

LIETUVIŲ KALBOS GARSŲ TRUKMĖS MODELIAVIMAS KLASIFIKAVIMO IR REGRESIJOS MEDŽIAIS, NAUDOJANT DIDELĖS APIMTIES GARSYNĄ

Giedrius Norkevičius, Gailius Raškinis

Vytauto Didžiojo universitetas, K. Donelaičio g. 58, 3000 Kaunas

Šio tyrimo tikslas - sukurti modelį gebantį prognozuoti lietuvių kalbos garsų trukmes pagal kontekstinę informaciją. Tyrimui naudotas, garsą bei jo kontekstą aprašantis, 15-os požymių rinkinys, svarbiausi jų: prognozuojamo garso identifikatorius, gretimų garsų identifikatoriai, garsų skaičius skiemenyje. Darbe aprašomas eksperimentas, kurio metu lietuvių kalbos garsų trukmės buvo prognozuojamos naudojant klasifikavimo ir regresijos medžius. Pateikiami eksperimentų rezultatai keičiant šiuos modelio parametrus: medžio mazgų dalinimo kriterijų - kvadratinės arba santykinės klaidos, minimalų leidžiamą mazgo dydį - 1, 10, 20 ar 30 pavyzdžių. Taip pat pateikiami medžio genėjimo minimalios klaidos - sudėtingumo algoritmo pagalba gauti rezultatai. Tyrimo duomenų bazę sudaro 300 tūkst. balsių ir 400 tūkst. priebalsių pavyzdžių paimtų iš VDU-AB20 garsyno. Sudarytasis modelis leidžia prognozuoti lietuvių kalbos garsų trukmes su ~25% santykinę klaida.

1 Įvadas

Garsų trukmės modelis, o tuo pačiu ir žinios kokie faktoriai ar jų sąveikos ir koku būdu įtakoja garsų trukmę yra aktualūs tokiems kalbos technologijos darbams kaip: šnekos sintezė, automatinis šnekos atpažinimas, automatinis šnekos signalo anotavimas. Kalboje garsų trukmės yra priklausomos nuo konteksto, pvz.: garso /a/ (naudojamas SAMPA žymėjimas [1]) žodyje „savo“ trukmė yra 30ms, o žodyje „avertė“ 115ms. Šio tyrimo tikslas – sukurti modelį gebantį prognozuoti lietuvių kalbos garsų trukmes pagal kontekstinę informaciją.

Garsų trukmė modeliuojama įvairiais metodais: sudarant garsų trukmes nusakančias taisykles [2], naudojant sandaugų sumų modelį [3,4], taikant sprendimo medžius[5, 6].

Populiariausias taisyklinis metodas yra D. Klatt [2] sudarytas modelis apibendrinamas formule:

$$DUR = ((INHDUR - MINDUR) * PRCNT) / 100 + MINDUR \quad (1)$$

kur $INHDUR$ yra segmento savaiminė trukmė ms, $MINDUR$ segmento minimali trukmė ir $PRCNT$ trukmės padidėjimas/sumažėjimas procentais. Kiekvienam garsui ar jų grupei yra nustatoma savaiminė ir minimali trukmė bei apibrėžiamos taisyklės (nuo konteksto priklausančios $PRCNT$ parametro reikšmės). Kiekviena taisyklė nusako kiekybės mažėjimą ar didėjimą procentais, o klasifikuotini segmentai negali būti trumpesni nei tam tikras minimumas. Modeliuoti pradeda nuo savaiminės garsų trukmės ir pridamas kiekvieną garso požymį atitinkantis trukmės santykis. Nors šio tipo modelis buvo gana sėkmingai pritaikytas švedų, anglų, prancūzų kalboms, pagrindinis metodo trūkumas yra per didelis apibendrinimas, bei staigus taisyklių augimas siekiant apdoroti išimtis [6].

Sandaugų sumų modelis, sukurtas Van Santeno [3, 4], apibendrinamas formule:

$$DUR(d) = \sum_{i \in K} \prod_{j \in I_i} S_{i,j}(d_j) \quad (2)$$

čia d yra parametrų vektorius, nusakantis prognozuojamą segmentą, K – indeksų, atitinkančių kiekvieną sandaugą, aibė, I_i – aibė parametrų, įeinančių į i -tąją sandaugą. Parametrai $S_{i,j}$ yra vadinami parametrų svoriais (*factor scales*).

Modeliavimas šiuo metodu vyksta trimis etapais:

- pagal jau žinomus garsų kiekybės santykius kalbininkai sudaro kategorijų medžius – vienas medžio lapas atspindi garsų grupę, kuriai turi įtakos tam tikri faktoriai/parametrai ar jų sąveikos,
- kiekvienam lapui/kategorijai sudaromas atskiras pavidalo (2) modelis,
- skaičiuojami modelių parametrai.

Daugelio nurodoma, kad tai vienas patikimiausių metodų, t.y. geriausiai prognozuojantis bei didžiausia koreliacija tarp prognozuojamos ir tikrosios reikšmės pasižymintis modelis. Tačiau, norint sudaryti kategorijų medį, bei priskirti šioms kategorijoms modelius reikia išsamių žinių apie faktorius ar jų sąveikas, kurie įtakoja garso trukmę. Kadangi nėra išsamių lietuvių kalbos garsų trukmių tyrimų šis metodas kol kas nėra tinkamas.

Šiuo metu lietuvių kalbos garsų trukmei modeliuoti galima pritaikyti vieną iš mašininio mokymo metodų – sprendimo medžio metodą. Sprendimo medžių vienas iš variantų – klasifikavimo ir regresijos medžiai. Garsų trukmės tyrimuose [3, 4, 8] ir bendrai literatūroje [7] galima sutikti nuorodų, jog naudojant šį modeliavimo metodą nėra gaunami tiksliausi rezultatai. Tačiau, nepaisant to, dėl savo gebėjimo gerai atskleisti duomenyse glūdinčią parametrų bei jas atitinkančios reikšmės sąryšio struktūrą, šis metodas labai dažnai

naudojamas pradinuose tyrimo etapuose. Šiame straipsnyje pristatomas garsų trukmės modeliavimo klasifikavimo ir regresijos medžiais tyrimas.

2 Modeliavimas klasifikavimo ir regresijos medžiais

Klasifikavimo ir regresijos medžiai – yra statistinio modeliavimo metodas, naudojamas prognozuoti kintamojo y reikšmei, atitinkančiai parametų vektorių X . Kaip ir kiekvienam mašininio mokymo algoritmui, taip ir regresijos medžiui reikalinga (X_n, y_n) pavidalo mokymo imtis L , kur y_n – nuo parametų vektoriaus X_n priklausomo objekto reikšmė, $n = 1, 2, \dots, N$, kur N – pavyzdžių skaičius. Modeliavimas susideda iš trijų etapų:

1. medžio konstravimo,
2. medžio genėjimo,
3. optimalaus medžio parinkimo.

Trumpai apibudinsime kiekvieną iš etapų.

2.1 Medžio konstravimas

Iš pradžių medis susideda iš vieno, vadinamojo šakninio, mazgo t_1 , kuriį sudaro visi aibės L mokymo pavyzdžiai. Užduotis yra surasti optimalų aibės L padalinimą į dvi dalis t_L, t_R . Su gautaisiais mazgais t_L, t_R kartojama tokia pati dalinimo į dvi dalis (visur ateityje skaitysime, kad regresijos medis yra binarinis) procedūra kaip ir su šakniniu mazgu. Toks iteracinis procesas vykdomas tol, kol pasiekiamas nutraukimo sąlyga (paprastai dalinama tol kol stebimas klaidos mažėjimas arba mazgo dydis didesnis už iš anksto apibrėžtą).

Optimalus aibės L padalinimas priklauso nuo to kokį pasirinksiame medžio daromos klaidos vertinimo būdą. Šiame tyrime medžio prognozės klaidai vertinti naudojome:

1. vidutinės kvadratinės klaidos kriterijų:

$$R_{KK}(T) = \frac{1}{N} \sum_n (y_n - y(t))^2 \quad (3)$$

2. vidutinės santykinės klaidos kriterijų:

$$R_{SK}(T) = \frac{1}{N} \sum_n \frac{|y_n - y(t)|}{y_n} \quad (4)$$

čia $R_{KK}(T)$ - medžio T daroma klaida (indeksas parodo kuris kriterijus taikomas), $y(t)$ - mazgo t prognozuojama reikšmė, aišku turi būti tenkinama sąlyga $X_n \in t$.

Pastebėsime, kad pirmuoju - vidutinės kvadratinės klaidos atveju $y(t)$ reikšmė, kuri minimizuoja $R_{KK}(T)$, yra visų y_n reikšmių, papuolančių į mazgą t , vidurkis, t.y:

$$\bar{y}(t) = \frac{1}{N(t)} \sum_{X_n \in t} y_n, \quad (5)$$

čia sumuojami visi y_n tenkinantys sąlygą: $X_n \in t$, o $N(t)$ yra mazge t esančių pavyzdžių skaičius. Kitaip, geriausia prognozė pirmuoju atveju – vidurkis. Sakykime, kad mazgo t visų galimų padalinių aibė yra S . Tada geriausiu mazgo t padalinimu vidutinės kvadratinės klaidos atveju vadinsime padalinimą $s^* \in S$:

$$\min_{s \in S} \left(\sum_{X_n \in t_L} (y_n - \bar{y}(t_L))^2 + \sum_{X_n \in t_R} (y_n - \bar{y}(t_R))^2 \right) \quad (6)$$

Antruoju, vidutinės santykinės klaidos atveju $y(t)$ reikšmė kuri minimizuoja $R_{SK}(T)$ yra visų y_n reikšmių, papuolančių į mazgą t , mediana, ją žymėsime $v(t)$. Tada geriausiu mazgo t padalinimu vidutinės santykinės klaidos atveju vadinsime padalinimą $s^* \in S$:

$$\min_{s \in S} \left(\sum_{X_n \in t_L} (y_n - v(t_L))^2 + \sum_{X_n \in t_R} (y_n - v(t_R))^2 \right) \quad (7)$$

Realaus tipo parametrams nustatomi visi $x^i < \tau$ padalinimai. Išvardijamojo tipo parametų padalinimo pavidalas yra: $x^i \in \Theta$, kur Θ gali būti bet koks aibės, sudarytos iš i-tojo požymio reikšmių, poaibis.

2.2 Medžio genėjimas

Kadangi medžio auginimui naudojama mokymo imtis, užauginto medžio T_{\max} įvertis klaidos įvertis $R(T_{\max})$ yra per daug optimistinis, t.y. šis įvertis neatspindės realios medžio daromos klaidos naudojant testavimo imtį. Šiame tyrime medžio genėjimui naudojamas minimalios klaidos, sudėtingumo genėjimo algoritmas (minimal cost complexity pruning) [7]. Šio algoritmo esmė yra iteracinis procesas, kurio metu nukertamos šakos turinčios mažiausią santykį:

$$\frac{R(t) - R(T_t)}{|\tilde{T}_t| - 1}, \quad (8)$$

čia $R(t)$ - mazgo t (šiuo atveju tai nėra lapas) daroma klaida, $R(T_t)$ medžio T_{\max} šakos, prasidedančios mazgu t , daroma klaida, $|\tilde{T}_t|$ - šakos T_t lapų skaičius. Tokiu būdu yra sudaroma vis mažesni mazgų skaičių turinti medžių seka $T_{\max} \supseteq \dots \supseteq T_k \supseteq \dots \supseteq T_K = t_1$. Šios sekos sudarymui naudojome mokymo imtį.

Taip pat tyrime naudojamas taip vadinamas išankstinis genėjimas. Šio genėjimo esmė – neleisti medžio konstravimo etape kurti mazgų kurie apima mažiau nei iš anksto užsibrėžtas skaičius pavyzdžių. Kitaip sakant mazgas yra kuriamas tik tada, kai yra pakankamas kiekis į jį papuolančių pavyzdžių.

2.3 Optimalaus medžio parinkimas

Šiame tyrime iš medžių sekos $T_{\max} \supseteq \dots \supseteq T_k \supseteq \dots \supseteq T_K = t_1$ išrinkti optimalų medį naudojome validavimo imtį. T.y. buvo išrenkamas mažiausia prognozavimo klaida pasižymintis medis naudojant validavimo duomenimis.

3 Eksperimentinis garsų trukmės modelio tyrimas

3.1 Eksperimento duomenys

Tyrimams buvo naudotas VDU-AB20 garsynas, sudarytas Vytauto Didžiojo universiteto Kompiuterinės Lingvistikos centre. Šis garsyną sudaro 20 diktorių skaitytinės rišlios šnekos įrašai. Kiekvieno diktoriaus šneka trunka apie 1 valandą, o bendra garsyno apimtis – 20 valandų. VDU-AB20 garsynas yra automatiškai anotuotas garsų, skiemenų, žodžių, jų junginių, frazių ir sakinių lygmenyse. Garsų lygmuo yra svarbiausias šiam tyrimui. Iš viso VDU-AB20 garsyne yra apie 300 tūkst. balsių ir 400 tūkst. priebalsių, o bazinį skirtingų garsų rinkinį sudaro apie 200 fonetinės transkripcijos ženklų. Jie atsižvelgia į garso kirtį, minkštumą, dalyvavimą dvigarsyje ir kt. savybes. Detalus bazinio garsų rinkinio aprašymas pateiktas [1], o procedūra, kuria ortografinis tekstas paverčiamas fonetine transkripcija aprašyta [9]. Anotavimui, t.y. pradžios ir pabaigos laiko momentų arba trukmės priskyrimui fonetinės transkripcijos ženklui, buvo naudota tikimybinė anotavimo procedūra, grindžiama Paslėptųjų Markovo modelių (PMM) metodika ir realizuota panaudojant HTK programų paketą [10]. Automatiniam garsų ribų nustatymui buvo naudota diskretinė gardelė, kurios žingsnis 10 ms.

3.2 Požymiai naudojami trukmei prognozuoti

Garso trukmės prognozavimui naudojami požymiai gali būti suskirstyti į lygmenis pagal nagrinėjamo garso konteksto plėtėjimą:

Garso lygmuo – šiame lygmenyje visi požymiai yra kategoriniai, reikšmės - SAMPA [1] kodai

- garso pavadinimas
- dviejų gretimų garsų iš kairės ir iš dešinės pavadinimai. Klaustuko simboliu buvo žymimas tolimesnis konteksto garsas jeigu artimesnis buvo tyla arba trumpa pauzė.

Skiemens lygmuo – šiame lygmenyje visi požymiai skaitiniai, mato vienetas - garsų skaičius

- skiemens ilgis
- nagrinėjamo garso atstumas nuo skiemens pradžios
- nagrinėjamo garso atstumas iki skiemens pabaigos

Žodžio lygmuo – šiame lygmenyje visi požymiai skaitiniai, mato vienetas - skiemenų skaičius

- žodžio ilgis
- nagrinėjamo skiemens atstumas nuo žodžio pradžios
- nagrinėjamo skiemens atstumas iki žodžio pabaigos

Frazės lygmuo – šiame lygmenyje visi požymiai skaitiniai, mato vienetas - žodžių skaičius

- frazės ilgis

- nagrinėjamo žodžio atstumas nuo frazės pradžios
 - nagrinėjamo žodžio atstumas iki frazės pabaigos
- Papildomi požymiai – loginio tipo kintamasis, reikšmės: TAIP/NE
- garsas pirmas žodyje
 - garsas paskutinis žodyje

3.3 Modelio vertinimo kriterijai

Garsų trukmės modelių vertinimo kriterijai paprastai yra:

- Modelio prognozavimo klaida. Kaip minėta skyriuje 2.1 nuo šio kriterijaus vertinimo būdo tiesiogiai priklauso medžio konstravimas, o tuo pačiu ir geriausios prognozės lapo viduje parinkimas. Literatūroje dažniausiai šis kriterijus vertinamas šaknimi iš vidutinės kvadratinės klaidos (toliau ŠVKK), šiuo atveju geriausia prognozė lapo viduje - vidurkis. Kadangi, neteko aptikti šio kriterijaus naudojimo pagrįstumo, išbandėme prognozavimo klaidą vertinti vidutine santykine klaida ir atitinkamai trukmę lapo viduje prognozuoti mediana. Nepaisant to, kad medžio konstravimo etape yra naudojamas pasirinktinai vienas iš įvardintų metodų (toliau juos vadinsime tiesiog prognozavimu naudojant vidurkį ir prognozavimu naudojant medianą) rezultatuose pateiksime abu jau sukonstruoto medžio prognozavimo klaidos įverčius.
- Koreliacija tarp prognozuotos ir tikrosios trukmių.

3.4 Eksperimentų eiga

Visuose eksperimentuose buvo naudota kryžminio patikrinimo metodika (cross validation), mokymui buvo skirta 90% visų duomenų, o validavimui ir testavimui po 5%. Reikia pažymėti, kad garsyno anotacijos, gautos automatinio būdu, nėra validuotos ekspertų. Validuoti tik vieno diktoriaus įrašai ir visų diktorių kirčiavimas. Dėl šios priežasties eksperimentai buvo atliekami su dviem duomenų rinkiniais: vieno diktoriaus duomenų rinkinys, toliau vadinsime eksperimentu „VD“ ir su visų 20-ties diktorių duomenų rinkinys, toliau vadinsime eksperimentu „DD“. „VD“ duomenų rinkinį sudarė 26940 balsių ir 34281 priebalsių pavyzdžių, atitinkamai „DD“ duomenų rinkinį sudarė 314114 ir 409927 pavyzdžių. Kadangi balsiai ir priebalsiai yra artikuliaciniu, akustiniu ir funkciniu požiūriu dvi visiškai skirtingos garsų klasės, buvo sudaromi du medžiai: atskirai balsiams ir priebalsiams. Eksperimentuose buvo tiriami tokie modelio parametrai:

- Išankstinio genėjimo parametras, bandomos šios minimalaus mazgo dydžio reikšmės: 1 (išankstinis genėjimas nenaudojamas), 10, 20, 30
- Prognozavimas naudojant mediana arba vidurkį

Naudojant skirtingus modelio parametrus, iš viso tyrimo metu buvo atlikti 32 eksperimentai. Tyrimo eigoje taip pat buvo nagrinėjama genėjimo įtaką modelio vertinimo kriterijams (ž.r. skyrių 3.3).

3.5 Eksperimentų rezultatai

Tiriant išankstinio genėjimo parametro daromą įtaką modelio vertinimo kriterijams, paaiškėjo, kad geriausi rezultatai pasiekiami neleidžiant kurti mazgų apimančių mažiau nei 10 pavyzdžių. Genėjimas tokiu būdu sukonstruoto medžio modelio vertinamų kriterijų apčiuopiamai neįtakojo, tačiau tuo pačiu reikia pažymėti, kad medžio mazgų skaičius (4 lentelė) buvo sumažinamas ~12 kartų. Dėl šios priežasties visi čia pateikiami rezultatai yra po genėjimo bei naudojant išankstinio genėjimo parametą 10. Kaip ir galima buvo tikėtis, visuose eksperimentuose įverčio ŠVKK prasme geriausi rezultatai buvo pasiekti prognozavimui naudojant vidurkį (1 lentelė), o įverčio VSK prasme - naudojant medianą (2 lentelė). Didžiausia koreliacija tarp tikros ir prognozuotos trukmės pasižymėjo modeliai sukurti prognozavimui naudojant vidurkį (3 lentelė).

1 lentelė. Vidutinės įverčio ŠVKK reikšmės po genėjimo, prognozuojant vidurkiu ir mediana. Išankstinio genėjimo parametras – 10.

ŠVKK				
	VD, Balsiai	VD, Priebalsiai	DD, Balsiai	DD, Priebalsiai
Prognozuojant vidurkiu	0.0314	0.0257	0.0367	0.0315
Prognozuojant mediana	0.0326	0.027	0.038	0.0333

2 lentelė. Vidutinės įverčio VSK reikšmės po genėjimo, prognozuojant vidurkiu ir mediana. Išankstinio genėjimo parametras – 10.

VSK				
	VD, Balsiai	VD, Priebalsiai	DD, Balsiai	DD, Priebalsiai
Prognozuojant vidurkiu	0.3441	0.3243	0.3131	0.2772
Prognozuojant mediana	0.2904	0.2957	0.2587	0.2478

3 lentelė. Vidutinės koreliacijos reikšmės, tarp tikrosios ir prognozuotos trukmių, po genėjimo, prognozuojant vidurkiu ir mediana. Išankstinio genėjimo parametras – 10.

Koreliacija				
	VD, Balsiai	VD, Priebalsiai	DD, Balsiai	DD, Priebalsiai
Prognozuojant vidurkiu	0.7298	0.7087	0.6508	0.5968
Prognozuojant mediana	0.7171	0.6938	0.6471	0.5698

4 lentelė. Medžio mazgų skaičius, prognozuojant vidurkiu ir mediana. Išankstinio genėjimo parametras – 10.

Mazgų skaičius				
	VD, Balsiai	VD, Priebalsiai	DD, Balsiai	DD, Priebalsiai
Prognozuojant vidurkiu	322.2	573.1	1253.2	1647.1
Prognozuojant mediana	445.4	249.5	579.3	706.6

4 Išvados

Iš šio tyrimo rezultatų lieka neaišku koks prognozavimo būdas – naudojant vidurkį ar medianą yra pranašesnis. Į šį klausimą galima atsakyti tik taikomojo pobūdžio tyrimu, pvz.: šnekos sintezėje tai galėtų būti natūralumo ar suprantamumo testai, šnekos atpažinime teisingai atpažįstamų žodžių kiekis ir pan.

Preliminari modelio daromų klaidų analizė parodė, kad garso trukmės skiriasi net ir esant visiems 15-ai prognozavimui skirtų parametru identiškiems. Pagrindinė to priežastis – skirtingas kalbėjimo tempas. Todėl norint tiksliau prognozuoti trukmes būtina atsižvelgti į tempo pokyčius.

Literatūros sąrašas

1. **Raškinis A., G. Raškinis, A. Kazlauskienė.** SAMPA (Speech Assessment Methods Phonetic Alphabet) for Encoding Transcriptions of Lithuanian Speech Corpora. *Information technology and control*. **Kaunas: Technologija**, 2003, No. 4(29), p. 52–55.
2. **D H Klatt**, Synthesis by rule of Segmental Durations in English Sentences, in *Frontiers of Speech Communication Research* edited by Lindblom & Ohman, Academic Press 1979 (pp 287-299)
3. **Jan P. H. van Santen**, Prosodic modeling in Text-To-Speech Synthesis, Lucent Technologies – Bell Labs, 600 Mountain Ave., Murray Hill, NJ 07974, U.S.A.
4. **Jan P. H. van Santen**, Quantitative modeling of segmental duration, Bell Labs, 600 Mountain Ave., Murray Hill, NJ 07974, U.S.A.
5. **Robert Batušek**, A Duration Model for Czech Text-To-Speech Synthesis, Laboratory of Speech and Dialogue, Faculty of Informatics, Masaryk University, Brno, Czech Republic
6. **Sridhar Krishna & Hema A. Murthy**, Duration modeling of Indian languages Hindi and Telugu, Indian Institute of Technology, Madras, Chennai – 60003
7. **L. Breiman, J. Friedman, R. Olshen, and C. Stone.** Classification and Regression Trees. Wadsworth and Brooks, 1984.
8. **Olga Gaubanova**, Predicting segment duration using sums of products model, Centre for Speech Technology Research, University of Edinburgh
9. **Norkevičius G., Raškinis G., Kazlauskienė A.**, Knowledge-based grapheme-to-phoneme conversion of Lithuanian words. In Proceedings of the 10th International Conference on Speech and Computer - Specom. Patras, Greece, 2005
10. **S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland**, "The HTK Book", *Entropic, Cambridge*, 2000.

Annotation

Modeling phone duration of Lithuanian by classification and regression trees, using very large data set

The goal of this research is building a model capable of predicting phonemes duration of Lithuanian. Set of 15 parameters characterizing phoneme and its context were selected for duration prediction. Data set consisting of 300 thousand vowels and 400 thousand consonants was used in this research. The influence of and minimal cost complexity pruning and different values of pre pruning are investigated. Models were built using two different data sets: one speaker and 20 speakers. Also prediction by average leaf duration vs. prediction by and median leaf duration are compared. The overall performance of ~25% average relative error was obtained.