Detektiranje pogrešanih značilk v šumnem okolju

Damjan Vlaj*, Zdravko Kačič, Bogomir Horvat****

* Center za interdisciplinarne in multidisciplinarne raziskave in študije Univerze v Mariboru
Razlagova 22, 2000 Maribor, Slovenija
damjan.vlaj@uni-mb.si
** Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Smetanova 17, 2000 Maribor, Slovenija
kacic@uni-mb.si, bogo.horvat@uni-mb.si

Povzetek

Pravilnost razpoznavanja se v procesu avtomatskega razpoznavanja govora občutno zmanjša, kadarkoli je prisoten šum. Zaradi dejstva, da je šum v ozadju v svetu resničnih aplikacij skorajda vedno prisoten, je bilo razvitih precej tehnik, ki poskušajo zmanjšati vpliv šuma v procesu razpoznavanja govora. Uporaba tehnike predprocesiranja govora predstavlja eno od možnosti za zmanjšanje vpliva šuma na uspešnost razpoznavanja.

V članku podajamo teoretične osnove in možnosti uporabe teorije pogrešanih značilk v primeru avtomatskega razpoznavanja govora v šumnem okolju. Problem pogrešanih značilk se pojavi, ko je govorni segment filtriran, prekinjan ali maskiran s šumom. Teste smo izvedli na studijski bazi TIDIGITS, ki smo ji dodali različne šume, in na slovenski telefonski bazi SpeechDat II, ki vsebuje šum iz naravnega okolja. Dosežena uspešnost razpoznavanja slovenske telefonske baze SpeechDat II je bila pri opravljenih testih za 23% boljša od klasičnih metod predprocesiranja govora.

Abstract

The problem of the automatic speech recognition is that whenever noise is present, the accuracy level of the recognition process is significantly reduced. Owing to the fact that in the world of real applications noise almost always exists in the background, several techniques have been developed to reduce the influence of noise on the speech recognition process. One of the possibilities to reduce the influence of noise on speech signal is to use pre-processing speech enhancement techniques.

In this article we discuss the theoretical base and the possibilities for the use of the Missing Feature Theory for automatic speech recognition in noisy environment. Missing feature problem occurs when speech segments are filtered, interrupted or masked by noise. We made our tests using the studio database TIDIGITS, to which different noise types were added, and Slovenian telephone database SpeechDat II, which contained noise from the natural environment. Using the Slovenian database SpeechDat II and our test methods, we managed to achieve the recognition accuracy level, which was by 23 % better than the one achieved by the use of classical methods.

1. Uvod

Človekova želja je, da bi komuniciral s strojem na enak način kot z ljudmi, torej z govorom in ne preko kombinacije tipk, stikal ali ročic. Področje avtomatskega razpoznavanja govora postaja z razvojem računalništva vse pomembnejše, zato se v tovrstne raziskave vlaga čedalje več sredstev. Cilj raziskav je zagotoviti, da bi v prihodnosti človek komuniciral z računalnikom kot s sočlovekom. Problem avtomatskega razpoznavanja govora je interdisciplinarne narave, saj zajema različna področja družboslovnih (fonetika, lingvistika, ...) in tehniških znanosti (procesiranje signalov, računalništvo, ...). Za uspešno reševanje problemov na področju avtomatskega razpoznavanja govora je neizbežno potrebno njihovo tesno medsebojno sodelovanje.

Pri avtomatskem razpoznavanju govora signal najprej spremenimo v stroju razumljivo obliko. Ker avtomatsko razpoznavanje govora izvajamo na digitalnih računalnikih, govor najprej filtriramo (da zagotovimo izpolnjevanje Nyquistovega kriterija) in nato vzorčimo z ustrezno frekvenco (8-16 kHz) ter kvantiziramo na določeno število diskretnih nivojev (256, 1024 ali 65536, kar ustreza 8, 10 ali 16 bitni kvantizaciji). Tako dobljen signal po potrebi še nadalje filtriramo, kar je odvisno od okolja, v katerem zajemamo govor (na primer filtriranje na kvaliteto telefonskega govora v pasu od 300 do 3400 Hz ali kompenzacija vpliva telefonskega voda). Zaradi časovnega spreminjanja značilnosti govora digitaliziran signal razdelimo v manjše odseke, za katere predpostavimo konstantne značilnosti. Ta proces imenujemo oknjenje govornega signala. Velikost oziroma dolžina okna znaša običajno od 25 do 50 ms, kar zagotavlja določeno stacionarnost značilnosti signala v oknu, saj za normalni govor velja, da ne nastopijo večje spremembe v signalu od približno 30 do 50 ms.

Iz vsakega majhnega segmenta govornega signala, ki ga dobimo pri oknenju, skušamo izločiti določene značilke. Faza izločanja značilk ima dva namena. Primarni namen je, da skušamo odpraviti razlike med posameznimi govorci. Neposredna primerjava posameznih segmentov govornega signala je namreč neprimerna. Dva segmenta, ki sta jih izgovorila dva različna govorca, sta si lahko slušno zelo podobna, se pa njuna časovna poteka močno razlikujeta. Sekundarni pomen imajo same metode, ki jih uporabljamo za izločanje značilk, saj zmanjšajo količino podatkov, potrebnih za nadaljnje procesiranje govora. Izločene značilke nato v procesu razvrščanja primerjamo z referenčnimi. Pri prikritih modelih Markova (HMM) ugotavljamo, kateri model, naučen z referenčnimi značilkami, je z največjo verjetnostjo generiral zaporedje značilk.

Postopek izločanja značilk ima v procesu avtomatskega razpoznavanja govora velik pomen in zelo vpliva na uspešnost sistemov avtomatskega razpoznavanja govora. Še tako dobre metode za razvrščanje značilnosti so neučinkovite, če so metode za izločanje značilnosti govora nejasne in nezmožne izločiti ustrezne karakteristične značilnosti posameznih segmentov govornega signala. Od metod izločanja značilk zahtevamo naslednje lastnosti:

- jasno morajo razlikovati različne segmente govornega signala,
- izločiti morajo le informacijo, potrebno za razpoznavanje oziroma razumevanje govora,
- ne smejo razlikovati enakih segmentov govornega signala, ki so jih izgovorili različni govorci.

V praksi pa je problem veliko bolj pereč, saj metode za izločanje značilk govora, ki jih trenutno uporabljajo v svetu, ne izpolnjujejo v celoti nobenega od zgoraj naštetih pogojev. Težave pri obstoječih metodah za izločanje značilk so naslednje:

- razlikovanje posameznih različnih segmentov govornega signala ni zmeraj natančno, zato lahko pride do prekrivanja značilnih vektorjev različnih segmentov govornega signala,
- značilni vektorji enakih segmentov govornega signala različnih govorcev se lahko med seboj razlikujejo,
- značilni vektorji različnih segmentov govornega signala različnih govorcev se lahko prekrivajo.

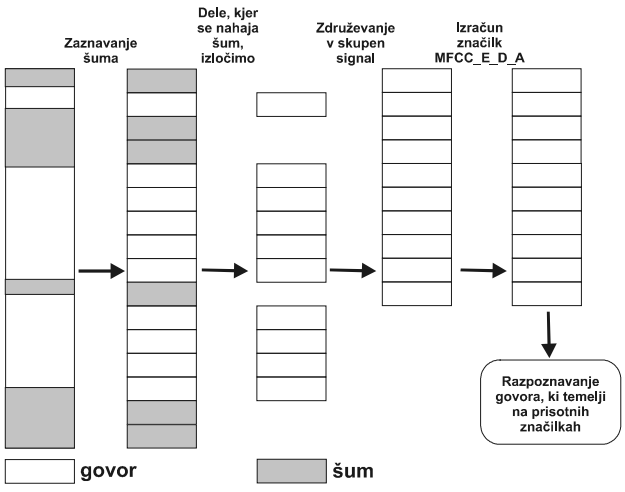
Težava pri raziskavah na področju izločanja značilk je tudi v tem, da ne vemo natančno, kaj so pravzaprav poglavitne značilnosti govornega signala, ki bi nam zagotovile prej navedene lastnosti.
Pri avtomatskem razpoznavanju govora lahko neugodni akustični pogoji povzročijo popačenje ene ali več komponent pri vhodnih vektorskih značilkah. Ko značilka zavzame neobičajne vrednosti in če ne ukrepamo, da bi obravnavali te motene značilke na drugačen način kot nemotene, lahko pričakujemo, da bo razpoznavanje neuspešno oziroma nemogoče.
Ljudje lahko razpoznavamo govor z običajno pojavljajočimi se motnjami, ki jih povzročajo senca glave ali šumi v okolju. Razpoznavamo lahko tudi govor, ki so mu bili dodani šumi, ki jih povzročajo moderne komunikacijske naprave.
Problem zmanjšanja kakovosti avtomatskega razpoznavanja govora v prisotnosti dodanega šuma raziskujejo že več let. V splošnem so si raziskovalci prizadevali, da bi zagotovili odpornost sistemov za avtomatsko razpoznavanje govora proti šumu na tri načine (Kermorvant & Morris, 1999):

- z uporabo značilk, odpornih proti šumu (primer so sistemi, ki temeljijo na odštevanju spektrov),
- z uporabo razpoznavalnih statističnih modelov za šum (primer je uporaba vzporedne modelne kombinacije),
- z uporabo dodeljenega merila, ki je robusten do šuma.

Da bi izboljšali avtomatsko razpoznavanje govora v neugodnih akustičnih pogojih, so na področju robustnega procesiranja govora uvedli teorijo, imenovano teorija pogrešanih značilk (Missing Feature Theory). Z uporabo le zanesljivih delov akustične informacije in neupoštevanjem nezanesljivih vektorjev značilk lahko mnogokrat uspešnost avtomatskega razpoznavanja govora skoraj ohranimo na istem nivoju, kot je v nemotenih pogojih.
Pri običajni metodi avtomatskega razpoznavanja govora najprej izračunamo značilni vektor za signal, ki vsebuje tako govor kot šum. Nato ta značilni vektor uporabimo za avtomatsko razpoznavanje govora z modeliranjem govora s prikritimi modeli Markova.
Pri naših raziskavah smo uporabili dve metodi za odkrivanje šuma v signalu. Dele signala, ki so vsebovali šum, smo izločili iz signala. Tako smo za signale, ki so vsebovali govor, izračunali vektorje značilk, ki smo jih imenovali *prisotne značilke*. Za dele signala, ki smo jih izločili zaradi prisotnosti šuma, vektorjev značilk nismo izračunali, zato smo jih imenovali *pogrešane značilke*. Nato smo vektor značilk, ki temelji samo na prisotnih značilkah, uporabili za avtomatsko razpoznavanje govora z modeliranjem govora s prikritimi modeli Markova.

2. Teorija pogrešanih značilk

Problem pogrešanih značilk se pojavi, ko je govorni segment filtriran, prekinjan ali maskiran s šumom. Večina raziskovalcev predpostavlja, da signal, ki je maskiran s šumom, poslabša pravilnost avtomatskega razpoznavanja govora. Zato raziskovalci (Lippmann & Carlson, 1997; El-Maliki & Drygajlo, 1999) z raznimi metodami določijo, kateri deli govornega signala so maskirani s šumom in kateri ne. Po izračunu značilnih vektorjev celotnega govornega signala in detektiranju segmentov signala, ki so maskirani s šumom, predpostavijo, da so značilke, dobljene v delu govornega signala, ki je maskiran s šumom, pogrešane. Za nadaljnje delo uporabljamo samo prisotne značilke, torej avtomatsko razpoznavanje govora temelji samo na prisotnih značilkah.
V naših raziskavah smo se lotili nekoliko drugačnega pristopa. V prvem koraku smo detektirali šum s količino energije v govornem signalu. Energijo smo izračunali vsakih 10 ms skozi celotni signal. Za odseke govornega signala, ki so vsebovali majhno količino energije, smo predpostavili, da vsebujejo šum. Tako smo odseke s šumom odstranili iz govornega signala. Za del signala, ki je vseboval samo govor, smo izračunali mel kepstralne vektorje značilk z dodatkom koeficientov energije ter prvega in drugega odvoda značilk (MFCC_E_D_A). Nato smo izvedli avtomatsko razpoznavanje govora z modeliranjem govora s prikritimi modeli Markova (Young, 1997), ki temelji samo na prisotnih značilkah.



Slika 1: Postopek izločanja govornega signala, ki je maskiran s šumom.

3. Detektiranje šuma

Kot smo zapisali že v predhodnem poglavju, smo segmente govornega signala, v katerih se nahaja šum,

določili s pomočjo energije, izmerjene v posameznem segmentu. Energijo smo izračunali z enačbo:

$$E = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} x^2[i]} \ , \qquad (1)$$

kjer je $x[i]$ segment govornega signala in N število vzorcev v tem segmentu.
Količino energije, izmerjene v posameznih segmentih, smo detektirali na dva načina. V prvem primeru smo izmerili energijo na začetku signala v dolžini 100 ms, pri tem pa smo predpostavili, da se v prvih 100 ms nahaja le šum in ne govor. Za nadaljnje delo smo uporabili le signal, ki je zadostil naslednji enačbi:

$$E = \begin{cases} E_{samp} & , \quad E_{samp} > E_{noise} \\ 0 & , \quad E_{samp} \le E_{noise} \end{cases} , \qquad (2)$$

pri tem je E_{samp} izmerjena energija v dolžini 10 ms in E_{noise} izmerjena energija v dolžini 100 ms na začetku govornega signala. Pri tej metodi smo iz nadaljnjega postopka izločili segmente signala, ki vsebujejo manj oz. enako energije, kot jo vsebuje segment, za katerega smo predpostavili, da je šumen.
Pri drugi metodi smo izmerili količino energije na začetku govornega signala in skozi celotni govorni signal. Nato smo izračunali razmerje med energijo celotnega govornega signala in energijo, ki smo jo izmerili na začetku. Z izpolnitvijo posameznih pogojev iz enačbe (3) smo energijo šuma povečali za pripadajoče faktorje.

$$E_{tmp} = \begin{cases} 15 \cdot E_{noise} & , \quad E_{whole} / E_{noise} > 100 \\ 8 \cdot E_{noise} & , \quad 50 < E_{whole} / E_{noise} \le 100 \\ 4 \cdot E_{noise} & , \quad 25 < E_{whole} / E_{noise} \le 50 \\ 2 \cdot E_{noise} & , \quad 12 < E_{whole} / E_{noise} \le 25 \\ 1{,}5 \cdot E_{noise} & , \quad 5 < E_{whole} / E_{noise} \le 12 \\ 1{,}25 \cdot E_{noise} & , \quad 2 < E_{whole} / E_{noise} \le 5 \\ E_{noise} & , \quad E_{whole} / E_{noise} \le 2 \end{cases} , \qquad (3)$$

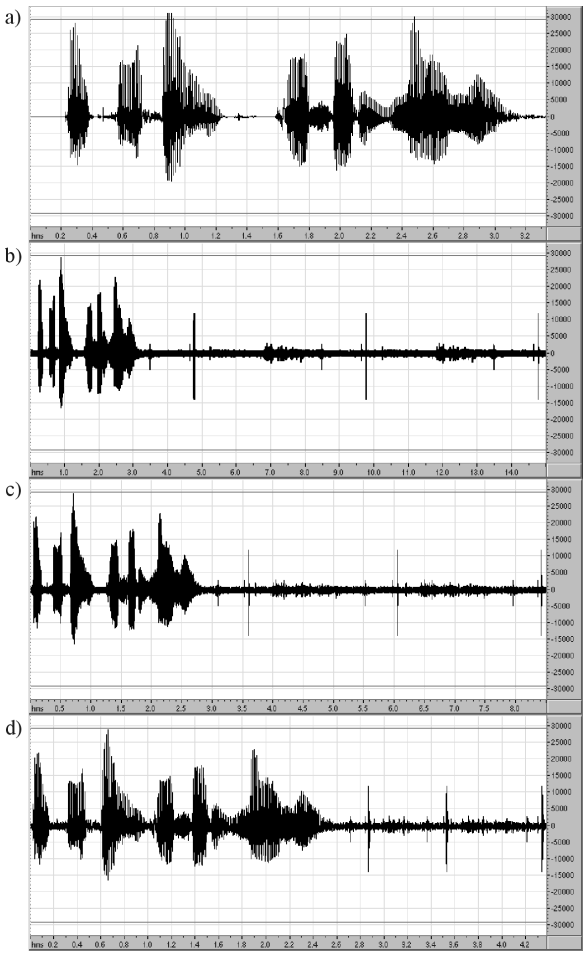
kjer je E_{samp} izmerjena energija v dolžini 10 ms, E_{whole} izmerjena energija celotnega signala, E_{noise} izmerjena energija v dolžini 100 ms na začetku govornega signala in E_{tmp} trenutno izmerjena energija, ki jo uporabimo v naslednji enačbi:

$$E = \begin{cases} E_{samp} & , \quad E_{samp} > E_{tmp} \\ 0 & , \quad E_{samp} \le E_{tmp} \end{cases} . \qquad (4)$$

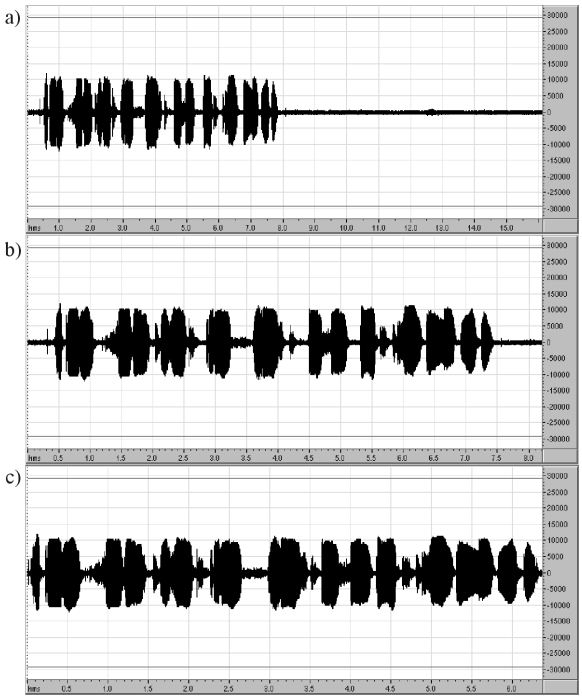
Pri drugi metodi smo iz nadaljnjega postopka izločili segmente signala, ki vsebujejo manj oz. enako energije, kot jo vsebuje šumen segment pomnožen z določenim faktorjem. Kakšen faktor je izbran, je odvisno od razmerja med energijo celotnega signala in energijo šumnega segmenta, kar je razvidno iz enačbe (3).

3. Testi z bazama TIDIGITS in SpeechDat II

Teste smo izvedli na dveh bazah. Baza TIDIGITS (Leonard & Doddington, 1991) je studijska in vsebuje nize števk v ameriški angleščini. Za teste smo uporabili niz sedmih števk. Za učenje in testiranje smo uporabili dvakrat po 1242 govorcev. Posnetkom smo dodali dvajset različnih šumov, ki smo jih predhodno izločili iz slovenske baze SpeechDat II. Izbrali smo šest razmerij signal šum med 16 dB in –4 dB. Šum smo dodali na takšen način, da bi se najbolj približali razmeram, ki so prisotne v bazi SpeechDat II. Na sliki 2 so prikazani primeri posnetkov, ki smo jih uporabljali pri testih z bazo TIDIGITS.
Slovenska baza SpeechDat II (Kaiser & Kačič, 1997) je bila posneta preko telefonskega omrežja. Za teste smo uporabili le del baze, ki vsebuje nize števk. Pri tem smo za učenje uporabili 300 govorcev in za testiranje 150 govorcev. K slovenski bazi SpeechDat II nismo dodali dodatnih šumov, saj ima baza že sama po sebi razmerje signal šum 25 dB. Na sliki 3 so prikazani primeri posnetkov, ki smo jih uporabljali pri testih s slovensko bazo SpeechDat II. Slika 3.a) prikazuje originalni signal iz baze, slika 3.b) predstavlja signal, ki smo ga dobili po izvedbi predprocesiranja s prvo metodo, in slika 3.c) signal, ki smo ga dobili po izvedbi predprocesiranja z drugo metodo.



Slika 2: Posnetki iz baze TIDIGITS: a) original, b) dodan šum – razmerje signal šum je 8 dB, c) izločanje šuma s prvo metodo in d) izločanje šuma z drugo metodo.



Slika 3: Posnetki iz baze SpeechDat II: a) original, b) izločanje šuma s prvo metodo in c) izločanje šuma z drugo metodo.

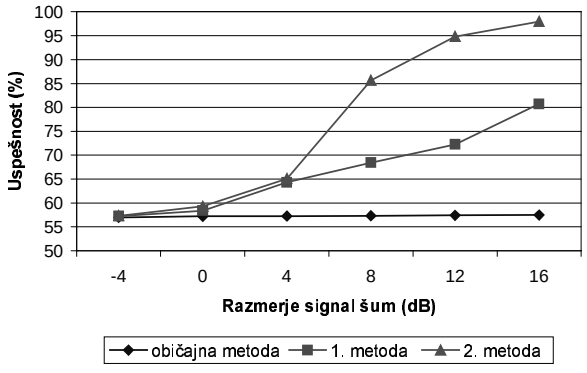
4. Rezultati

Uspešnost razpoznavanja smo določili z enačbo:

$$Uspešnost = \frac{H - I}{N} \times 100\% \ , \qquad (5)$$

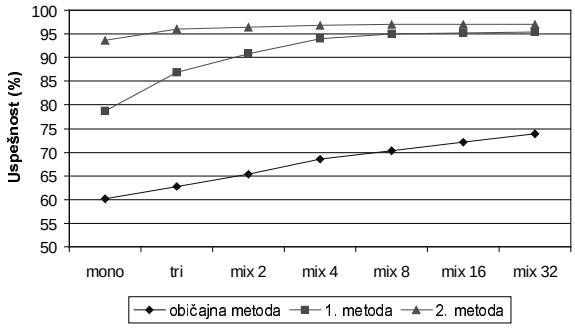
pri tem je H število pravilno razpoznanih besed, I število vrinjenih besed in N število vseh besed, ki so bile uporabljene pri testu.

Uspešnost razpoznavanja pri bazi TIDIGITS je 99,08% v primeru, ko ni dodanega šuma. S slike 4 je razvidno, da se uspešnost razpoznavanja zmanjšuje, kakor hitro se zmanjšuje razmerje signal šum. Zaželeno je, da bi bila uspešnost razpoznavanja čim dalj časa na visokem nivoju ob zmanjševanju razmerja signal šum. Z drugo metodo smo to tudi dosegli, kot je razvidno iz slike 4.



Slika 4: Uspešnost razpoznavanja pri testih z bazo TIDIGITS z različnimi razmerji signal šum.

Naslednji testi so bili narejeni s slovensko bazo SpeechDat II. Z njimi smo želeli ugotoviti učinkovitost metod, kadar uporabimo bazo, ki vsebuje veliko različnih vplivov iz okolice. Na sliki 5 je razvidna uspešnost razpoznavanja testov, ko smo uporabili monofone, trifone in trifone z različnimi Gaussovimi porazdelitvenimi verjetnostmi (mix 2 - 32). Oznaki *mono* in *tri* na sliki 5 predstavljata monofone in trifone z eno Gaussovo porazdelitveno verjetnostno funkcijo. Obe metodi doprineseta k boljši uspešnosti razpoznavanja besed. Tako na primer z običajno metodo dosežemo uspešnost razpoznavanja 74% pri trifonih z 32-imi Gaussovimi porazdelitvenimi verjetnostnimi funkcijami. Z uporabo prve metode dobimo 95,33%, z drugo pa 97,08%, kar pomeni, da smo povečali uspešnost razpoznavanja za več kot 23%. Druga metoda še posebej doprinese k uspešnosti razpoznavanja besed, ko uporabljamo monofone z eno Gaussovo porazdelitveno verjetnostno funkcijo.



Slika 5: Uspešnost razpoznavanja pri testih s slovensko bazo SpeechDat II z različnimi Gaussovimi porazdelitvenimi verjetnostmi.

5. Zaključek

V članku smo analizirali možnosti uporabe teorije pogrešanih značilk pri razpoznavanju govora v šumnem okolju. Deli signala, ki so maskirani s šumom, pokvarijo uspešnost razpoznavanja govora. Šumni deli signala vsebujejo značilke, ki smo jih imenovali pogrešane. Pri raziskavi smo uporabili dve metodi za detektiranje pogrešanih značilk v šumnem okolju. Del govornega signala, ki je vseboval šum, smo iz nadaljnjih raziskav odstranili. Kot so pokazali rezultati, je druga metoda prinesla boljše rezultate, vendar je njen problem v tem, da je ne moremo uporabiti pri avtomatskem razpoznavanju govora, ki bi potekalo v realnem času. Pri prvi metodi je to možno.

V prihodnosti želimo narediti teste na celotni slovenski bazi SpeechDat II, tako da bi dobili rezultate, ki bi nam pokazali, kako uspešni sta metodi v razmerah, ki ponazarjajo resnično govorno okolje.

Literatura

El-Maliki M. & Drygajlo A., 1999. Missing Features Detection and Handling for Robust Speaker Verification, *6th European Conference On Speech Communication And Technology - Eurospeech'99*, str. 975-978, Budimpešta, Madžarska.

Kaiser J. & Kačič Z., 1997. SpeechDat(II) Slovenian Database for the Fixed Telephone Network, Version 1, *University of Maribor*, Maribor, Slovenija.

Kermorvant C. & Morris A., 1999. A Comparison of Two Strategies for ASR in Additive Noise: Missing Data

and Spectral Subtraction, *6th European Conference On Speech Communication And Technology - Eurospeech'99*, str. 2841-2844, Budimpešta, Madžarska.

Leonard R. G. & Doddington G. R., 1991. A Speaker-Independent Connected-Digit Database, *Texas Instruments Inc., Central Research Laboratories*, Dallas, ZDA.

Lippmann R. & Carlson B. A., 1997. Using Missing Feature Theory to Actively Select Features for Robust Speech Recognition with Interruptions, Filtering and Noise, *5th European Conference On Speech Communication And Technology - Eurospeech'97*, str. 37-40, Rhodos, Grčija.

Young S., 1997. The HTK Book - version 2.1, *Cambridge University*, Cambridge, Velika Britanija.