Istraživanje tehnika normalizacija dužine vokalnog trakta baziranih na ML kriterijumu

Nikša Jakovljević, Dragiša Mišković i Marko Janev

***Sadržaj* — U ovom radu su prikazani preliminarni rezultati primene normalizacije dužine vokalnog trakta (VTN) na osnovu ML kriterijuma. Pored standardnog metoda za određivanje koeficijenta normalizacije predloženo je nekoliko alternativnih metoda. Da bi se procenile mogućnosti primene VTN izvršeni su tzv. nadgledani testovi. Najbolji sistem za prepoznavanje govora u kom je primenjen VTN je rezultovao relativnim poboljšanjem performansi od oko 20 % u odnosu na sistem nezavisan od govornika odnosno 17 % u odnosu na sistem zavisan od pola govornika.**

***Ključne reči* — Automatsko prepoznavanje govora, normalizacija dužine vokalnog trakta, robustna estimacija**

I. UVOD

SAVREMENI sistemi za automatsko prepoznavanje govora (ASR) su osetljivi na razlike koje postoje između akustičkih uslova u kojima su snimani iskazi namenjeni obuci sistema i iskazi nad kojima se treba vršiti prepoznavanje. Ove razlike se manifestuju kroz degradaciju performansi samog ASR sistema, pa tako sistemi zavisni od govornika postižu grešku na nivou reči od oko 1% a sistemi nezavisni od govornika od 5 i više procenata [1]-[3]. Uzroci ovih varijabilnosti su različiti mikrofoni, kanali, govornici, ambijentalni zvukovi i šumovi u kanalu.

Uzroci varijacija među govornicima se mogu podeliti na dve grupe: spoljašnje i fiziološke. Spoljašnji uzroci su različiti kulturološki faktori kao i emocionalno stanje govornika, a ispoljavaju se kao varijacije u načinu izgovora, akcetnovanju reči, prozodiji i brzini govora. Sa druge strane fiziološki uzroci su posledica anatomskih razlika u obliku i veličini vokalnog trakta, a manifestuju se kao varijacije u boji glasa odnosno položajima maksimuma u spektralnoj obvojnici govornog signala.

Postupci koji imaju za cilj smanjenje akustičke varijabilnosti koja postoji između iskaza na kojima se vrši obuka i iskaza na kojima se vrši prepoznavanje se u zavisnosti od prostora u kojem se realizuju dele na postupke normalizacije i adaptacije. Ako prilagođavanje podrazumeva transformaciju vektora obeležja tada je u pitanju normalizacija, a ako podrazumeva transformaciju parametara akustičkih modela onda je u pitanju adaptacija. Brojna istraživanja su pokazala da efekti normalizacije i adaptacije nisu isključivi već komplementarni [2],[4].

Tehnike normalizacije se klasifikuju u dve grupe: tehnike koje su zasnovane na nekom fizičkom modelu i tehnike zasnovane na raspodeli podataka. Mnoge pojave koje prouzrokuju akustičku varijabilnost mogu se predvideti a samim tim i korigovati na osnovu poznavanja odgovarajućih fizičkih modela, što se koristi u metodama koje su zasnovane na nekom fizičkom modelu. U ovu grupu metoda spada i normalizacija dužine vokalnog trakta (VTN), koja koristi činjenicu da je dužina vokalnog trakta obrnuto srazmerna frekvencijama formanata. Postoje i pojave koje nisu predvidljive ili su suviše složene tako da ne postoji bilo kakav fizički model kojim bi bile opisane. Da bi se izborili sa varijabilnostima koje su posledica takvih pojava transformacioni parametri se dobijaju na osnovu poređenja raspodela obeležja u skupu za obuku i testiranje. Tehnike normalizacije koje spadaju u ovu grupu su transformacija prostora obeležja i uklapanje prostora obeležja [3].

U ovom radu će biti razmotrene VTN tehnike bazirane na ML kriterijumu. Pored tehnika koje su bazirane na ML kriterijumu postoje i tehnike bazirane na estimaciji položaja formanata koje neće biti predmet razmatranja ovog rada.

U narednom odeljku je dat kratak opis govorne baze nakon koje sledi odeljak sa opisom VTN metoda i motivacijom za uvođenjem istih. U IV poglavlju su navedeni rezultati eksperimenata nakon čega sledi zaključak.

II. GOVORNA BAZA

Svi testovi i obuke koji su realizovani u ovom radu su koristili redukovanu SpeechDat(E) govornu bazu za srpski jezik. Ova redukovana govorna baza se razlikuje od standardne baze po tome što su svi govornici za koje ne postoji bar 30 s govornog materijala izostavljeni. Govorni materijal ne obuhvata tišinu i oštećene govorne segmente. Iako je pri snimanju svaki od govornika imao zadatak da izgovori približno istu količinu teksta pojedini govornici su neadekvatno artikulisali pojedine glasove, koji su stoga izbačeni iz skupa za obuku, što za posledicu ima da količina govornog materijala varira od govornika do govornika. Za veliki broj govornika ima bar 60 s govornog materijala na raspolaganju.

Ovaj rad je podržan od strane Ministarstva za nauku i tehnološki razvoj Republike Srbije, u okviru projekta "Govorna komunikacija čovek-mašina" (TR-11001)

N. M. Jakovljević, Fakultet tehničkih nauka u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (telefon: 381-21-485-2521; fax: +381-21-475-2997 e-mail: jakovnik@uns.ns.ac.yu).

D. Mišković, Fakultet tehničkih nauka u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (telefon: 381-21-485-2521; fax: +381-21-475-2997 e-mail: dragisa.miskovic@alfanum.co.rs).

M. Janev, Fakultet tehničkih nauka u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (telefon: 381-21-475-0080; fax: +381-21-475-2997 e-mail: marko.janev@alfanum.co.rs).

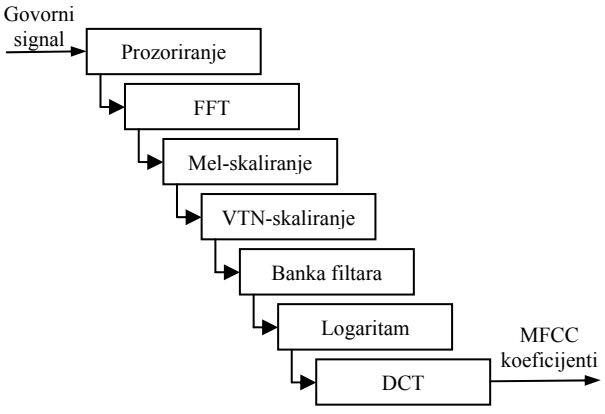
Ovu redukovanu SpeechDat (E) bazu čini oko 5.5 sati govornog materijala izgovorenog od strane 376 muškaraca i 6 sati govornog materijala izgovorenog od strane 323 žene. Ovo smanjenje količine materijala za obuku je dovelo do neznatne degradacije performansi ASR sistema, ali je bilo neophodno da bi se eliminisala zavisnost performansi kreiranih sistema za prepoznavanje govora od količine materijala za obuku.

Test skup čini oko 14 minuta govornog materijala izgovorenog od strane 107 muškaraca i oko 14 minuta govornog materijala izgovorenog od strane 77 žena. Pošto je cilj eksperimenata koji su opisani u ovom radu pružanje uvida koliki dobitak može da se očekuje primenom VTN trebalo je i u ovom slučaju isključiti govornike za koje ne postoji dovoljna količina govornog materijala, ali to ipak nije urađeno da bi se ovi rezultati mogli uporediti sa rezultatima realnih testova gde je situacija sa nedovoljnom količinom govornog materijala moguća.

III. OPIS PROCEDURE ZA NORMALIZACIJU DUŽINE VOKALNOG TRAKTA

A. Skaliranje frekvencijske ose

Dužina vektora obeležja je 26, a čine ga 12 MFCC koeficijenata, normalizovana energija i njihovi prvi izvodi po vremenu. Procedura izdvajanja MFCC obeležja je delimično modifikovana ubacivanjem bloka koji vrši skaliranje frekvencijske ose DFT spektra. Kao što je ilustrovano na slici 1. U samoj implementaciji banka filtara i blokovi koji vrše MEL i VTN skaliranje su objedinjeni u jedan.



Sl. 1. Blok šema za izdvajanje MFCC obeležja

U ovom radu skaliranje frekvencijske ose je realizovano pomoću deo po deo linearne funkcije. U slučaju deo po deo linearne funkcije zavisnost „nove“ učestalosti ω_α od stvarne učestalosti ω je data izrazom:

$$\omega_\alpha = \begin{cases} \alpha\omega & \omega \le \omega_0 \\ \alpha\omega - \dfrac{\pi - \alpha\omega_0}{\pi - \omega_0} & \omega > \omega_0 \end{cases} \quad (1)$$

gde je ω_0 granična učestalost za koju se obično uzima vrednost od $7\pi/8$, a α nagib krive odnosno VTN koeficijent.

Razlozi zašto je u ovom radu izabrana deo po deo linearna funkcija su njena jednostavnost implementacije i performanse koje su približno iste kao u slučaju da se skaliranje frekvencijske ose realizuje bilinearnom ili kvadratnom funkcijom [1]-[6].

B. Obuka

Obuka ASR sistema, zasnovanih na statističkom modelu, podrazumeva estimaciju parametara skrivenih Markovljevih modela i Gausovih mešavina, koji maksimizuju verodostojnost na skupu za obuku. Ako sa λ označimo skup parametara koje treba estimirati, sa X skup vektora obeležja koji su na raspolaganju za obuku i W odgovarajuće transkripcije, cilj obuke se može prikazati kao:

$$\lambda = \arg\max_{\lambda'} p(X \mid W; \lambda') \quad (2)$$

U slučaju sistema koji primenjuju VTN cilj obuke se modifikuje tako što se lista parametara koju treba estimirati (λ) proširuje VTN koeficijentima (α) za svakog od govornika u bazi. Zajednička estimacija parametara statističkog modela i VTN koeficijenata u velikoj meri bi usporila proceduru estimacije parametara. Alternativno rešenje predstavlja iterativna metoda u kojoj se u 1. koraku određuje vrednost VTN koeficijenta (α) za svakog od govornika za neki inicijalni skup parametara (λ_i) na osnovu jednačine:

$$\alpha = \arg\max_{\alpha'} p(X_g^{\alpha'} \mid W_g; \lambda_i) \quad (3)$$

gde su transkripcije koje pripadaju govorniku g obeležene sa W_g, vektori obeležja koji pripadaju govorniku g na kojima je primenjen VTN sa koeficijentom α obeleženi sa X_g^α (iz analize su eliminisani vektori obeležja koji pripadaju oštećenim segmentima i segmentima tišine), a u 2. koraku se estimiraju parametri statističkog modela λ kao u jednačini (2), s tom razlikom da se umesto originalnih vektora obeležja koriste vektori obeležja nad kojima je primenjen VTN ali sa optimalnim koeficijetom α za njemu pripadajućeg govornika. Ove korake je moguće ponoviti više puta.

Da bi se ubrzala estimacija VTN koeficijenata, skup mogućih vrednosti VTN koeficijenata je ograničen i diskretizovan. Obično je to skup vrednosti od 0,88 do 1,12, sa korakom 0,02, što je slučaj i u ovom radu. Izabran je ovaj interval pošto su merenja pokazala da su razlike u položaju formanata kod muškaraca i žena oko 20 % [7].

U ovom radu je implementirano rešenje koje je predloženo u radu [8], koje za inicijalni skup modela (skup koji se koristi za određivanje VTN koeficijenata) koristi HMM modele sa po jednom Gausovom raspodelom po stanju, a prethodna dva koraka obuke se izvršavaju samo jednom. Motivacija da se za inicijalne model koriste HMM modeli sa po jednom Gausovom raspodelom po stanju leži u činjenici da modeli sa više Gausovih raspodela po stanju uče i karakteristike govornika koje želimo da eliminišemo primenom VTN.

Mana predloženog načina izbora VTN koeficijenta datog izrazom (3) je da favorizuje duže i zastupljenije foneme. U ovom radu su razmotreni i neki alternativni načini izbora VTN koeficijenta i to:

1. Za VTN koeficijent se bira koeficijent za koji je prosečna vrednost verodostojnosti po instanci fonema maksimalna. (Pod instancom fonema se podrazumeva konkretna realizacija tog fonema u bazi) Verodostojnost instance fonema se računa kao uzoračka sredina verodostojnosti pridruženih mu

vektora obeležja. Na dalje u tekstu ova metoda nosi oznaku M1.

2. Slična kao metoda M1 s tom razlikom da se verodostojnost instance fonema računa kao uzoračka mediana verodostojnosti pridruženi mu vektora obeležja. Na dalje u tekstu ova metoda nosi oznaku M2.
3. Za vrednost VTN koeficijenta se bira koeficijent za koji je prosečna vrednost verodostojnosti po fonemu maksimalna. Verodostojnost fonema se računa kao uzoračka sredina svih vektora obeležja tog fonema. Na dalje u tekstu oznaka ove metode je M3.
4. Slična kao metoda M3 s tom razlikom da se verodostojnost fonema računa kao uzoračka mediana svih vektora obeležja tog fonema. Na dalje u tekstu oznaka ove metode je M4.
5. Za vrednost VTN koeficijenta se bira koeficijent za koji je uzoračka mediana verodostojnosti fonema maksimalna. Verodostojnost fonema se računa kao i u slučaju metode M3. Na dalje u tekstu oznaka ove metode je M5.
6. Slična kao metoda M5 s tom razlikom da se verodostojnost fonema računa kao u slučaju metode M4. Na dalje u tekstu oznaka ove metode je M6.

Metoda koja je predložena u radu [7] će nadalje u tekstu nositi oznaku M0. Kao i u slučaju M0 način određivanja VTN koeficijenta i predloženi alternativni metodi M1-M6, ne uzimaju u obzir negovorne i oštećene segmente, a za inicijalni skup modela koriste isti skup sa po jednom Gausovom raspodelom po stanju.

Metoda sa oznakom M1 ima za cilj da eliminiše uticaj dužine trajanja instance, ali ne i frekvencije pojavljivanja fonema na izbor VTN koeficijenta. Motivacija favorizovanja frekventnijih fonema pri izboru VTN koeficijenta je posledica želje da na što većem broj instanci izabrana vrednost VTN koeficijenta daje maksimalnu moguću verodostojnost. Ova metoda ne obezbeđuje povećanje verdostojnosti reči odnosno sekvence reči, pošto pri prepoznavanju (dekodovanju) duži segmenti imaju veći uticaj na ukupnu vrednost verodostojnosti.

Motivacija za metodu M2 je slična kao i za metodu M1, s tom razlikom da se izborom uzoračke mediane umesto uzoračke sredine za estimaciju verodostojnosti instance, želeli obezbediti da izabrani VTN koeficijent bude dobar za većinu vektora obeležja koji pripadaju toj instanci fonema, a ne za nekolicinu koji možda svojm velikim vrednostima verodostojnosti mogu da povećaju stvarnu prosečnu vrednost verodostojnosti. Ni ova metoda pri prepoznavanju ne obezbeđuje povećanje verodostojnosti reči odnosno sekvence reči.

Metoda M3 je uvedena s ciljem da se pri određivanju VTN koeficijenta eliminiše kako dužina trajanja tako i zastupljenost fonema, odnosno da svi fonemi ravnopravno učestvuju. Kao i u prethodne dve metode, ni ova metoda pri prepoznavanju ne obezbeđuje povećanje verodostojnosti reči odnosno sekvence reči.

Metoda M4 predstavlja modifikovanu metodu M3 kod koje se za estimaciju prosečne vrednosti verodostojnosti fonema koristi robustna estimacija (uzoračka mediana).

Metode M5 i M6 predstavljaju robustne varijante metoda M3 i M4. Osnovni cilj uvođenja uzoračke mediane umesto uzoračke sredine je smanjenje uticaja ekstremno velikih i ekstremno malih vrednost verodostojnosti pojedinih fonema (ako takvi postoje).

Nakon što se odrede VTN koeficijenti, prisutpa se obuci sistema koji ima više Gausovih raspodela po stanju koji će se potom koristiti pri prepoznavanju. Pri izdvajanju obeležja se koriste VTN koeficijenti koji su određeni u prethodno opisanom procesu.

C. Test

Proceduri prepoznavanja kod sistema koji koriste VTN prethodi procedura određivanja VTN koeficijenata. Cilj ovog rada je procena dobitka koji je rezultat primene VTN. Da bi se eliminisao uticaj greške usled pogrešnog određivanja vrednosti VTN koeficijenta primenjuje se tzv. nadgledano (*supervised*) testiranje. Nadgledatno testiranje ima za cilj da obezbedi da se vrednosti VTN koeficijenta koje se odrede pri testiranju ne razlikuju drastično od onih koje bi se dobile da se taj fajl nalazi u skupu za obuku, ali ne garantuje ispravnost estimiranih VTN koeficijenata. Ispravnost estimacije VTN koeficijenata prvenstveno zavisi od načina njihovog izbora. Primenjena procedura se može sistematizovati u sledeća dva koraka:

1. Na osnovu postojećih ispravnih transkripcija primenom odgovarajuće metode (M0-M6) odrediti vrednosti VTN koeficijenata za svakog od govornika. Metoda određivanja VTN koeficijenata koja se primenjuje u testu treba da bude identična metodi koja se koristi pri obuci.
2. Na osnovu vrednosti VTN koeficijenta određenih u koraku 1, transformisati obeležja i sa tako transformisanim obeležjima pristupiti prepoznavanju.

Princip određivanja VTN koeficijenata u nadgledanom testiranju je sličan kao onaj koji se primenjuje pri obuci. Razlika je da se umesto sistema sa po jednom smešom po stanju (inicijalnog skupa modela) koriste sistemi koji su dobijeni u VTN obuci svake od metoda. Na primer ako je u toku obuke metodom M0 dobijen sistem S_{M0} (ima više Gausovih raspodela po stanju) tada se taj isti sistem koristi i u testovima za metodu M0 kako pri određivanju VTN koeficijenata tako i pri samom prepoznavanju. Razlog ovakvog pristupa su rezultati prethodnih testova koji su publikovani u [9], u kojima je određvanje VTN koeficijenata na osnovu inicijalnog skupa modela dalo lošije rezultate od sistema koji je dobijen VTN obukom.

IV. Rezultati

Uporedni prikaz performansi analiziranih sistema je dat u tabeli 2. Prva kolona u tabeli predstavlja oznaku VTN metode koja je korišćena pri njegovom formiranju. Oznake odgovaraju oznakama koje su uvedene u prethodnom odeljku. Druga treća i četvrta kolona predstavljaju standardne veličine preko kojih se izražavaju performanse ASR sistema i to broj zamena, broj umetanja i broj brisanja respektivno. Peta kolona pedstavlja grešku na nivou reči iskazanu u procentima. Poslednje dve kolone predstavljaju relativno poboljšanje performansi sistema u odnosu na referentne sisteme čije su performanse navedene u tabeli 1 i to kolona *RI1* za prvi, a *RI2* za drugi referentni sistem. Relativno poboljšanje performansi (*RI*) je definisano jednačinom:

$$RI = \frac{WER_{base} - WER_{new}}{WER_{base}} \times 100\,\% \quad (4)$$

gde su sa WER_{base} i WER_{new} označene greške na nivou reči referentnog i novo-predloženog sistema respektivno.

Detaljan opis procedure kojom su formirani referentni sistemi je dat u [9]. Prvi referentni sistem predstavlja sistem koji je nezavisan od govornika. Složenost ovog sistema je 34,8k Gausovih mešavina, što je ujedno i složenost svih sistema koji su dobijeni nekom od predloženih VTN metoda. Drugi referentni sistem je sistem koji je zavisan od pola govornika, odnosno za svaki pol je formiran poseban skup akustičkih modela. Složenost ovog sistema je nešto manja i iznosi 32,3k Gausovih mešavina. Idealno bi bilo da se porede sistemi iste, a ne slične složenosti, ali trenutni proces obuke ne omogućava specificiranje tačnog broja mešavina.

TABELA 1: PERFORMANSE REFERENTNIH SISTEMA

Oznaka	*br. zamena*	*br. umetanja*	*br. brisanja*	*WER*
Ref1	94	56	9	5,49
Ref2	94	51	8	5,28

TABELA 2: PERFORMANSE ANALIZIRANIH SISTEMA

Oznaka	*broj zamena*	*broj umetanja*	*broj brisanja*	*WER*	*RI1*	*RI2*
M0	74	46	10	4,48	18,4	15,2
M1	72	46	9	4,38	20,2	17,0
M2	69	54	10	4,59	16,4	13,1
M3	72	54	7	4,59	16,4	13,1
M4	75	53	7	4,66	15,1	11,7
M5	69	53	9	4,52	17,7	14,4
M6	75	63	8	5,04	8,2	4,5

Kao što se iz priloženog može videti sve predložene VTN metode su dale određeno unapređenje performansi u odnosu na referentne sisteme. Ako posmatramo WER (grešku na nivou reči) vidimo da je najbolje rezultate dao sistem M1 nakon čega sledi M0, a kao najlošiji se pokazao sistem M6. Na osnovu ovoga bi mogli zaključiti da ima smisla eliminisati uticaj dužine fonema pri izboru VTN koeficijenta, ali treba dopustiti i da frekventniji fonemi imaju veći uticaj. Korišćenje uzoračke mediane umesto uzoračke sredine se pokazalo kao loš izbor jer ako se uporese parovi M1 i M2, M3 i M4 odnosno M5 i M6 vidimo da su metode koje su koristile uzoračku medianu za estimaciju verodostojnosti instance (M2) odnosno fonema (M4 i M6) dale lošije rezultate. Drastičniji primer je ako uzoračku medianu posmatramo na nivou fonema što je u slučaju metoda M4 i M6.

V. ZAKLJUČAK

U ovom radu su dati preliminarni rezultati eksperimenata koji su vezani za primenu normalizacije dužine vokalnog trakta na osnovu ML kriterijuma u ASR sistemima. Primenom odgovarajuće strategije pri izboru VTN koeficijenta moguće je očekivati relativno unapređenje performansi od oko 17 %.

Pored standardne metode određivanja vrednosti VTN koeficijenata predloženo je i implementirano nekoliko jednostavnih robustnih metoda za procenu prosečne verodostojnosti na nivou fonema i instance fonema. Rezultati su pokazali da robustne metode daju lošije rezultate. Jedno od mogućih obrazloženja ovakvog ponašanja bi mogla biti njihova neusklađenost sa principima dekodovanja (prepoznavanja).

Opravdanim se pokazala eliminacija uticaja dužine fonema pri određivanju VTN koeficijenta iako to nije u skladu sa primenjenim principom dekodovanja.

Naredni koraci treba da ispitaju performanse predloženih metoda u realnim nenadgledanim testovima. Ovaj korak podrazumeva i pronalaženje brzih i pouzdanih metoda za estimaciju VTN koeficijenta.

LITERATURA

[1] M. Pitz, "Investigation on Linear Transformations for Speaker Adaptation and Normalization", Ph.D. Thesis University Aachen, Germany, 2005.

[2] D. Pye, P. C. Woodland: "Experiments in Speaker Normalization and Adaptation for Large Vocabulary Speech Recognition" *In Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* Vol. II, Munich, Germany, 1997, pp: 1047-1050.

[3] S. Molau, "Normalization in the Acoustic Feature Space for Improved Speech Recognition", Ph.D. Thesis University Aachen, Germany 2003.

[4] L. Uebel, P. Woodland, "An Investigation into Vocal Tract Length Normalization" *in Proceedings*. *EUROSPEECH99,* 1999. pp: 2527-2530.

[5] P. Zhan, M. Westphal, "Speaker Normalization based on Frequency Warping", *IEEE Int. Conf. on Acoustics, Speech and Signal Processing Proceedings*, 1997 pp: 1039-1042.

[6] P. Zhan, A. Waibel, "Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition" *Language Technologies Institute Technical Report CMI-LTI-97-150*, Pittsburgh 1997.

[7] S. Jovičić, "Govorna komunikacija fiziologija, psihoakustika i precepcija", Nauka, Beograd, 1999, pp: 37-43.

[8] L. Welling, s. Kanthak, H. Ney, "Improved Methods for Vocal Tract Normalization" *In Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* Vol. II, Phoenix, AZ, March 1999, pp. 761-764

[9] N. Jakovljević, D. Mišković, D. Pekar, "Poboljšanje performansi sistem za automatsko prepoznavanje govora primenom modela zavisnih od govornika" Zbornik DOGS2008, Kelebija, Srbija, 2008, pp. 24-27.

ABSTRACT

In this paper preliminary test results of use vocal tract length normalization (VTN) based on ML criterion are presented. Beside the standard method a few alternative methods are proposed and tested. In order to estimate effects of use of VTN only supervised tests are done. Performances of the best system with VTN method obtain about 20 % relative improvement comparing to speaker independent system and 17 % comparing to gender dependent system.

An Investigation into Vocal Tract Length Normalization based on ML criterion

N. Jakovljević, D. Mišković and M. Janev