# Primena neuralnih mreža u prepoznavanju govora u šapatu

Đorđe T. Grozdić, Branko Marković, Jovan Galić, Slobodan T. Jovičić

***Sadržaj* — U radu su prezentovani eksperimentalni rezultati istraživanja prepoznavanja šapata, kao specifičnog oblika verbalne komunikacije, primenom veštačkih neuronskih mreža (ANN). Prikazana je govorna baza reči izgovorenih normalnim govorom i šapatom, posebno formirana za ovo istraživanje, čiji deo je upotrebljen za obuku i testiranje ANN. Testiran je slučaj prepoznavanja zavisno od govornika a rezultati su pokazali 100% prepoznavanja u slučaju govora i 99,3% u slučaju šapata. U slučaju prepoznavanja šapata, kada je ANN obučavana za govor, rezultat prepoznavanja je 59% i obrnuto, kada je ANN obučavana za šapat prepoznavanje govora je bilo 66,4%.**

***Ključne reči* — prepoznavanje govora, šapat, neuralne mreže, govorna baza.**

## I. UVOD

ŠAPAT predstavlja specifičan vid govora koji se često koristi u verbalnoj komunikaciji. Šapuće se u različitim situacijama, na primer kada se želi ostvariti diskretna ili intimna atmosfera u razgovoru. Tako se recimo u čitaonici šapuće kako se ne bi ometali drugi, ili se pak šapuće u cilju sakrivanja nekih poverljivih informacija od ušiju ostalih. Šapat se često koristi i u kriminalnim aktivnostima, najčešće tokom telefonskih razgovora kada se pokušava maskiranje identiteta govornika. Pored navedene upotrebe šapat može biti i posledica zdravstvenih problema, gde se javlja kao posledica prehlada i kijavica (*laringitis* i *rinitis*), dok se kod nekih ljudi javlja i kao hronično oboljenje laringealnih struktura.

Po svojoj prirodi i mehanizmu produkcije šapat se značajno razlikuje od uobičajenog govora. Odlikuje ga odsustvo laringealnih vibracija i šumna struktura govornog signala. Utvrđeno je da su formanti na nižim frekvencijama u šapatu pomereni ka višim frekvencijama, a spektralni nagib je dosta ravniji nego u normalnom govoru [1], [2], [3]. F1-F2 polje (granice formanata) kod samoglasnika u šapatu su drugačije nego u govoru [4]. Pored navedenog, šapat ima dosta nižu energiju u odnosu na govor [5]. Usled odsustva glotalnih vibracija, šapat ne poseduje osnovnu frekvenciju glasa, intonacione konture a samim tim ni mnogo drugih prozodijskih informacija [6].

Šapat zbog svega navedenog predstavlja značajan problem u govornim tehnologijama, posebno u prepoznavanju i sintezi govora, kao i u identifikaciji govornika. Zbog toga jeste aktuelna tema najnovijih istraživanja [1], [7]. Sa druge strane interesantno je da se ovakav tip govorne komunikacije i pored nešto povećanog napora u percepciji obavlja potpuno razumljivo sa praktičnog komunikativnog aspekta. Postavlja se pitanje kako se postiže tako visoka razumljivost šapata s obzirom na njegove bitne razlike u odnosu na govor (najnovija istraživanja su u toku a inicijalni rezultati su dati u [6]).

Postoje razni pristupi, tehnike i metode prepoznavanja govora. Te tehnike su najčešće zasnovane na algoritmima HMM (*Hidden Markov Model*), DTW (*Dynamic Time Warping*), neuralnim mrežama i njihovim hibridnim rešenjima [8]. Zbog sličnosti veštačkih neuralnih mreža sa strukturom čovekovog mozga i njegovim načinom percepcije govora, postavljena je hipoteza da neuralne mreže mogu dati dobre rezultate i u prepoznavanju šapata. U cilju analize ove hipoteze započeta su istraživanja u primeni neuralnih mreža, ali i drugih algoritama, u prepoznavanju šapata. U radu su prikazani preliminarni rezultati ovih istraživanja.

Rad je koncipiran na sledeći način. U odeljku 2 dat je opis govorne baze posebno formirane za ovu vrstu istraživanja, odeljak 3 sadrži prikaz postupka izdvajanja akustičkih obeležja iz govornih stimulusa kao ulaz u ANN, odeljak 4 opisuje karakteristike ANN, odeljak 5 daje prikaz eksperimentalnih rezultata i u zaključku, odeljak 6, biće rezimirani rezultati ovih analiza i naznačeni pravci daljih istraživanja.

## II. OPIS GOVORNE BAZE

Za istraživanje pepoznavanja šapata bilo je neophodno najpre sačiniti govornu bazu. Baza je snimana u prethodnih godinu dana i ona je pri kraju formiranja a za ova preliminarna istraživanja prepoznavanja šapata sa neuralnim mrežama (*Artificial Neural Networks - ANN*) upotrebljen je deo baze.

Govorna baza je označena sa **Whi-Spe** (*Whispered Speech*) i sastoji se od 50 reči: 14 brojeva, 6 boja i 30 reči. Reči su preuzete iz govorne baze GEES [9], te zadovoljavaju osnovne jezičke kriterijume srpskog jezika (distribucija fonema, slogovna kompozicija, akcenatska

Đorđe T. Grozdić, Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11120 Beograd, Srbija (telefon: +381-64-3757595, e-mail: djordjegrozdic@gmail.com). Centar za unapređenje životnih aktivnosti, Gospodar Jovanova 35, 11000 Beograd, Srbija.

Branko Marković, Visoka škola tehničkih strukovnih studija, Svetog Save 65, 32000 Čačak, Srbija (telefon +381-64-3373299, e-mail: branko333@open.telekom.rs).

Jovan Galić, Elektrotehnički fakultet, Univerzitet u Banja Luci, Patre 5, 78000, Banja Luka, Republika Srpska, BiH (telefon +387 65 426629, e-mail: jgalic@etfbl.net).

Slobodan T. Jovičić (autor za kontakte), Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11120 Beograd, Srbija (telefon +381-60-5377689, e-mail: jovicic@etf.rs). Centar za unapređenje životnih aktivnosti, Gospodar Jovanova 35, 11000 Beograd, Srbija.

struktura, konsonantski skupovi). U literaturi se mogu naći podaci o sledećim vrstama govornih baza u šapatu korišćenih u istraživanjima: u [1] su korišćene fonetski balansirane rečenice na japanskom jeziku, u [2] i [7] su korišćene čitane rečenice i rečenice u spontanom govoru na engleskom jeziku, u [4] su korišćeni vokali na švedskom, dok su na srpskom jeziku u [3] korišćeni kontinualno izgovarani vokali, u [6] logatomi tipa /CVC/ gde je C-konsonant i V-vokal i u [5] logatomi tipa /aCa/ (u [6] je dat širi pregled).

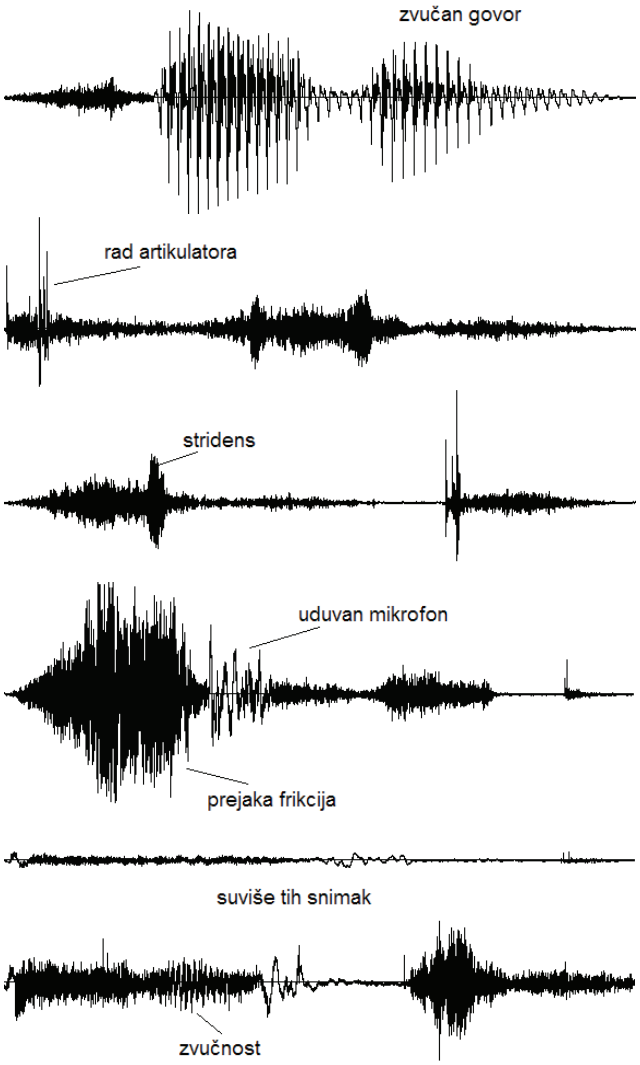
Celu bazu je 10 govornika (5 muških i 5 ženskih) izgovaralo po 10 puta sa normalnim izgovorom i sa šapatom. Svaki izgovor (svaka reč) je segmentiran početak/kraj i memorisan kao poseban fajl. Na taj način bazu **Whi-Spe** čini 5000 fajlova (izgovorenih reči) normalnog govora i 5000 fajlova šapata. Ustanovljena je posebna oznaka fajlova koja pogodno može da se iskoristi u kasnijim eksperimentima.

Baza je snimana na Visokoj školi tehničkih strukovnih studija u Čačku, u laboratorijskim uslovima, kvalitetnim mikrofonom i u adaptiranoj akustičkoj komori smeštenoj u prostoriji laboratorije. U svakoj sesiji subjekti su izgovarali celu bazu u kontinuitetu a sesije su bile razdvojene dužim vremenskim intervalima i po nekoliko dana. Baza je dosnimavana u više navrata jer su kontrolama kvaliteta snimaka ustanovljene različite greške, subjektivne i objektivne prirode. Kod normalnog govora najčešće je dolazilo do pogrešnog izgovora date reči ili pogrešne artikulacije pojedinog glasa u trenutku izgovora reči (što je česta pojava i u svakodnevnom govoru), dok je kod snimanja bio izražen efekat akustičkog udara u mikrofon koji je bio u datom trenutku preblizu usta govornika. Ipak, najveći broj loših snimaka se odnosio na reči izgovorene šapatom. Pošto se ovakva baza šapata po prvi puta formira kod nas, a cilj je bio dobijanje regularno izgovorenih reči šapatom, treba navesti najtipičnije greške kod izgovora i snimanja šapata:

- suviše tih izgovor šapata (značajno maskiran šumom),
- suviše naglašen (izforsiran do izobličenja) šapat,
- probijanje zvučnosti kod izgovora šapata,
- omisija glasova,
- direktno duvanje u mikrofon (akustički udar),
- neregularnost rada artikulatora (pojava stridensa, trenutak odlepljivanja jezika od nepca i sl.).

Na slici 1 prikazan je primer normalnog (zvučnog) izgovora jedne reči i nekoliko tipičnih loših snimaka reči izgovorenih šapatom.

Za ovo preliminarno istraživanje upotrebljen je deo baze koji sadrži sve izgovore brojeva jednog govornika, odnosno 140 izgovora normalnim govorom i 140 šapatom. Dakle, reč je o slučaju prepoznavanja govora/šapata zavisno od govornika.




Slika 1. Signal normalno izgovorene reči i nekoliko primera loših snimaka reči izgovorenih šapatom.

## III. IZDVAJANJE GOVORNIH OBELEŽJA

Mel frekvencijski kepstralni koeficijenti (*Mel Frequency Cepstral Coefficients - MFCC*) se koriste kao standardna govorna obeležja u mnogim sistemima za automatsko prepoznavanje govora. Oni se često dopunjuju sa dinamičkim obeležjima i u tu svrhu se koriste njihov prvi i drugi izvod koji modeluju dinamiku govora.

U ovom radu smo primenili specifičnu formu izračunavanja MFCC koeficijenata. Naime, svaki govorni signal iz formiranog uzorka govorne baze je segmentiran po čitavoj svojoj dužini sa jedanaest *Hamming* prozora koji se međusobno preklapaju 50%. To znači da je i širina *Hamming*-ovog prozora bila promenljiva od reči do reči i da je zavisila od trajanja analizirane reči. Na ovaj način smo pokušali da uskladimo različita trajanja u izgovorima jedne iste reči. Broj jedanaest *Hamming*-ovih prozora po reči je određen na bazi analize raspodele broja glasova u svim rečima u govornoj bazi. Raspon broja glasova ide od 3 do 9 (sa izuzetkom dve reči od 12 i 13 glasova) i sa srednjom vrednošću 5,58 glasova po reči, odnosno sa najčešćim brojem glasova po reči 4, 5, i 6. Usvojili smo 11 *Hamming*-ovih prozora po reči kako bi smo kod najdužih reči u proseku imali po jedan *Hamming*-ov prozor

efektivno po glasu a kod kraćih reči dva do tri prozora po glasu. Sa druge strane, pretpostavili smo da ovako finija vremensko-spektralna rezolucija kraćih reči treba da omogući njihovo bolje prepoznavanje, dok takav dobitak duge reči ostvaruju na račun bogatijeg fonetskog sadržaja.

Iz svakog prozorovanog segmenta je izdvojeno po 12 prvih MFCC koeficijenata i formiran je vektor dužine od 132 koeficijenta. Na sličan način su formirani vektori prvog i drugog izvoda MFCC, takođe dužine 132 koeficijenta. Ovi vektori su potom združeni u jedan vektor dužine 396 koeficijenata. Ovaj postupak je izvršen za svaku stimulus-reč i formirane su matrice govornih obeležja – jedna za govor a druga za šapat, svaka dimenzija 396x140. Ove matrice su potom korišćene kao ulaz u ANN radi njihove obuke i testiranja.

## IV. NEURALNE MREŽE

U ovom istraživanju za klasifijaciju govora su korišćene *Feedforward* neuralne mreže sa *Back Propagation* algoritmom u procesu obuke mreže. Neuralne mreže su realizovane pomoću MATLAB *Neural Network Toolbox* [10]. Napravljene su dve mreže – jedna za govor i jedna za šapat. Mreže su iste strukture kako bi se uporedile njihove preformanse u prepoznavanju govora i u prepoznavanju šapata.

Mreže su sačinjene iz tri sloja: jednog ulaznog, jednog skrivenog i jednog izlaznog sloja. Kako se na ulaz mreže dovode vektori dužine 396 koeficijenata, ulazni sloj mreže ima 396 neurona. Broj neurona u izlaznom sloju mreže je 14 kao i broj reči za prepoznavanje, dok je broj neurona u skrivenom sloju menjan tokom eksperimenta kako bi se postigle željene performanse mreže. U mreži su korišćene *tansig* (hiperbolički tangens sigmoid) transfer funkcije.

Ukupna baza podataka na kojoj je vršena obuka mreže je podeljena na tri manje celine: 70% baze je korišćeno za trening, 15% za validaciju i 15% za testiranje. Mreže su obučavane pomoću *trainscg* funkcije koja je zasnovana na algoritmu skaliranih konjugovanih gradijenata. Ovaj algoritam je razvio *Martin Moller* i predstavlja kombinaciju *Levenberg-Marquardt* algoritma i principa skaliranih konjugovanih gradijenata [10]. *Trainscg* obuka mreže zahteva nešto veći broj iteracija ali zato znatno smanjuje ukupan broj računskih operacija i potrebno vreme za obuku mreže, što čini ovu funkciju pogodnom za velike neuralne mreže. Kao kriterijumi za zaustavljanje obuke mreže korišćeni su: definisani maksimalni broj iteracija (1000), srednja kvadratna greška (0,00), maksimalni broj uzastopnih grešaka u validaciji tzv. *early stopping* metoda (6) i prag gradijenta ($10^{-6}$).

Izmene topologije mreže u pogledu broja skrivenih slojeva nisu imale bitnijeg uticaja, pa se ostalo pri prvobitnoj strukturi mreže sa tri sloja (jednim skrivenim slojem). Isprobavanjem i menjanjem broja neurona u skrivenom sloju dobijeni su zadovoljavajući rezultati. Upotrebom opcije *re-train* u MATLAB-u mreže su dodatno obučene i optimizovane po svojim performansama u prepoznavanju govora, pre svega u pogledu srednje kvadratne greške (*Mean Squared Error - MSE*) i ukupne greške napravljene u klasifikaciji tokom obuke, validacije i testiranja mreže.

## V. REZULTATI EKSPERIMENTA

Prvi korak u analizi ANN je bio da se odredi struktura neuralne mreže koja daje najbolje rezultate u pogledu uspešnosti prepoznavanja govora i šapata. Formirane su dve neuralne mreže sa po 10 neurona u skrivenom sloju, a potom je broj neurona postepeno povećavan. Mreže su obučavane odvojeno jedna na neutralnom govoru, a druga na šapatu. Praćene su performanse obe mreže.

TABELA 1: USPEH U PREPOZNAVANJU GOVORA I ŠAPATA U ZAVISNOSTI OD BROJA NEURONA U SKRIVENOM SLOJU ANN.

| Broj neurona | Prepoznavanje govora (%) | Prepoznavanje šapata (%) |
|---|---|---|
| 10 | 67,44 | 42,46 |
| 15 | 79,19 | 80,81 |
| 20 | 87,05 | 90,35 |
| 25 | 89,44 | 89,92 |
| 30 | 99,29 | 95,70 |
| 35 | 96,43 | 79,42 |

Dobijeni rezultati su prikazani u tabeli 1. Mreže sa po 10 neurona u skrivenom sloju su pokazale najmanji uspeh u prepoznavanju ulaznog signala. U slučaju govora prepoznavanje je bilo 67,40% a kod šapata 42,46%. Sa povećanjem broja neurona u skrivenom sloju performanse mreža su bivale sve bolje. Šapat je već u mreži sa 15 neurona bio prepoznat sa 80,8% tačnosti dok je govor bio prepoznat sa 79,19%. Mreže sa 20 neurona su pokazale uspeh u prepoznavanju ulaznog signala od oko 90%. Povećanjem broja neurona na 25 nije imalo bitnijeg uticaja na performanse mreže. Mreže su najbolje prepoznavanje imale sa 30 skrivenih neurona. Mreža te topologije obučavana govorom je ostvarila uspeh u prepoznavanju od 99,29%, dok je mreža obučavana šapatom imala prepoznavanje od 95,70%. Sa daljim povećanjem broja neurona u skrivenom sloju performanse mreže su opadale.

Mreže sa 30 neurona u skrivenom sloju su dodatno obučavane kako bi se optimizovale njihove performanse. Postignut je maksimum prepoznavanja od 100% za govor, i 99,29% za šapat. Ovako obučene mreže su potom testirane tako što je na ulaz mreže obučene govorom dovođen šapat i obrnutno. Dobijeni rezultati su prikazani u tabeli 2.

TABELA 2: PREPOZNAVANJE U ZAVISNOSTI OD TIPA MODALITETA ODNOSA OBUKA/TEST.

| Obuka / Test | Prepoznavanje (%) |
|---|---|
| Govor / Govor | 100,0 |
| Govor / Šapat | 59,0 |
| Šapat / Šapat | 99,3 |
| Šapat / Govor | 66,4 |

Kada se na ulaz mreže obučene govorom dovede šapat, njegova verovatnoća da bude tačno prepoznat je 59,0%. Interesantno je da kada se na mrežu obučenu šapatom dovede govor da se dobijaju bolji rezultati (66,4%) u prepoznavanju nego u obrnutom slučaju. Slični rezultati su dobijeni i u testiranju GMM (*Gaussian Mixture Models*) sistema za identifikaciju govornika [7], sa istim

kombinacijama obuka/test govora i šapata. Razlika nije velika (7,4%) ali jeste indikativna u korist obuke ANN šapatom. Ako se tome doda da se sa obukom ANN govorom - šapat lošije prepoznaje za 41%, i obrnuto sa obukom ANN šapatom - govor lošije prepoznaje za 29,9%, tada ova nesimetrija ukazuje da je ANN bolji prepoznavač za bimodalnu formu obuka/test - *šapat/govor*. Za dublje tumačenje ove pojave potrebna su dodatna istraživanja.

## VI. DISKUSIJA I ZAKLJUČAK

Prikazani eksperimentalni rezultati prepoznavanja šapata pomoću ANN su prvi preliminarni rezultati istraživanja u oblasti prepoznavanja šapata kao specifičnog oblika verbalne komunikacije. Motiv za ovu vrstu istraživanja jeste činjenica da šapat sadrži veoma visok nivo razumljivosti. Činjenica da šapat ne sadrži laringealne vibracije čine šapat potpuno devokalizovanim modom govora. Utoliko je intrigantnije pitanje – kako čovek sa tako visokom razumljivosti percipira šapat? Tekuća istraživanja šapata [6] ukazuju da se na nivou glasova najveća konfuzija u percepciji šapata dešava na nivou zvučnih/bezvučnih glasovnih parova a da najveću razumljivost nose intenzitetsko-vremensko-spektralne informacije u signalu šapata. Sa druge strane, zvučnost kao osnovni nosilac prozodijskih informacija u najvećoj meri nosi nelingvističke informacije. U ovom kontekstu treba istaći činjenicu da se u šapatu teško može prepoznati pol govornika ili emocija.

Iznete činjenice ukazuju na veliki prostor fundamentalnih istraživanja ne samo na akustičkom i jezičkom planu nego i kognitivnom. Posebno u domenu govornih tehnologija na algoritamskom planu postaje interesantno koliko su postojeći algoritmi kao što su DTW, HMM, ANN i drugi, efikasni u prepoznavanju šapata. Posebno ako se ima u vidu govornik koji ispred ASR (*Automatic Speech Recognition*) sistema iz govornog moda pređe u šapat.

Prikazani rezultati, i ako su preliminarni, ukazuju na neophodnost daljih istraživanja u prepoznavanju šapata sa ANN. Prethodna diskusija ukazuje na veliki značaj izbora i načina dobijanja akustičkih obeležja. Jer ako govor posmatramo trodimenzionalno u formi spektrograma tada možemo zaključiti da je izuzetan značaj ne samo dugovremenih već i kratkovremenih informacija, koje ostaju maskirane zvučnošću u normalnom govoru, dok se u šapatu pojavljuju kao izuzetni markeri pojedinih fonetskih obeležja ili značenja [6]. Rezultat u tabeli 2 o prepoznavanju šapata odnosno govora, kada je ANN obučena obrnuto govorom odnosno šapatom, respektivno, ide u prilog prethodnim razmatranjima u vezi ulaznih informacija u ANN.

Takođe, buduća istraživanja treba da obuhvate i komparativnu analizu gore pomenutih algoritama, i/ili njihovih kombinacija.


## ZAHVALNICA

Rad je finansiran sredstvima Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije u projektima No. 32032 i No. 178027.


## LITERATURA


[1] T. Ito, K. Takeda, F. Itakura, "Analysis and Recognition of Whispered speech," *Speech Communication*, 2005, pp.129-152.

[2] C. Zhang, J.H.L. Hansen, "Analysis and classification of Speech Mode: Whisper through Shouted," *Interspeech 2007*, 2007, pp. 2289-2292.

[3] S.T. Jovičić, "Formant feature differences between whispered and voiced sustained vowels," *ACUSTICA - Acta Acoustica*, vol. 84, no. 4, 1998, pp. 739-743.

[4] I. Eklund, H. Traunmuller, "Comparative Study of Male and Female Whispered and Phonated Versions of the Long Vowels of Swedish," *Phonetica*, 1996, pp. 1-21.

[5] S.T. Jovičić, Z. M. Šarić, „Acoustic analysis of consonants in whispered speech," *Journal of Voice*, Vol. 22, No. 3, 2008, pp. 263-274.

[6] S. Jovičić, M. Đorđević, "Percepcija fonema u šapatu: identifikacija i konfuzija," poglavlje u knjizi S.T. Jovičić, M. Sovilj (ed.): *Govor i jezik: interdisciplinarna istraživanja*, II; Centar za unapređenje životnih aktivnosti i Institut za eksperimentalnu fonetiku i patologiju govora, 2008, str. 80-95.

[7] X. Fan, J.H.L. Hansen, "Speaker identification within Whispered Speech Audio Stream," *Transactions on Audio, Speech and Language Processing*, Vol. 19, No. 5, 2011, pp. 1408-1421.

[8] J. Holms, W. Holms, "Speech Synthesis and Recognition," Taylor & Francis, London, 2001.

[9] S.T. Jovičić, Z. Kašić, M. Đorđević, M. Vojnović, M. Rajković, J. Savković, „Korpus psiho-emotivnog govora u srpskom jeziku," poglavlje u knjizi S.T. Jovičić, M. Sovilj (urednici): *Govor i jezik: interdisciplinarna istraživanja srpskog jezika*, I; 2004, str. 20-45.

[10] H. Demuth, M. Beale, "Neural Network Toolbox User's Guide," The MathWorks, Inc, 2002.


## ABSTRACT


This paper presents the results of an experimental research of recognition of whispered speech, as a specific form of verbal communication, based on application of artificial neural networks (ANN). The paper also describes the speech database of words that were spoken in whispered and normal manner, which was especially created for this study. Part of this database was used for preliminary training and testing the ANN. The case of the speaker dependent recognition was tested, and the results showed 100% accuracy in the case of speech recognition and 99,3% in the case of whisper recognition. In the case of whisper recognition, when ANN was trained for the normal speech the score of whisper recognition was 59%, and vice versa, when the ANN was trained with whisper the speech recognition was 66.4%.


## APPLICATION OF NEURAL NETWORKS IN WHISPERED SPEECH RECOGNITION


Đorđe T. Grozdić, Branko Marković, Jovan Galić, Slobodan T. Jovičić