Ocenjevanje govornih modelov: Razdalja med zaporedji govornih odsekov

Simon Dobrišek, Janez Žibert, France Mihelič, Nikola Pavešić

Univerza v Ljubljani, Fakulteta za elektrotehniko
Tržaška 25, SI-1000 Ljubljana
{simond,mihelicf,nikolap}@fe.uni-lj.si

Povzetek

Pri ocenjevanju govornih modelov se srečamo z zahtevo po primerjavi dveh zaporedij govornih enot. Zaporedji predstavljata različni razčlenitvi istega govora. Pri primerjavi želimo ugotoviti, na katerih mestih so razlike med zaporedjema v smislu zamenjav, vrivanj in izbrisov govornih enot, oziroma kolikšno je skupno število teh operacij na govornih enotah. Pri ocenjevanju primerljivosti samodejnega razčlenjevanja govora z razčlenjevanjem, ki bi ga izvedel človek, te operacije štejemo za napake. Ugotavljanje razlik med zaporedjema večinoma izvajamo s posebnim postopkom iskanja najboljše poravnave dveh zaporedij simbolov. Pri tem govorne enote obravnavamo le kot simbole. Postopek temelji na omenjenih operacijah na simbolih. Vsaki operaciji se priredi osnovna cena in najboljša poravnava je tista, za katero je potrebno izvesti zaporedje operacij za najmanjšo skupno ceno. V primeru, ko ima funkcija osnove cene le dve vrednosti, nič in ena, je skupna cena enaka najmanjšemu številu potrebnih operacij za poravnavo obeh zaporedij. Slabost takšne poravnave je v tem, da je pri istem najmanjšem skupnem številu operacij možnih več različnih zaporedij operacij. Zaradi tega je postopek neprimeren, če nas zanima, točno katere operacije so potrebe za poravnavo enega zaporedja simbolov v drugega. V članku govorne enote ne obravnavamo le kot simbole, temveč kot simbole s pripisanimi mejami časovnih odsekov. Tem govornim enotam pravimo govorni odseki. Postopek poravnave dveh zaporedij simbolov razširjamo v postopek poravnave dveh zaporedij simbolov s pripisanimi mejami časovnih odsekov. Cene določimo kot funkcijo simbolov in simbolom pripisanih mej časovnih odsekov. Uporaba razširjenega postopka je pokazala, da so rezultati primerjav glede na izvirni postopek skladnejši z rezultati, ki jih pri primerjavah istih zaporedij podaja človeški opazovalec. Prednosti predstavljene razširitve postopka zagovarjamo s tolmačenjem sprememb v podanih dendrogramih rojenja govornih enot in matrikah verjetnosti njihovih medsebojnih zamenjav pri samodejnem razpoznavanju.

1. Uvod

Določanju zaporedja govornih enot iz govora pravimo *razpoznavanje govora*. Uporaba računalniških govornih modelov, sestavljenih iz modelov govornih enot, omogoča samodejno razpoznavanje govora. Razpoznavanje izvajamo z iskanjem zaporedja modelov govornih enot, s katerim se najbolje poravna dani posnetek govora. S katerim zaporedjem modelov se bolje poravna dani govorom ocenjujemo z metriko, ki podaja količinsko sodbo o tem, kako blizu je model posamezne govorne enote odseku govornega signala.

Vsako zaporedje govornih enot predstavlja določeno razčlenitev govora. Samodejno razpoznavanje govora je tako tudi samodejno *razčlenjevanje govora* na zaporedje govornih enot. Govor samodejno razčlenjujemo z iskanjem najboljše proste ali vsiljene poravnave zaporedja modelov govorni enot z govorom. Pri prosti poravnavi dovoljujemo vsa zaporedja modelov govornih enot, ki so možna v danem govorjenem jeziku. Pri vsiljeni poravnavi pa vnaprej predvidimo določeno zaporedje modelov govornih enot za dani govor in nas zanima le še, kakšna točno je poravnava.

Govorne modele ocenjujemo z ugotavljanjem napak, storjenih pri samodejnem razčlenjevanju govora. Na prvi pogled je naloga preprosta. Referenčno razčlenitev govora moramo primerjati s samodejno razčlenitvijo tega istega govora. Pri primerjavi ugotavljamo, katere govorne enote se med zaporedjema ujemajo in katere so morebiti zamenjane, vrinjene ali izbrisane.

Do informacijo o razlikah med zaporedjema govornih enot pridemo z iskanjem najboljše poravnave obeh zaporedij. Postopek iskanja poravnave se od tistega pri poravnavi posnetkov govora z zaporedjim modelov govornih enot razlikuje v drugačni metriki. V postopek uvedemo nekaj osnovnih operacij na govornih enotah v obeh zaporedjih. Operacijam pravimo kar zamenjava, vrivanje in izbris govornih enot. Vsaki operaciji priredimo neko osnovno ceno (Rice et al., 1997), ki jo določimo iz mere podobnosti med govornimi enotami. Operaciji vrivanja in izbrisa obravnavamo kot operaciji zamenjave govorne enote z ničelno govorno enoto. Najmanjši skupni ceni poravnave dveh zaporedij pravimo tudi razdalja med zaporedjema.

Pogosto se zatečemo k poenostavitvi in izberemo ceno operacij, ki ima zgolj dve vrednosti. Vrednost nič pomeni, da sta govorni enoti enaki in vrednost ena, da nista. Najboljša poravnava je potem tista, ki zahteva najmanjše skupno število operacij s ceno ena, torej najmanjše skupno število zamenjav, vrivanj in izbrisov. Temu najmanjše skupno število opracij pravimu tudi Levenshteinova razdalja (Levenshtein, 1966). Razdalja ima znano slabost, da za več različnih poravnav dveh zaporedij govornih enot daje enako najmanjše skupno število potrebnih operacij.

Očitno je, da je mera podobnosti govornih enot, iz katere določamo osnovno ceno posameznih operacij, ključna za postopek iskanje najboljše poravnave dveh zaporedij. Osnovna cena je matematična funkcija dveh govornih enot in funkcijo lahko določimo šele potem, ko se odločimo, kaj sploh obravnavamo za govorne enote.

2. Govorne enote

Izhodiščna predpostavka pri tehniškem modeliranju govora je, da je vir govora vsaj v svojem jedru diskretne narave. To nam na svoj način potrjuje diskretni črkovni simbolni zapis govorjenih sporočil v govorjenem jeziku. Jedro vira govora tako razumemo kot diskretni informacijski vir, ki v zaporedju oddaja simbole iz končne abecede. Zaporedja oddanih simbolov predstavljajo sporočila, ki se na

izhodu govornega vira uresničujejo v zvočnem govornem signalu.

Uresničevanje posameznih zaporednih simbolov se odraža kot zaporedje razpoznavnih zvokov oziroma odsekov zvočnega govornega signala. Simboli se ne uresničujejo ponovljivo, torej kot vedno enako določeni zvoki, temveč se uresničujejo kot kratkočasovne nastavitve nekih simbolom pripisanih značilnosti signala. Zvočni govorni signal namreč poleg govorjenega sporočila nosi še informacijo o govorcu in njegovih govorilih. Govorci in delovanje njihovih govoril pa je neponovljivo, torej tudi zvočni govorni signal v naravi ni nikoli ponovljiv. Zato pravimo, da je zvočni govorni signal naključen.

Simboli predstavljajo *govorne enote* - gradnike sporočil govorjenega jezika. Čemu točno pravimo govorne enote je odvisno od tega, s kakšnim namenom in kako podrobno analiziramo govor. Pri jezikovnih analizah govorne enote večinoma obravnavamo le kot simbole, ki v zaporedjih tvorijo zapise sporočil z nekim pomenom, pri čemer pogosto odmislimo njihovo povezavo s funkcijo govoril in zvočnim govornim signalom.

Pri podrobnejši analizi, ki jo zahteva tehniško modeliranje govora, simbolom pridružujemo še pripadajočo zvočno plat. Govorjena sporočila ne zapisujemo le z zaporedji simbolov, temveč posameznim simbolom v zaporedju pripisujemo še odseke govornega signala ali vsaj časovne meje teh odsekov. Govornim enotam v obliki simbolov s pripisanimi časovnimi mejami odsekov pravimo *govorni odseki*.

Časovne meje med odseki ponavadi postavi človek ali samodejni računalniški sistem na mestih, za katera se predpostavi, da je najverjetneje že prišlo do konca uresničevanja prejšnjega in začetka uresničevanja naslednjega simbola v zaporedju. Zaradi koartikulacijskega fenomena in zveznosti govora smo bolj natančni, če trdimo, da so meje med odseki tam, kjer se predpostavi, da je najverjetneje že prišlo do prevladovanja uresničevanja naslednjega simbola nad zamirajočim uresničevanja prejšnjega simbola v zaporedju.

Za postopek iskanja najboljše poravnave dveh zaporedij govornih enot ni pomembno s točno kakšnimi simboli predstavljamo govore enote. Ponavadi imamo opravka s končno množico nekih osnovnih simbolov, s katerimi tvorimo sestavljene simbolne zapise govorjenih sporočil. Pri jezikovnih analizah so osnovni simboli *grafemi* ali črke. S temi osnovnim simboli tvorimo sestavljene simbolne zapise, kot so zlogi, besede, povedi in stavki.

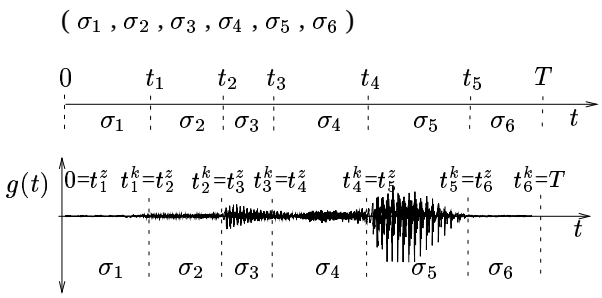
Pri analizi govora so osnovni simboli *fonemi*. Fonemska abeceda izvira iz grafemske in je določena predvsem po pomensko razločevalnem kriteriju v povezavi s slušno zaznavnimi akustičnimi kriteriji. Pri podrobnejši analizi govora, kjer upoštevamo koartikulacijski fenomen govora in spektralne značilnosti zvočnega govornega signala, množico fonemov razširimo v množico *alofonov*. Zvočno uresničenim fonemom in alofonom pravimo *foni*. To so s simbolom označeni odseki zvočnega govornega signala.

Govorna enota je torej, ali simbol, ali simbol s pripisanima mejama časovnega odseka, ali simbol s pripisanim odsekom govornega signala. Iskana funkcija mere podobnosti dveh govornih enot oziroma funkcija osnovne cene operacije na govornih enotah je lahko funkcija nečesa od naštetega. Za določitev funkcije moramo govorne enote in njihova zaporedja najprej podati v matematični obliki.

2.1. Zaporedje govornih enot

Najprej obravnavajmo govorne enote le kot simbole. Vzemimo, da je govorec izgovoril sporočilo, katerega zapis podamo v obliki zaporedje simbolov $(\sigma_1, \sigma_2, \cdots, \sigma_i \cdots, \sigma_N)$. Predpostavimo, da simbole izbiramo iz končne abecede simbolov $\sigma_i \in \Sigma$. Zaporedje simbolov predstavlja eno od možnih razčlenitev govora na zaporedje govornih enot.

Na splošno je lahko zapis sporočila poljubno vendar končno dolg. Končna zaporedja govornih enot zapišemo kot končna zaporedja simbolov iz abecede Σ. Vsa možna končna zaporedja simbolov tvorijo množico Σ^*.



Slika 1: Razčlenitve govora v zaporedja govornih enot.

Vzemimo dalje, da je v času s trajanjem T govorec z izgovorjavo sporočila zvočno uresničil prej zapisano zaporedje simbolov. Simbole je uresničil v zvočnem govornem signalu, ki ga zapišimo kot realno funkcijo časa $g(t)$, definirano za časovni odsek $[0, T]$.

Zamislimo si obstoj $N-1$ časovnih trenutkov $\{t_i\}$, oziroma točk na časovnem odseku $[0, T]$, ob katerih je najverjetneje prišlo do prevladovanja uresničevanja naslednjega simbola σ_{i+1} nad zamirajočim uresničevanjem predhodnega simbola σ_i. Časovni trenutki $\{t_i\}$ delijo časovni odsek $[0, T]$ in govorni signal $g(t)$ na zaporedje N pododsekov.

S simbolom označeni pododsek časa je govorni odsek, ki ga zapišimo kot trojico (σ_i, t_i^z, t_i^k). Trojico tvori simbol iz končne abecede Σ ter dve realni vrednosti t_i^z in t_i^k, ki predstavljata začetni in končni čas govornega odseka. Govorni odseki tvorijo množico

$$\Gamma = \left\{ (\sigma_i, t_i^z, t_i^k) \mid \sigma_i \in \Sigma\,,\ \ t_i^z, t_i^k \in \mathbb{R}_0^+\,,\ \ t_i^k > t_i^z \right\}.$$

V zapisu množice smo upoštevali, da je končni čas odseka vedno večji od začetnega in da ima čas le nenegativne vrednosti.

Za dani govor s trajanjem T se začetni in končni časi govornih odsekov lahko nahajajo le znotraj časovnega odseka $[0, T]$. Možni govorni odseki tako tvorijo manjšo množico

$$\Gamma_T = \left\{ (\sigma_i, t_i^z, t_i^k) \in \Gamma \mid\ t_i^z, t_i^k \in [0, T] \right\}.$$

Govor razčlenjujemo na zaporedje govornih odsekov. V zaporedju je začetek vsakega odseka je enak koncu predhodnega odseka. Zaporedje govornih odsekov zapišemo kot zaporedje trojic iz množice Γ_T. V zaporedju je N trojic, ki predstavljajo zaporedje N-tih govornih odsekov s skupnim

trajanjem T. Začetek prvega odseka je pri nič in konec zadnjega N-tega odseka pri T. Možna zaporedja N govornih odsekov v skupnem trajanju T tako tvorijo množico zaporedij

$$G_T^N = \{a_i = (\sigma_i, t_i^z, t_i^k) \in \Gamma \ , \ i = 1, \cdots, N \mid t_1^z = 0 \ ; \ t_i^z = t_{i-1}^k \ , \ 2 \leq i \leq N \ ; \ t_N^k = T\} .$$

Začetni in končni časi odsekov t_i^z in t_i^k se nanašajo na prej predpostavljene časovne trenutke t_i.

Govor s trajanjem T je lahko razčlenjen na zaporedje s poljubnim končnim številom govornih odsekov. Množico vseh možnih razčlenitev govora s trajanjem T je

$$G_T^* = \bigcup_{N \in \mathbb{N}} G_T^N \ . \tag{1}$$

Množico bi lahko razširili še v množico vseh možnih razčlenitev govora s poljubnim končnim trajanjem.

$$G^* = \bigcup_{T \in \mathbb{R}_0^+} G_T^* \ . \tag{2}$$

Podani matematični zapis zaporedja govornih odsekov zadošča za podajanje funkcije mere podobnosti med temi govornimi enotami. Smiselno podobno bi lahko matematično zapisali tudi zaporedje govornih enot v obliki s simbolom označenih odsekov govornega signala. Govorno enoto bi zapisali kot

$$(\sigma_i, g(t)|_{[t_i^z, t_i^k]}) \ . \tag{3}$$

V članku se ne posvečamo govornim enotam v tej obliki. Zapis smo podali zato, ker bomo pri podajanju funkcije mere podobnosti nakazali možnost posplošitve tudi na te govorne enote.

3. Razčlenjevanje govora

Zaporedja govornih enot pridobivamo z razčlenjevanjem posnetkov govora. Govor razčlenjuje človek na podlagi znanja o govoru in jeziku, ali računalnik, s katerim izvajamo postopek iskanja najboljše poravnave zaporedja modelov govorni enot s posnetim govorom.

3.1. Ročno razčlenjevanje posnetkov govora

Vsako zapisovanje govora je že njegovo razčlenjevanje. Bolj natančno razčlenjevanje zahteva poleg tvorjenja simbolnega zapisa izgovorjenega sporočila še določanje mej časovnih odsekov oziroma odsekov govornega signala, ki jih pripisujemo posameznim simbolom.

Pri predvajanju in poslušanju odsekov posnetega govora ter pri pregledovanju časovnega poteka odsekov posnetega govornega signala opazimo prisotnost značilnih kratkih odsekov signala. Opazovalcu bi mogli zadati nalogo, da s pomočjo primernega računalniškega orodja za urejanje zvočnih posnetkov poskuša ročno razčleniti dani govorni signal na zaporedje govornih enot.

Pri ročnem razčlenjevanju ugotovimo, da se bodo odločitve, kje v danem govornem signalu so razpoznavni odseki, kakšne so njihove značilnosti, katere govorne enote predstavljajo, predvsem pa kje so meje med zaporednimi odseki, pogosto precej razlikovale od opazovalca do opazovalca.

Najprej je to posledica omenjene naključne narave govornega signala. Razlike pa se pojavljajo tudi zaradi različnih vidnih predstavitev govornega signala, različnih slušnih in vidnih zaznavnih sposobnosti opazovalcev ter njihovega različnega znanja o zvočni plati govora. Najbolj pomemben dejavnik pri opazovalcih pa je njihovo poznavanju govorjenega jezika govorca, ki je tvoril opazovani posneti govor.

Praviloma pričakujemo, da vsi opazovalci posnetega govora poznajo jezik govorca, in jim vnaprej predložimo množico simbolov, med katerimi lahko izbirajo za predstavitev govornih enot. Vendar bodo preostale prej omenjene razlike povzročile še vedno znatne razlike med opazovalci v rezultatu ročnega razčlenjevanja.

Pogosto želimo analizirati razlike med opazovalci, ki ročno razčlenjujejo iste posnetke govora. Za takšno analizo moramo priskrbeti primerno računalniško orodje za pregledovanje in predvajanje posnetkov govora in za vnašanje časovnih mej med zaporednimi govornimi odseki. Orodje mora omogočati enostavno vnašanje simbolov, ki predstavljajo govorne enote. Za tovrstno analizo ter za čim večjo učinkovitost raziskovalnega dela smo v laboratoriju razvili primer takšnega orodja.

Razlike med dvema opazovalcema ugotavljamo s primerjavo razčlenitev istih posnetkov govora. Primerjati moramo časovne meje zaporednih odsekov govornega signala ter pripisane simbole. Potrebovali bi torej nek postopek, ki bi podal mero o tem, koliko se dve razčlenitvi istega govora na zaporedji govornih odsekov razlikujeta, in podatek o tem, kje v zaporedjih so razlike.

3.2. Samodejno razčlenjevanje govornega signala

Iz zbirke posnetkov govornih signalov tvorimo računalniški govorni model, ki ga sestavljajo modeli govornih enot. Za tvorjenje modelov potrebujemo poleg signalov vsaj še simbolne zapise posnetega govora. Tvorjeni govorni model nam omogoča izvajanje postopka iskanja najboljše poravnave zaporedja modelov govorni enot s poljubnim posnetim govorom. Rezultat poravnave je razčlenitev govora na zaporedje govornih enot.

Govorne modele primerjamo in ocenjujeno s statistično analizo napak samodejnega razčlenjevanja govora. Napake ugotavljamo s primerjavo samodejnih razčlenitev z referenčnimi. Oceno skupne napake podamo v obliki odstotka števila zamenjav, vrivanj in izbrisov, potrebnih za poravnavo referenčnih zaporedij govornih enot s samodejno pridobljenimi, proti vsoti števila vseh govornih enot v referenčni razčlenitvi in števila potrebnih vrivanj za poravnavo (Gibbon et al., 1997).

Pri iskanju čim boljšega govornega modela, tega ponavadi izboljšujemo korakoma. To pomeni, da poskušamo model izboljšati s spremembami postopkov ter s posameznimi posegi v strukturo in parametre modela. V vsakem koraku model ocenimo z izbrano kriterijsko funkcijo, ki podaja količinsko sodbo o tem, kako podoben je model viru, ki je rodil posnete govorne signale, torej kako podoben je model enemu ali več govorcem, ki so tvorili posneti govor.

Kakšno razčlenitev lahko pri danem govornem signalu pričakujemo je odvisno od predobdelave govornega signala, od strukture in parametrov modelov govornih enot ter nenazadnje tudi od postopka iskanja najboljše porav-

nave. Pričakovati je, da se bodo spremembe česarkoli od naštetega, odražala tudi v spremembah razčlenitev istega govornega signala.

Po vsakem posegu v model bi lahko izvedli postopek iskanja najboljše poravnave zaporedja modelov govornih enot z izbranimi posnetki govornih signalov ter analizirali njihove razčlenitve. Pri vsaki spremembi govornega modela lahko pričakujemo nekoliko drugačno razčlenitev. Primerjava razčlenitev istega govora po posegih v govorni model bi podala določen vpogled v trenutno stanje modela.

4. Primerjava dveh razčlenitev istega govora

Primerjavo dveh razčlenitev istega govora izvajamo s primerjavo dveh zaporedij govornih enot, ki se nanašata na isti govor. Razlike med zaporedjema določamo z iskanjem najboljše poravnave obeh zaporedij, pri čemer za poravnavo dovoljujemo omenjene operacije zamenjave, vrivanja in izbrisa govorne enote v enem ali drugem zaporedju. Postopek, ki bo predstavljen v nadaljevanju, je izpeljanka znanega postopka poravnave zaporedij simbolov. Rezultatu te poravnave pravimo tudi razdalja med zaporedjema simbolov.

4.1. Razdalja med zaporedjema simbolov

Razdaljo določata abecedi A in B ter funkcija osnovne cene $c : E \rightarrow \mathbb{R}_0^+$, kjer je $E = E_z \cup E_v \cup E_i$ množica operacij na simbolih in $\mathbb{R}_0^+$ množica nenegativnih realnih števil. Množica $E_z = A \times B$ je množica operacij zamenjav, $E_v = \{\epsilon\} \times B$ je množica operacij vrivanj in $E_i = A \times \{\epsilon\}$ množica operacij izbrisov. Simbol ϵ je ničelni simbol.

Obe abecedi in funkcija osnovne cene uvajajo funkcijo razdalje $d_c : A^* \times B^* \rightarrow \mathbb{R}_0^+$, ki preslika poljubni dve zaporedji simbolov ene in druge abecede v nenegativno realno število.

Določimo zdaj razdaljo med zaporedji $a^n \in A^n$ ter $b^m \in B^m$. Naj a_i označuje i-ti simbol v zaporedju a^n. Razdalja $d_c(a^n, b^m)$ je določena po rekurzivnem izrazu

$$d_c(a^n, b^m) = \min \left\{ \begin{array}{l} c(a_n, b_m) + d_c(a^{n-1}, b^{m-1}), \\ c(a_n, \epsilon) + d_c(a^{n-1}, b^m), \\ c(\epsilon, b_m) + d_c(a^n, b^{m-1}) \end{array} \right\}. \quad (4)$$

Rekurzivni izračun začnemo z $d_c(\epsilon, \epsilon) = 0$. S povratnim sledenjem vmesnih rezultatov rekurzivnega izračuna ugotavljamo potrebne operacije najboljše poravnave iz argumentov funkcije osnovne cene. Funkcija razdalje se lahko izračuna z dinamičnim programiranjem (Masek and Paterson, 1980). Izračun ima časovno zahtevnost $O(n \cdot m)$.

Iz podanega izraza je mogoče razbrati, da je funkcija razdalje odvisna od funkcije osnovne cene operacij. Glede na to da imamo le en ničelni simbol in dve končni abecedi, lahko funkcijo osnovne cene podamo tudi kot matriko nenegativnih realnih vrednosti cen. Matrika cen določa različice funkcije razdalje med zaporedjema simbolov.

4.1.1. Levenshteinova razdalja

Uvodoma omenjeno Levenshteinovo razdaljo (Levenshtein, 1966) določa matrika, v kateri so cene enake ena, če se simbola ne ujemata, in nič, če se simbola ujemata. Funkcija osnovne cene operacije je

$$c(a_n, b_m) = \left\{ \begin{array}{ll} 0, & a_n = b_m \\ 1, & a_n \neq b_m \end{array} \right. , \quad (5)$$

kjer sta $a_n \in A \cup \{\epsilon\}$ in $b_m \in B \cup \{\epsilon\}$. Za zamenjavo štejemo le tiste operacije, pri katerih se simbola razlikujeta. Razdalja je potem enaka najmanjšemu številu operacij zamenjave, vrivanj ali izbrisov, potrebnih za poravnavo dveh zaporedij simbolov.

Ta razdalja se večinoma uporablja v računalniških orodjih, s katerimi ocenjujemo govorne modele. Dejstvo, da ne dela razlik med cenami posameznih operacij, jo naredi neuporabno in celo zavajajočo, ko nas ne zanima le najmanjše skupno število operacij, temveč točno katere operacije so potrebne pri poravnavi enega zaporedja simbolov z drugim. Pri poravnavi lahko namreč dobimo enako najmanjše skupno število operacij pri različnih zaporedjih operacij.

Za ponazoritev razlik med različnimi postopki poravnav zaporedij govornih enot bomo v nadaljevanju uporabili v vseh primerih alofone slovenskega govorjenega jezika kot osnovne simbole, ki bodo določali oziroma predstavljali govorne enote. Primeri razčlenitev govora se bodo nanašali na govorno zbirko GOPOLIS (Dobrišek et al., 1998) in govorne modele, tvorjene v našem laboratoriju (Dobrišek, 2000).

```
LAB:    s   t O   p
REC: ts s p o sil t

LAB:    s t O     p
REC: ts s p o sil t
```

Tabela 1: Različni poravnavi dveh zaporedij alofonov, z enakim skupnim številom osnovni operacij.

Denimo, da moram izvesti primerjavo dveh razčlenitev govora govorca, ki je izgovoril ukaz *„Stop!“*. Referenčna razčlenitev posnetka govora, označena z „`LAB:`“, je zaporedje simbolov `LAB: s t O p`. Z uporabo našega modela govora pa smo dobili samodejno razčlenitev `REC: ts s p o sil t`. Primerjava podanih razčlenitev pravi, da je za poravnavo enega zaporedja z drugim potrebno izvesti pet osnovnih operacij. Takoj ugotovimo, da obstaja cela množica različnih možnih zaporedij petih operacij poravnave.

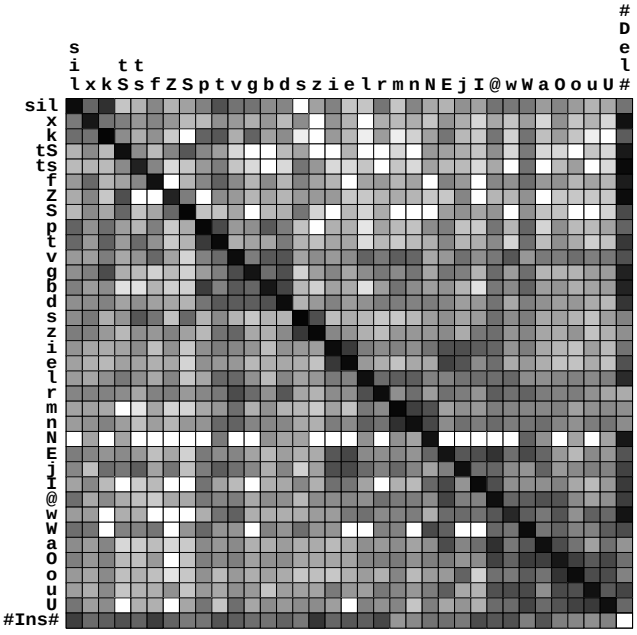
V tabeli 1 sta podana primera poravnav dveh zaporedij alofonov. Prva poravnava je rezultat, dobljen s programomom za podajanje ocen o napakah razpoznavalnika govora „HResults“ uveljavljenega računalniškega orodje HTK (Young et al., 1997) firme Entropic. V ta program je vgrajen postopek iskanja najboljše poravnave, ki temelji na Levenshteinovi razdalji.

Druga poravnava v tabeli 1 je prileganje z operacijami, ki bi jih verjetno izbral glasoslovec, ker bi ocenil, da je cena zamenjave zapornika `[t]` s samoglasnikom `[o]` in zamenjave samoglasnika `[O]` z premorom `[sil]` bistveno večja kot pa cena zamenjave zapornika `[t]` z zapornikom `[p]` in zamenjave samoglasnikom `[O]` s samoglasnikom `[o]`. Ob tem bi še ocenil, da sprememba vrivanje zapornika `[p]` v vrivanje premora `[sil]` ne vnaša posebne napake v poravnavo, ker je vrivanje premora itak verjetno posledica nekaj daljše zapore pred zadnjim zapornikom.

Omenili smo, da govorne modele ocenjujemo s primerjavo dveh razčlenitev istega govora. Ena razčlenitev je refe-

renčna, druga pa pridobljena z uporabo govornega modela. Vse osnovne operacije, potrebne za poravnavo dveh razčlenitev istega govora, štejemo za napako.

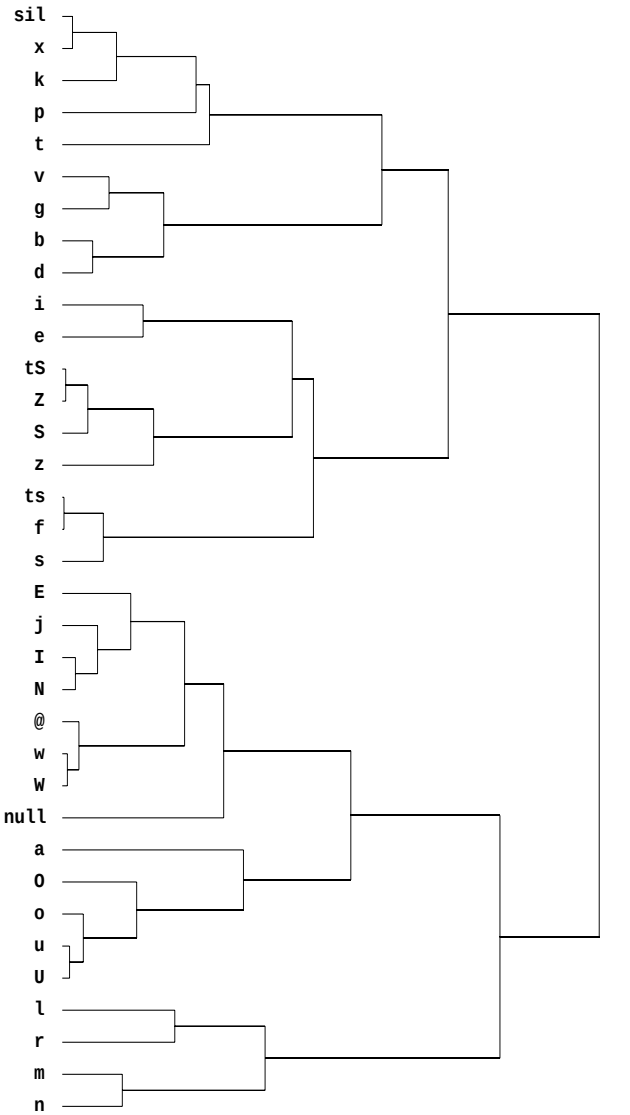
Večinoma nas res zanima le skupno število operacij in s tem napak. Iz tega števila določamo statistično verjetnost napačnega razpoznavanja govora z danim govornim modelom. Neredko pa potrebujemo podrobnejšo predstavitve rezultatov primerjav razčlenitev govora. V tem primeru nas ne zanima le skupno število napak, temveč podrobnejša informacija o zamenjavah, vrivanjih in izbrisih govornih enot. Tovrstne predstavitve rezultatov primerjav so *matrike verjetnosti zamenjav, vrivanj in izbrisov* govornih enot in *dendrogrami rojenja* govornih enot. Matrike vsebujejo podatke o verjetnosti zamenjav enih govornih enot z drugimi govornimi enotami v vseh kombinacijah ter podatke o verjetnosti izbrisov in vrivanj posameznih enot. Dendrograme rojenja govornih enot tvorimo iz matrike verjetnosti zamenjav. Dendrogram rojenja alofonov podajajo zelo pomembno informacijo, na podlagi katere denimo tvorimo t.i. fonetična vprašanja, ki so ključni del postopka vezave parametrov govornih modelov (Young et al., 1997).



Slika 2: Matrika logaritmov verjetnosti zamenjav, pridobljena na podlagi Levenshteinove razdalje.

Verjetnosti v matriki zamenjav, vrivanj in izbrisov določamo statistično, torej s štetjem osnovnih operacij poravnave zaporedjih govornih enot. Slabost obravnavanega postopka primerjave zaporedij simbolov se odraža v tej matriki. Zato je lahko matrika sama kot tudi iz matrike tvorjeni dendrogram rojenja zavajajoč.

Na sliki 2 je prikazan primer matrike verjetnosti zamenjav izbrisov in vrivanj. Verjetnosti so podane grafično v logaritemskem merilu. Temnejša sivina predstavlja večjo vrednost logaritma verjetnosti Takšno prestavitev uporabljamo zato, da so iz slike razvidne tudi zamenjave z zelo majhnimi verjetnostmi. Matrika se nanaša na naš osnovni alofonski govorni model, tvorjen in ocenjevan s pomočjo govorne zbirke GOPOLIS. Namenoma smo uporabili najbolj osnovni model, ki je sposoben razpoznavati alofone slovenskega govora s povprečno verjetnostjo napake razpoznavanja okoli 55 odstotkov. Zaradi večjega števila napak so namreč statistične ocene verjetnosti posameznih napak v matriki natančnejše kot pa bi bile z uporabo najboljšega govornega modela, za katerega ocenjujemo, da je sposoben razpoznavati alofone slovenskega govora s povprečno verjetnostjo napake razpoznavanja okoli 87 odstotkov (Dobrišek, 2000). Iz slike je kljub logaritemskemu merilu razvidna nepričakovano visoka stopnja naključnosti zamenjevanja vsake govorne enote s skoraj vsako drugo govorno enoto. Naša ocena je, da je ta navidezna naključnost zamenjav v bistvu odraz opisane slabosti Levenshteinove razdalje.



Slika 3: Dendrogram rojenja alofonov, pridobljen iz matrike na sliki 2.

Na sliki 3 je prikazan primer dendrograma rojenja alofonov, tvorjen iz matrike na sliki 2. V dendrogram smo vključili še ničelno govorno enoto `[null]`. Isto enoto smo pri podajanju funkcije razdalje označevali z ϵ. Vrivanje in izbrise govornih enot lahko predstavimo tudi kot zamenjave z ničelnim simbolom. Rojenje govornih enot z ničelnim simbolom kaže na njihovo „nagnjenost“ k izbrisom in vrivanjem.

Podani sliki bomo primerjali z rezultati, pridobljenimi

z predlagano razširitvijo postopka poravnave dveh zaporedij simbolov v postopek poravnave dveh zaporedij govornih odsekov.

4.1.2. Utežna poravnava dveh zaporedij simbolov

Slabost Levenshteinove razdalje do neke mere odpravimo, če uporabimo matriko cen osnovnih operacij, ki nima samo vrednosti nič in ena, tako kot je podano v izrazu (5), temveč določimo ceno za vsako operacijo posebej. Funkcija osnovne cene je lahko poljubna funkcija dveh simbolov, ki vrne nenegativno realno vrednost. Poravnavi s poljubno funkcijo osnovne cene pravimo utežna poravnava dveh zaporedij simbolov.

Funkcijo osnovne cene določamo glede na nek pomen simbolov. Pri alofonih bi lahko določili cene operacij na posameznih enotah ročno kot nenegativne realne vrednosti na podlagi izkustvenega glasoslovnega znanja o tem, kako so si posamezni alofoni med sabo podobni, ali kako pomemben je izbris oziroma vrivanje posameznega alofona. Nižjo vrednost cene bi pripisali parom alofonov, za katere sodimo, da so si zvočno bolj podobni in da sodijo v isto skupino, višjo ceno pa parom, za katere menimo, da se ne bi smeli zamenjevati oziroma da so lahko le redko izbrisani ali vrinjeni.

Po drugi strani bi lahko matriko cen pridobili tudi s postopkom ugotavljanja rojenja zvočnih vzorcev, pridobljenih iz odsekov govornih signalov. Matriko cen bi v tem primeru določili iz povprečnih razdalj med množicami zvočnih vzorcev, ki pripadajo posameznim alofonom. Zvočni vzorci nekaterih alofonov so si namreč zanesljivo nekaj bližji in spet drugih bolj narazen, odvisno seveda od izbrane metrike ter načina določanja zvočnih vzorcev.

Opisane rešitve pa imajo spet slabost, da zahtevajo za pridobivanje ustreznih vrednosti cen vrsto dodatnih analiz obsežnih zbirk podatkov. Polega tega pa bo na koncu dobljena matrika verjetnosti zamenjav, izbrisov in vrivanj v precejšnji meri odražala samo matriko cen. In ne nazadnje tudi nimamo vedno jasnih zagotovil, da bo pri izbranih cenah rekurzivna funkcija poravnave res funkcija razdalje.

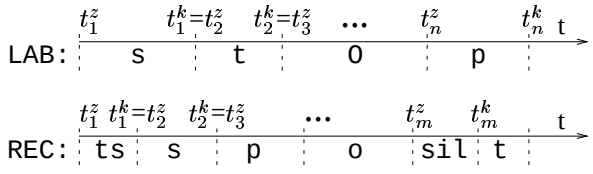
Pri besednih govornih enota so možnosti za določitev ustreznih cen operacij nekaj večje (Hunt, 1990). Uteženo poravnavo dveh zaporedij besed lahko izvedemo tako, da enačimo ceno zamenjave dveh besed z Levenshteinovo razdaljo med njunima dvema grafemskima zapisoma ali kanoninčnima fonetičnima prepisoma. Ceno vrivanja in izbrisa pa enačimo s številom simbolov v zapisu besede.

V okviru ameriških Defense Advanced Research Projects Agency (DARPA) ovrednotenj samodejnih razpoznavalnikov govora so pri ameriškem National Institute of Standards and Technology (NIST) razvili programski paket *„Score“* za ocenjevanje govornih modelov. Paket vsebuje orodje, ki med drugim omogoča uporabo poljubne matrike cen oziroma matrike, ki se določi iz razdalj med simbolnimi zapisi besed v besednjaku. Ta rešitev je zanesljivo mnogo bolj primerna kot uporaba zgolj Levenstheinove razdalje. Programski paket se nenehno dopolnjuje v nove različice in postaja svojevrsten standard pri ocenjevanju in primerjavah samodejnih razpoznavalnikov govora (Gibbon et al., 1997).

Po našem mnenju postopek poravnave z opisanimi načini uteževanja še vedno ni najboljša možna rešitev. Primerna je kvečjemu za besedne govorne enote in ne za podbesedne, kot so denimo alofoni. Predvsem pa govorne enote še vedno obravnava le kot simbole v zaporedju in pri iskanju najboljše poravnave ne upošteva časovnega ozadje simbolov, uresničenih v govornem signalu.

4.2. Razdalja med zaporedji govornih odsekov

Govorni odseki niso le simboli, temveč simboli s pripisanimi časovnimi mejami pripadajočih odsekov govornega signala. V primer iz tabele 1 moramo tako vključiti tudi časovno informacijo. Razčlenitev na govorne odseke je ponazorjena na sliki 4. Razčlenitvi sta različni, vendar se nanašata na isti govorni signal. Simbolni prikaz razčlenitev je takšen, ko smo ga uporabili na sliki 1.



Slika 4: Poravanava dveh zaporedij govornih odsekov.

Rezultat uporabe govornega modela za razčlenjevanje govora je lahko po potrebi vedno tudi zaporedje govornih odsekov in ne le zaporedje simbolov, ki predstavljajo govorne enote. Tako pa ni tudi z referenčnimi razčlenitvami testnih posnetkov govora. Govorne zbirke namreč ponavadi vsebujejo le simbolne zapise govora. To pa zato, ker natančnejše ročno razčlenjevanje zahteva preveč dela in ker tvorjenje sodobnih govornih modelov iz govornih zbirk ne zahteva nujno bolj natančne razčlenitve. Do razčlenitve referenčnih posnetkov govora pa lahko vseeno pridemo iz simbolnih zapisov tako, da najprej tvorimo modele govornih enot iz teh posnetkov in nato z uporabo postopka vsiljene poravnave razčlenimo posnetke na zaporedja govornih odsekov.

4.2.1. Funkcija osnovne cene operacij na govornih odsekih

Razširitev doslej obravnavanega postopka primerjave dveh zaporedij simbolov v postopek primerjave dveh zaporedij govornih odsekov smo izvedli tako, da smo funkcijo osnovne cene operacij na simbolih naredili odvisno ne le od simbolov, temveč tudi od časovnih mej govornih odsekov. Ideja razširitve je v tem, da bi morala biti cena zamenjave dveh govornih enot tem višja čim bolj sta govorni enoti v času oddaljeni ena od druge. Tudi vrivanje ali izbris govorne enote bi moralo imeti tem višjo ceno čim bolj je v času oddaljeno mesto vrivanja ali izbrisa v nasprotnem zaporedju od brisane ali vrivane govorne enote. Cene bi morale biti odvisne tudi od trajanja govornih enot, na katerih se izvaja operacija. Daljše je trajanje govorne enote dodatno višja naj bi bila cena.

Govorne enote pa zdaj niso več le simboli, zato funkcijo razdalje ne določata več končni abecedi simbolov, temveč množici govornih odsekov Γ in Λ. Množici določimo tako, kot je bilo podano v poglavju 2.1. Pri funkciji osnovne cene $c : E \to \mathbb{R}_0^+$ je množica $E = E_z \cup E_v \cup E_i$ zdaj

množica operacij na govornih odsekih in ne več le na simbolih. Množica $E_z = \Gamma \times \Lambda$ je množica operacij zamenjav, $E_v = \Upsilon \times \Lambda$ je množica operacij vrivanj in $E_i = \Gamma \times \Upsilon$ množica operacij izbrisov. Množica Υ je množica parov, ki predstavljajo ničelne enote oziroma ničelne govorne odseke

$$\Upsilon = \left\{ \varepsilon_\tau = (\epsilon, \tau) \in \{\epsilon\} \times \mathbb{R}_0^+ \right\} .$$

Funkcijo razdalje med zaporedjema govornih odsekov zapišemo kot

$$d_c : \Gamma^* \times \Lambda^* \to \mathbb{R}_0^+ ,$$

pri čemer sta množici zaporedij določeni tako kot v izrazu (2). V podanem primeru lahko imata obe zaporedji različno končno število zaporednih govornih odsekov z različnim končnim skupnim trajanjem.

Pri dveh zaporedjih govornih odsekov, ki se nanašata na isti govor s končnim trajanjem T, lahko posplošitev odpravimo in vzamemo v obravnavo manjši množici možnih govornih odsekov Γ_T in Λ_T, določeni po izrazu (1). Množico vseh operacij E zdaj uvedemo iz teh dve množic in za operacije izbrisov in vrivanj uvedemo zmanjšano množico možnih ničelnih govornih odsekov

$$\Upsilon_T = \{ \varepsilon_\tau = (\epsilon, \tau) \in \{\epsilon\} \times [0, T] \} .$$

Funkcijo razdalje med zaporedjema govornih odsekov istega govora zdaj zapišemo kot

$$d_c : \Gamma_T^* \times \Lambda_T^* \to \mathbb{R}_0^+ .$$

Podana funkcija razdalja se zdaj nanaša na zaporedji z poljubnima končnima številoma govornih odsekov z istim skupnim trajanjem, vendar ne nujno na isti govorni signal.

Potrebno je določiti še funkcije osnovne cene operacij na govornih odsekih. V skladu z razmišljanjem iz začetka podpoglavja predlagamo, da je funkcija cene zamenjave enega govornega odseka $a_n = (\sigma_n, t_n^z, t_n^k) \in \Gamma_T$ z drugim $b_m = (\lambda_m, t_m^z, t_m^k) \in \Lambda_T$ enaka

$$\begin{aligned} c(a_n, b_m) \quad = \quad & |t_n^z - t_m^z| \ + \ |t_n^k - t_m^k| \ + \\ & t_n^k \ + \ t_m^k \ - \ t_n^z \ - \ t_m^z , \qquad \sigma_n \neq \lambda_m . \end{aligned}$$

Podana funkcija osnovne cene zamenjave velja v primeru, ko se govorna odseka razlikujeta v simbolih, ki predstavljata primerjani govorni enoti. Če se simbola ujemata potem funkcijo cene zamenjave spremenimo v funkcijo cene ujemanja

$$c(a_n, b_m) = |t_n^z - t_m^z| \ + \ |t_n^k - t_m^k| , \quad \sigma_n = \lambda_m ,$$

ki podaja le zamik primerjanih govornih odsekov. Iz obeh izrazov je razvidno, da ceno zamenjave za razliko od cene ujemanja dodatno povišamo s trajanjem obeh zamenjanih govornih enot. V izvirnem postopku iskanja najboljše poravnave dveh zaporedij simbolov smo Funkcijo osnovne cene izbrisa smo izbrali

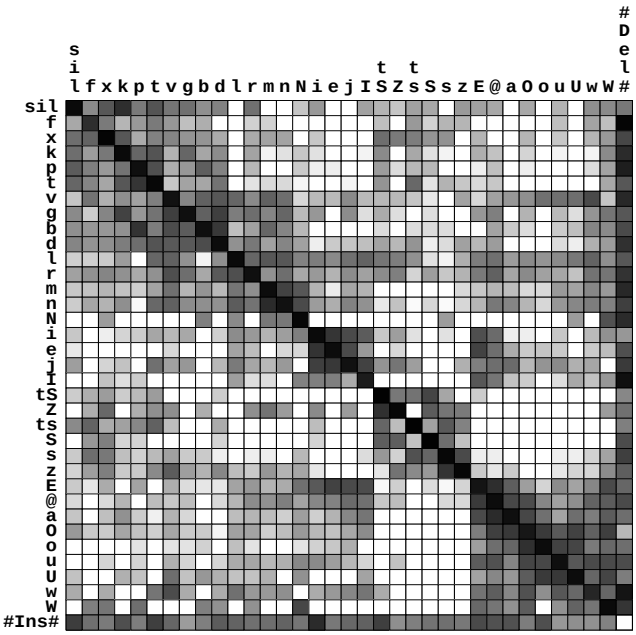
$$c(a_n, \varepsilon_\tau) = |t_n^z - \tau| \ + \ |t_n^k - \tau| \ + t_n^k - t_n^z .$$

Izbris je dejansko zamenjava govornega odseka z ničelno govorno enoto $\varepsilon_i \in \Upsilon_T$ brez trajanja v trenutku t_m^ϵ. Podobno lahko določimo še funkcijo osnovne cene vrivanja

$$c(\varepsilon_\tau, b_m) = |\tau - t_m^z| + |\tau - t_m^k| \ + t_m^k - t_m^z .$$

V izvirnem postopku iskanja najboljše poravnave zaporedij simbolov smo dejansko spremenili le funkcijo osnovne cene posamezne operacije. Po izvedbi spremenjenega postopka sicer dobimo drugačno skupno najmanjšo ceno poravnave zaporedij govornih odsekov, vendar lahko iz rezultata razberemo, katero točno zaporedje operacij je potrebno za poravnavo enega zaporedja z drugim. Uvedba predstavljene funkcije osnovne cene povzroči, da se v poravnavi izbere tisto zaporedje operacij, ki najbolj upošteva pridano časovno informacijo.

Skupna najmanjša cena poravnave se določa po istem rekurzivnem izračunu kot pri izvirnem postopku (4). Ali je rezultat poravnave dejansko funkcija razdalje, bi morali posebej dokazovati. Dokaz presega okvir tega članka. Zadostitve prvim trem pogojem za funkcijo razdalje so razvidne že iz načina, kako smo posplošili izvirni postopek. Pri trikotniškem pogoju za postopek poravnave velja, da moramo dokazati, da ne obstaja par dveh govornih enot, pri katerih bi bilo zaporedje operacij enega vrivanja in enega izbrisa cenejše od njune zamenjave.

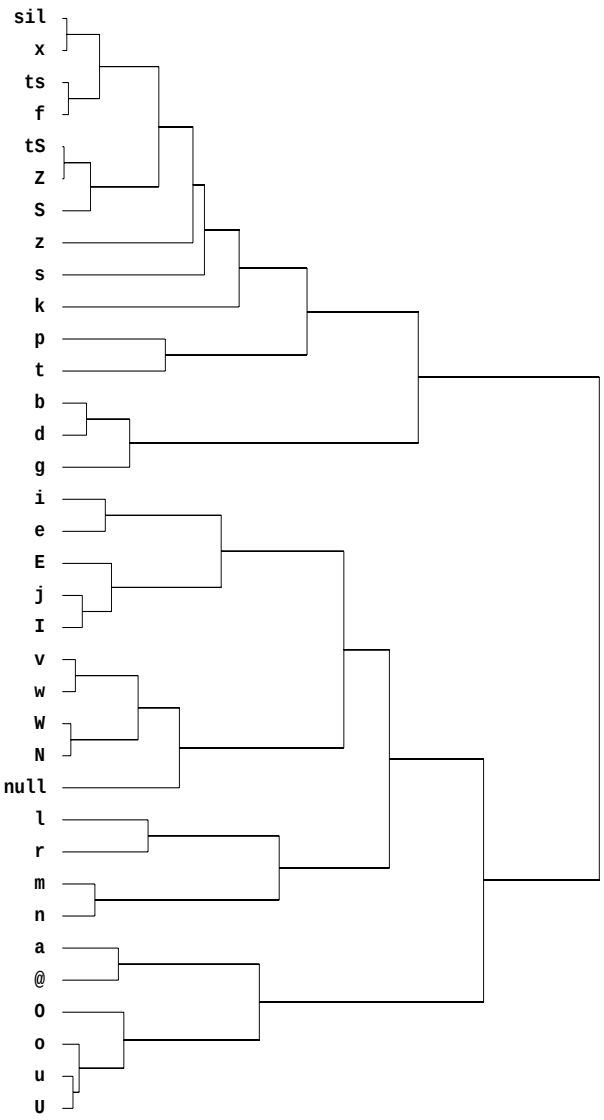


Slika 5: Matrika, pridobljena iz rezultatov iskanja najboljših poravnav zaporedij govornih odsekov.

Za primerjavo smo izvedli ugotavljanje napak samodejnega razčlenjevanja istega govora z istimi alofonskimi modeli, kot je bilo storjeno pri izvirnem postopku. Na sliki 5 je prikazana matrika verjetnosti zamenjav, izbrisov in vrivanj, pridobljene z ugotavljanjem poravnave istih razčlenitev istega govora kot v primeru na sliki 2. Kljub temu, da dobimo povsem enako skupno potrebno število operacij poravnav kot pri izvirnem postopku, je razvidna znatno manjša naključnost medsebojnih zamenjav govornih enot. Zamenjave, ki so ostale, pa so nekako bolj v skladu s pričakovanji, ki bi jih imel glasoslovec pri opazovanju primerjanih razčlenitev. Tudi iz obeh dendrogramov so razvidne posledice vključitve časovne informacije v obravnavani postopek. Morda je vredno omeniti, da so za natančnega glasoslovca tudi v drugem dendrogramu opazne določene nelogičnosti. Te so posledica redkosti določenih govornih

enot v govorni zbirki GOPOLIS. Mislimo predvsem na alofona [W] in [N].

Verodostojnost ugotavljanja razlik in napak v primerjanih razčlenitvah istega govora s postopkom iskanja najboljše poravnave zaporedij govornih odsekov bi morali dokazovati s statistično primerjavo s človeškim opazovalcem, ki je na podlagi svojega znanja ocenil, kje so dejansko razlike med zaporedji in kakšne so te razlike v smislu zamenjav, vrivanj in izbrisov govornih enot. Zanimivo je, da bi tudi pri človeških opazovalcih ugotovili znatne razlike pri odločitvah, katere točno so razlike med razčlenitvami istega govora (Oppermann et al., 2000).



Slika 6: Dendrogram rojenja alofonov, pridobljen iz matrike na sliki 5.

Žal zaenkrat ne premoremo nobene ustrezne podatkovne zbirke, ki bi poleg različnih razčlenitev istega govora vsebovala še preverjene podatek o tem, na katerih mestih v zaporedjih so razlike oziroma napake. Naše izkušnje pri opazovanju in primerjavah rezultatov nekaj desetin poravnav kažejo na to, da je predstavljeni postopek nedvomno skladnejši s tem, kakšne poravnave bi sami določili na podlagi lastnega znanja o jeziku in govoru. Menimo pa, da je tudi iz samega opisa postopka mogoče razbrati prednosti v primerjavi z izvirnim postopkom.

5. Zaključek

Predstavljeni prehod iz iskanja poravnave zaporedij simbolov na iskanje poravnave zaporedij govornih odsekov je le ena od možnih posplošitev. Pri besednih enotah bi lahko podani funkciji osnovne cene operacij na govornih odsekih pridružili še Levenshteinovo razdaljo med simbolnimi zapisi besed. Funkcijo osnovne cene operacij bi lahko naredili odvisno tudi od primerjanih s simboli označenih odsekov govornega signala (izraz (3)). V tem primeru bi funkcijo cene izpeljali iz mere podobnosti dveh odsekov signala, ki jo ponavadi določamo z znanim postopkom dinamičnega ukrivljanja časa (DTW).

Nedvomno je smiselno ocenjevati govorne modele z uporabo predstavljenega postopka takoj, ko nas ne zanima le skupno število napak samodejnega razčlenjevanja govora, temveč tudi katere točno napake so bile storjene. Takšno informacijo potrebujemo pri ugotavljanju rojenja govornih enot, iskanju izgovorjevalnih različic besed, iskanju morebitnih pravil v napakah in podobno.

V laboratoriju smo razvili program „IHResults“, ki je naša različica programa „HResults“ v paketu HTK firme Entropic (Young et al., 1997). Program v celoti vključuje predstavljeni postopek in omogoča izbiro med različnimi poravnavami, vključno z nalaganjem predoločene matrike osnovih cen. S tem programom smo tvorili prikazane slike matrik verjetnosti zamenjav in dendrograme. Morebitnim uporabnikom bo program v kratkem na voljo na laboratorijskem spletnem strežniku.

6. Literatura

S. Dobrišek, J. Gros, F. Mihelič, and N. Pavešić. 1998. Recording and labelling of the gopolis slovenian speech database. In *Proc. 1st Int. Conf. on Language Resources & Evaluation*, zvezek 2, strani 1089–1096. ELRA.

S. Dobrišek. 2000. *Analiza in razpoznavanje glasov v govornem signalu*. Doktorska disertacija (v pripravi), Univerza v Ljubljani, Fakulteta za elektrotehniko.

D. Gibbon, R. Moore, and R. Winski, editors. 1997. *EAGLES Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, Walter de Gruyter Publishers, Berlin & New York.

M. Hunt. 1990. Figures of merit for assessing connected-word recognizers. *Speech Communication*, 9:329–336.

A. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phy. Dokl.*, 10:707–710.

W. Masek and M. Paterson. 1980. A faster algorithm computing string edit distances. *Journal of Comput. System Sci.*, 20:18–31.

D. Oppermann, S. Burger, and K. Weilhammer. 2000. What are transcription errors and why are they made. In *Proc. 2nd Int. Conf. on Language Resources & Evaluation*, zvezek 1, strani 409–441. ELRA.

S. V. Rice, H. Bunke, and T. A. Nartker. 1997. Classes of cost functions for string edit distance. *Algoritmica*, 18:271–280.

S. Young, J. Odell, D. Ollason, V. Vatchev, and P. Woodland. 1997. *The HTK Book*. Cambridge University, Entropic Cambridge Research Laboratory Ltd.