Primerjava robustnosti metod določanja višine govora za različna šumna okolja in razmerja signal/šum

Vladimir Hozjan, Zdravko Kačič

Univerza v Mariboru
Fakulteta za elektrotehniko, računalništvo in informatiko
Smetanova 17, SI-2000 Maribor, Slovenija
{vladimir.hozjan, kacic}@uni-mb.si

Povzetek

Metode določanja višine govora dajejo v okoljih, ki vsebujejo šum,običajno slabše rezultate kot v čistih okoljih. V članku podajamo analizo metod določanja višine govora. Metode za določanje višine govora so največkrat sestavljene iz treh delov: detekcije zvočnosti govora, ocenitve višine govora in korekcije višine govora. Metode predstavljene v članku uporabljajo tri različne algoritme za ocenjevanje višine govora: avtokorelacijski algoritem, algoritem spremenljivih period in algoritem subharmoničnega seštevanja. Dve metodi ne uporabljata korekcijskega algoritma, tako smo lahko naredili primerjavo med algoritmi s korekcijo in tistimi brez nje. Šumno okolje smo simulirali. Izbrali smo osem različnih tipov šuma, ki se pogosto pojavljajo v realnem okolju, in jih dodali govornemu signalu. Primerjavo smo izvedli z različnimi postopki vrednotenja in ugotovili, da je algoritem spremenljive periode najmanj občutljiv na šum. Z uporabo korekcijskega algoritma se delovanje metode izboljša tako v primeru signala brez šuma kakor tudi v primeru signala z majhnim razmerjem signal/šum.

1. Uvod

Postopki s področja procesiranja govora se vedno pogosteje vključujejo v aplikacije, ki se uporabljajo v realnem okolju. To okolje vsebuje veliko zvokov, ki so govoru bolj ali manj podobni in pri procesiranju govora povzročajo veliko težav.

Metode za določanje višine govora se uporabljajo v sistemih za analiziranje prozodije govorca. S pomočjo analize prozodije govorca lahko razvijemo aplikacijo za razpoznavanje emocij ali spola v realnem času.

Pri uporabi metod za določanje osnovne harmonske frekvence govora oziroma višine govora (ang. »pitch«) v aplikacijah, ki morajo delovati v realnem okolju moramo zagotoviti, da bo uporabljena metoda za določanje višine govora čim bolj neobčutljiva na šum iz okolice.

V literaturi so obravnavane metode določanja višine, ki so prirejene za določeno šumno okolje (Picone et. al., 1985; Wu et. al., 2002). V literaturi pa nismo zasledili veliko primerjalnih raziskav, ki bi pokazale, katere metode so bolj in katere manj občutljive na šum in v kolikšni meri so posamezni algoritmi, ki so implementirani v metodah za določanje višine govora, občutljivi oziroma odporni na šum.

V tem prispevku smo se osredotočili na primerjavo metod za določanje višine govora. Primerjali smo jih glede na njihovo robustnost oziroma občutljivost na šum. Izbrali smo pet metod, ki smo jih že uporabljali v različnih aplikacijah, vendar doslej še nismo naredili primerjave med njimi.

V drugem razdelku bomo bolj podrobno predstavili posamezne metode za določanje višine govora, v tretjem bomo opisali izvedbo eksperimenta. Opisali bomo pripravo referenčnih podatkov, testnih podatkov in uporabljene metode vrednotenja. V četrtem poglavju bomo podali rezultate analize, v petem poglavju smo izvedli analizo rezultatov in v zadnjem podali zaključke.

2. Metode za določanje višine govora

Metode za določanje višine govora so pogosto sestavljene iz dveh ali treh algoritmov. Skoraj zmeraj vsebujejo algoritme za detekcijo višine govora in algoritme za ocenitev vrednosti višine govora. Za izboljšanje rezultata pa se dodajo še različni algoritmi v postprocesiranju.

Algoritmi za detekcijo višine govora poskušajo določiti zvočne in nezvočne segmente v govoru. Lastnost zvočnega segmenta je, da je periodičen. Ti algoritmi delujejo ali v časovnem ali v frekvenčnem prostoru in največkrat ne zahtevajo obsežnih računskih operacij.

Metode za določanje višine govora so ponavadi poimenovane po ocenitvenih algoritmih. Kakor algoritmi za detekcijo tudi ti algoritmi delujejo v časovni ali v frekvenčni domeni, vendar so računsko v primerjavi z algoritmi za detekcijo časovno in prostorsko bolj zahtevni.

Algoritmi v postprocesiranju poskušajo popraviti napake določanja višine govora, ki nastanejo v fazi detekcije in ocenitve. V tej fazi uporabljamo algoritme za sledenje in glajenje, saj poskušamo s temi algoritmi odstraniti nepravilne hipne spremembe in skoke v določeni višini.

V članku bomo predstavili in med seboj primerjali pet metod določanja višine govora. Te so:

- avtokorelacijska metoda s sledenjem (AKSFS)
- avtokorelacijska metoda brez sledenja (AK)
- metoda spremenljive periode (SP)
- metoda subharmoničnega seštevanja (SHS)
- metoda spremenljive periode z dinamičnim časovnim sledenjem (SPDČS)

2.1. Avtokorelacijska metoda s sledenjem

Avtokorelacijska metoda določanja višine govora je ena izmed najstarejših in verjetno največkrat uporabljena metoda do sedaj (Martino, 1999). V članku nam bo metoda služila kot referenčna metoda. Avtokorelacijska metoda s sledenjem je že implementirana v orodju »Speech Filling System« (SFS) (Speech Filing System Home Page, 2002). To orodje smo uporabili za izračun referenčnih podatkov.

2.2. Avtokorelacijska metoda brez sledenja

Kot drugo smo uporabili avtokorelacijsko metodo za določanje višine govora, ki ne vključuje algoritma sledenja v postprocesiranju. To metodo smo vključili v analizo, ker smo jo zaradi njene majhne računske zahtevnosti že uporabljali v sistemu za razpoznavanje emocij.

2.3. Metoda spremenljive periode

Avtokorelacijski algoritmi določajo višino govora v časovni domeni. Enako deluje v časovni domeni metoda spremenljive periode (Cosi et. al., 1998; Qian in Kimaresan, 1996). Metoda je dobro znana in velikokrat uporabljena ter v primerjavi z avtokorelacijsko metodo bolj natančna. Za povečanje natančnosti smo v postprocesiranju dodali še sledenje. Sledenje se izvaja le za eno predhodno vrednost višine. Zaradi takšnega načina sledenja je ta metoda primerna za uporabo v sistemih, ki delujejo v realnem času.

2.4. Metoda subharmoničnega seštevanja

Metoda subharmonične seštevanja (SHS) (Hermes, 1988) je edina izmed izbranih metod za določanje višine govora, ki deluje v frekvenčni domeni. Metodo SHS smo uporabili, ker je znana kot metoda, ki je robustna na šum. Višino govora izračuna tudi v primerih, ko osnovna frekvenca govora ni prisotna v signalu. Do takega primera pride pri procesiranju govora preko telefonske linije. Telefonski govor je frekvenčno omejen na področje med 300 in 3400 Hz. Tako v signalu največkrat ni prisotne osnovne harmonske frekvence govora. Algoritem subharmoničnega seštevanja določi višino govora iz višjih harmonikov. Uporabljena metoda za določanje višine govora SHS ne uporablja nobenega algoritma v postprocesiranju.

2.5. Metoda spremenljive periode z dinamičnim časovnim sledenjem

Kot zadnjo smo uporabili znano metodo spremenljive periode. Dodali smo ji algoritem dinamičnega časovnega sledenja v postprocesiranju (Kačič, 1995). Dinamično časovno sledenje se izvaja v razponu od začetka do konca vsakega zvočnega segmenta. Za določitev trenutne vrednosti višine govora potrebujemo vse ocenjene vrednosti višine govora v trenutnem zvočnem segmentu. Zato ta metoda ni najbolj primerna za delovanje v realnem času, saj v danem trenutku ne poznamo vrednosti višin govora, ki se bodo pojavile. Ker je ta metoda najbolj zapletena in kompleksna, smo pričakovali, da bo dala najboljše rezultate.

3. Izvedba eksperimenta

Primerjava robustnosti metod za določanje višine govora zahteva primerjavo natančnosti delovanja metod v različnih šumnih okoljih. Rezultate delovanja v šumnih okoljih smo primerjali z rezultati, dobljenimi iz referenčnih podatkov. Da smo lahko določili najbolj robustno metodo, smo rezultate ocenili z metodami vrednotenja, ki so bolj podrobno opisane v nadaljevanju.

3.1. Določitev referenčnih podatkov

V eksperimentu smo uporabili bazo izgovarjav Interface (Hozjan et. al., 2002). To je baza emocionalnega govora in vsebuje govor 9 govorcev in signal laringografa 5 govorcev. V bazi Interface sta govorni signal in signal laringografa shranjena v isti datoteki na dveh kanalih. Na prvem kanalu je govorni signal, na drugem signal laringografa. Iz signala laringografa smo določili referenčne podatke, ki so služili za primerjavo. Višino govora smo izračunali s pomočjo orodja SFS, ki vsebuje modul za določitev višine govora iz signala laringografa.

3.2. Določitev testnih podatkov

Ker je baza Interface posneta v studijskem okolju, vsebuje le zanemarljivo malo šuma. Zato smo morali šumno okolje simulirati. V ta namen smo uporabili šume, posnete v projektu AURORA 2 (Hirsh et. al., 2000). Poimenovanje šumov smo povzeli po zvočnih okoljih, ki so jih uporabljali v projektu AURORA 2. Uporabili smo naslednjih 8 šumov:

- letališče
- kramljanje
- avto
- razstava
- restavracija
- podzemna železnica
- cesta
- vlak

Teh osem šumov smo dodali govornemu signalu v različnih razmerjih signal/šum. Določili smo naslednjih šest razmerij signal/šum: 20 dB, 15 dB, 10 dB, 5 dB, 0 dB in -5 dB.

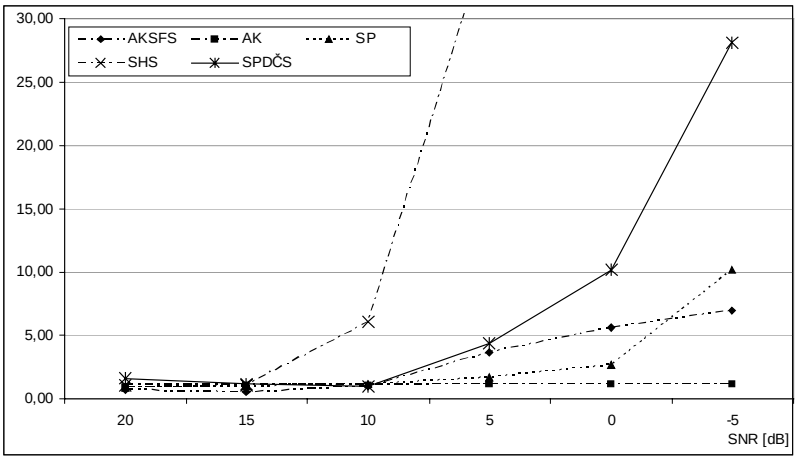
Testni nabor je vseboval 100 posnetkov govornega dela baze. Izbrani nabor vsebuje govor dveh govorcev: 50 stavkov govora moškega govorca in 50 stavkov ženske govorke. Tem 100 stavkom smo dodali vseh 8 šumov v vseh 6 razmerjih signal/šum. Tako smo dobili skupaj s posnetki brez šuma 4900 testnih posnetkov. Posnetke brez

	AKSFS	AK	SP	SHS	SPDČS
Napaka zvočnosti [%]	11,82	21,71	15,51	31,44	15,81
Napaka nezvočnosti [%]	16,23	45,02	14,64	4,99	0,63
Gross high [%]	3,66	0,209	4,45	1,88	1,21
Gross low [%]	0,83	74,73	1,49	9,89	3,32
Absolutna razlika v srednji vrednosti [Hz]	8,73	92,27	9,81	1,02	0,76
Absolutna razlika v standardni deviaciji [Hz]	4,20	24,077	0,75	30,86	7,58

Tabela 1: Ocenitev natančnosti vseh metod za določanje višine govora z izbranim naborom metod vrednotenja.

SNR[dB]	AKSFS	AK	SP	SHS	SPDČS
20	6,02	85,34	8,90	1,13	1,24
15	4,58	98,45	9,92	1,08	0,89
10	8,95	100,78	11,40	6,15	0,77
5	31,92	104,62	16,29	37,47	3,29
0	49,27	105,31	26,25	94,24	7,73
-5	60,98	106,66	100,27	100,44	21,37

Tabela 2: Absolutna razlika srednje vrednosti višine, povprečena na vse vrste šuma, izražena v Hz v odvisnosti od razmerja signal/šum (SNR).



Slika 1: Relativno poslabšanje absolutne razlike srednje vrednosti višine v odvisnosti od razmerja signal/šum (SNR) za vseh pet metod določanja višine.

šuma smo poimenovali čisti posnetki.

3.3. Metode vrednotenja

Za ocenitev rezultatov smo uporabili že uveljavljene metode vrednotenja pravilnosti izračuna višine govora. Uporabili smo pet metod. Napaki »Gross high« in »Gross low« je predlagal Goangshiuan (Goangshiuan, 1998), napako zvočnosti, napako nezvočnosti, absolutno razliko v srednji vrednosti in absolutno razliko v standardni deviaciji pa Martino (Martino, 1999).

Napako Gross uporabljamo za grobo primerjavo med različnimi metodami določanja višine govora. Gross predstavlja število otipkov zvočnega segmenta, ki se po vrednosti od reference razlikujejo za več kot 20 %. Napako Gross običajno izrazimo v odstotkih. Gross high predstavlja število otipkov zvočnega segmenta, ki so od reference večje za več kot 20 %, Gross low pa število otipkov zvočnega segmenta, ki so od reference manjše za več kot 20 %.

Napaki zvočnosti in nezvočnosti ocenjujeta delovanje algoritma za detekcijo. Napaka zvočnosti predstavlja število otipkov zvočnega segmenta, ki so zvočni in so bili detektirani kot nezvočni. Napaka nezvočnosti pa predstavlja število otipkov nezvočnega segmenta, ki so nezvočni in so bili detektirani kot zvočni. Obe napaki običajno izrazimo v odstotkih.

Napaki absolutnih razlik določata napako, ki jo naredi algoritem za oceno višine govora. Absolutna razlika srednje vrednosti je definirana kot absolutna razlika med srednjo vrednostjo višine reference in srednjo vrednostjo testnih posnetkov. Absolutna razlika standardne deviacije je definirana kot absolutna razlika med standardno deviacijo višine v referenci in standardno deviacijo višine v testnih posnetkih. Ti napaki sta izraženi v Hz.

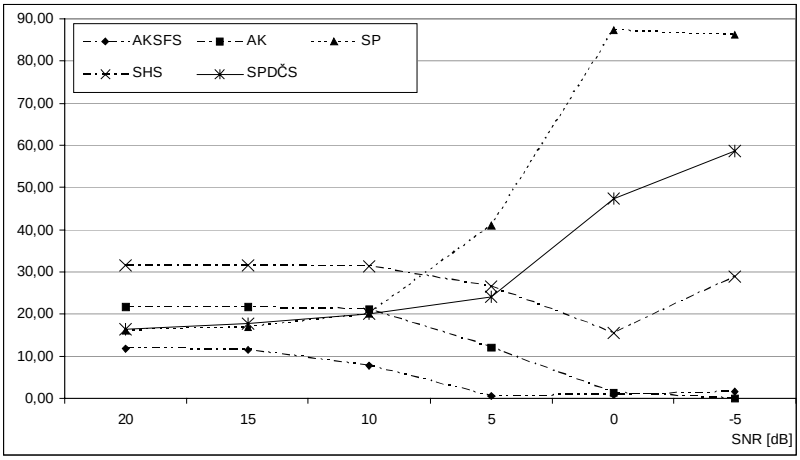
4. Rezultati

Za primerjavo postopkov smo najprej določili referenčne podatke, nato pa z vsemi metodami za določitev višine govora izračunali višino čistih posnetkov. V tabeli 1 so prikazane vrednosti vseh šestih metod vrednotenja višine govora za čiste posnetke.

Nadalje smo z vsemi metodami za določanje višine govora izračunali višine za vse vrste šuma v vseh razmerjih signal/šum.

Rezultate smo ocenili z vsemi metodami vrednotenja in tako dobili množico rezultatov za analizo. Iz te množice smo izbrali nekaj primerov, ki najbolj nazorno prikazujejo obnašanje metod za določanje višine govora v različnih okoljih, ki jih podrobneje predstavljamo v članku. Ker smo želeli analizirati rezultate različnih metod glede na velikost šuma in ne na tip šuma, smo vse rezultate povprečili na vse vrste šuma. Odvisnost rezultatov različnih metod od razmerja med signalom in šumom prikazuje tabela 2.

Da bi ugotovili, kako se napaka veča, smo iz absolutne razlike vrednosti višine govora in iz srednje vrednosti za čiste posnetke izračunali relativno poslabšanje absolutne razlike srednje vrednosti višine govora. Relativno



Slika 2: Napaka zvočnosti za vsako metodo določanja višine govora, povprečena na vse vrste šuma, v odvisnosti od razmerja signal/šum, izražena v odstotkih.

poslabšanje v odvisnosti od razmerja signal/šum prikazuje slika 1.

Rezultati, dobljeni iz napake zvočnosti v povezavi z napako nezvočnosti, kažejo odvisnost posameznih metod od razmerja signal/šum. Ti napaki se navezujeta na točnost delovanja detekcijskega algoritma. Slika 2 prikazuje napako zvočnosti in slika 3 napako nezvočnosti.

Nobena od metod vrednotenja ne ocenjuje delovanja algoritmov v postprocesiranju, kar pa bi bilo tudi zanimivo analizirati.

5. Razprava

Metode za določanje višine govora se med seboj razlikujejo po uspešnosti že pri analizi iz čistih posnetkov. Pri čistih posnetkih sta se pokazali kot najboljši metodi AKSFS in SPDŠC.

Iz rezultatov, dobljenih iz čistih posnetkov, lahko sklepamo na delovanje oziroma pomanjkljivosti, ki jih imajo posamezne metode za določanje višine govora. Detekcijski algoritem pogosteje določi nezvočni govor kot zvočnega. To sklepamo iz nezvočne napake v tabeli 1. Izračunana vrednost višine je za AKSFS večkrat določena previsoko kot prenizko. To nakazujeta Gross high in Gross low.

Metoda AK je najslabša metoda za določanje višine govora v našem eksperimentu, saj vrednost absolutne razlike srednje vrednosti pri čistem signalu znaša nad 90 Hz. Tako ostale napake niso zanimive za nadaljnjo obravnavo.

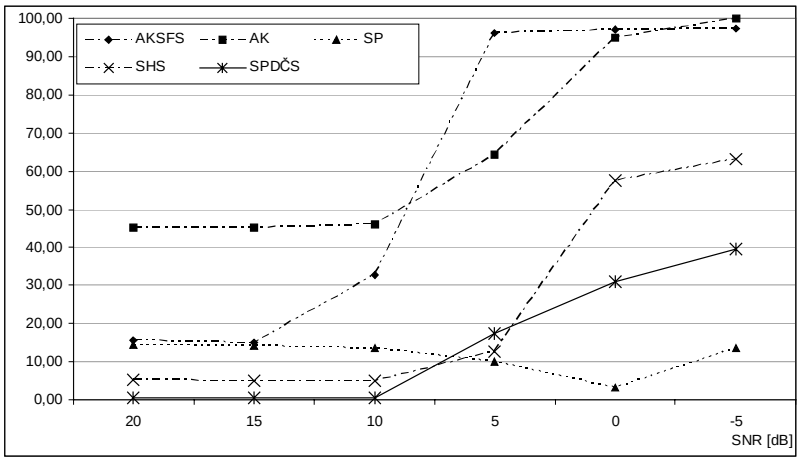
Detekcijski algoritem pri metodi SP se približno za enak odstotek zmoti pri določitvi zvočnih in pri določitvi nezvočnih delov govora. Algoritem spremenljive periode večkrat previsoko oceni višino govora. Absolutna razlika srednje vrednosti višine je le za odstotek večja kot pri metodi AKSFS, absolutna razlika v standardni deviaciji pa je manjša kot pri drugih metodah.

SHS ima najslabši detekcijski algoritem. Kar 30,86 % zvočnih segmentov je detekcijski algoritem v metodi SHS označil kot nezvočne segmente. Algoritem za ocenjevanje ocenjuje višino govora pogosteje prenizko kot previsoko. Metoda SHS pa ima najmanjšo absolutno razliko standardne deviacije, kar pomeni, da je dinamika določene višine govora podobna dinamiki, ki jo ima referenca.

Vse napake metode SPDČS so približno v istem razredu kot pri metodi AKSFS, razen absolutne razlike standardne deviacije, ki je dvakrat večja pri metodi SPDČS. Lahko trdimo, da je določena višina govora z metodo AKDČS manj dinamična od referenčne višine govora.

Pri tej raziskavi nas je zanimala predvsem robustnost metod določanja višine govora v šumnem okolju. Tabela 2 in slika 1 prikazujeta natančnost določanja višine govora v odvisnosti od razmerja singal/šuma. Metoda SHS je najbolj občutljiva na šum. Krivulja na sliki 1 je za metodo SHS najbolj strma. Pri razmerju signal/šum 5 dB znaša absolutna razlika srednje vrednosti za metodo SHS že kar 37,47 Hz. Nekoliko boljši rezultat je za metodo AKSFS. Pri razmerju signal/šum 5 dB znaša absolutna razlika srednje vrednosti višine 31,92 Hz, krivulja relativnega poslabšanja pa je veliko nižja kot pri metodi SHS. Vrednost absolutne razlike srednje vrednosti višine govora pri čistem signalu je za metodo AKSFS višja kot za metodo SHS. Tako so v primerjavi z metodo AKSFS vrednosti relativnega poslabšanja za iste vrednosti absolutne razlike srednje vrednosti višje za metodo SHS. Najboljše rezultate smo dosegli z metodama, ki delujeta na osnovi algoritma spremenljive periode. Krivulji relativnega poslabšanja nista najmanj strmi. Metoda SP ima nižjo krivuljo kot metoda SPDČS, saj je absolutna razlika srednje vrednosti za čiste posnetke z metodo SPDČS osemkrat nižja od absolutne razlike srednje vrednosti, dobljene z metodo SP. Glede na kriterij absolutne razlike srednje vrednosti višine govora bi lahko zadovoljivo uporabili metodo SP v šumnem okolju, v katerem znaša razmerje signal/šum 5 dB, metodo SPDČS pa celo v šumnem okolju z razmerjem signal/šum 0 dB.

Sliki 2 in 3 prikazujeta uspešnost delovanja algoritmov za detekcijo višine govora pri različnih razmerjih signal/šum. Metoda AK že pri 5 dB razmerja signal/šum detektira več kot 90 % otipkov nezvočnih segmentov kot zvočne, metoda AKSFS pa doseže enak rezultat pri 0 dB razmerja signal/šum (slika 3). Metodi AK in AKSFS detektirata šum kot zvočni govor. Detekcijski algoritem v metodi SHS se obnaša podobno kot detekcijska algoritma v metodah AK in AKSFS. Metodi SHS se poveča napaka nezvočnosti, vendar v manjši meri kot metodama AK in AKSFS (slika 3).



Slika 3: Napaka nezvočnosti za vsako metodo določanja višine govora, povprečena na vse vrste šuma, v odvisnosti od razmerja signal/šum, izražena v odstotkih.

Nasprotno metoda SP detektira zvočni govor v šumnem okolju kot šum, in ne kot zvočni segment govora, saj znaša napaka zvočnosti čez 90 % pri razmerju signal/šum 0 dB (slika 2). SPDČS je edina metoda, pri kateri se v odvisnosti od razmerja signal/šum povečujeta napaki zvočnosti in nezvočnosti. Šum vpliva na algoritem detekcije v metodi SPDČS tako, da nepravilno določi približno 60 % otipkov v zvočnem segmentu in približno 40 % odtipkov v nezvočnem segmentu govora.

6. Zaključek

Šum vpliva na metode za določanje višine govora na različne načine. Metode AK, AKSFS in SHS detektirajo pri dovolj velikem razmerju signal/šum šum kot zvočni del govora, metoda SP pa detektira v šumnem okolju zvočni del govora kot šum.

Šum vpliva tudi na zanesljivost delovanja posameznih metod. Najbolj vpliva na metodo SHS, kjer je krivulja poslabšanja najbolj strma.

Kot najmanj občutljiva se je izkazala metoda SPDČS, ki ima najmanjšo absolutno razliko srednje vrednosti višine govora pri vseh izbranih razmerjih signal/šum. Po naši oceni bi jo lahko uporabili tudi pri 0 dB razmerja signal/šum, kjer znaša absolutna razlika srednje vrednosti višine govora 7,73 Hz. Tudi vrednosti ostalih metod vrednotenja za metodo SPDČS so nizke, v večini primerov so najnižje. Če primerjamo rezultate te metode z metodo SP, lahko sklepamo, da k takemu rezultatu veliko prispeva uporabljeni algoritem v postprocesiranju, saj ima metoda SP skoraj vse rezultate slabše od metode SPDČS.

S stališča uporabe metod v realnem času najboljši kompromis, glede na izvedljivost, računsko in pomnilniško zahtevnost na eni strani ter uspešnost na drugi strani, predstavlja metoda SP.

SP smo uporabili v razpoznavalniku emocij in spola, kjer potrebujemo delovanje v realnem času in v realnem okolju. SP metoda ima dovolj malo časovno in računsko zahtevnost, saj je potrebno poleg računanja višine zagotoviti nekaj računskega časa in prostora za ostale metode in algoritme v končni aplikaciji.

7. Literatura

Cosi P., Pasquin S., Zovato E. 1998. Auditory Modeling Techniques for Robust Pitch Extraction and Noise Reduction. *ICSLP-98 Proceedings*.

Goangshiuan. S.Y., Leah. H.J., Carl. D.M. 1998. A probabilistic Approach to AMDF Pitch Detection. ICSLP *Proceedings*.

Kačič, Z. 1995. *Komunikacija človek – stroj*. Fakulteta za elktrotehniko, računalništvo in informatiko v Mariboru. Maribor.

Hermes. D.J. 1988, Measurement of Pitch by Subharmonic Summation. J.Acoust.Soc.Am. 83. 257-264.

Hirsh. H.G., in Pearce. D., 2000. The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Condition. *ISCA ITRW ASR2000*.

Hozjan. V. Kačič. Z., Moreno A., Bonafonte A., Nogueiras A. 2002. INTERFACE Databases: Design and Collection of a Multilingual Emotional Speech Database. *LREC-02 Proceeding*.

Maidment. J.A., in Lecumberri M.L.G. 1996. Pitch Analysis Methods for Cross-Speaker Comparision. *ICSLP1996 Proceeding*.

Martino. J., in Laprie. Y. 1999. An Efficient F0 Determination Algorithm Based on the Implicit Calculation of the Autocorrelation of the Temporal Excitation Signal. *EUROSPEECH'99 Proceeding.*

Qian X. in Kimaresan R., 1996. "A variable Frame Pitch Estimator and Test Results*", IEEE Int. Conf. on Acous., Speech, and S. Proc.*

Picone J., Doddington G.R., Secrest B. G.. 1985. *Robust Pitch Detection in a Noisy Telephone Environment.* IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-33,

Speech Filing System Home Page. 2002. Speech Filing System. *http://www.phon.ucl.ac.uk/resource/sfs/.*

Wu M., Wang D., Guy J. Brown. 2002. *A Multi-Pitch Tracking Algorithm for Noisy Speech*. ICASSP 2002.