Liet. matem. rink., T. 41, spec. nr., 584–590
© 2001 Matematikos ir informatikos institutas

Laiko skalės išlyginimas kalbos ir kalbančiojo atpažinime

Antanas LIPEIKA, Joana LIPEIKIENĖ (MII)
el. paštas: lipeika@ktl.mii.lt

1. Įvadas

Sprendžiant atskirai pasakytų žodžių atpažinimo [1] arba kalbančiojo atpažinimo pagal raktines frazes [2] uždavinį reikia sutapatinti laike ir palyginti, kiek atpažįstamas balso pavyzdys yra panašus į etaloninį balso pavyzdį. Atpažįstamo ir etaloninio balso pavyzdžių trukmės ir atskirų garsų trukmės juose paprastai skiriasi, todėl jų palyginimas yra gana sudėtingas uždavinys.

Balso pavyzdžiai atpažinime dažniausiai yra atstovaujami trumpalaikių spektrinių požymių sekomis, apibūdinančiomis pavyzdžiuose esančios garsus (fonemas). Atpažįstant pirmiausia reikia sutapatinti laike dviejų pavyzdžių tuos pačius skirtingo ilgio garsus atitinkančius spektrinius požymius, o po to paskaičiuoti atstumą tarp jų. Paprasčiausias atvejis, kai garsų trukmės pavyzdyje yra proporcingos kalbos pavyzdžio trukmei. Todėl nagrinėjimą ir pradėsime nuo šios situacijos [1].

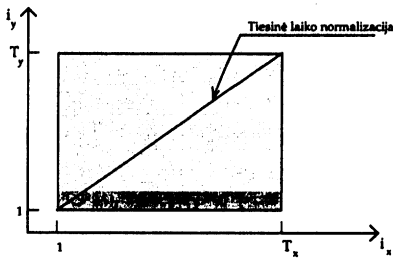
2. Tiesinis laiko skalės išlyginimas

Nagrinėkime du kalbos pavyzdžius X ir Y, atstovaujamus spektrų sekomis x_{i_x}, $i_x = 1,,\ldots,T_x$ ir y_{i_y}, $i_y = 1,,\ldots,T_y$, kur x_i ir y_i yra trumpalaikių akustinių (spektrinių) požymių vektoriai, o i_x ir i_y yra atitinkamai X ir Y laiko indeksai. Laiko trukmės T_x ir T_y nebūtinai turi būti vienodos. Skirtumas tarp X ir Y yra apibrėžiamas kokia nors trumpalaikių spektrų iškraipymo (atstumo) funkcija $d(x_{i_x}, y_{i_y})$, kurią supaprastintai žymėsime $d(i_x, i_y)$. Kadangi garsų eilės tvarka yra svarbi, yra būtina, kad lyginamų spektrų porų indeksai (x_{i_x}, y_{i_y}) tenkintų tam tikrus apribojimus.

Pagrindinis uždavinys yra kalbos pavyzdžių laiko skalės suvienodinimas, kad paskui galima būtų paskaičiuoti iškraipymus (atstumą) tarp atpažįstamo ir etaloninio žodžio arba frazės ištarimų.

Paprasčiausias laiko išlyginimo būdas yra tiesinė laiko skalės normalizacija. Esant tiesinei laiko skalės normalizacijai, skirtumas tarp X ir Y yra apibrėžiamas taip:

$$d(X,Y) = \sum_{i_x=1}^{T_x} d(i_x, i_y), \tag{1}$$



1 pav. Dviejų skirtingos trukmės požymių sekų tiesinis laiko skalės išlyginimas.

kur i_x ir i_y tenkina priklausomybę

$$i_y = \frac{T_y}{T_x} i_x. \tag{2}$$

Kadangi i_x ir i_y yra sveiki skaičiai, (2) išraiškoje gauname tam tikrą apvalinimo klaidą. Sumavimas (1) išraiškoje taip pat gali būti atliekamas ir pagal i_y. Tiesinės laiko normalizacijos metode daroma prielaida, kad kalbėjimo greičio kitimas yra proporcingas frazės trukmei ir nepriklauso nuo pasakomo garso. Tokiu būdu, iškraipymų mato paskaičiavimas yra atliekamas pagal stačiakampio įstrižainę (i_x, i_y) plokštumoje, kaip parodyta 1 pav.

Kiekvienas taškas (i_x, i_y) plokštumoje išilgai įstrižainės vaizduoja iškraipymus $d(i_x, i_y)$, t.y., atstumą tarp atitinkamų X ir Y spektrinių požymių vektorių, paskaičiuotų iš kalbos signalo segmentų (kadrų) i_x ir i_y. Tačiau prielaida apie kalbėjimo greičio nepriklausomumą nuo pasakomo garso yra neadekvati realiai situacijai ir reikalingas realesnis laiko normalizacijos būdas.

3. Netiesinis laiko skalės išlyginimas

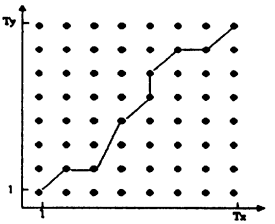
Daug bendresnis laiko skalės išlyginimo ir normalizavimo būdas naudoja dvi normalizavimo funkcijas Φ_x ir Φ_y, kurios suveda dviejų kalbos pavyzdžių laiko indeksus i_x ir i_y į bendrą laiko ašį, t.y.,

$$i_x = \Phi_x(k),\ i_y = \Phi_y(k), \quad k = 1, 2, \ldots, T. \tag{3}$$

Globalus pavyzdžių nepanašumo (atstumo) matas $d_\Phi(X, Y)$ gali būti apibrėžtas naudojant kraipymo funkcijų porą $\Phi = (\Phi_x, \Phi_y)$ kaip visam ištarimui akumuliuoti iškraipymai, būtent,

$$d_\Phi(X, Y) = \sum_{k=1}^{T} d\left(\Phi_x(k), \Phi_y(k)\right) m(k)/M_\Phi, \tag{4}$$

kur $d(\Phi_x(k), \Phi_y(k))$ vėl yra trumpalaikio spektro iškraipymai, apibrėžti požymiams $x_{\Phi_x(k)}$ ir $x_{\Phi_y(k)}$, $m(k)$ yra neneigiamas trajektorijos svorio koeficientas ir M_Φ yra tra-



2 pav. Dviejų pavyzdžių laiko skalės normalizacijos į bendrą laiko indeksą pavyzdys. Laiko skalės kraipymo funkcijos Φ_x ir Φ_y atvaizduoja laiko indeksus i_x ir i_y į bendrą laiko indeksą k.

jektoriją normalizuojantis daugiklis. 2 pav. parodytas aukščiau pateiktos laiko normalizacijos schemos pavyzdys. Linija vaizduoja trajektoriją, pagal kurią $d_\Phi(X,Y)$ yra paskaičiuotas. Tinklelio taškai ant trajektorijos yra pažymėti didėjančia tvarka nuo $k=1$ iki $k=T$, kur T yra dviejų pavyzdžių normalioje laiko skalėje „normali" trukmė.

Reikalavimas išlaikyti spektrinių pavaizdavimų X ir Y laikinę tvarką reiškia, kad kraipymo funkcijos Φ_x ir Φ_y turi būti monotoniškai nemažėjančios. Mūsų uždavinys yra nusakyti trajektoriją $\Phi=(\Phi_x,\Phi_y)$ kaip parodyta išraiškoje (4).

Akivaizdu, kad yra labai didelis galimų kraipymo funkcijų porų skaičius. Pagrindinis klausimas yra, kuri trajektorija turėtų būti parinkta, kad lyginami kalbos signalo pavyzdžiai būtų geriausiu būdu sutapatinti.

Geriausia sutapatinimo trajektorija bus ta, kuriai iškraipymai bus mažiausi [3]

$$d(X,Y)\equiv\min_{\Phi} d_\Phi(X,Y), \tag{5}$$

ieškant pagal visas galimas trajektorijas Φ, tenkinančias tam tikrą apribojimų aibę.

Tipiniai apribojimai galimoms trajektorijoms Φ yra:

- Galinių taškų apribojimai
- Monotoniškumo apribojimai
- Lokalaus tolydumo apribojimai
- Globaliniai trajektorijos apribojimai
- Svorių suteikimo trajektorijos nuolydžiui apribojimai.

Galinių taškų apribojimai. Kai lyginami kalbos pavyzdžiai yra izoliuoti žodžiai, kuriuos reikia atpažinti, jie paprastai turi gerai apibrėžtus galo taškus, kurie žymi pradinį ir galinį pavyzdžio kadrus. Ši galinio taško informacija yra paprastai gaunama kaip kalbos detektavimo operacijos rezultatas. Šia prasme yra laikoma, kad kalbos pavyzdžio galo taškai yra žinomi ir laikinės variacijos atsiranda diapazone, apibrėžtame galo taškais. Taigi, laiko normalizacijai galo taškai yra fiksuotos laikinės žodžių ribos, vedančios prie apribojimų kraipymo funkcijoms, kurie yra tokie:

$$\text{pradinis taškas}\quad \Phi_x(1)=1,\quad \Phi_y(1)=1, \tag{6a}$$

$$\text{galinis taškas}\quad \Phi_x(T)=T_x,\quad \Phi_y(T)=T_y. \tag{6b}$$

Tais atvejais, kai galo taškai negali būti patikimai nustatyti, anksčiau nurodyti galinio taško apribojimai turi būti modifikuoti, kad būtų atsižvelgta į neapibrėžtumą.

Monotoniškumo apribojimai. Laikinė spektrinių požymių sekos tvarka kalbos pavyzdyje yra lemiamos svarbos lingvistinei prasmei. Kad palaikytume laikinę tvarką atliekant laikinę normalizaciją, prasminga priimti tokius monotoniškumo apribojimus:

$$\Phi_x(k+1) \geqslant \Phi_x(k), \tag{7a}$$

$$\Phi_y(k+1) \geqslant \Phi_y(k). \tag{7b}$$

Monotoniškumo apribojimai, kaip parodyta 2 pav., reiškia, kad kiekviena trajektorija, išilgai kurios yra paskaičiuotas $d_\Phi(X, Y)$, neturės neigiamo nuolydžio. Šis apribojimas pašalina apgręžto laiko kraipymą išilgai laiko ašies, netgi trumpame laiko intervale.

Lokalaus tolydumo apribojimai. Kalbos ištarimuose, konkretaus garso buvimas kartais yra vienintelis išskirtinis faktorius, kuris padeda teisingam atpažinimui. Laiko normalizacija, randant geriausią laikinį atitikimą, kaip apibrėžta (5) išraiškoje, neturėtų praleisti jokio svarbaus informaciją nešančio garso segmento. Kad užtikrintume tinkamą laiko skalės išlyginimą, sumažinant potencialų informacijos praradimą iki minimumo, kraipymo funkcijai yra apibrėžiama lokalaus tolydumo apribojimų aibė. Lokalaus tolydumo apribojimai gali būti įvairūs [4]. Vienas pavyzdys, pasiūlytas Sakoe ir Chiba [3], yra

$$\Phi_x(k+1) - \Phi_x(k) \leqslant 1, \tag{8a}$$

$$\Phi_y(k+1) - \Phi_y(k) \leqslant 1. \tag{8b}$$

Tokiu būdu mes apibrėžiame trajektoriją P kaip perėjimų seką, kurių kiekvienas perėjimas yra nusakomas koordinačių padidėjimo pora,

$$\mathrm{P} \rightarrow (p_1, q_1)(p_2, q_2)\ldots(p_T, q_T). \tag{9}$$

Globaliniai trajektorijos apribojimai. Dėl globalinių trajektorijos apribojimų tam tikros (i_x, i_y) plokštumos dalys yra pašalinamos iš srities, kurią gali perkirsti laiko skalės kraipymo trajektorijos. Tai leidžia susiaurinti paieškos sritį ir tuo pačiu sumažinti skaičiavimų apimtį.

Svorių suteikimo trajektorijos nuolydžiui apribojimai. Šie apribojimai per svorio funkciją $m(k)$ valdo kiekvieno trumpalaikio iškraipymo $d(\Phi_x(k), \Phi_y(k))$ indėlį. Naudojant svorio funkciją galima, pvz., suteikti mažesnį svorį žodžio pradžiai ir pabaigai, kai žodžio galų taškai surandami netiksliai. Šią svorio funkciją galima panaudoti panašių žodžių atskiriamumui padidinti, suteikiant mažesnį svorį panašioms žodžio dalims.

4. Optimalios laiko skalės kraipymo trajektorijos paieška

Optimalios laiko skalės kraipymo trajektorijos paieška yra atliekama naudojant dinaminį programavimą [5]. Dėl galinio taško apribojimų (6), mes perrašome (5) išraišką per T_x ir T_y

$$M_\Phi d(X,Y) \equiv D(T_x,T_y) = \min_{\Phi_x,\Phi_y} \sum_{k=1}^{T} d\left(\Phi_x(k), \Phi_y(k)\right) m(k), \tag{10}$$

kadangi X ir Y pasibaigia atitinkamai prie T_x ir T_y, o M_Φ yra lygus

$$M_\Phi = \sum_{k=1}^{T} m(k).$$

Minimalūs daliniai akumuliuoti iškraipymai išilgai trajektorijos, sujungiančios taškus $(1,1)$ ir (i_x,i_y) yra

$$D(i_x,i_y) \equiv \min_{\Phi_x,\Phi_y,T'} \sum_{k=1}^{T'} d\left(\Phi_x(k),\Phi_y(k)\right) m(k), \tag{11}$$

kur yra laikoma, kad

$$\Phi_x(T') = i_x \quad \text{ir} \quad \Phi_y(T') = i_y. \tag{12}$$

Taigi, dinaminio programavimo rekursija su apribojimais tampa

$$D(i_x,i_y) = \min_{(i'_x,i'_y)} \left[D(i'_x,i'_y) + \zeta\left((i'_x,i'_y),(i_x,i_y)\right)\right], \tag{13}$$

kur ζ yra pasverti akumuliuoti iškraipymai (lokalinis atstumas) tarp taško (i'_x,i'_y) ir taško (i_x,i_y)

$$\zeta\left((i'_x,i'_y),(i_x,i_y)\right) = \sum_{l=0}^{L_s} d\left(\Phi_x(T'-l),\Phi_y(T'-l)\right) m(T'-l), \tag{14}$$

kur L_s yra perėjimų skaičius trajektorijoje nuo (i'_x,i'_y) iki (i_x,i_y) sutinkamai su kraipymo funkcijomis Φ_x ir Φ_y. Taip pat, iš (12) seka, kad

$$\Phi_x(T'-L_s) = i'_x \quad \text{ir} \quad \Phi_y(T'-L_s) = i'_y. \tag{15}$$

Priaugantys iškraipymai ζ yra paskaičiuojami tiktai išilgai leistinų trajektorijų, kaip apibrėžta parinktais lokalaus tolydumo apribojimais.

Tokiu būdu gauname dinaminio programavimo algoritmą geriausios trajektorijos radimui tinklelyje T_x, T_y, pradedant nuo taško $(1,1)$ ir baigiant tašku (T_x,T_y):

1. Inicializacija

$$D_A(1,1) = d(1,1)m(1).$$

2. Rekursija

Reikšmėms $1 \leqslant i_x \leqslant T_x$, $1 \leqslant i_y \leqslant T_y$, tokioms kad i_x ir i_y būtų leistiname tinklelyje, paskaičiuoti

$$D_A(i_x, i_y) = \min_{(i'_x, i'_y)} \left[D_A(i'_x, i'_y) + \zeta \left((i'_x, i'_y), (i_x, i_y) \right) \right],$$

kur $\zeta((i'_x, i'_y), (i_x, i_y))$ yra apibrėžti (14) išraiška.

3. Pabaiga

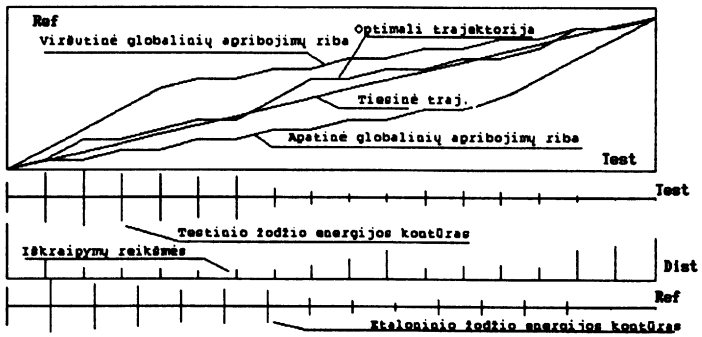
$$d(X,Y) = \frac{D_A(T_x, T_y)}{M_\Phi}.$$

Pagrindinė algoritmo idėja yra, kad rekursijos žingsnis yra atliekamas visoms lokalioms trajektorijoms, kurios pasiekia (i_x, i_y) tiesiai per vieną žingsnį (iš (i'_x, i'_y)) naudojant parinktus lokalius trajektorijos apribojimus. Tiktai (i_x, i_y) reikšmės, kurios gali būti pasiektos iš $(1,1)$ ir galiausiai pasibaigti prie (T_x, T_y) yra paskaičiuojamos rekursijos žingsnyje.

5. Laiko skalės išlyginimo realizacija

Remiantis netiesiniu laiko skalės išlyginimo būdu buvo sukurta programinė įranga šiam uždaviniui spręsti. Realizuojant buvo panaudoti Itakuros lokalaus tolydumo apribojimai [6]. Nuolydžio svorio koeficientai buvo prilyginti vienetui, galinių taškų apribojimai – (6), monotoniškumo apribojimai – (7). Globaliniai trajektorijos apribojimai buvo parinkti taip, kad paieškos ribos pagal etaloninį ištarimo pavyzdį neviršytų pusės etaloninio ištarimo pavyzdžio ilgio. Laiko skalės išlyginimo pavyzdys yra pateiktas 3 pav.

Kaip parodyta 3 pav., yra lyginami testinis ir etaloninis žodžiai „saulė“, kurių trukmės yra atitinkamai 18 ir 16 spektrinių požymių vektorių. Viršuje yra pavaizduotos viršutinė ir apatinė paieškos trajektorijos ribos, atitinkančios globalinius apribojimus. Gauta optimali laiko skalės išlyginimo trajektorija yra išsibarsčiusi tiesinio laiko skalės išlyginimo atžvilgiu. Trajektorija prasideda taške (1.1) ir baigiasi (18.16). Suradę prie priimtų apribojimų optimalią trajektoriją, mes kartu paskaičiuojame ir vidutinius iškraipymus, kurie yra šių dviejų žodžių panašumo kriterijus. Kuo vidutinių iškraipymų reikšmė mažesnė, tuo didesnė tikimybė, kad lyginami žodžiai yra tie patys. Vadinasi, mes galime konstruoti sprendimo taisyklę izolioutų žodžių atpažinimui. Priklausančiame nuo teksto kalbančiojo atpažinime sutapatinami yra ne žodžiai, o frazės ir vidutiniai iškraipymai yra lyginami su slenksčiu. Jeigu vidutiniai iškraipymai mažesni už slenkstį, daroma išvada, kad abiejų ištartų frazių autorius yra tas pats asmuo.



3 pav. Laiko skalės išlyginimo pavyzdys žodžiui „saulė".

3 pav. apačioje yra pavaizduoti testinio ir etaloninio žodžių energijos kontūrai, iš kurių vizualiai galima spręsti apie šių žodžių požymių laiko skalės nevienodumą. Taip pat grafiškai yra pavaizduota, kokie buvo gauti iškraipymai kiekvienam testiniam požymių vektoriui suradus tinkamiausią etaloninį požymių vektorių. Iš iškraipymų grafiko galima spręsti, kaip gerai yra sutapatinamos atskiros žodžio dalys. Šiame tyrime buvo naudojami tikėtinumo santykio iškraipymai [7].

Šių tyrimų rezultatus naudojame kurdami lietuvių kalbos atskirai sakomų žodžių atpažinimo ir balso rakto sistemas. Šių sistemų modeliavimui yra sukurta programinė įranga ir atliekamas sistemų darbingumo tyrimas. Dvylika lietuvių kalbos žodžių, skaičiai 0–9 ir žodžiai „pradžia", „pabaiga", buvo naudojami atpažinimui. 10 kalbėtojų ištarė šiuos žodžius po 10 kartų, esant geroms įrašymo sąlygoms. Pirmas pasakymas buvo panaudotas žodžių etalonų sukūrimui. Iš pasakytų $10 \times 12 \times 9 = 1080$ žodžių klaidingai buvo atpažinti 9. Tai sudaro 0,83 procento klaidų.

Literatūra

[1] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall (1993).
[2] R.L. Klevans and R.D. Rodman, *Voice Recognition*, Artech House (1997).
[3] H. Sakoe and S Chiba, Dynamic programming optimization for spoken word recognition, *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-26 (1): 43–49, February (1978).
[4] C. Myers, L.R. Rabiner, and A.E. Rosenberg, Performance tradeoff in dynamc time warping algorithms for isolated word recognition, *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-28 (6): 623–635, December (1980).
[5] R.E. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, New Jersey, USA (1957).
[6] F. Itakura, Minimum prediction residual principle applied to speech recognition, *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-23 (1): 57–72, February (1975).
[7] A.H. Gray and J. D. Markel, Distance measures for speech processing, *IEEE Trans. on Acoust. Speech and Signal Processing*, vol. ASSP-24, 5, 380–391 (1976).

Time alignment in speech and speaker recognition

A. Lipeika, J. Lipeikienė

Time scale alignment problems in speech and speaker recognition by voice are investigated. Dynamic programming approach to solution of this problem is discussed.