Mezuak egiaztatzeko tresna baten diseinua, bigarren hizkuntzak ikasteko sistemetan inplementatzeko

Igor Odriozola, Inma Hernaez, Eva Navas, Jon Sanchez

Aholab Seinale Prozesaketako Laborategia, UPV/EHU
{igor,inma,eva,ion}@aholab.ehu.es

Abstract

In this article a Computer-Assisted Language Learning system (CALL) development is presented. It is useful on the learning of second languages. The aim of the system is to check in real-time that the words pronounced by the student are right, presenting them in a screen as soon as they are spoken. First of all, it is necessary to fix a threshold for the confidence parameter (PS, Phoneme Score), artificially simulating the student's errors. Next, the ASR system decodification net which has a phoneme loop between words is described, to cancel the impact of unexpected speech. Finally, the experiment designed to test the whole system is described, simulating potential errors of the user. The experiment results show how the system appears to be valid in the message verification task.

Laburpena

Artikulu honetan, mezuak egiaztatzeko sistema baten garapena aurkezten da, CALL (CALL, Computer-Assisted Language Learning) edo bigarren hizkuntzak ikasteko sistemetan inplementatzeko balio duena. Sistemaren helburua da erabiltzaileak esandako hitzak denbora errealean egiaztatzea, hitzez hitz, atzeman bezain laster pantailan agerraraziz. Lehendabizi, konfiantza-parametroaren (PS, Phoneme Score) erabaki-atalasea finkatu da, erabiltzaileen ebakera-akatsak artifizialki simulatuz. Ondoren, ASR sistemarako proposatutako deskodetze-sarea azaldu da, hitz artean fonema-begizta bat duena espero ez den hizketaren eragina xurgatzeko. Azkenik, sistema osoaren funtzionamendua aztertzeko baliatu den esperimentua azaldu da, erabiltzailearen balizko akatsak artifizialki simulatzen dituena. Esperimentuen emaitzek erakusten dute sistema egokia dela mezuak egiaztatzeko atazetarako.

Keywords: CALL systems, Utterance Verification, Message Verification, PS scores, L2 acquisition

Gako hitzak: CALL sistemak, hizketa-egiaztapena, mezu-egiaztapena, PS parametroak, L2ren jabetzea

1. Sarrera

Hizkuntzak ikasteko hizketa-teknologiak erabiltzeak 1970eko hamarkadaren amaieran ditu hastapenak. Orduan eman zitzaion CALL (*Computer-Assisted Language Learning*) izena ere. CALL eremuak ibilbide luzea badu ere, azken hamarkadan izan du garapenik handiena, hizketa-teknologiek izan duten hobekuntzaren eta nolabaiteko egonkortasunaren eraginez. CALL eremuaren baitan, askotariko aplikazioak daude, eta gaur egun, *smartphone* eta taulen hedapenaren ondorioz, MALL (*Mobile-Assisted Language Learning*) kontzeptua ere sortu da (Shield et al, 2008).

Bigarren hizkuntzaren jabetze-prozesuaren arloko teorien arabera, ikasketa naturalista edo inplizitua ez da beti nahikoa helduek L2ko goi-mailako hizkuntza-gaitasuna eskuratu dezaten; instrukzio esplizituaren bidez gainditzen omen dira hala sortzen diren zenbait ikasketa-eragozpen (Norris et al, 2000) (Ellis et al, 2007). Hori dela-eta, CALL sistemak mota askotako materialez osaturik daude gaur egun, gehienbat ikus-entzunezko materialez. Dena dela, CALL sistemak erabilgarriago eta sofistikatuago bihurtu baldin badira eta haien erabilera asko hedatu baldin bada, haietan hizketa-teknologiak inplementatu direlako izan da, batez ere hizketa-ezagutze automatikoa (ASR, *Automatic Speech Recognition*). Hizketa-teknologiek aukera ematen dute sistemaren eta erabiltzailearen arteko komunikazioa bi noranzkotan izateko, eta hori ezinbestekoa da hizkuntzak ikasteko prozesuan. Ikaslearekiko harremana baliatzen duten aplikazioen artean, honako hauek dira ezagunenak: ebakera-akatsak hauteman eta ebaluatzea, pertzepzioa lantzea eta doinu edo prosodia hauteman eta zuzentzea (Eskenazi, 2009). ASR sistemak, batik bat, ebakera ebaluatzeko erabiltzen dira; hala ere, badira ASRan oinarritutako beste hainbat aplikazio bigarren hizkuntzaren alderdi garrantzitsu batzuk lantzen laguntzen dutenak, hala nola morfologia eta sintaxia (Strik et al, 2012) (Van Doremalen et al, 2009). Aplikazio horiek egiaztapenean oinarritzen dira, ariketa bakoitzerako aurrez zehaztutako balizko erantzunen (zuzenen eta okerren) zerrenda batekin batera.

Artikulu honetan, mezuak egiaztatzeko euskararako sistema bat aurkezten da, ASRan oinarritua. Sistema horretan, erabiltzaileak esandakoa hitzez hitz aztertzen da, denbora errealean. Hitz bat hauteman bezain laster, pantailan erakusten du sistemak, eta espero duen hurrengo hitzaren egiaztapen-prozesuari ekiten dio. Hitzik hautematen ez bada, aldiz, begizta itxian jarraitzen du sistemak, harik eta hauteman beharreko hitza egiaztatzen den arte. Halako sistemak oso lagungarriak dira erabiltzaileak hitz-ordena finkoa duten esaldi edo erantzunak eman behar dituen ariketetarako, hala nola galderei erantzuteko, esaldiak berrordenatzeko

eta antzeko atazetarako. Egiaztapen-prozesua hitz-mailan nahiz fonema-mailan egin daiteke; hala ere, fonema-mailan egiteak zenbait eragozpen sortzen ditu behe-mailako ikasleentzat, batez ere ama-hizkuntzan existitzen ez diren fonemak egiaztatzean. Horrenbestez, lehendabiziko hurbilketa gisa, hitz-mailako egiaztapena hautatu da artikulu honetarako.

Artikulu hau honela dago antolaturik: sarreraren ondoren, mezuak egiaztatzeko sistemaren oinarri teorikoak azaldu dira. Jarraian, zenbait esperimentu eta haien emaitzak aurkeztu dira. Azkenik, garapenari, hobekuntzari eta etorkizuneko lanari buruzko zenbait konklusio eta hausnarketa azaldu dira.

2. Sistemaren oinarria

2.1. Datu-basea

Gaur egun, ez dago CALL sistemak garatzeko euskararako berariazko datu-baserik. ASR-rako datu-baseei dagokienez, eskuratu daitekeen datu-base bakarra *SpeechDat_eu* datu-basea da, telefono finkoko sarea baliatuz grabatua, 8 kHz eta 16 bitetan (Hernaez et al, 2003). Datu-base horren grabazio-ezaugarriak eta CALL sistemetarako espero direnak ez dira bateragarriak; hortaz, ez da egokia gure esperimentuetarako. Artikulu honetako lanak burutzeko hautatu den datu-basea ikerketarako soilik erabil daitekeen datu-basea da, hein batean *Speecon* espezifikazio estandarrei jarraitzen diena. 16 kHz eta 16 bitetan dago grabatua, bi distantziatara: aurikularreko mikrofonoaren bidez eta mahai gaineko mikrofonoaren bidez. Hemengo esperimentuetarako, aurikularreko mikrofonoaren bidez grabatutako audio-fitxategiak baino ez dira erabili. Datu-baseak 230 hizlariren grabazioak ditu, euskaldun zaharrak nahiz euskaldun berriak; halaber, euskalkien ezaugarriak nahiz euskara batuarenak ageri dira. Euskaldun zaharren azpicorpusa 149 hizlarik osatzen dute; gainerako 81en artean, euskara-gaitasun desberdineko hizlariak daude. Informazio hori guztia jasota dago, eta erraz erauz daiteke testuzko datu-fitxategietatik.

Audio-fitxategi bakoitzari dagokien transkripzio ortografikoa fitxategi banatan dago gorderik. Transkripzio fonetikoak, berriz, laborategiko G2P transkribatzailea erabiliz sortu dira. Erabili diren eredu akustikoak HMMak (*Hidden Markov Models*) dira, eta euskaldun zaharren azpicorpusaz soilik entrenatu dira ereduok (euskaldun berrien atala etorkizuneko ikerketa-lanetarako utzi da). Eredu akustikoak sortzeko, berez, azpicorpusaren bi heren erabili dira, gainerakoa esperimentuak egiteko utzi da. Batez beste, 170 fitxategi erabili dira hizlari bakoitzeko; horietariko 60tan, bat-bateko erantzuna ematen dute hizlariek (datak, zenbakiak eta abar); gainerako 110ak, berriz, komandoak dira, irakurriak, batik bat hitz solteez osatuak.

2.2. ASRa eta PS parametroen kalkulua

Mezuak egiaztatzeko sistemak Aholab Seinale Prozesaketako Laborategian garatutako hizketa-egiaztapenerako sistema baliatzen du (Odriozola et al, 2012). Egiaztapen-sistema ASR estandar batean oinarriturik dago, eta HMMak erabiltzen ditu eredu akustiko gisa. 16 kHz-eko seinaleak prozesatzen ditu 16 biteko laginez, eta 39 MFCCko (*Mel-frequency cepstral coefficients*) bektoreak erauzten ditu 10 ms-ero, 25 ms-ko iraupeneko tramen bidez. Deskodeketa-prozesua Viterbiren algoritmoaren bitartez gauzatzen da, HMMez osatutako sarean zehar. Hautatu diren HMMek fonema testuingurudunak modelatzen dituzte (trifonemak).

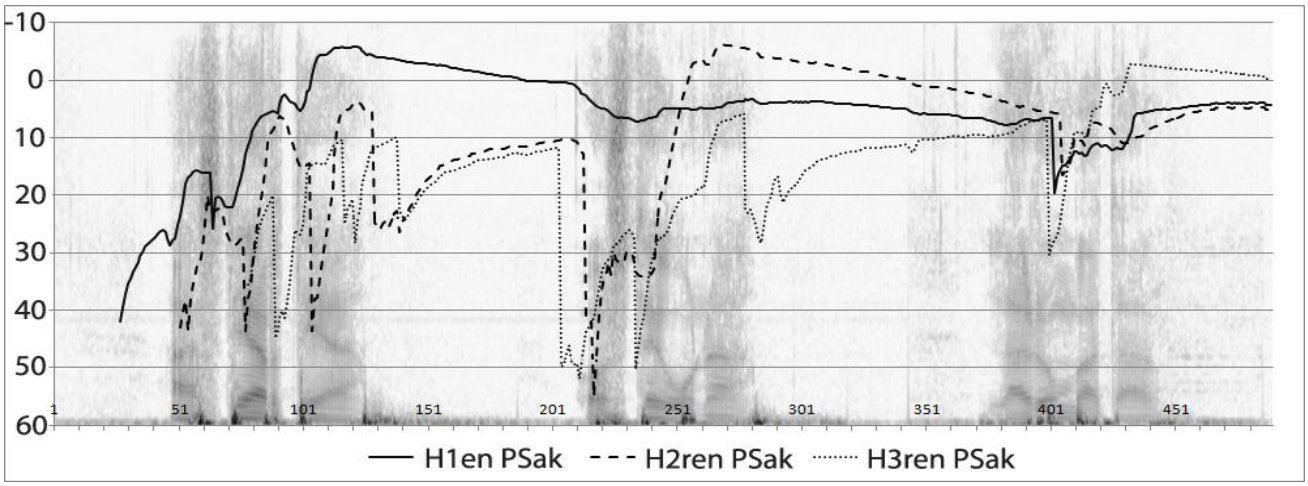
Egiaztapen-prozesuak beste HMM sare bat eskatzen du, paraleloan lanean. Hala, y_u fonemaren *GOP* (*Goodness of Pronunciation*) parametroa lortzen da, zeina X_u segmentu akustikoan zehar kalkulatzen baita haren a posteriori probabilitate gisa, (1) ekuazioaren bidez (fonemaren indizea da u). Ekuazio horretan, T_u da y_u fonemak irauten duen trama kopurua; j_{max}, berriz, segmentu horretarako probabilitaterik altuena ematen duen fonema-eredua, sare paraleloaren bidez kalkulatua.

$$GOP(y_u) = \log Pr(y_u|X_u) \simeq \frac{1}{T_u} \log\left[\frac{p(X_u|y_u)}{p(X_u|y_{jmax})}\right] \quad (1)$$

Hitz-mailan, hitzaren osoko konfiantza-parametroa hura osatzen duten fonemen *GOP* koefizienteen batura haztatu gisa kalkula daiteke. Hala, (2) ekuazioaren bidez kalkulatzen da *PS* (*Phoneme Score*) koefizientea. Ekuazio horretan, w_k da hitza osatzen duten N fonemen arteko k.aren haztapena. Haztapenak berdinak izan ohi dira fonema guztientzat (Mak et al, 2003).

$$PS(word) = \sum_{k=1}^{N} w_k \cdot GOP(phoneme_k) \quad (2)$$

1. irudian, *PS* koefizienteen jokaeraren adibide bat ageri da. Analizatutako audio-fitxategia isiluneen bidez banandutako hiru hitzez osaturik dago. *PS* koefizienteak hitz bakoitzaren azken HMMaren azken egoerako irteeran kalkulatu dira, trama guztietarako. Hortaz, hiru PS koefizienteen segida daude irudian marraztuta, bakoitza lerro mota desberdin batez. Gainera, hitz bakoitzaren iraupena zein den hobeto irudikatzearren, seinalearen adierazpen espektralaren gainean ezarri dira hiru lerroak. Irudian ageri denez, hitzaren amaieran du lerro bakoitzak bere maximoa (-*PS* kurbarena); horrek pentsarazten digu hitz bat egiaztatzeko aski izan daitekeela atalase jakin baten gainean *PS* kurbaren maximoari antzematea.



1. irudia: Isilunez banandutako hiru hitzen (H1: "asteartea"; H2: "osteguna"; H3: "larunbata") *PS* koefizienteen kurbak, seinalearen adierazpen espektralaren gainean erakutsia.

2.3. Erabaki-atalaseak

PS koefizienteekin jardutean izan ohi den arazo nagusia erabaki-atalaseak kalkulatzea da; izan ere, bi banaketa behar dira horretarako: batetik, hitz zuzenak egiaztatzean lortzen diren *PS* koefizienteen banaketa; bestetik, okerreko hitzak egiaztatzean lortzen dena. Bi banaketa horiekin, ebaketa-puntu bat ezartzen da. Puntu hori, eskuarki, EER (*Equal Error Rate*) izaten da; alegia, oker onartutako eta oker baztertutako hitzen probabilitatea berdina den puntua. Atalase hori, beraz, doitu egin daiteke L2 ikaslearen beharrak kontuan izanda. Adibidez, behe-mailetako ikasleentzat, errore kopuru gehiago onar daitezke; sistema, aldiz, zorrotzagoa izan daiteke goi-mailako ikasleekin.

Hitz zuzenen *PS* koefizienteak kalkulatzeko, ASRa lerrokatze behartu moduan erabiltzen da, eta hala lortutako emaitzen segmentazioa erabiltzen da. Okerreko hitzen *PS*ak kalkulatzea, ordea, ez da hain sinplea, nekeza baita ikasleen erroreez osatutako datu-baseak aurkitzea (Odriozola et al, 2012) (Kanters et al, 2009). Hortaz, aukera bat da hiztegian akatsak txertatzea, hau da, erroreak simulatzea. Artikulu honetan, hau da erroreak simulatzeko erabili den prozedura: fitxategi bakoitzaren transkripzioan, hitz baten ordez hiztegiko ausaz hautatutako beste edozein ezartzea. Hala lortutako *PS* koefizienteen banaketarekin, % 2.12ko EER lortzen da, zeina oso emaitza itxaropentsua baita; izan ere, horrek esan nahi du oso txikia dela hitz berri bat egiaztatzean lortuko dugun klasifikazio-errorea.

3. Esperimentuak eta emaitzak

Emaitzak ebaluatzeko hautatu den neurria SA (*Scoring Accuracy*) koefizientea da, (3) ekuazioaren bidez kalkulatzen dena

$$SA = \left(\frac{(CA+CR)}{(CA+CR+FA+FR)} \right) \cdot 100 \qquad (3)$$

Ekuazioan, *CA*: *Correctly Accepted* (zuzen onartua); *CR*: *Correctly Rejected* (zuzen baztertua); *FA*: *Falsely Accepted* (oker onartua); *FR*: *Falsely Rejected* (oker baztertua).

Hasierako bi esperimentu egin dira kalkulatutako erabaki-atalasearen sendotasuna aztertzeko: lehen esperimentuan, datu-basearen *test* ataleko 2.218 fitxategiak erabili dira (7.296 hitz guztira). Proba honetan, jo da hitz guztiak zuzenak direla. Bigarren esperimentuan, datu-basearen atal bereko hitz isolatuak baino ez dira erabili: 1.174 fitxategi; proba horretan, errore artifizialak simulatu dira, fitxategi bakoitzari dagokion transkripzioko hitza aldatuta. Hortaz, jo da hitz guztiak okerrekoak direla.

Bi esperimentuotan lortutako emaitzak 1. taulan ageri dira. Lehen esperimentuan, % 97,18 da SA, edo, errorea kontuan hartuz gero, % 2,82; hau da, EER puntua kalkulatzean lortutakoaren oso antzekoa. Bigarren esperimentuan, % 100eko SA lortu da; alegia, hitz guztiak jo dira okertzat. Errorea, beraz, zero da kasu horretan.

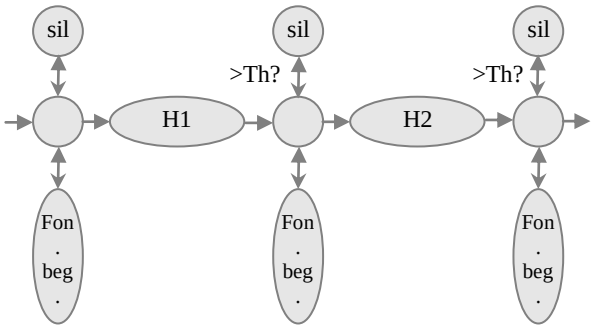
	CA	**CR**	**FA**	**FR**	**SA**
1. esp.	7.090	---	---	206	% 97,18
2. esp.	---	1.174	0	---	% 100,00

1. taula: 1. eta 2. esperimentuen emaitzak

Lehen esperimentuko emaitzetako akatsak azaletik aztertuta, ikusi da *FR* hitzen % 75,24k hiru fonema edo gutxiago dituztela. Hortaz, bistakoa da ezen, zenbat eta luzeagoa hitza, hainbat sendoagoa izango dela emaitza. Horren azalpena ertzetako fonemen eragina izan liteke, zeina nabarmenagoa da hitz laburren kasuan. Etorkizunerako lan-ildo gisa utzi da oraingoz.

4. Sistemaren diseinua

Inguru errealean, ikasleek akatsak egiten dituzte. Hortaz, egiaztapen-sistemak gai izan behar du soberako ahots-segmentu horiek maneiatzeko; bestela, ezin izango litzateke aurrez ikusi nolakoa izango den Viterbi deskodetzailearen segmentazioa, eta, beraz, ezin izango lirateke behar bezala kalkulatu egiaztapenerako koefizienteak. Hori saihesteko, sistemaren azken diseinuan, fonema-begiztak erantsi zaizkio ASRaren deskodeketa-sareari lehen hitzaren aurrean, hitz-bitarteetan eta azken hitzaren ondoren. Diseinuaren eskema 2. irudian ageri da.



2. irudia: Bi hitzeko (H1 eta H2) esaldi batentzako deskodeketa-sarea, fonema-begizta erantsiak dituena

Sistemak, lehendabizi, lehen hitzaren (H1) *PS* koefizientea kalkulatzen du, trama oro (10 ms-ero), koefizientearen balioak 2.3 puntuan adierazitako moduan kalkulatutako atalasea (*Th*) gainditu eta maximoa lortzen duen arte. Une horretan, hitza pantailan ezartzen da, eta bigarren hitza (H2) egiaztatzeko prozesuari ekiten dio era berean.

Sistema hori inguru errealago batean ebaluatzearren, beste esperimentu bat egin da, hitz zuzenak eta okerrekoak nahasten dituena. Horretarako, analizatu beharreko esaldi edo hitz segida bakoitzaren transkripzioan, kendu egin da hitz bat. Hartara, honako egoera hau simulatzen da: erabiltzaileak zenbait hitz zuzen esaten ditu, ondoren okerreko hitz bat, eta, jarraian, hitz zuzenak berriro ere.

886 esaldi erabili dira esperimentu honetan, 5.080 hitz guztira. Guztira, beraz, 5.080 *PS* koefiziente kalkulatu dira; horietatik 4.194 hitz zuzenei dagokie, eta gainerako 886ak okerrekoei. Esperimentuan lortutako *SA* parametroa 2. taulan ageri da.

CA	4.157
CR	752
FA	134
FR	37
SA	**% 96,63**

2. taula: Errore simulatuen esperimentuaren emaitzak

5. Konklusioak eta etorkizuneko lana

Artikulu honetan, mezuak egiaztatzeko sistema baten diseinua aurkeztu da, CALL edo ordenagailu bidez bigarren hizkuntzak ikasteko sistemetan inplementatzeko balio duena. Sistema horretan, hitz zuzenak onartzea bezain garrantzitsua da okerreko hitzak baztertzea, eta, baldintza hori kontuan izanda, erabaki-atalasea kalkulatzeko teknika bat aurkeztu da artikuluon, erroreak artifizialki txertatzea, alegia. Esperimentuen emaitzek erakusten dute sistemaren portaera pixka bat hobea dela hitz zuzenak onartzean okerreko hitzak baztertzean baino, nahiz eta erabaki-puntua kalkulatzeko EER balioa erabili den. Etorkizuneko lan gisa, probak inguru errealean egitea ikusten da. normala hemen 10eko tamainan.

6. Aipamenak

Ellis, N.C., Bogart, P.S.H. (2007): *Speech and Language Technology in Education: the perspective from SLA research and practice*, In Proceedings ISCA ITRW SLaTE, Farmington (AEB).

Eskenazi, M. (2009): *An overview of spoken language technology for education*. Speech Communication 51(10): 832–844.

Hernaez, I., Luengo, I., Navas, E., Zubizarreta, M., Gaminde, I., Sanchez, J. (2003): *The Basque speech_dat (II) database: a description and first test recognition results*, In Eurospeech-2003, 1549-1552.

Kanters, S., Cucchiarini, C. and Strik, H. (2009): *The Goodness of Pronunciation algorithm: a detailed performance study*, In Proceedings of SLaTE 2009, Birmingham.

Mak B., Siu M., Ng M., Tam Y., Chany Y., Chan K., Leung K., Ho S., Chong F., Wong J., Lo J. (2003): PLASER: Pronunciation Learning via Automatic Speech Recognition. In Proc. of HLT-NAACL, May, 2003, Edmonton, Canada.

Norris, J.M., Ortega, L. (2000): *Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis*, Language Learning, vol. 50, pp. 417-528.

Odriozola, I., Navas, E., Hernáez, I., Sainz, I., Saratxaga, I., Sánchez, J., Erro, D. (2012): *Using an ASR database to design a pronunciation evaluation system in Basque*. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pp. 4122-4126.

Shield, L., Kukulska-Hulme, A. (2008): *Special edition of ReCALL* (20, 3) on Mobile Assisted Language Learning.

Strik, H., Colpaert, J., Doremalen, J., Cucchiarini, C (2012): *The DISCO ASR-based CALL system: practicing L2 oral skills and beyond*. LREC 2012, pp. 2702-2707.

Van Doremalen, J., Strik, H., Cucchiarini, C. (2009): Utterance Verification in Language Learning Applications. Proceedings of the SLaTE-2009 workshop, Warwickshire, England.