Euskarazko ahoskera ebaluatzeko sistema baten garapena, ASR datu-basearen bidez

Igor Odriozola, Eva Navas, Inma Hernaez, Iñaki Sainz, Ibon Saratxaga, Jon Sánchez, Daniel Erro

University of the Basque Country
{igor, eva, inma, inaki, ibon, ion, derro}@aholab.ehu.es

Abstract

The aim of this paper is to explain how to use an ASR database to develop Computer-Assisted Pronunciation Training (CAPT) systems for under resourced languages, to evaluate then how 'Goodness of Pronunciation' (GOP) scores perform in this environment, like when we use a database not specifically designed for pronunciation evaluation.

Laburpena

Artikulu honen helburua baliabide mugatuetako hizkuntzatarako CAPT sistemak garatzeko ASR bat erabilera, eta giro honetan (adibidez, ahoskerarako bereiziki diseinatu ez den datu base batez batean) GOP puntuazioen portaera erakustea da.

Keywords: Basque, Pronunciation evaluation, ASR database

Gako hitzak: Euskara, Ahozkeraren ebaluaketa, ASR datu-basea

1. Sarrera

Ahotsaren teknologiak gero eta garrantzitsuago bihurtzen ari dira bigarren hizkuntza bat ikasterakoan[1], ahoskera lantzeko erabiltzen diren CAPT[2] izeneko sistemetan gehien bat. Interes handia sortu dute teknika hauek, bai ikertzaileen artean, baita L2 tutoretzarako teknologia berrien sistemetan lan egiten duten konpainietan eta L2 irakasle eta ikasleengan ere. Sistemaren abantailen artean, ikasleek haien kabuz, haien erritmoan lan egin ahal izatea dago, ikasgelan ikasitakoaren osagarria eta hobekuntza moduan.

90eko hamarkadan garatutako lehenengo puntuaketa metodoak berba eta esaldi mailetan oinarritzen ziren (Hamada et al., 1993; Hiller et al., 1993; Rogers et al., 1994). Hamarkadaren bukaeran errore fonemikoak saihesten irakasteko diseinatutako sistemak garatu ziren (Kawai and Hirose, 1997; Kim et al., 1997; Ronen et al., 1997). Bertan, hiztun ez-natiboen datu-baseak erabili ziren. Garai hartan deskribatu zen ASR-ren erabilera ere, HMMekin batera, esaldi bateko fonema bakoitza puntuatzeko (Eskenazi, 1996).

Azken urteotan, CAPT sistemen garapena multimedia teknologiak bateratzean enfokatu da, erabiltzaileari erantzun osoagoa eman ahal izateko. Giro honetan, PCetan oinarritutako ahoskera irakasteko sistemen artearen egoera *Pronunciation Power*, *American Sounds*, *Phonics Tutor* eta *Eyespeak* programetan datza. (LearningVillage, 2009). Duela gutxi, *Euronounce* europar proiektua garatu da irakaste aplikazio multimodala sortzeko (Demenko, G. et al., 2009). Sistema hauetan guztietan , ikasleak ahozko sarreraren analisia eta errekonozimendua erabil dezake, bere ahotsa grabatuz. Ondoren, grabaketaren espektroa erakusten da, beste erreferentzia ahotsekin batera, soinu- eta ikuste-konparaketa edukitzeko. Sistema gehienetan, hiru koloretako sistema erabiltzen da fonema baten zuzentasun maila adierazteko, *Plaser*en adibidez (Mak et al., 2003).

Aplikazio gehienetan erabiltzen den ahoskera automatikoki puntuatzeko metodoa, GOP[3] izenekoa, fonema baten traman oinarritutako atzeko probabilitatea da y_u fonemaren X_u zati akustiko bakoitzeko (u fonemaren indizea delarik), (1) ekuazioan ikusten dugu. Bertan, N fonema kopurua da eta j_{max} emandako zatirako antzekotasun handiena ematen duen fonemaren eredua. Praktikan, (1) ekuazioaren denominadorea fonema begizta batek emandako Viterbiren probabilitatearekin aldatzen da. Lan batzuen arabera neurri ona da hau (Witt & Young, 2000; Franco et al., 2000), baina hobekuntzaren bat erabiltzen da

$$GOP(y_u) = \log \Pr(y_u \mid X_u) \approx \frac{1}{T_u} \cdot \log\left[\frac{p(X_u \mid y_u)p(y_u)}{\sum_{k=1}^{N} p(X_u \mid y_k)p(y_k)}\right] \approx \frac{1}{T_u} \cdot \log\left[\frac{p(X_u \mid y_u)}{p(X_u \mid y_{j_{max}})}\right] \quad (1)$$

[1] *L2* arloa deritzona, *Second Lenguage Learning*

[2] *Computer-Assisted Pronunciation Training*, Ordenagailuz lagundutako ahoskeraren entrenamendua

[3] *Goodness of Pronunciation,* ahoskeraren kalitatea

puntuaketak hobetzeko. GOPa atari bat ezartzeko erabiltzen da, fonema ondo ahoskatu den erabakitzeko, baina normalean, GOP globala erabili beharrean, puntuazio desberdinak erabiltzen dira fonema edo fonema talde bakoitzerako.

GOP horiek kalkulatzeko ahoskera hiztegi estandar bera erabiltzen da, bai segmentu zuzenetarako, eta baita akastunentzat ere (hemendik aurrera, segmentu zuzenak eta erroredunak). Testuinguru erroredunak, normalean aurrez aurretik definitzen dituzte linguistek, eta datu-baseak analizatu behar diren fonemen gehieneko adibide kopurua edukitzeko diseinatzen dira. Ondoren, bi puntuazio banaketa desberdin garatzen dira aztertu behar diren fonementzat (edo fonema taldeentzat): bata segmentu zuzenekin, eta beste bat sarrera erroredunekin, horrela EER[4] atari bat ezarri ahal izateko.

Euskararen kasuan gertatzen den moduan, hizkuntzarako holako datu-base bat ez dagoenez, artikulu honetan ebazpide bat proposatzen da, ASR orokor bat erabiliz, segmentu erroredunen identifikazioaren puntuazioaren banaketa lortzeko. Hau lortzeko, aldaketak (errore kontrolatuak) sartzen dira ahoskera hiztegian bertan. Fonema bat aldatzen da, hiztegi osoan, mota bereko beste fonema batekin (bokalak, herskariak, dardarkariak, etab.) eta fonema horietarako lortutako puntuaketa fonema erroredunen puntuazio banaketa moduan erabiltzen da.

Artikulu honen helburua baliabide mugatuetako hizkuntzatarako CAPT sistemak garatzeko ASR bat erabilera, eta giro honetan (adibidez, ahoskerarako bereiziki diseinatu ez den datu-base batean) GOP puntuazioen portaera erakustea da.

2. Datu-basea

CAPT sistema guztietan erabiltzen dira grabaketa natiboak eta ez-natiboak ikaslearen grabaketak hobeto klasifikatzeko, beraz guztiak oinarritzen dira datu-base berezietan, bi soinu mota dutenak bi soinu mota dituztenetan. Hala ere, sistema gehienetan indar handiagoa ematen zaie L1-L2 hizkuntza parean adituek hautatutako fonema problematiko batzuei. Bereiziki garatzen dira, beraz, datu-base batzuk L1-L2 hizkuntza pare hedatuenentzat (Cylwik et al., 2008).

Euskararen kasuan, baliabideak mugatuak dira eta ez dago hizkuntza hedatuenen ahots teknologien beste datu-baserik. Adibidez, euskararako, ASR datu-base publiko bakarra dago (Hernaez, I., 2003), telefonia sare finkoaren bidez grabatua. Beraz, gaur egungo beharra ez da CAPT datu-base bat grabatzea; ahoskera irakasteko sistemak beste informazio mota batez garatu behar dira.

Esperimentu hauetan erabilitako datu-basea ahotsaren ezagutzarako grabatu da. Speecon-en antzekoa da, hiztun natiboak eta ez-natiboak ditu, eta euskara batua eta baita dialektala ere. 230 informanteren ahotsak grabatu dira, Euskal Herriko leku desberdinetakoak (euskararen egoera desberdinekin, ikuspuntu ofizialean, erabileran eta inguruko beste hizkuntzetan, frantsesa eta gaztelania gehien bat). Beraz, irregulartasun desberdin ugari agertzen dira hiztunen artean fonema berdinetan. Aldaera horiek ez dira transkribapenetan agertzen: beharrezkoa da hiztunaren edo sesioaren ezaugarriak begiratzea ezagutzeko. Audio fitxategiekin batera transkribapen ortografikoa dago, eta arauetan oinarritutako P2G transkriptorea erabili da euskara batuerazko transkribapenak lortzeko.

3. Esperimentuak

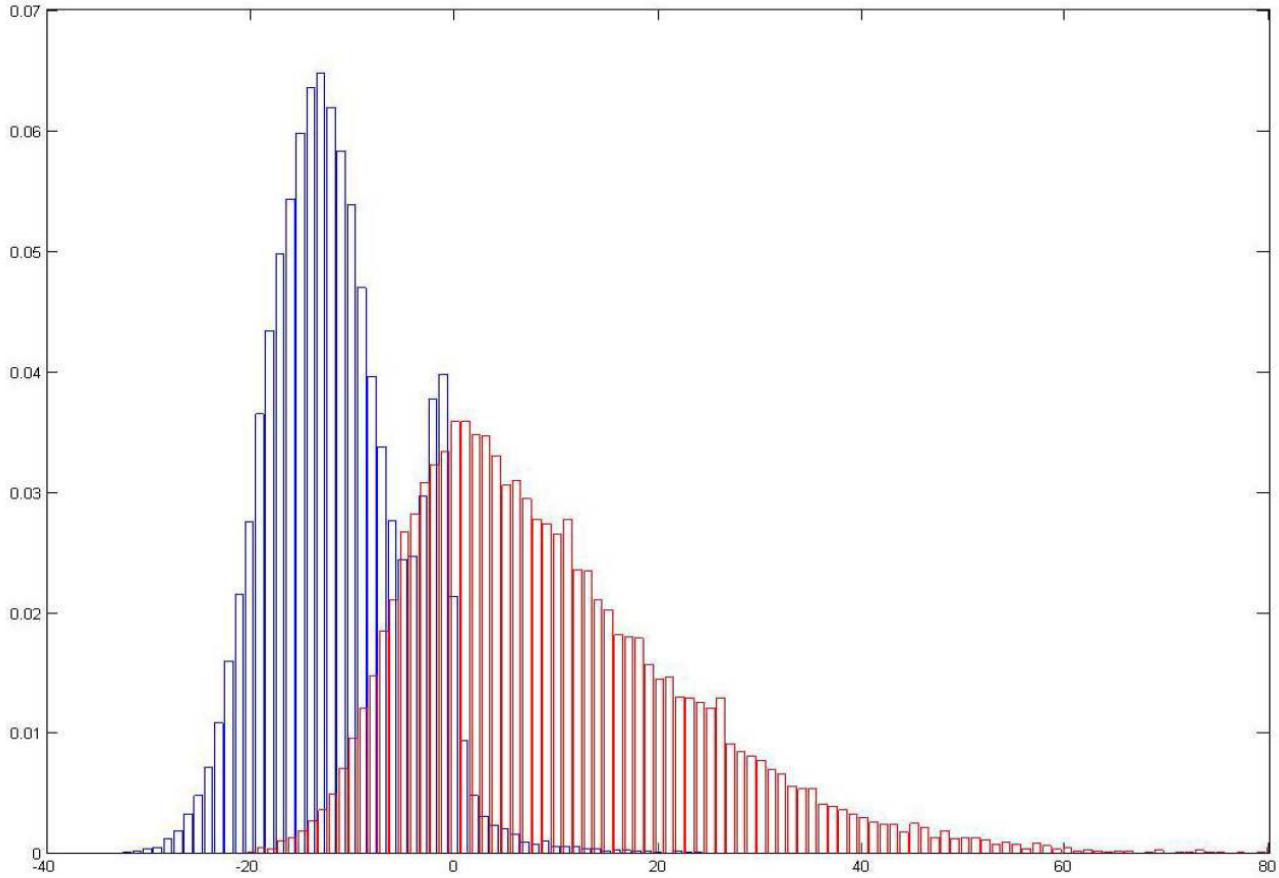
Esperimentu hauetarako aukeratutako estrategia ondorengoa da: lehenengoa, trifonemen HMMak entrenatu dira, testuinguruarekiko independenteak, datu-basearen bi herenekin. Eta monofonemen begizta inplementatu da paraleloan abiarazteko, GOP balioa (1) ekuazioan definitu den moduan lortzeko. Lehenengo esperimentuan fonema guztien GOP puntuazioak lortu dira, fonemen begiztarekin lerrokadura paraleloa bultzatuz. Lerrokadura bultzatzeko, bi aldaera hartu dira kontuan ondoko berbetarako: koartikulazioa eta haien arteko isilunea.

Fonema bakoitzerako puntuazio multzo bat lortu ostean, transkripzio erroreak sartu dira hiztegi transkripzio bakoitzean, ausazko trifonema bat hautatuz, eta mota bereko beste batekin aldatuz, ausazkoa hau ere. Testuinguruarekiko dependentzia ere aldatu da, baina transkribapenen koherentzia mantenduz. Ondoren, lerrokadura paralelo bultzatuaren esperimentu berria gauzatu da, txertatutako fonema bakoitzaren GOP puntuazioa kalkulatuz, hau da, testuinguru erroredun simulatuan dagoen fonemaren GOP puntuazioa kalkulatuz.

Azkeneko esperimentu hau birritan errepikatu da, informazio gehiago edukitzeko: testuinguru erroreduneko puntuazio kopurua zuzenekoa baino askoz ere txikiagoa baita. Ondoren, banaketa bien histogramak kalkulatzen dira fonema bakoitzeko, fonema bereko trifonema puntuazio guztiak bateratuz. 1. Irudian *a* fonemaren adibidea ikus daiteke, bi dentsitate banaketekin. Ataria funtzio bietatik kalkulatutako EERaren bidez kalkulatzen da, eta fonema bakoitzerako desberdina izango da.

Lortutako atariak ebaluatu behar dira, eta horretarako garapenean dago beste esperimentu bat datu-basetik gordetako testerako zatiaz, lerrokadura hiztegi transkribapenekin mantenduz. Testetako fitxategien transkribapen automatikoak adituek ebaluatu dituzte, eta hiztegi kanonikotik mugitzen diren aldaerak markatu dira. Atariak lortzeko, eskuz prestatutako datuekin konparatuz egiten da.

[4] *Equal Error Rate*



1. Irudia: *a* fonemaren puntuazioen histograma normalizatuak: barra urdinak testuinguru erroreduneko puntuazioen banaketa irudikatzen dute; gorriak, berriz, testuinguru zuzenekoena

4. Ondorioak

Baliabide mugatuen hizkuntzatarako ASR-rako ahozko datu-base orokorra erabiliz CAPT sistemak diseinatzeko metodo berria deskribatu da, eta GOP mailaren ataria automatikoki ezartzeko sistema proposatu da.

5. Aipamenak

Cylwik, N, Demenko, G, Jokisch, O., Jackel, R., Rusko, M., Hoffmann, R., Ronzhin,A., et al. (2008). The use of CALL in acquiring foreign language pronunciation and prosody–general specifications for Euronounce Project. Proc. SASR, 123-130.Retrieved from http://www.ptfon.pl/files/11SLTRB06.pdf

Cylwik, N., Wagner, A., Demenko, G, 2009. The euronounce corpus of non-native polish for asr-based pronunciation tutoring system, in SLaTE.

Demenko, Grazyna, Cylwik, Natalia, & Wagner, A. (2009). Applying speech and language technology to foreign language education. 2009 International Multiconference on Computer Science and Information Technology, 2, 457-463. Ieee. doi:10.1109/IMCSIT.2009.5352682

Eskenazi, M., 1996. Detection of foreign speakers' pronunciation errors for second language training – preliminary results. In: ICSLP'96. Philadelphia, PA, USA.

Franco, H. (2000). Combination of machine scores for automatic grading of pronunciation quality. Speech Communication, 30(2-3), 121-130. doi:10.1016/S0167-6393(99)00045-X

Hamada, H., Miki, S., Nakatsu, R., 1993. Automatic evaluation of English pronunciation based on speech recognition techniques. IEICE Trans. Inform. Syst. E76-D (3), 352-359.

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.

Hiller, S., Rooney, E. Laver, J., Jack, M., 1993. SPELL: An automated system for computer-aided pronunciation teaching. Speech Communication 13, 463-473.

Kawai , G., Hirose, K., 1997. A call system using speech recognition to train the pronunciation of Japanese long vowels, the mora nasal and mora obstruent. In: Proceedings EUROSPEECH'97. Rhodes, Greece.

Kim, Y., Franco, H., Neumeyer, L., 1997. Automatic pronunciation scoring of specific phone segments for language instruction. In: Proceedings EUROSPEECH'97. Rhodes, Greece.

Learning Village. Educational Software Review, Retrieved on 15th July 2008 from http://www.learningvillage.com/html/guide.html

Mak, B., Siu, M., Ng, M., Tam, Y.-cheung, Chan, Y.-chung, Chan, K.-wah, Leung, K.-yee, et al. (2003). PLASERı: Pronunciation Learning via Automatic Speech Recognition. HLT-NAACL Workshop on Building Educational Applications using Natural Language Processing.

Rogers, C., Dalby, J., DeVane, G., 1994. Intelligibility training for foreign-accented speech: A preliminary study. J. Acoust. Soc. Amer. 96 (4), pt. 2.

Ronen, O., Neumeyer, L., Franco, H., 1997. Automatic detection of mispronunciation for language instruction. In: Proceedings EUROSPEECH'97. Rhodes, Greece.

Witt, S., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. Speech Communication, 30(2-3), 95-108.doi:10.1016/S0167-6393(99)00044-8