Ahots iruzurtiak detektatzen: Spoofing Challenge 2015

Jon Sanchez[1], Ibon Saratxaga[1], Inma Hernaez[1], Eva Navas[1], Daniel Erro[1,2]

[1]Aholab, UPV/EHU - [2]Ikerbasque
{ion,ibon,inma,eva,derro}@aholab.ehu.eus

Abstract

This paper introduces the Synthetic Speech Detection system developed by Aholab for the Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015). The detector is a classifier based on Gaussian Mixture Models that are created using the Relative Phase Shift (RPS) transformation for the phase information. Different strategies have been evaluated: modeling the specific attacks using the information provided by the ASVspoof 2015 organizers, and modeling the vocoders possibly used in the spoofing signals, using data from previous works. The evaluation results show that attack specific models work for known attacks but they do not cope with the unknown attacks correctly. When using vocoder models build with other databases, the results suggest that the followed strategy do not take advantage of the available data and thus model adaptation should be explored.

Laburpena

Artikulu honetan Aholabek garatutako Ahots Sintetikoaren Detektorea (Synthetic Speech Detector, SSD) deskribatzen da, eta bere erabilera Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015) nazioarteko norgehiagokan. GMMtan oinarritutako detektorea da, non ereduak Relative Phase Shift (RPS) fasearen informaziorako transformazioa erabiliz osatzen diren. Estrategia desberdinak ebaluatzen dira: eraso espezifikoen ereduak sortu norgehiagokaren antolatzaileek emandako informazioaren bidez, edo mehatxu-seinaleak sortzeko erabili omen den bokoderraren ereduak sortu, aurreko lanetan garatutako informazioa erabiliz. Ebaluazioaren emaitzetan ikusten denez, eraso espezifikoen ereduak ondo aritzen dira eraso ezagunekin, baina ez ezezagunekin. Bokoderren ereduak erabiltzen direnean, informazioa ez da guztiz erabiltzen eta ereduen moldaketa landu behar da.

Keywords: synthetic speech detection, phase information, spoofing

Gako hitzak: Ahots sintetikoaren detekzioa, fasearen informazioa, spoofing

1. Sarrera

Gaur egun, ezinbestekoa da aplikazio batzuetarako leku edo informazio batzuk erabiltzeko baimena nork daukan kudeatzea. Azken urteotan, tendentzia bat indartzen ari da: txartelak, giltzak edo klabeak erabili beharrean, ezaugarri biometrikoak erabiltzea aldarrikatzen duena (Jain et al., 2006). Ezaugarri biometrikoen abantailarik handiena, ahazteko edo osteko ezintasunean datza.Eta ezaugarri biometrikoen artean erabilgarrienetariko bat ahotsa da: identifikazioa gauzatzeko beste informazio dauka, eta erabiltzailearentzako erosoa den eran lor daiteke.

Hizlariaren egiaztaketa sistemek (*Speaker Verification*, SV*)* ahotsa erabiltzen dute bektore biometriko moduan (Furui, 1981)(Reynolds et al., 2000)(Campbell, 1997)(Kinnunen y Li, 2010), eta posible da ezaugarri hori antzeratzea sistema iruzurtzeko, *spoofing* izeneko teknikak erabiliz.

Gaur egungo ahots bihurketa eta sintesi teknikek hizlariaren egiaztaketa egiten duen sistema bat iruzurtzeko beste kalitate lortu dute, eta SV sistemen segurtasunarekiko kezka handiagotzen ari da (Evans et al., 2013)(Wu et al., 2015).

Sistema biometrikotarako mehatxuak diren itxurazko ahotsak detektatzeko bi estrategia desberdin garatu dira: lehenengoan, SV sisteman bertan inplementatu behar dena, erabiltzaileen ereduak landu behar dira, blokeatu ahal izateko bai giza-iruzurtiak eta baita sintetikoak ere (Masuko et al., 2000)(Kons y Aronowitz, 2013)(Kinnunen et al., 2012). Bigarrenean, aparteko Ahots Sintetikoaren Detektagailu bat (Synthetic Speech Detector, SSD) garatzen da, SV sistema baino lehen edo ondoren erabiltzeko. SSD moduluan parametro eta detekzio teknika espezifikoak erabiltzen dira, ahots naturala eta sintetikoaren artean dauden aldeetan jarrita arreta: pitch-aren aldaketak (Steward et al., 2012), tramen arteko parametro batzuen antzekotasunak (De Leon et al., 2012)(Alegre et al., 2013), fase informazioa (De Leon et al., 2012)(Wu et al., 2012), denbora modulazioa (Wu et al., 2013), etab.

Bigarren hurbilketa hau arrakastaz erabili da bokoderdun ahotsa detektatzeko (Sanchez et al., 2015) lanean. Bokoderrak egungo ahots-sintesi eta ahots-bihurketa sistema gehienetan erabiltzen dira, eta horiek dira hain zuzen SV sistema bat iruzurtzeko ahotsak sortzeko erabiltzen direnak. Beraz, bokoderdun seinaleen detekzioa ahots iruzurtien aurkako neurri eraginkorra izan daiteke.

Moduluan oinarritutako parametroak (MFCC adibidez) askotan erabiltzen dira SSD sistemak garatzeko, baina gure sistema, (Sanchez et al., 2015) lanean deskribatutakoa, fasearen informazioa besterik ez du erabiltzen detekzioa egiteko: RPS parametroak

hain zuzen. Bokoderrik ezagunenetan ez da fasearen informazioa erabiltzen. Seinale natural eta faltsifikatuen faseen arteko aldea nabarmena dela frogatu da. Hurbilketa honen gaitasuna ahots sintetikoen erasoei aurre egiteko(Sanchez et al., 2015) lanean adierazi da. Sistemaren eraginkortasuna ona ze, seinale iruzurtiak TTS sistema ezezagunekin sortzen zirenean ere.

Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015) norgehiagokan, SSD modulu independenteak ebaluatzen dira (Wu et al., 2014). Parte-hartzaileek iruzurtien detekzio-lanabesak aplikatu behar dute, emandako datu-base baten. Horren barruan ahots iruzurtien teknika desberdinen adibideak daude, ahots sintesia edo ahots bihurketa adibidez.Sistema desberdinen eraginkortasuna norgehiagokaren antolatzaileek baieztatzen dute, metrika estandarrak erabiliz. Gehienez, erabiltzaile bakoitzeko 6 sistema desberdinen emaitzak bidal daitezke, bi mota desberdinetakoak: 'common' izenekoak garatzeko emandako datu-basearen entrenamendu-zatia besterik ezin daiteke erabili; 'flexible' izenekoetan, berriz, edozein datu-base erabil daiteke. Mota bakoitzerako, bidalketariko bat nagusia izango da, eta besteak egiaztatze moduan erabiltzen dira. Guztira, gure sistemaren 4 bertsio desberdin bidali dira, eredu desberdinak erabiliz.

Artikuluaren ondorengo atalean sistemaren deskribapena gauzatzen da, antolatzaileek emandako datu-baseari eredu desberdinak sortzeko egindako prozesaketarekin batera. Horren ostean, eredu desberdinetako emaitzak komentatzen dira. Bukatzeko, ondorioak laburbiltzen dira.

2. Sistemaren deskribapena

2.1. Arkitektura orokorra

Asvspoof2015 norgehiagokarako (Sanchez et al., 2015) lanean aurkeztutako sistema erabili da. Sailkagailu bitarra da, GM ereduetan oinarrituta. Bere helburua sarrerako seinale bat sintetikoa den erabakitzea da. 1. Irudian sistemaren arkitektura orokorra deskribatzen da.

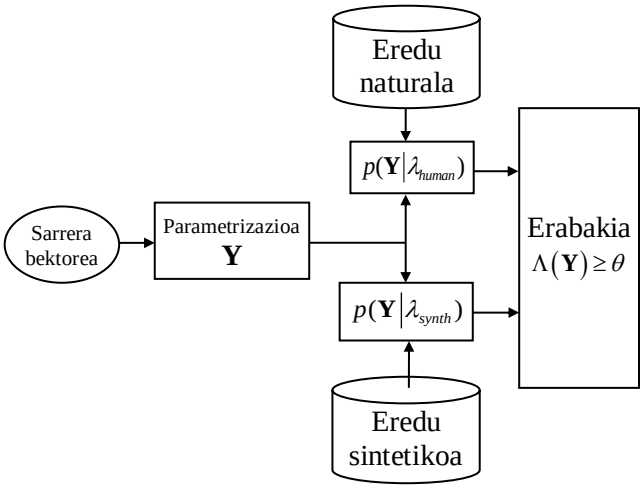
SSD sistemak bi GM eredu desberdin erabiltzen ditu: ahots naturalari dagokiona (λ_{human}) eta ahots sintetikoarena (λ_{synth}). Ereduak entrenamendu fasean sortzen dira, fase harmonikoan oinarritutako bektoreen bidez. Horiek sortzeko RPS transformazioa aplikatu da fase harmonikoetan.

Sintetikotasunaren detekzioa gauzatzeko, sistemak sarrerako **Y** bektore-sekuentzia bat hartzen du, N luzerakoa, eta eredu naturala eta sintetikoekin frogatzen da, bakoitzaren egiantza balioak $p(\mathbf{Y}|\lambda_{human})$ eta $p(\mathbf{Y}|\lambda_{synth})$ kalkulatzeko.

$$\log p(\mathbf{Y}|\lambda) = \frac{1}{N}\sum_{n=1}^{N} \log p(\mathbf{y}_n|\lambda) \qquad (0)$$

$$\Lambda(\mathbf{Y}) = \log p(\mathbf{Y}|\lambda_{human}) - \log p(\mathbf{Y}|\lambda_{synth}) \qquad (0)$$

Ondoren, (0) ekuazioaren bidez, Λ egiantza tasa kalkulatzen da, sarrera gizakiak sortuta bezala hartuz θ emandako atari bat pasatzen bada. Ataria, esperimentuetarako, *Equal Error Rate* (EER) puntuan jartzen da. ASVspoof2015 norgehiagokarako sarrera seinale bakoitzeko Λ egiantza tasa bidaltzen da.



1. Irudia: SSD sistemaren egitura

2.2. Seinalearen aurre-prozesaketa

Parametrizazioa burutu baino lehenago, seinaleen laginketa maiztasuna 8kHz-etara murrizten da konputazio-karga mugatzeko, eta korronte zuzena ezabatzen da. Horretaz gain, seinaleen polaritatea bateratzen da (Saratxaga et al., 2009a) RPS parametrizazioa asko aldatzen delako polaritate aldaketekin.

2.3. RPS parametrizazioa

Ahozko seinaleak RPS transformazioaren bidez parametrizatzen dira. RPS ahotsaren fase harmonikoaren irudikapen bat da, (Saratxaga et al., 2009b) lanean deskribatzen dena. Analisi harmonikoaren bidez seinalearen segmentuak modelatzen dira sinusoideen batuketa moduan, sinusoide bakoitza oinarrizko maiztasunarekin erlazio harmonikoa izanik.

$$h(t) = \sum_{k=1}^{N} A_k \cos\left(\varphi_k(t)\right) \qquad \varphi_k(t) = 2\pi k f_o t + \theta_k \qquad (0)$$

(3) ekuazioan, N banda kopurua da, A_k anplitudeak dira, $\varphi_k(t)$ uneko faseak, f_0 oinarrizko maiztasuna eta θ_k k-garren harmonikoaren fasearen desplazamendua. RPS irudikapena oinarrizko maiztasuneko osagarria (k=1) eta gainontzeko sinusoide guztien arteko fase aldea kalkulatuz osatzen da, oinarrizko periodoaren puntu zehatz baten, adibidez φ_o=0 betetzen duen puntua.

$$\psi_k(t_a) = \varphi_k(t_o) = \varphi_k(t_a) - k\varphi_1(t_a) \quad (0)$$

(4) ekuazioak RPS transformazioa definitzen du. Horren bidez, RPS balioak(ψ_k) kalkulatzen dira uneko faseetatik, seinalearen edozein puntuan (t_a). RPS balioak biribilkatzen dira [-π, π] tartera.

Horrela kalkulatutako RPS balioak, berriz, ezin dira erabili eredu estatistikoak garatzeko. Beraz, ereduak sortu eta frogatzeko DCT-mel-RPS izendatutako parametrizazioa erabiltzen da aurrekoaren ordez. Parametro hauek sakonean azaldu dira (Saratxaga et al., 2010) lanean,eta emaitza onekin erabili dira eredu estatistikoak behar duten lanetan: ahotsaren ezagutzan, (Saratxaga et al., 2010), hizlariaren ezagutzan (Hernáez et al., 2011), edo ahots sintetikoaren detekzioan (Sanchez et al., 2015).

Parametroak lortzeko, RPS balioen diferentziak mel iragazkien banku batetik pasatzen dira (48 iragazkikoa), eta kosinuaren transformatu diskretu bat (DCT) aplikatzen da irteera sekuentzian. DCTa 20 baliotan mugatzen da eta Δ eta ΔΔ balioak kalkulatzen dira.

Esperimentutarako, ahozko seinaleak 10ms-ro leihokatzen dira (3 pitch periodotako hamming leihoak erabiliz) eta RPSak ahostun trametan besterik ez dira kalkulatzen. Ondoren, DCT-mel-RPS parametrizazioa aplikatzen da eta batezbesteko RPS parametroen malda sartzen da ere. Horrela, 63 parametro erabiliko dira, fasetik kalkulatuak, trama bakoitzeko.

2.4. Eredutzea

Norgehiagokarako 4 eredu multzo desberdin erabili dira, aurreko lanetako datuak eta antolaketak emandako datu-baseekin. Multzo bakoitzean ahots naturaleko eredu bat (λ_{human}) eta ahots iruzurtirako beste bat (λ_{synth}) egongo dira.

Lehenengo bi multzoak (M1 eta M2 izenekin) antolaketak emandako entrenamendu eta garapen seinaleak erabiliz entrenatu dira, bai seinale naturalak gizaki eredua sortzeko eta baita seinale sintetikoak eredu sintetikoa sortzeko.

Lehenengo bidalketako eredu naturala eta sintetikoa garatzeko antolaketak emandako datu-basearen entrenamendu zatia besterik ez da erabili. Seinale kopurua 1. taulan ikus daiteke. Iruzur metodoak ondorengoak dira:

- Ahots bihurketa, bi inplementazio desberdinetan, STRAIGHT (Zen et al., 2007) erabiliz.
- Ahots bihurketa MLSA (Yoshimura et al., 1999) erabiliz.
- Ahots moldatuen sintesiaren bi inplementazio desberdin, STRAIGHT erabiliz.

Multzoko ereduak 1024 gaussiarekin entrenatu dira eta ‘primary common’ motako bidalketa egiteko baldintzak betetzen ditu.

M2 izeneko eredu multzoa sortzeko hurbilketa M1enaren antzekoa izan da, baina erabili daitekeen informazioa sartuz ereduetan: bai entrenamendu-zatia eta baita garapen-zatia ere, 7247 seinale natural eta 62500 seinale iruzrtiekin. Metodoak, M1en erabilitako berak dira. Ereduak 1024 gaussiarekin entrenatu dira eta ‘primary flexible’ moduan bidali dira.

M3 eredu multzoan (Sanchez et al., 2015) lanean erabilitako bokoder anitzeko eredua erabili da. M1 eta M2en kasuen desberdina da, ahots seinaletako bokoderra detektatzeko modelatu delako, eta ez eraso teknika edo algoritmo espezifikoak, benetako egoera baten ezezagunak izango direnak. Bokoder anitzeko eredua WSJ datu-basea (Paul y Baker, 1992) erabiliz sortu zen. Gizaki-eredua sortzeko WSJ datu-baseko 8599 seinale erabili ziren, 283 hizlaritakoak. Seinale sintetikoak sortzeko copy-synthesis teknika erabili zen, hiru bokoder desberdin erabiliz: MLSA, STRAIGHT eta AHOCODER (Erro et al., 2014) (Erro et al., 2011). Horrela 25797 seinale sortu ziren iruzurtien eredua sortzeko. M3 eredua entrenatzeko ez zen ASVSpoof2015 norgehiagokaren antolakuntzak emandako seinalerik erabili. Ereduak 512 gaussiarekin entrenatu ziren, eta ‘flexible contrastive’ bidalketarako erabili zen.

Azkeneko eredua, M4 zenekoa, aurretik deskribatutako estrategia desberdin biak bateratzen ditu: emandako seinaleak erabiltzen dira eraso ezagunak detektatzeko, eta bokoder anitzeko hurbilketa eraso ezezagunak detektatzeko. Ondorioz, ereduak sortzeko emandako entrenamendu-zatia erabili da M3ko WSJ datu-basetik ekarritako seinaleekin batera: guztira 12349 seinale natural eta 38422 iruzurti. 1024 gaussiarrekin entrenatu da eredua eta ‘flexible contrastive’ bezala bidali da.

	M1	M2	M3	M4
Naturala	3750	7247	8599	12349
VC STRAIGHT	5050	12500	-	5050
VC MLSA	2525	12500	-	2525
SS STRAIGHT	5050	12500	-	5050
CS STRAIGHT	-	-	8599	8599
CS MLSA	-	-	8599	8599
CS AHOCODER	-	-	8599	8599
Iruzurtiak guztira	12625	62500	25797	38422

1. taula: Eredu bakoitza entrenatzeko erabilitako seinale kopurua, bokoder era eraso metodoaren arabera sailkatuta.

3. Emaitzak

Ebaluatzeko ematen den seinale-sortan 9404 seinale natural eta 184000 iruzurti daude, 2. taulan ikusten den moduan. Seinale iruzurtien erdia metodo ezagunak erabiliz garatu dira, hau da, entrenamendu era garapen zatietako berak (ahots bihurketa STRAIGHT eta MLSA erabiliz, sintesi moldatua STRAIGHT erabiliz) . Beste erdia, berriz, eraso teknika ezezagunak erabiliz lortu ziren: STRAIGHT-en oinarritutako lau ahots-bihurketa algoritmo berri eta ahots sintetizadore berri bat, MaryTTS (DFKI, 2002), bokoderrik erabiltzen ez duena.

Zatia	Ezaugarriak	Ezaguna?	Seinale kopurua
N	Naturala	Bai	9404
S1	VC/STRAIGHT	Bai	18400
S2	VC/STRAIGHT	Bai	18400
S3	SS/STRAIGHT	Bai	18400
S4	SS/STRAIGHT	Bai	18400
S5	VC/MLSA	Bai	18400
S6	VC/STRAIGHT	Ez	18400
S7	VC/STRAIGHT	Ez	18400
S8	VC/STRAIGHT	Ez	18400
S9	VC/STRAIGHT	Ez	18400
S10	Bokoder gabekoa	Ez	18400

2. taula: Ereduak frogatzeko erabilitako seinale kopurua

Deskribatutako froaga seinaleekin Equal Error Rate (EER) balioa kalkulatuz ondorengo emaitzak lortu ziren:

Eredu sorta	Eraso ezagunak	Erazo ezezagunak	Eraso guztiak
M1 (Common Primary)	0.210	8.883	4.547
M2 (Flexible primary)	0.154	8.918	4.536
M3 (Flexible Contrastive 1)	9.845	17.371	13.608
M4 (Flexible Contrastive 2)	2.042	11.291	6.667

3. taula: Aholab-ek bidalitako 4 sistemen EER balioa, portzentajean.

Hirugarren taulako zenbakien arabera, M1 eta M2 eredu sortekin nahiko emaitza onak lortu dira, EER balioak %0.25en azpitik daudelarik seinaleak sortzeko metodoak ereduetan daudenean. Baina eraso ezezagunekin errorea %9ko ingurura igotzen da. Faktore desberdinekin azaltzen da efektu hori:

- Eraso mota desberdinen RPS parametroak oso desberdinak izan daitezke. Beraz, eraso espezifikoen ereduak erabiltzen direnean ezezagunen detekzioa ezin da bermatu. Hain zuzen, MaryTTS sistemako seinaleekin lortutako emaitza txarrekin eraginkortasuna lastatu egiten da sistemaren emaitza orokorra. MaryTTS sistema unitate hautaketa teknikan oinarritzen da eta ez du bokoderrik erabiltzen, beraz, entrenamendu materialen esparrutik kanpo zegoen bere sistema-mota.
- RPS parametrizazioak behar duen seinaleen analisi harmonikoak gutxieneko kalitatea maila behar du. Eraso metodo batzuek seinalearen kalitatea murrizten dute, RPS parametrizazioa ezegokia izatearen punturaino.
- Erabilitako GMM sailkatzaile sinpleak ez dauka eraso ezezagunak detektatzeko gainontzeko mekanismorik.

M3 eredu sortaren kasuak aparteko analisia behar du. Sortzeko erabili zen datu-basea zeharo desberdina zen, eta ez zegoen entrenamendu- edo garapen-zatiko seinalerik. Beraz, frogetako seinale guztiak eraso ezezagunetakoak etortzen ziren, “ezagunak” moduan sailkatzen direnak barne. Sistemaren etekina txara da eredu sorta honekin, %10eko inguruko EER balioekin. Interesgarria da honi buruzko puntu batzuk azpimarratzea:

- Ereduek (Sanchez et al., 2015) lanean lortutako orokortzeko gaitasuna ez da agertu esperimentu honetan.
- Behin-behineko esperimentu batzuetan ASVSpoof2015 datu-baseko seinale naturalen puntuazioak, M3 ereduarekin frogatzen direnean, espero baino txikiagoak zirela topatu da. Honen analisiak datu-base bietako seinale naturalen artean alde handia dagoela pentsatzera darama. Aldea grabaketa egoera desberdinetatik etor daiteke, eta fasearen egituran eragina handia izan, eta analisia sakonagoa behar da.
- Ebaluaketarako erabiltzen diren eraso gehienak ahots bihurketa erabiliz sortu dira. Sistemaren gaitasuna ahots sintetikoa detektatzeko (Sanchez et al., 2015) lanean egiaztatu da, baina ahots-bihurketaren bidez sortutako seinaleekin ez zen aurretik ebaluatu, eta analisi sakonagoa behar du ere.
- Berriro ere, MaryTTS sistemaren presentzia garrantzitsua da , sistemak ezin baitu bokoderrik gabe sortutako ahots iruzurtirik detektatu.
- Azken urteotan MLSA eta STRAIGHT bokoderren bertsio desberdinak garatu dira. Praktikan bertsioen arteko aldaketak handiak izan daitezke eta prtaera bokoder

desberdinetakoen antzekoa da, errorea handituz.

M4 eredu sortarekin lortutako emaitzak bat datoz eredu bateratuarekin espero daitekeenarekin: EER balioa aurreko bien artean dago.

Taldea	Eraso ezagunak	Eraso Ezezagunak	Eraso guztiak
A	0.408	2.013	1.211
B	0.008	3.922	1.965
C	0.058	4.998	2.528
D	0.003	5.231	2.617
E	0.041	5.347	2.694
F	0.358	6.078	3.218
G	0.405	6.247	3.326
H	0.67	6.041	3.355
I	0.005	7.447	3.726
J	0.025	8.168	4.097
K	0.21	8.883	4.547
L	0.412	13.026	6.719
M	8.528	20.253	14.391
N	7.874	21.262	14.568
O	17.723	19.929	18.826
P	21.206	21.831	21.518

4. taula: ASVSpoof2005 norgehiagokan parte hartu duten talde guztien emaitzak..

Parte-hartzaile guztien emaitzak laburbiltzen dira 4. taulan. Aholabek bidalitakoa 'K' izendatutakoa da. Zerrenda ordenatzeko irispidea "eraso guztiak" zutabea izan bada ere, benetako erronka eraso ezezagunetan egon da, bere zailtasun handiagoa dela eta. Aholaben kasuan, horien lasta oso handia izan da eta ikerkata sakonagoa behar da orokortzeko gaitasun osoa duen sistema burutzeko, ahots bihurketarekin sortutako seinaleen kasurako gehien bat.

4. Ondorioak

Aholabek ASVSpoof2015 norgehiagokarako prestatutako bidalketa sailkatzaile bitar batean oinarritzen da. Bere funtzionamendua GM eredu desberdinetan datza, gizaki- eta iruzurti-ahotzetarako. DCT-mel-RPS parametroekin osatuak. Bi estrategia desberdin erabili dira: eraso espezifikoak eredutzea antolaketak emandako seinaleen bidez, eta bokoderrak eredutzea aurreko lanetako informazioa erabiliz.

Eraso espezifikoen ereduak erabiltzen direnean etorkizun handiko emaitzak lortzen dira proposatutako arkitekturarekin. Eraso ezezagunen emaitzak MaryTTSrekin lastatuak daude, eta antza ezagunak baino txarragoak dira. Hau azaltzeko GMM sailkatzailera eta RPS parametrizazioak eraso espezifikoa eredutzera jo behar dugu.

Sistemaren eraginkortasuna apala da bokoderra detektatzeko garatu diren ereduak erabiltzen direnean. Detekzioa txarragoa da gizaki-seinaleen kasuan, ziur asko datu-base desberdinak grabatzerakoan fasearen trataera ere desberdina izan delako. Hau konpontzeko eredu moldaketa landu behar da.

5. Esker onean

Lan hau, zati baten baino ez bada ere, Eusko Jaurlaritzaren sostengua jaso du (ELKAROLA proiektua, KK-2015/0098).

6. Aipamenak

Alegre, F., Amehraye, A., Evans, N., 2013. Spoofing countermeasures to protect automatic speaker verification from voice conversion, en: ICASSP. pp. 3068-3072.

Campbell, J.P., 1997. Speaker recognition: a tutorial. Proc. IEEE 85, 1437-1462. doi:10.1109/5.628714

De Leon, P.L., Pucher, M., Yamagishi, J., Hernáez, I., Saratxaga, I., 2012. Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech. IEEE Trans. Audio. Speech. Lang. Processing 20, 2280-2290. doi:10.1109/TASL.2012.2201472

DFKI, 2002. MaryTTS – Introduction [WWW Document]. URL http://mary.dfki.de/ (accedido 3.9.15).

Erro, D., Sainz, I., Navas, E., Hernáez, I., 2014. Harmonics Plus Noise Model Based Vocoder for Statistical Parametric Speech Synthesis. IEEE J. Sel. Top. Signal Process. 8, 184-194. doi:10.1109/JSTSP.2013.2283471

Erro, D., Sainz, I., Navas, E., Hernáez, I., 2011. Improved HNM-Based Vocoder for Statistical Synthesizers., en: Interspeech. Florence, Italy, pp. 1809 - 1812.

Evans, N.W.D., Kinnunen, T., Yamagishi, J., 2013. Spoofing and countermeasures for automatic speaker verification, en: INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association. London.

Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. 29, 254-272. doi:10.1109/TASSP.1981.1163530

Hernáez, I., Saratxaga, I., Sanchez, J., Navas, E., Luengo, I., 2011. Use of The Harmonic Phase in Speaker Recognition, en: INTERSPEECH 2011, 12 th Annual Conference of the International Speech Communication Association. Florence, Italy, pp. 2757-2760.

Jain, A.K., Ross, A., Pankanti, S., 2006. Biometrics: A Tool for Information Security. IEEE Trans. Inf. Forensics Secur. 1, 125-143. doi:10.1109/TIFS.2006.873653

Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: From features to supervectors. Speech Commun. 52, 12-40. doi:10.1016/j.specom.2009.08.009

Kinnunen, T., Wu, Z., Lee, K.A., Sedlak, F., Chng, E.S., Li, H., 2012. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech, en: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4401-4404. doi:10.1109/ICASSP.2012.6288895

Kons, Z., Aronowitz, H., 2013. Voice Transformation-Based Spoofing of Text-Dependent Speaker Verification Systems, en: INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association. pp. 945-949.

Masuko, T., Tokuda, K., Kobayashi, T., 2000. Imposture Using Synthetic Speech Against Speaker Verification Based On Spectrum And Pitch, en: ICSLP. pp. 302-305.

Paul, D.B., Baker, J.M., 1992. The design for the wall street journal-based CSR corpus, en: Proceedings of the workshop on Speech and Natural Language - HLT '91. Association for Computational Linguistics, Morristown, NJ, USA, p. 357. doi:10.3115/1075527.1075614

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker Verification Using Adapted Gaussian Mixture Models. Digit. Signal Process. 10, 19-41. doi:10.1006/dspr.1999.0361

Sanchez, J., Saratxaga, I., Hernáez, I., Navas, E., Erro, D., Raitio, T., 2015. Toward a Universal Synthetic Speech Spoofing Detection using Phase Information. IEEE Trans. Inf. Forensics Secur. PP, 1-1. doi:10.1109/TIFS.2015.2398812

Saratxaga, I., Erro, D., Hernáez, I., Sainz, I., Navas, E., 2009a. Use of harmonic phase information for polarity detection in speech signals., en: Interspeech. pp. 1075 - 1078.

Saratxaga, I., Hernáez, I., Erro, D., Navas, E., Sanchez, J., 2009b. Simple representation of signal phase for harmonic speech models. Electron. Lett. 45, 381. doi:10.1049/el.2009.3328

Saratxaga, I., Hernáez, I., Odriozola, I., Navas, E., Luengo, I., Erro, D., 2010. Using harmonic phase information to improve ASR rate., en: Proc. Interspeech 2010. Makuhari, Japan, pp. 1185 - 1188.

Steward, B., De Leon, P.L., Yamagishi, J., 2012. Synthetic speech discrimination using pitch pattern statistics derived from image analysis, en: Interspeech. pp. 370-373.

Wu, Z., Chng, E.S., Li, H., 2012. Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition. Interspeech 2-5.

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., Li, H., 2015. Spoofing and countermeasures for speaker verification: A survey. Speech Commun. 66, 130-153. doi:10.1016/j.specom.2014.10.005

Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., 2014. ASVspoof 2015 : Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan [WWW Document]. URL http://www.spoofingchallenge.org/asvSpoof.pdf

Wu, Z., Xiao, X., Chng, E.S., Li, H., 2013. Synthetic speech detection using temporal modulation feature, en: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 7234-7238. doi:10.1109/ICASSP.2013.6639067

Yoshimura, T., Tokuda, K., Kobayashi, T., Masuko, T., Kitamura, T., 1999. Simultaneous Modeling Of Spectrum, Pitch And Duration In HMM-Based Speech Synthesis, en: Eurospeech. pp. 2347-2350.

Zen, H., Toda, T., Nakamura, N., Tokuda, K., 2007. Details of the Nitech HMM-Based Speech Synthesis System for the Blizzard Challenge 2005. IEICE Trans. Inf. Syst. E90-D, 325-333. doi:10.1093/ietisy/e90-1.1.325