Nepriklausomas nuo kalbėtojo interneto naršyklės valdymas balsu

Živilė Ringelienė

Matematikos ir informatikos instituto doktorantė
Institute of Mathematics and Informatics, PhD student
Goštauto g. 12–204, LT-01108 Vilnius
El. paštas: zivile.ringeliene@gmail.com

Straipsnyje aprašomas lietuvių kalbai sukurtas sistemos, skirtos nepriklausomam nuo kalbėtojo naršyklės valdymui balsu, prototipas: veikimo principai, sistemoje naudojamų akustinių modelių savybės, vartotojo sąsaja ir valdymas. Valdymo sistema sukurta naudojant atskirų žodžių atpažinimo sistemos prototipą, grįstą paslėptaisiais Markovo modeliais. Naršyklės valdymas realizuotas integravus į atpažinimo sistemos programą kodą, kurio paskirtis – susieti atpažintą balso komandą su klavišų kombinacija. Atpažinus žodį, valdymo sistema imituoja atitinkamo klavišo spustelėjimą, į kurį naršyklė atsako veiksmu, pavyzdžiui, atidaromas naujas naršyklės langas, adresų peržiūros langas ir pan. Pažymima, kad tokiu būdu atpažinimo sistemos pagrindu galima sukurti programinę įrangą ir kitoms kompiuterinėms programoms valdyti balsu.

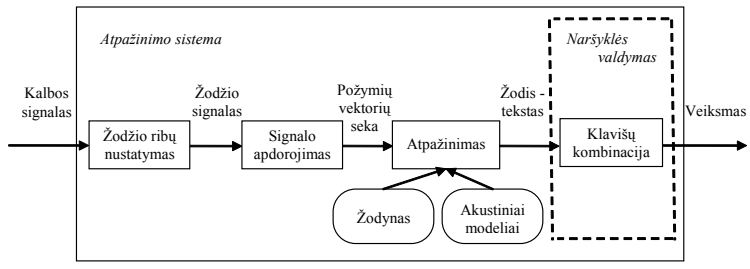
2007 metais Matematikos ir informatikos institute lietuvių kalbai buvo sukurtas interneto naršyklės valdymo balsu sistemos prototipas (Tamulevičius, 2007). Jo veikimas pagrįstas pavienių žodžių ir frazių atpažinimo sistema, kai naudojamas dinaminio laiko skalės kraipymo atpažinimo metodas (Tamulevičius, Lipeika, 2003; 2004). Sistema atlieka priklausomą nuo kalbėtojo žodžių atpažinimą.

Šio straipsnio tikslas – pristatyti nepriklausomo nuo kalbėtojo naršyklės valdymo balsu sistemos prototipą, kuris sukurtas atskirų lietuvių kalbos žodžių atpažinimo sistemos prototipo „Žodžių atpažintuvas“ (Filipovič, 2007) pagrindu. Atpažinimo sistemoje realizuotas paslėptaisiais Markovo modeliais grįstas atpažinimo metodas.

Naršyklės valdymo balsu sistemos struktūra

Naršyklės valdymo balsu sistemos darbo schema pateikiama 1 paveiksle. Pagrindinė sistemos dalis – atpažinimo sistema. Naršyklės valdymą balsu realizuoja kodas, integruotas į atpažintuvo programą.

Kai kalbos signalas patenka į atpažinimo sistemą, t. y. kai vartotojas į mikrofoną pasako žodžius, tarp kurių turi būti ~1–2 s pauzės, automatiškai



1 p a v. ***Naršyklės valdymo balsu sistemos darbo schema***

nustatomos žodžio ribos. Kalbos signalas dalijamas į 10–30 ms trukmės kadrus su 5–15 ms žingsniu tarp gretimų kadrų. Tada skaičiuojamas kiekvieno kadro normuotas signalo intensyvumo įvertis, tekste vadinamas energijos įverčiu. Nustačius, kad iš anksto parinktame gretimų kadrų skaičiuje įvertis yra didesnis už tam tikrą iš anksto parinktą slenkstį, pirmojo kadro pradžioje fiksuojama žodžio pradžia. Tada analogiškai nustatoma žodžio pabaiga, tik šiuo atveju tikrinama, ar įvertis yra mažesnis už pasirinktą slenkstį.

Nustačius žodžio ribas, apdorojamas kalbos signalas, reprezentuojantis žodį. Atpažinimo sistemoje realizuoti standartiniai kalbos signalo apdorojimo ir požymių išskyrimo metodai (Rabiner, 1989; Young ir kt., 2005).

Pirmiausia iš kalbos signalo pašalinama nuolatinė dedamoji – iš signalo atskaitų reikšmių atimamas jų vidurkis. Tada signalas dalijamas į 25 ms trukmės kadrus su 10 ms žingsniu tarp gretimų kadrų. Kiekvieno kadro signalas filtruojamas pirmos eilės ribotos impulsinės reakcijos filtru ir dauginamas iš Hammingo lango funkcijos. Sistemoje naudojamo filtro koeficientas – 0,97.

Vėliau atliekama kalbos signalo analizė – išskiriami akustiniai požymiai. Gaunama požymių vektorių seka, kitaip vadinama stebėjimų arba akustinių vektorių seka. Šioje atpažinimo sistemoje naudojami kalbos suvokimu, kai imituojamas žmogaus klausos aparato veikimas, grįsti požymiai – melų dažnių skalės kepstriniai koeficientai. Gauta seka pateikiama atpažinti ištartą žodį.

Sistemoje „Žodžių atpažintuvas“ realizuotas paslėptaisiais Markovo modeliais grįstas atskirų žodžių atpažinimas. Kad nustatytų, kurį ištartą žodį reprezentuoja atpažinimui pateikta požymių vektorių seka, sistema naudoja akustinių modelių aibę – paslėptuosius Markovo modelius.

Kai atliekamas pavienių žodžių atpažinimas, kiekvieną k-ąjį duoto žodyno žodį atitinka atskiras modelis M_k. Žodžių atpažinimas atliekamas pagal formulę:

$$j = \arg\max_k P(M_k \mid \mathbf{O}); \qquad (1)$$

čia M_k – k-ojo žodžio modelis, $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T\}$ – kalbos signalo požymių vektorių seka, j – atpažinto žodžio modelio, su kuriuo gauta didžiausia aposteriorinė tikimybė $P(M_k \mid \mathbf{O})$, indeksas.

Aposteriorinė tikimybė $P(M_k \mid \mathbf{O})$ skaičiuojama naudojant Bayeso formulę:

$$P(M_k \mid \mathbf{O}) = \frac{P(\mathbf{O} \mid M_k) P(M_k)}{P(\mathbf{O})}; \qquad (2)$$

čia $P(\mathbf{O} \mid M_k)$ – tikėtinumas, kad požymių vektorių seką $\mathbf{O}$ generavo modelis M_k, $P(M_k)$ – apriorinė modelio M_k tikimybė, $P(\mathbf{O})$ – požymių vektorių sekos $\mathbf{O}$ tikimybė. Vardiklyje esantis dydis $P(\mathbf{O})$ yra pastovus visiems modeliams ir neturi įtakos formulės (1) rezultatui, todėl į tolesnius skaičiavimus neįtraukiamas. Apriorinė tikimybė $P(M_k)$ vadinama *kalbos* modeliu, o tikėtinumas $P(\mathbf{O} \mid M_k)$ – *akustiniu* modeliu. Kadangi šioje sistemoje kalbos modelis nenaudojamas, t. y. sistemoje visų modelių apriorinės tikimybės vienodos, žodžio atpažinimas sistemoje priklauso tik nuo tikėtinumo $P(\mathbf{O} \mid M_k)$, kuris apskaičiuojamas naudojant Viterbi algoritmą.

Sistema skaičiuoja kiekvieno akustinio modelio tikėtinumo, kad gauta seka buvo generuota to modelio, įvertį ir iš žodyno pateikia kaip atpažintą tą žodį, kurį reprezentuoja modelis, turintis didžiausią tikėtinumo įvertį.

Kad atpažinimo sistema būtų pritaikyta naršyklei valdyti balsu, reikėjo sukurti žodyną – balso komandų sąrašą, ir akustinių modelių rinkinį.

Žodyną sudaro 55 naršyklės valdymo komandos, pavyzdžiui, komanda „baik darbą“ skirta išeiti iš programos, ir 16 konkrečių tinklalapių atvėrimo komandų. Komandą sudaro vienas arba du žodžiai.

Sistemoje naudojami tolydieji paslėptieji Markovo modeliai. Kiekvienam žodyno žodžiui sukurtas akustinio modelio prototipas, kuris buvo mokomas pagal skirtingus žodžių ištarimo pavyzdžius. Mokymo duomenis sudarė 8-ių kalbėtojų – 4 moterų ir 4 vyrų, amžius nuo 25 iki 56 m. – įrašai. Kiekvienas kalbėtojas ištarė visus žodžius po 10 kartų. Tad kiekvieno modelio mokymui buvo skirta po 80 pavyzdžių. Žodžiai buvo įrašyti naudojant vieną kanalą, 16 kHz signalo diskretizavimo dažnį, 16 bitų kvantavimo tikslumą. Modelių prototipuose nurodyti šie parametrai:

- požymių vektorius: 13 melų dažnių skalės kepstro koeficientų, 13 išvestinių pirmos ir 13 antros eilės koeficientų;

- viena Gauso mišinio komponentė;
- būsenų skaičius priklauso nuo žodyje esančių raidžių skaičiaus – kiekviena raidė modeliuojama trimis būsenomis;
- visos būsenų vidurkių vektorių reikšmės lygios 0, dispersijų – teigiamos;
- visų leistinų būsenų perėjimų tikimybių reikšmės lygios 0,5, išskyrus perėjimą iš pirmos į antrą būseną – lygi 1.

Modeliai buvo mokomi naudojant priemones iš HTK 3.3 priemonių rinkinio, sukurto Kembridžo universitete (Young ir kt., 2005). Mokymo procedūrą sudarė trys etapai: signalo apdorojimas ir požymių išskyrimas, modelio parametrų pradinių pagrįstų įverčių apskaičiavimas, parametrų įverčių tikslinimas naudojant Baumo ir Velčo (Baum–Welch) algoritmą (Young ir kt., 2005).

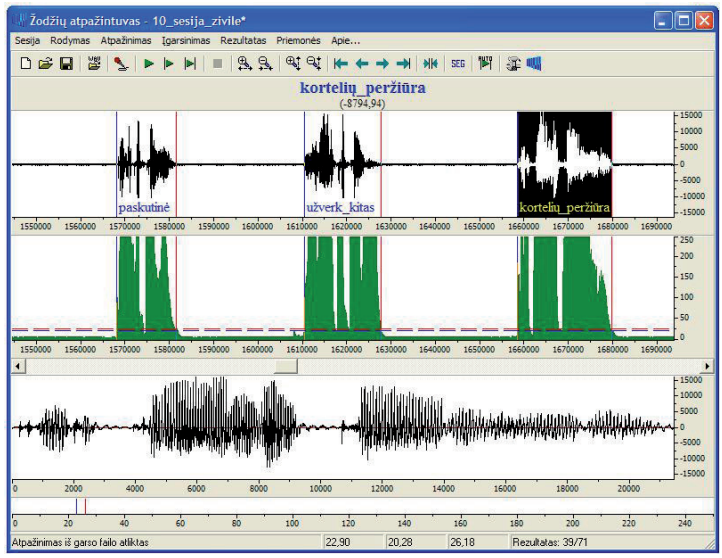
Sistemoje naudojami paprasčiausi visą žodį reprezentuojantys akustiniai modeliai. Reikia paminėti, kad atpažintuve galima naudoti fonetinių vienetų, kurie atpažinimo metu sujungiami į žodžius, modelius. Tada naršyklės valdymo sistemą galima patobulinti – suteikti vartotojui galimybę į sistemoje naudojamą žodyną įtraukti naujas jam tinkamesnes komandas ir pašalinti nereikalingas.

Naršyklei „Windows Internet Explorer 7“ valdyti pasirinktos komandos, kurias galima įvykdyti spustelėjus klaviatūros klavišą arba jų kombinaciją. Programoje naudojamos *Win32 API* funkcijos, kurios persiunčia aktyviam programos langui atitinkamus klavišų kodus. Kiekvienas žodyno žodis susietas su atitinkamu kodu arba konkretaus tinklalapio adresu. Kai atpažintuvas atpažįsta balso komandą, imituojamas reikiamo klavišo spustelėjimas, į kurį naršyklė atsako veiksmu, pavyzdžiui, kai atpažįstama komanda „žinynas“, atveriamas žinynas. Jei naršyklės langas aktyvus, ji vykdo atitinkamus veiksmus. Jei naršyklė uždaroma ir atpažintuvo langas tampa aktyvus, tada jis reaguoja atitinkamais veiksmais, pavyzdžiui, jei pasakoma komanda „baik darbą“, atpažintuvas baigs darbą. Tokiu būdu atpažinimo sistemos pagrindu galima sukurti programinę įrangą, kuri valdytų ir kitas kompiuterines programas. Jei atpažįstama komanda, kuri susieta su konkrečiu tinklalapiu, naudojama *Win32 API* funkcija, kuriai perduodamas tinklalapio adresas.

Naršyklės valdymo balsu sistemos prototipo sąsaja ir valdymas

Sistemos grafinę vartotojo sąsają sudaro keli langai. Pagrindiniame lange (2 pav.) yra meniu ir mygtukų juostos; žodžio laukas, kuriame rodomas atpažintas žodis ir jo atpažinimo tikimybės logaritmo įvertis; signalo sritis, kurioje matyti kalbos signalo bangos forma ir aptiktos žodžių ribos; signalo energijos sritis, kurioje vaizduojama kalbos signalo energija ir nustatytos žodžių ribos bei žodžio ribų aptikimo energijos slenksčiai; žodžio signalo sritis, kurioje vaizduojama signalo srityje pažymėto atpažinto žodžio signalo bangos forma; energijos indikatorius, kuriame rodoma įgarsinamo arba įvedamo iš mikrofono signalo energija bei ribų aptikimo energijos slenksčiai.

Nuostatų dialogo lange (3 pav.) naujos sesijos pradžioje vartotojas gali keisti programos



2 pav. ***Pagrindinis sistemos langas***

nuostatas. Energijos skaičiavimo kortelėje galima keisti signalo energijos įverčio skaičiavimo parametrus: signalo pradinio apdorojimo koeficientą, signalo kadro trukmę, kadro žingsnį, nurodyti, ar taikyti signalo kadrui Hammingo lango funkciją prieš skaičiuojant jo energijos įvertį.

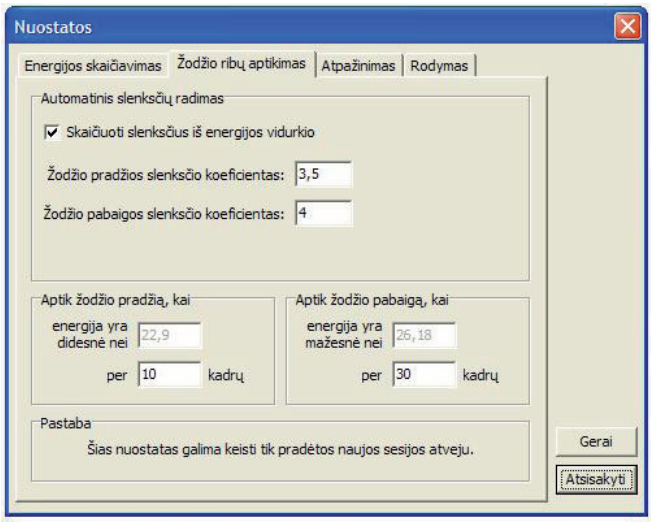
Žodžio ribų aptikimo nuostatų kortelėje (3 pav.) galima nurodyti žodžio pradžios ir pabaigos slenksčius arba nurodyti, kad slenksčiai būtų randami automatiškai. Tada slenksčių reikšmės apskaičiuojamos iš pirmųjų 10 signalo kadrų – energijos vidurkis dauginamas iš žodžio pradžios arba pabaigos slenksčio koeficiento. Žodžio pradžia fiksuojama tada, kai signalo energijos įvertis yra didesnis už žodžio pradžios slenkstį iš anksto parinktame gretimų kadrų skaičiuje. Analogiškai nustatoma žodžio pabaiga, tik šiuo atveju tikrinama, ar įvertis yra mažesnis už žodžio pabaigos slenkstį. Slenksčių reikšmės didinimo koeficientus ir kadrų skaičių vartotojas gali keisti.

Atpažinimo nuostatų kortelėje galima keisti modelio paieškos spindulį. Kuo mažesnis šis spindulys, tuo daugiau mažiausiai tikėtinų modelių analizuojant kiekvieną požymių vektorių pašalinama iš paieškos atpažinimo metu. Dėl to gali labai pagreitėti žodžio atpažinimas, tačiau pablogėti atpažinimo tikslumas.

Eksperimentinių tyrimų metu žodžių atpažinimas gali būti atliekamas įvedant žodžius iš failo arba mikrofono. Atpažinimas įjungiamas pasirinkus meniu komandą „Atpažinimas“ arba atitinkamą priemonių juostos mygtuką pagrindiniame lange (2 pav.). Kai kalbos signalas įvedamas iš mikrofono, pirmą kartą įjungus atpažintuvą turi būti tylu, nes programa apskaičiuoja naujus žodžio ribų aptikimo energijos slenksčius. Vykstant jų skaičiavimui žodžio lauke rodomas užrašas „Palaukite“. Kai slenksčiai nustatyti, pasirodo užrašas „Kalbėkite“ ir galima pradėti tarti žodžius į mikrofoną.

Naršyklės valdymo balsu langas (4 pav.) iškviečiamas pagrindinio lango priemonių juostoje spustelėjus atitinkamą mygtuką.

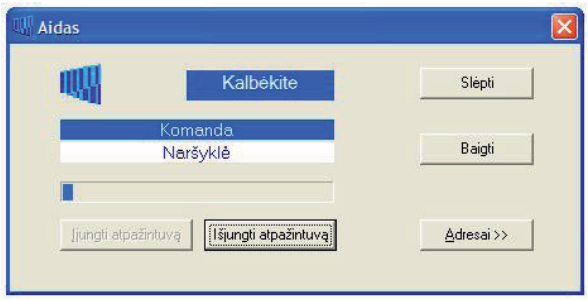
Valdymo lange yra atpažintuvo įjungimo ir išjungimo mygtukai. Spustelėjus mygtuką „Adresai“, langas padidinamas – atsiranda adresų keitimo laukas, kuriame vartotojas gali pakeisti tinklalapių, kurie turi būti atveriami ištarus atitinkamą komandą, adresus. Sistemos langus galima paslėpti. Tada naršyklės valdymo balsu sistema įjungiama arba išjungiama pasirinkus atitinkamą komandą iš sistemos piktogramos, esančios sisteminėje juostelėje, kontekstinio meniu.



3 pav. ***Žodžio ribų aptikimo nuostatų kortelė su numatytomis nuostatų reikšmėmis***

Sistemos darbingumo tyrimas

Eksperimentinio tyrimo metu buvo įvertintas sistemos darbingumas atpažįstant komandas iš įrašų, kurie nebuvo naudojami mokant modelius. Testavimo duomenų rinkinį sudarė 6 kalbėtojų – 3 vyrų ir 3 moterų, amžius nuo 17 iki 45 m. – įrašai. Kiekvienas kalbėtojas kiekvieną komandą ištarė po 10 kartų. Tad kiekvienam modeliui įvertinti buvo skirta 60 įrašų. Tyrimo rezultatai parodė, kad laboratorinėmis sąlygomis sistemos atpažinimo tikslumas siekia 98 procentus.



4 pav. ***Naršyklės valdymo langas***

Išvados

Atskirų žodžių atpažinimo sistemos, grįstos paslėptaisiais Markovo modeliais, pagrindu sukurtas nepriklausomo nuo kalbėtojo naršyklės valdymo balsu sistemos prototipas. Valdymo sistema gali atpažinti 71 komandą: 55 naršyklės valdymo komandas, 16 komandų, atveriančių konkrečius sistemoje nurodytus tinklalapius. Valdymas realizuotas kiekvieną žodyno žodį susiejant su konkrečiu klavišu arba jų kombinacija, arba konkretaus tinklalapio adresu. Kai atpažįstama balso komanda, imituojamas atitinkamo klavišo spustelėjimas, į kurį naršyklė atsako veiksmu. Analogiškai galima realizuoti ir kitų kompiuterinių programų valdymą balsu.

Sistemoje naudojami atskirų žodžių akustiniai modeliai. Tai riboja vartotojo galimybes – naršyklę galima valdyti tik numatytomis komandomis. Valdymo sistemą ketinama tobulinti – vartotojui suteikti galimybę į sistemoje naudojamą žodyną įvesti naujus žodžius. Tai galima realizuoti fonetinių vienetų akustiniais modeliais.

LITERATŪRA

TAMULEVIČIUS, G.; LIPEIKA, A. (2003). Žodžių atpažinimo sistemos kūrimas. *Lietuvos matematikos rinkinys*, 43 (spec. nr.), p. 292–296.

TAMULEVIČIUS, G. (2007). Interneto naršyklės valdymas balsu. Iš *Informacinės technologijos 2007. Konferencijos pranešimų medžiaga* [žiūrėta 2009 04 13]. Kaunas: Technologija, p. 67–70. Prieiga per internetą: http://www.ktu.lt/lt/apie_renginius/konferencijos/2007/k7_01/IT-2007/it%202007-II.pdf.

RABINER, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, vol. 77, no. 2, 257–286. Prieiga per internetą: http://www.cs.ubc.ca/~murphyk/Bayes/rabiner.pdf [žiūrėta 2009 04 13].

TAMULEVIČIUS, G.; LIPEIKA, A. (2004). Dynamic time warping based speech recognition system. *Human language technologies. The Baltic perspective*, p. 156–161.

YOUNG, S.; EVERMANN, G.; GALES, M.; HAIN, T.; KERSHAW, D.; MOORE, G.; ODELL, J.; OLLASON, D.; POVEY, D.; VALTCHEV, V.; WOODLAND, P. (2005). *The HTK Book* (for HTK Version 3.3) [žiūrėta 2009 04 13]. Cambridge University Engineering Department. Prieiga per internetą: http://htk.eng.cam.ac.uk/docs/docs.shtml.

SPEAKER-INDEPENDENT WEB BROWSER CONTROL BY VOICE

Živilė Ringelienė

Summary

The paper presents a Lithuanian prototype of the system for web browser control by voice. The program, which implements control by voice commands, is integrated in the hidden Markov models based isolated word recognition system. Every command is linked with some key combination or internet address. When the word is recognized, the appropriate key combination is simulated and the browser acts according to the command, for example, goes to the previous or next page, opens a new tab, etc. The engine of the prototype is a speaker-independent Lithuanian word recognition system and it can recognize 71 voice commands: 55 commands – for browser control, and 16 commands – to open various user predefined websites.