Uticaj klipovanja govornog signala na formantne frekvencije vokala

Milan Vojnović

***Sadržaj* — Kvalitet snimaka glasa koji se koriste u forenzičkoj identifikaciji govornika je uglavnom loš. Najčešće su u pitanju telefonski razgovori koje karakteriše uzak frekvencijski opseg (do 4 kHz) i veliki broj različitih vrsta smetnji i izobličenja. Odsecanje vršnih vrednosti signala (klipovanje) je vrlo česta pojava u telefonskom govornom signalu. U radu je analizirano kako klipovanje utiče na tačnost estimacije formantnih frekvencija vokala. Rezultati analize pokazuju da samo umereno klipovanje (do 5 dB) može da se toleriše u postupku forenzičke identifikacije govornika.**

***Ključne reči* — formantne frekvencije vokala, klipovanje, identifikacija govornika.**

I. UVOD

U forenzičkoj identifikaciji govornika egzistiraju dve vrste snimka govora: nesporni i sporni. Nesporni snimak govora podrazumeva da je identitet govornika poznat, za razliku od spornog snimka gde on nije poznat. Najčešći zahtev u forenzičkoj identifikaciji govornika je da se utvrdi stepen sličnosti glasova govornika na spornom i nespornom snimku. U praksi ne postoji samo jedan sporni i jedan nesporni snimak glasa već više njih. Postupak identifikacije govornika podrazumeva međusobno poređenje svih spornih sa svim nespornim snimcima glasa.

Sporni snimci su najčešće telefonski razgovori dobijeni tokom istražnog postupka. Pojam "telefonskog razgovora" je u današnje vreme nešto proširen i podrazumeva sve vidove govorne komunikacije na daljinu: klasična stacionarna telefonska mreža, mobilna telefonska mreža, govorna komunikacija preko internet mreže i dr. Drugi metod za dobijanje spornog snimka je pomoću prislušnih uređaja. Misli se na one prislušne uređaje koji se ne koriste u telefonskim centralama i na telefonskim linijama.

Dakle, u forenzičkoj identifikaciji govornika postoje dva tipa spornog snimka govora:

- snimak govora osumnjičenog kada on koristi komunikaciona sredstva i
- snimak govora osumnjičenog u neposrednoj govornoj komunikaciji sa drugim osobama.

U prvu grupu snimaka spadaju telefonski razgovori, a u drugu grupu snimci dobijeni pomoću prislušnih sredstava.

Ovaj rad je nastao u okviru projekta "E-medicine sistem za procenu kvaliteta sluha" broj 13011 koji je finansiran od strane ministarstva za nauku i zaštitu životne sredine Republike Srbije.

Milan Vojnović, Centar za unapređenje životnih aktivnosti – Inovacioni centar, Gospodar Jovanova 35, Beograd, Srbija (e-mail: vojnovicmilan@yahoo.com).

Postoji nekoliko bitnih činjenica zbog kojih je napravljena gornja podela snimaka. Kod korišćenja komunikacionih sredstava (najčešće telefona) govornik govori "u mikrofon" koji je relativno blizu njegovih usta. Zbog toga je uticaj ambijentalne buke mali pa je odnos signal/šum uglavnom dobar. Kada se koristi telefon menja se stil govora [1]: govor je glasniji, intonacija je viša i postoji želja govornika da bude jasan u svom izražavanju. Nasuprot ovome, snimci govora dobijeni prislušnim sredstvima su uglavnom lošijeg kvaliteta, jer se mikrofon nalazi na relativno velikoj udaljenosti od govornika. Odnos signal/šum je uglavnom nezadovoljavajući zbog značajnog uticaja ambijentalne buke. Takođe ne treba zanemariiti ni uticaj akustičkih karakteristika prostorije (reverberacija, flater i dr.) u kojoj se obavlja snimanje.

Nesporni snimci govora se dobijaju postupkom intervjua. Prvo je potrebno da osumnjičeni pristane na veštačenje glasa, odnosno na pravljenje intervjua sa njim. Intervju se snima i on kasnije služi kao nesporni snimak glasa osumnjičenog. Treba primetiti da nesporni snimak nije tipa telefonskog razgovora (kao što je to sporni snimak) već je dobijen korišćenjem "prislušnih sredstava". Naravno, ne radi se o korišćenju klasičnih prislušnih sredstava već o korišćenju mikrofona i uređaja za snimanje koji nisu skriveni. Međutim, mikrofon je udaljen od usta govornika tako da je u snimku najčešće prisutan značajan nivo ambijentalne buke kao i uticaj akustičkih karakteristika prostorije. Nije redak slučaj da je nesporni snimak govora lošijeg kvaliteta od spornog. Najčešći razlozi su korišćenje neadekvatne i nekvalitetne opreme za snimanje govora (diktafoni), neadekvatna prostorija za snimanje (prostorije sa velikom reverberacijom) i sl. Postoje pokušaji da se i nesporni snimak govora napravi u obliku telefonskog razgovor.

Bez obzira na prisustvo ovih problema u postupku pravljenja nespornog snimka govora, oni nisu toliko kritični jer se intervju uvek može ponoviti, a mogu se angažovati i stručnjaci iz domena studijskog snimanja zvuka. Mnogo ozbiljniji problemi se javljaju kod spornih snimaka govora jer su oni jedinstveni i ne mogu se ponoviti, osim u izuzetno retkim situacijama. Izobličenja i smetnje koje se jave u spornom snimku govora su trajna i jedino što može da se učini je da se proceni njihov uticaj na tačnost analiziranih parametara govora. Tip smetnji i izobličenja govornog signala je uglavnom takav da se ne mogu postići značajna poboljšanja obradom signala. Obrada govornog signala može da poboljša razumljivost, ali sa aspekta identifikacije govornika nema bitnijih

poboljšanja. Zbog toga je najbolje koristiti samo elementarne postupke obrade govornog signala koji nemaju uticaja na parametre relevantne za postupak identifikacije govornika. Osim toga, primena kompleksnih procedura obrade i predobrade govornih signala daje povod za sumnju (od strane odbrane) da su rezultati analize govora "namešteni". Na samom početku postupka identifikacije govornika mora se proceniti kvalitet govornih signala u smislu da li se njima mogu dobiti validni rezultati veštačenja. To je jedan od načina da se utvrdi uticaj izobličenja govornog signala na validnost postupka identifikacije govornika: da se odredi koja vrsta izobličenja i koji stepen tih izobličenja dovodi do značajnih promena u analizi parametara govora. Drugim rečima, da se odrede granice tolerancije pojedinih vrsta izobličenja u postupku identifikacije govornika.

Izobličenja koja se javljaju u telefonskom signalu mogu se podeliti prema tipu izobličenja i prema mestu nastajanja izobličenja. Prema tipu izobličenja razlikuju se izobličenja frekvencijskih karakteristika elektro-akustičkih pretvarača, izobličenja prenosne karakteristike telefonskog kanala, klipovanje, šumne i impulsne smetnje i dr. Kada se govori o mestu nastajanja izobličenja onda se razlikuju izobličenja na mestu predaje, u prenosnom kanalu i na mestu prijema. Podrazumeva se da su izobličenja na mestu predaje i u prenosnom kanalu značajnija jer se na njih ne može bitno uticati. Izobličenja na mestu prijema se mogu dovoljno redukovati tako da se mogu zanemariti.

Jedno od čestih izobličenja u telefoniji je odsecanje vršnih vrednosti, odnosno klipovanje. Praksa pokazuje da se klipovanje skoro redovno javlja u mobilnoj telefoniji. Ovo je posledica kompromisa između dozvoljenih izobličenja i razumljivosti govora. Osnovni zahtev u telefoniji je prenos informacija. Kada su u pitanju govorne komunikacije onda je osnovni zahtev razumljivost. Ostale informacije, koje se prenose zajedno sa govorom, nisu primarne. Zbog toga se mogu dozvoliti određena izobličenja ako se njima ne narušava razumljivost komunikacije. Istraživanja pokazuju da umereno klipovanje (do 10 dB) ne utiče bitno na razumljivost. Čak i za ekstremne vrednosti klipovanja (beskonačni kliper) razumljivost govora je velika 70% [2]. Sa druge strane, klipovanjem govornog signala postiže se veći i ujednačeniji nivo govora. Povećan nivo i redukovana dinamika govora povoljno utiču na čujnost i razumljivost govora.

U situacijama kada iz govora treba "izvući" i neke druge podatke (biometrijski podaci o govorniku) svaka vrsta izobličenja, pa i klipovanje, može imati značajan uticaj na validnost analize.

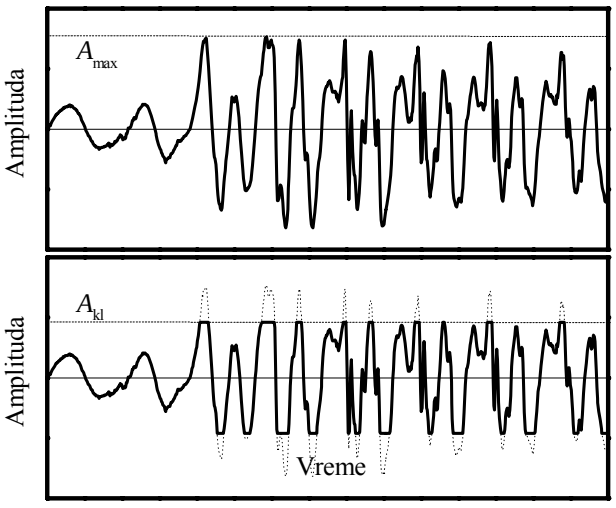
U ovom radu analizirano je kako odsecanje vršnih vrednosti (klipovanje) utiče na tačnost estimacije formantnih frekvencija vokala.

II. POSTUPAK ANALIZE

Da bi se estimirale formantne frekvencije vokala, za neki definisani govorni materijal, najpre se mora izvršiti markiranje izgovora pojedinih vokala. Najčešće je govorni materijal u standardnom Microsoft-ovom audio fajlu sa ekstenzijom WAV. Zbog toga se ovi snimci kratko nazivaju "WAV fajlovi". Krajnji rezultat procesa markiranja izgovora vokala je klasični ASCII (tekstualni) fajl u kome se nalaze podaci o početku i kraju izgovora svakog vokala pojedinačno. Par koji čini WAV i njemu odgovarajući ASCII fajl omogućuje automatizaciju procesa estimacije formantnih frekvencija vokala. U principu, ova dva fajla su povezana po vremenskoj dimenziji.

ASCII fajl sa podacima o markaciji vokala će uvek odgovarati WAV fajlu ako nisu izvršene izmene nad njim u vremenskoj dimenziji: ako WAV fajl nije produžen ili skraćen, ako nije vremenski komprimovan ili ekspandovan i sl. Klipovanje početnog govornog signala, smeštenog u WAV fajlu, neće remetiti harmoniju sa ASCII fajlom jer izobličenje nije po vremenskoj dimenziji. Dakle, početni govorni signal (ili više govornih signala) se klipuje različitim nivoima pa se za svaki ovaj slučaj estimiraju formantne frekvencije vokala. Proces estimacije formantnih frekvencija je automatizovan jer se uvek koristi isti ASCII fajl u kome su smešteni podaci o trenucima izgovora vokala.

Sl. 1 može da posluži za definisanje nivoa klipovanja. Na gornjem dijagramu je prikazan neklipovan, normalizovani govorni signal. Normalizacija podrazumeva da je signal maksimalno pojačan, ali da nije došlo do klipovanja. ovo je tzv. normalizacija signala na vršnu vrednost. Za 16-bitnu rezoluciju kvantovanja maksimalna amplituda signala normalizovanog na vršnu vrednost iznosi 32767 bita.



Sl. 1. Definicija nivoa klipovanja.

Maksimalna vrednost govornog signala iznosi A_{max} (gornji dijagram na Sl. 1). Ako se sve vrednosti signala veće od A_{kl} odseku (donji dijagram) doći će do klipovanja. Nivo klipovanja definiše se kao decibelski odnos maksimalne amplitude i amplitude na kojoj se vrši odsecanje:

$$L_{kl}[\mathrm{dB}] = 20\log\frac{A_{max}}{A_{kl}} \qquad (1)$$

Tako na primer, za nivo klipovanja od 20 dB sve vrednosti iznad desetog dela maksimalne amplitude će biti odsečene.

U radu su korišćeni snimci govora iz GEES baze [3] (Govorne Ekspresije, Emocije i Stavovi). Analizom su obuhvaćena samo dva govornika: jedan ženski i jedan muški. Iz GEES baze su uzeti dugi i kratki iskazi kao i diskurs izgovoreni neutralno (bez emocija). Dugi i kratki iskazi imaju po 30 rečenica. Ukupno trajanje govornog materijala, za jednog govornika, iznosi oko 250 s. U okviru ovako odabranog govornog materijala nalazi se oko 900 vokala.

Formantne frekvencije vokala su estimirane programom PRAAT [4]. Parametri analize su bili sledeći:

- gornja granična frekvencija analize : 5500 Hz
- broj estimiranih formanata : 5
- širina prozora analize : 25 ms
- dinamički opseg analize : 30 dB
- pretpojačanje signala (6 dB/oktavi) : iznad 50 Hz
- metod estimacije formanata : Burg

Gornja granična frekvencija od 5500 Hz je korišćena kod estimacije formantnih frekvencija ženskog govornika, a kod muškog govornika ona je iznosila 5000 Hz.

Odabrani govorni materijal je klipovan različitim nivoima: 5, 10, 15 i 20 dB. Za ova četiri slučaja estimirane su formantne frekvencije vokala i upoređene sa formantnim frekvencijama vokala neklipovanog govora. U svim slučajevima estimacije formantnih frekvencija parametri analize su bili isti.

Ovim postupkom su praktično izmerene razlike u formantnim frekvencijama vokala koje su rezultat klipovanja govornog signala. Svi ostali parametri i uslovi estimacije formantnih frekvencija nisu menjani.

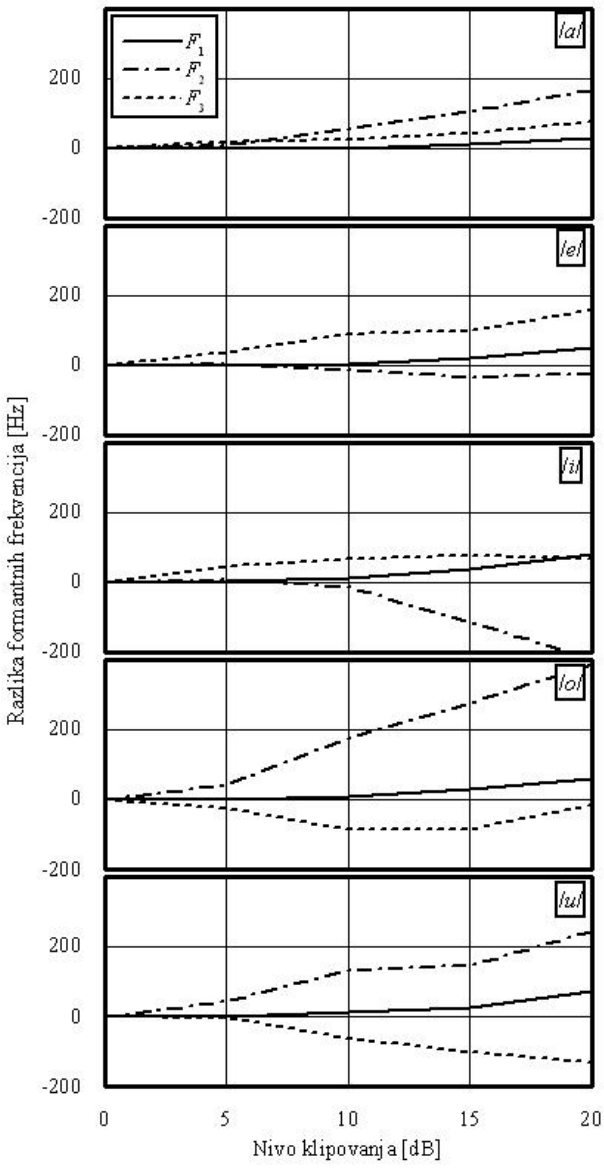
III. REZULTATI ANALIZE

Na Sl. 2 prikazane su razlike formantnih frekvencija srpskih vokala u funkciji nivoa klipovanja. Slika se odnosi na muškog govornika. Rezultati analize za ženskog govornika su vrlo slični tako da nisu posebno prikazani.

Prema Sl. 2, opseg promena formantnih frekvencija vokal je od -220 do +400 Hz kada se govor klipuje nivoom od 0 do 20 dB. Opseg promena frekvencije prvog formanta iznosi od 0 do 80 Hz, drugog formanta od -220 do 400 Hz i trećeg od -130 do 160 Hz. Drugi formant vokala je najosetljiviji na klipovanje, zatim treći i na kraju prvi formant. Najviše se menjaju frekvencije vokala /o/, zatim vokala /*u*/, /*i*/, /*a*/ i na kraju vokala /*e*/.

Prikaz uticaja nivoa klipovanja na formantne frekvencije vokala prema Sl. 2 možda nije adekvatan. Tako na primer, promene od par desetina herca kod prvog i trećeg formanta ne mogu se isto tretirati. Frekvencija prvog formanta je niska pa je promena od par desetina herca bitna. Nasuprot ovome, frekvencija trećeg formanta je visoka tako da je promena od par desetina herca zanemarljiva. Zbog svega toga, možda je adekvatnije posmatrati procentualne promene formantnih frekvencija vokala koje su prikazane na Sl. 3.

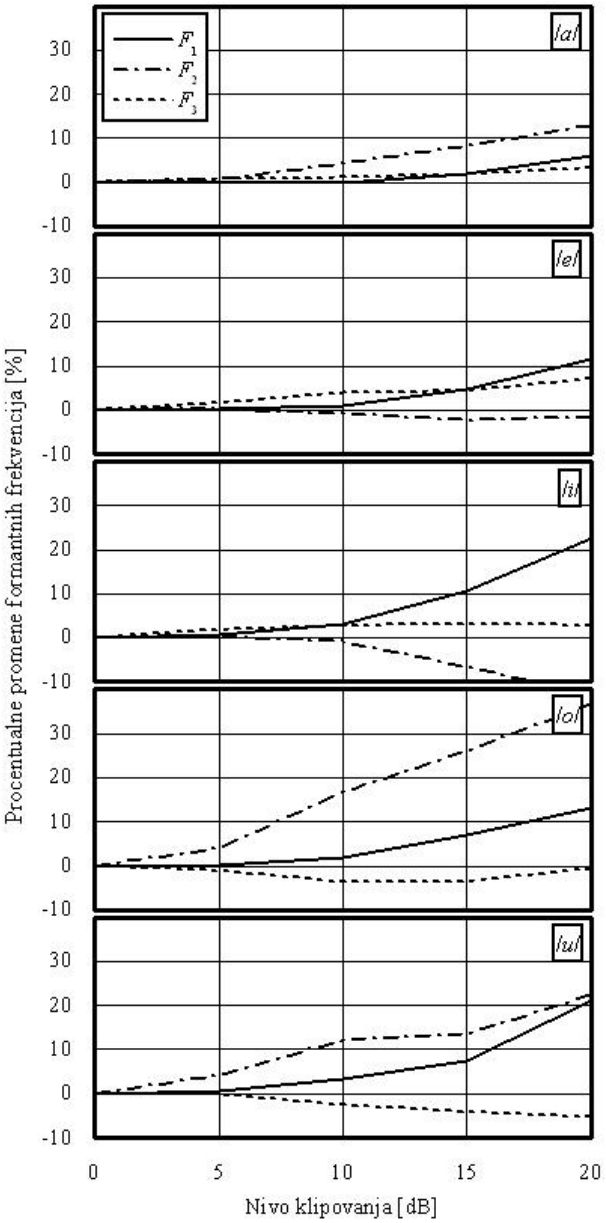
Procentualni prikaz uticaja klipovanja na formantne frekvencije daje sasvim drugačiju sliku. Ovo se pre svega odnosi na frekvenciju prvog formanta čije su promene došle do većeg izražaja. Promene formantnih frekvencija vokala su u opsegu od -13 do +37%. Opseg promena frekvencije prvog formanta je od 0 do +23%, drugog formanta od -13 do +37% i trećeg od -6 do +7%. Prema procentualnim promenama, najosetljivija je frekvencija drugog formanta, zatim prvog i na kraju trećeg formanta. Redosled osetljivosti vokala je praktično isti kao i u slučaju promena izraženih u hercima: /o/, /u/, /i/. /e/ i /a/. Prosečna procentualna promena sve tri formantne frekvencije u funkciji nivoa klipovanja iznosi: 1% za nivo klipovanja 5 dB, 4% za nivo klipovanja 10 dB, 7% za nivo klipovanja 15 dB i 12% za nivo klipovanja 20 dB. Kod ženskog govornika ove procentualne promene su nešto niže i iznose: 0,5%, 2%, 6% i 10%, respektivno.



Sl. 2. Razlike formantnih frekvencija srpskih vokala u funkciji nivoa klipovanja.

Spektrogramska analiza takođe pokazuje promene u formantnoj strukturi vokala kada je govor klipovan. Dolazi do "zamućivanja" viših formanata, vidi se pomeranje

formanata, a česta je i pojava stvaranja novih, virtuelnih formanata. Klipovanje signala ima za posledicu povećanje šuma u višem frekvencijskom području [6]. To je razlog "zamućivanja" viših formanata, jer sa povećanjem nivoa klipovanja opada odnos signal/šum u ovom frekvencijskom području.



Sl. 3. Procentualne promene formantnih frekvencija srpskih vokala u funkciji nivoa klipovanja.

Govor nije deterministički proces pa se zbog toga ne dobijaju egzaktne vrednosti formantnih frekvencija. U analizi formantnih frekvencija vrši se njihova estimacija u smislu određivanja statističke raspodele formantnih frekvencija. To znači da i kod normalnog (neklipovanog) govora dolazi do "rasturanja" formantnih frekvencija. Estimirane formantne frekvencije imaju normalnu (Gauss-ovu) raspodelu [5]. Zbog toga je razumno usvojiti granice tolerancije promena formantnih frekvencija. Razumno je usvojiti toleranciju od ±5%. Sa ovako usvojenim kriterijumom može se zaključiti da umereno klipovanje (do 5 dB) nema veliki uticaj na formantne frekvencije vokala jer se oni menjaju za manje od 4%. Klipovanje od 10 dB izaziva veće promene formantnih frekvencija kod vokala /*o*/ i /*u*/. Kod preostala tri vokala čak i sa ovim nivoom klipovanja procentualne promene nisu velike i nalaze se ispod 5%.

IV. ZAKLJUČAK

Klipovanje govornog signala ima uticaja na formantne frekvencije vokala. Veći je uticaj na više formante, ali se ne sme zanemariti da prvi formant može da bude vrlo nisko tako da male promene u hercima uzrokuju velike procentualne promene.

Zadnji vokali (/*o*/ i /*u*/) su osetljiviji na klipovanje nego preostala tri. Kod ova dva vokala najveće su promene kod drugog formanta, a kod prednjih i srednjeg vokala (/*i*/, /*e*/ i /*a*/) frekvencija prvog formanta.

Kod umerenog klipovanja govornog signala (manje od 5 dB) procentualne promene formantnih frekvencija vokala su manje od 4%. Ovo nisu velike promene tako da se mogu tolerisati u postupku forenzičke identifikacije govornika.

LITERATURA

[1] Catherine Byrne, and Paul Foulkes, "The 'Mobile Phone Effect' on Vowel Formants", *Speech, Language and the Law*, 11(1), pp.83-102, 2004.

[2] Lukatela G., Drajić D., Petrović G., *Digitalne telekomunikacije*, Građevinska knjiga, Beograd, 1978.

[3] Jovičić S., Kašić Z., Đorđević M., Vojnović M., Rajković M., "Formiranje korpusa govorne ekspresije emocija i stavova u srpskom jeziku - GEES", *XI Telekomunikacioni forum TELFOR*, Sekcija 7.10, 2003, Beograd.

[4] Boersma P., Weenink D., "PRAAT: A system for doing phonetics by computer", 1992-2005, http://www.praat.org/.

[5] Vojnović M., "Neki praktični problemi estimacije formantnih frekvencija vokala", *Šesta konferencija: DIGITALNA OBRADA GOVORA I SLIKE - DOGS2006*, Zbornik radova, str. 18-21, 2006, Vršac.

[6] Vojnović M., "Prepoznavanje govornika pomoću dugovremenog usrednjenog spektra", *Nauka Tehnika Bezbednost*, br. 2, str. 53-66, 2005, Beograd.

ABSTRACT

The quality of speech recordings used in forensic speaker identification is usually poor. The most often telephone speech conservations are used which are characterized narrow frequency band (up to 4 kHz) and numbered different disturbances and distortions. Cutting the peaks of signal (clipping) is very usual in telephony. In this paper the influence of the clipping on accurate of vowel formant frequency estimation. The results show that only moderate clipping (up to 5 dB) can be tolerated in the forensic speaker identification procedure.

THE INFLUENCE OF THE CLIPPING ON THE VOWEL FORMANT FREQUENCIES

Milan Vojnović