ÄÄNIELEET VUOROVAIKUTTEISISSA SOVELLUKSISSA JA NIIDEN AUTOMAATTINEN TUNNISTAMINEN

Antti Jylhä

Aalto-yliopiston sähkötekniikan korkeakoulu
Signaalinkäsittelyn ja akustiikan laitos
PL 13000, 00076 AALTO
antti.jylha@aalto.fi

1 JOHDANTO

Ääniperusteinen vuorovaikutussuunnittelu (sonic interaction design, SID, `http://www.cost-sid.org/`) on nuori tutkimusala, joka tarkastelee ääntä informatiivisen, emotionaalisen ja esteettisen sisällön välittäjänä. Yhtenä SID:n tutkimuskohteena on äänen käyttö ihmisen ja tietokoneen välisessä vuorovaikutuksessa, missä on valtaosin keskitytty äänen avulla käyttäjälle välitettävään informaatioon. Sen sijaan äänen käyttö toiseen suuntaan, ihmiseltä tietokoneelle, on saanut vähemmän huomiota osakseen. Osittain tämä johtuu siitä, että tyypillisesti tämänsuuntaisessa ääniviestinnässä on keskitytty puheentunnistukseen ja SID on määritelmällisesti jättänyt puheen tutkimuskentän ulkopuolelle. Puhe ei kuitenkaan ole ainoa käyttökelpoinen äänimuoto, jonka avulla informaatiota voidaan kuljettaa.

Tässä tutkimuksessa keskitytään äänieleisiin (eng. *sonic gestures*), jotka ovat ihmisen tuottamia merkityksellisiä ja tunnistettavia ääniä. Esimerkkejä tällaisista äänistä ovat käsien taputukset, sormien napsutukset, viheltäminen ja hyräily, sekä erilaiset äännähdykset, jotka eivät ole puhetta. Vaikka äänieleitä on käsitelty viime aikoina yksittäisissä tutkimuksissa osana sovelluksia ja käyttöliittymiä, systemaattista katsausta niiden luokitteluun ei ole aiemmin esitetty.

Äänieleiden käyttöä vuorovaikutteisissa rajapinnoissa ja sovelluksissa puoltaa moni seikka. Äänieleiden tunnistamiseen ei vaadita yleensä mitään erityistä laitteistoa, sillä useimmissa relevanteissa laitteissa on mikrofoni sisäänrakennettuna ja riittävästi laskennallista tehokkuutta reaaliaikaiseen vuorovaikutukseen. Haasteet liittyvätkin enemmän laskennallisesti tehokkaiden algoritmien toteutukseen [1]. Lisäksi äänieleillä tapahtuva viestintä ei edellytä kosketusta tai näköyhteyttä laitteen kanssa, mikä on erityisesti tarpeen tilanteissa, joissa käyttäjän kädet tai huomio ovat keskittyneet muuhun toimintaan. Äänieleet voivat myös mahdollistaa näkörajoitteisten tai motorisesti rajoitteisten käyttäjien vuorovaikutuksen sovellusten kanssa, joiden käyttöön heillä ei muulla tavoin olisi mahdollisuutta.

Tämä tutkimus jakautuu kahteen osaan. Ensin esitellään äänieleiden määritelmä, joukko erilaisia äänieleitä ja mahdollisuuksia näiden luokitteluun niiden ominaisuuksien pohjalta. Tämän jälkeen osoitetaan äänieleiden käyttömahdollisuuksia käymällä lyhyesti läpi joukko rajapintoja ja sovelluksia sekä niiden yhteydessä käytettyjä äänten tunnistusmenetelmiä. Lopuksi esitetään johtopäätöksiä.

2 ÄÄNIELEET

Eleiden käyttö ihmisen ja tietokoneen välisessä vuorovaikutuksessa on tyypillisesti keskittynyt visuaalisesti tai haptisesti havaittaviin eleisiin, joiden avulla ohjataan jotain toimintoa perustuen kameran tai antureiden avulla tapahtuvaan eleiden tunnistukseen. Äänen ja eleiden suhdetta on aiemmin tarkasteltu enimmäkseen näkökulmasta, jossa ele on fyysinen toiminto, joka tuottaa tai muokkaa jotain ääntä [2]. Äänieleet, joissa ele itsessään välittää informaatiota äänen avulla, ovat sen sijaan jääneet vähemmälle huomiolle. Musiikillisissa yhteyksissä tosin on tutkittu "äänieleitä", jotka kuitenkaan eivät eleinä itsessään välttämättä ole ääntä tuottavia [3, 4].

Tässä tutkimuksessa äänieleet määritellään ihmisen aiheuttamiksi ääntä tuottaviksi toiminnoiksi, jotka välittävät informaatiota. Esimerkiksi käsien taputus on ihmisen tuottama ääniele, jolla on sangen tunnistettava ääni. Äänieleiden kirjo on varsin laaja, sillä niiden voidaan katsoa sisältävän kaikki ihmisen tuottamat äänet, jotka eivät ole puhetta. Ääntöväylän avullakin voidaan kuitenkin tuottaa äänieleitä mm. hyräilemällä.

Äänieleiden luokittelua voidaan lähestyä useista näkökulmista. Perustuen aiempaan eleisiin ja akustiseen typologiaan ja morfologiaan liittyvän tutkimukseen, äänieleet voidaan jaotella kolmeen ryhmään: impulsiivisiin, toistuviin eli iteratiivisiin ja jatkuviin äänieleisiin [2, 4]. Impulsiivinen ääniele on yksittäinen, impulssimainen ele, kuten käsien taputus, sormien napsutus, pöydän taputus tai jalan iskeminen maahan; toisin sanoen impulsiivinen ääniele on pelkistettävissä yhdessä ajanhetkessä tapahtuvaksi eleeksi. Jatkuva ääniele puolestaan on pitempiaikainen ele, kuten vihellys, hyräily tai kynsien hankaaminen karkeaa pintaa vasten. Iteratiivinen ääniele koostuu peräkkäisistä äänieleistä, joita toistetaan mahdollisesti säännöllisin aikavälein. Esimerkiksi käsien taputtaminen tietyllä tempolla on iteratiivinen ääniele. On huomattava, että iteratiivinen ääniele ei välttämättä koostu impulsiivisten äänieleiden sarjasta, vaan voi myös koostua toisiaan seuraavista jatkuvista äänieleistä.

Äänieleet voidaan myös jaotella soinnillisiin ja soinnittomiin niiden spektriominaisuuksien mukaan. Taulukossa 1 on luokiteltu joukko äänieleitä niiden soinnillisuuden ja typologian mukaan. Esitetty jaottelu ei ole tyhjentävä tai ainoa mahdollinen, vaan äänieleillä on muitakin luokitteluun soveltuvia ulottuvuuksia. Toisistaan voidaan erotella esimerkiksi äänieleet, joiden spektrikomponentit pysyvät jokseenkin muuttumattomina äänitapahtuman ajan (esim. vihellys vakiosävelkorkeudella) ja eleet, joiden spektri muuttuu (esim. nouseva/laskeva sävelkorkeus tai muuttuva äänenväri).

Äänieleet voidaan myös jaotella niiden tuottomekanismin mukaan instrumentaalisiin ja ei-instrumentaalisiin (ns. tyhjäkätisiin) äänieleisiin [1]. Instrumentaalisissa äänieleissä ääntä tuottava vuorovaikutus tapahtuu ihmisen ja jonkin toissijaisen ääntä tuottavan esineen välillä, kuten taputtaessa sormilla pöydän pintaa. Tyhjäkätiset äänieleet puolestaan ovat täysin ihmisen itsensä tuottamia ilman välikappaleita. Näitä ovat esimerkiksi käsien taputukset ja ääntöväylän avulla tuotetut eleet.

Lisäksi äänieleitä voidaan tarkastella sen perusteella, minkälaiseen vuorovaikutukseen ne soveltuvat. Esimerkiksi impulsiiviset äänieleet soveltuvat paremmin diskreettien komentojen antamiseen kuin jatkuva-asteikkoisten parametrien ohjaamiseen, johon soinnilliset äänieleet ovat luonnollisempi valinta. Soveltuvuuteen vaikuttaa se, mitä para-

Taulukko 1: *Äänieleiden esimerkkijoukon luokittelu akustisen morfologian perusteella. Mitä tahansa impulsiivista tai jatkuvaa äänielettä voidaan toistaa sarjassa iteratiivisen eleen muodostamiseksi.*

	impulsiivinen	iteratiivinen	jatkuva
soinniton	käsien taputus sormien napsutus askeläänet koputusäänet impulsiiviset äänteet	käsien taputussarjat sormien napsutussarjat askeläänisarjat koputusäänisarjat imp. äännesarjat raaputusäänisarjat frikatiivisarjat hengitysäänisarjat	kitka- ja raaputusäänet suhinaäänteet / frikatiivit hengitys- ja puhaltamis-äänet
soinnillinen		vihellyssarjat hyräilysarjat vokaalisarjat	vihellys hyräily vokaaliäänteet

metreja (diskreettejä/jatkuvia) kustakin äänieleestä voidaan riittävän luotettavasti ja luontevasti tunnistaa. Esimerkiksi iteratiiviset äänieleet ovat erinomaisia silloin, kun tarvitaan jatkuvia aikapiirteitä sinänsä impulssimaisten äänten tunnistukseen perustuvissa rajapinnoissa vaikkapa rytmisen vuorovaikutuksen sovelluksissa.

3 ÄÄNIELEIDEN SOVELLUKSIA JA TUNNISTUSMENETELMIÄ

Äänieleiden kirjo ja sovellusmahdollisuudet ovat mittavat. Vaikka yhtenäistä katsausta äänieleiden käyttöön ja tunnistukseen ei aiemmin ole esitetty, useat yksittäiset sovellukset ja tutkimukset osoittavat niiden monipuolisuuden ja käyttökelpoisuuden ihmisen ja tietokoneen välisessä vuorovaikutuksessa. Keskeisenä tutkimuskohteena äänielevuorovaikutuksen toteuttamisessa on reaaliaikaisten ja tehokkaiden tunnistusalgoritmien kehittäminen. Tässä esitetään lyhyt katsaus erilaisiin sovelluskohteisiin ja tunnistusmenetelmiin aiempien tutkimusten pohjalta.

Vesa ja Lokki ovat kehittäneet sormien napsautuksilla ohjattavan mediasoitinkäyttöliittymän, jossa kahden mikrofonin avulla tunnistetaan paitsi napsautus itsessään, myös se, millä puolella päätä napsautus tapahtuu [5]. Vasemmalla puolella päätä tapahtuva napsautus ohjaa edelliseen ja oikealla puolella tapahtuva seuraavaan raitaan siirtymistä. Pään edessä (tai takana) tapahtuva napsautus ohjaa toisto/pysäytys -toimintoa. Napsautusten tunnistaminen perustuu signaalienergian vertaamiseen kynnysarvoon ja näin tunnistettujen äänitapahtumien luokitteluun kaistaenergiasuhteiden perusteella. Napsautusten paikantamisessa käytetään korvien välistä ristikorrelaatiota, jonka avulla lasketaan korvien väliset aika- ja tasoerot (ITD ja ILD).

Jylhä ja Erkut ovat kehittäneet perkussiivisten äänieleiden käyttöön perustuvan rajapinnan, joka on suunniteltu erityisesti käsien taputusäänille [6]. Käsien asento taputtaessa vaikuttaa syntyvään ääneen, ja näin syntyvät erilaiset äänet on mahdollista erotella toisistaan käyttäen yksinkertaisia hahmontunnistusmenetelmiä [7]. Tätä informaatiota

hyödyntäen rajapinnan avulla käyttäjä voi erilaisia taputuksia käyttäen antaa erilaisia komentoja ohjattavalle sovellukselle. Äänten tunnistamiseen käytetään algoritmia, joka laskee signaalille tehoestimaattia usealla taajuuskaistalla, havaitsee perkussiivisen äänitapahtuman tehoestimaatin perusteella ja kykenee oppimaan kullekin äänelle ominaisen mallin kaistojen tehoestimaateista [8]. Lisäksi rajapinta pystyy seuraamaan taputusten tempoa ja rytmiikkaa, mitä voidaan myös käyttää vuorovaikutteisissa sovelluksissa. Osoituksena tästä rajapinnan ympärille on kehitetty vuorovaikutteinen flamencoopettajasovellus, jonka avulla käyttäjä voi harjoitella flamencomusiikin taputusrytmejä ja saa palautetta virtuaaliselta opettajalta suorituksestaan ja oppimisestaan [9].

Vocal Joystick on rajapinta, jonka avulla käyttäjä voi ohjata käyttöjärjestelmän kursoria vokaaliäänteillä [10]. Vokaaliäänteet kuvataan kursorin liikkeen suunnaksi kaksiulotteisen vokaalikartan avulla ja äänen voimakkuus vaikuttaa kursorin liikenopeuteen. Lisäksi konsonanttiäänteillä on mahdollista antaa diskreettejä komentoja. Äänisignaalia tarkastellaan kehyksittäin ja jokainen kehys luokitellaan aluksi joko hiljaisuudeksi, esiaktiiviseksi (mahdollinen äänitapahtuman alku) tai aktiiviseksi (osa äännettä) kehykseksi perustuen signaalin energian ja nollanylitystaajuuden tarkkailuun. Aktiiviset kehykset luokitellaan vokaaliäänteiksi käyttäen tunnistuspiirteinä mel-kepstrikertoimia (MFCC) ja luokittelualgoritmina monikerrosperseptroniverkkoa. Konsonanttiäänteet tunnistetaan Markovin piilomallin (HMM) avulla. Lisäksi vokaaliäänteen sävelkorkeus tunnistetaan käyttäen dynaamista bayesilaista todennäköisyysmallia.

Hiiren korvikkeeksi on esitetty myös hyräilyyn ja suhinaäänteisiin perustuva rajapinta [11]. Tässä käyttöliittymässä kursoria liikutellaan perustuen nelisoluiseen hilaesitykseen, jossa liikutaan hyräilemällä korkealta tai matalalta. Valinta voidaan tehdä suhisemalla. Hyräilyn ja suhinan tunnistaminen perustuu signaalienergian laskentaan kahdella eri taajuuskaistalla sekä autokorrelaatiopohjaiseen sävelkorkeuden estimointiin.

Sporka on esitellyt useita soinnillisten äännähdysten tunnistamiseen perustuvia käyttöliittymiä ja sovelluskohteita perustuen sävelkorkeuden tunnistamiseen [12]. Käyttöliittymät mahdollistavat mm. kursorin ohjaamisen viheltämällä tai hyräilemällä, näppäimistösyötteen emuloinnin hyräiltävistä äännähdyssarjoista muodostetun aakkoston avulla sekä tietokonepelien ohjaamisen käyttäen sävelkorkeuseleitä. Sävelkorkeus tunnistetaan käyttäen nopeaa Fourier-muunnosta (FFT) ja/tai autokorrelaatiota.

Hämäläinen on tarkastellut äänieleitä osana tietokonepelien käyttöliittymää [13]. Sävelkorkeutta käytetään pelihahmojen ohjaamiseen ja eräässä lapsille suunnatussa pelissä huutamalla saa lohikäärmeen syöksemään tulta.

Billaboop [14] on alkujaan kehitetty rumpuäänten ohjaamiseen beatboxing-tyyppisellä syötteellä, mutta soveltuu myös muunlaisille perkussiivisille syötteille, kuten pöytärummutukselle. Äänten tunnistus perustuu useisiin aika- ja taajuustason piirteisiin sekä päätöspuutyyppiseen hahmontunnistusalgoritmiin. Tältä pohjalta on kehitetty mobiilisovellus, BoomClap (`http://billaboop.com/en/boomclap`), joka mahdollistaa reaaliaikaisen rumpusyntetisaattorin ohjaamisen perkussiivisten äänieleiden avulla.

Scratch Input [15] on rajapinta, joka hyödyntää pinnan raaputuksessa syntyviä runkoääniä. Äänten havaitsemiseen käytetään kontaktimikrofoneja ja erilaiset, akustisesti uniikit äänieleet erotellaan toisistaan yksinkertaisten hahmontunnistusalgoritmien avulla. Koska rakenneäänet ovat varsin epäherkkiä ympäristön äänten aiheuttamille häiriöille,

järjestelmän on todettu toimivan myös meluisissa ympäristöissä.

Smule Ocarina [16] on sovellus, joka mallintaa okarina-soittimen ääntä ja soittotapaa käyttäen hyödyksi nykyaikaisen matkapuhelimen syötemahdollisuuksia. Heräteäänieleenä järjestelmässä käytetään laitteen mikrofoniin puhaltamista, joka synnyttää ääntä turbulentin ilmavirtauksen kautta.

Äänieleiden käyttö vuorovaikutteisissa sovelluksissa on siis varsin monipuolinen tutkimuskohde ja uusia innovaatiota äänisyötteiden käytössä ja tunnistuksessa tehdään jatkuvasti. On myös huomattava, että monet käytetyistä tunnistusmenetelmistä eivät ole sinänsä uusia, vaan tunnettuja esimerkiksi musiikillisen informaation louhinnan ja puheentunnistuksen aloilta. Niinpä sovelluksia kehitettäessä on usein mahdollista käyttää valmiita, toisaalla hyviksi todettuja ratkaisuja.

4 JOHTOPÄÄTÖKSET

Äänieleillä on monia potentiaalisia sovelluskohteita ihmisen ja tietokoneen välisessä vuorovaikutuksessa. Niiden avulla voidaan välittää sekä diskreettiä että jatkuvaa informaatiota, jota voidaan käyttää sovellusten ohjaamiseen eri tavoin. Tähän mennessä äänieleitä on tarkasteltu lähinnä yksittäisten sovellusten ja tutkimusten osana, joten yhtenäistä katsausta niiden käyttöön ja käytettävissä oleviin tunnistusmenetelmiin ei ole esitetty. On kuitenkin huomattava, että monet tässä tutkimuksessa mainituista tunnistusmenetelmistä ovat tunnettuja audiosignaalinkäsittelyn muilta osa-alueilta, kuten musiikki-informaation käsittelystä ja puheentunnistuksesta. Äänieleiden tunnistus voidaan nähdä puheentunnistusta helpompana ongelmana siinä mielessä, että äänieleiden "kieli"on vähemmän sidoksissa kulttuurillisiin ulottuvuuksiin. Näin ollen äänieleiden käyttö voidaan ainakin teoriassa nähdä yleismaailmallisempana.

Äänieleiden laajamittainen käyttö käytännön sovelluksissa on toistaiseksi vähäistä, mutta viimeaikainen kehitys esimerkiksi mobiililaitteiden ja viihde-elektroniikan saroilla on esittänyt viitteitä äänieleiden yleistymisestä sovellusten ja rajapintojen osana. Onkin varsin uskottavaa, että lähitulevaisuudessa kohtaamme monia äänieleisiin perustuvia uudenlaisia vuorovaikutusmuotoja ja sovelluksia.

4.1 Kiitokset

Tätä tutkimusta ovat rahoittaneet Aalto-yliopiston sähkötekniikan tutkijakoulu sekä Suomen Akatemia (projekti SCHEMA-SID). Kiitokset Cumhur Erkutille tuesta artikkelin kirjoittamisessa.

VIITTEET

[1] MIRANDA E & WANDERLEY M, *New Digital Musical Instruments: Control and interaction beyond the keyboard*, AR Editions, Inc., 2006.

[2] GODØY R & LEMAN M, *Musical gestures: Sound, movement, and meaning*, Taylor & Francis, 2009, ISBN 0415998867.

[3] CADOZ C & WANDERLEY M, Gesture-music, *Trends in Gestural Control of Music*, (2000), 71–93.

[4] VAN NORT D, Instrumental Listening: sonic gesture as design principle, *Organised Sound*, **14**(2009) 02, 177–187, ISSN 1469-8153.

[5] VESA S & LOKKI T, An Eyes-Free User Interface Controlled by Finger Snaps, in *Proc. 8th Intl. Conf. Digital Audio Effects (DAFx)*, pages 262–265, Madrid, Spain, 2005.

[6] JYLHÄ A & ERKUT C, A hand clap interface for sonic interaction with the computer, in *Proc. Conf. Human Factors in Computing Systems (CHI)*, pages 3175–3180, Boston, MA, USA, April 2009, presented in interactivity.

[7] JYLHÄ A & ERKUT C, Inferring the hand configuration from hand clapping sounds, in *Proc. 11th Intl. Conf. Digital Audio Effects (DAFx-08)*, pages 300–304, Espoo, Finland, 2008.

[8] PUCKETTE M, APEL T, & ZICARELLI D, Real-Time Audio Analysis Tools for Pd and MSP, in *Proc. Intl. Computer Music Conference*, pages 109–112, Ann Arbor, MI, USA, October 1998.

[9] JYLHÄ A, EKMAN I, ERKUT C, & TAHIROĞLU K, Design and Evaluation of Rhythmic Interaction with an Interactive Tutoring System, *Computer Music Journal*, (2011), accepted for publication.

[10] BILMES J, MALKIN J, LI X, HARADA S, KILANSKI K, KIRCHHOFFI K, WRIGHT R, SUBRAMANYA A, LANDAY J, DOWDEN P, ET AL., The vocal joystick, in *Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages I–625–I–628, Toulouse, France, 2006.

[11] CHANJARADWICHAI S, PUNYABUKKANA P, & SUCHATO A, Design and evaluation of a non-verbal voice-controlled cursor for point-and-click tasks, in *Proc. 4th Intl. Conv. Rehabilitation Engineering & Assistive Technology*, pages 48:1–48:4, 2010.

[12] SPORKA A, *Non-Speech Sounds for User Interface Control*, Faculty of Electrical Engineering, Czech Technical University, 2008.

[13] HÄMÄLÄINEN P, *Novel Applications of Real-Time Audiovisual Signal Processing Technology for Art and Sports Education and Entertainment*, Helsinki University of Technology, 2007.

[14] HAZAN A, Performing Expressive Rhythms with BillaBoop Voice-Driven Drum Generator, in *Proc. 8th Intl. Conf. Digital Audio Effects (DAFx)*, pages 254–257, Madrid, Spain, 2005.

[15] HARRISON C & HUDSON S E, Scratch Input: creating large, inexpensive, unpowered and mobile finger input surfaces, in *Proc. 21st annual ACM Symp. on User Interface Software and Technology*, UIST '08, pages 205–208, 2008.

[16] WANG G, Designing Smule's iPhone ocarina, in *Proc. Intl. Conf. New Interfaces for Musical Expression*, Pittsburgh, PA, USA, 2009.