

KALBOS GARSŲ ATPAŽINIMAS

Priebalsių atpažinimas difonuose priebalsis–balsis

Vytautas Rudžionis

Kauno Technologijos universitetas

Kaunas University of Technology

Studentų 56-302, 3031 Kaunas

El. paštas: vyrud@mmlab.ktu.lt

Ivadas. Kalbos atpažinimo problemos pasaulyje tyrinėjamos jau daugiau nei 30 metų. Sprendžiant šį klausimą buvo įdėta daug pastangų bei padaryta nemaža pažanga (Young, 1996). Pastaruoju metu didelės apimties žodynų automatinės atpažinimo sistemos veikia 5–10 proc. klaidų tikslumu (ARPA testų 1994, 1996 metais rezultatų duomenimis).

Tačiau, nepaisant akivaizdžios pažangos, net ir geriausių automatinių atpažinimo sistemų efektyvumas yra žymiai prastesnis negu žmogaus (klausytojo) atpažinimo tikslumas (Lippman, 1996), o nepalankioje aplinkoje – dar blogiau.

Daugelis šiuolaikinių atpažinimo sistemų naudoja pjūvių (angl. *frames*) analizės principą – skaido signalą į nedidelius 10–20 ms trukmės fragmentus, kurie apdorojami ir klasifikuojami nepriklausomai vienas nuo kito. Kadangi net ir to paties žodžio ištara gali būti nevienodos trukmės ir vidinės struktūros, projektuojant kalbos atpažinimo sistemas tenka naudoti metodus, leidžiančius modeliuoti ištaramų trukmės variatyvumą. Tai savo ruožtu riboja gerai išvystytą statistinės klasifikavimo teorijos algoritmų naudojimą atpažįstant kalbą. Daugelis šiuolaikinių sistemų automatinių kalbos atpažinimo ištaramų trukmės variatyvumą laiko atžvilgiu modeliuoja tolydžiosiomis paslėptomis Markovo grandinėmis (angl. *Hidden Marko Model – HMM*) arba jų modifikacijomis (Rabiner, 1989). Tačiau šis modelis turi keletą

esminių trūkumų taikant jį kalbos atpažinimo tikslams. Pirma, bazinė Markovo modelių prielaida yra stebėjimo vektorių (signalų pjūvių atitinkančių požymių vektorių) nepriklausomumo prielaida. Tačiau yra žinoma, kad kalbos signalų pjūviams daro įtaką greta einantys signalo pjūviai, todėl ši prielaida yra neadekvati (Blackburn, 1996). Antra, dažniausiai HMM yra apmokomi naudojant iteracinį Baum–Velčo algoritmą, neturintį diskriminacinių savybių (Poritz, 1988).

Šių trūkumų galima išvengti naudojant fonetinio atpažinimo principą. Fonetiniu atpažinimu vadinamas toks kalbos atpažinimo būdas, kai stengiamasi rasti atskirų fonetinių vienetų ribas kalbos įrašė, parinkti jiems tinkamą aprašymo formą ir, naudojant statistinės klasifikavimo teorijos principus, priskirti juos vienai iš fonetinių vienetų klasių. Toks atpažinimo organizavimo principas turi keletą pranašumų. Pirma, fonetiškai segmentuotas ištara suteikia galimybę skirtingoms fonemoms naudoti skirtingus metodus. Tai reiškia, kad mes galime panaudoti iš anksto sukaupias akustines ir fonetines žinias, gautas statistinės analizės arba percepcijos eksperimentų metu. Antra, toks metodas leidžia modeliuoti koartikuliacijos – garsų akustinės aplinkos – efektus ir panaudoti juos kaip veiksmingą priemonę klasifikavimo metu. Trečia, skirtingiems fonetiniams vienetams atpažinti gali būti naudojami skirtingi požymiai. Ketvirta, fonetinių vienetų skyrimą tobulinti dažniausiai galima nepriklausomai nuo kitų fonetinių vienetų bei nepermokant visos sistemos.

Nepaisant akivaizdžių pranašumų, fonetinio atpažinimo principai šiuo metu naudojami nedažnai. Viena iš didžiausių kliūčių fonetiniam atpažinimui plisti – reikalingas efektyvus fonetinių vienetų išskyrimo (akustinio segmentavimo) algoritmas, nes šioje stadijoje padarytų klaidų vėliau praktiškai jau nebeįmanoma neutralizuoti (Schmid, 1996). Todėl ir darbų, skirtų tokio atpažinimo principo problemoms, nėra daug.

Dengas (1996), naudodamas TIMIT duomenų bazę ir tolydžiąsias HMM, pasiekė 54–59 proc. daugiadiktorinio balsių atpažinimo tikslumą. Valtčevas (1995), operuodamas BTL duomenų bazės E aibei priklausančiais garsais ir apmokydamas HMM maksimalios tarpusavio informacijos principu, gavo 84,5 proc. atpažinimo tikslumą. Spanias su kolegomis (1996), tyrinėdamas R aibei priklausančių garsų skyrimo problemas, pažymėjo, kad ypač

komplikuotas yra garsų m ir n atpažinimas bei kad tarp šių garsų spektrų nedideli skirtumai matomi tik formantėms pereinant iš vieno garso į kitą.

Prieš keletą metų, naudodami analoginį filtrinį analizatorių ir tiesinės diskriminantinės analizės principus, kruopščiai apmokę ir tinkamai parinkę požymius, Rudžionis ir Domatas sukūrė akustinio segmentavimo algoritmą. Dėl to pavyko gana tiksliai atpažinti nosines priebalses net ir sudėtingoje uždaro balsio i aplinkoje (Domatas, 1991). Fonetinio atpažinimo principus Rudžionis ir Noreika panaudojo 200 žodžių apimties žodyno atpažinimui (Noreika, 1991). Pastaraisiais metais, naudojant tik skaitmeninius signalų apdorojimo algoritmus, fonetinio segmentavimo ir atpažinimo principai pritaikyti platesniam klasifikatorių rinkiniui (Rudžionis, 1996).

Šiame darbe stengiamės palyginti filtrinės ir kepstrinės analizės būdu gautų požymių efektyvumą fonemų skyrimo tikslams bei panaudoti pažangius statistinės klasifikavimo teorijos pateikiamus diskriminavimo metodus, nenaudotus kalbos atpažinimo uždaviniams. Antrame skirsnyje supažindinama su atpažinimo metu naudotais požymiais ir jų skaičiavimo metodika, trečiame skirsnyje pristatoma eksperimentuose naudota duomenų bazė, o ketvirtame skirsnyje pateikiami eksperimentų rezultatai.

Teorinis pagrindimas. Fonetinio atpažinimo efektyvumą sąlygoja trys veiksniai: geras fonetinis segmentavimas, tinkamų požymių klasifikavimui parinkimas bei efektyvus apmokymo ir klasifikatoriaus panaudojimas. Sutrikimai bet kurioje iš šių duomenų apdorojimo fazių gali iš esmės sutrikdyti visą sistemos funkcionavimą.

Požymiai ir segmentavimo procedūra. Fonetinio atpažinimo požymiais vadiname parametrus, aprašančius tam tikrą fonetinį įvykį. Šiame darbe naudojami dvi skirtingas klases atstovaujantys požymiai – filtriniai ir kepstriniai. Tokį pasirinkimą lėmė du veiksniai. Mūsų anksčiau atliktuose fonemų diskriminavimo eksperimentuose filtriniai požymiai savo efektyvumu pranoko visus kitus spektrinės analizės būdu gautus požymius (Rudžionis, 1995). Kita vertus, dauguma šiuolaikinių kalbos atpažinimo sistemų naudoja kepstrinio tipo požymius (MFCC arba rečiau LPCC). Todėl yra svarbu palyginti šių tipų požymių efektyvumą skiriant fonemas.

Prieš pristatydami požymių skaičiavimo būdus, turime paaiškinti, kad skiriame du terminus: analizės parametrai ir atpažinimo požymiai. Analizės

parametrai gaunami apdorojus signalo pjūvį skaitmeninio signalų apdorojimo algoritmu, (pvz., atlikus filtraciją skaitmeniniais filtrais arba įvykdžius Furjė transformaciją). Atpažinimo požymiai gaunami iš tų pjūvių, kurie priklauso duotam fonetiniam vienetai, analizės parametrai.

Atliekant filtrinę analizę, naudotas 12 skaitmeninių trečios eilės Batervorto filtrų analizatorius. Filtrų centriniai dažniai ir juostų pločiai išdėstyti melų skalėje. Kiekvieno filtro išėjime signalas papildomai filtruotas žemo dažnio filtru su 15 Hz pjūvio dažniu nuolatinės dedamosios pašalinimui. Analizė atlikta 6,4 ms žingsniu (arba 64 taškų intervalu, nes signalo diskretinimo dažnis 10 kHz) (f12 parametrai).

Atliekant kepstrinę analizę, naudotas 25,6 ms dydžio Hemingo analizės langas. Skaiciuojant LPCC parametrus buvo gaunami 12 eilės tiesinės prognozės koeficientai Levinsono–Durbino algoritmu (Markel, 1976). Toliau jie buvo transformuoti į 12 kepstro koeficientų naudojant formulę:

$$c_n = a_n + \sum_{i=1}^{n-1} c_i a_{n-i} \frac{i}{n} . \quad (1)$$

MFCC parametrai skaičiuoti pagal Mermelsteino (Davis, 1980) pasiūlytą metodą. 256 taškų signalo pjūvis buvo papildomas nuliais iki 1024 taškų siekiant padidinti dažnio rezoliuciją. Tuomet pjūviui uždedamas Hemingo langas ir atliekama Furjė transformacija. Gautą ir logaritmuotą 512 taškų amplitudinį spektrą filtruoja 20 trikampių filtrų, išdėstytų melų skalėje. Gautai 20 koeficientų sekai atliekama antro tipo diskretinė kosinusinė transformacija ir randama 12 MFCC koeficientų pagal formulę (c12 parametrai):

$$c_n = \frac{2}{N'} \sum_{i=1}^{N'} Y(i) \cos \left(in \frac{2\pi}{N'} \right), \quad (2)$$

kur $N'=20$, o $Y(i)$ – atitinkamo trikampio filtro išėjimas.

Apskaiciuoti kalbos signalų analizės parametrai naudoti požymių vektoriams sudaryti. Savo ruožtu skiriame du požymių tipus: paprastieji ir *delta*

požymiai. Paprastieji požymiai gaunami suvidurkinant analizės parametrus per 7 arba 11 pjūvių taško, gauto naudojant fonetinio segmentavimo funkciją, aplinkoje, o *delta* požymiai gaunami įvertinant parametru kitimo greitį tuose pjūviuose. Pjūvių skaičius priklauso nuo balsio, esančio priebalsio kontekste, tipo: atvirų balsių kontekste suvidurkinama per 11 pjūvių, o uždarų balsių – per 7 pjūvius.

Fonetinius įvykius žymintys taškai gauti iš segmentavimo funkcijos ekstremumų, o ši gaunama įvertinant f_{12} parametru kitimo greitį. Šiuo atveju parametru kitimo greitis buvo įvertintas per 4 pjūvius. Pirmasis segmentinės funkcijos ekstremumas žymi stacionarios priebalsio dalies vidurį. Kiti du ekstremumai – atitinkamai maksimumas ir minimumas – žymi perėjimo iš priebalsio į balsį ir stacionarios balsio dalies vidurinius taškus. Toks pasirinkimas remiasi tuo, kad spektriniai kalbos signalo parametrai greičiausiai kinta vykstant perėjimui iš vieno garso į kitą (vykstant artikuliatorių padėties pasikeitimui balso trakte), o lėčiausiai – tariant stacionarią garso dalį (šiuo atveju artikuliatorių judėjimas nežymus).

Klasifikavimo metodų parinkimas. Statistinio atpažinimo teorijoje yra aprašyta ir išanalizuota daug įvairių objektų klasifikavimo algoritmų. Tačiau kalbos atpažinimo uždaviniams ši patirtis nėra labai plačiai taikyta, nes pjūviais paremtoje atpažinimo koncepcijoje daugelio šių metodų panaudojimas yra komplikotas.

Mūsų atliktuose viendiktoriniuose ir daugiadiktoriniuose balsių atpažinimo tyrimuose parodytas tiesinėmis diskriminantinėmis funkcijomis paremtų klasifikatorių (Fišerio, dichotominio) bei neuroninio tinklo pranašumas prieš kitus klasifikavimo metodus (Rudžionis, 1996). Vėliau ši patirtis patvirtinta ir atliekant daugiadiktorinį priebalsių diskriminavimą (Rudžionis, 1995).

Projektuojant atpažinimo sistemas, dažnai iškyla apmokymo imties dydžio ir požymių vektoriaus matiškumo santykio problema. Neretai yra sunku padidinti turimos duomenų bazės apimtį, o esant nedidelėms apmokymo duomenų imtims klasifikatoriaus parametru įverčiai gali būti nutolę nuo asimptotinių ir gaunamas atitinkamai mažesnis atpažinimo tikslumas.

Yra pasiūlyta daug būdų sumažinti tokio tipo problemų įtaką: požymių vektoriaus matiškumo sumažinimas išrenkant informatyviausius požymius, kovariacinės matricos struktūrizavimas arba jos reguliarizavimas.

Paprasčiausias ir populiariausias kovariacinės matricos reguliarizavimo būdas gaunamas pridėdant nedidelę konstantą prie jos įverčio:

$$S^R = S + \lambda I, \quad (3)$$

Tuomet tiesinė diskriminantinė funkcija užrašoma taip:

$$W^R = (S + \lambda I) \left(x^1 - x^2 \right), \quad (4)$$

kur x^1 ir x^2 yra abiejų klasių vidurkių vektoriai.

Siekiant išvengti skaičiavimo problemų, kovariacinės matricos apvertimo metu buvo pasiūlyta naudoti kovariacinės matricos pseudoinversiją:

$$S^P = TD^{-1}T,$$

kur $S = TDT'$ yra matricos S dekompozicija singulariniais skaičiais.

Neseniai Raudys (1998) pasiūlė naudoti pastumtą posūkio reguliarizavimą siekiant gauti geresnį kovariacinės matricos įvertį:

$$S^{SR} = T^\alpha (D + \lambda I) T^\alpha, \quad (5)$$

čia parametras α valdo matricos pasukimą, o parametras λ valdo įverčio postūmį.

Šiame darbe naudojamos reguliarizuotos diskriminantinės analizės idėjos sprendžiant fonetinio atpažinimo uždavinius.

Duomenų bazė. Yra žinoma, kad priebalsių akustinės charakteristikos yra stipriai veikiamos fonetinio konteksto, kuriame priebalsis buvo ištartas. Ypač didelę įtaką daro priebalsio kaimynystėje esantis balsis. Todėl formuojant fonetinių duomenų bazę tenka įvertinti koartikuliacijos reiškinį.

Šiuose eksperimentuose naudota nosinių (m , n) ir sonorinių (v , l) priebalsių duomenų bazė. Jie, o ypač m ir n , sudaro tarpusavyje panašių garsų grupes. Fonetinio konteksto modeliavimas atliktas kiekvieną iš šių priebalsių tariant trifone balsis–priebalsis–balsis, kur antrasis balsis visuomet kirčiuotas. Abu balsiai trifonuose vienodi, pvz., *ama*, *ini* ir t. t. Mūsų bazėje priebalsiai įrašyti 6 balsių (a , o , e , $ė$, u , i) aplinkoje, tačiau šiuose eksperimentuose naudoti tik

ištarimai balsių *a* ir *i* aplinkose. Šiuos balsius pasirinkome todėl, kad jie sudaro opoziciją vienas kitam – balsis *a* yra atviriausias, o balsis *i* – uždariausias.

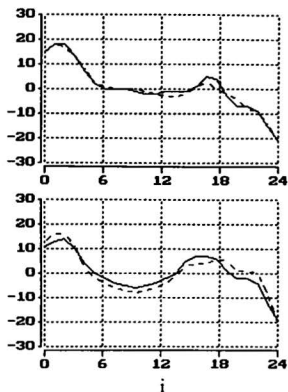
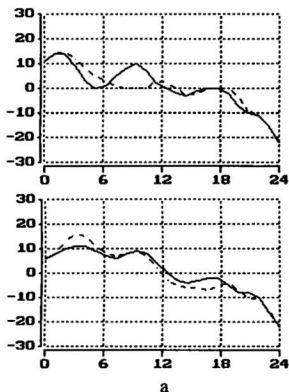
Įrašyti trifonai buvo transformuoti į difonus po ekspertinės analizės, atmetant pirmą balsį po to, kai buvo nustatytos jo ribos. Likęs difonas būdavo fonetiškai segmentuojamas į tris segmentus: stacionarias priebalsio ir balsio dalis bei perėjimą iš priebalsio į balsį.

Sudarant duomenų bazę dalyvavo 20 įvairaus amžiaus vyrų diktorių. Visi jie po 15 kartų ištardavo kiekvieną iš trifonų vienos įrašymo sesijos metu. Kadangi ne visi trifonai buvo ištarti gerai, tik 10 geriausių kandidatų palikta tolesnei analizei. Įrašai padaryti 10 kHz diskretinimo dažniu, naudojant standartinę personalinių kompiuterių garso plokštę. Kiekvienas atskaitymas yra 8 bitų, nes tai maksimalus naudotos garsinės plokštės leidžiamas skiltiškumas. Sudarant duomenų bazę, naudotas paprastas elektretinis mikrofonas.

Eksperimentų rezultatai. Minėjome, kad sonoriniai ir ypač nosiniai priebalsiai sudaro garsų, turinčių panašias spektrines charakteristikas, grupes. Iliustruojant tai, 1 pav. parodyti suvidurkinti per visus ištarimus priebalsių *m* ir *n* stacionarių dalių spektrai balsių *a* ir *i* kontekstuose.

Balsio *a* kontekste dažninių priebalsių charakteristikų tarpusavio skirtumai akivaizdesni, o balsio *i* kontekste spektrinės stacionarių dalių charakteristikos yra labai panašios. Šis panašumas buvo aprašytas dar Fujimura'os darbe (Fujimura, 1962). Žinia, kad ir atskirti priebalsius tokiu atveju sudėtinga. Kita ypatybė, kurią iliustruoja šis grafikas – akustinė priebalsio realizacija priklauso nuo jo konteksto. Įvairių balsių aplinkoje skirtingų priebalsių charakteristikos skiriasi nevienodai, todėl galima tikėtis ir nevienodo atpažinimo tikslumo. Spektrinių savybių priklausomybė nuo konteksto sukelia poreikį apmokymo ir klasifikavimo procedūras organizuoti atskirai kiekvieno balsio aplinkoje.

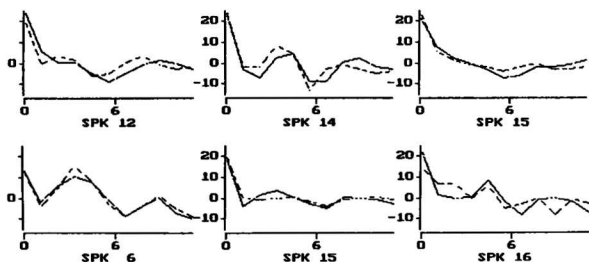
Suvidurkinti per visus ištarimus stacionarių dalių spektrai neparodo kito atpažinimą apsunkinančio veiksnio – didelės to paties garso spektrinių charakteristikų įvairovės tarp skirtingų diktorių. 2 pav. parodyti trijų tarpusavyje labiausiai besiskiriančių diktorių to paties garso ištarimų suvidurkinti spektrai. Analizė parodė, kad to paties kalbėtojo skirtingų ištarimų parametrų išsibarstymas nėra didelis, tuo tarpu vidutinis standartinis



1 pav. Suvidurkinti priebalsių m (ištisinė linija) ir n (brūkšninė linija) spektrai balsių a ir i kontekstuose. Apatiniuose grafikuose – diferencijuoto signalo spektrai

to paties garso parametrų išsibarstymas tarp skirtingų diktorių visuomet yra 2,5–3 kartus didesnis.

Ieškodami būdų, kaip pagerinti šių priebalsių diskriminavimą, atlikome skirtingų požymių tipų efektyvumo lyginamąjį tyrimą. Tyrimas apėmė tiek



2 pav. Trijų labiausiai besiskiriančių diktorių priebalsių m (ištisinė linija) ir n (brūkšninė linija) spektrai (viršuje – a kontekste, apačioje – i kontekste)

filtrinių ir keptrinių požymių efektyvumo palyginimą, tiek skirtingų požymių tipų sujungimo galimybes, tiek skirtingus fonetinius reiškinius aprašančių požymių vektorių panaudojimo galimybes. Eksperimentams naudota perklasifikavimo metodika, kai visi duomenys naudojami apmokymui, o po to – testavimui. Tai leidžia greitai įvertinti potencialias požymių galimybes atpažinimui ir planuoti tolesnius eksperimentus. Klasifikuota tiesinės diskriminančios analizės metodu.

1 lentelėje pateikti šio eksperimento rezultatai – diskriminuojant gautų klaidų skaičius procentais. Čia keptrinius požymius atstovavo LPCC tipo požymiai. C12 ir F12 požymiuose naudotas tik stacionarios priebalsių dalies aprašymas, C12+F12 požymiams vektoriai buvo sudaryti iš keptrinių ir filtrinių stacionarios dalies požymių, o C12T ir F12T požymiams naudoti vektoriai, sudaryti atitinkamai iš keptrinių arba filtrinių stacionarios dalies ir perėjimo požymių.

1 lentelė. Skirtingų tipų požymių lyginimas priebalsių diskriminavimui įvairiuose fonetiniuose kontekstuose

	a	o	e	ė	u	i	vidurkis
C12	13,0	17,0	10,3	7,7	19,1	14,8	13,9
F12	6,3	6,5	8,9	10,1	12,3	18,5	12,1
C12+F12	6,2	6,0	6,2	6,8	10,8	12,0	8,3
C12T	9,5	8,2	4,8	5,0	11,5	6,7	7,6
F12T	3,2	4,5	2,8	3,1	6,8	11,5	5,8

Analizuojant pateiktus rezultatus gali būti padaryta keletas pastabų. Savo efektyvumu filtrinės analizės būdu gauti požymiai pranoko keptrinius požymius. Atvirųjų balsių (*a, o, e*) aplinkoje tikslumo padidėjimas panaudojus filtrinius požymius yra ypač ryškus (neretai daugiau nei 2 kartus). Išimtį čia sudaro tik uždaras balsis *i*, kai keptriniai požymiai savo efektyvumu pranoko filtrinius. Kita svarbi išvada – perėjimo tarp garsų panaudojimas klasifikuojant leidžia gerokai sumažinti klaidų skaičių. Šis sumažėjimas matyti visų balsių kontekste ir visiems požymių tipams. Klasifikavimo klaidų skaičių šiek tiek leido sumažinti ir filtrinių bei keptrinių požymių sujungimas, o tai

rodo, kad šie abu požymių tipai turi tarpusavyje nekoreliuotos naudingos informacijos.

[vertinę šio tyrimo rezultatus, atlikome daugiadiktorinio priebalsių atpažinimo testus, taikydami perstumto posūkio maksimizavimo algoritmą. Daugiadiktorinis atpažinimas reiškia, kad buvo taikyta slenkančio egzamino metodika: apmokymo metu vieno diktoriaus ištariai laikinai būdavo pašalinami iš duomenų bazės, likę naudoti klasifikatoriaus parametrų nustatymui (mokymui), o po to nedalyvavusio diktoriaus medžiaga naudojama testavimui. Atskiriems diktoriams gauti atpažinimo tikslumo rezultatai suvidurkinti. Eksperimentams naudoti filtriniai ir keprstiniai (tiek MFCC, tiek LPCC) požymių vektoriai sudaryti iš stacionarios dalies ir perėjimų iš priebalsio į balsį, remiantis ankstesnio eksperimento nustatytais dėsningumais. Eksperimento rezultatai pateikti 2 lentelėje.

2 lentelė. Priebalsių diskriminavimo rezultatai panaudojant įvairius klasifikavimo metodus (FSH – Fišerio, LH – dichotominis klasifikatoriai, RM – įprastinė reguliarizuota matrica, PPM – perstumto posūkio maksimizacija)

	a			i		
	FLT	MFCC	LPCC	FLT	MFCC	LPCC
FSH	14,7	-	32,5	22,2	-	24,2
LH	14,7	27,5	27,4	26,9	33,5	23,2
RM	9,2	19,5	24,6	18,5	25,1	20,6
PPM	6,6	13,1	17,2	12,1	17,6	14,1

Atliekant daugiadiktorinio atpažinimo eksperimentus, naudojant filtrinius požymius taip pat gauti geriausi rezultatai. Vertinant klasifikavimo metodų efektyvumą matyti, kad pastumto posūkio maksimizavimo algoritmas leidžia žymiai sumažinti atpažinimo klaidą skirtingų balsių kontekste ir visiems požymių tipams. Balsio *a* kontekste klaida sumažėja daugiau nei 2 kartus, o balsio *i* kontekste klaida sumažėja beveik 2 kartus lyginant su įprastu Fišerio klasifikatoriumi. Tokį patį uždavinį sprendėme ir naudodami tolydžiąsias paslėptas Markovo grandines. Naudojant 3 būsenų HMM struktūrą, išėjimo tikimybes ir MFCC tipo požymius modeliuojant Gausiniu skirstiniu bei padarius prielaidą, kad kovariacinė matrica yra įstriža, gauta 44 proc. klaidų balsio *a* kontekste ir 57 proc. klaidų balsio *i* kontekste. Galima teigti,

kad naudojant filtrinius požymius, įvertinant koartikuliacinius požymius ir taikant pažangius klasifikavimo metodus diskriminavimo klaida fonetiškai sudėtinguose kontekstuose gali būti žymiai sumažinta.

Išvados. Fonetinis atpažinimas turi daug pranašumų lyginant su labiau įprastais atpažinimo metodais, paremtais pjūvių analize. Organizuojant atpažinimą fonetiniu principu galima: 1) modeliuoti koartikuliacinius efektus ir panaudoti juos aprašančius požymius; 2) fonetinių vienetų diskriminavimo metu naudoti progresyviuosius statistinio klasifikavimo algoritmus. Fonetinių vienetų atpažinimą pagerinti galima dviem kryptimis: parenkant efektyvesnius požymius, geriau aprašančius fonetinius reiškinius, ir panaudojant geresnius klasifikavimo algoritmus. Parodyta, kad koartikuliacinių požymių naudojimas leidžia sumažinti atpažinimo klaidą. Naudojant pastumto posūkio maksimizacijos algoritmą, diskriminavimo klaida gali būti dar labiau sumažinta. Gautas priebalsių *m*, *n*, *v*, *l* atpažinimo tikslumas (6,6 proc. balsio *a* kontekste ir 12,1 proc. balsio *i* kontekste) yra žymiai didesnis negu naudojant bazinės struktūros HMM modelį.

Literatūra

1. Blackburn Ch. S. Articulatory Methods for Speech Production and Recognition. Ph.D. thesis, CUED, December 1996.
2. Davis B., Mermelstain P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences // IEEE Transactions on Acoustic, Speech and Signal Processing, Vol. 28, No 4, August 1980.
3. Deller J., Proakis J., Hansen J. Discrete-Time Processing of Speech Signals. Macmillan, N. Y., 1993.
4. Deng L., Aksmanovic M. Speaker-independent Phonetic Classification Using Hidden Markov Models with Mixtures of Trend Functions // IEEE Transactions on Speech and Audio Processing, Vol. 5, No 4, July 1996.
5. Domatas A., Rudzionis A. Towards More Reliable Automatic Recognition of the Phonetic Units // Proceedings of the XIIth International Congress of Phonetic Sciences. Aix-En-Provence, France, 1991, Vol. 4.
6. Fujimura O. Analysis of Nasal Consonants // Journal of Acoustic Society of America, Vol. 34, No 12, December 1962.
7. Lippman R. Speech Perception by Humans and Machines // Proceedings of NATO ASI Workshop, King Island, July 1996.
8. Loizou P. C., Spanias A. S. High Performance Alphabet Recognition // IEEE Transactions on Speech and Audio Processing, Vol. 4, No 6, November 1996.
9. Markel J., Gray A. Linear Prediction of Speech. Springer, N. Y., 1976.

10. Noreika S., Rudžionis A. Phoneme-like Model of Speech Signal // Proceedings of XIIth International Congress of Phonetic Sciences. Vol. 4, Aix-en-Provence, 1991.
11. Poritz A. Hidden Markov Models: A Guided Tour // Proceedings of ICASSP, 1988.
12. Rabiner L. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition // Proceedings of IEEE, Vol. 77, February 1989.
13. Raudys Š. Scaled Rotation Maximization // Proceedings of 14th Conference on Pattern Recognition. Brisbane, August 1998.
14. Rudžionis A. Recognition by Averaged Template // 1st COST249 MC Meeting. Brussels, September 1994.
15. Rudžionis A., Rudžionis V. Consonant and speaker discrimination by consonant-vowel diphone components // Minutes of COST250 4th MC meeting, Stockholm, June, 1996.
16. Rudžionis V. Įvairių klasifikatorių ir analizatorių palyginimas daugiadiktoringame fonemų atpažinime // Informacinės technologijos, Konferencijos pranešimų medžiaga, KTU, 1995.
17. Schmid P. Explicit N-Best Formant Features for Segment-Based Speech Recognition // Ph. D. thesis, OGI, October 1996.
18. Valtchev V. Discriminative Methods in HMM-based Speech Recognition // Ph. D. thesis, CUED, March 1995.
19. Young S. Large Vocabulary Speech Recognition // Signal Processing Magazine, Nr. 4, 1996.

[teikta 1998 m. liepos mėn.]

Consonant Discrimination in Consonant-Vowel Diphones

Summary

The paper addresses problem of improvement of nasal and sonorant discrimination in different vowels environments. Both selection of more appropriate features for discrimination of phonemes and more advanced training has been investigated. Selection of more appropriate features has been performed by checking different signal analysis algorithms (comparing filterbank analysis and cepstrum procedures) and combining different acoustic events describing features (evaluation of coarticulation). Advanced pattern classification algorithms such as scaled rotation maximization and regularized discriminant analysis were applied to phoneme discrimination task also. It has been shown that recognition accuracy depends strongly on the acoustic environment of consonant. Best results were observed in the environment of open vowel a while closed vowel i is more complicated one. The evaluation of these factors allowed to reduce discrimination error rate from 25–30 % to 6.6% in the context of open vowel a and 12.1% in the context of closed vowel i.