

TILASTOLLISTEN MALLIEN AVULLA TEHOKKAAMPIA KUUNTELUKOKEITA

Petteri Hyvärinen^{1,2}

¹ Festnets Oy, Helsinki

petteri.hyvarinen@iki.fi

² Aalto-yliopiston sähkötekniikan korkeakoulu

Informaatio- ja tietoliikennetekniikan laitos

Otakaari 5, 02150 Espoo

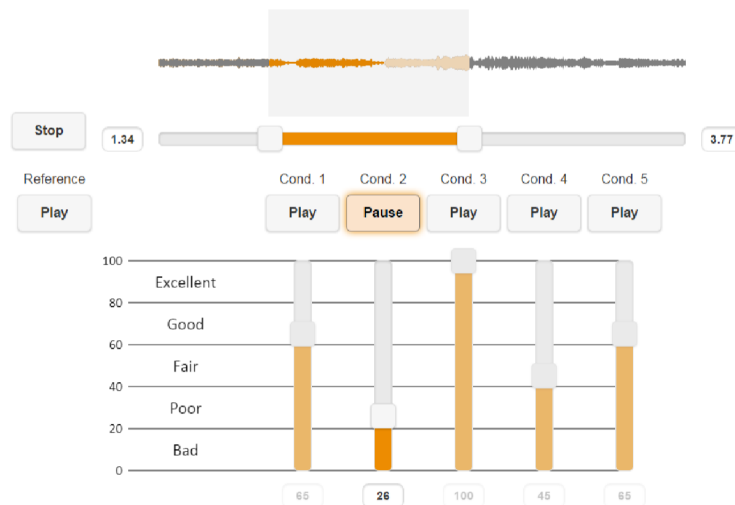
Tiivistelmä

Kuuntelukokeet ovat olennainen osa audio- ja akustiikkajärjestelmien kehitystyötä. Kuuntelukokeiden tulosten tilastollinen analyysi näyttyy kuitenkin toisinaan epäintuitiivisena, välttämättömänä pahana, josta saatava hyöty jää monesti akateemiselle tasolle. Tässä esitelmässä tarkastellaan nk. ranking & selection -menetelmiä, jotka mahdollistavat psykoakustiikan ja tuotekehityksen kannalta mielekkäämpiä kysymyksenasetteluita, sekä tulosten tulkinnan kannalta intuitiivisempia muotoiluja. Esimerkkinä käytetään MUSHRA -testin (ITU-R BS.1534-3) tulosten tilastollista mallintamista. Käyttämällä tarkoituksenmukaisempia tilastollisia malleja, kuuntelukokeissa tarvittavia kuuntelijamääriä ja/tai ääninäytteiden toistokertoja on mahdollista kohdentaa tarkemmin ja keventää kuuntelukokeisiin vaadittavia resursseja. Eri vaiheissa tutkimus- ja tuotekehityskaarta kuuntelukokeiden tehostamiselle voi olla erilaisia tarpeita; alussa seulotaan potentiaalisimmat ehdokkaat ja loppupäässä keskitytään vain lupaavimpiin vaihtoehtoihin. Asianmukaisella tilastollisella mallintamisella voidaan kuuntelukokeista saada paras mahdollinen hyöty.

1 JOHDANTO

Kun kuuntelukokeita käytetään aistinvaraisen arvioinnin välineenä, tavoitteena on usein arvioitavien järjestelmien (esimerkiksi vaihtoehtoisten signaalinkäsittelymenetelmien, audiokoodekkien, laitteistokokonaisuuksien tai kaiutinasetteluiden) asettaminen keskinäiseen paremmuusjärjestykseen. Kuuntelukokeiden tuloksissa esiintyy kuitenkin aina epävarmuutta; eri kuuntelijat saattavat antaa hieman erilaisia vastauksia, tai jopa sama kuuntelija voi toistetussa kokeessa antaa eri koekerroilla eri vastauksen, ja tämä epävarmuus huomioidaan tulosten tulkintavaiheessa tilastollisten menetelmien ja mallien avulla. Harmillisesti, ja ehkä jopa yllättävästi, laajasti käytössä olevat tilastollisiin testeihin perustuvat klassiset menetelmät eivät kuitenkaan sovellu kovin hyvin aistinvaraisen äänenlaadun arvioinnin kaltaisiin asetelmiin, vaan ne pahimmillaan pakottavat epäintuitiivisiin muotoiluihin ja hankaloittavat tulosten tulkintaa [1]. Vaihtoehtona näille tyypillisille menetelmille, tässä artikkelissa esitellään nk. *ranking & selection* (R&S)

Copyright ©2025 Petteri Hyvärinen. Tämä on avoimesti julkaistu teos, joka noudattaa Creative Commons NIMEÄ 4.0 Kansainvälinen –lisenssiä (CC BY 4.0). Teosta saa kopioida, levittää, näyttää ja esittää julkisesti ja siitä saa luoda johdannaisteoksia, kunhan tekijän nimi ja lähde mainitaan asianmukaisesti.



Kuva 1: webMUSHRA-käyttöliittymä [3], jossa järjestelmien pisteytys on toteutettu liukuvalitsimilla. Kutakin sokkoutettua järjestelmää voi kuunnella yksi kerrallaan ja verrata referenssiin. Ääninäytteen kuuntelua voi myös rajoittaa tiettyihin osioihin.

-lähestymistapa yhdistettynä Bayesilaiseen tulosten mallintamiseen ja kuuntelukokeiden optimointiin.

Esimerkkinä käytetään MUSHRA-testin tulosten tilastollista mallintamista. ITU-R BS.1534-3 suositus [2] määrittelee “**MU**lti **S**timulus test with **H**idden **R**eference and **A**nchor (**MUSHRA**)” -menetelmän, joka on tarkoitettu keskilaatuisten (engl. *intermediate quality*) järjestelmien kuuntelukokeisiin. Testissä vertailtavana ovat korkealaatuinen referenssi, huonolaatuinen ankkuri, sekä varsinaiset vertailtavana olevat audiojärjestelmät, jotka kuuntelijat pisteyttävät arviontiskaalalla 0–100, jossa 100 vastaa referenssijärjestelmän laatua ja 0 vastaa huonolaatuisen ankkurin laatua. Esimerkkinä testin toteutukselta vaadittavista toiminnallisuuksista, kuvassa [1] selainpohjainen webMUSHRA-käyttöliittymä [3].

Tulosten tilastollinen mallintaminen etenisi tämän jälkeen tyypillisesti varianssianalyysin kautta (engl. *analysis of variance*, ANOVA). Varianssianalyysi on esimerkki klassisesta tilastotieteellisestä menetelmästä, joka perustuu tilastollisiin hypoteeseihin ja niiden testaamiseen; ANOVA:n tapauksessa muodostettaisiin nollahypoteesi tutkittavien järjestelmien saamien arvioiden keskiarvoista $H_0 : \mu_1 = \mu_2 = \dots = \mu_N$, joka vastaa tilannetta jossa kaikki vertailtavat järjestelmät MUSHRA-testissä kuulostaisivat keskimäärin täysin samoilta. Kuuntelukokeiden tulokset tulkitaan näyttönä nollahypoteesia vastaan, vastahypoteesin $H_1 : \mu_i \neq \mu_j$ joillekin $i \neq j$ puolesta, ja mikäli näyttö on valitun tilastollisen testin perusteella riittävän vahvaa, hylätään H_0 ja hyväksytään H_1 . Tässä vaiheessa ollaan toivon mukaan saatu tukea sille tulkinnalle, että vähintäänkin yksi vertailtavista järjestelmistä *ei* kuulosta keskimäärin täysin samalta kuin muut. Mutta edelleenkin ei ole tiedossa, mikä järjestelmä tai mitkä järjestelmät ovat kyseessä, vaan tuota kysymystä varten täytyy tehdä ns. *post-hoc* testejä, ja niiden myötä muodostaa uusia nollahypoteeseja, vastahypoteeseja ja tilastollisia testejä. Siis vaikka H_0 voitaisiinkin hylätä, tarkkaan ottaen ei voida vielä sanoa mitään kovin kiinnostavaa H_1 :stä; ainoastaan että kyseiset kuuntelukoetulokset olisivat hyvin epätodennäköisiä (esim. $p < 0.05$) H_0 :n

mukaisessa tilanteessa. Tällaisen tuloksen informaatioarvo tuntuu kuitenkin häviävän pieneltä, kun otetaan huomioon että käytännössä jo kuuntelukoetta suunniteltaessa on selvää että järjestelmien välillä on ainakin *jonkinlaisia* havaittavia eroja, ja kuuntelukokeen tarkoituksena on ennemminkin systemaattisesti kvantifioida nämä erot kuin selvittää, löytyykö eroja lainkaan. Nollahypoteesi vaikuttaa siis hyvin keinotekoiselta, ja sitä myöten hankalalta lähtökohdalta tulosten tulkinnalle. Kärjistäen voisikin sanoa, että tilastolliset testit antavat paljon vastauksia sellaisiin kysymyksiin, joihin MUSHRA-testin tulosten tulkitsija ei vastauksia kaipaa.

Ranking & Selection (R&S) -menetelmät sen sijaan keskittyvät vastaamaan sellaisiin kysymyksiin, kuten “mikä järjestelmä on keskimäärin paras, ja millä todennäköisyydellä?”, “mitkä järjestelmät ovat 97%:n todennäköisyydellä keskimäärin enintään 20:n pisteen päässä referenssistä?” tai “millä todennäköisyydellä meidän menetelmämme sijoittuu top-3:een?”. [4, 5, 6] R&S-menetelmissä lähtökohdaksi otetaan se, että järjestelmien välillä on havaittavia eroavaisuuksia, ja tilastollisen tarkastelun tavoitteena on määrittää kuuntelukokeiden tulosten perusteella, millä varmuudella järjestelmiä voidaan laittaa paremmuusjärjestykseen (*ranking*) tai muodostaa niistä haluttujen kriteerien mukaisia osajoukkoja (*selection*). Vaikka tämänkaltainen näkökulman vaihto saattaa vaikuttaa kosmeettiselta, sillä on merkittäviä vaikutuksia tilastollisiin malleihin ja käytännön analyysiin. Esimerkiksi, jos fokus asetetaan nimenomaan vain parhaan järjestelmän löytämiseen, voidaan huomommin sijoittuvan järjestelmän testaaminen lopettaa heti, kun käy riittävän selväksi (ts. todennäköiseksi), että se ei ainakaan tule olemaan paras järjestelmä vaikka datan keruuta jatkettaisiin kuinka kauan — mutta varsinaisesti sillä ei ole väliä, sijoittuisiko kyseinen järjestelmä viimeiseksi vai toiseksi viimeiseksi, eikä siis tämän asian selvittämiseen ole syytä kuluttaa kuuntelukooeresursseja. Tässä artikkelissa yhdistetään R&S-lähestymistapa bayesilaiseen mallintamiseen. MUSHRA-testistä saatavat laatuarviot mallinnetaan bayesilaisena hierarkkisena mallina, jonka parametrien jakaumista lasketaan R&S-tyyppisiä todennäköisyyksiä testattaville järjestelmille.

Bayesilaisten mallien hyödyt eivät rajoitu yksittäisen kuuntelukokeen tulosten analyysiin tai koejärjestelyiden optimointiin. Esimerkiksi suurten kielimallien (LLM, *large language model*) kehittämisessä oleellinen vaihe on kielimallin vastausten sovittaminen ihmisarvioijien preferensseihin (ns. *alignment*-vaihe), jossa hyödynnetään vahvistusoppimista (RLHF, *reinforcement learning from human feedback*). [7] Vastaavaa lähestymistapaa on käytetty myös audiosovelluksissa, esimerkkinä Cideron *et al.* (2024) [8], jossa n. 300 000:n parivertailun tuloksia hyödynnettiin *text-to-music*-järjestelmän opettamiseen. Viime vuosina myös bayesilaisia malleja on hyödynnetty RLHF:n optimointipalkkion mallintamisessa — arvioitsijoiden vastausten ja preferenssien perusteella muodostetaan siis bayesilainen malli, jota käytetään varsinaiseen neuroverkon opettamiseen yksittäisten vastausten sijaan. [9]

2 METODI

Tarkastellaan erään tilaänen laatua tutkineen kuuntelukokeen tuloksia, jossa viisitoista kuuntelijaa $k = 1 \dots 15$ on arvioinut kuusi järjestelmää $j = 1 \dots 6$ (joista järjestelmä 1 on referenssi ja järjestelmä 6 ankkuri) ja antanut niille arviot $r_{kj} \in [0, 100]$. Skaalataan arviot välille $[0, 1]$ ja mallinnetaan näin saadut arvot y_{kj} beta-jakautuneina satunnaislukuina.

Hierarkkinen malli on kokonaisuudessaan:

$$\begin{aligned}
 y_{kj} &\sim \text{Beta}(\mu_j \phi_j, (1 - \mu_j) \phi_j), \\
 \text{logit}(\mu_j) &= \mu_0 + \alpha_j, \\
 \mu_0 &\sim \mathcal{N}(0, 1), \\
 \alpha_j &\sim \mathcal{N}(0, \sigma_\alpha^2), \\
 \sigma_\alpha^2 &\sim \text{Inv-Gamma}(3, 1), \\
 \phi_j &\sim \text{Gamma}(2, 0, 1),
 \end{aligned} \tag{1}$$

jossa μ_j on järjestelmäkohtainen beta-jakauman keskiarvo välillä $[0, 1]$. μ_0 ja α_j ovat normaalijakautuneita muuttujia reaaliplanalla, jotka liitetään ns. linkkifunktiona toimivan $\text{logit}(\cdot)$ -funktion avulla välille $[0, 1]$. ϕ_j määrittää beta-jakauman “tarkkuuden”, eli käytännössä sen, kuinka tiiviisti jakauma on keskittynyt keskiarvon μ_j läheisyyteen. Korkea ϕ_j arvo tarkoittaa, että eri kuuntelijoiden antamat arviot järjestelmästä j ovat hyvin lähellä toisiaan; matala ϕ_j arvo puolestaan tarkoittaa, arviot ovat levittäytyneet laajemmin välille $[0, 1]$. σ_α^2 kuvaa järjestelmien keskiarvojen epävarmuutta logit-skaalalla.

Bayesiläisittain mallin (1) parametreja θ ovat μ_j , μ_0 , α_j ja ϕ_j , jotka määrittävät havaintojen $\mathbf{y} = \{y_{kj}\}$ ehdollisen todennäköisyyden $p(\mathbf{y}|\theta)$. Parametrien θ jakaumia kutsutaan priorijakaumiksi, ja ne puolestaan voivat riippua hyperparametreista η (esim. σ_α^2 mallissa (1)). Bayesiläisessä päättelyssä parametreille ja hyperparametreille johdetaan yhteinen posteriorijakauma $p(\theta, \eta|\mathbf{y})$ Bayesin kaavan avulla:

$$p(\theta, \eta|\mathbf{y}) \propto p(\eta)p(\theta|\eta)p(\mathbf{y}|\theta), \tag{2}$$

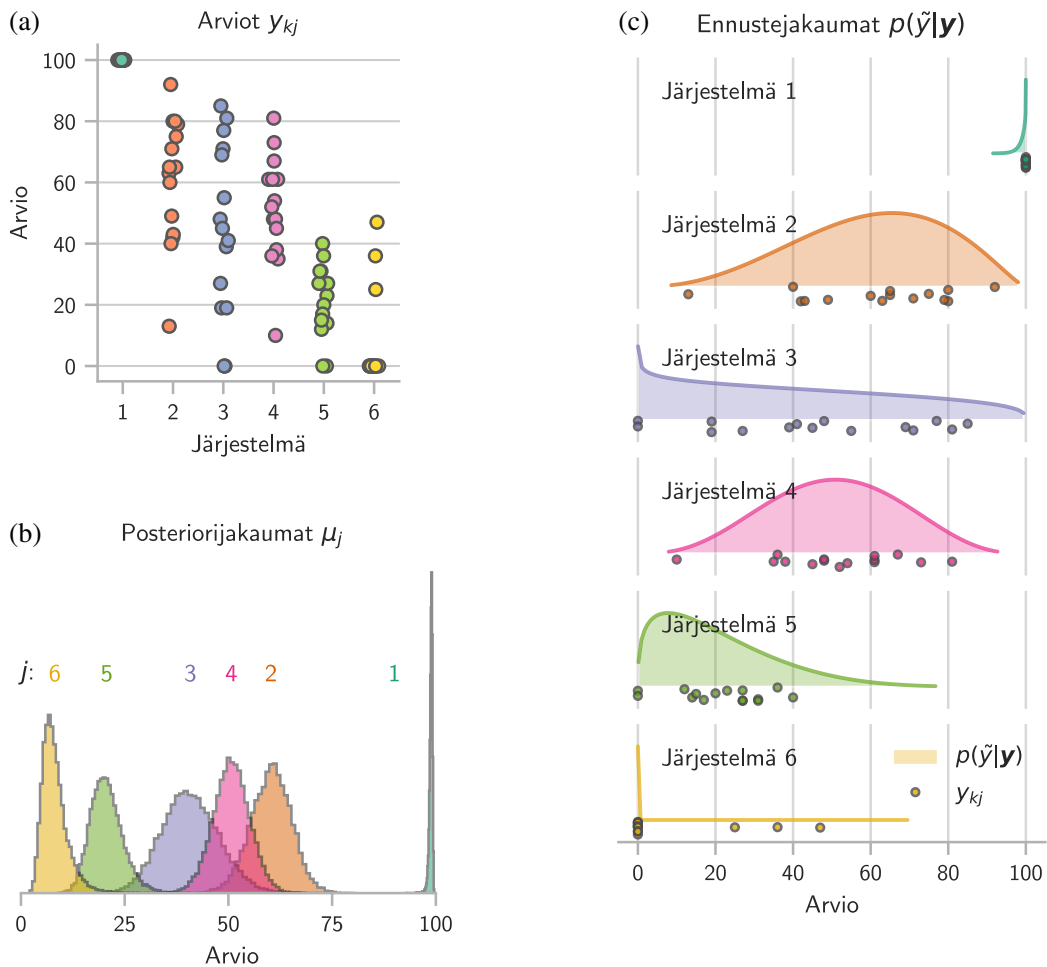
joka siis kuvaa, kuinka todennäköisiä eri θ ja η arvot ovat, havaintojen \mathbf{y} perusteella. “Priori” ja “posteriori”-termillä viitataan siihen, että priori-jakauma vastaa odotuksia ja oletuksia parametrien todennäköisistä arvoista *ennen* kuin ainuttakaan havaintoa on vielä tehty; posteriori-jakauma taas vastaa päivitettyä tietoa parametrien mahdollisista arvoista *sen jälkeen* kun tutkittavasta ilmiöstä on tehty havaintoja.

Malli (1) toteutettiin bayesiläiseen päättelyyn tarkoitetun STAN-ohjelmiston `CmdStanPy` Python-rajapinnalla. STAN käyttää MCMC-menetelmää (*Markov Chain Monte Carlo*) posteriorijakauman numeeriseen ratkaisemiseen — simuloimalla riittävä määrä näytteitä mallin mukaisesta posteriorijakaumasta, voidaan jakaumaa approksimoida halutulla tarkkuudella. Tällä menettelyllä voidaan mallin (1) parametreille määrittellä bayesiläisittain todennäköisyysjakaumat, ja näistä jakaumista laskea R&S-tunnuslukuja kuten esimerkiksi millä todennäköisyydellä järjestelmä 1 on parempi kuin järjestelmä 2 ($\Pr(\mu_1 > \mu_2)$), tai millä todennäköisyydellä järjestelmä 3 on paras kaikista ($\Pr(\mu_3 > \max_{j \neq 3} \mu_j)$).

Jo kerättyjen tulosten analysoinnin lisäksi bayesiläisellä mallilla on myös mahdollista ennakoita miten tulevat havainnot voisivat muuttaa parametrien jakaumia, ja valita seuraavaksi testattavat järjestelmät siten, että niistä saatava lisätieto olisi kuuntelukokeen tutkimuskysymyksen kannalta optimaalinen. Esimerkiksi, jos tavoitteena on tunnistaa vertailtavista järjestelmistä paras, ja kuuntelukokeen hetkellä t kerättyjen vastausten \mathbf{y} perusteella näyttäisi siltä, että järjestelmä j^* on paras, ja siihen liittyvä todennäköisyys on $P_{CS}^t = \Pr(\mu_{j^*} > \max_{i \neq j^*} \mu_i | \mathbf{y})$. Tällöin, jos on mahdollista kerätä vielä yksi arvio \tilde{y}_{kj} lisää jostain järjestelmästä j , voidaan valita testata sitä järjestelmää, jonka tuloksen odotetaan eniten lisäävän tällä hetkellä parhaaksi arvioidun järjestelmän j^* todennäköisyyttä

olla paras järjestelmä $\max_j P_{CS}^{t+1}(j) = \max_j \Pr(\mu_{j^*} > \max_{i \neq j^*} \mu_i | (\mathbf{y}, \tilde{y}_{kj}))$. Tällaista yhden askeleen optimaalista valintaa kutsutaan kirjallisuudessa 0–1-loss tai knowledge gradient (KG_1) -lähestymistavaksi [10, 11]. Vaihtoehtoisesti voitaisiin optimoitavana kriteerinä pitää parhaan ja toiseksi parhaan järjestelmän keskiarvojen välistä marginaalia $\mathbb{E}[\mu_{[1]} - \mu_{[2]} | (\mathbf{y}, \tilde{y}_{kj})]$, jolloin testattaisiin sitä järjestelmää, joka oletettavasti kasvattaisi marginaalia eniten. Tätä optimointiperiaatetta kutsutaan LL_1 :ksi (LL : *linear loss*). [12] Bayesilaisella mallilla edellä mainitut optimointikriteerit lasketaan parametrien μ_j posteriorijakaumista, joiden oletettuja muutoksia voidaan arvioida simuloimalla näytteitä \tilde{y}_{kj} prediktiiivisestä jakaumasta $p(\tilde{y}_{kj} | \mathbf{y})$. KG_1 ja LL_1 kriteerit ovat vain yksittäisiä esimerkkejä ns. bayesilaisesta koesuunnittelusta (BED, *Bayesian experimental design*) [13].

3 TULOKSET



Kuva 2: MUSHRA-kuuntelukokeessa annetut arviot järjestelmille $j = 1 \dots 6$ kuvassa (a). Mallin [1] parametrien μ_j , eli eri järjestelmien keskiarvojen jakaumat kuvassa (b). Mallin [1] havaintojen ennustejakaumat eri järjestelmille kuvassa (c).

Kuvassa 2(a) ovat kuuntelukokeessa kerätyt arviot, eli mallin (1) havainnot $\mathbf{y} = \{y_{kj}\}$; 15 arviota kullekin 6:lle järjestelmälle. Kuvasta nähdään, että referenssijärjestelmä on saanut aina arvion 100, ja ankkurijärjestelmä kolmea poikkeusta lukuunottamatta arvion 0. Muiden järjestelmien arviot jakautuvat hyvin laajalle skaalalle välillä 0–100.

Kuvassa 2(b) on esitetty hierarkkisen mallin parametrien μ_j MCMC-näytteistettyjen, havainnoille y_{kj} ehdollisten, posteriori-jakaumien histogrammit, skaalattuna välille [0, 100]. Kuvaajasta nähdään, että referenssijärjestelmän 1 parametri μ_1 on hyvin tiiviisti keskittynyt skaalan yläpäähän, kun taas muiden järjestelmien μ_j parametreissa on enemmän epävarmuutta. Ankkuri 6 on kuitenkin selvästi huonoin ja järjestelmä 5 toiseksi huonoin; parametrien μ_6 ja μ_5 jakaumissa ei ole juuri lainkaan päällekkäisyyksiä toistensa tai muiden järjestelmien jakaumien kanssa. Kiinnostavimmat kysymykset liittyvätkin järjestelmien 2, 3 ja 4 väliseen keskinäiseen paremmuusjärjestykseen, eli parametrien μ_2 , μ_3 ja μ_4 suuruusjärjestykseen. Posteriorijakaumista voidaan laskea, että 91,7%:n todennäköisyydellä μ_2 :n arvo on korkein, ja μ_3 ja μ_4 todennäköisyydet sijoittua kolmikon kärkeen ovat vastaavasti 1,0% ja 7,3%.

Kuvassa 2(c) ovat parametrien μ_j ja ϕ_j posteriorijakaumien keskiarvojen mukaiset ennustejakaumat $p(\tilde{y}|\mathbf{y})$. Kuvaajassa ovat myös havainnot y_{kj} , jotta voidaan visuaalisesti arvioida, ovatko analyysin tuloksina saadut jakaumat uskottavia; epämuodollisen laadullisen tarkastelun perusteella vaikuttaa siltä että havainnot olisivat hyvinkin voineet muodostua näistä jakaumista, eikä todennäköisyysmassa ole keskittynyt esimerkiksi vain skaalan ääripäihin tai hyvin kapealle alueelle. Koska ennustejakaumissa otetaan huomioon sekä keskiarvo μ_j että jakauman tarkkuutta kuvaava parametri ϕ_j , huomataan, että vaikka esimerkiksi järjestelmän 3 keskiarvo asettuu järjestelmän 4 alapuolelle, järjestelmään 3 liittyvissä havainnoissa esiintyy huomattavasti enemmän hajontaa, ja käytännössä kaikki arviot välillä 0–100 saavat verrattain saman suuruusluokan todennäköisyyden. Järjestelmän 4 kohdalla taas arviot ovat selkeämmin keskittyneet keskiarvon ympärille.

Mikäli keskitytään vain järjestelmiin 2–5 (referenssijärjestelmä 1 on ilmiselvästi paras ja ankkuri 6 huonoin, kuten oletettua), simuloimalla mahdollisia uusia havaintoja ennustejakaumista KG_1 -optimaalinen valinta seuraavaksi testattavaksi järjestelmäksi olisi järjestelmä 3, jonka mittaamisen odotetaan parantavan todennäköisyyttä valita tällä hetkellä toiseksi sijoittuva järjestelmä 2 seuraavallakin askeleella toiseksi parhaaksi 0,36%:lla. LL_1 -optimaalinen valinta olisi taas järjestelmä 2, josta vielä yhden havainnon keräämisen odotetaan nostavan μ_2 :ta 0,47:llä (skaalalla 0–100).

4 DISKUSSIO

Tarkastelemalla bayesilaisen hierarkkisen mallin posteriorijakaumia R&S-näkökulmasta, toiseksi paras järjestelmä (referenssin jälkeen) voitiin tunnistaa 91,7%:n varmuudella. Jää tutkijan arvioitavaksi, onko tämä varmuustaso riittävä, vai tuleeko kuuntelukoetta jatkaa paremman varmuuden saavuttamiseksi. On myös mahdollista todeta, että tähän mennessä kerättyjen havaintojen perusteella järjestelmien arvioissa on liikaa hajontaa, eikä kokeen jatkamisesta todennäköisesti ole odotettavissa dramaattista muutosta isoon kuvaan.

Tulosten tulkinnan kannalta posteriorijakaumat antavat paljon monipuolisemman kuvan kuin yksittäiset tilastolliset testit — sen sijaan että yksi kerrallaan poissuljettaisiin

keinotekoisia nollahypoteeseja, voidaan tehdä esimerkiksi tuotekehityksen kannalta relevantteja kysymyksenasetteluita ja arvioida niiden toteutumisen todennäköisyyttä suoraan. Mikäli esimerkiksi kiinnostuksen kohteena ei olisikaan järjestelmien keskiarvojen paremmuusjärjestys, vaan se, kuinka suurella todennäköisyydellä kukin järjestelmä saisi yli 80:n arvioita, voitaisiin tätä arvioida ennustejakaumien kautta, ja tällä mittarilla järjestelmä 3:n voitaisiin olettaa suoriutuvan järjestelmää 4 paremmin, johtuen siitä että arviot järjestelmälle 3 näyttäivät menevät tasaisesti laidasta laitaan, verrattuna järjestelmän 4 skaalan keskivaiheille keskittyvään jakaumaan.

Kokeen jatkoon kannalta optimaalisen testauskohteen valinnassa, kummassakin tapauksessa odotettavissa olevat muutokset toiseksi parhaan järjestelmän suhteen olivat verrattain pieniä. Tämä heijastelee sitä tosiasiaa, että havainnoissa on paljon hajontaa, eikä yksittäisellä lisähavainnolla vielä liikauteta jakaumia merkittävästi. Kuitenkin, mikäli kuuntelukoea jatkettaisiin tästä pisteestä siten, että jokainen arvioitava järjestelmä valittaisiin KG_1 tai LL_1 periaatteella, olisivat tutkittavat järjestelmät parhaan järjestelmän tunnistamisen kannalta tehokkaampia valintoja (esimerkiksi ajankäytöllisesti) kuin kaikkien järjestelmien yhtäläinen testaaminen. Laajentamalla tätä näkökulmaa, voidaan suunnitella kuuntelukokeita, joissa testattavia järjestelmiä onkin viiden tai kahdeksan sijaan kymmeniä, ja jokaisessa testauksen vaiheessa valitaan kunkin kuuntelijan arvioitaviksi vain tietty, optimaalinen alijoukko kaikista testattavista järjestelmistä, mikäli tavoitteena olisi esimerkiksi tunnistaa yksittäinen paras järjestelmä tai top- k kärki testattavien järjestelmien joukosta.

Esitetty malli (I) on verrattain yksinkertainen, sillä se ottaa huomioon ainoastaan eri järjestelmien väliset erot ja tarkastelee vain yhdelle ääninäytteelle annettuja arvioita. Mallia voitaisiin laajentaa huomioimalla myös kuuntelijakohtaiset erot sekä mallintamalla vastauksia useampiin ääninäytteisiin — hierarkkiseen malliin siis lisättäisiin kuulijakohtaiset ja ääninäytekohtaiset termit. Tällöin optimointivaiheessa voitaisiin valita paras vaihtoehto kaikkien järjestelmien sekä ääninäytteiden joukosta.

5 YHTEENVETO

Analysoimalla MUSHRA-testin tulokset bayesilaiseen tilastolliseen päättelyyn perustavalla mallilla, voitiin mallin parametrien posteriorijakaumista sekä havaintojen ennustejakaumista laskea tulosten tulkinnan kannalta monipuolisesti hyödyllisiä indikaattoreita. R&S-lähestymistavalla muotoillut tutkimuskysymykset soveltuvat kuuntelukokeisiin luontevammin kuin tilastolliset testit ja hypoteesit ja mahdollistavat tutkimuksen optimoimisen bayesilaisia työkaluja hyödyntäen. Bayesilaisilla malleilla on lupaavia sovel-luskohteita myös koneoppimisessa, jossa arvioitsijoiden preferenssit voidaan mallintaa, ja tätä mallia vuorostaan käyttää koneoppimisalgoritmien opettamiseen.

VIITTEET

- [1] Catarina Mendonça and Symeon Delikaris-Manias. Statistical tests with MUSHRA data. In *Proceedings of the 144th Audio Engineering Society Convention*, 2018.
- [2] Rec. ITU-R BS.1534-3. Method for the subjective assessment of intermediate quality level of audio systems. Standard, International Telecommunications Union, Geneva, CH, 2015.

- [3] Michael Schoeffler, Sarah Bartoschek, Fabian-Robert Stöter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jürgen Herre. webMUSHRA — a comprehensive framework for web-based listening tests. *Journal of Open Research Software*, Feb 2018. doi: 10.5334/jors.187.
- [4] Jean D Gibbons, Ingram Olkin, and Milton Sobel. An Introduction to Ranking and Selection. *The American Statistician*, 33(4):185–195, November 1979. URL <https://www.jstor.org/stable/2683731>
- [5] L. Jeff Hong, Weiwei Fan, and Jun Luo. Review on ranking and selection: A new perspective. *Frontiers of Engineering Management*, 8(3):321–343, September 2021. ISSN 2095-7513, 2096-0255. doi: 10.1007/s42524-021-0152-6. URL <https://link.springer.com/10.1007/s42524-021-0152-6>.
- [6] Edward J. Dudewicz. Ranking (Ordering) and Selection: An Overview of How to Select the Best. *Technometrics*, 22(1):113, February 1980. ISSN 00401706. doi: 10.2307/1268390. URL <https://www.jstor.org/stable/1268390?origin=crossref>
- [7] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.
- [8] Geoffrey Cideron, Sertan Girgin, Mauro Verzetti, Damien Vincent, Matej Kastelic, Zalán Borsos, Brian McWilliams, Victor Ungureanu, Olivier Bachem, Olivier Pietquin, Matthieu Geist, Léonard Hussenot, Neil Zeghidour, and Andrea Agostinelli. MusicRL: aligning music generation to human preferences. In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org, 2024.
- [9] Adam X. Yang, Maxime Robeyns, Thomas Coste, Zhengyan Shi, Jun Wang, Haitham Bou-Ammar, and Laurence Aitchison. Bayesian reward models for LLM alignment, 2024. URL <https://arxiv.org/abs/2402.13210>.
- [10] Peter I. Frazier, Warren B. Powell, and Savas Dayanik. A Knowledge-Gradient Policy for Sequential Information Collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, January 2008. ISSN 0363-0129, 1095-7138. doi: 10.1137/070693424. URL <http://epubs.siam.org/doi/10.1137/070693424>
- [11] Stephen E. Chick, Jürgen Branke, and Christian Schmidt. Sequential Sampling to Myopically Maximize the Expected Value of Information. *INFORMS Journal on Computing*, 22(1):71–80, February 2010. ISSN 1091-9856, 1526-5528. doi: 10.1287/ijoc.1090.0327. URL <https://pubsonline.informs.org/doi/10.1287/ijoc.1090.0327>

- [12] Chun-Hung Chen, Stephen E. Chick, Loo Hay Lee, and Nugroho A. Pujowidianto. Ranking and Selection: Efficient Simulation Budget Allocation. In Michael C Fu, editor, *Handbook of Simulation Optimization*, volume 216, pages 45–80. Springer New York, New York, NY, 2015. ISBN 978-1-4939-1383-1 978-1-4939-1384-8. doi: 10.1007/978-1-4939-1384-8_3. URL https://link.springer.com/10.1007/978-1-4939-1384-8_3. Series Title: International Series in Operations Research & Management Science.
- [13] Tom Rainforth, Adam Foster, Desi R. Ivanova, and Freddie Bickford Smith. Modern Bayesian experimental design. *Statistical Science*, 2024.