INFORMACINĖS TECHNOLOGIJOS IR KALBA

Žodžių atpažinimo, grįsto paslėptaisiais Markovo modeliais, vizualizavimo ir analizės programinė įranga

Živilė Ringelienė

Vilniaus universiteto Matematikos ir informatikos instituto doktorantė
Vilnius University Institute of Mathematics and Informatics, PhD student
Akademijos g. 4–217, LT-08663 Vilnius
El. paštas: zivile.ringeliene@gmail.com

Mark Filipovič

Informatikos ir ryšių departamento prie Lietuvos Respublikos vidaus reikalų ministerijos vyriausiasis specialistas, daktaras
Information Technology and Communications Department under the Ministry of the Interior of the Republic of Lithuania, specialist, PhD
Šventaragio g. 2–248, LT-01510 Vilnius
El. paštas: mark.filipovic@gmail.com

Straipsnyje aprašomas atpažinimo, grįsto paslėptaisiais Markovo modeliais, sistemos prototipo veikimas. Ši programinė įranga skirta lietuvių kalbos žodžių atpažinimui tirti. Nagrinėjama, kaip sistemos pateikiama informacija apie žodžių atpažinimo procesą ir rezultatus padeda analizuoti klaidų priežastis. Žodžio atpažinimas priklauso nuo žodžio ribų nustatymo tikslumo. Signalo, energijos, žodžio ribų vizualizavimas leidžia lengviau įvertinti, ar sistema teisingai nustatė ribas. Jei žodis atpažintas klaidingai dėl to, kad buvo blogai nustatytos ribos, galima keisti sistemos parametrų, darančių įtaką ribų nustatymo tikslumui, reikšmes. Tam tikrais atvejais tai pagerina atpažinimo rezultatus. Žodžio paieškos vaizdavimas padeda įvertinti kiekvieno fonemos modelio įtaką žodžio atpažinimui ir parinkti žodžių transkripcijas, kurios pagerina atpažinimo rezultatus.

Įvadas

Aštuntąjį praėjusio amžiaus dešimtmetį kalbai atpažinti buvo pradėtas taikyti paslėptųjų Markovo modelių metodas. 1975 m. Bakeris aprašė kalbos atpažinimo sistemą *Dragon*, grįstą paslėptaisiais Markovo modeliais, o 1976 m. Jelinekas paslėptųjų Markovo modelių metodą pristatė kaip veiksmingą būdą kalbai atpažinti (Baker, 1975; Jelinek, 1976). Savo darbuose Bakeris ir Jelinekas rėmėsi Baumo ir jo kolegų klasikiniais darbais, kuriuose pateikiami modelių parametrų įvertinimo metodai (Baum ir kt., 1967; 1970). Devintąjį dešimtmetį paslėptieji Markovo modeliai tapo vienu iš dažniausiai naudojamų kalbos signalo modeliavimo metodų (Huang ir kt., 2001). 1989 m. paskelbtame straipsnyje Rabineris apžvelgė teorinius paslėptųjų Markovo modelių aspektus ir aprašė šių modelių taikymo galimybes kuriant realias kalbos atpažinimo sistemas (Rabiner, 1989). Pastaraisiais dešimtmečiais paslėptieji Markovo modeliai – dažniausiai naudojamas metodas kuriant eksperimentines ir komercines automatinio kalbos atpažinimo sistemas.

Matematikos ir informatikos institute buvo sukurtas atskirų lietuvių kalbos žodžių atpažinimo sistemos prototipas „Žodžių atpažintuvas“, kuriame įgyvendintas paslėptaisiais Markovo modeliais grįstas atpažinimo metodas. Sistema

skirta lietuvių kalbos žodžių atpažinimo eksperimentiniams tyrimams atlikti, kompiuterio valdymo balso komandomis programinei įrangai kurti. Jos pagrindu buvo sukurtas interneto naršyklės valdymo balsu sistemos prototipas (Ringelienė, 2009). Atpažinimo sistemoje naudojami kalbos suvokimu, kai imituojamas žmogaus klausos aparato veikimas, grįsti požymiai – melų dažnių skalės kepstriniai koeficientai (Davis, Mermelstein, 1980). Šiuo metu tai dažniausiai naudojami požymiai kalbai atpažinti. Sistemoje galima naudoti skirtingus žodynus, visą žodį arba fonetinius vienetus reprezentuojančius modelių rinkinius. Sistema atlieka nepriklausomą nuo kalbėtojo atpažinimą. Modeliams mokyti naudojamos *HTK* (angl. *The Hidden Markov Model Toolkit*) priemonės (Young ir kt., 2005).

HTK – Kembridžo (Cambridge) universitete sukurtas programinių priemonių rinkinys, plačiai naudojamas paslėptaisiais Markovo modeliais grįsto kalbos atpažinimo tyrimams. Rinkinį sudaro priemonės, skirtos duomenims rengti, akustiniams ir kalbos modeliams mokyti ir testuoti, akustiniams modeliams pritaikyti konkrečiam kalbėtojui, rezultatams analizuoti. Naudojantis šia programine įranga galima modeliuoti pavienių žodžių ir ištisinės kalbos atpažinimo sistemas, pasirinkti skirtingus paslėptųjų Markovo modelių parametrus ir topologijas, signalą reprezentuojančių požymių tipą. *HTK* priemonėms valdyti naudojama tekstinė sąsaja. *HTK* priemonėmis gauti kalbos atpažinimo rezultatai pateikiami tik skaitine forma. Todėl analizuojant atpažinimo klaidas nėra paprasta išsiaiškinti jų priežastis. Atpažinimo eigos ir rezultatų vizualizavimas leistų greičiau nustatyti klaidų priežastis ir numatyti kalbos atpažinimo tikslumo didinimo kryptis.

Kuriant žodžių atpažintuvą daugiausia dėmesio skirta jo valdymo paprastumui ir vaizdžiam atpažinimo rezultatų pateikimui. Naudotojui pateikiama tekstinė, skaitinė, grafinė, garsinė informacija apie žodžių atpažinimo eigą ir rezultatus. Remiantis šia informacija galima lengviau ir greičiau atlikti klaidų analizę ir numatyti žodžių atpažinimo tikslumo gerinimo kryptis.

Šio straipsnio tikslas – aprašyti sukurtos programinės įrangos „Žodžių atpažintuvas“ veikimą ir parodyti, kaip jos pateikiamą informaciją galima naudoti žodžių atpažinimo tikslumui gerinti.

Atpažinimo sistemos struktūra

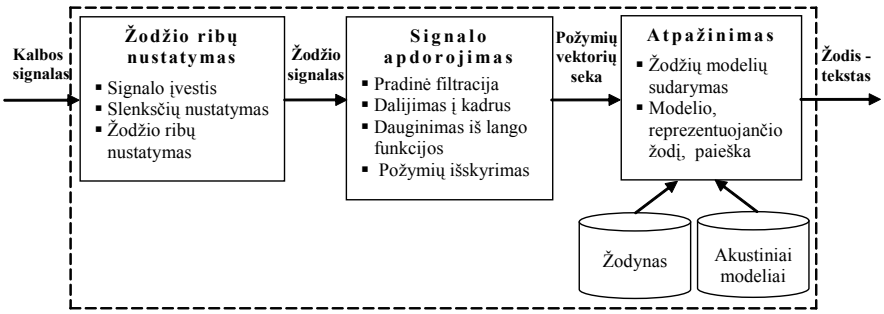
Atpažinimo sistemos struktūra vaizduojama 1 paveiksle. Žodžių atpažinimas gali būti atliekamas įvedant žodžius iš failo arba mikrofono. Tarp ištartų žodžių turi būti ~1 s pauzės. Sistema gali apdoroti kalbos signalą, kurio diskretizavimo dažnis – 11025 Hz, kvantavimo tikslumas – 16 bitų.

Žodžių ribos nustatomos automatiškai naudojant signalo energijos slenksčių metodą. Įjungus atpažintuvą, prieš įvedant kalbos signalą, skaičiuojamos žodžio pradžios sl_{pr} ir pabaigos sl_{pab} slenksčių reikšmės pasitelkiant šias formules:

$$sl_{pr} = a_{pr}\overline{E}\,, \tag{1}$$

$$sl_{pab} = a_{pab}\overline{E}\,, \tag{2}$$

čia a_{pr}, a_{pab} – žodžio pradžios ir pabaigos slenksčių reikšmių koeficientai. Signalo intensyvumo



1 p a v. ***Žodžių atpažinimo sistemos struktūra***

įverčio vidurkis $\overline{E}$ skaičiuojamas iš pirmųjų 10 signalo kadrų:

$$\overline{E} = \frac{1}{10} \sum_{i=1}^{10} E_i \ . \quad (3)$$

Kadro signalo intensyvumo įvertis E_i, tekste vadinamas energijos įverčiu, skaičiuojamas formule

$$E_i = \frac{1}{N} \sum_{j=1}^{N} | s_j | , \quad (4)$$

čia s_j – signalo *i*-ojo kadro *j*-osios atskaitos reikšmė, N – signalo atskaitų skaičius kadre.

Kalbos signalas, patekęs į atpažinimo sistemą, dalijamas į 10–30 ms trukmės kadrus su 5–15 ms žingsniu tarp gretimų kadrų. Tada skaičiuojamas kiekvieno kadro signalo energijos įvertis. Nustačius, kad iš anksto parinktame gretimų kadrų skaičiuje M įvertis E_i yra didesnis už žodžio pradžios slenkstį sl_{pr}, pirmojo kadro pradžioje fiksuojama žodžio pradžia. Analogiškai nustatoma žodžio pabaiga, tik šiuo atveju tikrinama, ar įvertis E_i gretimų kadrų skaičiuje L yra mažesnis už žodžio pabaigos slenkstį sl_{pab}.

Atpažinimo sistemoje kadro trukmę ir žingsnį, slenksčių koeficientų a_{pr} ir a_{pab} reikšmes, kadrų skaičių M žodžio pradžiai nustatyti ir kadrų skaičių L žodžio pabaigai nustatyti naudotojas gali keisti. Atliekant eksperimentus buvo nustatytos šios pradinės parametrų reikšmės: *kadro trukmė* = 25 ms, *kadro žingsnis* = 10 ms, $a_{pr} = 3{,}5$, $a_{pab} = 4$, $M = 10$, $L = 30$. Šias reikšmes vadinsime numatytosiomis sistemos parametrų reikšmėmis.

Nustačius žodžio ribas, apdorojamas žodį reprezentuojantis kalbos signalas. Atpažinimo sistemoje realizuoti standartiniai kalbos signalo apdorojimo ir požymių išskyrimo metodai (Rabiner, 1989, 1993; Young ir kt., 2005).

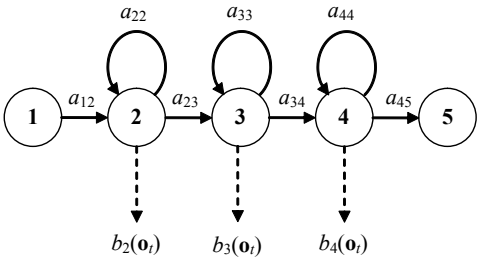
Pirmiausia iš kalbos signalo pašalinama nuolatinė dedamoji – iš signalo atskaitų reikšmių atimamas jų vidurkis. Signalas dalijamas į 25 ms trukmės kadrus su 10 ms žingsniu tarp gretimų kadrų. Kiekvieno kadro signalas filtruojamas pirmos eilės ribotos impulsinės reakcijos filtru ir dauginamas iš Hamingo (Hamming) lango funkcijos. Sistemoje naudojamo filtro koeficientas $\alpha = 0{,}97$.

Atlikus pradinį apdorojimą išskiriami signalą reprezentuojantys akustiniai požymiai. Šioje sistemoje žodžiams atpažinti naudojami melų dažnių skalės kepstriniai koeficientai. Taikant greitosios Furjė transformacijos algoritmą apskaičiuojami kiekvieno kadro signalo amplitudinio spektro įverčiai. Gauti įverčiai transformuojami į melų skalę naudojant trikampius juostinius filtrus, kurie išdėstomi pagal melų dažnių skalę. Filtrų išėjimuose gauti spektro įverčiai transformuojami į kepstrinius koeficientus taikant diskrečiąją kosinuso transformaciją šių įverčių logaritmams. Kiekvieno kadro signalo požymių vektorius sudaromas iš 39 elementų: 13 kepstro koeficientų, 13 jų išvestinių pirmos eilės ir 13 antros eilės koeficientų. Gauta požymių vektorių seka, kitaip vadinama stebėjimų arba akustinių vektorių seka, pateikiama, kad būtų atpažintas ištartas žodis.

Kad nustatytų, kurį ištartą žodį reprezentuoja atpažinti pateikta požymių vektorių seka $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T\}$, sistema naudoja akustinių modelių aibę ir žodyną. Sistema skaičiuoja kiekvieno žodžio paslėptojo Markovo modelio tikėtinumo, kad gauta požymių vektorių seka buvo sukurta to modelio, įvertį ir iš žodyno pateikia, kaip atpažintą tą žodį, kurį reprezentuoja modelis, turintis didžiausią tikėtinumo įvertį.

Markovo modelį galima apibūdinti kaip baigtinio skaičiaus būsenų sistemą. Sakoma, kad stebimą požymių vektorių seką sukūrė žodį reprezentuojantis paslėptasis Markovo modelis.

Atpažintuve kiekviena fonema modeliuojama kairės-dešinės tipo tolydžiuoju paslėptuoju Markovo modeliu, kurio struktūra vaizduojama 2 paveiksle. Kaip matyti, kiekvienos fonemos



2 p a v. ***Kairės-dešinės paslėptasis Markovo modelis***

modelį sudaro trys būsenos, kurios sukuria stebėjimus, ir dvi stebėjimų nesukuriančios būsenos, kurios skirtos modeliams jungti į sekas.

Kiekvienu diskrečiojo laiko momentu t paslėptasis Markovo modelis pereina iš būsenos i į būseną j esant tikimybei a_{ij}:

$$a_{ij} = P(s_t = j \mid s_{t-1} = i),$$
$$i, j = 1, ..., N, a_{ij} \geq 0, \sum_j a_{ij} = 1, \quad (5)$$

čia s_t – būsena laiko momentu t, N – būsenų skaičius.

Tikimybė laiko momentu t patekti į j-ąją būseną priklauso tik nuo to, kokioje būsenoje modelis buvo laiko momentu $t - 1$. Kaip matyti iš 2 paveikslo, modelis gali likti toje pačioje būsenoje arba pereiti į gretimą būseną, kiti perėjimai negalimi. Matrica, sudaryta iš perėjimų tikimybių, vadinama perėjimo matrica. Kai įvyksta perėjimas į j-ąją būseną, sukuriamas stebėjimo vektorius $\mathbf{o}_t$ su tikėtinumu $b_j(\mathbf{o}_t)$. Funkcija $b_j(\mathbf{o}_t)$ yra stebėjimų tikimybinio tankio funkcija, priskirta j-ajai būsenai:

$$b_j(\mathbf{o}_t) = p(\mathbf{o}_t \mid s_t = j), j = 2, ..., N - 1. \quad (6)$$

Sistemoje stebimi požymių vektoriai aproksimuojami Gauso tankio funkcijų mišiniu:

$$b_j(\mathbf{o}_t) = \sum_{k=1}^{K} c_{jk} \mathrm{N}(\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}), \quad (7)$$

čia K – mišinio komponenčių skaičius, c_{jk} – mišinio komponenčių svoriai, tenkinantys sąlygas $c_{jk} \geq 0$ ir $\sum_{k=1}^{K} c_{jk} = 1$, $\mathrm{N}(\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk})$ – daugiamatė tolydžioji Gauso tankio funkcija, $\boldsymbol{\mu}_{jk}$ – k-osios mišinio komponentės vidurkio vektorius, $\boldsymbol{\Sigma}_{jk}$ – kovariacinė matrica. Atpažintuve naudojama diagonalioji kovariacinė matrica. Sumažėja skaičiavimų ir reikalingos atminties mastas, nes reikia įvertinti tik stebėjimų dispersijas. Kiekvienos būsenos stebėjimai modeliuojami keturių Gauso tankio funkcijų mišiniu.

Perėjimo matrica, Gauso tankio funkcijų svoriai, vidurkių ir dispersijų vektoriai sudaro paslėptojo Markovo modelio parametrus, kurie įvertinami iš kalbos įrašų mokymo metu naudojant Baumo ir Velčo (Baum–Welch) algoritmą, kuris grindžiamas tikėtinumo maksimizavimo kriterijumi (Young ir kt., 2005). Sistemoje naudojamas 73 fonemų, kurios nepriklauso nuo konteksto, modelių rinkinys. Modeliai buvo mokomi pagal skirtingus žodžių tarimo pavyzdžius. Mokymo žodyną sudarė šimtas žodžių. Mokymo duomenis sudarė penkių kalbėtojų – trijų moterų ir dviejų vyrų – įrašai. Kiekvienas kalbėtojas ištarė visus žodžius po keliolika kartų. Iš viso mokymo duomenis sudarė 12 650 pavyzdžių.

Kai atliekamas pavienių žodžių atpažinimas, kiekvieną k-ąjį duoto žodyno žodį atitinka atskiras modelis M_k. Sistema pagal žodyne esančias žodžių fonetines transkripcijas sukuria žodžio modelį, sujungdama fonemų modelius. Sistema iš žodyno pateikia kaip atpažintą tą žodį, kurį reprezentuoja modelis, turintis didžiausią tikimybės įvertį:

$$j = \arg\max_k P(M_k|\mathbf{O}), \quad (8)$$

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 $P(M_k|\mathbf{O})$, indeksas. Aposteriorinė tikimybė skaičiuojama naudojant Bajeso (Bayes) formulę:

$$P(M_k|\mathbf{O}) = \frac{p(\mathrm{O}|M_k)P(M_k)}{P(\mathbf{O})}, \quad (9)$$

čia $p(\mathbf{O}|M_k)$ – tikėtinumas, kad požymių vektorių seką $\mathbf{O}$ generavo modelis M_k, $P(M_k)$ – apriorinė modelio tikimybė, $P(\mathbf{O})$ – požymių vektorių sekos $\mathbf{O}$ tikimybė. Vardiklyje esantis dydis $P(\mathbf{O})$ yra pastovus visiems modeliams ir neturi įtakos formulės (8) rezultatui, todėl į tolesnius skaičiavimus neįtraukiamas:

$$j = \arg\max_k P(M_k|\mathbf{O}) = \arg\max_k p(\mathbf{O}|M_k)P(M_k). \quad (10)$$

Apriorinė tikimybė $P(M_k)$ vadinama kalbos modeliu, o tikėtinumas $p(\mathbf{O}|M_k)$ – akustiniu modeliu. Kadangi šioje sistemoje kalbos modelis nenaudojamas, t. y. laikoma, kad sistemoje visų modelių apriorinės tikimybės vienodos, žodžio atpažinimas sistemoje priklauso tik nuo tikėtinumo $p(\mathbf{O}|M_k)$:

$$j = \arg\max_k P(M_k|\mathbf{O}) = \arg\max_k p(\mathbf{O}|M_k). \quad (11)$$

Kadangi būsenų seka S nežinoma, tikėtinumas skaičiuojamas sumuojant pagal visas galimas būsenų sekas $S = s_1, \ldots, s_T$:

$$p(\mathbf{O}|M_k) = \sum_S a_{s_0 s_1} \prod_{t=1}^{T} b_{s_t}(\mathbf{o}_t) a_{s_t s_{t+1}}, \quad (12)$$

čia s_0 ir s_{T+1} atitinkamai žymi modelio įėjimo ir išėjimo būsenas, $s_0 = 1$, $s_{T+1} = N$.

Alternatyviai tikėtinumo $p(\mathbf{O}|M_k)$ įvertį galima apskaičiuoti pagal vieną labiausiai tikėtiną būsenų seką:

$$\hat{p}(\mathbf{O}|M_k) = \max_S \{a_{s_0 s_1} \prod_{t=1}^{T} b_{s_t}(\mathbf{o}_t) a_{s_t s_{t+1}}\}. \quad (13)$$

Tikėtinumo skaičiavimas pagal šias formules reikalauja labai daug skaičiavimų, todėl kalbos atpažinimo sistemose įverčiui skaičiuoti naudojami efektyvūs skaičiavimo algoritmai. Žodžių atpažintuve kiekvieno žodžio modelio tikėtinumo $\hat{p}(\mathbf{O}|M_k)$, kad gauta seka $\mathbf{O}$ buvo generuota to modelio M_k, įvertis apskaičiuojamas naudojant dinaminio programavimo principą realizuojantį Viterbio (Viterbi) algoritmą.

Viterbio algoritmas randa seką $\boldsymbol{S}^*$, kuri maksimizuoja tikėtinumą $p(\mathbf{O},\boldsymbol{S}^*|M_k)$. Kitais žodžiais tariant, randa būsenų seką, kuri galėtų sukurti stebėjimų seką su didžiausiu tikėtinumu. Tarkime, $\varphi_j(t)$ reprezentuoja tikėtinumą, kad laiko momentu t modeliui esant j-ojoje būsenoje stebima požymių vektorių seka $\mathbf{o}_1$, $\mathbf{o}_2$, ..., $\mathbf{o}_t$, o prieš tai buvusios $t-1$ būsenos sudaro labiausiai tikėtiną seką. Dalinis tikėtinumas $\varphi_j(t)$ skaičiuojamas naudojant rekursiją:

$$\varphi_j(t) = \max_{1<i<N} \{\varphi_i(t-1) a_{ij}\} b_j(\mathbf{o}_t), \quad 1 < j < N, 2 \le t \le T, \quad (14)$$

kai pradinės sąlygos

$$\varphi_1(1) = 1, \ \varphi_j(1) = a_{1j} b_j(\mathbf{o}_1), \ 1 < j < N. \quad (15)$$

Tikėtinumo $\hat{p}(\mathbf{O}|M_k)$ įvertis gaunamas, kai apskaičiuojamas visos požymių vektorių sekos, kurią galėjo sukurti labiausiai tikėtina būsenų seka, tikėtinumas:

$$\hat{p}(\mathbf{O}|M_k) = \varphi_N(T) = \max_{1<i<N} \{\varphi_i(T) a_{iN}\}. \quad (16)$$

Kad būtų galima atkurti geriausią būsenų seką, kiekvienu laiko momentu t kiekvienai j-ajai būsenai išsaugoma geriausios būsenos laiko momentu $t-1$ indeksas i. Kadangi atliekama daug daugybos operacijų su labai mažais skaičiais, jie gali netilpti į jiems skirtą atminties vietą. Todėl skaičiavimai atliekami naudojant logaritmus. Skaičiavimų mastas taip pat sumažėja, nes daugyba keičiama sudėtimi.

Informacijos apie atpažinimo procesą ir rezultatus vaizdavimas

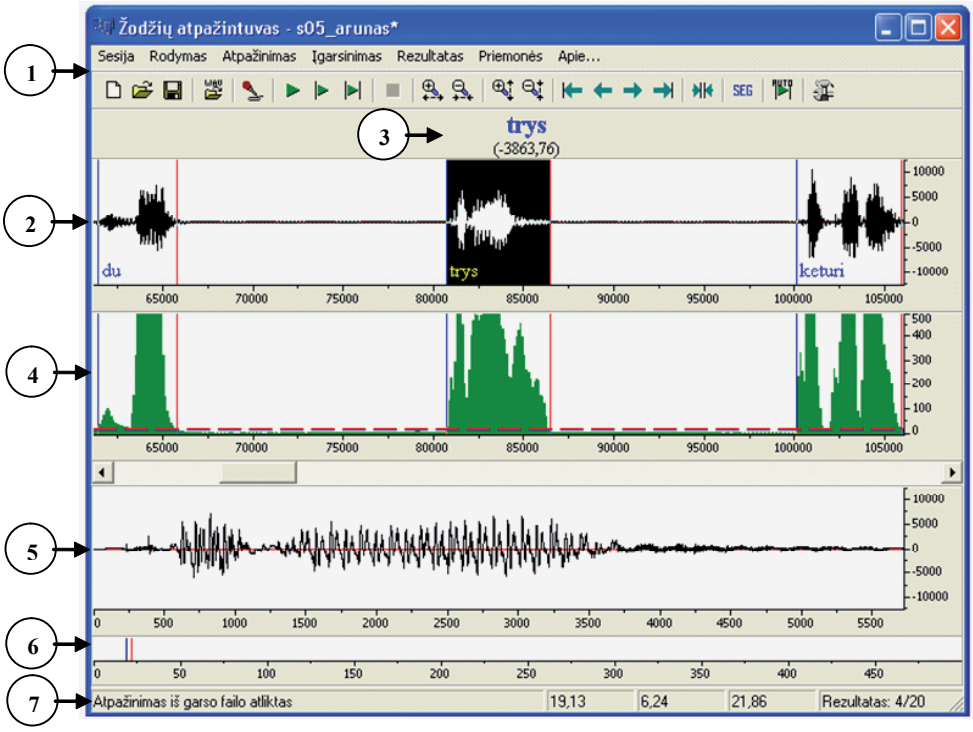
Programinė įranga „Žodžių atpažintuvas“ turi grafinę sąsają, kurioje yra priemonės atpažintuvui valdyti, galima keisti anksčiau minėtų sistemos parametrų reikšmes. Informacijai apie atpažinimo procesą ir gautus rezultatus vaizduoti skirti du langai. Pagrindiniame lange yra meniu ir mygtukų juostos (3 pav., 1) atpažinimo sistemai valdyti.

Signalo srityje vaizduojama kalbos signalo bangos forma, skirtingos spalvos vertikaliomis linijomis žymimos aptiktos žodžių ribos, įrašomi atpažinti žodžiai (3 pav., 2). Pažymėjus norimą žodį, rodomas jo atpažinimo tikėtinumo logaritmo įvertis (3 pav., 3). Šį žodį galima įgarsinti. Signalo energijos srityje vaizduojama signalo energija ir nustatytos žodžių ribos (3 pav., 4). Skirtingų spalvų trūkios horizontalios linijos žymi žodžio ribų aptikimo energijos slenksčius. Žodžio signalo srityje (3 pav., 5) vaizduojama signalo srityje pažymėto atpažinto žodžio signalo bangos forma. Žemiau rodoma įgarsinamo, įvedamo iš mikrofono arba failo signalo energija ir skirtingų spalvų vertikaliomis atkarpomis vaizduojami žodžio ribų aptikimo energijos slenksčiai (3 pav., 6). Būsenos juostoje pateikiama informacija apie atpažinimo eigą, skaitinės energijos slenksčių reikšmės, įvedamo signalo energijos reikšmė, pažymėto žodžio numeris ir visų atpažintų žodžių skaičius (3 pav., 7).

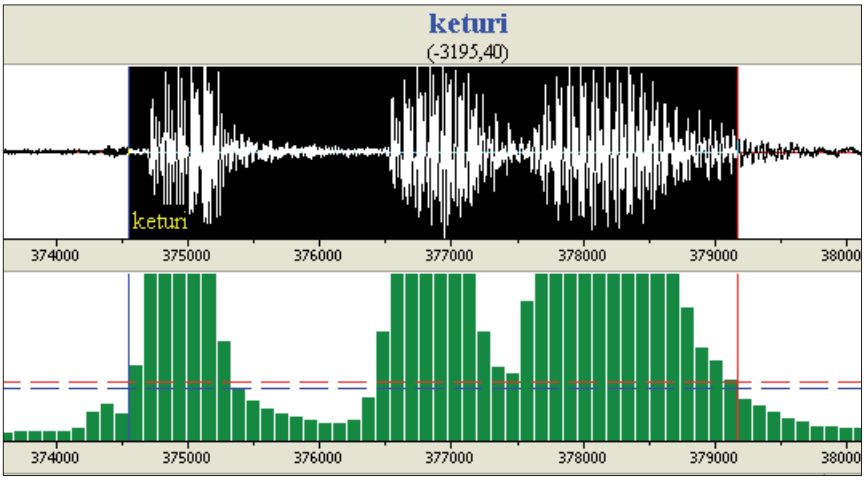
Naudotojas gali pasirinkti signalo ir energijos vaizdo rodymo mastelį. 4 paveiksle matyti žodžio „keturi“ signalo vertikaliai ir horizontaliai padidintas vaizdas. Pradinis vaizdas pateiktas 5 paveikslo b dalyje.

Kai sistema atpažįsta žodį klaidingai, viena iš galimų klaidos priežasčių yra neteisingai nustatytos žodžio ribos. Signalo, energijos, žodžio ribų vizualizavimas ir galimybė įgarsinti žodį

leidžia greitai patikrinti, ar sistema žodį atpažino teisingai, ir įvertinti, ar teisingai nustatė žodžio ribas. Ši informacija svarbi siekiant pagerinti žodžių atpažinimo tikslumą. Jei lange matyti, kad ribos nustatytos klaidingai, galima keisti minėtus parametrus, darančius įtaką nustatant ribas. Tam tikrais atvejais tai padidina atpažinimo tikslumą. Toliau pateikiamas pavyzdys, kuris iliustruoja kadrų skaičiaus žodžio pradžiai nustatyti įtaką.



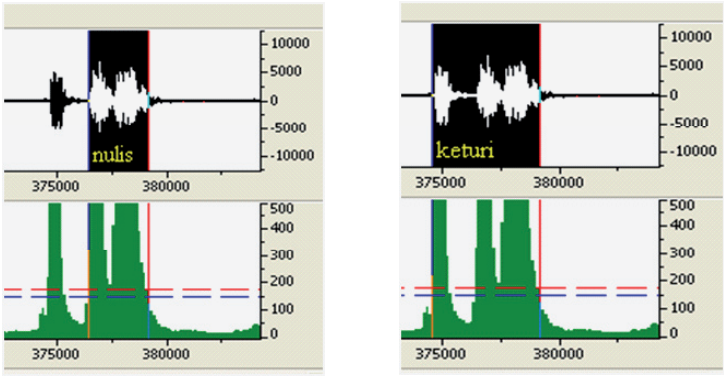
3 p a v. ***Pagrindinis „Žodžių atpažintuvo“ langas***



4 p a v. ***Signalo ir jo energijos vaizdas padidinus mastelį***

Penktame paveiksle pavaizduoti žodžio „keturi“ atpažinimo rezultatai. Abiem atvejais sistemai buvo pateiktas tas pats garso įrašas. Pirmuoju atveju žodžio pradžia buvo nustatyta iš dešimties kadrų. Matyti, kad sistema žodį „keturi“ atpažino klaidingai – kaip „nulį“ (5 pav., a). Antruoju atveju žodžio pradžia buvo nustatyta iš aštuonių signalo kadrų (5 pav., b). Matyti, kad pakeitus kadrų skaičių sistema nustatė žodžio pradžią teisingai ir žodis buvo atpažintas. Taigi galima tvirtinti, kad pirmuoju atveju klaidingo atpažinimo priežastis buvo neteisingai nustatytos žodžio ribos.

Žodžio ribų nustatymo tikslumui atpažintuve daro įtaką ne tik žodžio pradžios ir pabaigos kadrų skaičiai. Žodžio pradžios (1) ir pabaigos (2) slenksčių koeficientai, garso įrašymo įranga ir aplinka, kurioje atliekamas atpažinimas, taip pat turi įtakos ribų nustatymo tikslumui. Informacijos vizualizavimas leidžia lengviau išsiaiškinti žodžių ribų nustatymo klaidų priežastis, kai žodžių atpažinimas atliekamas skirtingomis sąlygomis. Atliekant didesnio masto eksperimentinius tyrimus galima rasti optimalias žodžio ribų nustatymo parametrų reikšmes ir taip pagerinti žodžių atpažinimą.

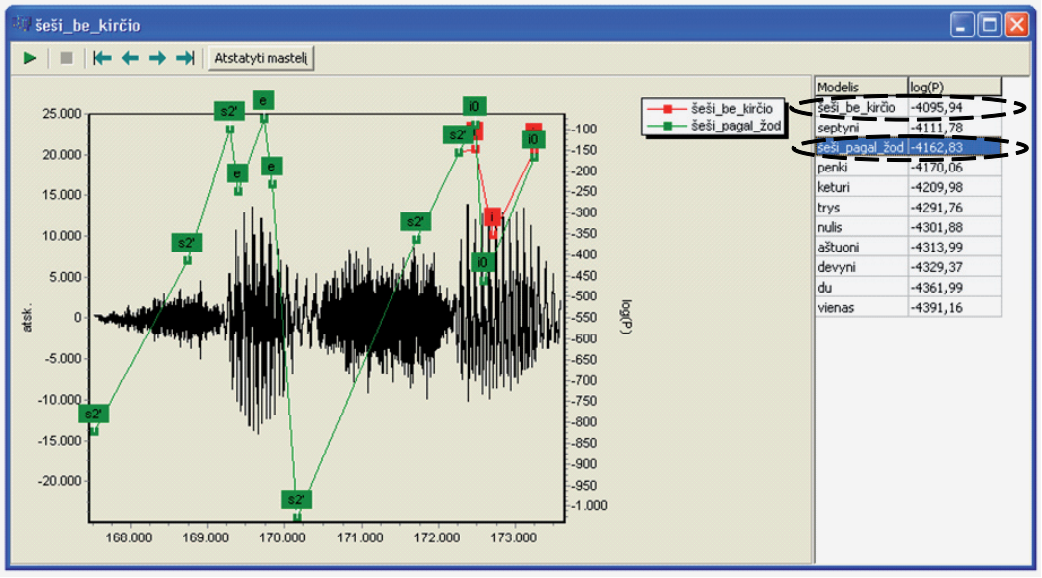


(a) (b)

5 p a v. ***Žodžio „keturi“ atpažinimo rezultatai: a) kadrų skaičius žodžio pradžiai nustatyti M = 10; b) kadrų skaičius žodžio pradžiai nustatyti M = 8***

Žodžio paieška, kuri atliekama naudojant Viterbio algoritmą, vaizduojama kitame lange (6 pav.).

Šiame lange pateikiamas žodžių, tekste vadinamų žodžiais kandidatais, kurių tikėtinumas didžiausias, ir jų tikėtinumo logaritmo



6 p a v. ***Žodžio paieškos vaizdavimas***

įverčių sąrašas. Atpažintuve žodžiams atpažinti naudojamas paieškos su spinduliu algoritmas. Kiekvienu laiko momentu *t* Viterbio algoritmu apskaičiuotas kiekvieno modelio dalinio tikėtinumo įvertis (14) lyginamas su slenksčiu, kurio reikšmė priklauso nuo paieškos spindulio reikšmės. Kuo mažesnis šis spindulys, tuo daugiau mažiausiai tikėtinų modelių, reprezentuojančių žodyne esamus žodžius, kuriuos turi atpažinti sistema, pašalinama iš paieškos atpažinimo metu. Dėl to gali labai pagreitėti žodžio atpažinimas, tačiau pablogėti atpažinimo tikslumas. Žodžių kandidatų sąrašo ilgis priklauso nuo paieškos spindulio. Paieškos spindulio reikšmę galima keisti ir atlikus eksperimentinius tyrimus parinkti reikšmę, kuri labai susiaurina paieškos erdvę, tačiau atpažinimo tikslumas nesumažėja.

Žodžių paieškos vaizdavimo lange kartu su žodžių kandidatų sąrašu pateikiama diagrama (6 pav.), kurioje vaizduojama pasirinkto žodžio signalo bangos forma. Ar žodis buvo atpažintas teisingai, galima patikrinti įgarsinus kalbos signalą. Diagramoje vaizduojama, kuriems signalo segmentams Viterbio algoritmas priskyrė žodį sudarančių fonemų modelius. Kaip buvo minėta, kiekvieną fonemos modelį sudaro trys būsenos (2 pav.). Kairėje esančioje vertikalioje ašyje vaizduojamos signalo atskaitų reikšmės, dešinėje – tikėtinumo logaritmo įverčiai. Spustelėjus sąraše esantį žodį, diagramoje braižoma tą žodį reprezentuojanti būsenų seka. Turint kelias kreives galima įvertinti kiekvieno fonemos modelio įtaką žodžio atpažinimui ir numatyti tolesnių tyrimų kryptis. Kaip buvo rašyta, sistema sukuria žodžio modelį sujungdama fonemų modelius pagal žodyne esamas žodžių fonetines transkripcijas. Kai žodžiai transkribuojami pagal tarties žodyną, kiekvieno žodžio tarimą nurodo viena fonetinė transkripcija, kuri atspindi taisyklingą lietuvių bendrinės kalbos tartį. Tačiau dėl tarmių, svetimų kalbų įtakos ir kitų priežasčių atsiranda tarties klaidų (Pakerys, 2003). Todėl, kai žodis dažnai atpažįstamas klaidingai, gali būti tikslinga atpažintuvo žodyne naudoti šiek tiek pakeistas kai kurių žodžių transkripcijas arba kelias to paties žodžio fonetines transkripcijas. Tam tikrais atvejais tai padidina atpažinimo tikslumą. Parinkti transkripcijas padeda šiame lange pateikiama informacija.

Buvo atliktas skaičių nuo nulio iki devynių atpažinimo eksperimentinis tyrimas, kuris parodė, kad atpažinimo tikslumas priklauso nuo žodyne naudojamų žodžio transkripcijų. Naudojant straipsnyje aprašytą fonemų modelių rinkinį ir žodžius transkribuojant pagal žodyną (Vaitkevičiūtė, 2001), skirtingų kalbėtojų ištartų skaičių bendras atpažinimo tikslumas buvo 80 procentų (Ringelienė, Lipeika, 2010). Pakeitus kai kurių žodžių transkripcijas atpažinimo tikslumas padidėjo iki 92 procentų.

Toliau pateikiamas pavyzdys, kuris parodo, kaip pateikta informacija, atliekant skaičių atpažinimo eksperimentą, padėjo įvertinti transkripcijų įtaką žodžio „šeši“ atpažinimui, kuris vaizduojamas 6 paveiksle. Diagramoje nubrėžtos tik dvi kreivės, kurios vaizduoja žodį sudarančių fonemų modelių būsenų seką, kai naudojamos dvi žodžio „šeši“ transkripcijos. Pirmoji – pagal Lietuvių kalbos tarties žodyną, antroji – kai žodžio transkripcijoje naudojamas nekirčiuotas balsis „i“. Iš 6 paveikslo matyti, kad žodžio tikėtinumo įvertis yra didesnis, kai naudojamas nekirčiuotas balsis „i“. Jei žodyne paliekama tik pirmoji žodžio „šeši“ transkripcija, kai balsis yra kirčiuotas (diagramoje žymimas „i0“), sistema žodį „šeši“ atpažįsta klaidingai – kaip „septynis“. Kad būtų aiškiau, lentelėje pateikti žodžio ir kiekvieną fonemos modelį sudarančių būsenų tikėtinumo logaritmo įverčiai. Matyti, kad skiriasi tik fonemų „i“ ir „i0“ įverčiai.

Reikia paminėti, kad skaičių atpažinimo eksperimentiniai tyrimai parodė, kad ne visada geriausia naudoti transkripciją, pagal kurią sukuriamo žodžio modelio tikėtinumo įvertis atpažinimo metu gaunamas didžiausias. Šiuo atveju žodis gali būti atpažintas teisingai, tačiau kitų žodžių atpažinimo tikslumas gali pablogėti (Ringelienė, Lipeika, 2010).

Kaip minėta, atpažinimo sistemos pagrindu buvo sukurta interneto naršyklės valdymo balsu sistema. Tyrimas parodė, kad 71 komandos atpažinimo tikslumas – 77 procentai, kai sistemoje naudojamas straipsnyje aprašytas fonemų modelių rinkinys ir atliekamas nepriklausomas

L e n t e l ė. ***Skirtingų transkripcijų įtaka žodžio „šeši“ atpažinimui***

Šeši /s2‘e s2‘ i/, (s2 žymi š), ***i*** nekirčiuotas. Žodžio modelio įvertis: **-4095,94**											
Fonemos š‘ modelis			Fonemos e modelis			Fonemos š‘ modelis			Fonemos i modelis		
s2‘	s2‘	s2‘	e	e	e	s2‘	s2‘	s2‘	i	i	i
-820,75	-414,15	-101,02	-249,51	-74,72	-233,35	-1026,95	-365,38	-157,27	**-148,42**	**-353,15**	**-151,27**
Šeši /s2‘e s2‘ i0/. Žodžio modelio įvertis: **-4162,83**											
Fonemos š‘ modelis			Fonemos e modelis			Fonemos š‘ modelis			Fonemos i0 modelis		
s2‘	s2‘	s2‘	e	e	e	s2‘	s2‘	s2‘	i0	i0	i0
-820,75	-414,15	-101,02	-249,51	-74,72	-233,35	-1026,95	-365,38	-157,27	**-89,49**	**-463,19**	**-167,05**

nuo kalbėtojo atpažinimas (Ringelienė, Lipeika, 2010). Preliminarių tyrimų duomenimis, galima rasti tam tikrų žodžių transkripcijas, kurias naudojant pagerėja konkretaus kalbėtojo ištartų žodžių atpažinimo tikslumas. Taigi teigtina, kad žodžių paieškos lange pateikiama informacija gali padėti pagerinti žodžių atpažinimo, kuris nepriklauso nuo kalbėtojo, tikslumą ir pritaikyti atpažinimo sistemą konkrečiam kalbėtojui.

Išvados

Programine įranga „Žodžių atpažintuvas“ galima tirti atskirų lietuvių kalbos žodžių atpažinimą, grįstą paslėptaisiais Markovo modeliais. Jos pagrindu galima kurti programinę įrangą kompiuteriui valdyti balsu.

Naudotojui pateikiama tekstinė, skaitinė, grafinė, garsinė informacija apie žodžių atpažinimo eigą ir rezultatus padeda analizuoti klaidų priežastis. Signalo, energijos, žodžio ribų vizualizavimas ir galimybė įgarsinti žodį leidžia greičiau patikrinti, ar sistema žodį atpažino teisingai, įvertinti, ar teisingai nustatė žodžio ribas, ir lengviau rasti žodžių ribų nustatymo klaidų priežastis, kai žodžių atpažinimas atliekamas skirtingomis sąlygomis. Jei sistema suklydo atpažindama žodį ir ribos nustatytos klaidingai, galima keisti sistemos parametrus, darančius įtaką ribų nustatymo tikslumui. Tam tikrais atvejais tai padidina žodžių atpažinimo tikslumą.

Preliminarūs tyrimai parodė, kad nuo žodžių transkripcijų priklauso atpažinimo tikslumas, kai sistemoje naudojami fonemų modeliai. Žodžių paieškos vaizdavimo lange pateikiama informacija padeda įvertinti kiekvieno fonemos modelio įtaką žodžio modelio atpažinimui, leidžia palyginti žodžių skirtingų transkripcijų įtaką žodžių atpažinimo tikslumui. Remiantis šia informacija galima parinkti žodžių transkripcijas, kurios pagerina nuo kalbėtojo nepriklausomo žodžių atpažinimo rezultatus. Tai taip pat gali padėti pritaikyti atpažinimo sistemą konkrečiam kalbėtojui.

LITERATŪRA

BAKER, J. K. (1975). The DRAGON system – an overview. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP–23, no. 1, p. 24–29.

BAUM, L. E.; EAGON J. A. (1967). An inequality with applications to statistical estimations for probabilistic functions of Markov processes and to a model of ecology. *Amer. Math. Soc. Bull.,* vol. 73, p. 360–362.

BAUM, L. E.; PETRIE, T.; SOULES, G.; WEISS, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, vol. 41, no. 1, p. 164–171.

DAVIS, S.; MERMELSTEIN, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP–28, no. 4, p. 357–366.

HUANG, X.; ACERO, A.; HON, H-W. (2001). *Spoken Language Processing*. New Jersey: Prentice-Hall, Inc. 980 p. ISBN 0130226165.

JELINEK, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, vol. 64, no. 4, p. 532–556.

PAKERYS, A. (2003). *Lietuvių bendrinės kalbos fonetika*. Vilnius: Enciklopedija. 244 p. ISBN 9986433320.

RABINER, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, vol. 77, no. 2, p. 257–286.

RABINER, L. R., JUANG, B.-H. (1993). *Fundamentals of speech recognition*. New Jersey: Prentice-Hall, Inc. 496 p. ISBN 0132858266.

RINGELIENĖ, Ž. (2009). Naršyklės valdymas balsu. *Informacijos mokslai*, t. 50, p. 223–227.

RINGELIENĖ, Ž.; LIPEIKA, A. (2010). Development of the hidden Markov models based Lithuanian speech recognition system. *Proc. of SPIE*, vol. 7745, 774512-1.

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). Cambridge University Engineering Department [žiūrėta 2011 m. balandžio 18 d.]. Prieiga per internetą: <http://htk.eng.cam.ac.uk/docs/docs.shtml>.

VAITKEVIČIŪTĖ, V. (2001). *Lietuvių kalbos tarties pagrindai ir žodynas.* Vilnius: Pradai. 1241 p.

A TOOL FOR VISUALIZATION AND ANALYSIS OF ISOLATED WORD RECOGNITION BASED ON THE HIDDEN MARKOV MODELS

Živilė Ringelienė, Mark Filipovič

S u m m a r y

The paper presents a prototype of the isolated word recognition system based on hidden Markov models. The developed prototype of the speaker-independent Lithuanian isolated word recognition system is handy for recognition experiments and the analysis of their results. The user is provided with numeric and visual recognition information on the results. The word recognition pivots on the precision of the determination of the word limits. The main window contains a recognized word and its logarithmic likelihood, a visible waveform of the speech signal, the depicted energy of the speech signal, the identified word boundaries and energy detection thresholds. If the system misrecognized the word, such visualization enables to identify easier whether it resulted from wrong end-point detection. The segmentation window provides with a list of words which acoustic models to the given speech signal are the best, the scores of their likelihood and a diagram of the most likely sequence of the phoneme models aligned with the speech signal. Such visualization helps to analyze recognition errors and the impact of each phoneme model on the recognition accuracy. Results of preliminary experiments have shown that by changing the transcription of some words the recognition accuracy can be increased.