Razpoznavanje tekočega angleškega govora z bazo Broadcast News

Andrej Žgank, Zdravko Kačič, Bogomir Horvat

Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Smetanova 17, 2000 Maribor, Slovenija
e-mail: andrej.zgank@uni-mb.si

Povzetek

V članku bomo predstavili izdelavo razpoznavalnika tekočega govora neodvisnega govorca z velikim slovarjem besed za ameriško angleščino. Pri izdelavi sistema smo uporabili bazo Broadcast News 1996, ki vsebuje posnetke televizijskih in radijskih oddaj. Sistem razpoznavalnika govora smo zasnovali na osnovi zveznih prikritih modelov Markova. Po obdelavi učne baze smo začeli z učenjem modelov monofonov, ki smo jih kasneje razširili na trifone. Modele trifonov smo vezali z uporabo odločitvenega drevesa. Modele trifonov smo razširili na dva nabora trifonov. Število kombinacij porazdelitev gostot verjetnosti smo iz ene povečali na 32. Na osnovi teksta za izdelavo jezikovnega modela so bili izdelani bigramski in trigramski jezikovni modeli. Razpoznavali smo stavke iz referenčne baze. Dobljeni rezultati so primerljivi z rezultati dobljenimi na testu v organizaciji ameriške vladne agencije ARPA.

Abstract

In this paper, we will present development of speaker independent continuous speech recogniser with large vocabulary of words for American English. The system is based on Broadcast New 1996 audio database, which contains recordings of television and radio shows. The recogniser was designed with continuous hidden Markov models. After preparation of training database the development of monophones was done. Triphones were built from monophones and tied with use of the decision tree clustering. Two different sets of triphones were built. The number of Gaussian mixtures was increased from one to 32. With text corpora, different bigram and trigram language models were built. Recognition was performed on benchmark test database. The results achieved are comparable with the results from ARPA tests.

1. Uvod

V današnji dobi informatike in komunikacij dobivajo govorne tehnologije vedno večji pomen. Ena izmed teh tehnologij je tudi avtomatsko razpoznavanje govora. Področje uporabe razpoznavanja govora se vedno bolj širi iz laboratorijev v prakso. Sistemi za razpoznavanje govora se uporabljajo v telefoniji, računalništvu, nudenju informacij ... V zadnjih letih je bil dosežen velik napredek pri razvoju teh sistemov, obstaja pa še veliko možnosti za nadaljnje raziskovanje in izboljšave.

Baze govora, ki so bile v preteklosti uporabljene za razvoj sistemov razpoznavanja tekočega govora, so bile posebej narejene v ta namen. Izbran je bil korpus tekstov, ki so ga nato različni govorci prebrali največkrat v studijskem okolju ali ob uporabi telefona. V letu 1995 je ameriška vladna agencija ARPA (Garofolo, 1998; Pallet, 1998) prvič uporabila "najdeni" govor za izdelavo baze. Posneli so različne televizijske in radijske oddaje in jih pripravili v obliki baze za učenje razpoznavalnikov. Namen uporabe takšnih posnetkov je izboljšati robustnost razpoznavalnikov tekočega govora, saj je govor v takšni bazi v primerjavi s klasično bazo precej bolj raznolik in posnet v zelo različnih okoljih. Zaradi uporabe posnetkov se tudi precej skrajša čas, ki je potreben za dokončanje baze, saj na podlagi izkušenj z bazo SpeechDat(II) (Kaiser, 1998) traja snemanje govorcev približno polovico časa potrebnega za dokončanje baze. Takšen razpoznavalnik govora je možno uporabljati tudi za podnaslavljanje televizijskih oddaj, kar je bilo v obliki pomoči gluhim osebam izvedeno na Japonskem (Imai, 2000).

Pri izgradnji v članku predstavljenega sistema razpoznavanja tekočega govora neodvisnega govorca z velikim slovarjem besed smo uporabili bazo angleškega govora Broadcast News 1996, ki vsebuje posnetke enajstih različnih ameriških televizijskih in radijskih postaj. Akustični modeli so bili narejeni z zveznimi prikritimi modeli Markova, jezikovni modeli pa so N-gramski. Za izdelavo akustičnih modelov smo uporabili orodje HTK na HP delovnih postajah z operacijskim sistemom HP-UX. S to bazo je bil narejen referenčni test v organizaciji agencije ARPA, v katerem je sodelovalo devet različnih laboratorijev, njihovi rezultati pa nam bodo služili za primerjavo kakovosti sistema.

V drugem razdelku bomo na kratko predstavili splošno delovanje sistema razpoznavanja tekočega govora. V tretjem razdelku bomo opisali postopek izgradnje razpoznavalnika in podatke, ki smo jih uporabili pri tem. Rezultate in njihov komentar podajamo v četrtem razdelku.

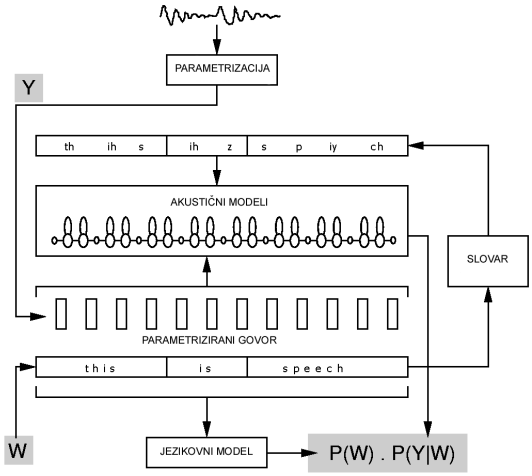
2. Razpoznavanje tekočega govora

Na področju razpoznavanja govora ločimo med dvema osnovnima pristopoma v načinu uporabe razpoznavalnikov. V prvem primeru sistem deluje za odvisnega govorca. To pomeni, da mora vsak uporabnik prilagoditi razpoznavalnik na značilnosti svojega glasu in način branja. Ta prilagoditev se izvede z branjem teksta. V drugem primeru je razpoznavalnik sposoben razpoznavati govor neodvisnega govorca. Takšen sistem je lahko tudi osnova za sistem razpoznavanja govora z odvisnim govorcem, seveda pa je potrebno izvesti adaptacijo na govorca.

Najtežja naloga na področju avtomatskega razpoznavanja govora je razpoznavanje tekočega govora neodvisnega govorca z velikim slovarjem besed. Takšen sistem omogoča razpoznavanje stavkov, ne da bi bili med besedami potrebni presledki. V slovarju besed takšnega razpoznavalnika je več kot 1.000 besed, ponavadi celo več

kot 20.000. Besede, ki je v slovarju ni, razpoznavalnik seveda ne more razpoznati. V takšnem primeru bo sistem razpoznal tisto besedo, ki ji je akustično najbolj podobna.

Osnovna ideja delovanje razpoznavalnika je, da sistem naučimo na določeni bazi govora, ki vsebuje posnetke, za katere pričakujemo, da so dovolj podobni tistim, ki jih želimo razpoznavati (Jelinek, 1997). Pri pristopu z uporabo prikritih modelov Markova (PMM) je osnovna zgradba razpoznavalnika prikazana na sliki 1 (Young, 1995).



Slika 1: Diagram razpoznavalnika govora s prikritimi modeli Markova. Parametrizirani govorni signal je označen z Y, zaporedje besed je označeno z W.

Kot vidimo na sliki 1 govorni signal najprej parametriziramo v vektor značilk Y v katerem so shranjene vse informacije o govornem signalu. Vektorje značilk potrebujemo za nadaljnjo obdelavo govora. S pomočjo akustičnih modelov, ki so narejeni s PMM in predstavljajo foneme jezika, dobimo najbolj verjetni zapis izgovarjave, glede na uporabljene akustične modele. Pri tem s pomočjo slovarja besed tvorimo možne kombinacije zaporedja besed W. Postopek, kjer bi stalno upoštevali vse možne kombinacije, bi bil prezahteven, zato manj verjetne možne stavke stalno opuščamo. Na nivoju akustičnega modeliranja se izkaže, da so rezultati boljši, če pri uporabi modelov fonemov upoštevamo tudi kontekst v katerem se ti nahajajo. Če upoštevamo predhodni in naslednji fonem dobimo tako imenovane trifone. Obstajata dva načina upoštevanja konteksta. V prvem primeru upoštevamo kontekst samo znotraj ene besede (notranjebesedni trifoni), v drugem pa upoštevamo tudi sosednje besede (medbesedni trifoni). Drugi pristop daje boljše rezultate, saj je v modelih upoštevan tudi učinek koartikulacije, vendar je takšen postopek zaradi večjega števila možnih trifonov računsko zahtevnejši.

Jezikovni model služi za oceno verjetnost zaporedja nastopajočih besed, najpogosteje pa uporabljamo N-gramske jezikovne modele (Rabiner, 1993). Na področju tekočega govora dosežemo najboljše rezultate s trigramskimi ali štirigramskimi jezikovnimi modeli (Pallet, et.al., 1997). Produkt ocene akustičnega in jezikovnega modela da najbolj verjetno hipotezo o stavku, ki je bil izgovorjen.

Prikrite modele Markova, ki jih poleg nevronskih mrež najpogosteje uporabljamo v takšnih sistemih, moramo pred uporabo naučiti z uporabo ustrezne baze izgovarjav. Pri uporabi baze posnetkov televizijskih in radijskih oddaj za učenje je vključen tako spontani, kot brani govor. Zaradi različnih javljanj s terena so posnetki v bazi posneti v različnih akustičnih okoljih. Te razlike napram klasični studijsko posneti bazi izboljšajo robustnost razpoznavalnika.

3. Izgradnja razpoznavalnika govora

3.1. Baza Broadcast News

Baza Broadcast News 1996 (Stern, 1997), ki smo jo uporabili pri razvoju in testiranju našega sistema, vsebuje:

- 35 ur posnetkov iz 11 oddaj: ABC Nightline, ABC World News Now, ABC World News Tonight, CNN Early Edition, CNN Early Primetime News, CNN Headline News, CNN Primetime News, CNN The World Today, CSPAN Washington Journal, NPR All Things Considered in NPR Marketplace,
- posnetke za testiranje med razvojem - development,
- posnetke za končno testiranje - evaluation,
- ortografske transkripcije vseh oddaj,
- dodatni tekstovni zapisi oddaj namenjeni izgradnji jezikovnega modela,
- informacije o bazi BN 1996.

Posnetki oddaj, ki jih je pripravil NIST, so bili prenešeni preko kabelskega omrežja in so jih simultano snemali na S-VHS videorekorder in DAT kasetofon. Radijske oddaje so bile posnete s pomočjo Hifi – stereo digitalnega FM sprejemnika in DAT kasetofona. Posnetki na DAT kasetah so bili posneti s frekvenco vzorčenja 32kHz. Frekvenca vzorčenja teh posnetkov je bila s pomočjo DAT-linka znižana na 16kHz. Ker razpoznavalnika govora ne moremo učiti s stereo posnetki, je bil uporabljen samo levi kanal, ki je bil ob 16 bitni kvantizaciji zapisan v PCM obliki. Datotekam so dodali glave za program Sphere, nakar so bile shranjene na CD-ROM.

Posnetki vsebujejo oddaje v celoti, tako da je v nekaterih primerih govoru v ozadju dodana glasba. Nekatere oddaje vsebujejo tudi dele pogovorov v živo, kjer se prepletata govora dveh ali več govorcev. Takšni kategoriji posnetkov sta zelo težavni za razpoznavanje govora, zato smo takšne posnetke izločili iz baze posnetkov. Prav tako smo izločili posnetke, pri katerih je bila kvaliteta govora slaba. Na podlagi teh kriterijev je bilo iz baze izločenih 41,3% posnetkov. Ker smo za razvoj našega razpoznavalnika uporabili orodje HTK, ki je namenjeno učenju prikritih modelov Markova, smo morali ustrezno pripraviti vse podatke. Za tvorbo fonetičnega zapisa oddaj smo ustrezno priredili ortografske zapise, pri pretvorbi pa uporabili fonetični slovar za ameriško angleščino, ki je bil izdelan na Univerzi Carnegie-Mellon.

Posnetke oddaj smo segmentirali v dele, ki niso bili daljši od 30 sekund, znotraj enega posnetka pa so bili akustični pogoji nespremenjeni (en govorec, enaka kvaliteta posnetka, ...). Govor smo nato ob uporabi Hammingovega okna pretvorili v vektorje značilk, ki jih je sestavljalo 12 kepstralnih koeficientov in energija (Young, 1997). Za boljše modeliranje dinamičnih sprememb v govornem signalu smo vektorju značilk dodali tudi prvi in drugi odvod. V vektorju značilk je bilo tako 39 elementov. Zaradi učne baze s posnetki različne kvalitete smo uporabili normalizacijo kepstralnih koeficientov.

Na enak način, kot smo obdelali ortografske transkripcije posnetkov, smo obdelali tudi tekstovni material, ki je namenjen izgradnji jezikovnega modela. K temu tekstu smo dodali pet kopij ortografskega zapisa posnetkov iz učne baze. Iz tega tekstovnega materiala smo izločili tudi 10.000 in 32.000 najbolj pogostih angleških besed, iz katerih sta bila nato generirana dva različno velika fonetična slovarja besed za razpoznavanje.

3.2. Kontekstno neodvisni modeli

Razpoznavalnik smo zgradili na osnovi zveznih prikritih modelov Markova. Na podlagi fonetičnega slovarja besed smo uporabili 40 različnih fonemov za ameriško angleščino. Dodali smo tudi poseben model "sil", ki označuje začetek in konec stavka, ter model "sp", ki označuje premor med dvema besedama. Uporabili smo levo – desni model monofona s tremi stanji in eno Gaussovo porazdelitvijo verjetnosti na stanje. Za večjo odpornost sistema na šum v posnetkih smo model "sp" vezali na model "sil" (Young, 1997).

V prvem koraku učenja monofonov smo uporabili fonetični zapis oddaj, brez časovnih oznak, od kdaj do kdaj je trajal posamezni fonem. Nato smo v modelih povečali število Gaussovih porazdelitev verjetnosti na 32. To je potrebno zaradi dejstva, da so si izgovarjave istega fonema celo pri enem govorcu dokaj različne (Rabiner, 1993). S takšnimi razširjenimi modeli smo izvedli ponovno razvrščanje, katerega rezultat je fonetični zapis s časovnim trajanjem posameznega fonema.

Z novim časovnim zapisom smo ponovno od začetka izvedeli učenje monofonov. V nasprotju s pričakovanji se je pri testiranju modelov izkazalo, da so rezultati s temi modeli slabši, kot s prvotnimi. Do tega je verjetno prišlo zaradi preslabih monofonskih modelov, ki smo jih pri tem uporabili. Zaradi tega smo za nadaljevanje učenja uporabili prvotne modele z eno Gaussovo porazdelitvijo verjetnosti na stanje.

3.3. Kontekstno odvisni modeli

Kontekst v katerem nastopi posamezni fonem smo upoštevali s pomočjo trifonov. Modele monofonov smo najprej razširili na vse možne trifone, ki so se pojavili v učnem besedilu. S temi modeli smo izvedli prehodno učenje.

Težava pri učenju s trifoni je, da se kar nekaj trifonov v učni bazi pojavi samo enkrat ali dvakrat. To je premalo, da bi lahko učinkovito izračunali vrednosti modela (Jelinek, 1997). Zaradi tega smo uporabili združevanje trifonov s pomočjo odločitvenega drevesa (Young, et.al., 1994). Pri tem postopku na osnovi fonetičnih razredov, v katere sodi posamični fonem, poiščemo akustično podobne foneme, katerih stanja združimo. Nato za izračun vrednosti takšnih združenih stanj uporabimo skupne posnetke govora.

Ker ne vemo, kateri trifoni se bodo pojavili v govoru, ki ga hočemo razpoznavati, smo vzeli iz fonetičnega slovarja za razpoznavanje vse notranjebesedne trifone in jih na osnovi združevanja z odločitvenim drevesom vezali na obstoječe modele. Tako smo dobili 9.695 notranjebesednih trifonov. Drugi nabor trifonov, ki je bil naučen s to učno bazo, je vseboval vse možne kombinacije monofonov, saj ni znan vrstni red besed pri razpoznavanju. Tako smo dobili 70.562 možnih medbesednih trifonov.

Pri dosedanjem postopku učenja smo v posameznem stanju uporabili samo eno Gaussovo porazdelitev na stanje. Sedaj smo ob prehodnem učenju modelov, v koraku po dva, povečali število porazdelitev verjetnosti do števila 32. Tako smo zagotovili boljše modeliranje trifonov. Ta dva nabora trifonskih modelov smo uporabili za razpoznavanje govora.

3.4. Jezikovni model

Za izgradnjo jezikovnih modelov smo uporabili orodje CMU-SLM Toolkit V2.0 (Clarkson, 1997). Za oba fonetična slovarja za razpoznavanje smo izgradili ločena N-gramska jezikovna modela. Pri uporabi jezikovnih modelov za razpoznavanje govora se v testnem naboru stavkov pojavijo zaporedja besed, ki jih v učni bazi ni bilo. Da bi takšnemu zaporedju besed s pomočjo jezikovnega modela pripisali neko verjetnost, smo pri izgradnji uporabili Good–Turingovo metodo za glajenje verjetnosti v jezikovnem modelu (Jelinek, 1997).

Vplivu zaporedja besed, ki se v velikem učnem materialu pojavijo zelo redko, smo se izognili tako, da smo vse N-grame, ki so se pojavili manj kot trikrat, izločili iz jezikovnega modela. Tako smo dobili bigramski in trigramski jezikovni model. V prvem smo upoštevali zaporedje dveh besed, v drugem pa zaporedje treh besed. Velikost in perpleksnost jezikovnih modelov vidimo v tabeli 1.

Model	1g	2g	3g	Perpl.
10k-2g	9.929	886.555	-	364,5
10k-3g	9.929	1.440.274	5.284.659	279,2
32k-2g	32.406	1.858.556	-	298,3
32k-3g	32.406	1.600.931	2.125.313	233,6

Tabela 1: Podatki o velikosti in lastnostih uporabljenih jezikovnih modelov

Kot vidimo v tabeli 1, je število bigramov različno za bigramske in trigramske jezikovne modele. Do tega je prišlo zaradi dejstva, da smo zaradi velikosti trigramskih modelov, pri njihovi izdelavi uporabili nekoliko drugačne parametre, kot pri gradnji bigramskih jezikovnih modelov. Perpleksnost za trigramske jezikovne modele je manjša kot za bigramske, kar je pričakovano. Razpoznavanje s trigramskimi jezikovnimi modeli je sicer računsko

zahtevnejše kot v primeru uporabe bigramskih modelov, vendar praviloma daje boljše rezultate (Graff, 1997).

4. Rezultati

Referenčna baza, ki smo jo uporabili za testiranje razpoznavalnika, je vsebovala 565 različnih stavkov. Posnetke in ortografsko transkripcijo teh stavkov smo obdelali na enak način, kot podatke za učno bazo. Pri razpoznavanju tekočega govora ločimo tri vrste napak: zamenjavo, brisanje in vrinjanje. Napako razpoznavanja besed izračunamo kot:

$$WER(\%) = \left(1 - \frac{N - S - D - I}{N}\right) \cdot 100 \ , \qquad (1)$$

pri tem je S število zamenjanih besed, D število izbrisanih besed, I število vrinjenih besed in N število vseh besed.

Za lažje optimiranje parametrov razpoznavalnika smo prve teste izvedli z notranjebesednimi trifoni in fonetičnim slovarjem velikosti 10.000 besed. Takšen sistem je deloval hitreje in tako omogočal lažje eksperimentiranje. Rezultate za takšno konfiguracijo vidimo v tabeli 2.

PMM	JM	Prav. (%)	Zam. (%)	Izb. (%)	Vrin. (%)	*WER (%)*
NB	2g	59,27	30,24	8,40	4,40	*44,81*
NB	3g	63,66	28,20	8,15	3,99	*40,33*
MB	2g	63,20	29,84	6,46	6,25	*42,05*

Tabela 2: Rezultati razpoznavanja za slovar 10.000 besed (PMM–prik. mod. Markova., NB–notranjebesedni trifoni, MB-medbesedni trifoni, JM-jezikovni model, Prav.-pravilne besede, Zam.-zamenjane besede, Izb.-izbrisane besede, Vrin.-vrinjene besede).

Kot lahko vidimo iz tabele 2 je bil rezultat najslabši za notranjebesedne trifone z uporabo bigramskega jezikovnega modela. Pri uporabi trigramskega jezikovnega modela in notranjebesednih trifonov je prišlo do izboljšanja rezultata za 4,5%, vendar je bila hitrost delovanja sistema zaradi tega manjša. To izboljšanje je bilo pričakovano, saj lahko s trigramskimi jezikovnimi modeli bolj natančno modeliramo stavke v tekočem govoru.

Pri uporabi medbesednih trifonov in bigramskega jezikovnega modela je prav tako prišlo do izboljšanja rezultata za 2,8% v primerjavi z enakimi notranjebesednimi trifoni. Tudi z uporabo medbesednih trifonov se je hitrost delovanja razpoznavalnika zmanjšala. Do izboljšanja rezultata z medbesednimi trifoni je prišlo zaradi dejstva, da so takšni trifoni zmožni boljše modelirati vplive koartikulacije v tekočem govoru.

Razpoznavalnik ne more pravilno razpoznati besede, ki je nima v slovarju. Pri velikosti slovarja 10.000 besed je v testnih posnetkih kar 23% besed, ki jih ni bilo v slovarju. Zaradi tega smo velikost slovarja povečali na 32.000 najbolj pogostih angleških besed. S tem se je število besed, ki jih ni bilo v slovarju zmanjšalo na 3,1%. Rezultate razpoznavanja za slovar z 32.000 besedami podaja tabela 3.

PMM	JM	Prav. (%)	Zam. (%)	Izb. (%)	Vrin. (%)	*WER (%)*
MB	2g	65,11	29,25	5,64	6,37	*41,27*
MB	3g	67,07	27,47	5,46	5,83	*38,76*

Tabela 3: Rezultati razpoznavanja za slovar 32.000 besed z medbesednimi trifoni (PMM–prik. mod. Markova., JM-jezikovni model).

Z uporabo večjega slovarja besed pri bigramskem jezikovnem modelu je prišlo do majhnega izboljšanja rezultatov. Zaradi boljše pokritosti besed v slovarju je bilo pričakovano večje izboljšanje rezultatov (Jelinek, 1997). Pri uporabi medbesednih trifonov s trigramskim jezikovnim modelom in slovarjem velikim 32.000 besed smo dosegli najboljše rezultate. WER za takšen sistem je znašal 38,76%. Slaba stran najboljše kombinacije uporabljene za razpoznavanje je, da je takšen sistem časovno najbolj zahteven, poleg tega pa zahteva tudi računalnike z veliko količino prostega spomina.

Na testu, ki ga je organizirala agencija ARPA (Pallet, et.al. 1997) in je služil za primerjavo rezultatov je bil najboljši sistem z WER 27,8%, najslabši sistem pa je dosegel rezultat 54,8%. Primerjava naših rezultatov pokaže, da so doseženi rezultati primerljivi s tistimi, ki so bili doseženi na testu ARPEe.

5. Zaključek

Predstavljen je bil sistem za razpoznavanje tekočega angleškega govora z velikim slovarjem besed. Doseženi rezultati so primerljivi z rezultati doseženimi na testu ARPEe. Glavna možna izboljšava sistema bi bila uporaba ločenih akustičnih modelov za posamezna okolja iz katerih so posnetki. Prav tako bi bilo možno uporabiti ločene akustične modele za ženski in moški spol. Težava takšne izboljšave bi bila, da bi sistem, ki že sedaj deluje počasi, postal še počasnejši in obsežnejši. Ena izmed možnosti za izboljšavo rezultatov, bi bila tudi povečanje akustične baze za učenje modelov. V primeru, da bi želeli razpoznavalnik uporabljati za narekovanje ali podnaslavljanje, bi bilo potrebno predvsem izboljšati hitrost delovanja. To bi bilo možno doseči z uporabo hitrejšega dekodirnika.

6. Literatura

Clarkson P.R., R. Rosenfeld., 1997, Statistical Language Modeling Using the CMU-Cambridge Toolkit. *Proc. Europ. Conf. Speech Proc. and Techn. (EUROSPEECH).*

Garofolo J. S. 1998, Design and Preparation of the 1996 Hub 4 Broadcast News Benchmark Test Corpora, *Proc. 1997 DARPA Speech Recognition Workshop*, Chantilly, Virginia, ZDA.

Graff D., 1997, The 1996 Broadcast News speech and language-model corpus, *Proc. 1997 DARPA Speech Recognition Workshop*, Chantilly, Virginia, ZDA.

Imai T., A. Kobayashi, S. Sato, H. Tanaka, A. Ando, 2000, Progressive 2-pass Decoder for Real-time Broadcast News Captioning. *Proc. International Conf.*

on Acoustic, Speech and Signal Processing, Istanbul, Turkey.

Jelinek, Frederic, 1997, *Statistical Methods for Speech Recognition*, MIT Press.

Kaiser, J., Z. Kačič, 1998, Development of the Slovenian SpeechDat database, *First International Conference on Language Resources and Evaluation*. Granada, Spain.

Pallet D.S., 1998, The NIST role in automatic speech recognition benchmark tests, *First International Conference on Language Resources and Evaluation,* Granada, Spain.

Pallet D.S., J. G. Fiscus, M. A. Przybocki, 1997, 1996 Preliminary Broadcast News Benchmark Tests, *Proc. 1997 DARPA Speech Recognition Workshop*, Chantilly, Virginia, ZDA.

Rabiner L., B. Juang, 1993, *Fundamentals of Speech Recognition*, Prentice Hall.

Stern R.M., 1997, Specification of the 1996 Hub 4 Broadcast News evaluation, *Proc. 1997 DARPA Speech Recognition Workshop*, Chantilly, Virginia, ZDA.

Young S., 1995, Large Vocabulary Continuous Speech Recognition: a Review, *Proc. IEEE Workshop on Speech Recognition*, Snowbird, Utah, ZDA,.

Young S., 1997, *The HTK Book Version 2.1*, Cambridge University.

Young S., J. Odell, P. Woodland, 1994, Tree-based state tying for high accuracy acoustic modelling. *Proc. ARPA Human Language Technology Conference,* Plainsboro, ZDA.