Razpoznavanje govora v domeni dnevno-informativnih oddaj

Speech Recognition in the Broadcast News Domain

Mirjam Sepesy Maučec*, Andrej Žgank

Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru / Smetanova 17, 2000 Maribor

E-mails: mirjam.sepesy@uni-mb.si ; andrej.zgank@uni-mb.si

* Avtor za korespondenco; Tel.: +386-2-220-72-25; Fax: +386-2-220-72-72

Povzetek: V članku bomo predstavili splošno zgradbo razpoznavalnika govora. Na kratko bomo opisali vse njegove ključne komponente.V nadaljevanju se bomo posvetili razpoznavalniku slovenskega tekočega govora UMB BN, ki trenutno predstavlja najkompleksnejši razpoznavalnik slovenskega govora, katerega razvoj poteka že nekaj let. Namenjen je razpoznavanju govora za domeno dnevno-informativnih oddaj. V članku se bomo posvetili tudi jezikovnim virom, ki so ključnega pomena ne samo pri razvoju razpoznavalnika ampak pri razvoju poljubnih jezikovnih oz. govornih tehnologij.

Ključne besede: tekoči govor; razpoznavanje govora; pregibni jezik.

Abstract: In this paper the structure of a speech recognition system is presented. We focus on Slovenian speech recognition. UMB BN system is described. Currently, it is the most complex large vocabulary continuous speech recognition system for Slovenian language. It is designed for transcribing TV-news shows. In the paper language resources are also described being crucial in developing such systems.

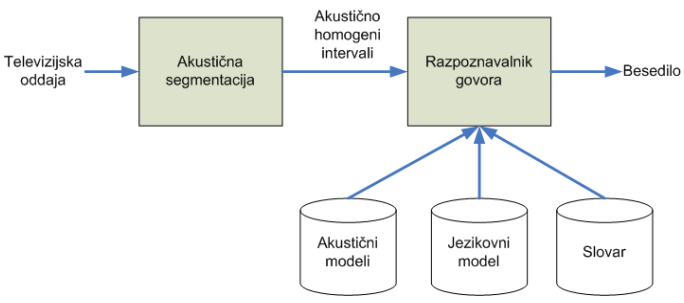
Key words: speech recognition; highly inflective language; language resources

1. Uvod

Telekomunikacijske storitve danes ponujajo uporabniku veliko količino multimedijskega gradiva. Kadar želimo v takšni količini gradiva poiskati želeno informacijo, smo prisiljeni uporabiti metode avtomatskega iskanja po vsebini. Iskanje po vsebini pa ni omejeno le na iskanje po besedilu, ampak tudi po zvočnih zapisih. Drug vidik je na primer tudi spremljanje televizijskega programa, ki bi ga želeli omogočiti tudi gluhim osebam. Ena izmed ključnih funkcionalnosti, ki to omogoča, je avtomatsko razpoznavanje tekočega govora. V članku bomo najprej predstavili splošno zgradbo razpoznavalnika. Sledila bo predstavitev jezikovnih virov in karakteristik razpoznavalnika govora, ki nastaja pod imenom UMB BN (kratica je okrajšava za University of Maribor, speech recognition system for Broadcast News domain).

2. Zgradba razpoznavalnika govora

Razpoznavalnik govora ima modularno zasnovo (slika 1). Sestavljajo ga modul za predprocesiranje govora, modul za akustično modeliranje, jezikovni model in modul za iskanje najbolj verjetnega zaporedja besed.



Slika 1. Modularna zasnova razpoznavalnika govora

2.1. Predprocesiranje govora

Razpoznavalnik zajema govor v digitalni obliki in ga zaznava v obliki akustičnega signala. Zajet signal pa ni čisti govor, ampak mu je prištet tudi šum. V prvem koraku razpoznavanja, pravimo mu akustični analizi, skušamo ločiti govor od šuma in predstaviti le informacijo, vsebovano v govornem signalu. Informacija v govornem signalu mora biti predstavljena na način, ki dobro razločuje med posameznimi fonemi. Danes najpogostejše metode za akustično analizo govornega signala temeljijo na t. i. mel frekvenčnih kepstralnih koeficientih (MFCC). Postopek MFCC matematično opisuje govorni signal na osnovi ocene kratkočasovne spektralne energije v različno definiranih frekvenčnih področjih. Poleg tako dobljenih t. i. statičnih značilk navadno uporabljamo še dinamične značilke: prve in druge odvode statičnih koeficientov po času. Le-ti skupaj s statičnimi koeficienti sestavljajo končni vektor značilk, ki ga uporabimo v nadaljnjem postopku razpoznavanja govora.

2.2. Akustično modeliranje

Naloga akustičnega modela je razpoznavanje fonemov v akustičnem signalu oz. čim bolj učinkovito predstaviti akustično-fonetične parametre glasu. Danes se za ta namen uporabljajo predvsem prikriti modeli Markova (HMM). Za tvorjenje (oz. učenje) akustičnih modelov HMM potrebujemo govorne baze, ki vsebujejo posnetke govora in natančen (fonemski) prepis tega, kar je bilo izgovorjeno. Pri učenju akustičnih modelov uporabljamo predvsem Baum-Welchev algoritem. Največkrat se uporabljajo podbesedni akustični modeli, kot so monofoni (ti so bolj ali manj ekvivalentni fonemom). Ker se izgovorjava istega fonema razlikuje glede na njegova sosednja fonema, se danes za akustično modeliranje najpogosteje uporabljajo trifonski akustični modeli, ki upoštevajo tudi levi (predhodni) in desni (sledeči) fonem oz. kontekst fonema. Podobna stanja akustičnih modelov med seboj združujemo, da zmanjšamo kompleksnost sistema.

2.3. Jezikovno modeliranje

Pri razpoznavanju tekočega govora so meje med besedami zabrisane, saj jih ne označujejo premori. Tudi nabor besed (tj. slovar), ki jih razpoznavalnik mora razlikovati, močno naraste, zato samo akustični model ni dovolj uspešen. Potreben je dodaten vir znanja o jeziku, ki mu pravimo jezikovni model. Medtem ko akustični model računa verjetnosti na ravni fonemov oz. trifonov, jezikovni model računa verjetnosti na ravni besed. Za učenje jezikovnega modela ne potrebujemo več baze izgovorjav, ampak velike elektronske zbirke besedil – korpuse, ki obsegajo več milijonov besed. Iz teh besedilnih zbirk jezikovni model računa verjetnosti zaporedij besed. Slovenski jezik sodi v skupino bolj kompleksnih jezikov, predvsem zaradi bogatega pregibanja besed in relativno nedoločenega vrstnega reda besed.

2.4. Iskanje najbolj verjetne hipoteze

Naloga razpoznavalnika govora je poiskati najbolj verjetni niz besed za zajeti vhodni govor. Iskanje izvedemo s pomočjo iskalnih algoritmov, ki vhodni govorni signal predstavljen z značilkami, modelirajo s pomočjo informacij iz akustičnega in jezikovnega modela. Pri iskanju najbolj verjetnega niza besed ni moč pregledati celotnega iskalnega prostora, ampak ga z različnimi heurističnimi metodami omejujemo. Razlikujemo statično omejevanje (npr. drevesna predstavitev slovarja) in dinamično omejevanje iskalnega prostora (npr. snopovno omejevanje, pogled-naprej v jezikovni model ipd).

3. Jezikovni viri

Preden se lahko lotimo snovanja razpoznavalnika govora potrebujemo različne vire znanja o jeziku. Za učenje akustičnih modelov potrebujemo govorne baze, za učenje jezikovnih modelov pa besedilne korpuse. V tem poglavju bomo na kratko predstavili vire, ki smo jih uporabili pri gradnji razpoznavalnika UMB BN.

3.1. Govorna baza

Osnovno učenje akustičnih modelov sistema UMB BN smo izvedli z govornim korpusom slovenske baze BNSI Broadcast News (Žgank et al., 2004), ki obsega 36 ur zapisanega govornega materiala iz obdobja 1999-2003. V korpus so vključene različne dnevno-informativne oddaje RTV Slovenija (TV Dnevnik, Odmevi). Govorni posnetki so bili v celoti ročno segmentirani, označeni in zapisani. Slovenska baza BNSI je dostopna pri evropski organizaciji ELRA/ELDA.

Postopek priprave na učenje akustičnih modelov zahteva dodatno ročno delo za pripravo in uskladitev vseh vključenih jezikovnih virov, predvsem če želimo doseči visoko kvaliteto razpoznavanja govora.

3.2. Besedilni korpusi

Za učenje jezikovnih modelov smo uporabili štiri besedilne korpuse: BNSI-Speech, BNSI-Text, Večer in FidaPLUS. BNSI-Speech vključuje transkripcije govornega korpusa, ki smo ga uporabili za učenje akustičnih modelov. BNSI-Text je besedilni korpus baze BNSI in vsebuje scenarije dnevno-informativnih oddaj

RTV SLO. Večer je korpus člankov časopisa Večer in obsega 100 mio besed. Najobsežnejši je korpus FidaPLUS. FidaPLUS korpus predstavlja razširitev korpusa FIDA. FIDA je referenčni korpus slovenskega pisanega jezika in obsega 100 mio besed iz različnih tekstovnih virov iz obdobja 1990-1999. FidaPLUS obsega 621 mio besed in dodaja besedila iz tekstovnih virov iz obdobja 1999-2006. Večji del korpusa predstavljajo časopisni in revijalni članki ter knjige. Nekaj besedil izvira tudi iz Spleta. Med govorjenim (predvsem spontano govorjenim) in pisanim jezikom je velika razlika. Stavki v korpusih pisanega jezika so daljši od izjav v korpusih govorjenega jezika. Številnih pojavov, ki so značilni za govor (npr. uporaba mašil, ponavljanje, napačni starti ipd.), v pisanem jeziku ne zasledimo. Pri gradnji jezikovnega modela je zato treba poskrbeti za uravnoteženi vpliv različnih jezikovnih virov.

tam ne kje konec maja oziroma sredini maja

Slika 2. Akustični signal in njegov ortografski zapis.

4. Razpoznavalnik UMB Broadcast News

Sistem UMB Broadcast News trenutno predstavlja najkompleksnejši razpoznavalnik slovenskega govora (Žgank et al., 2008, 2010). Na univerzi v Mariboru ga razvijamo že več let. Na vzorcu, ki ga uporabljamo za vrednotenje rezultatov dosegamo 71% uspešnost. Rezultat je primerljiv z rezultati podobnih sistemov, ki so bili razviti za druge jezike. V tem poglavju bomo predstavili osnovne karakteristike razpoznavalnika.

Sistem temelji na eni iteraciji razpoznavanja. Slovar sistema je obsegal 64.000 besed. Pri vzorčenju smo uporabili okno 25 ms s korakom 10 ms. V naši raziskavi smo opravili primerjavo dveh najpogosteje uporabljenih postopkov izločanja značilk. To so mel frekvenčni kepstralni koeficienti (MFCC) in koeficienti perceptivnega linearnega napovedovanja (PLP) . Dimenzija vektorja značilk je bila 39 (12 MFCC koeficientov, energija ter 1. in 2. odvod). Ker je naš cilj delovanje sistema v realnem času, smo dimenzijo zmanjšali na 27. Za akustično modeliranje smo uporabili medbesedne trigrafemske modele, ki dobro modelirajo koartikulacijo na besednih mejah. Obseg besedilnih korpusov je upravičil uporabo trigramskega interpoliranega jezikovnega modela, katerega perpleksnost na vzorcu za vrednotenje je znašala 246. Upoštevaje vnaprej definiran slovar besed, je bilo 4,22% besed izven slovarja.

5. Zaključek

V članku smo predstavili zgradbo razpoznavalnika govora in podali nekaj karakteristik razpoznavalnika UMB BN. Z uporabo različnih nadgradenj akustičnih in jezikovnih modelov in dodatnih korakov predprocesiranja uspešno izboljšujemo rezultate razpoznavanja. V naslednjem koraku razvoja bomo sistemu UMB BN dodali dodatno iteracijo razpoznavanja govora.

Zahvala

Avtorja članka se zahvaljujeta vsem sedanjim in bivšim sodelavcem Laboratorija za digitalno procesiranje signalov, FERI, UMB, ki so prispevali svoje znanje pri razvoju razpoznavalnika UMB BN. Hvala tudi avtorjem korpusa FidaPLUS, ki so nam omogočili njegovo uporabo. Zahvaljujemo se tudi agenciji ARRS za delno sofinanciranje po pogodbi št. P2-0069. Zahvala gre tudi urednikom znanstvene revije PAZU, ki se trudijo v deželo ob Muri pripeljati na prvi pogled nedostopne dosežke znanosti.

Literatura

1. Žgank, A., Rotovnik, T., Sepesy Maučec, M., Verdonik, D., Kitak, J., Vlaj, D., Hozjan, V., Kačič, Z., Horvat, B., Acquisition and annotation of Slovenian broadcast news database. *Zbornik konference LREC,* Lizbona, Portugalska, 2004.
2. Žgank, A.; Kos, M.; Kotnik, B.; Sepesy Maučec, M., Rotovnik; T., Kačič, Z., Nadgradnja sistema za razpoznavanje slovenskega tekočega govora UMB Broadcast news. *Jezikovne tehnologije*, Ljubljana, Slovenija, 2008.
3. Žgank, A.; Sepesy Maučec, M. Razpoznavalnik tekočega govora UMB Broadcast News 2010: nadgradnja akustičnih in jezikovnih modelov. *Jezikovne tehnologije*, Ljubljana, Slovenija, 2010.