Izvirni znanstveni članek

TEHNIKA

ANALI PAZU
3/ 2013/ 2: 69-74
www.anali-pazu.si

Nenadzorovano učenje akustičnih modelov govora

Unsupervised training for acoustic models of speech

Gregor Donaj, Andrej Žgank, Mirjam Sepesy Maučec*

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko

E-Mails: gregor.donaj@um.si; andrej.zgank@um.si ; mirjam.sepesy@um.si

*Avtor za korespondenco: tel. +386 2 220 72 25

Povzetek: V članku je predstavljeno nenadzorovano učenje akustičnih modelov za razpoznavanje tekočega govora. Ključna prednost takega učenja je njegova hitrost in nizki stroški v primerjavi z izdelavo transkripcij govora, ki so potrebne za nadzorovano učenje. Predstavljeni sta dve metodi nenadzorovanega učenja, ki sta preizkušeni na razpoznavalniku tekočega govora z velikim slovarjem v domeni dnevno-informativnih oddaj.

Ključne besede: akustični modeli, razpoznavanje govora, nenadzorovano učenje.

Abstract: This paper presents unsupervised acoustical model training for automatic speech recognition. The main advantage of this training method is its speed and cost effectiveness compared to the manual transcription of speech, which is needed for supervised training. We present two methods of unsupervised training and test them on a large vocabulary continuous speech recognition system in the Broadcast News domain.

Key words: acoustical models, speech recognition, unsupervised training.

1. Uvod

Ena najpomembnejših komponent v razpoznavalniku govora so akustični modeli. To so statistični modeli, ki opisujejo akustične značilnosti fonemov. Akustični modeli so zelo kompleksni, saj vsebujejo veliko število parametrov. Zato za ocenjevanje teh parametrov – učenje modelov – potrebujemo velike količine učnih podatkov. To so pari zvočnih posnetkov govora in pripadajočih transkripcij.

S tehničnega stališča ni težavno dobiti dovolj veliko množico golega akustičnega materiala. Bolj zahtevna je izdelava pripadajočih transkripcij. Te je potrebno izdelovati skrbno in ročno. Ocenjujemo, da izdelava transkripcij za eno uro zvočnega posnetka lahko zahteva 20 do 40 ur ročnega dela [1,2]. S tem postane izdelava primerne učne množice dolgotrajna in draga.

Tudi kadar že imamo pripravljeno učno množico za učenje akustičnih modelov, jo morda želimo kasneje razširiti. Morda pa imamo pripravljen nek razpoznavalnik in ga hočemo prilagoditi na drugo domeno. V teh primerih lahko ponovno na enak način izdelujemo oz. povečujemo učno množici, kot je to običajno. Alternativna možnost je uporaba nenadzorovanega učenja, kjer uporabljamo razpoznavalnik za izdelavo transkripcij. Kadar smo v domeni razpoznavanja z velikim slovarjem (60.000 besed in več) je trajanje avtomatske izdelave transkripcij primerljivo z ročno izdelavo. Prednost razpoznavalnika pa je, da ga lahko uporabljamo 24 ur na dan v več procesih vzporedno. S tem lahko skrajšamo čas izdelave transkripcij. Razpoznavalnik tudi predstavlja manjši strošek delovanja, kot pa ročna izdelava transkripcij.

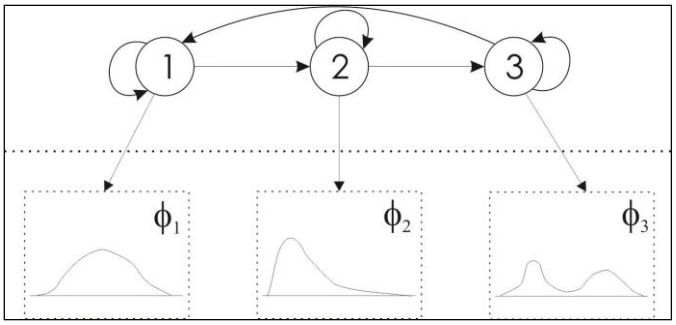
Osnovna ideja nenadzorovanega učenja [3,4] je uporaba razpoznavalnika govora na netranskribiranih zvočnih posnetkih in uporaba dobljenih rezultatov razpoznavanja kot transkripcij pri učenju akustičnih modelov. Dobra stran takšnega pristopa je zmanjšanje stroškov izdelave učnih podatkov, saj odpade večji del ročnega dela. Slaba stran pa je vnos napak v učno množico, ki so posledica napak pri razpoznavanju.

V članku bomo opisali postopek nadzorovanega in nenadzorovanega učenja akustičnih modelov. Predlagali bomo dva načina izvedbe nenadzorovanega učenja. Delovanje obeh načinov bomo primerjali z nadzorovanih učenjem. Pri tem bomo uporabljali bazo BNSI [2]. Za vrednotenje uspešnosti učenja bomo izdelane akustične modele uporabili v razpoznavalniku tekočega govora UMB Broadcast News [5].

V nadaljevanju bomo opisali splošne značilnosti akustičnih modelov ter splošne postopke nadzorovanega in nenadzorovanega učenja. Nato bomo opisali eksperimentalni sistem in naša predlagana načina nenadzorovanega učenja. Sledilo bo primerjanje uspešnosti naših načinov učenja z nadzorovanim načinom učenja.

2. Akustični modeli

Prevladujoči pristop za akustično modeliranje govora je uporaba prikritih modelov Markova (angl.: Hidden Markov Model – HMM) [6]. Definirani so z množico stanj, prehodnimi verjetnostmi med stanji in verjetnostmi izhodnih simbolov. Če imamo v splošnem HMM-u N stanj, lahko prehodne verjetnosti opišemo z N^2-N parametri. Število parametrov, ki jih potrebujemo za izhodne verjetnosti, je odvisno od vrste porazdelitve izhodne spremenljivke. Če ima ta na primer enodimenzionalno Gaussovo porazdelitev, potrebujemo 2 parametra (srednja vrednost in varianca). Primer enostavnega modela HMM je na sliki 1. Da ga natančno definiramo, moramo podati 6 prehodnih verjetnosti in vse parametre, ki so potrebni za definicijo treh izhodnih porazdelitev.



Slika 1. Primer modela HMM s tremi stanji in zveznimi porazdelitvami izhodne spremenljivke.

Vrednosti izhodnih spremenljivk HMM-a so vektorji značilk. Značilke se izračunajo iz zvočnega posnetka po določenem postopku. Najpogostejša postopka sta izračuna mel-frekvenčno kepstralnih koeficientov [7] (angl.: Mel-Frequency Cepstral Coefficients – MFCC) in koeficientov percepcijskega linearnega napovedovanja [8] (angl.: Perceptual Linear Prediction – PLP). Tipično računamo vrednosti tudi do 12 koeficientov ter energijo signala. Zraven samih vrednosti uporabljamo še njihove prve in druge odvode. Skupaj imamo do 39 parametrov v vektorju značilk. Za porazdelitve izhodnih spremenljivk uporabljajo mešane Gaussove porazdelitve značilk. Za opis take porazdelitve moramo podati za vsako komponento porazdelitve utež, srednji vektor in kovariančno matriko.

V razpoznavanju tekočega govora z velikim slovarjem se uporabljajo trifonski akustični modeli. Ti so sestavljeni iz treh različnih stanj. Za vsak fonem jezika v katerem koli kontekstu (predhodni in naslednji fonem) definiramo en akustični model. Za jezik z N fonemi pomeni to N^3 trifonskih modelov. K temu še se kasneje prištejejo modeli za tišino in kratke premore. Zaradi podobnosti med modeli glede na različni kontekst se nekateri modeli kasneje združijo. Kljub temu običajno ostane število različnih trifonskih modelov veliko.

Vse te modele lahko opišemo, če v vsakem modelu za vsako stanje podamo prehodne verjetnosti do ostalih stanj in parametre za opis večdimenzionalne verjetnostne porazdelitve vektorja značilk. Zaradi velikega števila posameznih trifonskih modelov in velike dimenzionalnosti značilk to pomeni zelo veliko število parametrov, ki so potrebni za opis celotnega akustičnega modela. Zato za dobro ocenjevanje teh parametrov potrebujemo čim večjo učno množico.

3. Učenje modelov

Učenje modelov pomeni, da na podlagi učnih podatkov statistično ocenjujemo parametre modelov. V primeru akustičnih modelov to pomeni, da vzamemo zvočni posnetek in iz njega izločimo značilke. Istočasno na podlagi pripadajoče transkripcije sestavimo nenaučen akustični model za posnetek. S pomočjo algoritmov učenja spreminjamo parametre modela tako, da maksimiramo verjetnost, da je akustični model kot izhodne spremenljivke tvoril zaporedje značilk, ki smo jih dobili iz posnetka. Opisan kriterij učenja imenujemo kriterij največje verjetnosti [9]. Poznamo tudi druge kriterije za učenje akustičnih modelov, kot sta največja skupna informacija [10] in najmanjša napaka klasifikacije [9], ki pa jih v tem članku ne bomo obravnavali.

Označimo z Λ množico vseh parametrov, ki opisujejo akustični model. Učni posnetki naj bodo $X=\{X_1,\ldots,X_T\}$, pripadajoče transkripcije pa $S=\{S_1,\ldots,S_T\}$. Pri kriteriju največje verjetnosti velja, da določimo parametre modela s predpisom

$$\begin{aligned}\Lambda &= \arg\max_{\lambda} p(X \mid S,\lambda) \\ &= \arg\max_{\lambda} \prod_{t=1}^{T} p(X_t \mid S_t,\lambda).\end{aligned} \tag{1}$$

To pomeni, da optimalne parametre izberemo tako, da maksimirajo kriterijsko funkcijo

$$F(\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log(p(X_t \mid S_t,\lambda)). \tag{2}$$

Za učenje HMM uporabljamo Baum-Welchev algoritem [9]. Ta v osnovi temelji na bolj splošnem EM (Expectation-Maximization) algoritmu. Pri njem se izmenično ponavljata dva koraka. V prvem izračunamo pričakovano vrednost verjetnosti, da je model tvoril

opazovane podatke. V drugem koraku pa spremenimo vrednosti parametrov tako, da poskušamo maksimirati verjetnosti, ki jih dobimo v prvem koraku. V algoritmu potrebujemo tudi začetne približke. Ker se na vsakem koraku le postopoma v (majhnih) korakih spreminjajo vrednosti parametrov, algoritem vodi le do nekega lokalnega ekstrema kriterijske funkcije, ki je odvisen od začetnih vrednosti. Oba koraka v algoritmu ponavljamo tako dolgo, dokler ne postane razlika med verjetnostma v dveh zaporednih iteracijah dovolj majhna.

Splošni postopek nadzorovanega učenja v praktični uporabi lahko opišemo v naslednjih korakih:

1. Zberemo zvočni material, ga segmentiramo in izdelamo transkripcije.
2. Naredimo seznam besed in vsaki pripišemo fonemsko transkripcijo.
3. Tvorimo začetne modele na zmanjšani učni množici.
4. Časovno poravnamo fonemske transkripcije segmentov z zvočnimi posnetki.
5. Izločimo segmente, ki jih ne moremo uspešno poravnati.
6. Izvedemo algoritem za učenje modelov.

3.1. Nenadzorovano učenje

Kadar na splošno govorimo o razliki med nadzorovanim in nenadzorovanim učenjem statističnih modelov, mislimo na razpoložljive podatke. V nadzorovanem učenju podamo algoritmu učne vhodne podatke in referenčne izhodne podatke. Primer je klasifikacija, kjer za vsak učni podatek podamo informacijo, kateremu razredu pripada. Pri tem morajo referenčni biti preverjeni. Pri nenadzorovanem učenju, kot je na primer grozdenje, pa podamo le gole učne podatke. Algoritem nato podatke sam razdeli na več razredov, brez da bi imel informacijo o pomenu posameznih razredov.

Kadar govorimo o nenadzorovanem učenju na primeru akustičnih modelov, imamo rahlo drugačno predstavo. Sam algoritem učenja modelov je še vedno enak kot pri nadzorovanem učenju. To pomeni, da uporabljamo zvočne posnetke in transkripcije. Razlika je v izdelavi transkripcij. V primeru nadzorovanega učenja so dobljene ročno, v primeru nenadzorovanega pa avtomatsko s pomočjo razpoznavalnika. Pojem nenadzorovano se torej nanaša na dejstvo, da transkripcije niso preverjene. Še vedno pa pred začetkom samega postopka nenadzorovanega učenja potrebujemo že izdelane akustične modele. S postopkom nenadzorovanega učenja modele le izboljšujemo.

Za ocenjevanje uspešnosti razpoznavanja govora uporabljamo delež napačno razpoznanih besed (angl.: Word Error Rate – WER). To je razmerje med številom napak v razpoznavanju in številom vseh besed. Primer »metrike«, ki jo lahko uporabljamo za ocenjevanje postopka nenadzorovanega je *WER Recovery* [11]. Ta je definirana kot razmerje med izboljšanjem deleža napačno razpoznanih besed pri nenadzorovanem in nadzorovanem učenju. Izračunamo ga z enačbo

$$WER\ Recovery = \frac{WER_I - WER_U}{WER_I - WER_S}, \qquad (3)$$

kjer predstavljajo posamezni indeksi: I – začetni model, U – model z nenadzorovanim učenjem in S – model z nadzorovanim učenjem.

Lamel in drugi [3] so pokazali, da je možno učenje akustičnih modelov z začetnim nadzorovanim učenjem na le 10 minutah transkribiranega materiala. Prav tako je predstavila postopek t.i. rahlo nadzorovanega učenja, kjer so uporabljali le približne transkripcije zvočnega materiala. V [8] je bilo opisano nenadzorovano učenje z jezikovnim modelom izdelanim na majhni količini teksta – 100.000 besed. Dosežen je bil 50 % WER Recovery. Novotney [11] uporablja postopek nenadzorovanega učenja tudi na jezikovnih modelih. Wessel in Ney [4] sta testirala nenadzorovano učenje s transkribiranim materialom v obsegu od 1 do 6 ur in 72 urami netranskribiranega materiala. Z iterativnim postopkom učenja, uporabo mere zaupanja in testiranjem na različnih testnih setih sta zmanjšala delež napačno razpoznanih besed za približno 50 % relativno in dosegla WER Recovery med 87 % in 89 %. V [12] je bil pred kratkim predstavljen tudi postopek nenadzorovanega učenja, ki namesto najboljše hipoteze uporablja besedne mreže, ki jih tvori razpoznavalnik.

Splošni postopek nenadzorovanega učenja v praktični uporabi lahko opišemo v naslednjih korakih:

1. Zberemo zvočni material, ga segmentiramo in izdelamo transkripcije.
2. Zberemo dodatni zvočni material brez transkripcij.
3. Naredimo seznam besed iz transkripcij in slovarja razpoznavalnika in vsaki besedi pripišemo fonemsko transkripcijo.
4. Tvorimo začetne modele na osnovni učni množici po postopku nadzorovanega učenja.
5. Razpoznavamo zvočni material iz dodatne učne množice.
6. Časovno poravnamo transkripcije segmentov dobljene z razpoznavalnikom z zvočnimi posnetki.
7. Izvedemo algoritem za učenje modelov na razširjeni množici.

Vsakega od obeh postopkov lahko priredimo na nekoliko drugačne različice delovanja. Natančni postopki za učenje, ki smo jih uporabljali v eksperimentih, so opisani v naslednjem poglavju.

4. Eksperimentalni sistem

4.1. Uporabljene baze in razpoznavalnik

Vse eksperimente smo izvajali na bazi BNSI [2] z razpoznavalnikom UMB Broadcast News [5]. Trenutna različica učne množice vsebuje 24 oddaj s skupno dolžino

21,6 ur. V zvočnem delu baze sta še razvojna in testna množica, ki obe vsebujeta po 4 oddaje s skupno dolžino približno 3 ure. Učno množico smo razdelili na dva dela. Prvi del predstavlja približno četrtino množice. Ta del bomo uporabljali za nadzorovanje učenje osnovnih modelov. Drugi del vsebuje preostanek učne množice. Na tem delu učne množice bomo izvajali nenadzorovano učenje. Testna množica je namenjena testiranju uspešnosti razpoznavanja z različnimi akustičnimi modeli, ki smo jih izdelali. Celotna učna množica je ročno segmentirana. Zraven zvočnega dela vsebuje baza tudi tekstovni del, ki obsega 11 milijonov besed. Ta del je bil uporabljen pri izdelavi jezikovnega modela.

Za delo s posnetki in transkripcijami ter za razpoznavanje smo uporabljali nabor orodij HTK [13].

Uporabljene značilke so koeficienti MFCC, izračunani na Hammingovih oknih dolžine 25 ms in v razmiku 10 ms. Izračunali smo 12 značilk. Uporabili smo 26 kanalov in 22 kepstralnih filtrov. Značilkam MFCC smo dodali še energijo. Pri značilkah smo uporabljali tudi prve in druge odvode. Tako smo imeli skupno 39 značilk. Za izračun smo uporabljali orodje *HCopy*.

Za razpoznavanje so uporabljali orodje *HDecode*, ki izvaja časovno sinhroni Viterbijev iskalni algoritem [14]. Uporabljeni jezikovni modeli so klasični trigramski modeli interpolirani na treh množicah: transkripcije učne množice BNSI, tekstovni del baze BNSI in slovenski jezikovni korpus FidaPLUS [15]. Tudi pri jezikovnem modeliranju je pomembna velikost učnega gradiva. Korpus FidaPLUS predstavlja največjo zbirko slovenskih besedil, ki nam je trenutno na voljo. Modele smo interpolirali z drugima dvema deloma, ker ta predstavljata besedila iz domene (dnevna poročila), v kateri uporabljamo razpoznavalnik. Interpolacijski koeficienti so optimizirani na razvojnem delu baze BNSI. Uporabljen je slovar velikosti 64.000 besed. Ker nimamo na voljo pravil za grafemsko-fonemsko pretvorbo slovenskih besed, smo v slovarju izgovorjav uporabljali grafemske transkripcije.

4.2. Postopek nadzorovanega učenja

Nadzorovano učenje smo uporabljali dvakrat. Prvič na četrtini učne množice. Dobljeni modeli so nam služili kot referenčni, t.i. *baseline*, modeli. Drugič smo nadzorovano učenje uporabili na celotni učni množici. S pomočjo teh rezultatov lahko kasneje ocenimo doprinos nenadzorovanega učenja na povečani učni množici v primerjavi z nadzorovanim učenjem na isti množici. Vse transkripcije so bile že v naprej pripravljene v poenoteni obliki. Natančen postopek, ki smo ga uporabljali pri učenju, je sledeči:

1. Uredimo slovar grafemskih transkripcij v oblike, ki jih potrebujemo za orodje HTK. Pri tem izdelamo tudi seznam vseh grafemov, ki se pojavijo v slovarju. Besedne transkripcije segmentov razširimo v grafemske transkripcije.
2. Izberemo podmnožico učne množice za učenje prvih modelov.
3. Izračunamo globalne srednje vrednosti in variance značilk na izbrani učni podmnožici. Tako izdelamo prototipni model. Naredimo njegove kopije za vsak grafem.
4. Na učni podmnožici naučimo začetne monofonske modele.
5. Dodamo model za tišino med besedami in ponovimo dve iteraciji učenja.
6. Na celotni učni množici izvedemo časovno poravnavo med monofonskimi transkripcijami in zvočnimi posnetki. Izločimo segmente, pri katerih poravnava ni uspešna. Ponovimo dve iteraciji učenja; tokrat na celotni učni množici.
7. Naučimo modele, ki uporabljalo mešane Gaussove porazdelitve z 2, 4, 8, 16 in 32 porazdelitvami. Tukaj modele z več porazdelitvami dobimo postopoma iz modela z manj porazdelitvami. Za vsak novi model ponovimo dve iteraciji učenja.
8. S pomočjo zadnjih dobljenih modelov ponovno izvedeno časovno poravnavo in izločimo segmente, kjer poravnava ni uspela. Iz seznama preostalih segmentov ponovno izberemo začetno učno podmnožico. Na novi podmnožici tvorimo nove prototipne modele, ki jim dodamo še model za tišino.
9. Ponovimo 4 iteracije učenja na celotni učni množici, brez izločenih segmentov.
10. Grafemske transkripcije pretvorimo v trigrafemske.
11. Monofonske modele kopiramo v trifonske in ponovimo dve iteraciji učenja.
12. Tvorimo modele z vezanimi stanji in ponovimo dve iteraciji učenja.
13. Naučimo modele, ki uporabljajo mešane Gaussove porazdelitve z 2, 4, 8, 16 in 32 porazdelitvami. Tukaj modele z več porazdelitvami dobimo postopoma iz modela z manj porazdelitvami. Za vsak novi model ponovimo dve iteraciji učenja.

Na koncu postopka dobimo medbesedne trifonske modele z vezanimi stanji in več Gaussovimi porazdelitvami, ki smo jih kasneje uporabljali pri vrednotenju na testni množici.

4.3. Postopek nenadzorovanega učenja

Postopek nenadzorovanega učenja smo preizkušali na dva nekoliko različna načina. Razlikujeta se v načinu učenja modelov med posameznimi iteracijami razpoznavanja. Najprej smo iz učne množice izbrali 6 oddaj. Te predstavljajo prvo četrtino. Uporabljali smo tudi ročno izdelane transkripcije teh oddaj. Od preostalih treh četrtin pa smo uporabljali le zvočne posnetke. V nadaljevanju sta opisana oba postopka, ki smo ju preizkušali.

Prvi postopek je učenje od začetka. Tukaj smo po vsaki iteraciji razpoznavanja na preostalih treh četrtinah

učne množici učili nove modele od začetka. To pomeni, da nismo uporabili nobenih že prej izdelanih modelov. Natančen postopek je bil:

1. Izvedemo nadzorovano učenje na prvi četrtini učnih podatkov. Uporabimo ročno izdelane transkripcije.
2. S pomočjo pravkar izdelanih modelov izvedemo razpoznavanje na preostalih treh četrtinah učne množice.
3. Rezultate razpoznavanja združimo s transkripcijami prve četrtine. Tako dobimo novo transkripcijo, ki bo v naslednjem koraku služila učenju.
4. S enakim postopkom, kot smo ga uporabljali za nadzorovano učenje, naučimo nove modele. Uporabimo pravkar dobljene transkripcije.
5. Korake 2 do 4 ponovimo še dvakrat.

Tako smo izvedli eno iteracijo postopka nadzorovanega učenja in tri iteracije nenadzorovanega ter s tem postopkom dobili 4 nabore modelov. Prvi je dobljen na eni četrtini učne množice, ostali pa na celotni učni množici. V vsakem razpoznavanju smo uporabljali modele s 16 Gaussovimi porazdelitvami.

Drugi postopek je dodatno učenje. Tukaj smo po vsaki iteraciji razpoznavanja učili modele tako, da smo vzeli modele iz prejšnje iteracije in jih dodatno učili na treh četrtinah učne množice. Natančen postopek je bil:

1. Izvedemo nadzorovano učenje na prvi četrtini učnih podatkov. Uporabimo ročno izdelane transkripcije (enako kot v prvem postopku).
2. S pomočjo izdelanih modelov izvedemo razpoznavanje na preostalih treh četrtinah učne množice.
3. Rezultate razpoznavanja uporabimo kot transkripcijo za novo iteracijo učenja.
4. Izvedemo dodatno učenje modelov.
5. Uredimo slovar transkripcij. Izdelamo seznam grafemov. Besedne transkripcije pretvorimo v fonemske.
   a. Izvedemo časovno poravnavo med transkripcijami in zvočnimi posnetki. Uporabimo modele monofonske modele iz prvega koraka.
   b. Grafemske transkripcije razširimo v trigrafemske.
   c. Ponovimo dve iteraciji učenja na modeli z mešanimi porazdelitvami in vezanimi stanji. Pri tem v prvi iteraciji izhajamo iz modela, uporabljena v razpoznavanju.
6. Korake 2 do 4 ponovimo še dvakrat.

Tudi tukaj smo dobili 3 nove nabore modelov (prvi modeli so enaki kot v prvem postopku) in prav tako smo pri vseh razpoznavanjih uporabljali modele s 16 Gaussovimi porazdelitvami.

Ker je količina učnih podatkov premajhna, da bi lahko naučili modele za vse možne trifone, dobimo manjše število modelov, med katerimi pa nekateri predstavljajo več trifonov, ki imajo podobne akustične značilnosti. Najpomembnejša razlika med predstavlje-nima postopkoma je število končnih modelov. V prvem postopku (učenju od začetka) se v vsaki iteraciji na podlagi učnih podatkov na novo določi nabor izdelanih modelov. V drugem postopku (doučenje) pa število modelov ostaja enako kot pri nadzorovanem učenju na četrtini učne množice.

4.4. Ocenjevanje modelov

Modele smo ocenjevali na ročno segmentirani testni množici BNSI. Uporabili smo modele s 16 Gaussovimi porazdelitvami. Skupno smo ocenili 8 modelov:

- nadzorovano naučen model na eni četrtini učne množice,
- trije nenadzorovano naučeni modeli po prvem postopku,
- trije nenadzorovano naučeni modeli po drugem postopku,
- nadzorovano naučen model na celotno učni množici.

V vseh primerih smo uporabljali isti jezikovni model in razpoznavanje smo izvajali na isti strojni in programski opremi.

5. Rezultati

Rezultati uspešnosti razpoznavanja so prikazani v tabeli 1. V prvi vrstici je podatek za uspešnost razpoznavanja z *baseline* modelom naučenim nadzorovano na eni četrtini učne množice. V naslednjih treh vrsticah so podatki za modele, ki smo jih dobili v treh iteracijah nenadzorovanega učenja po obeh postopkih. V zadnji vrstici je podatek za uspešnost z modeli, ki so bili naučeni nadzorovano na celotni učni množici. Za vsakim nenadzorovanim modelom smo dosegli rezultate boljše od *baseline* modela. V obeh postopki opazimo izboljšanja uspešnosti, ki pa ne naraščajo vedno s številom iteracij učenja. Vidimo tudi, da prvi postopek daje rahlo boljše rezultate od drugega. Vzrok za to vidimo v večjem številu naučenih modelov za trifone. Največje izboljšanje, 1,35 % absolutno, smo dosegli s prvim postopkom v prvi iteraciji učenja.

Tabela 1. Rezultati WER razpoznavanja na testni množici.

Iteracija	Prvi postopek	Drugi postopek
Baseline	32,28 %	
1	30,93 %	31,06 %
2	31,01 %	31,27 %
3	31,36 %	30,99 %
Nadzorovano	28,34 %	

Tabela 2 prikazuje vrednosti faktorjev realnega časa, razmerjem med trajanjem razpoznavanja in dolžino posnetkov, za vse testirane modele. Drugače kot pa pri uspešnosti razpoznavanja se ti rezultati vedno izboljšujejo z večanjem števila iteracij. Prav tako vidimo, da dobimo rahlo boljše rezultate pri prvem postopku.

V tabeli 3 smo podali še rezultate za WER Recovery. Iz njih vidimo, da smo z nenadzorovanim učenjem zvečali uspešnost za približno tretjino zvišanja, ki ga dosežemo z nadzorovanim učenjem. Dobljena izboljšanja so sicer slabša, če jih primerjamo z rezultati dobljenimi v [8] in [12], vendar pa so vseeno koristna za izboljšanje delovanja razpoznavalnika.

Tabela 2. Faktorji realnega časa razpoznavanja.

Iteracija	Prvi postopek	Drugi postopek
Baseline	20,91	
1	20,56	21,94
2	20,26	21,18
3	15,67	18,96
Nadzorovano	18,74	

Tabela 3. Rezultati *WER Recovery*

Iteracija	Prvi postopek	Drugi postopek
1	34,3 %	31,0 %
2	32,2 %	25,6 %
3	23,4 %	32,7 %

6. Zaključek

V članku smo predstavili osnovno idejo nenadzorovanega učenja akustičnih modelov ter predstavili dva postopka, ki smo ju tudi preizkusili na razpoznavalniku tekočega govora za slovenski jezik. Z obema postopkoma smo uspeli izboljšati uspešnost razpoznavanja na testni množici, s tem da je bil postopek s ponovnim učenjem modelov od začetka rahlo boljši od postopka z dodatnim učenjem.

Predstavljen način učenja bi lahko uporabljali v trenutnem sistemu razpoznavalnika tekočega govora UMB BN. Z dodatnim akustičnim materialom bi lahko izboljšali uspešnost razpoznavanja. Pri tem bi izhajali iz modelov, naučenih na celotni učni množici, in bi to postopoma širili na večjo množico.

Zahvala

Delo je bilo delno sofinancirano s strani ARRS po pogodbah P2-0069 in 1000-10-310131.

Literatura

1. Lamel, L.; Gauvain, J.-L.; Adda, G. Unsupervised Acoustic Model Training. V: International Conference on Acoustics, Speech, and Signal Processing, Orlando, Florida, **2002**, I-887–I-880.
2. Žgank A.; Verdonik D.; Kačič Z. Slovenska baza BNSI Broadcast News za razpoznavanje tekočega govora. *Elektrotehniški vestnik* **2008**, 3, 85–90.
3. Lamel, L.; Gauvain, J.-L.; Adda, G.; Lightly Supervised and Unsupervised Acoustic Model Training. *Computer Speech and Language* **2002**, 1, 115–129.
4. Wessel, F; Ney, H. Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition. *IEEE Transactions on Speech and Audio Processing* **2005**, 1, 23–31.
5. Žgank, A.; Sepesy Maučec, M. Razpoznavalnik tekočega govora UMB Broadcast News 2010: nadgradnja akustičnih in jezikovnih modelov. V: Jezikovne tehnologije, Ljubljana, Slovenija, **2010**, 28–31.
6. Rabiner, L.R.; A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **1989**, 2, 257–286.
7. Biem, A.; McDermott, E.; Katagiri, S. A Discriminative Filter Bank Model for Speech Recognition, V: Fourth European Conference on Speech Communication and Technology, EUROSPEECH, **1995**, Madrid, Španija, 545–548.
8. Hermansky, H. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America* **1990**, 4, 1738–1752.
9. Gales, M.; Young, S. The Application of Hidden Markov Models in Speech Recognition. *Foundations and Trends in Signal Processing* **2007**, 3, 195–304.
10. Valtchev, V. Discriminative Methods in HMM-based Speech Recognition, Ph.D. Dissertation. Cambridge Unversity, Cambridge, UK, 1995.
11. Novotney, S.; Schwartz, R.; Ma, J. Unsupervised Acoustic and Language Model Training with Small Amounts of Labelled Data. V: International Conference on Acoustics, Speech, and Signal Processing, **2009**, Tajpej, Tajvan, 4297–4300.
12. Fraga-Silva, T.; Gauvain, J.-L.; Lamel, L. Lattice-based Unsupervised Acoustic Model Training. V: International Conference on Acoustics, Speech, and Signal Processing, **2011**, Praga, Češka, 4656–4659.
13. Young, S.; Evermann, G.; Gales, M.; Hain, T.; Kershaw, D.; Moore, G.; Odell, J.; Ollason, D.; Povey, D.; Valtchev, V.; Woodland, P. The HTK Book, version 3.4; Cambridge University Engineering Department, Cambridge, UK, 2006
14. Aubert X.L. An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition. *Computer Speech and Language* **2002**, 1, 89–114.
15. Arhar, Š.; Gorjanc, V. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* **2007**, 2, 95–110.