Barbara Hajdarević i Ratko Amanović

Analiza uspešnosti identifikacije govornika korišćenjem obeležja u spektralnom i vremenskom domenu

Analizirana je uspešnost identifikacije govornika na zatvorenoj bazi, nezavisno od sadržaja govora, korišćenjem veštačkih neuronskih mreža. Korišćena karakteristična obeležja govornog signala su kepstralni koeficijenti mel skale, kepstralni koeficijenti linearne predikcije i linijski spektralni parovi. Korišćena je baza na srpskom jeziku koja sadrži govorne signale 22 muškaraca i 22 žene, gde za svaku osobu postoji 60 različitih govornih signala. Najveća dobijena tačnost prepoznavanja je 86%, gde identifikaciona karakteristika predstavlja kombinaciju kepstralnih koeficijenata mel skale i kepstralnih koeficijenata linearne predikcije, a prepoznaje se 10 govornika. Pokazano je da tačnost prepoznavanja ne zavisi od pola govornika, niti od broja slojeva neuronske mreže.

Uvod

Biometrijska obeležja su karakteristične osobine po kojima se ljudi mogu razlikovati. Dele se na fizičke karakteristike i karakteristike ponašanja. Fizičke karakteristike su otisak prsta, raspored vena, glas, dužica oka, crte lica itd. Neke od karakteristika ponašanja su intonacija govora, dinamika hoda, način kucanja na tastaturi i one nisu toliko pouzdane, jer se mogu voljno kontrolisati. Biometrijska obeležja su jedinstvena za svakog čoveka i kao takva su pogodna za korišćenje prilikom formiranja sistema za identifikaciju i verifikaciju ljudi.

Glas kao biometrijsko obeležje je pogodno za identifikaciju ljudi. Prednosti korišćenja glasa su te što je pristupačnost laka, za razliku od npr. DNK, međutim mana korišćenja glasa za prepoznavanje osoba je ta što se glas menja tokom vremena, odnosno osnovna frekvencija glasa opada starenjem. Jedne od osnovnih osobina koje karakterišu ljudski glas su intenzitet, visina i boja glasa.

Glas kao zvučni signal može se predstaviti i neparametarskim i parametarskim metodama. Neparametarska reprezentacija ljudskog glasa predstavlja odbirke zvučnog signala u vremenu, talasni oblik signala. Parametarska reprezentacija se deli na parametre eksitacije i parametre vokalnog trakta. U parametarsku reprezentaciju spadaju usrednjena energija, kepstralni koeficijenti, koeficijenti linearne predikcije, formanti itd.

Sistemi za prepoznavanje osoba koji koriste glas kao obeležje mogu biti zavisni ili nezavisni od reči koja se izgovara. Takođe ovi sistemi mogu da rade na zatvorenom skupu ljudi, kada je onaj ko govori neko od N unapred poznatih osoba ili na otvorenom skupu ljudi kada osoba koja se identifikuje ili verifikuje ne mora da pripada bazi od N poznatih osoba. Da bi se dobio podatak o kom je govorniku reč potrebno je najpre odrediti karakteristike glasa svakog od govornika na osnovu kojih se vrši klasifikacija. Neke od karakteristika glasa koje se koriste prilikom identifikacije ljudi su kratkovremenske osobine glasa u

Barbara Hajdarević (1998), Beograd, Radovana Šimića Cige 5, učenica 3. razreda Matematičke gimnazije u Beogradu

Ratko Amanović (1997), Smederevska Palanka, Đure Đakovića 15, učenik 4. razreda Palanačke gimnazije

MENTORI:

Natalija Todorčević, student Elektrotehničkog fakulteta Univerziteta u Beogradu

Vladimir Ranković, diplomirani inženjer elektrotehnike, Microsoft Development Center Serbia

koje spada kratkovremenska brzina prolaska kroz nulu i kratkovremenska energija signala, a pored toga se koriste i koeficijenti mel skale, kepstralni koeficijenti linearne predikcije itd.

U ovom radu korišćene osobine glasa su koeficijenti mel skale, kepstralni koeficijenti linearne predikcije i linijski spektralni parovi (Islam *et al*. 2013). Kao klasifikator korišćena je veštačka neuronska mreža. Ona se sastoji od velikog broja neurona, odnosno čvorova koji su povezani težinskim koeficijentima, i ima tu osobinu da prilikom obučavanja vrši prilagođavanje koeficijenata zadatom problemu. Cilj rada jeste da se analizira uspešnost identifikacije govornika na zatvorenoj bazi nezavisno od sadržaja govora. Uspešnost je ispitivana za različite kombinacije karakterističnih obeležja glasa kao i za različit broj slojeva i neurona u neuronskoj mreži. Takođe je izvršeno ispitivanje da li prepoznavanje zavisi od toga da li je govornik muška ili ženska osoba. Za svrhe istraživanja formirana je baza na srpskom jeziku koja sadrži govorne signale muškaraca i žena, 22 osobe muškog i 22 osobe ženskog pola. Za svaku osobu postoji po 60 različitih govornih signala.

Metod

Algoritam za identifikaciju osoba se sastoji iz tri dela (Campbell 1997):

1. Izdvajanje vektora karakterističnih obeležja govornog signala
2. Obučavanje neuronske mreže
3. Testiranje neuronske mreže

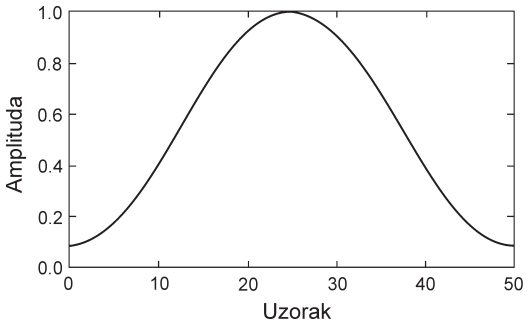
Izdvajanje vektora karakterističnih obeležja

U ovom radu su korišćena sledeća karakteristična obeležja govornog signala: kepstralni koeficijenti mel skale, kepstralni koeficijenti linearne predikcije i linearni spektralni parovi.

Kepstralni koeficijenti mel skale. Kepstralni koeficijenti mel skale se skraćeno nazivaju MFCC (eng. mel frequency cepstral coefficients). Upotreba ovih koeficijenata u prepoznavanju govora je česta. Mel skala je dobijena eksperimentalnim putem tako da najviše odgovara načinu na koji čovek percipira frekvencije. Čovek niže frekvencije bolje razlikuje od viših frekvencija (Masterton 1993). Mel skala je logaritamska i to je čini približnom ljudskom sluhu.

Proces određivanja MFCC se može podeliti u više faza.

1. Primena prozorske funkcije na signal govora: govorni signal je podeljen na prozore širine 25 ms, pri čemu svaki novi prozor počinje 10ms od početka prethodnog. Na taj način dolazi do preklapanja prozorskih funkcija. Svaki prozor govornog signala pomnožen je Hamingovom funkcijom (slika 1). Ova funkcija se koristi da bi se ublažilo spektralno curenje.



Slika 1. Hamingov prozor

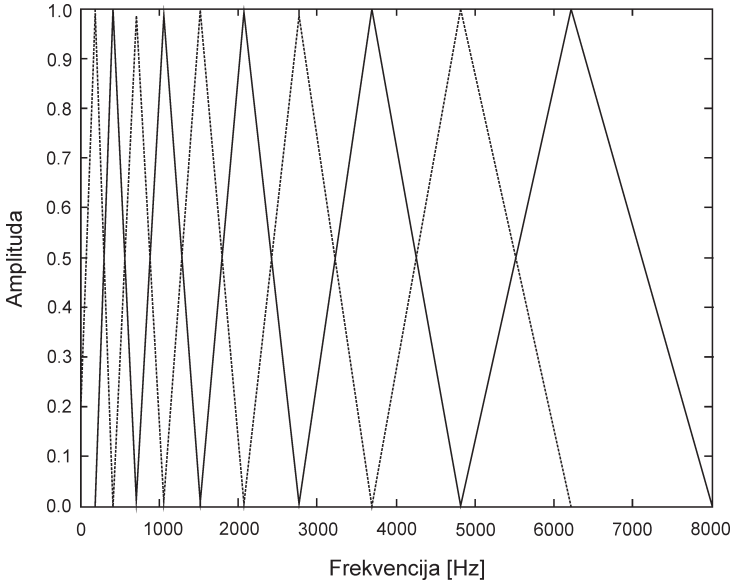
Figure 1. Hamming's window

2. Računanje brze Furijeove transformacije (eng. fast Fourier transformation – FFT): Unutar svakog prozora određena je brza Furijeova transformacija. Dobijene vrednosti su kvadrirane i na taj način je dobijen spektogram snage signala, na osnovu čega se može posmatrati promena snage signala tokom vremena.

3. Formiranje filtar banke: Formirana je filtar banka koja obuhvata frekvencijski opseg od 0 do 4000 Hz. Ove frekvencije su izabrane na osnovu raspona frekvencija koje čovek može da percipira. Frekvencije su preračunate u mel skalu pomoću formule:

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right)$$

Između odgovarajućih tačaka u mel skali je linearno raspoređeno onoliko tačaka koliko ima



Slika 2. Filtar banka

Figure 2. Filter bank

filtara unutar banke, a potom su te vrednosti preračunate u Hz sledećom formulom:

$$M^{-1}(m) = 1125 \ln\left(1 + \frac{m}{1125}\right)$$

Nad dobijenim tačkama u Hz formirana je filtar banka prikazana na slici 2. Svaki filtar je obuhvatao tri uzastopne tačke – u prvoj je počinjao i ima vrednost 0, u drugoj je dostizao maksimum i imao vrednost 1 i u trećoj se vraćao u nulu.

4. Filtriranje signala i izračunavanje MFCC: Svaki prozor govornog signal je filtriran filtar bankom. Informacija koju signal nosi unutar frekvencijskog opsega jednog filtra dobijena je sumiranjem i logaritmovanjem filtriranih vrednosti signala. Na taj način u okviru svakog prozora dobija se onoliko koeficijenata koliko ima filtara unutar filtar banke. Nakon toga nad signalom se vrši inverzna brza Furijeova transformacija i dobijaju se kepstralni koeficijenti mel skale.

Kepstralni koeficijenti linearne predikcije. Kepstralni koeficijenti linearne predikcije se skraćeno nazivaju LPCC (eng. linear prediciton cepstral coefficients). Oni se prvenstveno koriste jer određuju osnovne parametre govora uz minimalnu numeričku složenost. Kepstralni koeficijenti linearne predikcije se baziraju na principu da se trenutni odbirak signala može predvideti pomoću linearne kombinacije određenog broja odbiraka u prošlosti. Svaki odbirak iz prošlosti utiče na predikciju određenom merom, pa se pre sumiranja množe odgovarajućim težinskim koeficijentima. Upravo ti koeficijenti su kepstralni koeficijenti linearne predikcije.

Ovi koeficijenti se dobijaju na sledeći način:

1. Izdvajanje koeficijenata linearne predikcije (eng. linear prediction coefficients, LPC). Koeficijenti linearne predikcije na osnovu početnog signala predviđaju njegove buduće vrednosti. Na osnovu vrednosti N prethodnih odabiraka govornog signala, uz pomoć LPC moguće je odrediti vrednost $(N+1)$-og odabirka. LPC zapravo predstavljaju koeficijente kojima treba izmnožiti vrednosti svakog od N odabiraka kako bismo dobili vrednost $(N+1)$-og. Prilikom traženja ovih koeficijenata prolazi se kroz ceo govorni signal i računa se kojih N koeficijenata najbolje predviđa vrednost $(N+1)$-og odbirka u opštem slučaju za ceo signal.

2. Kepstar od LPC. Kepstralni LPC (LPCC) dobijaju se primenom kepstra na LPC. Prvi korak je računanje apsolutne vrednosti kvadrata brze Furijeove transformacije LPC-a. Potom se od dobijenih koeficijenata računa logaritam i primeni inverzna brza Furijeova transformacija.

Linijski spektralni parovi. Linijski spektralni parovi (Saha *et al*. 2010) se nazivaju i LSF (skraćeno od eng. line spectral frequencies). Oni predstavljaju drugačiju reprezentaciju koefici-

jenata linearne predikcije. Linearnu predikciju možemo predstaviti u obliku polinoma *N*-og stepena:

$$A(z) = 1 + \sum_{i=1}^{N} a_i z^{-i}$$

gde su a_i kepstralni koeficijenti linearne predikcije. Dekompozicija ovog polinoma na polinome *P* i *Q* se može zapisati na sledeći način:

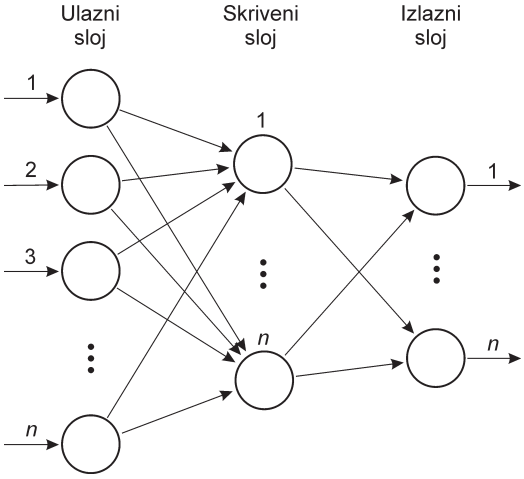
$$P(z) = A(z) + z^{-(p+1)} A(z^{-1})$$
$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1})$$

Argumenti pozitivnih rešenja polinoma *P* i *Q* predstavljaju LSF parove.

Neuronske mreže

Prilikom identifikacije govornika korišćene su veštačke neuronske mreže. One se sastoje od neurona, tj. čvorova, koji se mogu podeliti u tri sloja: ulazni, skriveni i izlazni sloj. Primer strukture neuronske mreže je prikazan na slici 3. Neuroni su međusobno povezani granama koje imaju određene težinske koeficijente. Neuronska mreža prilikom formiranja prolazi kroz tri faze: obučavanje, verifikacija i testiranje. Obučavanje predstavlja adaptiranje mreže na određeni problem u vidu podešavanja težinskih koeficijenata između neurona. Verifikacija se sprovodi tokom obučavanja i služi da se mreža ne adaptira više na skup podataka za obučavanje. Testiranje mreže se vrši na samom kraju.

Baza podataka je podeljena na tri dela: skup podataka za obučavanje mreže, skup podataka za verifikaciju i skup podataka za testiranje.



Slika 3. Neuronska mreža

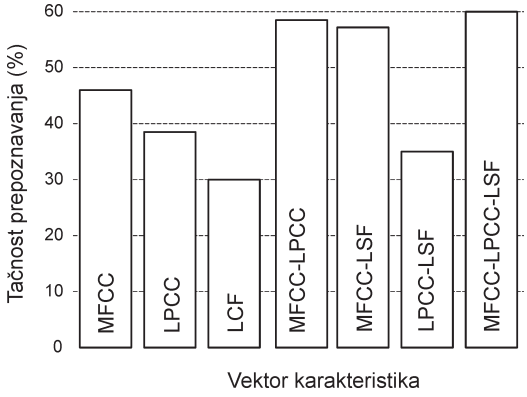
Figure 3. Neural network

Rezultati i diskusija

Prilikom testiranja algoritma za identifikaciju ljudi varirani su sledeći parametri:

1. kombinacija vektora karakterističnih obeležja na ulazu u mrežu
2. broj neurona i skrivenih slojeva unutar mreže
3. broj govornika
4. pol govornika prilikom identifikacije

Na slici 4 prikazan je grafik zavisnosti procenta tačne identifikacije od vektora karakterističnih obeležja, koji je dobijen na osnovu eksperimenta gde je korišćena veštačka neuralna mreža sa jednim slojem koji ima 60 neurona. Na osnovu datih rezultata došlo se do zaključka da od pojedinačnih vektora karakteristika MFCC daje najbolje rezultate koji iznose 46%. Kombinacijom dva različita karakteristična obeležja dobijaju se najbolji rezultati za MFCC i LPCC. Navedena kombinacija karakteristika daje tačnost od 57%. Dok kombinacija sva tri karakteristična obeležja daje najbolje rezultate, odnosno postiže se procenat tačnosti od 60%.



Slika 4. Grafik zavisnosti tačnosti prepoznavanja od vektora karakteristika

Figure 4. Accuracy of recognition as a function of features

Tabela 1. Zavisnost tačnosti prepoznavanja (u procentima) od broja neurona u jednom skrivenom sloju za različite vektore karakteristika

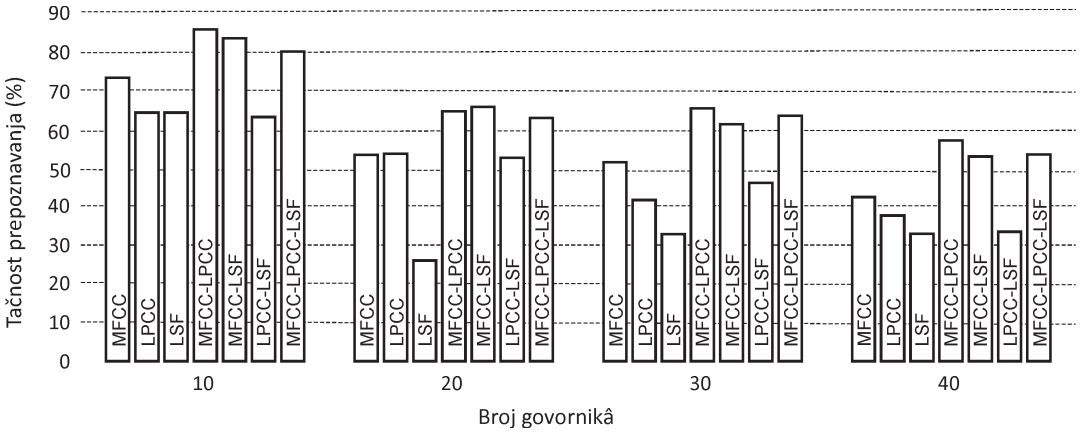
Vektor karakteristika	Broj neurona													
	5	10	15	20	25	30	35	40	45	50	55	60	65	70
MFCC	21	39	42	43	40	43	43	48	45	40	42	46	44	47
LPCC	19	27	26	30	30	35	32	37	34	37	31	38	36	39
LSF	15	23	28	27	29	26	26	27	31	33	32	30	35	32
MFCC-LPCC	32	44	47	52	56	55	52	50	53	53	54	58	58	55
MFCC-LSF	28	35	41	50	48	53	45	48	51	51	46	57	56	52
LPCC-LSF	22	32	32	34	34	34	34	36	40	34	36	35	39	37
MFCC-LPCC-LSF	27	39	46	44	53	46	56	55	49	54	52	60	52	54

Tabela 2. Zavisnost tačnosti prepoznavanja (u procentima) od broja neurona u dva skrivena sloja

Vektor karakteristika	Broj neurona													
	5	10	15	20	25	30	35	40	45	50	55	60	65	70
MFCC	18	30	35	39	40	42	42	45	45	46	43	48	47	47
LPCC	9	24	29	30	31	36	32	36	37	32	37	37	33	37
LSF	9	18	20	23	23	34	24	31	28	30	30	30	29	35
MFCC-LPCC	19	41	45	46	45	52	55	52	54	52	52	53	50	56
MFCC-LSF	19	37	42	48	45	48	51	47	53	52	52	55	52	57
LPCC-LSF	14	28	33	33	27	38	33	35	37	40	35	36	31	42
MFCC-LPCC-LSF	16	42	42	45	49	52	48	55	58	56	54	49	57	58

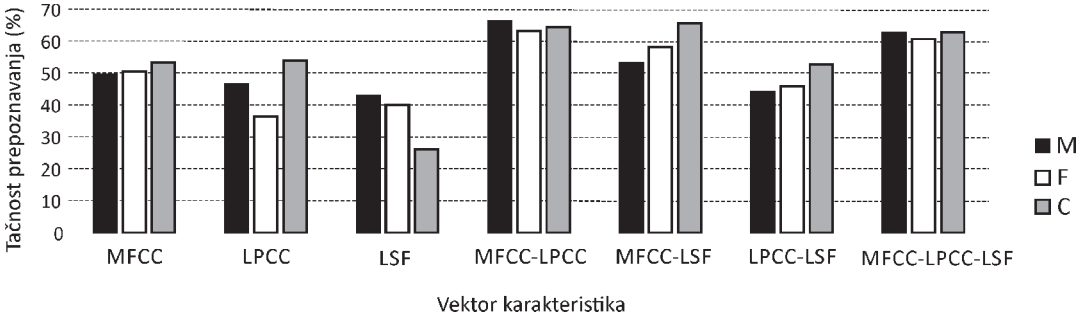
Tabela 3. Zavisnost tačnosti prepoznavanja (u procentima) od broja neurona u tri skrivena sloja

Vektor karakteristika	Broj neurona													
	5	10	15	20	25	30	35	40	45	50	55	60	65	70
MFCC	5	20	32	39	40	46	40	42	42	47	45	45	44	42
LPCC	2	13	16	21	30	35	34	33	32	30	34	31	37	32
LSF	6	8	24	29	26	32	24	24	27	26	30	24	32	32
MFCC-LPCC	6	27	34	41	43	42	46	48	55	52	49	54	52	59
MFCC-LSF	4	14	33	36	39	44	50	43	46	47	52	54	55	56
LPCC-LSF	5	17	23	30	28	31	36	35	38	32	32	38	36	34
MFCC-LPCC-LSF	6	28	34	39	41	46	53	49	49	49	50	47	55	52



Slika 5. Grafik zavisnosti tačnosti prepoznavanja od broja govornika

Figure 5. Accuracy of recognition as a function of number of speakers



Slika 6. Grafik zavisnosti tačnosti prepoznavanja od pola govornika: M – muški, F – ženski, C – kombinovano

Figure 6. Accuracy of recognition as a function of gender of speaker: M – male, F – female, C – both

U tabeli 1 je prikazan procenat tačnosti prepoznavanja govornika u zavisnosti od broja neurona u jednom skrivenom sloju za različite vektore karakteristika. Analizom dobijenih rezultata zaključuje se da procenat tačnosti ima tendenciju rasta sa povećanjem broja neurona dok se ne dostigne 15 neurona, nakon čega se ne uočava pravilnost i tačnost identifikacije stagnira unutar uskog opsega – smatra se da je neuralna mreža ušla u zasićenje svog kapaiteta.

U tabelama 2 i 3 prikazana je zavisnost procenta tačnog prepoznavanja u zavisnosti od broja neurona u skrivenim slojevima. U tabeli 2 su prikazani rezultati koji su dobijeni testiranjem mreže sa dva skrivena sloja, dok su u tabeli 3 rezultati dobijeni testiranjem mreže sa tri skrivena sloja. U skrivenim slojevima se nalazi isti broj neurona i kreće se od 5 do 70.

Posmatranjem dobijenih rezultata za neuronske mreže sa dva i tri skrivena sloja uočava se sličan efekat stagnacije povećanja tačnosti sa povećanjem broja neurona nakon određenog praga, s tim što je sada prag uočen na 30 neurona po sloju. Takođe, zaključuje se i da ne postoji značajna razlika u tačnosti identifikacije za mreže sa različitim brojem skrivenih slojeva.

Na slici 5 prikazana je tačnost prepoznavanja u procentima u zavisnoti od broja govornika. Analizom dobijenih rezultata dolazi se do zaključka da sa smanjenjem broja govornika unutar baze raste procenat tačnog prepoznavanja.

Na slici 6 prikazan je grafik zavisnoti procenta tačnosti od pola govornika i kombinacije različitih vektora karakterističnih obeležja. Došlo se do zaključka da opisani algoritam za identifikaciju ljudi, u konfiguraciji kada postiže maksimalnu tačnost, nije osetljiv na pol govornika i da se za mušku, žensku i kombinovnu bazu dobijaju isti rezultati.

Zaključak

Od pojedinačnih vektora karakterističnih obeležja MFCC ima najveći procenat tačnosti koji za 10 govornika iznosi 73%. Najveća dobijena tačnost prepoznavanja je 86%, gde identifikaciona karakteristika predstavlja kombinaciju kepstralnih koeficijenata mel skale i kepstralnih koeficijenata linearne predikcije, a prepoznaje se 10 govornika. Takođe je pokazano da sa porastom broja govornika opada procenat tačnosti prepoznavanja, kao i da pol govornika i broj slojeva neuronske mreže ne utiču na tačnost prepoznavanja.

Literatura

Campbell J. P. 1997. Speaker recognition: A tutorial. *Proceedings of the IEEE*, **85**: 1437.

Islam M., Khan F. H., Haque A. A. M. 2013. A novel approach for text-independent speaker identification using artificial neural network. *International journal of innovative research in computer and communication engineering*, **1** (4): 838.

Masterton B. 1993. Central auditory system. *ORL; Journal of otorhinolaryngology and related specialities*, **55** (3): 159.

Practical Cryptography. http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/.

Saha G., Chakroborty S., Sahidullah M. 2010. On the use of perceptual line spectral pairs frequencies and heigher-order residual moments for speaker identification. *International journal of biometrics*, **2** (4): 358.

Barbara Hajdarević and Ratko Amanović

Analysis of Speaker Identification through Features in Spectral and Time Domain

This paper analyzes the performance of speaker identification through the use of artificial neural networks. The features of the speech signals that were used are Mel frequency cepstral coefficients (MFCC), cepstral coefficients of linear prediction (LPCC) and line spectral pairs (LSF). The used database is in Serbian and it includes speech signals from 44 people (22 female and 22 male speakers). For every person there are 60 speech signals. The accuracy achieved for the database was 86%, for 10 speakers, and the features MFCC, LPCC and LSF. Results show that the accuracy of speaker identification does not depend on the sex of the speaker nor on the number of hidden layers in the neural network.