HUUDETUN PUHEEN ANALYYSI JA SYNTEESI

Tuomo Raitio[1], Antti Suni[2], Jouni Pohjalainen[1], Manu Airaksinen[1], Martti Vainio[2] ja Paavo Alku[1]

[1] Signaalinkäsittelyn ja akustiikan laitos, Aalto-yliopisto, Espoo
Otakaari 5 A, PL 13000, 00076 AALTO
etunimi.sukunimi@aalto.fi

[2] Käyttäytymistieteellinen tiedekunta, Helsingin yliopisto, Helsinki
Siltavuorenpenger 5 A, PL 9, 00014 Helsingin yliopisto
etunimi.sukunimi@helsinki.fi

Tiivistelmä

Tässä tutkimuksessa äänitettiin kaiuttomassa huoneessa sekä normaalia puhetta että huutoa 22 suomalaiselta koehenkilöltä. Puheen ja huudon analyysistä saadun tiedon perusteella tutkimuksessa etsitään keinoja joilla puhesyntetisaattori voi tuottaa luonnollisen kuuloista huutoa. Koska huudon perustaajuus on hyvin korkea, tulokset puoltavat sellaisten menetelmien käyttöä, jotka eivät ole herkkiä korkean perustaajuuden harmonisten aiheuttamille häiriöille estimoiduissa formanttitaajuuksissa. Tutkimuksessa rakennetaan tilastollinen puhesyntetisaattori, joka pystyy tuottamaan tekstistä sekä puhetta että huutoa käyttämällä datapohjaista puheen mallintamista ja tilastollisia adaptaatiometodeja. Tulokset osoittavat, että normaalin puheen synteesimallista saadaan tuotettua huutosynteesiä jo kohtalaisen pienellä adaptaatiohuutomateriaalilla. Subjektiivisissä kuuntelukokeissa synteettinen huuto arvioitiin, äänen laatua lukuunottamatta, hyvin lähelle aitoa huutoa.

1 JOHDANTO

Huuto on intensiteetiltään voimakkain puhekommunikaation muoto, jota käytetään usein meluisissa olosuhteissa nostamaan äänenvoimakkuutta häiritsevän melun yläpuolelle tai välittämään informaatiota puhujan tilasta tai aikeista. Huutaminen eroaa tavallisesta voimakkaasta tai Lombard-puheesta [1] siten, että se on sekä voimakkain että tahdonalainen äänenkäytön muoto. Huuto on kompromissi äänenpainetason (sound pressure level, SPL) sekä ymmärrettävyyden välillä; esimerkiksi Lombard-puhe on ymmärrettävämpää kuin tavallinen puhe, mutta huudossa äänen ominaisuudet muuttuvat siten että sen ymmärrettävyys heikkenee [2, 3].

Huutava ääni saadaan aikaan nostamalla glottiksen alapuolista painetta sekä kasvattamalla äänihuulten kireyttä. Tästä aiheutuen huudon SPL on huomattavasti korkeampi verrattuna tavalliseen puheeseen. Rostolland [4] raportoi huudon C-painotettujen SPL-erojen tavallisen puheen ja huudon välillä olevan 20 dB naisilla ja 28 dB miehillä. Huudon perustaajuus (F0) on myös korkeampi ja F0-käyrien variaatio on pienempää. Äänilähteessä glottiksen sulkuvaiheen suhteellinen pituus lyhenee [5], mikä näkyy taajuusalueessa korkeiden taajuuksien korostumisena. Myös huudon prosodiset piirteet eroavat

tavallisesta puheesta: huudossa vokaalien suhteelliset kestot ja voimakkuus kasvavat verrattuna soinnittomiin äänteisiin. Myös artikulointi on heikompaa; matalimmat formantit siirtyvät siten että vokaaleista tulee keskenään samankaltaisempia [6].

Huudetun puheen ominaisuuksia [2, 4, 6, 7] ja ymmärrettävyttä [2, 3] on tutkittu suhteellisen laajasti. Myös huudon tunnistusta ja luokittelua on tutkittu [8, 9], mutta huutosynteesistä ei löydy aikaisempia tutkimuksia. Huutoa käytetään onneksi harvemmin, mutta se on kuitenkin olennainen osa ihmisen puhekommunikaatiota. Syntetisaattoria joka osaa myös huutaa, tarvitaan mm. luotaessa emotionaalisia virtuaalisia agentteja, joita voidaan käyttää esim. ihmisen ja tietokoneen vuorovaikutuksessa, kommunikaatioapuvälineenä, tai virtuaalimaailmoissa.

2 PUHE- JA HUUTOMATERIAALI

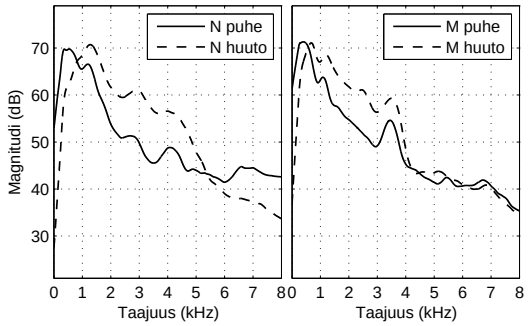
Tutkimuksessa äänitettiin kaiuttomassa huoneessa sekä normaalia puhetta että huutoa 22 suomalaiselta koehenkilöltä, joista 11 oli naisia. Kultakin koehenkilöltä äänitettiin 24 lausetta, jotka koehenkilöt pystyivät luontevasti sekä puhumaan että huutamaan. Koehenkilöitä pyydettiin käyttämään hyvin voimakasta ääntä huudettaessa. Jos koehenkilön huuto ei ollut tarpeeksi voimakasta, henkilöä pyydettiin toistamaan kyseinen lause. Yhteensä 528 lausetta äänitettiin sekä normaalilla puheella että huudettuna. Lisäksi yhdeltä mieheltä ja yhdeltä naiselta, joilta on aikaisemmin äänitetty puhetietokanta puhesynteesin kehitystä varten, äänitettiin 100 uutta huudettua lausetta kummallekin.

3 PUHEEN JA HUUDON ANALYYSI

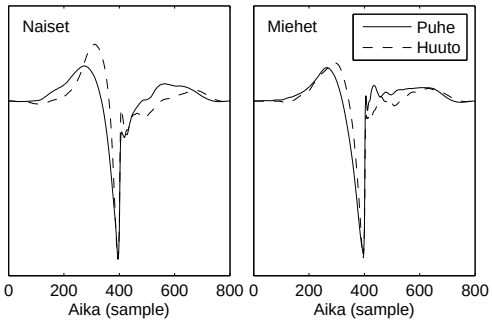
Äänitetystä materiaalista tutkittiin puheen ja huudon seuraavia akustisia ominaisuuksia: intensiteettiä, kestoa, perustaajuutta, spektriä sekä erinäisiä äänilähteen ominaisuuksia. Analyysin tulokset on esitetty taulukossa 1. Keskimääräinen äänenvoimakkuusero puheen ja huudon välillä oli 21 dB naisilla ja 22 dB miehillä. Huudettujen lauseiden kestot olivat naisilla 20% ja miehillä 24% pidempiä verrattuna puhuttuihin lauseisiin. Huudon perustaajuus oli naisilla keskimäärin 71% korkeampi ja miehillä 152%. Huudetun puheen spektri on korostunut erityisesti 1–4 kHz:n alueella äänilähteen spektrin kaltevuuden laskun takia, mikä nähdään kuvassa 1. Toisaalta energianormalisoiduissa puhenäytteissä 5–8 kHz:n alue on naisilla keskimäärin voimakkaampi puheella kuin huudolla, ja miehillä yhtä voimakas. Tämä johtuu siitä, että huudossa suurin osa energian lisäyksestä aiheutuu soinnillisten äänteiden energian kasvusta, jolloin soinnittomien äänteiden suhde soinnillisiin äänteisiin pienenee huomattavasti. Normaalin puheen ja huudon keskimääräiset glottispulssin muodot on esitetty kuvassa 2, josta nähdään, että äänilähde toimii hyvin eri tavalla puheessa ja huudossa. Näiden havaintojen lisäksi äänilähdettä

Taulukko 1: Puheen ja huudo keskimääräiset parametrit naisille ja miehille.

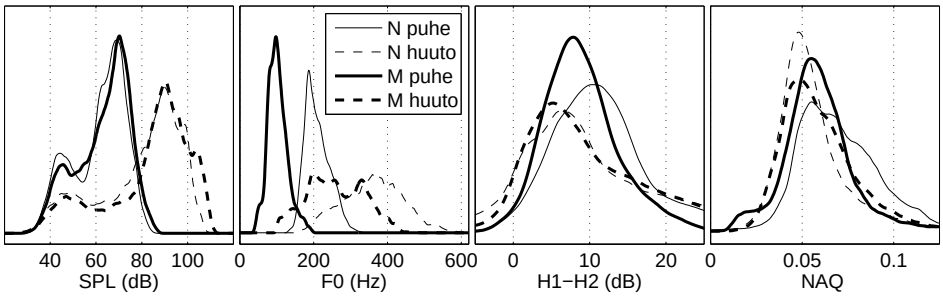
Param.	Yksikkö	Naisten puhe	Naisten huuto	Miesten puhe	Miesten huuto
Kesto	s	1.35 ± 0.03	1.62 ± 0.04	1.42 ± 0.04	1.76 ± 0.05
SPL	dB	61.6 ± 0.1	79.2 ± 0.1	63.0 ± 0.1	82.7 ± 0.1
F0	Hz	209.9 ± 0.3	359.7 ± 0.7	102.4 ± 0.2	259.4 ± 0.6
H1–H2	dB	11.56 ± 0.05	9.26 ± 0.07	9.01 ± 0.04	9.44 ± 0.06
NAQ	-	0.0729 ± 0.0003	0.0563 ± 0.0002	0.0607 ± 0.0002	0.0599 ± 0.0002



Kuva 1: Puheen ja huudon energianormalisoidut spektrit naisille (N) ja miehille (M).



Kuva 2: Puheen ja huudon keskimääräiset glottispulssit naisille ja miehille.



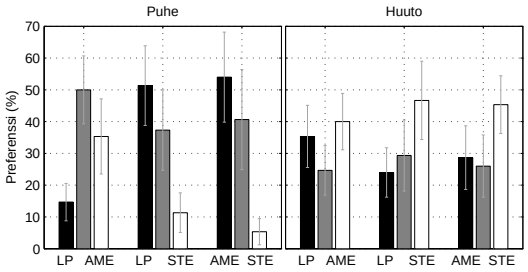
Kuva 3: *Puheen ja huudon parametrien jakaumat naisille (N) ja miehille (M).*

tutkittiin sitä kuvaavilla akustisilla parametreilla, kuten NAQ ja H1–H2, joiden jakaumat puheelle ja huudolle on esitetty kuvassa 3.

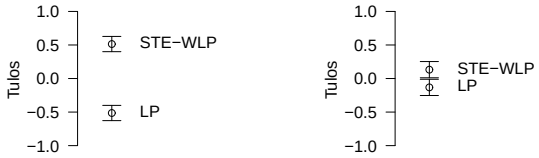
4 HUUDETUN PUHEEN SYNTEESI

Puheen ja huudon analyysistä saadun tiedon perusteella tutkimuksessa etsitään keinoja joilla puhesyntetisaattori voi tuottaa luonnollisen kuuloista huutavaa puhetta. Koska huudon perustaajuus on usein hyvin korkea, on huudon spektrin estimointi vaikeaa. Tämä johtuu siitä, että suuren F0:n omaavan herätesignaalin harvassa olevat harmoniset komponentit aiheuttavat formanttien vääristymistä kyseisiä harmonisia kohti. Etenkin tavallisesti käytetty lineaariprediktio (linear prediction, LP) on herkkä tällaiselle virheelle. Painotetulla lineaariprediktiolla (weighted linear prediction, WLP) [10] sen sijaan voidaan vähentää harmonisten vaikutusta spektriin ja näin ollen estimointivirhettä.

Tässä työssä verrataan LP:n kykyä estimoida puheen ja huudon spektriä kahteen eri WLP tekniikkaan: WLP lyhytaikaisen energian (short-time energy, STE) [11] painotuksella sekä WLP pääherätteen vaimennuksella (attenuation of the main excitation, AME) [12]. WLP-STE:ssä puheesta lasketaan lyhytaikainen energiafunktio, joka korostaa puheen kohtia joissa heräte on heikoimmillaan. Näin ollen tällä funktiolla voidaan vähentää äänilähteen harmonisten vaikutusta spektriestimointiin. WLP-AME:ssa taas vaimennetaan herätteen vaikutusta tunnistamalla puhesignaalista glottiksen sulkeutumisajankohdat, ja vaimentamalla näitä kohtia spektrin estimoinnissa. WLP-STE -tekniikassa myös käytetään suodinmallin stabilointia [11].



Kuva 4: ABX-testin osoittama analyysi-synteesin laatu kolmella eri spektriestimointimetodilla.

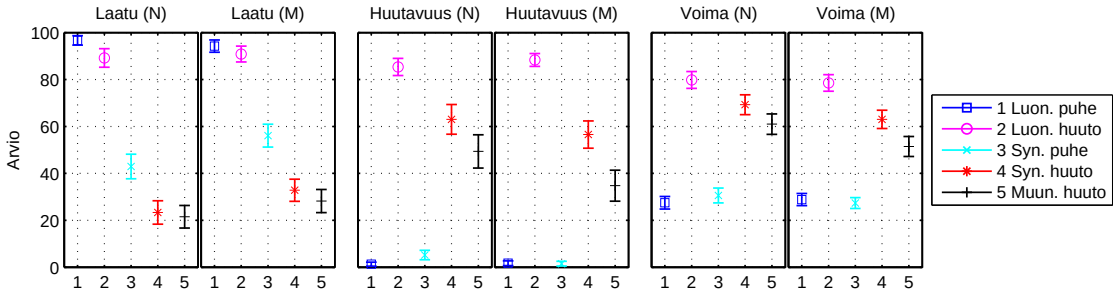


Kuva 5: CCR-testin osoittama adaptaation laatu naisäänelle (vasemalla) ja miesäänelle (oikealla).

Näitä kolmea eri metodia verrataan puheen ja huudon analyysi-synteesin avulla. Työssä käytetään GlottHMM vokooderia [13], joka hyödyntää äänilähteen käänteissuodatusta puheen analyysissä ja luonnollisia glottispulsseja äänilähteen synteesissä. GlottHMM vokooderia käytetään tilastolliseen parametriseen puhesynteesiin, ja se on osoittautunut hyväksi työkaluksi syntetisoitaessa mm. Lombard-puhetta [14]. Testissä puhe- ja huutosignaalit syötetään vokooderille, joka hajottaa ne parametreiksi ja rakentaa jälleen parametreista puhetta käyttäen edellämainittuja kolmea eri spektriestimointimenetelmää. Näitä rekonstruoituja signaaleita arvioidaan subjektiivisissa kuuntelukokeissa, joissa koehenkilön tehtävä on arvioida kahdesta rekonstruoidusta signaalista se, joka on lähempänä alkuperäistä signaalia (ns. ABX-testi). 15 suomenkielistä kuuntelijaa arvioi kukin 60 näyteparia, jotka koostuivat 20:sta kullakin metodilla rekonstruoidusta signaalista. Kokeen tulokset on esitetty kuvassa 5, jotka osoittavat, että AME-WLP suoriutuu parhaiten normaalilla puheella ja STE-WLP huudolla.

4.1 HMM-pohjainen synteesi

Tilastollinen parametrinen puhesynteesi [15], tai toisin sanoen Markovin piilomalleihin (hidden Markov model, HMM) perustuva puhesynteesi, on joustavin tapa tuottaa erilaisia puhetyylejä kuten huutoa. Tässä työssä käytetään adaptointimenelmää [16], jossa normaalista puheesta opetetaan tavallinen synteesi, joka tilastollisesti muutetaan (eli adaptoidaan) vastaamaan huutoa. Kaksi synteettistä ääntä (mies ja nainen) rakennettiin käyttäen standardia HMM-opetusta [17] ja adaptoitiin äänitettyjen 100:n lauseen avulla vastaamaan huutosynteesiä. Subjektiivisissa kokeissa testattiin, vaikuttaako spektriestimointimenetelmä adaptaation laatuun. Stabiloitua STE-WLP:tä ja tavallista LP:tä käytettiin adaptaatiomateriaalin estimoimiseen ja tällä saadulla datalla adaptoitiin normaali ääni huudoksi. Huutosynteesin laatua testattiin subjektiivisillä kuuntelukokeilla (ns. CCR-testi), joissa kuulija arvioi kahden ääninäytteen välillä olevan laatueron CMOS-asteikolla, joka on jatkuva asteikko ulottuen -3:sta (paljon huonompi) +3:een (paljon parempi). 11 koehenkilöä arvioi kukin 40 signaaliparia, joista laskettiin kummallekin metodille keskiarvot. Tulokset on esitetty kuvassa 5, jotka osoittavat että STE-WLP on huomattavasti parempi naisäänellä ja hieman parempi miesäänellä kuin LP.



Kuva 6: *MOS-testin tulokset eri puhe- ja huutomateriaaleilla.*

5 HUUTOSYNTEESIN EVALUOINTI

Tuotettua huutosynteesiä arvioitiin kuuntelukokeilla, joissa haluttiin selvittää mm. miten synteettinen huuto havaitaan suhteessa aitoon huutoon. Seuraavat puhetyyppit valittiin kokeisiin: 1) Luonnollinen puhe 2) Luonnollinen huuto 3) Synteettinen puhe 4) Synteettinen huuto (adaptoitu) 5) Synteettinen huuto (äänenmuuntelu). Referenssiksi (5) testiin otettiin synteettisestä normaalista puheesta äänenmuuntelulla (voice conversion) muokattu huutoa muistuttava puhe. Äänenmuuntelu on keino, jolla puhesignaalia voidaan keinotekoisesti muokata kuulostamaan joko toiselta puhujalta tai puhetyyliltä.

Ääniä arvioitiin MOS-tyyppisessä kokeessa, jossa kysyttiin seuraavia ominaisuuksia: 1) *Millainen on puheäänen laatu?* 2) *Kuinka paljon näyte muistuttaa huutoa?* ja 3) *Kuinka paljon voimaa puhuja käyttää äänen tuottamiseen?* Arviot annettiin verbaalisin kuvauksin ankkuroiduilla jatkuvilla asteikoilla skaalalla 1–5. Ääninäytteiden voimakkuus normalisoitiin käyttämällä standardoitua metodia [18], jotta koehenkilöt eivät havainneet huutoa äänenvoimakkuuden vaan muiden ominaisuuksien perusteella. 11 koehenkilöä arvioi kukin yhteensä 50 näytettä, 10 näytettä kustakin kategoriasta. Evaluoinnin tulokset on esitetty kuvassa 6. Tulokset osoittavat, että adaptoitu huutosynteesi on laadultaan heikompi kuin tavallinen puhesynteesi, mutta huudon vaikutelma ja voiman käyttö äänessä on kuitenkin hyvin säilynyt, toisin kuin äänenmuuntelulla tuotetulla huudolla.

6 DISKUSSIO JA YHTEENVETO

Huudon syntetisointi on haastavaa monestakin syystä. Ensiksi, on erittäin vaikeaa äänittää suuria määriä huutoa tasaisella laadulla. Toiseksi, puheen ja huudon erot ovat suuret, mikä aiheuttaa ongelmia puheenkäsittelymetodeissa. Tässä työssä paneuduttiin ehkäisemään suuren F0:n aiheuttamaa spektrivirhettä käyttämällä painotettuja lineaariprediktiomalleja. Tutkimuksessa tuotettiin synteettistä huutoa tavallisesta HMM-synteesistä sekä adaptoimalla että puheenmuunnoksella. Subjektiiviset kuuntelukokeet osoittavat, että metodit aiheuttavat erilaisia artefakteja aiheutuen mm. edellämainituista haasteista: huutodatan vähäisestä määrästä ja huudon erilaisista ominaisuuksista puheeseen verrattuna. Adaptaatiossa nämä ongelmat aiheuttivat artefakteja etenkin puheen prosodiassa, mutta vaikutelma huudosta ja voiman käyttö äänessä säilyi hyvin, toisin kuin puhujamuunnoksessa, jossa on konsistentimpi prosodia mutta vähemmän huudon piirteitä.

KIITOKSET

Tätä työtä on tukenut EU:n FP7 Simple4All -projekti ja Suomen Akatemia.

VIITTEET

[1] Lombard, E., “Le signe de l’elevation de la voix”, Ann. Maladies Oreille, Larynx, Nez, Pharynx, 37:101–119, 1911.
[2] Pickett, J., “Effects of Vocal Force on the Intelligibility of Speech Sounds”, J. Acoust. Soc. Am., 28(5):902–905, 1956.
[3] Rostolland, D., “Intelligibility of shouted voice”, Acustica 57(3):103–121, 1985.
[4] Rostolland, D., “Acoustic features of shouted voice”, Acustica 50(2):118–125, 1982.
[5] Alku, P., Airas, M., Björkner, E. and Sundberg, J., “An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity”, J. Acoust. Soc. Am., 120(2):1052–1062, 2006.
[6] Rostolland, D., “Phonetic structure of shouted voice”, Acustica 51(2):80–89, 1982.
[7] Elliott, J., “Comparing the acoustic properties of normal and shouted speech: a study in forensic phonetics”, Proc. SST-2000: 8th Int. Conf. Speech Sci. & Tech., 2000, pp. 154–159.
[8] Zhang, C. and Hansen, J., “Analysis and classification of speech mode: whispered through shouted”, Proc. Interspeech, 2007, pp. 2289–2292.
[9] Pohjalainen, J., Raitio, T., Yrttiaho, S. and Alku, P., “Detection of shouted speech in noise: human and machine”, J. Acoust. Soc. Am., 133(4), 2013 (accepted).
[10] Ma, C., Kamp, Y. and Willems, L., “Robust signal selection for linear prediction analysis of voiced speech”, Speech Commun., 12(1):69–81, 1993.
[11] Magi, C., Pohjalainen, J., Bäckström, T. and Alku, P., “Stabilised weighted linear prediction”, Speech Comm. 51(5):401–411, 2009.
[12] Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A.-M. and Story, B., “Improved formant frequency estimation from high-pitched vowels by downgrading the contribution of the glottal source with weighted linear prediction”, Proc. Interspeech, 2012.
[13] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M. and Alku, P., “HMM-based speech synthesis utilizing glottal inverse filtering”, IEEE Trans. on Audio, Speech, and Lang. Proc., 19(1):153–165, 2011.
[14] Raitio, T., Suni, A., Vainio, M. and Alku, P., “Analysis of HMM-based Lombard speech synthesis”, Proc. Interspeech, 2011, pp. 2781–2784.
[15] Zen, H., Tokuda, K. and Black, A.W., “Statistical parametric speech synthesis”, Speech Commun., 51(11):1039–1064, 2009.
[16] Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K. and Isogai, J., “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm”, IEEE Trans. on Audio, Speech, and Lang. Proc., 17(1):66–83, 2009.
[17] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W. and Tokuda, K., “The HMM-based speech synthesis system (HTS) version 2.0”, Sixth ISCA Workshop on Speech Synthesis, 2007, pp. 294–299.
[18] ITU, “Objective measurement of active speech level”, International Telecommunication Union, Recommendation ITU-T P.56, 2011.