

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Ingrida MAŽONAVIČIŪTĖ

LIETUVIŲ KALBOS ANIMAVIMO TECHNOLOGIJA TAIKANT TRIMATĮ VEIDO MODELĮ

DAKTARO DISERTACIJOS SANTRAUKA

TECHNOLOGIJOS MOKSLAI,
INFORMATIKOS INŽINERIJA (07T)



Vilnius LEIDYKLA
TECHNIKA 2012

Disertacija rengta 2008–2012 metais Vilniaus Gedimino technikos universitete.
Mokslinis vadovas

prof. habil. dr. Romualdas BAUŠYS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T).

Disertacija ginama Vilniaus Gedimino technikos universiteto Informatikos inžinerijos mokslo krypties taryboje:

Pirmininkas

doc. dr. Arnas KAČENIAUSKAS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T).

Nariai:

prof. habil. dr. Antanas ČENYS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T),

prof. habil. dr. Gintautas DZEMYDA (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07T),

prof. habil. dr. Rimvydas SIMUTIS (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija – 07T),

prof. dr. Olegas VASILECAS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T).

Oponentai:

prof. dr. Dalius NAVAKAUSKAS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T),

prof. habil. dr. Rimantas ŠEINAUSKAS (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija – 07T).

Disertacija bus ginama viešame Informatikos inžinerijos mokslo krypties tarybos posėdyje 2013 m. sausio 8 d. 10 val. Vilniaus Gedimino technikos universiteto senato posėdžių salėje.

Adresas: Saulėtekio al. 11, LT-10223 Vilnius, Lietuva.

Tel.: (8 5) 274 4952, (8 5) 274 4956; faksas (8 5) 270 0112;

el. paštas doktor@vgtu.lt

Disertacijos santrauka išsiuntinėta 2012 m. gruodžio 7 d.

Disertaciją galima peržiūrėti Vilniaus Gedimino technikos universiteto (Saulėtekio al. 14, LT-10223 Vilnius, Lietuva) ir Matematikos ir informatikos instituto (Akademijos g. 4., LT-08663 Vilnius, Lietuva) bibliotekose.

VGTU leidyklos „Technika“ 2081-M mokslo literatūros knyga.

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

Ingrida MAŽONAVIČIŪTĖ

LITHUANIAN SPEECH ANIMATION TECHNOLOGY FOR 3D FACIAL MODEL

SUMMARY OF DOCTORAL DISSERTATION

TECHNOLOGICAL SCIENCES,
INFORMATICS ENGINEERING (07T)



LEIDYKLA
Vilnius TECHNIKA 2012

Doctoral dissertation was prepared at Vilnius Gediminas Technical University in 2008–2012.

Scientific Supervisor

Prof Dr Habil Romualdas BAUŠYS (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T).

The dissertation is being defended at the Council of Scientific Field of Informatics Engineering at Vilnius Gediminas Technical University:

Chairman

Assoc Prof Dr Arnas KAČENIAUSKAS (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T).

Members:

Prof Dr Habil Antanas ČENYS (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T),

Prof Dr Habil Gintautas DZEMYDA (Vilnius University, Technological Sciences, Informatics Engineering – 07T),

Prof Dr Habil Rimvydas SIMUTIS (Kaunas University of Technology, Technological Sciences, Informatics Engineering – 07T),

Prof Dr Olegas VASILECAS (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T).

Opponents:

Prof Dr Dalius NAVAKAUSKAS (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T),

Prof Dr Habil Rimantas ŠEINAUSKAS (Kaunas University of Technology, Technological Sciences, Informatics Engineering – 07T).

The dissertation will be defended at the public meeting of the Council of Scientific Field of Informatics Engineering in the Senate Hall of Vilnius Gediminas Technical University at 10 a. m. on 8 January 2013

Address: Saulėtekio al. 11, LT-10223 Vilnius, Lithuania.

Tel.: +370 5 274 4952, +370 5 274 4956; fax +370 5 270 0112;

e-mail: doktor@vgtu.lt

The summary of the doctoral dissertation was distributed on 7 December 2012.

A copy of the doctoral dissertation is available for review at the Library of Vilnius Gediminas Technical University (Saulėtekio al. 14, LT-10223 Vilnius, Lithuania) and at the Library of Institute of Mathematics and Informatics (Akademijos str. 4, LT-08663 Vilnius, Lithuania).

Įvadas

Mokslo problemos aktualumas. Skaitymas iš lūpų judesių suprantamas kaip psichologinis procesas padedantis žmonėms suprasti akustinę šnekos informaciją. Tai aktualu ne tik klausos sutrikimų turintiems, bet ir gerai girdintiems žmonėms. Mokslininkai ištyrė, kad veido vaizdinės informacijos naudojimas 57 % padidina priebalsių, 30% – balsių, 39% – viensikiemenių žodžių ir 17% – trumpų frazių suprantamumą. Tai ypač svarbu esant nepalankioms akustinėms sąlygoms.

Kalbos animacija plačiai naudojama technikos įrenginiuose siekiant kurtiesiems, vaikams, vidutinio ir vyresnio amžiaus žmonėms sudaryti vienodas bendravimo galimybes. Ji yra labai svarbi žmogaus ir kompiuterio sąveikos mokslo dalis, todėl kelia didelį mokslininkų, technologų ir menininkų susidomėjimą. Dvimačiai ir trimačiai kompiuteriniai veido modeliai, skirti žmogaus kalbai animuoti sinchronizuojant akustinį ir regimąjį kalbos signalus („kalbančios galvos“), gali būti naudojami elektroninio konsultavimo versle (virtualūs pagalbininkai, interneto tinklalapių navigatoriai, agentai bendraujantys su žmogumi „protinguose“ namuose) arba kaip kompiuteriniai 3D modeliai, galintys padėti mokymo procese (tarti garsus, mokyti užsienio kalbų). Šiuo metu veido animacija labiausiai yra paplitusi filmuose, reklamose ir kompiuteriniuose žaidimuose.

Žmonės yra labai jautrūs veido išvaizdos pokyčiams, todėl kalbos animavimas yra sudėtingas procesas, kurio metu žmogaus kalboje atpažinta akustinė informacija (fonemos) yra vizualizuojama naudojant specialiai sumodeliuotas veido išraiškas vadinamas vizemomis. Didžiausią įtaką kalbos animacijos tikrovei turi teisingas fonemas atitinkančių vizemų identifikavimas, modeliavimas ir jų išrikiavimas laiko juostoje. Tačiau, norint užtikrinti kalbos animacijos natūralumą, būtina papildomai išnalizuoti vizemų įtaką kaimyninėms fonemoms ir atsižvelgiant į animuojamos kalbos fonetines savybes sukurti koartikuliacijos valdymo modelį. Koartikuliacijos efektas apibūdina situaciją, kai vieno izoliuoto garso vaizdinė išraiška yra priklausoma nuo prieš tai arba po to tariamų garsų.

„Kalbančios galvos“ jau sukurtos anglų, prancūzų, suomių, italų, kinų, mandarinų ir kitoms kalboms, tačiau iki šiol nėra „kalbančios galvos“ skirtos lietuvių kalbai animuoti. Naujos kalbos animavimo sistemos kūrimas ir programavimas yra daug laiko, žinių, techninių ir finansinių resursų reikalaujantis darbas. Kai kurie „kalbančių galvų“ programinės įrangos kūrėjai tvirtina, kad jų siūlomi modeliai yra nuo kalbos nepriklausomi, tačiau eksperimentiniai tyrimai rodo, jog praktikoje tokie modeliai turi būti kuriami kiekvienai naujai kalbai atskirai. To priežastis yra pasaulio kalbų unikalumas:

pasaulio kalbose yra apibrėžtas skirtingas fonemų ir vizemų kiekis, be to skirtingose kalbose taikomos specifinės fonetinės taisyklės.

Labai svarbi kalbos animacijos tyrimų probleminė sritis yra vizemų identifikavimas ir modeliavimas. Mokslininkai ištyrė, jog kai kurios fonemos tariamos atrodo beveik vienodai, todėl jų vizemos gali būti klasifikuojamos į vieną vizemų klasę, kuriai apibrėžti užtenka sumodeliuoti vieną dvimatę arba trimatę vizemą. Taip galima sumažinti kalbos animacijai reikalingų duomenų kiekį.

Didelė dalis lietuvių kalboje naudojamų fonemų yra apibrėžtos ir kitose pasaulio kalbose, tačiau tokių vizemų grupavimas į klases nėra standartizuotas, t. y. skirtingose kalbose tos pačios vizemos būna priskiriamos skirtingoms vizemų klasėms. 2005 metais Lietuvos mokslininkas Balbonas pasiūlė lietuvių kalbos vizemas klasifikuoti pagal fonemų skirstymą į sprogstamuosius, dantinius, gomurinius ir t. t. garsus. Naudojant šią metodiką, į vieną vizemų klasę grupuojamos vizualiai skirtingai atrodančios priebalsės /d, t, s, z, n, dz, l/. Toks vizemų grupavimo metodas sąlygoja kalbos animacijos nenatūralumą, todėl Lietuvių kalbos vizemų klasių aibėms apibrėžti reikia atlikti papildomus tyrimus. Tarpkalbinis vizemų parinkimo modelis yra siūlomas šiai problemai spręsti.

Viena iš svarbiausių kalbos animavimo problemų yra koartikuliacijos efekto – vizemų įtakos gretimų vizemų išvaizdai modeliavimas. Norint sukurti kuo natūraliau atrodančią lietuvių kalbos animaciją, reikėtų atsižvelgti į žodyje naudojamo kirčio tipą, kirčiuotų ir nekirčiuotų garsų padėtį žodyje, ilgų ir trumpų balsių naudojimo, dvibalsių tarimo ypatumus ir pan. Atsižvelgiant į lietuvių kalbos fonetikos taisykles, šiame darbe yra kuriamas Lietuvių kalbos koartikuliacijos valdymo modelis, kuriame didžiausias dėmesys skiriamas žodžių su dvibalsiais koartikuliacijai valdyti.

Lietuvių kalbos koartikuliacijos valdymo modelis ir tarpkalbinis vizemų parinkimo modelis yra naudojami lietuviškai „kalbančiai galvai“ „LIT“ sukurti.

Tyrimų objektas. Šios disertacijos tyrimo objektas – lietuvių kalbos animavimas.

Darbo tikslas ir uždaviniai. Pagrindinis šios disertacijos tikslas yra lietuvių kalbai animuoti skirtos technologijos, kuri taiko trimatį veido modelį, sukūrimas. Darbo tikslui pasiekti reikia išspręsti tokius uždavinius:

1. Nustatyti, koks kalbos animavimo metodas ir kokio tipo kalbančios galvos modelis yra perspektyviausi lietuvių kalbai animuoti. Parinkti atviro kodo kalbos animavimo sistemą, kuri turi nustatytas savybes.

2. Sukurti lietuvių kalbos animavimo karkasą atsižvelgiant į lietuvių kalbos akustinio ir vizualaus signalų apdorojimo bei sinchronizavimo ypatumus.
3. Pasiūlyti technologiją, leidžiančią bet kurios kalbos vizemų klasių aibę skaidyti į pastoviąją ir kintamąją dalis. Pastoviajai daliai apibrėžti identifikuoti tarptautinę vizemų klasifikavimo aibę dažniausiai pasaulyje tariamoms priebalsėms.
4. Atsižvelgiant į lietuvių kalbos fonetinius ypatumus, sukurti koartikuliacijos valdymo modelį lietuvių kalbos dvibalsiams animuoti.
5. Įvertinti pasiūlytos kalbos animavimo metodikos efektyvumą analizuojant sukurtos lietuvių „kalbančios galvos“ „LIT“ suprantamumą ir natūralumą.

Tyrimų metodika. Siekiant užsibrėžtų tikslų buvo taikomi tokie tyrimų metodai:

Lyginamosios analizės ir literatūros analizės metodai. Šie metodai buvo taikomi lyginant kalbos animavimo metodologijas, fonemų ir vizemų sužymėjimo metodikas, koartikuliacijos modelių kūrimo technologijas. Matematinės ir geometrinio modeliavimo žinios buvo taikomos trimačių lietuvių kalbos vizemų kūrimui. Pagrindinė programinė atviro kodo įranga, kurią naudojant buvo sukurti animacijos pavyzdžiai yra anglų kalbos animavimo variklis iFACE ir joje naudojama kalbos animavimo kalba FML (angl. – Facial Modelling Language).

Darbo mokslinis naujumas

1. Sukurta nauja lietuvių kalbos animavimo technologija, taikanti trimatę parametrinę lietuviškai „kalbančią galvą“ „LIT“ įrašytai šnekai animuoti ir leidžianti pagerinti animuotos lietuvių kalbos suprantamumą ir natūralumą.
2. Sukurta nauja kalbos animavimo technologija, leidžianti bet kurios kalbos vizemų aibę skaidyti į pastoviąją (tarptautinę) ir kintamąją dalis. Jos naudojimas pastebimai sumažina vizemų klasių, kurios turėtų būti apibrėžtos arba sumodeliuotos animuojant konkrečią kalbą, skaičių. Tarptautinė vizemų klasifikavimo technologija 22 priebalsėms (pastovioji dalis) gali būti taikoma nuo kalbos nepriklausančioms „kalbančioms galvoms“ kurti.
3. Sukurtas naujas lietuvių kalbos koartikuliacijos valdymo modelis, kuris apibrėžia 8 taisykles, padedančias vizualizuoti lietuvių kalbos

dvibalsius ir pritaiko originalią metodiką kalbos struktūroms BBPB (balsė-balsė-priebalsė-balsė) animuoti.

Darbo rezultatų praktinė reikšmė. Pasiūlyta lietuvių kalbos animavimo technologija, ypač vizemų klasifikavimas į pastoviąją (tarptautinę) ir kintamąją aibes, gali būti efektyviai taikoma kuriant naujus, nuo kalbos nepriklausančius, vizemomis pagrįstus „Kalbančių galvų“ modelius. Darbe pasiūlyta kalbos animavimo technologija panaudota, kuriant Lietuviškai kalbančią galvą „LIT“. Ji gali būti naudojama elektroninėse mokymo sistemose, kaip animuotas 3D modelis padedantis suprasti žodžių tarimą, arba pritaikoma filmų, reklamos ar žaidimų pramonėje. Pasiūlytas lietuvių kalbos koartikuliacijos modelis gali būti panaudotas ne tik lietuvių, bet ir kitų kalbų, kuriose yra daug dvibalsių, koartikuliacijai apibrėžti.

Ginamieji teiginiai. Formuluojami šie ginamieji teiginiai:

1. Sukurtas kalbos animavimo karkasas yra tinkamas įrašyti lietuvių kalbai vizualizuoti.
2. Pasiūlyta tarpkalbinė vizemų parinkimo technologija, skaidanti bet kurios kalbos vizemų klasių aibę į kintamąją ir tarptautinę (pastoviąją) dalis, gali būti naudojama kuriant naujas „kalbančias galvas“.
3. Sukurtas koartikuliacijos modelis yra tinkamas lietuvių kalbos žodžiams, kuriuose yra dvibalsių, animuoti.

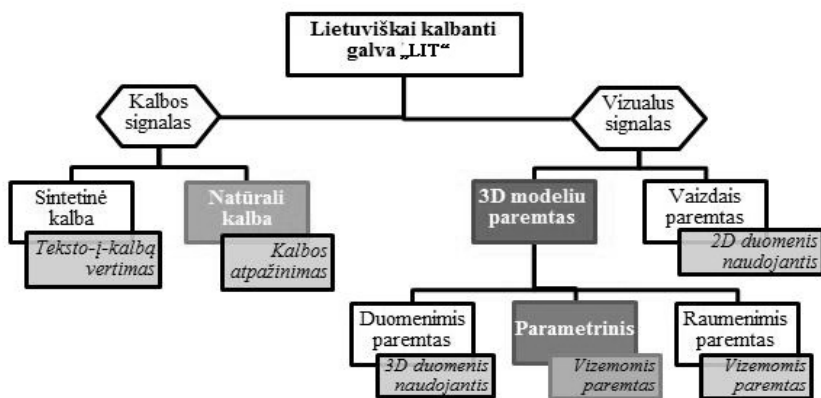
Darbo apimtis. Darbą sudaro įvadas, trys pagrindiniai skyriai, bendrosios išvados, literatūros sąrašas, publikacijų sąrašas. Darbo apimtis – 107 puslapiai, tekste panaudotos 10 numeruotų formulių, 30 paveikslų ir 9 lentelės. Rašant disertaciją panaudoti 144 literatūros šaltiniai. Darbas buvo pristatytas 3 konferencijose. Disertacijos tema publikuoti 3 straipsniai.

1. Esamų kalbos animavimo technologijų analizė

Skyriuje analizuojamos pasaulyje naudojamos kalbos animavimo technologijos. Kalbos signalas yra ir girdimas, ir matomas, todėl jos animacija yra sudėtinis procesas priklausantis nuo pasirinktos veido modeliavimo metodikos, kalbos signalo tipo, koartikuliacijos valdymo modelio.

Literatūros analizė rodo, jog „kalbančios galvos“ gali naudoti natūralią įrašytą žmogaus kalbą arba sintetinę kalbą, gautą įgarsinant parašytą tekstą. Moksliniai tyrimai patvirtina, jog natūralaus balso ir sintetinio veido junginių natūralumas yra 25% didesnis lyginant su sintetinio balso ir sintetinio veido junginių natūralumu.

Trimatį veido modelį arba 2D filmuotą (fotografuotą) vaizdo medžiagą naudojantys kalbos animavimo metodai gali būti taikomi „kalbančioms galvoms“ kurti. 2D duomenis naudojančios „kalbančios galvos“ pasižymi didesniu gautos animacijos realizmu, tačiau jų automatizmas ir lankstumas yra daug mažesni negu trimatčių modelių. Trimatėms „kalbančioms galvoms“ kurti gali būti naudojamos skirtingos veido modeliavimo metodikos. 3D skanuotais duomenimis pagrįstai metodikai reikia brangios techninės įrangos ir didelės įrašytų veido judesių duomenų bazės, tuo tarpu parametriniai ir raumenimis pagrįsti trimatčiai veido modeliai naudoja nedidelį kiekį iš anksto sumodeliuotų trimatčių vizumų kalbos animacijai kurti. Be to, turint vieną trimatį veido modelį, bet naudojant skirtingas tekstūras, galima sukurti vizualiai skirtingai atrodančias „kalbančias galvas“. 1 paveiksle pavaizduota klasifikavimo schema, nurodanti kokia kalbos animavimo technologija ir kokio tipo „kalbančios galvos“ modelis yra perspektyviausi lietuvių kalbai animuoti.



1 pav. „Kalbančių galvų“ kūrimui naudojamų metodų klasifikavimo schema. Tamsūs stačiakampiai apibrėžia metodus, naudojamus lietuvių „kalbančiai galvai“ „LIT“ kurti

Viena iš svarbiausių kalbos animavimo problemų yra koartikuliacijos efekto modeliavimas. Literatūros analizė rodo, jog koartikuliacijos valdymo modelio taikymas pastebimai padidina kalbos animacijos suprantamumą. Dažniausiai vizualios kalbos koartikuliacijos valdymui naudojami dirbtiniai neuronai tinklai, Cohen-Massaro, Ohman arba taisyklėmis pagrįsti modeliai. Tyrimais patvirtinta, jog taisyklėmis pagrįsti modeliai yra patys tiksliausi. Jie naudoja specifines analizuojamos kalbos fonetines savybes koartikuliacijai

valdyti. Antras pagal tikslumą yra duomenimis pagrįstas Ohman koartikuliacijos valdymo modelis koartikuliacijai tarp dviejų balsių valdyti. Lietuvių kalboje yra daug specifinių taisyklių, įtakojančių žodžių tarimo išraiškingumą (kirčio vieta ir tipas, dvibalsių tarimo ypatumai ir pan.). Jos turėtų atsispindėti ir Lietuvių kalbai animuoti skirname koartikuliacijos modelyje.

Vizemas naudojančios „Kalbančios galvos“ animacijos natūralumą ir suprantamumą stipriai įtakoja vaizdo ir garso signalų atitikimo ir jų synchronizavimo laike tikslumas. Skirtingų mokslininkų tyrimuose apibrėžiamas skirtingas vizemų klasifikavimas į grupes bei tų grupių skaičius, o tos pačios fonemos priskiriamos kitoms vizemų klasėms. Norint sudaryti galimybę kuo mažesnėmis pastangomis tą pačią veido animavimo technologiją ir trimatį veido modelį naudoti naujoms kalbos animuoti, mums reikia naujo sprendimo.

2. Lietuvių kalbai animuoti skirtas karkasas

Antrajame skyriuje pristatomas karkasas, skirtas lietuvių kalbai animuoti. Jis sukurtas atsižvelgiant į esamų kalbos animavimo technologijų privalumus ir trūkumus. Išanalizavus esamus sprendimus, skyriuje siūlomas teorinis modelis, kaip atviro kodo vizemomis paremta anglų kalbos animavimo sistema gali būti panaudota siūlomai lietuvių kalbos animavimo technologijai realizuoti. Pasirinkta sistema veido judesiams kontroliuoti naudoja hierarchinį trimatį galvos modelį.

Žodžių fonetinė transkripcija ir fonemų išsidėstymo laike informacija yra pirminiai kalbos animavimo proceso duomenys, įtakojantys animuotos kalbos natūralumą. Atlikus analizę, įrašytų lietuvių kalbos žodžių atpažinimui nuspręsta naudoti Lietuvos mokslininkų A. Lipeikos ir jo kolegų sukurtą kalbos atpažinimo sistemą. Naujai sukurtų tarpkalbinio vizemų parinkimo modelio bei lietuvių kalbos dvibalsių koartikuliacijos valdymo modelio integraciją į pasirinktą kalbos animavimo sistemą siūloma atlikti taikant veido modeliavimo kalbą FML (angl. – Facial Modelling Language).

3. Siūloma lietuvių kalbos animavimo technologija

Skyriuje detalai aprašomi du nauji modeliai, kurie naudojami lietuvių kalbos animavimo technologijai sukurti ir lietuvių kalbos animacijos suprantamumui ir tikroviškumui padidinti. Tarpkalbinis vizemų parinkimo modelis yra naudojamas lietuvių kalbos vizemoms skirstyti į pastoviąją (tarptautinę) ir kintamąją dalis. Atsižvelgiant į lietuvių kalbos fonetikos

taisykles lietuvių kalbos dvibalsiams animuoti sukuriamas naujas koartikuliacijos valdymo modelis. Siūlomos lietuvių kalbos animavimo technologijos kokybė įvertinama praktiniais eksperimentais.

3.1. Tarpkalbinis vizemų parinkimo modelis

Žmogaus akis negali atskirti akustiškai skirtingų, bet vizualiai panašių garsų, tokių kaip /p/ ir /b/, todėl šių fonemų vizemos gali būti grupuojamos į vieną vizemų klasę. Vizemų klasifikavimas į vizemų klases yra viena iš pagrindinių kalbos animacijos problemų. Skirtingi mokslininkai naudojo skirtingus metodus vizemų klasėms identifikuoti, tačiau iki šiol nėra vieno standarto kaip tai turėtų būti daroma, todėl tos pačios fonemos skirtingose kalbose yra priskiriamos skirtingoms vizemų klasėms.

Pastovios vizemų aibės apibrėžimas. Priebalsės yra nepozicionuojamos – t. y., jų tarimas nedaug priklauso nuo to kokiame kontekste jos buvo pasakytos. Tokios priebalsių savybės sudaro prielaidas tarpkalbinei (pastoviajai) vizemų klasių aibei egzistuoti. 11-oje skirtingų pasaulio kalbų eksperimentiškai apibrėžtos vizemų grupės buvo analizuojamos norint apibrėžti pastoviąją vizemų aibę 22 dažniausiai pasaulyje pasitaikančioms priebalsėms. Tarpkalbinė priebalsių panašumo matrica buvo apibrėžta norint nustatyti panašiausiai atrodančias vizemų poras. Vieną vizemų klasę gali sudaryti daugiau nei dvi vizemos. Norint patikslinti vizemų kiekį klasėse, skirtingų kalbų vizemų klasifikavimo duomenims apdoroti buvo panaudotas Grupinio vidurkinimo aglomeratyvaus klasterizavimo hierarchinis metodas, kurį naudojant buvo apibrėžta tarpkalbinė (pastovioji) vizemų aibė 22 priebalsėms. Ją sudaro 11 vizemų klasių.

Kintamos vizemų aibės apibrėžimas. Priebalsės galima apibrėžti pastovia vizemų klase su nekintančiu ekspresijos koeficientu, kadangi jų ekspresyvumas beveik nekinta. Balsės yra nepozicionuojamos, t. y., jų vizuali išraiška labai stipriai keičiasi priklausomai nuo konteksto, todėl jas animuojant turėtų būti naudojamas papildomas ekspresijos koeficientas. Apibūdinant šias vizemų klases nustatotas maksimalus ekspresijos koeficientas, tačiau kalbos sinchronizavimo algoritme jis kinta. Vizemų dinamiškumą apibūdina pasirinktas koartikuliacijos modelis. Dviejų ortogonalinių vaizdų kalibravimo technologija buvo panaudota modeliuojant lietuviškas trimates vizemas balsėms. Jų ekspresijos koeficientas ir priklausomybę anglų kalbos vizemų klasėms buvo nustatyta naudojant lietuviškų ir angliskų fonemų vizualų panašumą.

Tarpkalbinio vizemų parinkimo modelio taikymas. Tarpkalbinis vizemų parinkimo modelis gali būti taikomas animuoti naujas kalbas naudojant

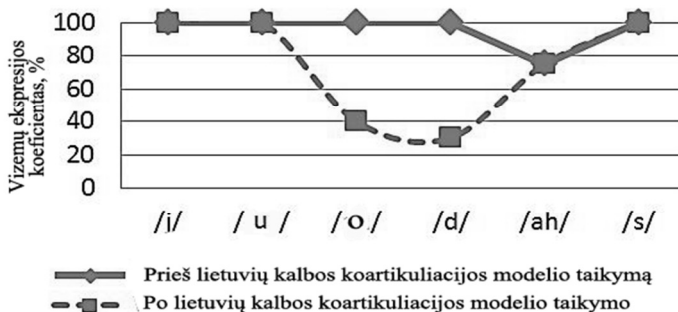
disertacijoje apibrėžtą pastoviąją vizemų klasių aibę. Modulyje detalizuota kaip elgtis, jeigu fonema nepriklauso pastoviajai vizemų klasių aibei. Šio modelio taikymas pastebimai sumažina vizemų skaičių, kurios turi būti papildomai sumodeliuotos arba apibrėžtos kalbai animuoti. Be to, jis suteikia galimybę naudoti esamų kalbos animavimo sistemų vizemų aibę naujoms kalboms animuoti.

3.2. Lietuvių kalbos dvibalsių koartikuliacijos valdymo modelis

Žmonės atpažįsta net patį mažiausią kalbos netobulumą, todėl animuotos kalbos išraiškingumui kontroliuoti yra taikomi koartikuliacijos valdymo modeliai. Dvibalsiai yra lietuvių kalbos žodžiuose dažnai pasitaikanti fonetinė struktūra, todėl jų vizualizacijos tikslumas padidina animuotos lietuvių kalbos suprantamumą.

Tyrimais patvirtinta, jog taisyklėmis pagrįsti vizualios kalbos koartikuliacijos valdymo modeliai yra patys tiksliausi. Jie naudoja specifines analizuojamos kalbos fonetines savybes koartikuliacijai valdyti. Antras pagal tikslumą yra duomenimis pagrįstas Ohman koartikuliacijos valdymo modelis koartikuliacijai tarp dviejų balsių valdyti. Situacijos analizė parodė, jog specialaus koartikuliacijos valdymo modelio fonetinėms kalbos struktūroms BBPB (balsė-balsė-priebalsė-balsė), BBPPB ir t. t. nėra.

Atsižvelgiant į lietuvių kalbos fonetines savybes, skiemenų kirčiavimą, kirčio tipą ir dvibalsio vietą žodyje, buvo sukurtas naujas lietuvių kalbos koartikuliacijos valdymo modelis, kuris apibrėžia 8 taisykles, padedančias vizualizuoti lietuvių kalbos dvibalsius ir pritaiko originalią metodiką kalbos struktūroms BBPB (balsė-balsė-priebalsė-balsė) animuoti.



2 pav. Vizemų ekspresijos koeficiento reikšmė prieš ir po lietuvių kalbos koartikuliacijos modelio taikymo

Pasiūlytas lietuvių kalbos koartikuliacijos modelis yra skirtas nustatyti analizuojamos vizmos ekspresijos koeficiento, apibūdinančio statinės vizmos išraiškingumą, reikšmę. Ji yra priklausoma nuo konteksto, kuriame tiriamą vizmą buvo ištarta. Žodyje „juodas“ naudojamų vizmų ekspresijos koeficiento reikšmės kitimas prieš ir po lietuvių kalbos koartikuliacijos modelio taikymą (2 pav.) patvirtina, kad siūlomas modelis pastebimai padidina dvibalsių turinčių lietuvių kalbos žodžių animacijos natūralumą (pavyzdžiui, nekirčiuotas dvibalsio „uo“ balsis yra mažiau išraiškingas negu kirčiuotas).

3.3. Lietuvių kalbos animavimo technologijos praktinis vertinimas

Siekiant nustatyti lietuviškai kalbančios galvos „LIT“ kokybę, buvo atlikti animuotos kalbos suprantamumo ir natūralumo vertinimai. Eksperimentų metu tie patys lietuvių kalbos garso failai buvo vizualizuojami naudojant anglų kalbos atviro kodo animavimo sistemą iFACE ir lietuvių kalbos animavimo sistemą „LIT“.

Lietuviškai kalbančios galvos suprantamumo vertinimas buvo atliekamas naudojant garso failus su papildomu triukšmu. Eksperimento dalyvių buvo prašoma identifikuoti triukšme vos girdimus žodžius. Subjektyvaus vertinimo rezultatai (1 lentelė) patvirtino, jog „LIT“ „Kalbančios galvos“ pateikiama animacija yra 13% suprantamesnė lyginant su anglų kalbos animavimo sistema iFACE. Be to, eksperimentas taip pat patvirtino, jog vizualaus signalo pridėjimas akivaizdžiai padidina akustinio signalo suprantamumo lygį.

1 lentelė. Vidutinis lietuvių kalbos žodžių atpažinimo dažnis skirtingose aplinkose

Aplinka	Žodžių atpažino dažnis, %
Tik girdimas signalas	11
„iFACE“ kalbos animavimo sistema	43
„LIT“ kalbos animavimo sistema	56

Lietuvių kalbos animacijos natūralumui įvertinti akustinis signalas buvo pateikiamas be triukšmo. Vertinamas vyko naudojant 5 balų Likert skalę, kurioje 5 reiškia, jog animacija yra labai natūrali. 3 skirtingų žodžių aibės buvo animuotos ir pateiktos eksperimento dalyviams.

2 lentelė. Kalbos natūralumo vertinimo eksperimento rezultatai

Vertintų žodžių aibė	„iFACE“ kalbos animavimo sistema	„LIT“ kalbos animavimo sistema	Kalbos natūralumo pagerėjimas
I	2,57	3,13	18,15%
II	3,04	3,43	11,39%
III	2,30	2,83	18,46%
Vidurkis	2,64	3,13	16,00%

Rezultatai pateikiami 2 lentelėje patvirtina, kad visais trim atvejais su „LIT“ „kalbančia galva“ sukurtos animacijos pavyzdžiai yra natūralesni už iFACE originaliu varikliu sukurtas lietuvių kalbos vizualizijas.

Bendrosios išvados

Šioje disertacijoje buvo pristatyta nauja įrašytos lietuvių kalbos animavimo technologija. Ji buvo panaudota kuriant trimatę lietuviškai „kalbančią galvą“ „LIT“. Analizuojant teorines žinias ir vykdant praktinius eksperimentus buvo apibrėžtos išvados:

1. Atlikus literatūros šaltinių analizę galima daryti išvadą, jog kuriama lietuvių kalbos animavimo sistema „LIT“ turi naudoti perspektyviu laikomą trimatį parametrinį veido modelį, leidžiantį nesunkiai integruoti įvairius veido judesius aprašančius metodus ir naudojančių vizemomis pagrįstą kalbos animavimo technologiją animuotos kalbos natūralumui užtikrinti.
2. Pasiūlyta nauja tarpkalbinė kalbos animavimo technologija, kurią taikant bet kurios kalbos vizemų klasių aibė gali būti skaidoma į pastoviąją ir kintamąją dalis. Pastovioji dalis atitinka tarpkalbinę priebalsinių vizemų aibę gautą hierarchinio klasterizavimo statistiniu metodu. Pasiūlytos 7 rekomendacijos, kaip 22 priebalsės turėtų būti klasifikuojamos į 11 tarptautinių vizemų klasių. Pastovioji aibė yra bendra 11-ai tirtų, labiausiai paplitusių pasaulio kalbų, todėl galime teigti, jog ši aibė galėtų būti taikoma bet kurios pasaulio kalbos animacijai.
3. Sukurtas lietuvių kalbai animuoti skirtas dvibalsių koartikuliacijos valdymo modelis, kuris leidžia integruoti 8 lietuvių kalbos fonetikos taisykles žodžiams su dvibalsiais vizualizuoti. Kalbos natūralumui pasiekti taikyta originali metodika, apibrėžianti fonetinių struktūrų

kaip BBPB (balsė-balsė-priebalsė-balsė) modeliavimą. Lietuvių kalbai sukurtas koartikuliacijos modelis padeda nustatyti vizemų ekspresyvumo koeficiento reikšmę įtakojamą gretimų vizemų, bei padidina kalbos išraiškingumą.

4. Pasiūlytas lietuvių kalbai animuoti skirtas karkasas, kurį naudojant tarpkalbinis vizemų parinkimo ir lietuvių kalbos dvibalsių koartikuliacijos valdymo modeliai yra integruojami į anglų kalbai animuoti sukurtą sistemą.
5. Lietuviškai „kalbančios galvos“ „LIT“ efektyvumas buvo nustatytas vertinant jos suprantamumą ir natūralumą. Palyginus animacijos pavyzdžius, gautus su „LIT“ ir pavyzdžius, gautus naudojant anglų kalbai animuoti skirtą sistemą „iFACE“, nustatyta jog darbe pasiūlyta technologija 13% pagerino lietuvių kalbos animacijos natūralumą ir 16% pagerino jos suprantamumą.

Autoriaus publikacijų disertacijos tema sąrašas Recenzuojamuose mokslo žurnaluose

Mažonavičiūtė, I.; Baušys, R. 2011. Translingual visemes mapping for Lithuanian speech animation, *Elektronika ir elektrotechnika* 5(111): 95–9. (ISI Web of Science)

Mažonavičiūtė, I.; Baušys, R. 2009. English talking head adaptation for Lithuanian speech animation, *Information technology and control* 3(38): 217–224. (ISI Web of Science)

Straipsniai kituose leidiniuose

Mažonavičiūtė, I.; Baušys, R. 2010. Framework for Lithuanian speech animation, *18th European Signal Processing Conference (EUSIPCO-2010)*, Aalborg, Denmark, Rugpjūčio 23–27, 2010. Aalborg : EUSIPCO, 781–785. ISSN 2076-1465.

Trumpos žinios apie autorių

Ingrida Mažonavičiūtė gimė 1984 m. kovo 8 d. Telšiuose. 2006 m. įgijo informatikos inžinerijos bakalauro laipsnį Vilniaus Gedimino technikos universiteto (VGTU) Fundamentinių mokslų fakultete. 2008 m. įgijo informatikos inžinerijos mokslo magistro laipsnį VGTU. 2008–2012 m. – VGTU doktorantė. Šiuo metu dirba lektore VGTU Fundamentinių mokslų fakulteto Grafinių sistemų katedroje.

LITHUANIAN SPEECH ANIMATION TECHNOLOGY FOR 3D FACIAL MODEL

Topicality of the problem. Human communicate using words and sentences. Lipreading is considered as a psychological process not only in people with hearing loss and deafness, but also in normal hearing people. Visual information (i.e. lips and tongue movements) significantly enhances the comprehension of the uttered audio signal. This is particularly true when the acoustic conditions are inadequate. Researchers revealed that addition of the visual information increases the speech perspicuity with 57% for consonants, 30% for vowels, 39% for monosyllabic words and 17% for short phrases.

Ever since visual speech improves the perception of acoustical signal, design and implementation of 2D or 3D synthetic head models that can produce naturally looking audio to video mapping („Talking heads”) became very important in Human – Computer Interaction area. Design of „Talking heads” indistinguishable from a real human is challenging task, since people are highly sensitive to facial behaviour. Moreover, casual viewer immediately perceives the slightest deviation from the natural facial appearance and defines it as inaccurate visualization.

During the last years, different „Talking heads” have been proposed, pursuing either a natural 3D facial mesh or video-realistic effects. Most of the existing models are dedicated to animate English language, although recently we also see talking heads driven by Finnish, Italian, Chinese Mandarin (Putonghua) and etc. However, there is not „Talking head” specified to visualize Lithuanian language.

Speech animation is widely used in technical devices to allow the growing number of hearing impaired persons, children, middle-aged and elderly equal participation in communication. Talking heads are playing considerably important role in human-computer communication and have caused significant scientific, technological and artistic interests in computer facial animation. Talking heads can be employed for e-consulting services: virtual secretary, WEB navigator or virtual agent who is responsible for information conveying to user in a Smart Ecological and Social Apartments (SESA). Also they are widely used in e-learning technologies as animated 3D models for the correct sound pronounsation presentation or applied in movie, advertising and computer game industries.

Talking heads are basically driven by speech phonetics and their visual representation – visemes. Many acoustic sounds of separate languages are visually similar and accordingly different phonemes can be classified using the same viseme. Phoneme to viseme mapping remains the key subject in bimodal

speech processing due to the fact that correctness of identified visemes considerably influence the realism of animated speech. Despite the fact, that various researchers attempted to classify phonemes according to their visual similarity, as of date, neither a single classification, nor a standardized viseme set exist for individual languages, as it exists for phonemes. For this reason audio to visual mapping should be revised for every new language, even if it is expensive and time consuming process.

In 2005, Lithuanian scientist Balbonas proposed classification of Lithuanian visemes by their phonemic distinction into explosive, flint, guttural, etc. In this approach, differently apparent Lithuanian consonants /d, t, s, z, n, dz, l/ belong to the same viseme class. It causes speech animation artificiality, thus additional research should be done to define Lithuanian viseme classes.

Since all humans are experts in lip reading and detects even the slightest errors during speech animation, coarticulation generation is crucial part of any speech animation system. Coarticulation is the physical phenomena which refers to the blurring of boundaries between separate speech units (both auditory and visibly). Word stressing, position of the accented phoneme, pronunciation of diphthongs, long and short vowels highly influence final visualization of the speech. Lithuanian coarticulation control model, which is based on Lithuanian phonetics, is proposed to animate words with pure diphthongs.

Proper visemes definition and their correct alignment are the crucial points for believable language animation. The proposed coarticulation control and translanguing viseme classification models are used to create Lithuanian talking head „LIT”.

The object of the research. The objective of the thesis is Lithuanian speech animation.

Main objective and tasks of the work. The main goal of the thesis is to create Lithuanian speech animation technology. In order to achieve the goal, the following tasks had to be solved:

1. To define Talking head modelling and speech animation generation methods, that are the most perspective for Lithuanian speech animation. To indicate open source speech animation engine that has these features.
2. To create Lithuanian speech animation platform, which pre-processes, synchronizes and integrates acoustic and visual features of Lithuanian speech.

3. To propose technology, that designates how to divide viseme set of any language into two parts: constant and variable. Translingual viseme set for globally the most common consonants should be determined to identify constant set of visemes.
4. To create Lithuanian coarticulation control model that concerns phonetic features of Lithuanian language and is specified to animate pure Lithuanian diphthongs.
5. To evaluate the effectiveness of the proposed speech animation technology considering intelligibility and naturalness of created Lithuanian Talking head „LIT“.

Research methodology. Theoretical analysis and practical implementation knowledge of mathematical analysis, speech animation methodologies, phoneme to viseme mapping techniques, coarticulation generation and 3D modelling by calibration from 2D pictures were used. The base software for practical implementation of the experiments was English speech animation engine iFACE and its Facial Modelling Language (FML).

Scientific novelty

1. A novel Lithuanian speech animation technology, that employs parameter-based, 3-dimensional, speech driven Lithuanian Talking head „LIT“, was created. It improves intelligibility and naturalness of animated Lithuanian speech.
2. The new speech animation technology was created, that allows division of viseme set of any language into constant (translingual) and variable parts. It considerably reduces the amount of visemes that should be analysed or modelled for specific language. Translingual classification of visemes for 22 internationally common consonants can be applied to produce language independent Talking heads.
3. New coarticulation control model, that defines eight rules for pure Lithuanian diphthong animation, was created. Original coarticulation generation methodology to animate structures VVCV (vowel-vowel-consonant-vowel) was proposed and incorporated into the model.

Practical significance of the achieved results. The proposed Lithuanian speech animation technology with integrated translingual viseme identification model can be successfully applied to model new language-independent viseme-driven „Talking heads“. Lithuanian speech animation platform can be used to

create Lithuanian off-line „Talking head“ that can be employed in e-learning technologies as animated 3D model for the correct sound pronounsation presentation or applied in movie, advertising and computer game industries. The presented Lithuanian coarticulation modelling technique can be used to increase naturalness of animated speech for other languages with many diphthongs.

The defended statements. The following defendable statements are defined:

1. The proposed speech animation platform is suitable to visualise recorded Lithuanian speech.
2. Translingual viseme classification technology, that allows division of viseme set of any language into constant (translingual) and variable parts, can be employed to create new viseme – driven “talking heads”.
3. The proposed Lithuanian coarticulation control model is suitable to animate Lithuanian words with pure diphthongs.

The scope of the scientific work. The dissertation consists of Introduction, three main chapters and general conclusions. The total dissertation scope is 107 pages, 10 equations, 30 pictures and 9 tables.

In Introduction topicality of the problem is analysed, work main objective and tasks are formulated, scientific novelty is proved, research methods used are listed and defended statements are presented.

Chapter 1 provides the analysis of the existing speech animation technologies. Audiovisual speech generation methods including analysis of data-driven and viseme-driven are compared to choose the most suitable speech animation technology for Lithuanian language. Viseme classification experiments across different languages are analysed in detail to identify variety of viseme classes in various languages. Speech processing for speech animation is also investigated on signal, phoneme and word level to choose what type of audio to visual mapping should be used for Lithuanian speech animation. Existing coarticulation control models including look-ahead, time-locked, Artificial Neural Networks and Rule-based models are analysed to compare them according to their perceptual intelligibility. Different facial modelling techniques are analysed as well to define the most suitable 3D „Talking head” modelling technique for Lithuanian speech animation. Also in this chapter current speech animation systems and facial animation evaluation methods are examined.

Chapter 2 describes theoretical framework for Lithuanian speech animation. Translingual visual speech synthesis methods, specifics of Lithuanian phonetics and their influence for the selection of automatic phonemes recognition system that will be used in the proposed framework are analysed. Features of speech animation system that is compatible with the requirements defined in the first chapter are also investigated.

Chapter 3 presents the details of two new models that help to improve intelligibility of animated Lithuanian speech. Translingual viseme-driven speech animation model divides viseme classes of any language into two parts: constant (translingual) and variable. New translingual viseme set is defined using statistical hierarchical clustering of visually similar consonants in 11 different languages. Modelling of 3D Lithuanian visemes is proposed to identify expressiveness coefficients of visemes that belong to variable set of visemes. The second model presented in this chapter is new coarticulation control model specified for pure Lithuanian diphthongs. This chapter is concluded with practical evaluation of the proposed framework for Lithuanian speech animation.

General conclusions summarises the present study. It is followed by an extensive list of 144 references and a list of 3 publications by the author on the topic of the dissertation.

General conclusions

In this thesis a novel framework to animate recorded Lithuanian speech was presented. Translingual speech animation technology was employed to create Lithuanian Talking head named „LIT“. Theoretical evaluation and practical experiments allowed the following conclusions to be made:

1. State of art analysis revealed, that Lithuanian talking head “LIT” should be viseme-driven speech animation system, which uses perspective parametric 3D head model to describe facial movements and to produce naturally looking animated speech.
2. The proposed new translingual speech animation technology allows division of viseme set of any language into constant (translingual) and variable parts. Based on the results of hierarchical clustering, 7 recommendations were offered to classify 22 internationally common consonants into 11 translingual viseme classes and to determine the constant part of visemes set. It is mutual for 11 analysed languages; therefore it can be used to animate other languages.

3. The proposed coarticulation control model allows simple integration of 8 Lithuanian phonetic rules to animate words with pure Lithuanian diphthongs. The original methodology that defines coarticulation modelling of the phonetic structures VVCV (vowel-vowel-consonant-vowel) was applied to attain speech naturalness. Coarticulation control model specified for Lithuanian language helps to determine expressiveness parameter of specific visemes and to increase naturalness and intelligibility of animated Lithuanian speech.
4. The new speech animation framework, that integrates Lithuanian coarticulation control and translingual visemes classification models into chosen English speech animation system, was proposed to produce Lithuanian „Talking head” „LIT”.
5. Lithuanian „Talking head” „LIT” was estimated in terms of intelligibility and naturalness during the experiments, where quality of video samples generated with „LIT” was compared with Lithuanian speech animation samples produced by English speech animation engine iFACE. The proposed Lithuanian speech animation technology improved speech animation naturalness by 13% and increased its intelligibility by 16%.

About the author

Ingrida Mažonavičiūtė was born in Telšiai, on the 8th of March in 1984. First degree in Informatics Engineering, Faculty of Fundamental Sciences, Vilnius Gediminas Technical University, 2006. Master of Science in Informatics Engineering, Faculty of Fundamental Sciences, Vilnius Gediminas Technical University, 2008. In 2008–2012 – PhD student of Vilnius Gediminas Technical University. At present – lecturer in Graphical Systems Department of Vilnius Gediminas Technical University.

Ingrida MAŽONAVIČIŪTĖ

LIETUVIŲ KALBOS ANIMAVIMO TECHNOLOGIJA TAIKANT TRIMATĮ VEIDO
MODELĮ

Daktaro disertacijos santrauka
Technologijos mokslai, informatikos inžinerija (07T)

Ingrida MAŽONAVIČIŪTĖ

LITHUANIAN SPEECH ANIMATION TECHNOLOGY
FOR 3D FACIAL MODEL

Summary of Doctoral Dissertation
Technological Sciences, Informatics Engineering (07T)

2012 00 00. 1,5 sp. I. Tiražas 70 egz.
Vilniaus Gedimino technikos universiteto
leidykla „Technika“,
Saulėtekio al. 11, 10223 Vilnius,
<http://leidykla.vgtu.lt>
Spausdino UAB „Ciklonas“
J. Jasinskio g. 15, 01111 Vilnius