Problemi pri vključevanju ekspertnega lingvističnega znanja v akustično modeliranje

Matjaž Rodman

Laboratorij za interaktivne sisteme
Naravoslovnotehniška fakulteta
Univerza v Ljubljani
Snežniška 5, 1000 Ljubljana, Slovenija
matjaz.rodman@ntf.uni-lj.si

Povzetek

Metoda odločitvenih dreves je zelo znana in pogosto uporabljena metoda v procesu gradnje sistemov za avtomatsko razpoznavanje govora. Hkrati pa je ta metoda že tako ustaljena, da je ponavadi ta del procesa obravnavan kot že zaključeno poglavje. V tem članku je metoda odločitvenih dreves ponovno vzeta pod drobnogled vključno s parametri, ki vplivajo na postopek izgradnje odločitvenega drevesa. Članek opozori na nekatere detajle, ki so pri uporabi te metode pogosto spregledani, vendar vplivajo na uspešnost sistemov razpoznavanja govora. Podana je tudi ena od možnih razlag, zakaj do teh vplivov pride.

Abstract

The method of decision trees is a very well known and a very frequently used method in the process of building automatic speech recognition systems. At the same time, the method is so accepted that this part of the process is usually dealt with as a finished chapter. In this article the method of decision trees is again taken under thorough consideration, including the parameters which influence the process of building the decision tree. The article points out some details which are frequently overlooked when using this method, but they influence the success of the speech recognition systems. There is also one possible explanation given why these influences appear.

1. Uvod

V naravnem govoru lahko velik del variabilnosti govornega signala pripišemo kontekstni variabilnosti. Ker naši govorni organi niso sposobni nenadnih in velikih sprememb v gibanju, se to odraža na akustični realizaciji posameznega fona, ki je zelo odvisna od pozicije govoril pred in po izgovorjavi le-tega. Temu pojavu pravimo koartikulacija. V sistemih za razpoznavanje govora, ki temeljijo na prikritih Markovih modelih (PMM), lahko to lastnost govora uspešno modeliramo. To storimo tako, da za vsak fon ustvarimo modele, ki predstavljajo fon v poljubnem levo-desnem kontekstu. Takšni modeli tipično vsebujejo tri oddajna stanja in jim pravimo trifonski modeli. Vendar z ustvarjanjem različnih modelov za vse možne kontekste postane število le-teh zelo visoko. S tem se pojavi problem natančne ocenitve parametrov, saj bi za tako veliko množico modelov potrebovali tudi ogromno količino dobro uravnoteženih učnih podatkov v obliki izgovorjav s pripadajočo transkripcijo, ki pa pogosto niso na voljo. Za reševanje tega problema se ponavadi poslužujemo tehnik povezovanja parametrov posameznih modelov. Največkrat uporabljeni metodi sta podatkovno vodena metoda in pa metoda odločitvenih dreves. Ti metodi nam omogočata, da se za ocenitev določenih parametrov "podobnih" modelov uporabljajo isti učni podatki. Vključitev teh metod v postopek gradnje sistema za avtomatsko razpoznavanje govora je tako pogosto uporabljena, da se zdi že skoraj samoumevna. Ta prispevek pa ponovno odpira analizo metode odločitvenih dreves ter predstavi probleme podajanja ekspertnega lingvističnega znanja pri tej metodi.

Tako nam članek v 2. poglavju predstavi tehnike povezovanja parametrov PMM in podrobno prikaže metodo odločitvenih dreves ter parametre, ki vplivajo na izgradnjo odločitvenega drevesa. V 3. poglavju so opisani načini določanja skupin fonov ter problemi, povezani s tem. V 4. poglavju so opisani poskusi, izvedeni z različnimi definicijami skupin fonov, ki jim sledijo rezultati v 5. poglavju. Poglavje 6 predstavi zaključke in postavi smernice za nadaljnje delo.

2. Tehnike povezovanja parametrov PMM

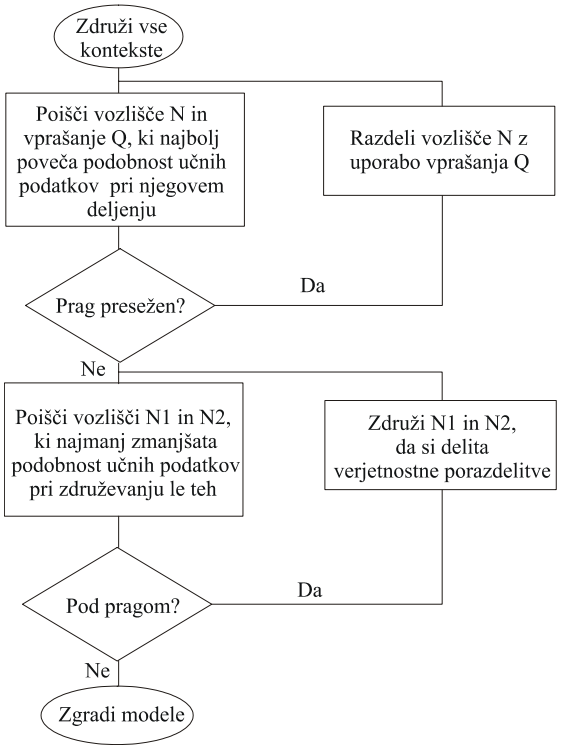
Kot je že omenjeno v uvodu, naletimo pri gradnji sistemov za razpoznavanje govora z velikim besediščem (slovarjem) na problem velikega števila modelov trifonov in ponavadi neuravnotežene in nezadostne količine učnih podatkov. V teh primerih se ponavadi poslužujemo ene izmed dveh metod vezanja parametrov. Pri podatkovno vodeni metodi najprej zgradimo ločene trifonske modele za vse kontekste, ki se pojavijo v učnih podatkih. Nato modele razbijemo tako, da vsako stanje predstavlja svojo skupino. Velikost skupine se definira kot največja razdalja med katerimakoli stanjema v določeni skupini. Sledi združevanje skupin. Najprej se združita tisti dve skupini, ki sta si najbolj podobni; to pomeni, da združeni ustvarita najmanjšo možno novo skupino. Ta postopek se ponavlja, dokler velikost največje skupine ne preseže postavljenega praga ali pa število skupin pade pod določeno vrednost. Slabost podatkovne metode je, da ustvari modele le za tiste trifone, ki se vsaj enkrat pojavijo v učnih podatkih. Metoda odločitvenih dreves pa v nasprotju s podatkovno vodeno metodo uporabi podano ekspertno lingvistično znanje za definiranje modelov tudi tistih trifonov, ki se nikoli ne pojavijo v učnih podatkih. To znanje je ponavadi podano v obliki datoteke, kjer so navedene skupine fonov za določen jezik.

2.1. Metoda odločitvenih dreves

Metoda odločitvenih dreves temelji na ustvarjanju binarnega drevesa, kjer se na vprašanje v vsakem vozlišču odgovarja le z "da" ali "ne". Torej so pri tej metodi na

začetku vsa "izbrana stanja" združena v korenu drevesa. Glede na postavljena vprašanja se skupine stanj razdeli. Ta postopek se nadaljuje, dokler ne pridemo do listov drevesa. Vsa stanja, ki se nahajajo v istem listu, se nato združijo in učijo iz istih učnih podatkov.

Zaradi lažje predstavitve samega postopka gradnje odločitvenega drevesa predpostavimo, da imamo trifonske modele, ki so zgrajeni iz treh stanj. Preden se postopek gradnje odločitvenega drevesa lahko začne, moramo zgraditi modele za vse trifone, ki se pojavijo v učnih podatkih. Kot je bilo že omenjeno, so na začetku gradnje drevesa vsa "izbrana stanja" združena v korenu drevesa. Izbrana stanja so tipično vsa začetna, vsa srednja ali vsa končna stanja določenega fona. To pomeni, da imamo na začetku N korenskih vozlišč, kjer je N število fonov pomnoženo s številom stanj, s katerimi je predstavljen vsak model. Iz teh N korenskih vozlišč se zgradi tudi N odločitvenih dreves.



Slika 1: Algoritem za izgradnjo odločitvenega drevesa (Woodland et al., 2000)

Pri gradnji dreves se izmed omejenega nabora vprašanj za vsako vozlišče izbere tisto vprašanje, ki razdeli vozlišče tako, da se lokalno najbolj poveča podobnost učnih podatkov. Za merjenje podobnosti podatkov se uporablja ocena logaritemske podobnosti (log likelihood), (Young et al., 2000). Deljenje vsakega vozlišča v dve novi povzroči povečanje logaritemske podobnosti podatkov v novih vozliščih, saj za opis (modeliranje) iste količine podatkov uporabimo dvakrat več parametrov. Postopek deljenja vozlišč v odločitvenem drevesu se ustavi šele, ko porast logaritemske podobnosti podatkov, pri deljenju določenega vozlišča, pade pod določen prag. Nazadnje se izračuna še, kakšen bi bil upad logaritemske podobnosti pri združevanju posameznih listov drevesa, ki pripadajo različnim staršem (vozliščem iz katerih izhajajo). Vsi pari listov, ki bi se jim z združevanjem zmanjšala logaritemska podobnost za manj kot je to določeno z že zgoraj omenjenim pragom, se nato združijo. Seveda se lahko tako združujejo le listi, ki pripadajo istemu drevesu. Algoritem, ki se uporablja pri izgradnji odločitvenega drevesa, je prikazan na sliki 1.

2.2. Parametri, ki vplivajo na izgradnjo odločitvenega drevesa s programskim paketom HTK

Preden se lahko začne avtomatski postopek gradnje odločitvenih dreves, je potrebno definirati posebno datoteko (za lažjo razlago jo poimenujmo `destree.hed`) z vsemi določenimi fonetičnimi vprašanji, ukazi za ustvarjanje korenskih vozlišč dreves in vrednostmi pragov. HTK uporablja določilo, da prikrit Markov model z imenom *a-b+c* predstavlja kontekstno odvisno verzijo fona *b*, ki se uporablja, ko je njegov levi kontekst fon *a* in desni fon *c*. Vsako vprašanje, postavljeno znotraj odločitvenega drevesa, ima obliko "Ali je levi (desni) kontekst pripadnik fonetične skupine M?", kjer kontekst pomeni neposrednega levega ali desnega soseda fona, na katerega se glasi vprašanje. Vsakemu vprašanju je zaradi pripravnosti dodeljeno ime. Tako na primer vrstica

```
QS "L_SL_Nasal" {m-*, n-*}
```

pomeni vprašanje "Ali je levi kontekst nosnik?", pri čemer levi kontekst fona pripada skupini slovenskih nosnikov, kjer je skupina določena z elementoma m in n.

V našem primeru izdelave sistema za razpoznavanje govora pa je dovolj, da ustvarimo datoteko (za lažjo razlago jo poimenujmo `broad.cls`), v kateri definiramo vse skupine fonov določenega jezika. Nadalje se iz teh skupin v procesu izgradnje odločitvenega drevesa vprašanja ustvarijo avtomatsko. Definicije skupin fonov izgledajo tako:

```
SL_Stop p b t d k g
SL_Nasal m n
SL_Vowel i: E: e: a: O: o: u: i e E a O o u @
```

Poleg vprašanj, če levi ali desni kontekst fona pripada katerikoli od teh skupin, pa se avtomatsko ustvarijo tudi vprašanja, kjer je vsak fon obravnavan kot samostojna fonetična skupina.

Ukazi za definiranje korenskih vozlišč dreves v datoteki `destree.hed` izgledajo tako:

```
TB 350.0 ST_a_2 {(a,*-a+*,a+*,*-a).state[2]}
```

Zgornji ukaz nam pove, da naj se v korenskem vozlišču združijo vsa začetna stanja vseh prikritih Markovih modelov fona *a*. To vozlišče se bo v procesu gradnje odločitvenega drevesa delilo glede na definirana vprašanja vse dokler porast logaritemske podobnosti, pri deljenju posameznega vozlišča, ne pade pod prag 350.0. Torej se z ukazom TB določi tudi minimalni porast logaritemske podobnosti, ki je potreben, da se določeno vozlišče sploh lahko deli. Hkrati pa je s tem določen tudi prag, ki vpliva v zadnji fazi izgradnje drevesa na odločitev, katere liste bomo združili. Za vsak par listov, ki pripada istemu drevesu, ne pa tudi istim staršem, se izračuna, kakšen bi bil upad logaritemske podobnosti pri

njunem združevanju. Če je ta upad manjši od vrednosti praga, določenega z ukazom TB, se ta dva lista združita.

Če bi odločitveno drevo gradili le na osnovi praga, določenega z ukazom TB, bi naleteli na problem, da bi večina zunanjih stanj modelov hotela ustvariti svoje liste, za katere pa ne bi obstajalo dovolj učnih podatkov za dobro ocenitev njihovih parametrov. To preprečimo z uporabo ukaza RO v datoteki `destree.hed`, kot nam prikazuje spodnji primer.

```
RO 100.0 state.cnt
```

Ukaz RO nam določa še en prag, ki se upošteva pri izgradnji odločitvenega drevesa. Določa nam, koliko učnih podatkov mora obstajati za določen list v drevesu, da lahko ta list samostojno predstavlja določeno skupino stanj. Oceno količine učnih podatkov (v zgornjem primeru datoteko z imenom `state.cnt`) pridobimo med procesom preračunavanja trifonskih modelov pred postopkom spajanja le-teh.

Ocena količine učnih podatkov se upošteva že pri izračunu porasta logaritemske podobnosti podatkov pri deljenju določenega vozlišča. Ta izračun je v programskem paketu HTK implementiran tako, da če je ocena količine učnih podatkov premajhna (pod določenim pragom), postane vrednost porasta logaritemske podobnosti podatkov enaka nič, kar prepreči nadaljnje deljenje določenega vozlišča (Odell, 1995).

Kot glavna prednost postopka odločitvenih dreves pred podatkovno vodenim postopkom povezovanja parametrov je bilo omenjeno dejstvo, da nam ta postopek omogoča natančno modeliranje trifonov, ki se ne pojavijo med učnimi podatki. To storimo npr. z ukazom

```
AU fulltri.lis
```

kjer v tem primeru predstavlja `fulltri.lis` datoteko, v kateri so napisana imena vseh trifonov, ki jih želimo uporabljati v določenem sistemu razpoznavanja govora. Če je v tem seznamu kakšen trifon, za katerega model še ni določen, se za ta trifon začne preiskovanje odločitvenega drevesa. Med preiskovanjem se odgovarja na zastavljena vprašanja o kontekstu tega trifona, dokler ne pridemo skozi drevo do končnih listov, ki nam tako predstavljajo stanja modela novega trifona.

3. Načini določanje skupin fonov za proces izgradnje odločitvenega drevesa

Z definiranjem fonetičnih skupin nam je v metodi odločitvenih dreves omogočeno vnašanje ekspertnega lingvističnega znanja v proces gradnje akustičnih modelov. Najbolje je, če lahko to delo opravi ekspert s področja jezikoslovja. Problem pa nastane, ko je potrebno določiti skupine fonov za nove jezike ali pa za podatkovne baze z različnim naborom fonov. Ena od možnosti reševanja tega problema je podatkovno voden postopek kreiranja teh skupin (Žgank et al., 2003). Tu pa se pojavi vprašanje, kako uporaben je ta postopek pri večjezikovnih sistemih za razpoznavanje govora ter pri podatkovnih bazah, kjer je nivo šuma visok (telefonski govor). V teh primerih lahko pride do velike zamenljivosti med akustičnimi modeli, ki predstavljajo različne fone, kjer predvsem izstopa model fona *s* (Iskra et al., 2001). Druga možnost reševanja tega problema je uporaba SPE (The Sound Pattern of English), (Chomsky et al., 1968) teorije pri definiranju skupin fonov (Rodman et al., 2002). Vendar ima tudi ta metoda svojo slabost, in ta je vprašanje razvrščanja diftongov.

Postavlja pa se tudi vprašanje, kako pomembno je sploh določiti "prave skupine fonov" in ali bi nam naključno generirane skupine fonov močno pokvarile rezultat pravilnega razpoznavanja govora (Žgank et al., 2003), (Rodman, 2002). Predpostavimo primer, kjer bi izdelali skupine, ki bi vsebovale vse možne kombinacije fonov določenega jezika, nato pa prepustimo postopku gradnje odločitvenega drevesa, da izbere najbolj primerne izmed njih po svojih lastnih kriterijih. Tako bi proces gradnje odločitvenega drevesa izbral za tvorbo vprašanj le zanj pomembne skupine in izpustil nepomembne. Takšna zamisel se porodi zaradi razlage, kako definirati vprašanja za odločitveno drevo v knjigi The HTK book (Young, 2000), ki pravi: "Nobene škode ni, če definiramo nekaj dodatnih nepotrebnih (nepomembnih) vprašanj, saj bodo tista, ki se ugotovijo kot nepomembna za podatke, ignorirana". Hkrati pa se je postavilo tudi vprašanje, ali je določitev vrstnega reda skupin fonov v datoteki `broad.cls` pomembna. Na ti dve vprašanji poskušam odgovoriti v naslednjih poglavjih.

4. Metodologija

4.1. Opis testnih sistemov za avtomatsko razpoznavanje govora

Kot del projekta COST 249 je bil ustvarjen sistem za razpoznavanje govora, ki je posebej namenjen uporabi na podatkovni bazi SpeechDat(II) (Lindberg et al., 2000). Zgrajen je bil z namenom, da služi kot referenčni sistem pri raziskavah večjezikovnih sistemov za razpoznavanje govora. Za izgradnjo sistema se uporabljajo HTK orodja (Young et al., 2000) ter tudi nekatera dodatna orodja, ki so bila napisana v okviru projekta SpeechDat(II). Vse procedure za izdelavo in testiranje sistema so napisane v PERL skriptnem jeziku in so prosto dosegljive na Refrecovi spletni strani (Refrec, 2000). Tu najdemo tudi rezultate testov, ki so bili narejeni na ostalih podatkovnih bazah SpeechDat(II).

Za testiranje domnev, navedenih v 3. poglavju, sem tako ustvaril referenčni sistem za razpoznavanje govora, ki pa se je šele pri procesu ustvarjanja odločitvenih dreves razdelil na dva različna sistema. Tako je bilo poskrbljeno, da so bili trifonski modeli in tudi vse ostale datoteke, ki se ustvarijo pred samim procesom gradnje odločitvenega drevesa, identične za oba sistema. Edina datoteka, ki se je razlikovala pri gradnji obeh sistemov, je bila datoteka `broad.cls`, v kateri so določene skupine fonov, ki jih v postopku gradnje sistemov podamo kot ekspertno lingvistično znanje. Tu je bilo uporabljenih 45 skupin fonov za slovenski jezik, ki so jih določili stokovnjaki s področja jezikoslovja, v okviru projekta COST 249. Vrstni red teh skupin sem v teh datotekah generiral naključno. Za fonetični zapis izgovorjav je bilo uporabljenih 46 slovenskih SAMPA simbolov. S tako določenima sistemoma za avtomatsko razpoznavanje govora so bili zagotovljeni pogoji, da je bil edini faktor, ki bi lahko vplival na razliko v uspešnosti razpoznavanja govora obeh sistemov, vrstni red določanja skupin fonov v datoteki `broad.cls`.

Na enak način sta bila zgrajena še dva sistema. Tu so bile v datoteki `broad.cls` določene skupine fonov, ki temeljijo na SPE teoriji. Tudi tu je bil vrstni red 174 skupin fonov naključno generiran. Zaradi narave teorije SPE se je pri fonetični transkripciji, v nasprotju z zgoraj opisanima sistemoma, uporabilo le 39 slovenskih SAMPA simbolov.

4.2. Testiranje

V okvirju projekta COST 249 je bilo ustvarjenih 6 testov, namenjenih ocenjevanju kvalitete razpoznavanja govora sistemov zgrajenih na osnovi podatkovne baze SpeechDat(II):

- Test da/ne (seznam vsebuje le 2 besedi)
- Test izoliranih števk (seznam vsebuje 10 besed)
- Test niza števk (seznam vsebuje 10 besed)
- Test ukaznih besed (seznam vsebuje 31 besed)
- Test imen krajev (seznam vsebuje 597 besed)
- Test fonetično uravnoteženih besed (seznam vsebuje 1491 besed)

Z namenom analize zgradbe čim večjega dela odločitvenega drevesa sem se odločil uporabiti le zadnja dva testa. Pri prvih štirih testih je število uporabljenih besed majhno in posledično je temu primerno tudi majhno število uporabljenih trifonov. Zato nam ti štirje testi dajo vpogled le v zelo ozko področje odločitvenega drevesa.

5. Rezultati

Testa imen krajev in fonetično uravnoteženih besed sem izvedel na vseh štirih sistemih v opisanih poglavju 4.1. Za ocene uspešnosti razpoznavanja govora sem uporabil akustične modele trifonov predstavljene z 32 Gausovimi verjetnostnimi porazdelitvami. V tabeli 1 so prikazani rezultati poskusov na referenčnih sistemih in na sistemih, ki temeljita na teoriji SPE.

	SPE-1	SPE-2	Ref-1	Ref-2
Test-FUB	16,31	15,78	17,51	18,04
Test-IK	7,65	6,63	6,12	7,14

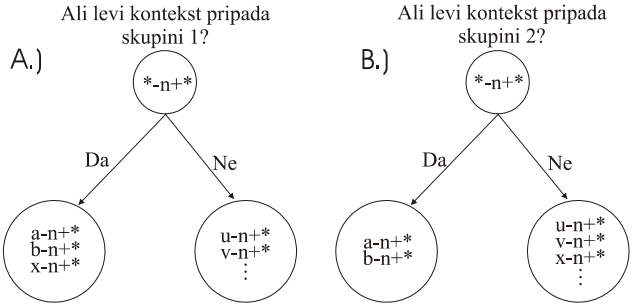
Tabela 1: Vrednosti napake razpoznavanja besed (%) za sistema, ki temeljita na teoriji SPE (SPE-1,2) ter za referenčna sistema (REF-1,2) pri testih fonetično bogatih besed (Test-FUB) in imen krajev (Test-IK)

Iz rezultatov v tabeli 1 je razvidno, da obstaja vpliv spremembe vrstnega reda skupin fonov na uspešnost sistemov za razpoznavanje govora. Vidimo da se ta vpliv prikaže pri testu fonetično uravnoteženih besed kot 0,53% spremembe absolutne napake razpoznavanja besed (3,25% relativno), pri testu imen krajev pa je ta sprememba celo 1,02% (13,3% relativno).

5.1. Analiza rezultatov

Analiziral sem dve, med postopkom gradnje odločitvenih dreves zgrajeni datoteki, `tied.lis` in pa `fulltri.des`. Datoteka `tied.lis` se ustvari, da nam po procesu gradnje odločitvenega drevesa poda seznam vseh zgrajenih modelov trifonov. V tej datoteki lahko vidimo, kateri trifoni si po postopku vezanja parametrov v celoti delijo isti akustični model. V datoteki `fulltri.des` pa je določena struktura odločitvenega drevesa za vsako oddajno stanje vsakega fona. Napisal sem program v skriptnem jeziku PERL, ki pregleda datoteko `fulltri.des` in nam pove, kolikokrat je bilo določeno vprašanje uporabljeno v drevesu ter kolikokrat na prvem, drugem, tretjem, četrtem in petem nivoju odločitvenega drevesa. Izkazalo se je, da se je vrsta vprašanj in njihova zastopanost po nivojih odločitvenih dreves spremenila pri spremembi vrstnega reda skupin fonov. Tudi v datoteki `tied.lis` je opaziti, da je v teh primerih prišlo do združevanja različnih trifonskih modelov. Tako je npr. v enem sistemu prišlo do popolne združitve trifona *E-k+o:* s trifonom *E-k+u:* v drugem sistemu pa s trifonom *i-k+o:*. Tudi primerjava zgrajenih modelov trifonov nam pokaže, da modeli niso identični.

Seveda se postavlja vprašanje, kako lahko le vrstni red definiranja skupin fonov tako močno spremeni zgrajene sisteme.



Slika 2: Vpliv vrstnega reda skupin fonov na odločitveno drevo

Glede na to, katera vprašanja se izbirajo v procesu gradnje odločitvenega drevesa in na kakšen način, sem našel možno razlago za takšno spremembo. Predpostavimo, da imamo vozlišče v drevesu, ki predstavlja prvo stanje prikritega Markovega modela generaliziranega[1] trifona **-n+** z vsemi možnimi levimi in desnimi konteksti. Denimo, da imamo v podatkovni bazi le učne primere trifona *n* z levimi konteksti *a, b, u, v* (*a-n+*, b-n+*, u-n+*, v-n+**) ter da imamo med skupinami fonov definirani tudi skupini skupina 1 = (*a, b, x*) ter skupina 2 = (*a, b*). Kot vidimo, je skupina 2 podmnožica skupine 1. Skupina 1 vsebuje še fon *x*, ki ga skupina 2 ne vsebuje. Ker pa vidimo, da se v učnih podatkih trifon *n* z levim kontekstom *x* ne pojavlja (*x-n+**), vemo, da imata ti dve skupini enako moč deljenja tega vozlišča, saj se ocena logaritemske podobnosti (Young et al., 2000) računa le za vzorce, ki se pojavijo med učnimi podatki. Torej ne glede na to, katera od teh dveh skupin razdeli to vozlišče, bo izračunan porast ocene logaritemske podobnosti v novih vozliščih enak. Predpostavimo še, da ti dve skupini med vsemi skupinami fonov najbolje delijo to vozlišče. Torej bo verjetno za delitev tega vozlišča izbrana tista izmed teh dveh skupin, ki bo prej definirana v datoteki `broad.cls`. Če je v datoteki skupina 1 definirana pred skupino 2, se bo vozlišče razdelilo na dve novi, kot je prikazano na sliki 2a, v nasprotnem primeru pa nastane situacija, prikazana na sliki 2b. V primeru, da skupina 1 deli to vozlišče, lahko iz

[1] Če en trifonski model predstavlja fon znotraj več različnih kontekstov, pravimo takim modelom generalizirani (*generalised*) oziroma posplošeni trifonski modeli.

slike 2a vidimo, da se bo prvo stanje modela trifona *n* z levim kontekstom *x* "učilo" iz istih učnih podatkov kot tudi prva stanja modelov trifonov *a-n+*, b-n+**, saj je levi kontekst *x* uvrščen v skupino 1. Če pa je skupina 2 definirana v datoteki `broad.cls` pred skupino 1, nam nastala situacija na sliki 2b prikazuje, da se bo prvo stanje modela trifona *n* z levim kontekstom *x* "učilo" iz istih učnih podatkov kot tudi prva stanja modelov trifonov *u-n+*, v-n+**. To torej pomeni, da se bo prvo stanje fona *n* z levim kontekstom *x* učilo iz popolnoma drugačnih podatkov kot bi se, če bi imeli ti dve skupini obrnjen vrstni red v datoteki `broad.cls`.

6. Zaključek

V prispevku je bilo pokazano, da obstaja vpliv določanja vrstnega reda skupin fonov na uspešnost sistemov za avtomatsko razpoznavanje govora, ki uporabljajo metodo odločitvenih dreves. Zaradi razlogov, opisanih v podpoglavju 5.1, je tudi jasno, da skupine fonov ne smejo biti definirane brez vsakršnega lingvističnega znanja in potem prepuščene avtomatičnemu procesu gradnje odločitvenega drevesa, da izbere najboljše med njimi. Tak način ne bi imel problemov s klasifikacijo trifonov, ki se pojavijo v učnih podatkih, vodi pa do nepravilne razvrstitve trifonov, ki se med učnimi podatki ne pojavijo. Ti trifoni bi bili na tak način razvrščeni v liste drevesa brez vsakršne lingvistične osnove. Na njihovo razvrstitev bi v tem primeru še najbolj vplival vrstni red definiranih skupin fonov v datoteki `broad.cls`. Zato bi bilo smiselno še enkrat premisliti o uporabi unije skupin fonov več jezikov za postopek izgradnje odločitvenega drevesa novega ciljnega jezika, ki se je uporabljal pri nekaterih večjezikovnih sistemih za razpoznavanje govora (Žgank et al., 2001), (Lindberg et al., 2000).

Potrebno pa bo bolj natančno oceniti, kako pomemben je lahko ta vpliv ob pojavljanju velikega števila trifonov, ki se ne pojavijo v učnih podatkih. Sprememba uspešnosti razpoznavanja govora v opisanih poskusih res ni velika. Vendar je treba upoštevati, da bi lahko ta sprememba ob drugačni razvrstitvi skupin še narastla. Hkrati pa, če upoštevamo dejstva, da gre ta napaka le na račun spremembe vrstnega reda skupin fonov v datoteki `broad.cls` ter da se pri testu fonetično uravnoteženih besed pojavi le 50 trifonov, ki se ne pojavijo med učnimi podatki (pri testu imen krajev celo manj), skupno pa imamo v odločitvenem drevesu definiranih 6954 trifonov, lahko vidimo, da ta del predstavlja le dobrih 0.7 % vseh trifonov. Iz rezultatov testov je razvidno, da to spremeni vrednost absolutne napake razpoznavanja besed za 0,53 % oz. 1.02 %. To pa ni zanemarljivo glede na tako majhno število trifonov, ki se ne pojavijo v postopku učenja sistema.

Vsekakor so ti rezultati bolj pomembni v raziskavah tistih jezikov, kjer nimamo dovolj učnih podatkov, saj je tam ekspertno vneseno lingvistično znanje eden ključnih faktorjev za uspešno razpoznavanje govora. Pri jezikih, kjer imamo kvalitetnih učnih podatkov dovolj, pa je tudi alternativna podatkovno vodena metoda povezovanja parametrov akustičnih modelov ustrezna in v takih primerih mogoče celo priporočljiva.

Najosnovnejši test, ki bo za potrditev omenjene teorije še nujno potreben, bo sestaviti testni set izgovorjav, kjer bodo besede vsebovale le tiste trifone, ki se pojavijo tudi znotraj učnih podatkov sistema za avtomatsko razpoznavanje govora. Nato bo potrebno ponoviti poskusa z naključno generiranima vrstnima redoma skupin fonov. Pri tako definiranem testu vrstni red skupin ne bi smel več vplivati na izgradnjo odločitvenega drevesa in s tem tudi na uspešnost takega sistema za avtomatsko razpoznavanje govora.

7. Literatura

Chomsky, N., in Halle, M. 1968. *The Sound Pattern of English.* Harper & Row, New York, Evanston, and London.

Iskra, A., Petek, B., in Brøndsted, T. 2001. Recognition of Slovenian Speech: Within and Cross–Language Experiments on Monophones using the SpeechDat(II), *V Proc. EUROSPEECH, European Conference on Speech Communication and Technology*. Aalborg, str. 2777-2780.

Lindberg, B., Johansen, F.T., Warakagoda, N., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., in Salvi, G. 2000. A Noise Robust Multilingual Reference Recognizer Based on SpeechDat(II). *V Proc. ICSLP, International Conference on Spoken Language Processing*, Beijing.

Odell, J.J. 1995. *The Use of Context in Large Vocabulary Speech Recognition.* Doktorsko delo, University of Cambridge. Queens' College.

Rodman, M., Petek, B., in Brøndsted, T. 2002. "SPE-Based Selection of Context-Dependent Units for Speech Recognition", *V Proc. LREC, Workshop on Portability Issues in Human Language Technologies*. Gran Canaria, str. 78-84.

Rodman, M., 2002. *Analiza izbire enot za razpoznavanje govora*. Magistrsko delo, Univerza v Ljubljani.

Žgank, A., Imperl, B., Johansen, F.T., Kačič, Z., in Horvat, B. 2001. Crosslingual Speech Recognition with Multilingual Acoustic Models Based on Agglomerative and Tree-Based Triphone Clustering. *V Proc. EUROSPEECH, European Conference on Speech Communication and Technology*. Aalborg, str. 2725-2728.

Žgank, A., Kačič, Z., in Horvat, B. 2003. Data driven generation of broad classes for decision tree construction in acoustic modeling. V *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*. Geneva.

Welcome to the Refrec homepage [online]. Telenor, 2000, obnovljeno 16.8.2000 [citirano 30. 6. 2004]. Dostopno na svetovnem spletu: <http://www.telenor.no/fou/prosjekter/taletek/refrec/>.

Woodland, P., in Evermann, G. 2000. *Speech Recognition for Information Access.* Elsnet TeSTIA Summer School. Chios. Greece.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., in Woodland, P. 2000. *The HTK Book (for HTK Version 3.0)*. Cambridge. Entropic Cambridge Research Laboratory.