kuultua puhetta puheeksi ja heräämään unitilasta. Esimerkiksi jos puheentunnistus-

järjestelmä kuulee 100 kappaletta herätyssanoja ja epäonnistuu tunnistamasta kymmentä kappaletta herätyssanoista, FRR-taso on suuruudeltaan 10 %. FAR kuvaa sitä lukumäärää, kuinka monta kertaa laite oli virheellisesti tunnistanut herätyssanan ja herännyt unitilasta. Suorituskykytesteissä käytettävissä 12 h:n kestoisissa FAR-testeissä tyypillisesti nähdään yksittäisiä kertoja, kun laite oli virheellisesti herännyt unitilasta. Laitteen ei pitäisi herätä unitilasta silloin, kun sen herätyssanaa ei ollut sanottu. Tarpeettomat heräämiset eivät ole toivottavia, jonka vuoksi heräämisten lukumäärä olisi minimoitava.

Näiden kahden metriikan välillä on molemminpuolinen yhteys liittyen siihen, kuinka herkästi puheentunnistusjärjestelmä konfiguroidaan hyväksymään heräämisiä. Näihin testeihin liittyy tietty konfiguraatioasetus, joka määrittää lasketun arvon perusteella, hyväksytäänkö tunnistus ja herätetäänkö laite. Liian korkea tunnistuksen herkkyyden konfiguraatio johtaa siihen, että laite ei lähes ikinä herää ilman syytä, mutta tunnistusten määrä voi laskea jopa huomattavasti riippuen akustisesta skenaariosta. Jos tunnistuksen herkkyys taas on liian alhainen, laite herää paljonkin ilman syytä sekä myös hyväksyy suurimman osan herätyssanoista.

Kuvassa 7 on esillä tyypillinen kuvaaja siitä, kuinka FRR (y-akseli) ja FAR (x-akseli) tasojen muuttumisella on vaikutus toisiinsa. Kuvaajasta nähdään, kuinka FRR-tasojen laskiessa FAR-tasot kasvavat. Tätä suhdetta käytetään järjestelmän säätämiseen, jotta saavutetaan ideaali valinta, tunnistusherkkyydelle, joka määrää tavoitellut FAR- ja FRR-tasot. Käytännössä tämä toteutetaan molempien metriikoiden erillisillä testeillä.



KUVA 7. Tyypillinen FRR-tason suhde FAR-tasoon tunnistusherkkyyden muuttuessa. (Benchmarking a Wake Word Detection Engine. 2019.)

4 SUORITUSKYKYTESTIEN TULOKSET: FRR & FAR

Tässä luvussa käydään läpi suorituskykytesteistä saadut tulokset. Luvussa 4.1, 4.2 ja 4.3 vertaillaan laitteisto- (kuvissa merkittynä HW DSP) ja ohjelmistopohjaisen puheenkäsittelyratkaisun (kuvissa merkittynä SW DSP) suorituskykyä referenssinä järjestelmä, joka konfiguroitiin niin, että minkäänlaista puheenkäsittelyratkaisua ei ollut käytössä (kuvissa merkittynä lyhenteellä REF).

Lisänä näille suorituskykytuloksille on luvun 4.4 suorituskykytestien tulokset. Näissä testeissä vertailtiin ohjelmistopohjaisen puheenkäsittelyratkaisun tuloksia, joissa käytössä oli kahden mikrofonin lineaarinen mikrofonimatriisi (kuvissa merkittynä n=2), ohjelmistopohjaisen puheenkäsittelyratkaisun tuloksiin, joissa oli käytössä kolmen mikrofonin pyöreä mikrofonimatriisi (kuvissa merkittynä n=3). Kaikissa FRR-testeissä käytettiin luvussa 3.3 esiteltyä testiraitaa, jonka näytteiden (herätyssanojen) lukumäärä oli kokonaisuudessaan 280. Jokainen FRR-prosentti tulos siis koostuu testistä, jossa on sanottu 280 kertaa herätyssana ”Hello Home”.

Testattavista laitteista huomioitavaa on, että molemmat laitteet ovat olleet testattavana täysin samaan aikaan ja näin ollen niiden akustiset olosuhteet ovat olleet täysin identtiset testien aikana. Lisäksi huomioitavaa on, että referenssijärjestelmän testit on tehty eri aikaan, sillä testattavia laitteita ei ollut kolmea kappaletta.

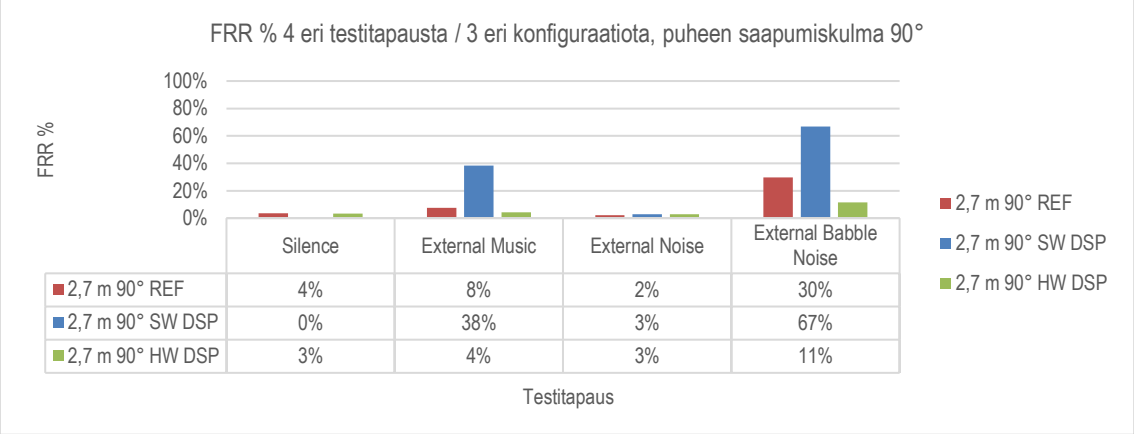
4.1 Laitteisto- vs. ohjelmistopohjainen puheenkäsittelyratkaisu FRR % (puheen saapumiskulma 90° testattavan laitteen näkökulmasta)

Kuvissa 8 ja 9 on esitetty FRR-testien tulokset testikonfiguraatiolla, jossa puhujalähde-kaiutin on asetettu 90°:n saapumiskulmaan testattavan laitteen näkökulmasta. Huomattava heikkenemä suorituskyvyssä paljastui 4. testitapauksessa, ”External Babble Noise”. Sinällään tämä on ymmärrettävää, sillä tiedetään jo luvusta 3, että skenaariona tämä on haastava puheentunnistusjärjestelmälle.

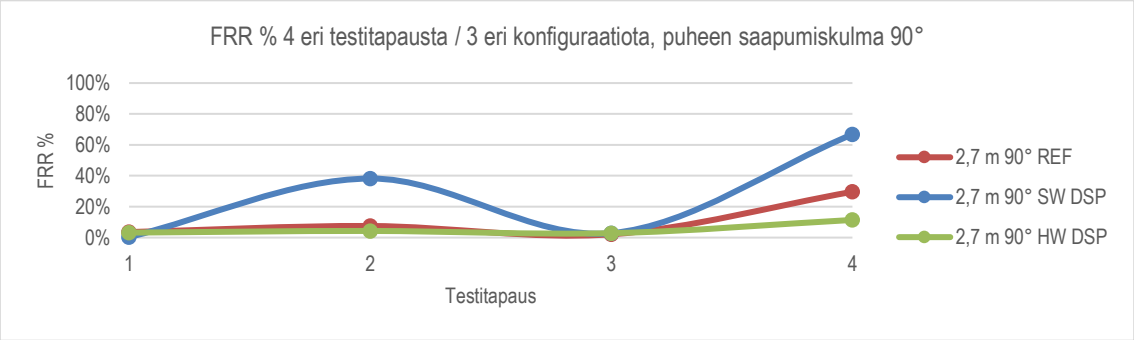
Ohjelmistopohjainen ratkaisu (SW DSP) suoriutui yllättävän oudolla tavalla dynaamista sisältöä sisältävissä akustisissa skenaarioissa tuloksilla ”External Music” (38 %) ja ”External Babble Noise” (67 %). Näissä testitapauksissa suorituskyky ei laskenut moninkertaisesti ainoastaan helpoimman

akustisen skenaarion tuloksen ”Silence” (0 %) alle, vaan myös verrattuna samojen testitapausten referenssijärjestelmän tuloksille. Näistä havainnoista voisi siis päätellä jonkun olevan pielessä ohjelmistopohjaisessa ratkaisussa. Syy selvisi kuitenkin mikrofonikonfiguraation geometrisesta asettelusta. Kaksi mikrofonia ei kykene luomaan kokonaista 360°:n ympyrää ja suoriutuu epävakaasti, sillä se ei pysty ylläpitämään monitorointia puhujalähteen sijainnista huoneen avaruudessa, vaan välillä menettää sen. Tämä menetys aiheuttaa sen, että keilanmuodostusalgoritmi ei kykene säännöllisesti vaimentamaan puheääneen kuulumattomien melulähteiden äänenpainetasoja ja säilyttämään puheen äänenpainetasoja.

Laitteistopohjainen ratkaisu suoriutui odotetulla tavalla kaikissa tapauksissa yltäen pahimmillaan tulokseen ”External Babble Noise” (11 %). Tulokset olivat odotusten mukaisia, sillä tulokset paranivat lähes kaikissa testitapauksissa referenssijärjestelmän tuloksiin verrattuna.



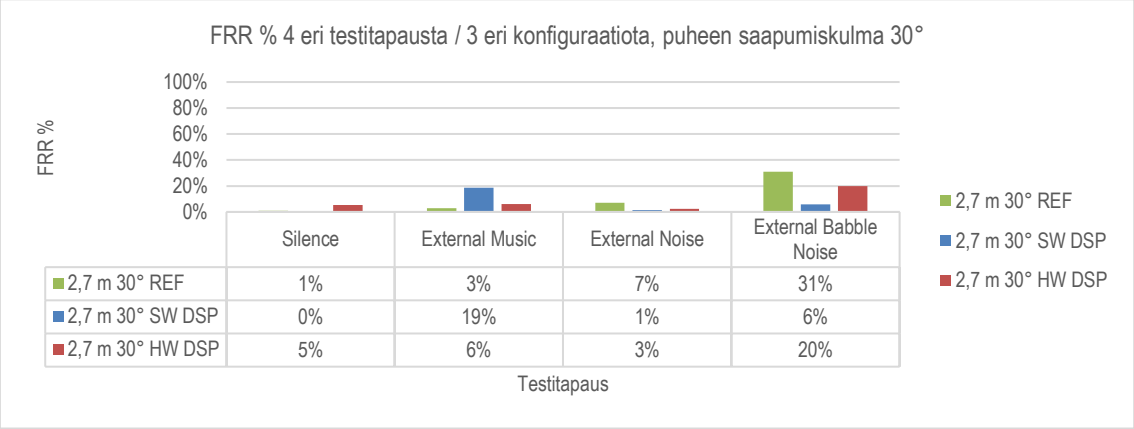
KUVA 8. FRR % neljässä eri testitapauksessa kolmella eri konfiguraatiolla: 90° REF, 90° SW DSP ja 90° HW DSP pylväskaaviolla kuvattuna.



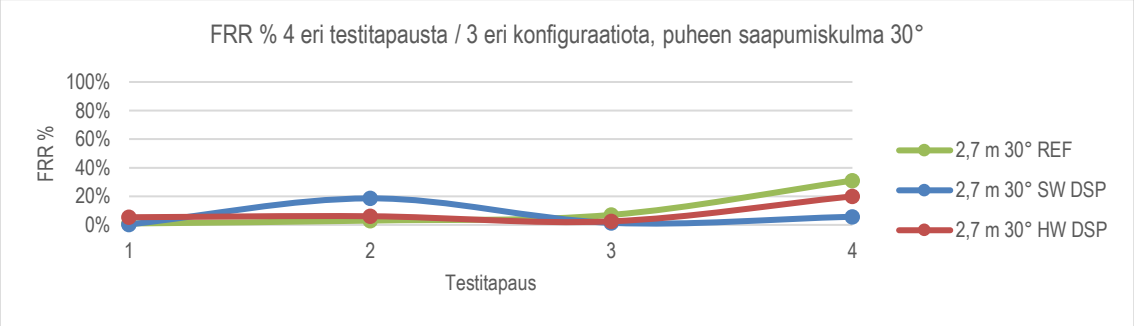
KUVA 9. FRR % neljässä eri testitapauksessa kolmella eri konfiguraatiolla: 90° REF, 90° SW DSP ja 90° HW DSP pistekaaviolla kuvattuna.

4.2 Laitteisto- vs. ohjelmistopohjainen puheenkäsittelyratkaisu FRR % (puheen saapumiskulma 30° testattavan laitteen näkökulmasta)

Kuvien 10 ja 11 tuloksissa testit olivat järjestelty siten, että puheen saapumiskulma testattaviin laitteisiin nähden oli 30°. Näissä tuloksissa oli myös yllättäviä havaintoja, kuten 90°:n ohjelmistopohjaisen ratkaisun tuloksissakin. Referenssijärjestelmän (kuvassa kuvaaja REF) tulokset olivat lähes identtisiä 90°:n tuloksiin verrattuna. Yllätykselliset havainnot ilmenivät ohjelmistoratkaisun tuloksissa tuloksella ”External Babble Noise” (6 %). Tämä tulos osoitti, että ohjelmistopohjaisella ratkaisulla oli potentiaalia kilpailla ei ainoastaan referenssijärjestelmän kanssa, vaan myös laitteistopohjaisen ratkaisun järjestelmän kanssa. Kuitenkin tulos ”External Music” (19 %) jälleen muistutti siitä, kuinka epävakaa ohjelmistopohjainen ratkaisu olikaan. Viimeinen havainto oli, että laitteistopohjaisen ratkaisun (HW DSP) suorituskyky heikkeni huomattavasti, joka nähtiin tuloksessa ”External Babble Noise” (20 %) verrattuna 90°:n tulokseen ”External Babble Noise” (11 %).



KUVA 10. FRR % neljässä eri testitapauksessa kolmella eri konfiguraatiolla 30° REF, 30° SW DSP ja 30° HW DSP pylväskaaviolla kuvattuna.



KUVA 11. FRR % neljässä eri testitapauksessa kolmella eri konfiguraatiolla 30° REF, 30° SW DSP ja 30° HW DSP pistekaaviolla kuvattuna.

4.3 Laitteisto- vs. ohjelmistopohjainen puheenkäsittelyratkaisu FAR

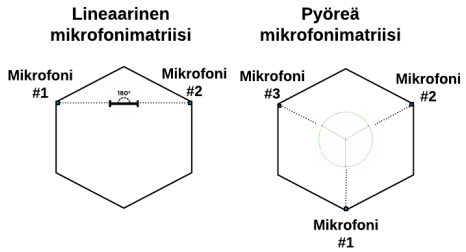
Taulukon 1 FAR-tulokset olivat odotettavia tunnistusherkkyyden säätämisen testien tulosten perusteella. Taulukon luvut kuvaavat sitä lukumäärää, kuinka monta kertaa herätyssana tunnistettiin virheellisesti. Virheellisyys siis perustuu siihen, että herätyssanaa ”Hello Home” ei sanottu 12 h nauhoituksessa, mutta se oli silti tunnistettu. Johtopäätöksiä yhden lukumäärän erosta ei voida varsinaisesti tehdä. Ymmärretään kuitenkin, että tunnistusherkkyydellä on merkittävä vaikutus FAR-tuloksiin ja tämä tulos jälleen verifioi tunnistusherkkyyden säädön pätevyyden.

TAULUKKO 1. Referenssijärjestelmän (REF) ja ohjelmistopohjaisen (SW DSP) sekä laitteistopohjaisen (HW DSP) ratkaisun FAR-tulokset 12 h kestoisessa testissä, joista nähdään virheellisesti tunnistettujen herätyssanojen lukumäärä.

Testin tyyppi	Konfiguraatio	Tunnistetut herätyssanat
FAR 12 h	REF	1
FAR 12 h	SW DSP	0
FAR 12 h	HW DSP	0

4.4 Ohjelmistopohjainen puheenkäsittelyratkaisu FRR %: 2 mikrofonin konfiguraatio vs. 3 mikrofonin konfiguraatio

Viimeisinä testeinä haluttiin selvittää, kuinka lineaarisen mikrofonimatriisin vaihtaminen (ULA, Uniform Linear Array) pyöreään mikrofonimatriisiin (UCA, Uniform Circular Array) vaikuttaa FRR-testien tuloksiin. Lineaarinen matriisi koostui vaakatasossa linjassa asetelluista mikrofoniparista, joiden etäisyys toisistaan oli 50 mm. Pyöreä mikrofonimatriisi taas muodosti kolmen mikrofonin välillä tasasivuisen kolmion ympyrän kehällä. Näiden mikrofonien etäisyys ympyrän keskelle oli 35 mm. Kuvassa 12 esiteltynä vasemmalla lineaarinen ja oikealla pyöreä mikrofonimatriisi.

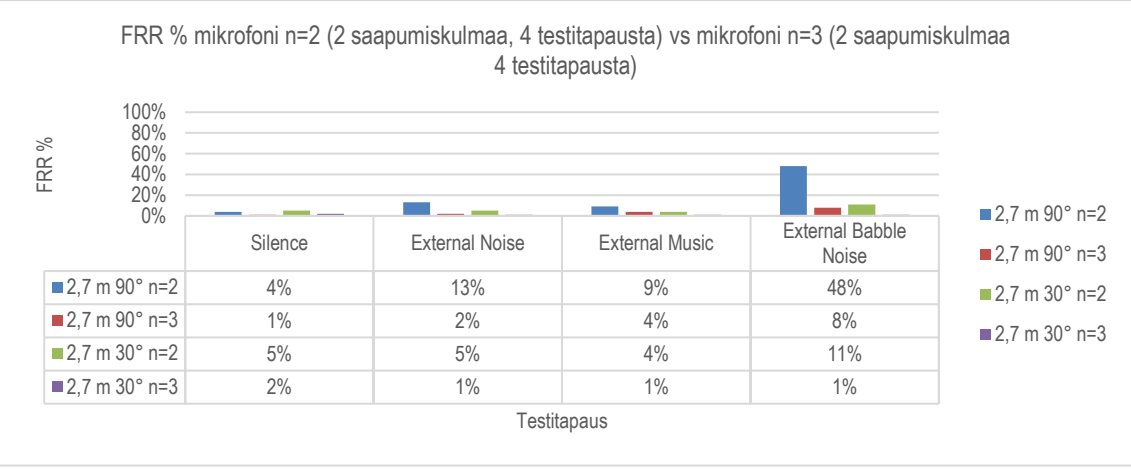


KUVA 12. Kahden mikrofonin lineaarinen mikrofonimatriisi ja kolmen mikrofonin pyöreä mikrofonimatriisi lintuperspektiivissä.

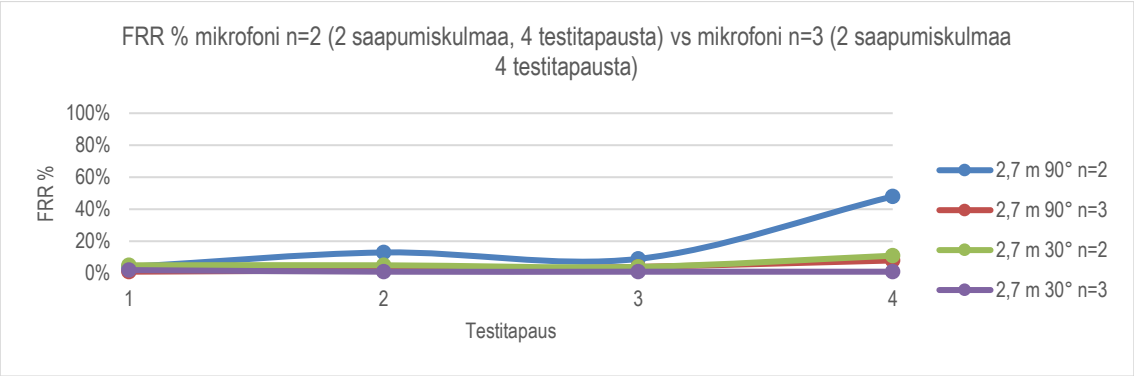
Vaihtamalla lineaarisesta mikrofonimatriisista pyöreään mikrofonimatriisiin päästään irti riippuvuudesta saapumiskulmaan ja saavutetaan kokonaisen 360°:n kattava tasainen tilallinen selektiivisyys missä tahansa ulottuvuudessa, joka on yhdensuuntainen mikrofonimatriisin kanssa. Tämä johtuu mikrofonimatriisin likimääräisestä pyörimisen vakioisuudesta. Tällöin teoriassa pystytään ylläpitämään lähteiden monitorointia tasaisesti ja säännöllisesti, jonka kautta voidaan säännöllisemmin vaimentaa melulähteiden äänenpainetasoja ja säilyttää puheenlähteen äänenpainetasoja. (Virtanen ym. 2018, luku 10.2)

Suorituskykytestien suoritettiin seuraavilla testikonfiguraatioilla: kahden mikrofonin lineaarisella mikrofonimatriisilla 90°:en ja 30°:en saapumiskulmilla (kuvaajat 2,7 m 90° n=2 ja 2,7 m 30° n=2) ja kolmen mikrofonin pyöreällä mikrofonimatriisilla 90°:en ja 30°:en saapumiskulmilla (kuvaajat 2,7 m 90° n=3 ja 2,7 m 30° n=3).

Kuvien 13 ja 14 havainnollistamat tulokset olivat hypoteesin oletuksien mukaisia. Ne ylsivät pahimmillaan kolmen mikrofonin konfiguraatiolla ja 90°:n saapumiskulmalla (kuvassa kuvaaja 2,7 m 90° n=3) tulokseen ”External Babble Noise” (8 %), joka oli moninkertaisesti parempi tulos verrattuna kahden mikrofonin tulokseen (kuvassa kuvaaja 2,7 m 90° n=2) ”External Babble Noise” (48 %). Tulokset tukevat hypoteesia, jonka mukaan kolmen mikrofonin pyöreä mikrofonimatriisi mahdollistaa kyseisellä keilanmuodostusalgoritmilla paremman selkeyden puheäänelle ja tällöin puheentunnistuksen hylkäykset vähenevät.



KUVA 13. FRR % neljässä eri testitapauksessa neljällä eri testikonfiguraatiolla 90° n=2, 90° n=3, 30° n=2 ja 30° n=3 pylväskaaviolla kuvattuna.



KUVA 14. FRR % neljässä eri testitapauksessa neljällä eri testikonfiguraatiolla 90° n=2, 90° n=3, 30° n=2 ja 30° n=3 pistekaaviolla kuvattuna.

5 SUORITUSKYKYTESTIEN TULOSTEN PÄÄTELMÄT

Kootusti näistä suorituskykytuloksista nähdään, että kahden mikrofonin ohjelmistopohjainen puheenkäsittelyratkaisu ei suoriudu tällä keilanmuodostusmenetelmällä joissain akustisissa skenaarioissa vakaasti. Laitteistopohjainen ratkaisu toimii vakaasti ja on edelleen luotettava ratkaisu, jota voidaan käyttää tuotteissa, mutta johdannossa esitellyn rajoituksen vuoksi tavoitellaan ohjelmistopohjaista ratkaisua.

Kolmen mikrofonin suorituskykytuloksista nähdään, että ohjelmistopohjainen puheenkäsittelyratkaisu kykenee kuitenkin toimimaan pyöreällä mikrofonimatriisilla tavoitellusti ja vakaasti sekä jopa päihittämään laitteistopohjaisen ratkaisun järjestelmän suorituskyvyn. Kuitenkin kolmen mikrofonin ratkaisu tuo oman rajoituksensa tuotteen suunnitteluun sen mekaaniselta puolelta, sillä mikrofonit on sijoitettava tasaisesti esimerkiksi 35 mm:n säteellä tasasivuisen kolmion keskipisteestä. Yksikkökustannusten hinta myös nousee osakustannusten osalta mikrofonien lukumäärän kasvaessa.

Ohjelmistopohjaisten ratkaisujen käyttämistä haluttaisiin edellä mainittujen rajoitusten vuoksi käyttää tuotteessa. Jatkotoimintana tämän opinnäytetyön tutkimuksen päätyttyä esitetään, että sovelletaan toisenlaista keilanmuodostusalgoritmia kahden mikrofonin lineaarisen mikrofonimatriisin järjestelmässä.

6 YHTEENVETO

Opinnäytetyössä tehtiin laitteisto- ja ohjelmistopohjaisen puheenkäsittelyratkaisun suorituskykyvertailu FRR- ja FAR-testein puheohjauskäyttöliittymän tehtävissä. Näiden testien järjestely ja valmistelu käytiin läpi. Lisäksi tehtiin suorituskykytestit ohjelmistopohjaisella ratkaisulla, jossa vertailtiin lineaarisen kahden mikrofonin mikrofonimatriisin suorituskykyä pyöreän kolmen mikrofonin mikrofonimatriisin suorituskykyyn. Tehtiin myös katsaus puheenkäsittelyalgoritmeihin, joita tyypillisesti nähdään puheenkäsittelyratkaisuissa.

Opinnäytetyön prosessi kauttaaltaan lopulta onnistui tavoitteissaan, mutta vaati yllättävän paljon resursseja, ennen kuin suorituskykytestit voitiin tehdä. Esimerkiksi testiaineiston puhuttujen herätyssanojen lukumäärä kerrottiin kymmenellä aloitettuun määrään verrattuna. Lisäksi yllättäviä vaihteluita tuloksissa esiintyi, joiden syytä ei saatu selville suoraan, vaan jouduttiin tutkimaan teoriaa ja tekemään lisätestit, jotta saatiin todisteet esitetylle hypoteesille.

Saatuja tuloksia pystytään hyödyntämään tuotekehityksessä varsinkin tulevien ohjelmistopohjaisten puheenkäsittelyratkaisujen soveltamisessa puheohjauskäyttöliittymissä. Ymmärrys sekä laitteisto- että ohjelmistopohjaisten puheenkäsittelyratkaisujen far field suorituskyvystä saavutettiin, joiden avulla voidaan paremmin suunnitella ja kehittää tuotteita sekä säätää niiden konfigurointeja.

Opinnäytetyö oli omasta näkökulmasta erittäin opettava ja uudenlainen kokemus, jota en ollut aiemmin saanut. Tutkimuksen tekemisen käytänteet harjaantuivat säännöllisesti ja vaativat kriittistä ajattelukykyä ja säntillisyyttä. Saatujen testien tuloksista tehtiin jatkuvasti uusia hypoteeseja lähes viikoittain ja näitä hypoteeseja analysoitaessa Markku Heiskarin ja Heikki Juntusen kanssa sain usein muistutuksen siitä, että hypoteeseista tehdyt päätelmät olivat virheellisiä. Tällöin jouduttiin joko tekemään uusia testejä varmistaaksemme hypoteesin päätelmät tai analysoimaan hypoteesi uudelleen ja luomaan uusia päätelmiä. Isot kiitokset koko Creoirin väelle mahtavasta ilmapiiristä ja ainutlaatuisen opettavasta kokemuksesta sekä yritysprojekteissa, että opinnäytetyössä.

LÄHTEET

Alexa Acoustic Testing Guide. 2020. Amazon. Saatavissa: https://developer.amazon.com/en-US/docs/alexa/alexa-voice-service/acoustic-testing-guide.html. Hakupäivä 6.11.2020.

Benchmarking a Wake Word Detection Engine. 2019. Picovoice. Saatavissa: https://picovoice.ai/blog/benchmarking-a-wake-word-detection-engine/. Hakupäivä 6.11.2020.

GSC Beamformer. 2020. MathWorks. Saatavissa: https://se.mathworks.com/help/phased/ref/gscbeamformer.html#d122e275602. Hakupäivä 6.11.2020

Habets, Emanüel A.P – Naylor, Patrick A. 2018. 15 Dereverberation. Teoksessa Vincent, Emmanuel – Virtanen, Tuomas – Gannot, Sharon (toim.). Audio Source Separation and Speech Enhancement. Hoboken, NJ: John Wiley and Sons. Saatavissa: https://learning.oreilly.com/library/view/audio-source-separation/9781119279891/c15.xhtml. Hakupäivä 7.11.2020.

Hyvärinen, Aapo – Karhunen, Juha – Oja, Erkki 2001. Independent Component Analysis. New York: John Wiley & Sons, Inc. Saatavissa: https://www.cs.helsinki.fi/u/ahyvarin/papers/bookfinal_ICA.pdf. Hakupäivä 6.11.2020.

Markovich-Golan, Shmulik – Kellermann, Walter – Gannot, Sharon 2018. 10 Spatial Filtering. Teoksessa Vincent, Emmanuel – Virtanen, Tuomas – Gannot, Sharon (toim.). Audio Source Separation and Speech Enhancement. Hoboken, NJ: John Wiley and Sons. Saatavissa: https://learning.oreilly.com/library/view/audio-source-separation/9781119279891/c10.xhtml. Hakupäivä 7.11.2020.

Virtanen, Tuomas – Vincent, Emmanuel – Gannot, Sharon 2018. 2 Time-Frequency processing: Spectral Properties. Teoksessa Vincent, Emmanuel – Virtanen, Tuomas – Gannot, Sharon (toim.). Audio Source Separation and Speech Enhancement. Hoboken, NJ: John Wiley and Sons. Saatavissa: https://learning.oreilly.com/library/view/audio-source-separation/9781119279891/c10.xhtml. Hakupäivä 7.11.2020.

Watanabe, Shinji – Virtanen, Tuomas – Kolossa Dorothea 2018. 17 Application of Source Separation to Robust Speech Analysis and Recognition. Teoksessa Vincent, Emmanuel – Virtanen, Tuomas – Gannot, Sharon (toim.). Audio Source Separation and Speech Enhancement. Hoboken, NJ: John Wiley and Sons. Saatavissa: https://learning.oreilly.com/library/view/audio-source-separation/9781119279891/c15.xhtml. Hakupäivä 7.11.2020.