PUHEEN ARTIKULATORINEN MALLINNUS JA PUHEINVERSIO

Heikki Rasilo, Unto K. Laine

Aalto-yliopiston Sähkötekniikan Korkeakoulu
Signaalinkäsittelyn ja Akustiikan laitos
PL 13000
00076 AALTO
etunimi.sukunimi@aalto.fi

1 JOHDANTO

Puheen monimuotoisuus tekee siitä ainutlaatuisen tiedonvälitystavan ihmisten välisessä kommunikaatiossa. Ihminen on muista eläimistä poiketen oppinut kontrolloimaan artikulatorista koneistoaan siten, että se mahdollistaa erittäin monipuolisen äännemaailman tuottamisen. Puheen kuvaaminen artikulatorisessa avaruudessa olisi oletettavasti monin tavoin sopivampi esitysmuoto kuin akustiset parametrit. Yksittäiset artikulaatioelimet liikkuvat hitaasti verrattuna puheen spektraalisiin ominaisuuksiin. Puhetta tuotettaessa artikulatoristen elinten suhteelliset liikeradat ovat verrattain samanlaisia eri puhujien välillä, mikä saattaa helpottaa ratkaisemaan monia ongelmia liittyen puhujariippuvaan vaihteluun akustisessa alueessa. Koartikulaation vuoksi myös akustisesti samanlaiset tilanteet saattavat erottua paremmin artikulatorisessa maailmassa, jos liikkeiden dynamiikka myös pidemmältä aikaväliltä saadaan huomioitua. Luotettava puheinversio artikulatorisiin parametreihin on yhä ratkaisematon ongelma, jonka ratkaisu johtaisi oletettavasti lukuisiin teknologisiin innovaatioihin, kuten uusiin tehokkaisiin menetelmiin puhekoodauksessa, puheanalyysissä ja –synteesissä, puheentunnistuksessa sekä mm. puheterapian tarpeisiin.

Puheen artikulatorisessa mallinnuksessa ja inversiossa käytetään usein apuna staattisia tai dynaamisia ääntöväylämalleja, joilla pyritään kuvaamaan ihmisen ääntöväylän tärkeimpiä ominaisuuksia vähäisellä määrällä parametreja. Mermelsteinin [1] ja Maedan [2] artikulatorisia malleja on käytetty paljon puheentutkijoiden keskuudessa. Mallien säädettävät parametrit ja niiden vaikutus ääntöväylän profiiliin on estimoitu analysoimalla ääntöväylän röngen-kuvasarjoja. 1990- ja 2000-luvulla on kehitetty yhä monipuolisempia ääntöväylä- ja kielimalleja. Esim. Dang ja Honda [3] ovat kehittäneet 3-ulotteisen artikulatorisen mallin, joka hyödyntää suun lihasten fyysisiä rajoitteita. Mallia käytettiin mm. vokaali-vokaali siirtymien inversioon.

1.1 Puheinversio ongelma

Puheinversio-ongelmassa koetetaan löytää äänitetyistä puhesignaaleista niiden tuottamiseen käytetyt artikulatoriset liikkeet. Ihmiset kykenevät suorittamaan puheinversiota jatkuvasti, joka mahdollistaa mm. puheen matkimisen. Yksi klassisen *puheen havaitsemisen motorisen teorian* [esim. 4] pääväitteistä on, että ihmiset käyttävät puheen artikulatorista esitystapaa apuna myös puhetta havaittaessa; puhesignaali ikään kuin matkitaan mielessä oman puheentuottomekanismin kautta, tuoden lisää luotettavuutta puhehavaintoon. Useat koejärjestelyt puhuvat tämän väitteen puolesta. Aivojen motoristen alueiden aktivoituminen puhehavaintojen aikana on varmistettu fMRI mittauksissa [5]. Kokeissa on myös huomattu että magneettinen stimulaatio huulten tai kielen liikkeistä vastaavalle motoriselle aivokuorelle vaikuttaa huulilla tai kielellä artikuloitujen foneemien erottelukykyyn [6].

Puheinversio-ongelma on ns. *ill-posed* –ongelma, joka tekee siitä vaikeasti lähestyttävän. Suora ongelma artikulatorisista parametreista akustisiin parametreihin voidaan kuvata yhtälöllä $\mathbf{Az} = \mathbf{u}$, $\mathbf{z} \in Z$, missä **A** on puheen tuottomekanismia kuvaava jatkuva operaattori (esim. artikulatorinen malli), **z** on vektori artikulatorisista parametreista ja **u** on näiden avulla syntetisoitu yksiselitteinen akustinen vektori. Käänteinen ongelma, eli vektorin **z** löytäminen käyttämällä mallia **A** ja vektoria **u** on huonosti käyttäytyvä. Eroavaisuus estimoidussa operaattorissa **A** ja todellisessa puhujan puheentuottomekanismissa saattaa aiheuttaa sen, että ratkaisua **z** ei välttämättä löydy mahdollisten ratkaisujen Z joukosta. Vaikka malli **A** olisikin täydellinen, ei ratkaisu silti yleisessä tapauksessa ole yksiselitteinen: monet artikulatoriset konfiguraatiot voivat tuottaa samat akustiset vektorit (esim. vatsastapuhujat käyttävät tätä ominaisuutta hyödykseen).

Huonosti käyttäytyvien ongelmien ratkaisua voidaan lähestyä käyttämällä rajoitteita, jotka rajoittavat ratkaisujen joukkoa jollakin järkevällä tavalla. Puheinversiossa huomioitavia rajoitteita ja ratkaisutapoja on käsitellyt töissään mm. Viktor N. Sorokin [esim. 7]. Rajoitteina on mainittu mm. puheen kielelliset ja akustiset ominaisuudet, samoin kuin anatomiset rajoitteet. Artikulatoristen parametrien dynamiikka, vaihtelualueet, niiden mahdollinen synkronismi, lihaksiin vaikuttavat maksimivoimat tai niitä ohjaavien motoristen komentojen luonne ovat esimerkkejä mahdollisista rajoitteista.

Täsmälliset artikulaatiota koskevat rajoitteet eivät ole selvillä, mikä vaikeuttaa inversio-ongelman ratkaisua. Sorokin esittää, että tasaisten vokaaliäänteiden tapauksessa artikulaatioelimet pyrkivät sijoittumaan siten, että ne ovat mahdollisimman vähän poikkeutettuina neutraaliasemistaan. Dynaamisen ääntöväylän inversioon puolestaan tarvitaan artikulaatioelinten nopeuksiin tai kiihtyvyyksiin liittyviä rajoitteita. Tällaisen parametreihin liittyvän optimointifunktion $\Omega(\mathbf{z})$ lisäksi on otettava huomioon estimoidun puheentuotto-operaattorin $\mathbf{A}_\mathrm{h}$ poikkeavuus alkuperäisen puhujan operaattorista **A**. On siis sallittava tietty määrä poikkeavuutta alkuperäisessä mitatussa akustisessa vektorissa $\mathbf{u}_\delta$ ja estimoidun mallin tuottamassa vektorissa $\mathbf{A}_\mathrm{h}\mathbf{z}$: $\rho(\mathbf{A}_h\mathbf{z}, \mathbf{u}_\delta) = \left\|\mathbf{A}_h\mathbf{z} - \mathbf{u}_\delta\right\|$. Inversio-ongelma voidaan siis kuvata kustannusfunktion

$$M(\mathbf{z}_{h\delta}) = \Omega(\mathbf{z}) + \rho^2(\mathbf{A}_h\mathbf{z}, \mathbf{u}_\delta)/\beta \tag{1}$$

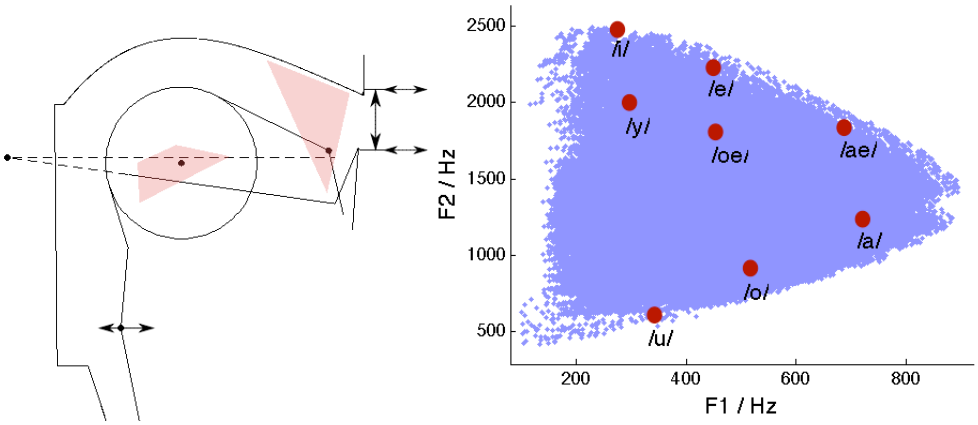
minimointiongelmana. β kuvaa kompromissia akustisen eron ja optimointifunktion välillä.

Inversio-ongelman ratkaisua voidaan tehostaa luomalla koodikirjoja, jotka sisältävät vastaavuuksia akustisten ja artikulatoristen vektorien välillä. Koodikirjojen avulla voidaan estimoida karkeita trajektoreja, joista voidaan iteratiivisesti lähestyä optimaalisempaa ratkaisua [esim. 8]. Ouni ja Laprie [9] käyttivät kehittyneempää koodikirjaa, jossa koodikirja on jaettu pieniin hyperkuutioihin joiden sisällä kuvaus artikulatorisesta alueesta akustiseen on lineaarinen. Löydetystä aloitusratkaisusta etsittiin lopullinen pehmeä trajektori ratkaisemalla kustannusfunktiosta formuloidut Euler-Lagrange yhtälöt variaatiolaskennan avulla.

Tässä työssä esitellään luomamme dynaaminen artikulatorinen malli, ja sen avulla luotu koodikirja inversiokokeita varten. Koodikirjahaussa olemme käyttäneet apuna kaksinapaista prediktoria, joka pyrkii säilyttämään artikulatoristen parametrien dynamiikan ja vähentämään koodikirjan karkeudesta johtuvia ongelmia. Testataksemme motorisen teorian hypoteeseja teemme sekä akustisilla vektoreilla että niiden inversioilla suomenkielisten tavujen luokittelukokeita.

2 ARTIKULATORINEN MALLI JA KOODIKIRJA

Käyttämämme artikulatorinen malli perustuu pääosiltaan Mermelsteinin malliin [1]. Tiettyjä muutoksia alkuperäiseen kuvaukseen tehtiin, tarkoituksenamme pystyä mallintamaan ihmispuhujalle tyypillinen vokaaliavaruus. Kuvassa 1 (vas.) näkyy ääntöväylämallin rakenne.



Kuva 1. Vasemmalla: Ääntöväylämalli ja sen liikkuvat parametrit. Oikealla: Ääntöväylämallilla luotu koodikirja esitettynä F1-F2-tasossa.

Ääntöväylän kahdeksan parametrin paikat on kuvattu keskisagittaalisessa tasossa koordinaatteina, ja niiden avulla ääntöväylän ääriviivat lasketaan tiettyjen geometristen funktioiden avulla. Parametreina toimivat kieliluun x-koordinaatti, kielen tyven x- ja y-koordinaatit, kielen kärjen x- ja y-koordinaatit, leuan kulma, huulten pituus ja huulten avoimuus.

Väylän pinta-alafunktio lasketaan jakamalla ilmatila tasaisesti 16 sektioon. Janojen pituuksia käytetään 16 sama-akselisten sylinterinmuotoisten putkien halkaisijoina. Pinta-alafunktio saadaan näiden putkien poikkileikkausten pinta-aloina liikuttaessa glottiksesta huulille. Kunkin segmentin pinta-ala skaalataan vielä vakiokertoimilla, jotta ääntöväylän ilmatilan vaihteleva leveys saadaan tarkemmin huomioitua. Parametrien säätö suoritettiin vertailemalla mallia MRI:llä mitattuihin keskisagittaalikuviin ja pinta-ala funktioihin [10]. Pinta-ala funktion avulla vastaavat äänet on syntetisoitu käyttämällä Kelly-Lochbaum tyyppistä siirtolinjaa [11], joka huomioi huulisäteilyn vaikutuksen. Myös dynaamisia siirtymiä ja *jokellusta* saadaan onnistuneesti syntetisoitua muuttamalla parametrien arvoja jatkuvasti.

Akustinen-artikulatorinen koodikirja luodaan vaihtelemalla parametreja tasavälisesti niille sallittujen rajojen sisällä ja syntetisoimalla äänet. Akustisina parametreina käytetään LPC-analyysilla saatavia kolmea ensimmäistä formanttitaajuutta. Koodikirja koostuu yhteensä 212,500:sta formantti-parametri paristä. Kuvan 1 oikea puoli esittää koodikirjan alkioita F1-F2-tasossa. Tunnettu *vokaalikolmio* on selkeästi havaittavissa. Kuvaan on lisätty suomenkielen vokaaliäänteiden F1 ja F2 arvot [12].

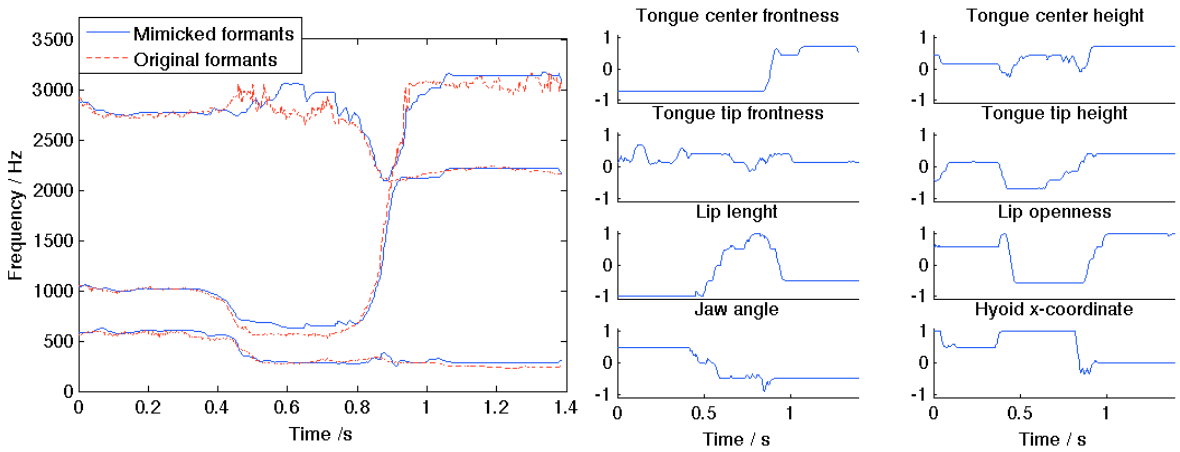
3 INVERSIOKOKEET

Tässä työssä inversiokokeita tehdään koodikirjan avulla ikkunoiduista äänisignaaleista artikulatorisen mallin parametreihin. Jokaiselle puhesignaalin ikkunalle etsitään kolme ensimmäistä formanttitaajuutta, joiden avulla koodikirjasta valitaan N lähintä artikulatorista

kandidaattia käyttäen formanttivektorien Euklidista etäisyyttä etäisyysmittana. Formanttien painotusta voidaan muuttaa kokeellisin skalaarein; kolmannelle formantille on syytä sallia enemmän vaihtelua kuin ensimmäiselle tai toiselle formantille. Tämän vaiheen jälkeen ratkaistavana on hankala optimaalisen polun etsimisongelma. Jokaiselle aikaikkunalle $t = 1 \ldots T$ on olemassa N kandidaattia, joten mahdollisia polkuja joukon läpi on N^T kappaletta. Koska kaikkien mahdollisten kombinaatioiden läpikäyminen on käytännössä mahdotonta, täytyy ongelmaan löytää approksimatiivinen ratkaisutapa.

Yritämme löytää ongelmaan mahdollisimman tehokasta lähestymistapaa, joka mahdollistaa suuren näytemäärän inversion siedettävässä ajassa inversiotulosten tilastollisia testejä varten. Aikaisemmin olemme käyttäneet ”ahnetta” (eng. *greedy*) dynaamista ohjelmointialgoritmia, joka valitsee seuraavaan ikkunaan aina edellistä ikkunaa lähimpänä olevan kandidaatin, lähtien liikkeelle jokaisesta ensimmäisen aikaikkunan kandidaatista [13]. Tämä metodi luo pehmeyttä artikulatorisille liikeradoille, mutta ei ota lihasten liikemääriä huomioon, ja saattaa aiheuttaa epätodenmukaisia kiihtyvyyksiä vierekkäisten ikkunoiden välille.

Uudempi inversiomenetelmä käyttää kaksinapaista, alipäästötyyppistä prediktoria, joka pyrkii ennustamaan seuraavan ikkunan kandidaattia kahden edellisen ikkunan parametrin avulla. Metodi käy läpi myös suuremman määrän mahdollisia polkukandidaatteja ja valitsee lopulliseksi ratkaisuksi polun, joka käyttää minimimäärän kiihtyvyyttä koko lausahduksen läpi. Tämä malli säilyttää paremmin parametrien liikkeiden dynamiikan ja vähentää myös koodikirjan karkeuden aiheuttamia hyppäyksiä trajektoreissa [14]. Kuva 2 esittää vokaalisiirtymän /aui/ inversiota. Alkuperäiset formantit ja inversiotuloksen avulla matkitut formantit näkyvät vasemmanpuoleisessa kuvassa. Oikea kuva näyttää löydetyt parametriarvot. Monien parametriarvojen voidaan intuitiivisesti todeta vastaavan todellisuutta: /i/ äänteessä kielen tyvi siirtyy eteen ja ylöspäin, leuan aukeamiskulma on pieni ja huulet ovat voimakkaasti avonaiset. /u/ äänteessä huulet ovat työntyneet eteenpäin ja lähes suljetut, sekä kurkunpää on laajentunut.



Kuva 2. Vokaalisiirtymän /aui/ inversio. Vasemmalla näkyy alkuperäisen äänteen formantit katkoviivalla ja matkitun äänen formantit yhtenäisellä viivalla. Oikealla inversion tulos parametriarvoina skaalattuna välille [-1, 1].

Tällä tavalla ratkaistuna inversio tuottaa sellaiset parametriarvot, joita käyttämällä artikulatorinen malli ”matkii” alkuperäisen äänen formanttitaajuuksien mielessä. Testeissä kuitenkin on huomattu, että inversiotulos ei ole stabiili pienille vaihteluille inversion parametreissa. Esimerkiksi muutettaessa formanttien painotusta, saattaa inversiotulos näyttää

parametriavaruudessa hyvin erilaiselta, ja kompensatorisia muutoksia parametrien välillä saattaa tapahtua.

Inversion laatua on pyritty testaamaan tilastollisella suomenkielen tavujen luokittelutestillä. Hypoteesinamme on ollut, että artikulatoristen parametrien mielessä eri puhujien lausumat tavut luokittuisivat luotettavammin annotaation mukaisiin luokkiin. Tavut on poimittu suomenkielisestä puhetietokannasta. Testeissä on käytetty yhteensä 2500 tavua: 20 kappaletta tavuja jokaista 125 annotoitua tavuluokkaa kohti. Eri tavuluokissa esiintyy kahdesta kuuteen eri miespuhujaa.

Tilastollisia luokittelutestejä tehdään uudelleen näytteistämällä inversiotulokset ja alkuperäiset formanttivektorit samaan pituuteen (60 ms) ja luokittelemalla erikseen inversiotuloksia sekä formanttitrajektoreja *k-means* klusterointialgoritmilla. Luokkien määräksi valitaan tavuluokkien lukumäärä 125. Kunkin klusterin selektiivisyys $S(c) = n_c^{\alpha_{\max}} / n_c$ kertoo klusterin yleisimmän tavun, $n_c^{\alpha_{\max}}$, osuuden kaikista klusteriin valikoituneista tavuista n_c. Keskimääräinen kokonaisselektiivisyys lasketaan painottamalla klusterien selektiivisyyttä klusterin alkioiden kokonaismäärällä ja keskiarvoistamalla:

$$S_{avg} = \sum_{c=1}^{125} n_c S(c) / \sum_{c=1}^{125} n_c \tag{2}$$

Ensimmäiset tulokset [14] puhuivat sen puolesta, että inversion avulla samat tavut luokittuivat hieman luotettavammin inversion, kuin formanttitrajektorien avulla (S_{avg} = 30.02 % ja 27.54 % vastaavasti). Painottamalla formanttitaajuuksia siten, että toiselle ja kolmannelle formantille sallitaan enemmän vaihtelua, nostaa kuitenkin myös formanttien avulla lasketun selektiivisyyden yli 30 %:iin.

Yllättäen uusimmat inversiotulokset näyttävät, että jos inversiovaiheessa jokaiselle ikkunalle löydettyä 100:a parametrivektorikandidaattia keskiarvoistetaan ja käytetään tätä tulosta luokittelussa, saadaan tähänastisista luokittelutesteistä kaikkein paras tulos. Tämä todennäköisesti viittaa siihen, että kun artikulatorisesti perusteltua koodikirjaa käytetään yksittäisten ikkunoiden formanttitaajuuksien inversiossa, suuri osa lähimmistä koodikirjan formanttikandidaateista sattuu parametriavaruudessa lähelle todellista ääntöväylämuotoa. Toisin sanoen, artikulatoriset trajektorit saattavat kulkea pääosin sellaisia ratoja, joiden tuottamat akustiset signaalit ovat *vähiten herkkiä pienille vaihteluille artikulaatiossa*. Vaihteluvälille [-1, 1] painotetuilla formanttitrajektoreilla ja edellä kuvatulla tilastollisella inversiometodilla saadut selektiivisyystulokset on esitetty taulukossa 1. Tulokset on keskiarvoistettu 30 *k-means* klusteroinnin tuloksesta, koska alkuperäisten klusterikeskusten valinta voi vaikuttaa klusteroinnin lopputulokseen.

Taulukko 1. *Klusterien selektiivisyydet käyttämällä normalisoituja alkuperäisiä formanttitrajektoreita, sekä yksinkertaista koodikirjan avulla suoritettua inversiota.*

	Klusterin selektiivisyys S_{avg}	Keskihajonta
Formanttitrajektorit (painotettu)	30.94 %	0.71 %
Tilastollinen inversiometodi	32.61 %	0.73 %

4 YHTEENVETO

Olemme luoneet dynaamisen artikulatorisen mallin, joka pystyy tuottamaan koko ihmispuhujalle tyypillisen vokaaliavaruuden käyttämällä parametreinaan kahdeksaa artikulaatioelimien paikkoja kuvaavaa koordinaattia. Mallia käyttämällä on kehitetty ja testattu uusia tehokkaita puheinversiomenetelmiä. Inversiotulosten evaluonnissa on käytetty suomenkielen tavujen klusteroitumistestejä hypoteesinamme, että eri puhujien lausumat samat tavut luokittuisivat luotettavammin artikulatorisia parametreja, kuin alkuperäisiä formanttitaajuuksia käyttäen.

Inversiossa on testattu liikkeiden dynamiikkaa mallintavaa prediktorirakennetta, mutta uusimmat testit näyttävät, että yksinkertainen koodikirjan lähimpiä kandidaatteja keskiarvoistava inversio johtaa vielä parempiin luokittelutuloksiin. Tulos viittaa siihen, että tavutasolla artikulatoriset liikeradat muodostavat sellaisia polkuja, joista vähäinen poikkeaminen luo mahdollisimman pienet muutokset akustisissa trajektoreissa.

5 VIITTEET

1. MERMELSTEIN, P., Articulatory model for the study of speech production, *J. Acoust. Soc. Am.*, 53(4), 1070-1082, 1973.
2. MAEDA, S., *Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model*, W. J. Hardcastle and A. Marchal, Speech production and speech modeling, 131-149, Kluwer Acad. Publ., 1990.
3. DANG, J., HONDA, K., Estimation of vocal tract shapes from speech sounds with a physiological articulatory model, *Journal of Phonetics*, 30, 511-532, 2002.
4. LIBERMAN, A., AND MATTINGLY, I., The motor theory of speech perception revised, *Cognition,* 21, 1-36, 1985.
5. WILSON, STEPHEN M., et al., Listening to speech activates motor areas involved in speech production, *Nature Neuroscience*, 7, 701–702, 2004.
6. D'AUSILIO A., ET.AL., The Motor Somatotopy of Speech Perception, *Current Biology*, 19, 381-385, 2009.
7. SOROKIN, VICTOR N., *Speech Inversion: Problems and Solutions, Dynamics of Speech Production and Perception*, 263-282, P. Divenyi et al. (Eds.), IOS Press 2006.
8. LAPRIE T, & MATHIEU B, A variational approach for estimating vocal tract shapes from the speech signal. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2, 929–932, May 1998.
9. OUNI, S. & LAPRIE, Y., Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion, *J. Acoust. Soc. Am*, 118 (1), 444-460, 2005.
10. STORY, B. H., TITZE, I. R., & HOFFMAN, E. A., Vocal tract area functions from magnetic resonance imaging, *J. Acoust. Soc. Am*., 100(1), 537-554, 1996.
11. KELLY, J. L. & LOCHBAUM, C. C., Speech Synthesis, *Proc 4th Int. Congr. Acoustics*, Copenhagen, 1-4, 1962.
12. WIIK, K., *Finnish and English vowels*, University of Turku, Turku, 1965.
13. RASILO H., et. al. Estimation studies of vocal tract shape trajectory using a variable length and lossy Kelly-Lochbaum model, *Proc. Interspeech'10*, Japan, 2414-2417, 2010.
14. RASILO H., et. al., Method of speech inversion with large scale statistical evaluation, submitted.