SAL 28/2016

Lietuviško balso sintezatorių kokybės vertinimas

Received 10/2015
Accepted 04/2016

Lietuviško balso sintezatorių kokybės vertinimas

Evaluation of Lithuanian Text-to-Speech Synthesizers

COMPUTATIONAL LINGUISTICS / KOMPIUTERINĖ LINGVISTIKA

Pijus Kasparaitis

Fizinių mokslų daktaras, Vilniaus universiteto Matematikos ir informatikos fakulteto Kompiuterijos katedros docentas, Vilniaus universitetas, Lietuva.

http://dx.doi.org/10.5755/j01.sal.0.28.15130

Anotacija

Kitų kalbų balso sintezė iš teksto plačiai naudojama jau kelis dešimtmečius, tuo tarpu kuriant lietuvių kalbos sintezatorius esminis proveržis įvyko tik pastaraisiais metais. Vien 2013–2015 metais pasirodė šeši nauji lietuviški sintetiniai balsai. Todėl atsirado poreikis įvertinti naujai atsiradusius lietuviško balso sintezatorius. Šiame darbe pateikta chronologinė esamų lietuviško balso sintezatorių apžvalga, išsamiau aprašytas naujausiuose sintezatoriuose *SINT.AS* ir LIEPA pritaikytas vienetų parinkimo algoritmas bei vykdyta diktorių atranka, nes tai turėjo lemiamos reikšmės sintezuoto balso kokybei. Pagrindinės sintetinio balso charakteristikos yra suprantamumas ir priimtinumas, jos vertinamos pasitelkiant žmones-klausytojus ir gautus duomenis apdorojant statistiniais metodais. Taigi šiame darbe tiriami devyni naujausi lietuviški sintetiniai balsai (*Regina*, *Edvardas*, *Aistė*, *Vladas*, *Laima*, *Marijus*, *Egidius*, *Aistis 2* ir *Gintaras*), bandoma įvertinti, koks iš tikrųjų sintetinių balsų suprantamumas pasiektas, palyginti sintezatorius tarpusavyje pagal įvairias charakteristikas, pateikti patarimų potencialiems sintetinio balso taikymų kūrėjams renkantis sintezatorių balsus ir galiausiai parodyti sintezatorių kūrėjams perspektyviausias tobulinimo kryptis.

REIKŠMINIAI ŽODŽIAI: balso sintezė iš teksto, vienetų parinkimo metodas, sintezuoto balso suprantamumas, sintezuoto balso priimtinumas.

Įvadas

Populiariausiomis pasaulio kalbomis parašytą tekstą kompiuteriai išmoko perskaityti balsu jau prieš kelis dešimtmečius. Pirmasis lietuviškas sintetinis balsas taip pat atsirado kiek daugiau nei prieš du dešimtmečius, tačiau ilgą laiką lietuviški sintezuoti balsai skambėjo nenatūraliai, todėl buvo naudojami tik siaurame vartotojų būryje – Lietuvos aklųjų ir silpnaregių. Pastaruoju metu lietuviško balso sintezė iš teksto sparčiai plėtojama, populiarėja, atsiranda vis naujų jos taikymo pavyzdžių. Vien 2013–2015 metais pasirodė šeši nauji lietuviški sintetiniai balsai. Todėl aktualu panagrinėti, ar suprantami šie naujieji balsai, ar žmogui malonu jų klausytis, kurį balsą rinktis vienu ar kitu taikymo atveju, kaip sukurti dar suprantamesnius ir natūraliau skambančius sintetinius balsus.

Sintetinio balso kokybė nusakoma dviem pagrindiniais parametrais: suprantamumu (angl. *intelligibility*) ir priimtinumu (angl. *acceptability*) (Schmidt-Nielsen, 1995). Suprantamumas

ktu 1922

Research Journal
Studies about Languages
No. 28/2016
ISSN 1648-2824 (print)
ISSN 2029-7203 (online)
pp. 80-91
DOI 10.5755/j01.sal.0.28.15130

nusako, kokią lingvistinių vienetų (fonemų, skiemenų ar žodžių) dalį gali suvokti klausytojas. Tai objektyvus parametras. Šiame darbe vertintas žodžių sakinyje ir sakinio visumos suprantamumas. Priimtinumas yra subjektyvus kriterijus: juo stengiamasi nustatyti, ar sintetinis balsas yra priimtinas klausytojui, ar malonu jo klausytis, kiek jis skiriasi nuo žmogaus balso. Kurį laiką daugiau pastangų buvo nukreipta suprantamumui gerinti, kadangi nesuprantama kalba praktiškai yra bevertė. Pasiekus pakankamai aukštą sintetinio balso suprantamumo lygį, stengiamasi pagerinti ir priimtinumą.

Šiame darbe trumpai aprašyti esami lietuviško balso sintezatoriai, jų pagrindinių komponentų (kirčiavimo-transkribavimo modulio bei garsų bazės) charakteristikos, taip pat paminėta, kur šie sintezatoriai naudojami. Atskiruose skyreliuose aptarti esminiai balso kokybę lemiantys veiksniai: diktoriaus atranka ir sintezės vienetų parinkimo algoritmas. Toliau pateikta ankstesniais metais atliktų sintetinio lietuviško balso kokybės vertinimų santrauka. Pagrindinis šio darbo tikslas – įvertinti devynių lietuviškų sintetinių balsų suprantamumą ir priimtinumą. Darbe aprašyta vertinimo procedūra, pateiktos rekomendacijos potencialiems sintetinių balsų naudotojams bei kūrėjams.

Esami lietuviško balso sintezatoriai

Pirmieji sintezatoriai *Apollo* ir *Aistis*

1994 m. Didžiosios Britanijos kompanija *Dolphin systems Inc.* į sintezatorių *Apollo* greta kitų kalbų įtraukė ir lietuvių kalbą. Tai buvo pirmasis Lietuvos aklųjų plačiai naudotas sintezatorius, veikęs operacinėje sistemoje *MS-DOS*. Vienas jo išskirtinių bruožų, kad tai buvo aparatūrinis įrenginys, visi kiti lietuviški sintezatoriai yra programiniai moduliai. Kitas išskirtinis bruožas – formantinis sintezės metodas, generuojantis grynai dirbtinį garsą. Išsamiau apie formantinį metodą žr. Klatt (1980). Visuose kituose lietuviškuose sintezatoriuose naudojamas konkatenacinis metodas, kuris remiasi natūralių diktoriaus balso įrašų segmentų jungimu.

Sintezatorius *Aistis* sukurtas 1996 m. VU Filosofijos fakulteto Specialiosios psichologijos laboratorijoje (Kasparaitis, 2001b), projekto vadovas doc. Albinas Bagdonas. Kadangi balso kokybę labiausiai lemia taisyklingas kirčiavimas ir transkribavimas bei naudojama garsų bazė, todėl *Aisčio* ir kitų sintezatorių šie du komponentai bus aprašyti išsamiau. Kirčiavimą ir transkribavimą atliko Arijaus Ketlėriaus sukurta programinė įranga, kuri naudoja kirčiavimo taisykles ir apie 3000 kamienų žodyną. Garsų bazę sudarė 480 įvairaus ilgio segmentų, iškirptų iš diktoriaus Juozo Šalkausko balso įrašų. Sintezuotas signalas gaunamas jungiant nemodifikuotus signalo segmentus.

Sintezatoriai *Apollo* ir *Aistis* praktiškai jau nebenaudojami, todėl jų balso kokybė netiriama.

Gintaras

Kompanija *RosaSOFT* iš Čekijos yra sukūrusi daugelio kalbų sintezatorius: čekų, slovakų, vengrų, serbų, kroatų ir kitų. 2000 metais į jų sukurtą programinį *Windows* aplinkoje veikiantį sintezatorių *WinTalker* pirmą kartą pabandyta įtraukti ir lietuvišką balsą. Kadangi vartotojai pageidavo aukštesnės kokybės, 2003 m. kompanija *RosaSOFT* kartu su šio straipsnio autoriumi į sintezatorių *WinTalker* įtraukė tris lietuviškus balsus: *Gintaras*, *Aistis* ir *Aistis 2* (išsamiau žr. *Aistis 2*). 2006 m. pasirodė patobulinta *WinTalker* versija, turinti tik du balsus: *Gintaras* ir *Aistis 2*.

Balso *Gintaras* signalas formuojamas konkatenaciniu metodu naudojant apie 1500 įvairaus ilgio natūralios kalbos segmentų, kuriuos kompanija *RosaSOFT* iškirpo iš diktoriaus Gintaro Deksnio įrašų. Garsų ilgio ir tono modifikavimui bei garsų jungimui naudojamas *RosaSOFT* pasiūlytas metodas, kuris viešai nepublikuotas.

Balsas Gintaras naudoja šio straipsnio autoriaus sukurtą kirčiavimo-transkribavimo modulį, kurio atskiri komponentai yra publikuoti (Kasparaitis, 1999; 2000; 2001a), taip pat publikuotas išsamus aprašymas (Kasparaitis, 2001c). Kirčiavimui naudojami kamienų žodynai (virš 60 000 daiktavardžių ir būdvardžių, virš 8000 veiksmažodžių, apie 3000 nekaitomų žodžių) ir morfologinės kaitymo taisyklės. Nuo 2006 m. versijos modulis papildytas klitikų paieška (Anbinderis, Kasparaitis, 2007) ir homografų vienareikšminimu (Anbinderis, Kasparaitis, 2009). Šis modulis naudojamas ir visuose vėlesniuose (išskyrus sintezatorių *Egidius*) sintezatoriuose, todėl ir kirčiavimo kokybė juose yra identiška.

Balsas *Gintaras* buvo plačiai naudojamas Lietuvos aklųjų ir silpnaregių kartu su ekrano skaitymo programa *JAWS*. Jo populiarumą galima paaiškinti tuo, kad tai pirmasis lietuviško balso sintezatorius, pritaikytas *Windows* aplinkai.

Aistis 2

Vardu *Aistis 2* vadinsime visą sintezatorių klasę. Visus juos vienija tai, kad signalas formuojamas konkatenaciniu metodu naudojant 5003 difonus, iškirptus iš diktoriaus Gintaro Deksnio balso įrašų. Difonų bazę sudarė prof. Aleksas Girdenis (VU Filologijos fak.). Jo pasiūlyta fonemų sistema (92 fonemos) buvo naudojama ir vėlesniuose sintezatoriuose *SINT.AS* bei *LIEPA* (išsamiau apie garsų bazę žr. Kasparaitį (2005)). Garsų trukmėms ir pagrindiniam tonui modifikuoti bei garsams jungti skirtinguose sintezatoriuose buvo naudojami skirtingi metodai:

- Kompanijos *RosaSOFT* pasiūlytas metodas;
- *MBROLA* algoritmas (apie šį algoritmą išsamiau žr. Dutoit ir kt., 1996);
- *TD-PSOLA* algoritmas (apie šį algoritmą išsamiau žr. Moulines, Charpentier, 1990).

Pirmasis metodas vis dar naudojamas sintezatoriuje *WinTalker*, antrasis iki 2010 metų buvo naudotas svetainės *Text-Talk* sintezatoriuje, trečiasis metodas šiuo metu nebenaudojamas.

Šiame darbe balso kokybės tyrimuose naudosime *Aistis 2* versiją su *MBROLA* metodu.

Egidius

2008 metais UAB *Etalinkas* ir kompanija *Sakrament* iš Baltarusijos sukūrė lietuviško balso sintezatorių, kalbantį vyrišku balsu. Sintezatoriaus balsas pavadintas *Egidius*, jis sukurtas buvusio sporto komentatoriaus Vasilijaus Kuzminsko balso pagrindu. Sintezatorių galima laisvai parsisiųsti iš *Etalinko* svetainės. Diktoriaus Andriaus Kavaliausko balso pagrindu minėtos kompanijos yra sukūrusios ir dar vieną sintezatorių, tačiau jis viešai neplatinamas, todėl apie jį nebus kalbama. Kompanija *Sakrament* dar yra sukūrusi rusų ir anglų kalbų sintezatorius.

Sintezės metodas konkatenacinis, naudojama apie 6500 kontekstinių fonemų, t. y. fonemos dydžio segmentų, paimtų iš įvairiausių kontekstų. Be to, iki kirčiuoto skiemens ir po kirčiuoto skiemens naudojami skirtingi garsai, t. y. skiriamos prieškirtinės ir pokirtinės fonemos. Visi dvibalsiai ir mišrieji dvigarsiai laikomi savarankiškomis fonemomis. Garsams modifikuoti ir jungti naudojamas *Sakrament* pasiūlytas metodas, kuris viešai nepublikuotas.

Tekstui kirčiuoti iš pradžių naudojamas kirčiuotų žodžių sąrašas, kurį sudarė kompanija *Sakrament* iš 1 mln. žodžių kirčiuoto teksto. Šitaip sukirčiuojama apie 75 % žodžių. Likusiems žodžiams kirčiuoti naudojamas Tomo Anbinderio sukurtas algoritmas (Anbinderis, 2010), kuris automatiškai sugeneravo kirčiavimo taisykles remdamasis tuo pačiu 1 mln. žodžių kirčiuotu tekstu.

SINT.AS

Sintezatorių *SINT.AS* sukūrė šio straipsnio autorius kartu su UAB *Algoritmų sistemos*. Pirmą kartą viešai apie sintezatoriaus sukūrimą paskelbta svetainėje *Inovacijų prizas 2013*. Sintezatorius geba kalbėti dviem balsais: vyrišku ir moterišku. Tai pirmasis lietuviškas moteriško balso sintezatorius. Balsai vadinami *Laima* (diktorė Laima Kybartienė) ir *Marijus* (diktorius Marijus Žiedas). Nuo ankstesnių sintezatorių *SINT.AS* skiriasi dviem ypatumais:

- specialiai atrinkti diktoriai (išsamiau žr. *Diktorių atranka balso sintezatoriui*);
- sintezei panaudotas vienetų parinkimo metodas (išsamiau žr. *Vienetų parinkimo metodo esmė*).

Vienetų parinkimo metodas naudoja fonemų lygmenyje anotuotus ištisinės šnekos įrašus. Jų apimtis moteriškam ir vyriškam balsui yra atitinkamai 271 ir 224 MB (1 val. 42 min. ir 1 val. 24 min.), arba 2000 sakinių, kuriuos sudaro apie 77 tūkst. garsų.

SINT.AS balsai nėra viešai prieinami.

Projektas LIEPA

Projektą LIEPA („Lietuvių šneka valdomos paslaugos“) vykdė Vilniaus universitetas kartu su partneriais 2013–2015 metais. Vykdant projektą buvo sukurtas keturių balsų sintezatorius. Specialiai stengtasi, kad balsai būtų kuo skirtingesni, todėl sukurtas jaunatviškas moteriškas, jaunatviškas vyriškas, vyresnio amžiaus moteriškas ir vyresnio amžiaus vyriškas balsai. Balsai vadinami atitinkamai: *Aistė* (diktorė Aistė Diržiūtė), *Edvardas* (Edvardas Kubilius), *Regina* (Regina Jokubauskaitė) ir *Vladas* (Vladas Bagdonas).

Sintezei naudojamas vienetų parinkimo algoritmas, t. y. sudarytos kiekvieno diktoriaus ištisinių įrašų bazės. Jų apimtis: nuo 483 iki 577 MB, t. y. nuo 3 val. 2 min. iki 3 val. 38 min. arba šiek tiek daugiau nei 5000 sakinių, kuriuose – daugiau kaip 161 tūkst. garsų.

Projekto LIEPA balsų galima pasiklausyti svetainėje *LIEPA – Teksto sintezatorius*. Balsai *Regina* ir *Edvardas* buvo specialiai pritaikyti akliesiems, juos galima parsisiųsti iš svetainės *LIEPA – Sintezatorius akliesiems* ir įsidiegti kompiuteryje. Be to, balsas *Regina* „skaito“ *Lietuvos žinias*, balsai *Aistė* ir *Edvardas* „vaidino“ trupės *Rimini Protokoll* spektaklyje *Remote Vilnius*, visi keturi balsai „įdarbinti“ paslaugoje *RoboBraille*, kurios esmė tokia: el. paštu nusiuntus tekstinį failą adresu regina@robobraille.org, edvardas@robobraille.org, aiste@robobraille.org arba vladas@robobraille.org gaunama nuoroda į atitinkamu balsu susintezuotą garso failą.

Diktorių atranka balso sintezatoriui

Kiekvienas sintezatorius kuriamas remiantis konkretaus žmogaus balso parametrais, t. y. stengiamasi, kad sintezatoriaus balsas būtų panašus į to žmogaus balsą. Taigi visų pirma reikia pasirinkti diktorių. Visuotinai sutariama, kad nuo diktoriaus parinkimo priklauso galutinė sintezuoto balso kokybė. Huang ir kt. (2001, p.800) teigia, kad diktoriaus parinkimas nulemia iki 0,3 balo subjektyvioje penkiabalėje MOS (angl. *mean opinion score*) skalėje. Nors diktoriaus parinkimas svarbus, tačiau literatūros apie tai yra labai nedaug (Syrdal ir kt., 1998; Braga ir kt., 2007).

Bragos ir kt. (2007) išsamiai aprašyta diktorės atranka portugalų kalbos sintezatoriui. Pirmiausiai iš 485 kandidačių atrinktos 74 remiantis anketiniais duomenimis (gimtoji kalba, išsilavinimas, diktoriaus darbo patirtis ir t. t.). Iš pastarųjų diktorių atsiųstų balso pavyzdžių remiantis subjektyviais testais atrinkta 12 kandidačių. Subjektyvių testų metu klausytojai penkiabalėje sistemoje vertino tokius parametrus kaip balso malonumas, suprantamumas, tartis, kirčiavimas, išraiškingumas, išskirtinumas, jausmingumas, tinkamumas skaityti nau-

jienas, instrukcijas, el. pašto laiškus ir pan. Galiausiai visos diktorės vienodomis įrašymo sąlygomis perskaitė tą patį tekstą (219 žodžių). Galutiniam vertinimui taip pat galima panaudoti subjektyvius testus, tačiau egzistuoja ir objektyviai apskaičiuojami parametrai, kurie koreliuoja su subjektyviais vertinimais, pavyzdžiui, žmonėms patinka moteriški balsai, kurių pagrindinio tono vidurkis yra nuo 186 iki 206 Hz, arba, kai dusliųjų segmentų (garsų /s/, /š/) energija yra didelė lyginat su skardžiųjų segmentų energija (Syrdal ir kt., 1998). Dar vienas būdas – tai atlikti bandomąją sintezę ir subjektyviai įvertinti rezultatą. Tam galima tiesiog modifikuoti signalo pagrindinį toną ir (arba) garsų trukmę (Braga ir kt., 2007) arba iškirpti tam tikrą aibę difonų ir juos perstatant susintezuoti kelias frazes (Syrdal ir kt., 1998). Subjektyviai įvertinus paaiškėja, kad vieni balsai tokioms modifikacijoms atsparesni už kitus, o sintezėje tai svarbu.

Tiek kuriant *SINT.AS*, tiek projekte LIEPA į atranką buvo kviečiami tik profesionalūs diktoriai ir aktoriai, taigi galima teigti, kad buvo atliekamas tik paskutinis atrankos etapas. Sintezatoriui *SINT.AS* moteriškas balsas atsirinktas iš 6-ių kandidatų, o vyriškas balsas – iš 3-jų. Į atranką buvo kviečiama visa diktorių grupė ir iš jų išsirenkamas geriausias. Projekte LIEPA diktoriai į atranką buvo kviečiami po vieną. Atranka buvo nutraukiama radus bent minimalius reikalavimus tenkinantį diktorių. Taip buvo siekiama sumažinti į atranką kviečiamų diktorių skaičių. Kaip vyresnis moteriškas balsas pakviesta *SINT.AS* atrankoje dalyvavusi diktorė, kaip vyresnis vyriškas balsas pasirinktas atrankoje ketvirtuoju dalyvavęs kandidatas, likusių dviejų balsų atrankose dalyvavo po 3-is kandidatus.

SINT.AS atrankoje diktoriai turėjo perskaityti specialų 13 sakinių tekstą (309 garsai), kuriame visi 92 fonemų sistemos garsai panaudoti bent po vieną kartą. Projekto LIEPA atrankoje šis tekstas dar papildytas 7 sakiniais (339 garsais) su balsių junginiais, taip pat 14 sakinių (491 garsu) su sprogstamųjų priebalsių junginiais. Diktoriai atrinkti pagal tokius kriterijus:

- Pagrindinio tono monotoniškumas;
- Sklandus balsių jungimas;
- Aiškus garsų artikuliavimas, ypač sakinių pabaigoje;
- Signalo formos paprastumas: ar lengva vizualiai įžiūrėti pagrindinio tono periodus, garsų ribas;
- Ar balsai pakankamai skirtingi (projekte LIEPA).

Vienetų parinkimo metodo esmė

Vienetų parinkimo metodas pirmą kartą pasiūlytas Hunt ir Black (1996). Jis naudoja anotuotus ištisinius diktoriaus balso įrašus. Sintezuojant idealiu atveju galima rasti ir visą įrašytą sakinį, o nepavykus imami mažesni segmentai, blogiausiu atveju sintezuotas įrašas suklijuojamas iš atskirų fonemų. Tinkamiausi įrašo segmentai parenkami remiantis garsų keitimo ir garsų jungimo kainomis, kainų prasmę iliustruoja toliau pateiktas pavyzdys. Kainos gali būti apskaičiuojamos remiantis įvairiais akustiniais arba fonologiniais požymiais, arba įvairiais jų deriniais (Taylor, 2009, p.512). Kuriant lietuviškus vienetų parinkimu grįstus sintezatorius buvo pasinaudota Yi ir Glass (2002) pasiūlytais fonologiniais požymiais. Vienetų parinkimo algoritmas išsamiau aprašytas Kasparaičio ir Anbinderio (2014).

Panagrinėkime pavyzdį. Tarkime, norime susintezuoti sakinį **Vãsara Palangojè**. Projekto LIEPA garsų bazėse yra toliau pateikti penki sakiniai, iš kurių ir paimami reikiami segmentai:

Vãsara Palangojè

Vãsaros jūra nuplóvė spalvàs

Ẽglė S**ara**páitė ir Giẽdrė Paugáitė

kad visi kiti jaũčia tikrą **pal**áimą

herefòrdų, aberdinų, **ang**ùsų

ar prisimeni nãgą žaliojè lank**ojè**

Galima pastebėti, kad žodis **Palangojè** yra paskutinis sakinyje, todėl ir segmentai jam imami iš paskutinių sakinių žodžių. Žodis **Vãsara** yra pirmas sakinyje, todėl pirmasis segmentas taip pat paimtas iš pirmo sakinio žodžio, o štai antrasis segmentas paimtas iš sakinio vidurio, nes nepavyko rasti reikiamo segmento sakinio pradžioje. Tai, kad šiame pavyzdyje paimti trijų fonemų dydžio segmentai, yra visiškas atsitiktinumas, segmentų dydis gali būti bet koks.

Jungimo kainų veikimą paaiškinsime žemiau pateiktu pavyzdžiu. Klausimas, kodėl imama po tris fonemas, o ne keturios ir dvi, t. y.

Vãsaros jūra nuplóvė spalvàs

Ẽglė S**ara**páitė ir Giẽdrė Paugáitė

o ne

Vãsaros jūra nuplóvė spalvàs

Ẽglė Sa**ra**páitė ir Giẽdrė Paugáitė

Atsakymas: **sa** jungimo kaina mažesnė nei **ar**, nes kuo garsai skirtingesni, tuo labiau tikėtina, kad žmogus garsų jungimo nepajus.

Keitimo kainų prasmę iliustruosim kitu pavyzdžiu. Klausimas, kodėl imama

herefòrdų, aberdinų, **ang**ùsų

ar prisimeni nãgą žaliojè lank**ojè**

o ne

herefòrdų, aberdinų, **ang**ùsų

galimýbes táikyti juõs Lietuv**ojè**

Atsakymas: k**o** keitimo į g**o** kaina mažesnė nei v**o** keitimo į g**o**, nes panašesnis kontekstas.

Dar vienas svarbus klausimas, kokio dydžio turėtų būti vienetų parinkimui naudojama garsų bazė. Taylor (2009, p.529) teigia, kad turėtų būti bent viena valanda įrašo. Priešingu atveju reikiami segmentai bus labai išbarstyti, todėl nebus išnaudojamas tas metodo privalumas, kai bazėje randami dideli ištisai įrašytos kalbos fragmentai. Daugelyje sistemų naudojamos bazės, apimančios 5 valandas ir daugiau. Taigi *SINT.AS* garsų bazių apimtis tik truputį viršija rekomenduojamą minimumą, o projekto LIEPA garsų bazių apimtis yra artima vidutinei.

Ankstesniais metais atlikti sintezuoto balso vertinimai

Pirmasis sintezatoriumi *Apollo* sintezuoto balso suprantamumo tyrimas išsamiai aprašytas Bagdono ir Laugalio (2002). Tyrime buvo vertintas atskirų garsų, atskirų žodžių, žodžių sakinyje ir sakinio visumos suprantamumas. Minėtas darbas svarbus tuo, kad buvo sudaryti trys rinkiniai po 30 sakinių, tie patys sakiniai ir tyrimo metodika buvo naudojami ir visuose vėlesniuose lietuviško sintezuoto balso suprantamumo vertinimuose. Sakinius galima rasti Kasparaičio (2001c) disertacijos priede (kompaktiniame diske).

Kasparaitis (2001c) vertino sintezatorių *Apollo*, *Aistis*, *Aistis* be kirčiavimo modulio ir diktoriaus balso suprantamumą. Parodyta, kad į sintezatorių *Aistis* įdiegus kirčiavimo modulį jo suprantamumas padidėjo nuo 78,6 % iki 93,2 %.

Kasparaitis (2011) lygino sintezatorių *Gintaras*, *Aistis 2* ir *Egidius* suprantamumą. Sintezuoto balso suprantamumui įvertinti buvo pasitelkti klausytojai – 8 vyrai ir 7 moterys, amžius nuo 10 iki 57 metų, anksčiau su balso sintezatoriais nesusidūrę, dešimties iš jų gimtoji kalba – lietuvių, likusių penkių gimtoji kalba – rusų. Klausytojų grupes sudarė nuo 1 iki 3 žmonių. Buvo vertinami tik reikšminiai sakinio žodžiai, o funkciniai žodžiai (jungtukai, prielinksniai) – ne. Žodžių sakinyje suprantamumas, sakinio visumos suprantamumas (vidurkis ir pasikliautinis intervalas su parametru $\alpha = 0{,}05$) ir subjektyvus priimtinumas 5 balų skalėje pateikti 1 lentelėje.

1 lentelė
Sintezuoto balso suprantamumas ir priimtinumas (Kasparaitis, 2011)

Sintezatorius	Žodžių sakinyje suprantamumas, %
Aistis 2	92,41 ± 3,34
Egidius	88,10 ± 3,45
Gintaras	81,65 ± 4,50
	Sakinio visumos suprantamumas, %
Aistis 2	82,00 ± 6,68
Egidius	71,33 ± 6,31
Gintaras	65,33 ± 6,59
	Subjektyvus priimtinumas 5 balų skalėje
Aistis 2	3,80
Egidius	3,58
Gintaras	3,33

Buvo vertintas ir naujai sukurtų sintezatorių *SINT.AS* suprantamumas. Vertino penkios 2-ojo kurso studentės (VU Filologijos fak.). Rezultatai (vidurkis ir pasikliautinis intervalas su parametru $\alpha = 0{,}05$) pateikti 2 lentelėje.

2 lentelė
Sintezatorių suprantamumo rezultatai, gauti projekto *SINT.AS* pabaigoje

Sintezatorius	Žodžių sakinyje suprantamumas, %
Marijus	97,58 ± 1,69
Laima	93,84 ± 1,04
Aistis 2	86,71 ± 6,77

Iš 1 ir 2 lentelių galima pastebėti, kad skirtingu laiku skirtingomis sąlygomis gauti rezultatai labai skiriasi, todėl jų lyginti negalima. Šiame darbe toliau aprašyti eksperimentai buvo atlikti su visais sintezatoriais iš naujo vienodomis sąlygomis.

Sintezuoto balso suprantamumo vertinimas

Sintezuoto balso suprantamumui vertinti buvo pasitelkti žmonės-klausytojai (VU MIF informatikos specialybės antro kurso studentai, amžius apie 20 metų, neturintys klausos sutrikimų, anksčiau nedirbę su testuojamų sintezatorių balsais). Visų klausytojų gimtoji kalba – lietuvių. Iš viso dalyvavo 46 klausytojai, suskirstyti į 4 grupes: 14, 8, 9 ir 15 klausytojų.

Suprantamumui vertinti buvo naudojami sintezuoti trumpi prasmingi sakiniai, sudaryti tik iš bendrinių lietuvių kalbos žodžių (jokių asmenvardžių, vietovardžių ar tarptautinių žodžių). Sakinių ilgis – 4–7 žodžiai, vidutinis sakinio ilgis – 5,5 žodžio. Kiekvienai klausytojų grupei buvo pateikti visų devynių sintezatorių po 10 sakinių, visi 90 sakinių skirtingi. Visoms keturioms klausytojų grupėms sakiniai buvo pateikti ta pačia tvarka, tačiau buvo keičiamas sintezatorių eiliškumas, tokiu būdu kiekvienas sintezatorius buvo testuotas su 40 skirtingų sakinių.

Sakiniai buvo iš anksto įrašyti į garsinius failus, o ne sintezuojami vertinimo metu. Prieš sintezuojant buvo apskaičiuotas kiekvieno sintezatoriaus numatytasis greitis ir pastebėta, kad greičiai skiriasi. Akivaizdu, kad kuo sintezatoriaus kalba lėtesnė, tuo lengviau ji suprantama. Buvo nuspręsta visų sintezatorių greičius suvienodinti iki 120 žodžių per minutę. Lėtinimo koeficientai pateikti 3 lentelėje. Sintezatoriai išrikiuoti nuo lėčiausio iki greičiausio.

Iš 3 lentelės matyti, kad trys seniausieji sintezatoriai pagreitinti, o šeši naujausieji – sulėtinti.

Taip pat buvo nuspręsta suvienodinti sintezatorių balsų garsumą, nes kuo balsas garsesnis, tuo suprantamesnis. Visų sintezatorių garsumas suvienodintas iki 73,3 dB. Toks garsumas užtikrina, kad balsas būtų pakankamai girdimas, tačiau signalo reikšmės neviršytų leistino diapazono (nuo – 2^{15} iki 2^{15}).

Klausytojai turėjo išklausyti po vieną sakinį ir užrašyti popieriaus lape tai, ką išgirdo. Kiekvienas sakinys buvo išklausytas tik vieną kartą. Klausytojams buvo suteikta pakankamai laiko sakiniui užrašyti.

Klausytojų užrašai buvo patikrinti rankiniu būdu. Suprantamumas nusakomas kaip teisingai suprastų žodžių skaičiaus ir visų išklausytų žodžių skaičiaus santykis. Sintezatoriaus balso suprantamumas apskaičiuojamas kaip visų klausytojų įvertinimų vidurkis. Apskaičiuotas ir pasikliautinis intervalas su parametru α=0,05. Rezultatai pateikti 4 lentelėje. Sintezatoriai išrikiuoti suprantamumo mažėjimo tvarka.

Iš 4 lentelės matyti, kad sintezatoriai, kuriuose naudojamas vienetų parinkimo metodas, gerokai lenkia ankstesnius sintezatorius, kuriuose naudojama po vieną kiekvieno garso egzempliorių. Projekto LIEPA balsai kiek suprantamesni už *SINT.AS* balsus, nes juose naudojamos didesnės garsų bazės. Pats suprantamiausias balsas – *Regina*. Iš vyriškų balsų suprantamiausias yra *Edvardas*. Būtent šiuos balsus rekomenduojama rinktis sintezatorių naudotojams tuo atveju, kai svarbu teisingai suprasti kiekvieną žodį.

3 lentelė
Sintezatorių lėtinimo koeficientai

Sintezatorius	Lėtinimo koeficientas
Egidius	0,81
Gintaras	0,82
Aistis 2	0,95
Aistė	1,01
Laima	1,04
Vladas	1,15
Edvardas	1,23
Marijus	1,25
Regina	1,28

4 lentelė
Žodžių sakinyje suprantamumas

Sintezatorius	Žodžių sakinyje suprantamumas, %
Regina	97.52 ± 0,69
Edvardas	96.22 ± 0,97
Aiste	95.11 ± 1,05
Marijus	95,01 ± 1,30
Vladas	92.33 ± 1,50
Laima	91,97 ± 1,91
Aistis 2	81,72 ± 2,82
Egidius	80,84 ± 2,85
Gintaras	71,84 ± 3,23

Kalbant apie sakinių suprantamumą, Bagdonas ir Laugalys (2002) atskirai skaičiavo visiškai suprastus sakinius (angl. *totally correct reproductions*) ir suprastus sakinius (angl. *correct reproductions*). Sakinys laikytas visiškai suprastu, jei visi žodžiai užrašyti be jokių iškraipymų. Visiškai suprastų sakinių skaičius nėra labai informatyvus, nes tiesiogiai priklauso nuo žodžių suprantamumo, todėl šiame darbe nebuvo skaičiuotas. Daugeliu atvejų, pavyzdžiui, kai sintezatorius naudojamas grožinei literatūrai skaityti, klausančiajam pakanka suvokti sakinio prasmę, todėl naudingesnis įvertis gali būti sakinio visumos suprantamumas, būtent jis ir vertintas šiame darbe. Sakinys laikytas suprastu, jei iš esmės teisingai perteikta sakinio mintis, nors tam tikri iškraipymai leistini. Keletas pavyzdžių (pakeistas ar praleis-

5 lentelė
Sakinio visumos suprantamumas.

Sintezatorius	Sakinio visumos suprantamumas, %
Regina	92,17 ± 2,35
Edvardas	87,83 ± 2,79
Marijus	86,74 ± 3,17
Aiste	84,57 ± 3,38
Vladas	79,57 ± 3,55
Laima	78,26 ± 4,49
Aistis 2	61,52 ± 5,27
Egidius	59,13 ± 4,67
Gintaras	51,09 ± 4,61

tas žodis nurodytas skliaustuose): „Filmą neigiamai vertino (įvertino) daugelis kritikų“, „Sakoma, (kad) norėti reiškia galėti“, „Klaidingai užrašytas adresas sutrukdė (sukliudė) gauti laišką“. Sakinio visumos suprantamumo vertinimo rezultatai pateikti 5 lentelėje. Sintezatoriai išrikiuoti suprantamumo mažėjimo tvarka.

Diktorių išsidėstymas pagal suprantamumą panašus, kaip ir 4 lentelėje, tik *Marijus* aplenkė *Aistę*.

Sintezuoto balso priimtinumo vertinimas

Tiems patiems 46-iems klausytojams buvo pateikta dar viena užduotis: subjektyviai dešimties balų skalėje įvertinti kiekvieno sintezatoriaus balso priimtinumą. Šios užduoties metu klausytojams buvo pateikta po vieną kiekvieno sintetinio balso sakinį. Įvertinimų vidurkiai ir pasikliautiniai intervalai su parametru α=0,05 pateikti 6 lentelėje. Sintezatoriai išrikiuoti priimtinumo mažėjimo tvarka. Priežastys, kodėl klausytojai būtent tokį balą skyrė vienam ar kitam sintezatoriui, nebuvo tirtos.

Iš 6 lentelės matyti, kad sintezatoriai, kuriuose naudojamas vienetų parinkimo metodas, vartotojui yra daug priimtinesni. Balso *Egidius* žemiausią priimtinumą galėjo lemti ir itin nevykęs diktoriaus parinkimas. Tai, kad *SINT.AS* balsai pasirodė priimtiniausi, galima būtų paaiškinti itin sėkmingu diktorių pasirinkimu. Nepasiteisino projekte LIEPA vykdyta diktorių atranka, kai diktoriai buvo kviečiami po vieną, nes šiuo atveju atranką gali įveikti ir diktoriai, kurie tenkina tik minimalius reikalavimus. Kaip matome, priimtinumas nėra tiesiogiai susijęs su suprantamumu. Taigi sintezatorių naudotojams rekomenduojama rinktis balsus *Marijus* arba *Laima*, jei svarbu, kad balsas būtų malonus, nevargintų, o suprantamumas nėra esminis dalykas, pvz., sintezuotu balsu ketinama skaityti ilgą grožinės literatūros kūrinį.

6 lentelė
Sintezatorių priimtinumas.

Sintezatorius	Priimtinumo įvertis 10 balų skalėje
Marijus	9,13 ± 0,24
Laima	8,68 ± 0,29
Regina	8,66 ± 0,27
Aistė	8,29 ± 0,32
Edvardas	8,14 ± 0,39
Vladas	7,25 ± 0,40
Aistis 2	5,34 ± 0,42
Gintaras	4,97 ± 0,46
Egidius	4,78 ± 0,49

Išvados

Šiame darbe buvo atliktas devynių naujausių lietuviško balso sintezatorių kokybės tyrimas pasitelkiant žmones-klausytojus. Vertintas žodžių sakinyje ir sakinio visumos suprantamumas bei subjektyvus sintetinio balso priimtinumas. Remiantis tyrimo rezultatais galima padaryti tokias išvadas:

1 Norint palyginti skirtingu laiku sukurtus sintezatorius, negalima pasinaudoti ankstesniais metais gautais vertinimo rezultatais, nes praktiškai neįmanoma užtikrinti vienodų eksperimento sąlygų. Visus vertinimo eksperimentus reikia atlikti iš naujo.

2 Sintezatoriai, kuriuose naudojamas vienetų parinkimo metodas, yra žymiai suprantamesni ir priimtinesni nei sintezatoriai, turintys tik po vieną kiekvieno garso egzempliorių. Sintezatorių kūrėjai turėtų rinktis metodus, kurie naudoja daug kiekvieno garso egzempliorių.

3 Sintetinio balso suprantamumas, kai naudojamas vienetų parinkimas, priklauso nuo garsų bazės dydžio. Kuriamo sintezatoriaus suprantamumas padidės padidinus garsų bazę.

4 Absoliučiai suprantamiausias balsas – *Regina*. Iš vyriškų balsų suprantamiausias yra *Edvardas*. Būtent šiuos balsus turėtų rinktis sintezatorių naudotojai, jei esminis dalykas yra suprantamumas.

5 Priimtinumas nėra tiesiogiai susijęs nei su suprantamumu, nei su garsų bazės dydžiu. Priimtinumą labiau lemia asmeninės diktoriaus savybės, todėl sintezatorių kūrėjams itin svarbu sėkminga diktoriaus atranka.

6 Absoliučiai priimtiniausias balsas – *Marijus*. Iš moteriškų balsų priimtiniausias balsas yra *Laima*, nors beveik tokie pat rezultatai gauti ir balsui *Regina*. Būtent šiuos balsus turėtų rinktis sintezatorių naudotojai, jei esminis dalykas yra priimtinumas.

Literatūra

1. Anbinderis, T., Kasparaitis, P., 2007. Klitikų paieškos lietuviškame tekste algoritmai. *Kalbų studijos/ Studies about languages*, nr. 10, pp.30–37. Prieiga per internetą: http://www.kalbos.lt/zurnalai/10_numeris/05.pdf [žiūrėta 2015 m. spalio mėn.].
2. Anbinderis, T., Kasparaitis, P., 2009. Lietuvių kalbos homografų vienareikšminimas remiantis leksemų ir morfologinių pažymų vartosenos dažniais. *Kalbų studijos/ Studies about languages*, nr. 14, pp.25–31. Prieiga per internetą: http://www.kalbos.lt/zurnalai/14_numeris/05.pdf [žiūrėta 2015 m. spalio mėn.].
3. Anbinderis, T., 2010. Automatic Stressing of Lithuanian Text Using Decision Trees. *Information Technology and Control*, 39(1), pp.61–67. Prieiga per internetą: http://www.itc.ktu.lt/index.php/ITC/article/download/12084/6732 [žiūrėta 2015 m. spalio mėn.].
4. Bagdonas, A., Laugalys, F., 2002. Evaluation of synthetic speech quality: A comparative study of several computer-based speech synthesizers. *Psichologija*, 25, pp.1–22.
5. Braga, D. ir kt., 2007. Subjective and Objective Assessment of TTS Voice Font Quality, Proc. of SPECOM, pp.306–311. Prieiga per internetą: http://download.microsoft.com/download/A/0/B/A0B1A66A-5EBF-4CF3-9453-4B13BB027F1F/SPECOM2007.pdf [žiūrėta 2015 m. spalio mėn.].
6. Dutoit, T. ir kt., 1996. The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes, *Proc. ICSLP 96*, pp.1393–1396. Prieiga per internetą: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.16.5411&rep=rep1&type=pdf [žiūrėta 2015 m. spalio mėn.].
7. Hunt, A., Black, A., 1996. Unit selection in a concatenative speech synthesis system using a large speech database. *In: ICASSP 1996*, pp.373–376. Prieiga per internetą: https://www.ee.columbia.edu/~dpwe/e6820/papers/HuntB96-speechsynth.pdf [žiūrėta 2015 m. spalio mėn.].
8. Yi, J.,Glass, J., 2002. Information-theoretic criteria for unit selection synthesis. *In: Interspeech 2002*, pp.2617–2620. Prieiga per internetą: https://groups.csail.mit.edu/sls/publications/2002/jonyi.pdf [žiūrėta 2015 m. spalio mėn.].
9. Kasparaitis, P., 1999. Transcribing of the Lithuanian Text Using Formal Rules. *Informatica*, 10(4), pp.367–376. Prieiga per internetą: http://www.mii.lt/informatica/pdf/INFO183.pdf [žiūrėta 2015 m. spalio mėn.].
10. Kasparaitis, P., 2000. Automatic Stressing of the Lithuanian Text on the Basis of a Dictionary. *Informatica*, 11(1), pp.19–40. Prieiga per internetą: http://www.mii.lt/informatica/pdf/INFO192.pdf [žiūrėta 2015 m. spalio mėn.].
11. Kasparaitis, P., 2001a. Automatic Stressing of the Lithuanian Nouns and Adjectives on the Basis of Rules. *Informatica*, 12(2), pp.315–336. Prieiga per internetą: http://www.mii.lt/informatica/pdf/INFO218.pdf [žiūrėta 2015 m. spalio mėn.].

12. Kasparaitis, P., 2001b. Lietuvių kalbos kompiuterinis sintezatorius „Aistis“. *Garso korta 2001 (CD)*, Republic Science-Technology Conference, KTU, Kaunas: Technologija.
13. Kasparaitis, P., 2001c. *Lietuvių kalbos kompiuterinė sintezė*. Daktaro disertacija (fiziniai mokslai, informatika (09P). Vilniaus universitetas: Vilnius. Prieiga per internetą: http://www.mif.vu.lt/~pijus/publikacijos/KaspDis.pdf [žiūrėta 2015 m. spalio mėn.].
14. Kasparaitis, P., 2005. Diphone Databases for Lithuanian Text-to-Speech Synthesis. *Informatica*, 16(2), pp.193–202. Prieiga per internetą: http://www.mii.lt/informatica/pdf/INFO583.pdf [žiūrėta 2015 m. spalio mėn.].
15. Kasparaitis, P., 2011. *Lietuviško sintezatoriaus sukūrimo galimybių studija*. (UAB *Epasas.lt* užsakymas). Vilnius.
16. Kasparaitis, P., Anbinderis, T., 2014. Building Text Corpus for Unit Selection Synthesis. *Informatica*, 25(4), pp.551–562. http://dx.doi.org/10.15388/Informatica.2014.29
17. Klatt, D. H., 1980. Software for a Cascade/Parallel Formant Synthesizer. *J. Acoust. Soc. Am.*, 67(3), pp.971–995. Prieiga per internetą: http://www.fon.hum.uva.nl/david/ma_ssp/doc/Klatt-1980-JAS000971.pdf [žiūrėta 2015 m. spalio mėn.].
18. Moulines, E., Charpentier, F., 1990. Pitch-synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. *Speech Communication*, 9, pp.453–467. http://dx.doi.org/10.1016/0167-6393(90)90021-Z
19. Schmidt-Nielsen, A., 1995. Intelligibility and Acceptability Testing for Speech Technology. In: Syrdal, A., Bennett, R., Greenspan, S. (eds.), *Applied Speech Technology*, CRC Press: Boca Raton/ Ann Arbor/ London/ Tokyo, pp.195–232. Prieiga per internetą: http://www.dtic.mil/dtic/tr/fulltext/u2/a252015.pdf [žiūrėta 2015 m. spalio mėn.].
20. Syrdal, A., Conkie, A., Stylianou, Y., 1998. Exploration of Acoustic Correlates in Speaker Selection for Concatenative Synthesis, *Proc. of ICSLP 1998*. Prieiga per internetą: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.54.3829&rep=rep1&type=pdf [žiūrėta 2015 m. spalio mėn.].
21. Taylor, P., 2009. *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511816338

Šaltiniai

1. *RosaSOFT* http://www.rosasoft.cz/
2. *JAWS* http://www.freedomscientific.com/Products/Blindness/JAWS
3. *Text-Talk* http://www.text-talk.com/
4. *Sakrament* http://sakrament.by/index.html
5. *Etalik (Lietuvių kalbos sintezatorius)* http://old.etalinq.com/pradzia/apie-mus/produktai/lietuviu-kalbos-sintezatorius/
6. *Inovacijų prizas 2013* http://www.inovacijuprizas.lt/index.php?1822664735
7. *LIEPA (teksto sintezatorius)* https://liepa.rastija.lt/Ieškotuvas/Teksto-sintezatorius
8. *LIEPA (Sintezatorius akliesiems)* https://www.raštija.lt/liepa/paslaugos-vartotojams/sintezatorius-akliesiems/7520
9. *Lietuvos žinios* http://lzinios.lt/lzinios/index.php
10. *Remote Vilnius* https://www.raštija.lt/naujienos/projekto-liepa-naujienos/projekto-liepa-sintezatorius-vaidino-spektaklyje-remote-vilnius-nuorodos/223?partner=11
11. *RoboBraille* http://www.robobraille.org/

Summary

Pijus Kasparaitis. Evaluation of Lithuanian Text-to-Speech Synthesizers

Text-to-speech synthesis of most popular languages is widely used for several decades, while the Lithuanian text-to-speech synthesis breakthrough occurred only in recent years. Six new Lithuanian synthetic voices appeared in 2013-2015. Therefore, there was a need to evaluate the newly created Lithuanian text-to-speech synthesizers. This paper presents a chronological review of the current Lithuanian text-to-speech synthesizers. Unit selection algorithm that was implemented in recent synthesizers *SINT.AS* and LIEPA and selection procedure of announcers are described in more detail because they were crucial to the synthesized voice quality. The main characteristics of synthetic voice are intelligibility and acceptability; they are assessed by involving human-listeners and the received

data are processed by statistical methods. Thus, this paper will investigate nine recent Lithuanian synthetic voices (*Regina*, *Edvardas*, *Aistė*, *Vladas*, *Laima*, *Marijus*, *Egidius*, *Aistis 2*, *Gintaras*) and evaluate what intelligibility of a synthetic voice is actually achieved, which synthesizer is better when comparing them with each other, to give advice to potential synthetic voice application developers when choosing a synthesizer voices and finally to show synthesizer developers the most promising areas of improvement.

Apie autorių

Pijus Kasparaitis

Fizinių mokslų daktaras, Vilniaus universiteto Matematikos ir informatikos fakulteto Kompiuterijos katedros docentas.

Mokslinių tyrimų sritis

Balso sintezė iš teksto ir kitos kompiuterinės lingvistikos sritys.

Adresas

Vilniaus universitetas Matematikos ir informatikos fakultetas Kompiuterijos katedra,
Didlaukio g. 47, LT-08303 Vilnius, Lietuva.

El. paštas:

pkasparaitis@yahoo.com