Musiikinhaku sisällön perusteella akustisesta signaalista

Jari Suominen

Tiivistelmä

Musiikinhaku akustisesta signaalista tarkoittaa tavallisesti hakua, jossa muutaman sekunnin mittaisen ääninäytteen perusteella haetaan useiden äänitteiden muodostamasta tietokannasta vastaava äänite. Toimiva tapa suorittaa haku on muodostaa ääninäytteestä äänenjälki *(engl. audio fingerprint)*, joka yksinkertaistaa ja tiivistää näytteen akustista informaatiota. Tämän jälkeen saatua äänenjälkeä verrataan tietokannan aineistosta muodostettuihin äänenjälkiin. Hyvin muodostettu äänenjälki muodostaa näytteestään yksiselitteisen esityksen, samoin kuin ihmistenkin sormenjälkien tapauksessa.

1 Johdanto

Tärkein sovellusalue musiikinhaulle akustisesta signaalista on kahden ääninäytteen vertaaminen toisiinsa. Tyypillisesti tietokantaan syötetään äänitteistä johdettuja äänenjälkiä, joiden perusteella tietokanta indeksoidaan. Äänenjälkiin liitetään erilaisia tunnistetietoja. Tunnistettavasta ääninäytteestä lasketaan äänenjälki, jolla haetaan tietokannasta. Äänenjälkien käyttö on järkevämpää kuin kokonaisten äänitiedostojen vertailu: Äänijäljet vievät vähän tilaa, jolloin tietokannat ovat pienempiä. Puhdas äänisignaali sisältää paljon informaatiota, joka ei ole relevanttia haun kannalta. Äänenjälkiä käytettäessä myöskin haku on mahdollista toteuttaa tehokkaammin. [2: 107]

Perussovelluksia äänenjälkimenetelmiin perustuville musiikinhakualgoritmeille ovat lähetysten monitorointi, kappaleiden tunnistus, tietoliikenteen monitorointi [2: 108]. Äänenjälkiä on myös mahdollista käyttää varmistamaan, että jokin äänitetiedosto on yhä alkuperäisessä tilassa [1: 392].

Lähetysten ja tietoliikenteen monitorointi ovat ääniteteollisuuden tarpeisiin kehitettyjä käyttötapauksia, jotka liittyvät nimenomaan saman äänitteen tunnistamiseen. Tällöin akustisen signaalin käsittely on luonteva lähtökohta. Sovelluksen tehtävänä on tällöin automaattisesti tarkkailla joko radiolähetystä, tai palvelimella kulkevaa tietoliikennettä ja tunnistaa siirrettävästä informaatiosta kappaleita, joita sen tietokantaan on siirretty. Näin voidaan valvoa soittokorvausten tilittämistä ja tekijänoikeuksien rikkomista.

Tavalliselle musiikinkuluttaja voi tunnistuttaa vaikkapa radiosta kuulemiaan kappaleita esimerkiksi lähettämällä äänitietoa matkapuhelimen välityksellä palveluntarjoajalle. Tällainen sovellus esitellään lyhyesti kappaleessa 5.

Tässä tutkielmassa pyrin esittellemään lyhyesti äänenjälkimenetelmien

ominaisuudet ja rajoitukset. Lisäksi esittelen pääpiirteittäin kaksi käytännön äänenjälkimenetelmää.

2 Äänenjäljet

Äänenjälkien käyttö on verrattavissa hajautusfunktioihin: Verrattavista ääninäytteistä muodostetaan tiivistykset, äänenjäljet, joita verrataan toisiinsa, jolloin varsinaisia tiedostoja ei ole tarpeen verrata toisiinsa. Varsinainen syy äänenjälkien käyttöön on kuitenkin eri kuin hajautusfunktioiden tapauksessa. Tunnistettava ääninäyte on usein äänenlaadultaan huomattavasti tietokannan vastaavasta äänitteestä poikkeava.

Tyypillisiä äänenmuokkaustapahtumia, joiden läpi tunnistettava signaali on viety käytännön sovelluksissa ovat muun muassa ad/da-muunnokset, äänenpakkaus (esim mp3), dynamiikan rajoittajat, taajuuskorjaimet, matkapuhelinten äänensiirron pakkaus ja säröytyminen vahvistimessa. Lisäksi tunnistettava ääninäyte voi olla meluisassa paikassa tallennettu. Vaikka ihmisen on helppo tunnistaa kappale edellä mainituista häiriöistä huolimatta, tekevät ne signaalien suorasta vertailuista mahdotonta. Äänenjäljet muodostetaankin yleensä sellaisten akustisten ominaisuuksien perusteella, jotka erilaisten äänenmuokkaustekniikoiden jäljiltä pysyvät mahdollisimman muuttumattomina.

Äänenjälki voidaan muodostaa suorittamalla hajautus äänenpätkän akustisten ominaisuuksien suhteen [1: 392]. Näitä ovat esimerkiksi signaalin voimakkuus, spektrin painokeskus ja sävelkorkeus [1: 392]. Tällöin puhutaan sitkeästä hajautuksesta *(engl. robust hashing)*. Suodinpankkien avulla on myöskin mahdollista suorittaa äänenjäljen muodostaminen [2]. Kappaleessa 5 esitetään tarkemmin tähtikuviokarttoihin *(engl. constellation map)* perustuva menetelmä. Myöskin signaalin musiikillisia ominaisuuksia, kuten rytmiä ja harmoniaa voidaan käyttää äänenjälkien luomisessa [1: 392].

Yleisesti äänenjäljen tulisi muodostaa kuulohavaintoa vastaava tiivistys äänitteestä, sekä olla häiriöille immuuni, pienikokoinen ja nopeasti laskettavissa [1: 390].

3 Äänenjälkimenetelmien omaisuuksia

Äänenjälkimenetelmiltä vaadittavat ominaisuudet riippuvat sovelluksesta, mutta pääparametreina voidaan pitää sitkeyttä *(engl. robustness)*, luotettavuutta, äänenjäljen kokoa, rakeisuutta *(engl. granularity)* ja hakujen nopeutta ja skaalattavuutta [2: 108]. Ominaisuudet ovat toisitaan riippuvaisia. Esimerkiksi tunnistus mahdollisimman lyhyestä signaalinpätkästä (rakeisuus) johtaa luotettavuuden heikkenemiseen [2: 108]. Ominaisuuksien painotus riippuu sovelluksen luonteesta.

Sitkeysvaatimuksen mukaan äänenpätkä pitäisi pystyä tunnistamaan häiriöisestä signaalista. Tällöin signaalista johdettu äänenjäljen tulisi olla aina samanlainen, riippumatta sen korruptoitumisasteesta. Lisäksi äänenjäljen tulisi olla yksikäsitteinen, eikä sitä pitäisi voida yhdistää väärään äänitteeseen. Sitkeyttä mitataan usein

tarkkailemalla virheellisesti epäonnistuneiden hakujen määrää *(the false negative rate)*. Hakua pidetään virheellisesti epäonnistuneena mikäli tunnistettavaa signaalia vastaava äänite löytyy tietokannasta. [2: 108]

Menetelmän luotettavuus liittyy sitkeysvaatimukseen. Mitä harvemmin tunnistettavasta signaalista johdettu äänenjälki yhdistetään väärään nauhoitteeseen *(the false positive rate)*, sitä luotettavampi menetelmä on [2: 108]. Musiikinhakusovelluksissa pidetään virheellisten tunnistusten välttämistä ensisijaisena, vaikka se johtaisi äänipätkän tunnistamatta jäämiseen, vaikka vastaava äänite olisi tietokannassa.

Äänenjäljen pitää tehokkaan haun näkökulmasta olla mahdollisimman pieni. Tämä on ristiriidassa sitkeys- ja luotettavuusvaatimuksen kanssa. [2: 108]

Menetelmän rakeisuus määrittää kuinka pitkä tunnistettavan signaalin tulee olla. Jos kyseessä on matkapuhelimen tai tietokoneen toistoohjelman välityksellä toimiva sovellus, on käyttäjäystävällisyyden näkökulmasta ajan syytä olla mahdollisimman pieni.

Haun nopeuteen vaikuttaa haun aikavaativuusluokka sekä hakuun kuluva aika. Kaupallisissa sovelluksissa haun pitäisi valmistua sekunnin murto-osissa, vaikka käytettävä tietokanta koostuisi sadoista tuhansista äänitteistä. [2: 108]

4 Haitsman ja Kalkerin sitkeä hakumenetelmä

Haitsma ja Kalker ovat esitelleet sitkeän äänenjälkiä käyttävän menetelmän [2], joka perustuu suodinpankkeihin. Menetelmässä äänenjäljet koostuvat 256 aliäänenjäljestä. Menetelmä kykenee tunnistamaan oikean äänitteen, mikäli se löytyy kannasta, vaikka signaali olisi pahoin korruptoitunut.

Äänenjälki muodostetaan laskemalla ääninäytteestä 256 aliäänenjäljen muodostama äänenjälkilohko. Tämä tehdään ikkunoimalla signaali 0,37 sekunnin mittaisilla Hanning-ikkunoilla jotka on asetettu päällekäin suhteen 31/32 mukaisesti. Tällöin aliäänenjälki tuotetaan 11,6 millisekunnin välein. Äänenjälkilohko kattaa tällöin tulee noin kolme sekuntia äänisignaalia. Jokaiselle ikkunaa kohti suoritetaan seuraavat toimenpiteet: Aluksi lasketaan Fourier-muunnos, ja saadun signaalin magnitudi-arvot jaetaan 32 taajuuskaistaan. Varsinaisessa aliäänenjäljessä yhtä kaistaa edustaa yksi bitti, jonka arvo valitaan funktiolla, jossa painoarvoa on kaistalla, sekä sen kahdella naapurikaistalla. Lopulta kukin aliäänenjälki siis koostuu 32 bitistä. Jokaista bittiä kohti lasketaan lisäksi varmuusluku, joka kertoo bitin luotettavuuden. Tällöin haun osumatarkuutta voidaan parantaa, kun haku suoritetaan myös epävarmojen bittien vaihtoehtoisella arvolla. Tällä on merkitystä varsinkin matkapuhelimen kautta siirrettyä signaalia tunnistettaessa. [2: 109]

Tietokanta on koostettu muodostamalla kustakin äänenjälkilohkoja koko kappaleen matkalta. Haku menetelmässä suoritetaan vertaamalla aluksi jotakin aliäänenjälkeä tietokannan vastaaviin. Aliäänenjälki hajautetaan sen arvon perusteella. Saadun indeksin kautta päästään listaan, johon on viety kaikki kyseisen äänenjäljen esiintymät eri kappaleissa. Jokainen sijaintipaikka käydän yksitellen läpi, ja sen naapurialiäänenjälkiä ja tunnistettavan äänenjälkilohkon aliäänenjälkiä verrataan

toisiinsa kunnes löydetään kiistaton osuma. Osumaa voidaan pitää oikeana jos äänenjälkilohkojen välisten bittivirheiden määrä on riittävän pieni. Keskimäärin yhtä tunnistusta varten tarvitaan noin 300 vertailuoperaatiota, jos käytössä on 10000 kappaleen tietokanta. Vertailuoperaatioiden kertautuu, mikäli myös bittien varmuustekijät otetaan huomioon. [2: 112]

Ainoastaan äänenjälkilohkon arvoja käytettäessä kevyesti korruptoituneet signaalit kyettiin tunnistamaan. Ongelmia aiheutui ylipakattu mp3 sekä matkapuhelimen kautta siirretty signaali. Kun luotettavuusluvut otettiin huomioon, myös edellämainitut tapaukset tuottivat onnistuneita hakuja. Nopeuden muutokset tekevät kuitenkin hausta mahdottomia, sillä aliäänenjälkiä tuotetaan kuitenkin tasaisin aikavälein. Menetelmä on aika- ja tilavaatimuksiltaan täysin toimintakelpoinen ja virheellisten tulosten todennäköisyys on pieni ($fpr = 3{,}6 * 10^{-20}$). [2: 115]

5 Shazam Entertainment, Ltd:n hakumenetelmä

Wang [3] on esitellyt tähtikuviokarttoihin perustuvan menetelmän. Menetelmä on käytössä myös Suomessa. Menetelmä on räätälöity matkapuhelimen avulla käytettäväksi. Tavallinen käyttötapaus on seuraava: Kun asiakas kuulee esimerkiksi radiossa soivan kappaleen, jonka nimen hän haluaisi selvittää, soittaa hän palveluntarjoajan numeroon ja puhelimellaan äänittää kappaletta joitakin sekunteja. Vastaus lähetetään tekstiviestinä asiakkaalle. Shazamin algoritmi on nopea, mutta lisäksi erittäin sitkeä. Tunnistettava äänite voi soida esimerkiksi mainoksessa puheen taustalla ilman että tunnistus epäonnistuu. [3]

Äänenjäljet muodostetaan Shazamin menetelmässä spektogrammin huippujen perusteella. Spektogrammin huiput ovat suhteellisen immuuneja häiriöille. Huippuja etsitään tarkkailemalla pisteen ympäristöä: jos pisteen arvo on suurempi kun sen kaikkien sen naapuruston pisteiden, valitaan se mahdolliseksi huipuksi. Varsinaisia huippuja pyritään valitsemaan siten, että äänitettä kuvaava aika-taajuus-taso tulee kattavasti täytetyksi. Valitut huiput muodostavat tähtikuviokartan, joka koostuu ainoastaan aika-taajuus koordinaattiparista, sillä amplitudien arvoja huippupisteissä ei säilytetä. Menetelmän etuna on se, että valitut huiput ovat havaittavissa kohinan ja häiriöiden takaa. Lisäksi tahallinen ja tahaton taajuusalueen korjaaminen tai vääristyminen ei vaikuta huippujen havaittavuuteen. Näin siksi, että edellämainituissa tilanteissa suodattimien taajuusvaste on melko tasainen. [3: 8]

Hakeminen tapahtuu menetelmässä intuitiivisesti kuljettamalla kalvolle piirrettyä tunnistettavan äänenpätkän tähtikuviokarttaa tietokannassa olevien karttojen yli, kunnes löydetään samankaltainen paikka. Käytännössä Shazamin algoritmi käyttää tehokkaampaa menetelmää. Tällöin jokaisesta tietokannan äänitettä kuvaavasta kartasta johdetaan joukko äänijälkihajautusarvoja. Tämä suoritetaan valitsemalla kartalta ankkuripiste, sekä kohdealue, jonka pisteiden kanssa hajautus suoritetaan. Hajautuksessa käytetään ankkuripisteen ja kohde-alueen pisteen välimatkaa, sekä molemman pisteen taajuuskomponenttia. Hajautusarvot lasketaan näin ankkuripisteestä jokaiseen kohde-aluen pisteeseen. Hajautusarvo liitetään kappaleen tunnisteeseen sekä ankkuripisteen aikakomponenttiin. Kartalta valitaan useita ankkuripisteitä ja käsitellään

ne vastaavasti. Yhtä karttaa kuvaavan taulukon kokoon vaikuttaa valittujen ankkuripisteiden määrä, sekä kohde-alueen koko, ja siitä valittujen pisteiden määrä. Ottamalla huomioon vähemmän pisteitä ja pienempi kohde-alue, voidaan tilaa säästää. [3: 8]

Haku suoritetaan tuottamalla tunnistettavasta ääninäytteestä joukko hajautusarvoja, jotka yhdistetään ajanhetkiin, täysin vastaavasti kuin kokonaisen kartankin tapauksessa. Hajautusarvojen perusteella suoritetaan hakuja tietokannasta, ja osumien sattuessa kirjataan muistiin löytyneen äänitteen tunniste ja ajanhetki sekä haussa käytetyn hajautusarvon ajanhetki. Kun kaikki tunnistettavan äänenpätkän hajautusarvoilla on suoritettu haku tietokannasta, analysoidaan tulokset. Jokaiseen äänitteeseen liitettyjen ajanhetki-parien arvojen tulisi vastata toisiaan sillä hetkellä kun tunnistettava äänipätkä on oikealla paikallaan. Vastaavuudessa pitää ottaa huomioon ainoastaan se, että tunnistettavan äänenpätkän ajankohdat ovat tietokannan arvoja edellä, riippuen äänenpätkän sijainnista kappaleessa. Osuma voidaan selvittää yksinkertaisesti laskemalla jokaisen ajanhetki-parin arvojen erotus, ja ottamalla näistä histogrammi. Mikäli histogrammissa on havaittavissa selkeä huippuarvo tai arvokeskittymä, on yhtenevyys todennäköinen. Histogrammeja lasketaan äänitteille, joihin on haussa tullut osumia, kunnes kiistaton yhtenevyys on löydetty. [3: 10]

Shazamin algoritmi löytää tehokkaasti tunnistettavan äänitteen tietokannasta huolimatta tunnistettavan äänitteen säröisyydestä. Se on kuitenkin herkkä ajankulun suhteen, joten se ei pysty tunnistamaan esimerkiksi saman kappaleen eri versioita samaksi, mihin sitä ei ole suunniteltukaan. Kaikkiin sovelluksiin, joissa riittää tunnistaa täsmälleen sama versio kappaleesta vaikuttaa Shazamin algoritmi melko optimaaliselta.

6 Johtopäätökset

Varsinkin ääniteteollisuuden tarpeisiin akustisesta signaalista tapahtuva musiikinhaku on tarpeellinen ja järkevin vaihtoehto. Sitä on kuitenkin vaikea verrata merkkijonoja käyttäviin menetelmiin, sillä lähtökohdat ovat aivan toisenlaiset. Äänenjälkimenetelmien rajoja on kuitenkin vielä vaikea nähdä. Äänenjälkiä on mahdollista muodostaa muillakin kuin tässä tutkielmassa esitetyillä tavoilla. Periaatteessa voisi siis pitää mahdollisena, että erilaisten melodioiden haku vaikkapa hyräilyn perusteella saattaisi olla toteutettavissa käyttämällä puhdasta akustista signaalia, jolloin tarvetta muuntaa musiikkia merkkijonoesitykseksi ei ole.

Viitteet

[1] Cano, Pedro - Battle, Eloi - Gómez, Emilia - Gomes, Leandro de C. T. - Bonnet, Madeleine 2002: Audio Fingerprinting: Concepts and applications. *Proceedings of the International Conference on Fuzzy Systems and Knowledge Discovery (FSKD): vol 2: 389-393.*

[2] Haitsma, Jaap - Kalker, Ton 2002: A Highly Robust Audio Fingerprinting System. *Proceedings of the International Conference on Music Information Retrieval (ISMIR): 107-115.*

[3] Wang, Avery 2003: An Industrial Audio Search Algorithm. *Proceedings of the International Conference on Music Information Retrieval (ISMIR): 7-13.*