Euskarazko datu base emoziodun baten ebaluaketa subjektiboa

Iñaki Sainz, Ibon Saratxaga, Eva Navas, Inmaculada Hernáez, Jon Sanchez, Iker Luengo, Igor Odriozola, Imanol Madariaga

Aholab – Elektronika eta Telekomunikazioak Saila. Ingenieritza Goi eskola. Euskal Herriko Unibertsitatea.
inaki, ibon, eva, inma, ion, ikerl, igor, imanol @aholab.ehu.es

Abstract

This paper describes the evaluation process of an emotional speech database recorded for standard Basque, in order to determine its adequacy for the analysis of emotional models and its use in speech synthesis. The corpus consists of seven hundred semantically neutral sentences that were recorded for the Big Six emotions and neutral style, by two professional actors. The test results show that every emotion is readily recognized far above chance level for both speakers. Therefore the database is a valid linguistic resource for the research and development purposes it was designed for.

Laburpena

Artikulu honetan, euskara batuaz grabatutako datu-base emoziodun baten ebaluaketa prozesua deskribatzen da, egiaztatzeko ea datu-basea egokia den emozioen ereduak aztertzeko eta ahotsaren sintesia lortzeko. Corpusean semantikoki neutroak diren zazpiehun esaldi daude, sei emozio nagusietan eta neutroan grabatuak, bi aktore profesionalen ahotsez. Proben emaitzen arabera, ausazko atalasearen gainetik antzematen zaio emozio bakoitzari, hiztun bientzat. Beraz, datu-basea hizkuntza baliabide baliagarria da bere ikerketa helburuak betetzeko.

Keywords: Emotional speech database, subjective evaluation

Gako hitzak: Datu-base emozioduna, ebaluaketa subjektiboa.

1. Sarrera

Ahots-sintesirako tekniken aurrerakuntza dela eta, Corpus bidezko TTS sistema gehienen ulergarritasuna giza hizkeraren ia parekoa da. Hala ere, ahots sintetikoaren naturaltasuna eta jariotasuna urruti daude gizakiarena bezala nabaritzetik. Ahots-sintesian, oraindik ere faltan dagoen ezaugarri garrantzitsua da emozioak egoki adieraztea. Emozioak ahots sintetikoaren monotonia gutxitzeko eta gizakiaren eta makinaren arteko komunikazioa hobetzeko bidea dira.

Azken urteotan, ahots-sintetizagailu emozioduna garatzeko hainbat saiakera izan dira (Ilda eta beste, 2003)(Murray eta Arnott, 1996)(Bulut eta beste, 2002). Baina azken emaitzak ez dira nahi bezain onak izan. Emozio sinesgarriak adierazteko, sakon ikertu behar dira ahots emoziodunaren prosodiaren ezaugarriak (tonu kurba, fonemen iraupena eta energia kurba), eta, hori egin ahal izateko, datu base emoziodun bat grabatu beharra dago.

Prosodiaren aldaketek eragin handia dute adierazten den emozioan (Vroomen eta beste, 1993)(Montero eta beste, 1999), baina bide horretatik sorturiko ahots emozioduna ez da oso naturala (Schroeder, 1999), benetako prosodia erabili arren ere (prosodiaren kopia) (Heuft eta beste, 1996). Benetan falta dena da ahots emoziodunaren espektro-ezaugarriak erauztea (Rank eta Pirker, 2006). Horretarako, ezaugarri akustikoak esplizituki modelatu beharrean (lan nekeza izan daitekeena), hori inplizituki egingo duten corpus bidezko teknikak erabil daitezke.

Corpus bidezko metodoetan, esaldi bakoitza datu-baseko unitate optimoak lotuz eratzen da. Unitateak aukeratzeko algoritmoak kostu-funtzio globala minimizatu behar du. Funtzio horrek helburu-kostua eta lotura-kostua hartzen ditu kontuan, 0 (kasurik onena) eta 1 (okerrena) arteko balioak dituztenak (Hunt eta Black, 1996).

Helburu-kostuak behar den unitatea (testu-ahots sintetizagailuaren prosodia moduluak aurreikusitakoa) eta datu-basean dauden unitateen arteko antzekotasuna neurtzen du. Lotura-kostuak unitate-loturaren kalitatea neurtzen du (0 da, datu-basean elkarren segidan baldin badaude). Unitateak aukeratzen dituzten sistemek emaitza onak dituzte helburu jakin bakoitzerako aukera aski daudenean, ez baitago, ondorioz, ahotsaren naturaltasuna distortsionatuko luketen uhin-formaren aldaketarik egin beharrik.

Emozio bakoitzerako aukerako aski unitate edukitzeko, datu-base handia behar da.

Artikulu honetan emozioen datu-base baten ebaluaketa aurkezten da. 2. atalak corpusaren diseinua eta grabaketa-prozesua azaltzen du. Ebaluaketa prozedura 3. atalean azaltzen da, eta emaitzak 4. atalean aurkezten eta eztabaidatzen dira. Azkenik, ondorio batzuk azaltzen dira.

2. Corpusaren diseinua eta grabaketa

Hemen deskribatutako datu-basea helburu nagusi birekin sortu da. Alde batetik, corpus emoziodunean oinarritutako Euskararako TTS sistema garatzeko erabili nahi dugu; beste aldetik, ahots emoziodunaren prosodia eta akustika aztertzeko baliagarria izatea nahi da.

2.1. Ahots emozioduna grabatzea

Ahots emozioduna grabatzeko, hainbat aukera daude (bat-batekoa, sorrarazia, antzeztua eta abar), eta bakoitzak bere abantailak eta desabantailak ditu:

- Bat-bateko emozioak: Argi dago benetako emozioak grabatzen dituela, baina, jendearen pribatutasunean duen eragina dela eta, moralki desegokia da. Gainera, edukia kontrolatu ezin denez, ia ezinezkoa da corpus bidezko sintesirako datu-base egokia biltzea, eremu fonetiko mugatua izango luke eta.
- Emozio sorraraziak: Hizlaria emozio bakoitza eragiten duen egoerara eramaten da. Egoera bera izanda ere hizlari bakoitzak desberdin erreakzionatzen duenez, grabatuko den emozioa ez dago guztiz ziurtatuta. Beste desabantaila bat da ez dela oso etikoa egoera negatiboak sorraraztea haserrea eta tristura grabatu ahal izateko.
- Emozio antzeztua: Teknika horretan, aktore profesionalek testu bat irakurtzen dute, dagokion emozioa ematen saiatuz. Horrela grabatzeak emozioak gehiegi nabarmentzea ekar dezake, eta igarriko litzateke ez direla benetakoak; halere, badirudi, azken finean, entzuleok ondo ezagutzen ditugula.

Lan honetan, hirugarren metodoa aukeratu da, bere abantailak direla eta: alde batetik, datu-basearen edukia zehaztu daiteke diseinatutako corpusaren oreka fonetikoa eta aldagarritasun akustikoa gordez; bestetik, erraztu egiten du emozio bakoitzaren ezaugarriak aztertzea eta konparatzea.

Horretarako, esaldi semantikoki neutralak erabili dira, emozio bakoitzari lotuta ez daudenak, eta testu multzo bera erabili da emozio guztiak grabatzeko. Aukera honen baliagarritasuna esperimentalki frogatuta dago (Navas eta beste, 2006).

Eduki espresiodunari dagokionez, sei emozio nagusiak hartu dira kontuan (Haserrea, Beldurra, Harridura, Nazka, Poza, eta Tristura) (Cowi eta Cornelius, 2003), horiek baitira, eskuarki, bereizgarrienak. Horiez gainera, estilo neutrala ere hartu da kontuan, TTSarekin emoziorik gabeko ahotsa sortu ahal izateko.

Emozio bakoitzeko ordu bete inguruko grabaketa lortzeko, 702 esaldi aukeratu ziren, corpusak aztertzeko tekniken bidez oreka fonetikoa eta difonema-estaldura, biak, ziurtatuz. Corpusa hemen (Saratxaga eta beste, 2006) deskribatzen da zehazki. Esatari profesional bik grabatu zuten: 40 urteko gizonezko bikoizketa-aktore batek eta 37 urteko emakumezko irrati-esatari eta aktore batek.

3. Ebaluaketa prozesua

Grabatutako datu-basearen emozio-edukia egokia den jakiteko, ebaluaketa subjektiboa egin da.

3.1. Azterketaren diseinua

Aukera finkoko azterketa egin zen, jakiteko ea entzuleak gai ziren grabatutako ahots zatiaren emozioa zuzen ezagutzeko. Aktore bakoitzaren 30 ahots zati aurkeztu zitzaien web bidezko ordenagailu ingurune batez. Ahotsak nahastu eta hamarkako formularioetan sailkatu ziren. Ebaluatzaileek sei emozioen artean aukeratu behar zuten. "Ezezaguna" aukera ez zegoen, emozioa argi ez zegoenerako ere ez. Esaldi guztiak deklaratzaileak ziren, galderazko bat izan ezik. Esaldien batez besteko luzera 8'61 berbakoa zen; laburrenak 4 eta luzeenak 14 berba zituen.

3.2. Ebaluaketa-protokoloa

Entzule bakoitzak azterketa bere aldetik egin zuen. Ahots zatiak ordenagailuko soinu-txartel arrunt baten bidez eta kalitate oneko aurikularrekin entzun zituzten. Entzuleei ez zaie entrenatzeko aukerarik eman azterketa saioaren aurretik, eta saio osoan ez dute jaso beren erantzunen gaineko baloraziorik. Formulario bakoitzeko 10 seinaleak izendatu beharra zuten, eta, behin beteta zutela, ezin ziren atzera bueltatu aldaketak egitera.

Guztira, 20 pertsonak hartu zuten parte (14 gizonek eta 6 emakumek) 10 eta 53 urte bitartekoak. Guztiek hitz egiten zuten ondo euskara batua, baina 11 baino ez ziren euskaldun zaharrak. Bereizi egin ditugu euskaldun zaharrak (A taldea) eta euskaldun berriak (B taldea), ebaluazioaren emaitzak aztertzeko orduan sailkatzeko irizpide bezala, beste batzuen artean.

4. Emaitzak

Azterketa subjektiboaren emaitzak 1. taulan agertzen dira. Matrizeko lerro bakoitzak aktoreak emandako benetako emozioa adierazten du; zutabeek, berriz, entzuleek zehaztutako emozioak adierazten dituzte. Balioak ehunekoak dira, eta emozio bakoitza letra bakarrarekin bereizten da: Haserrea (*Anger*, A), Beldurra (*Fear*, F), Harridura (*Surprise*, S), Nazka (*Disgust*, D), Poza (*Happiness*, H) eta Tristura (*Sadness*, X).

Taulan, *zehaztasuna* (*precision*, P) eta *memoria* (*recall*, R) parametroak ere jaso dira. Esleitutako emozioen zuzeneko kopurua neurtzen du zehaztasunak (identifikazio zuzenak / emozioarekin identifikatutako estimulu kopurua); memoriak, berriz, parametroa

neurtzeko emozioaren agerraldien zuzeneko kopurua hartzen du kontuan (Identifikazio zuzenak / Emozioaren estimulu kopurua).

Aktoreak	Entzuleak							
	A	F	S	D	H	X	P	R
A	81.5	2.5	5.5	9	-	1.5	0.78	0.82
F	0.5	64	3	-	1	31.5	0.68	0.64
S	6	2.5	73	1	17.5	-	0.80	0.73
D	15.5	4	3.5	67	2.5	-	0.86	0.67
H	0.5	0.5	5	-	94	-	0.81	0.94
X	-	20	1	0.5	1	77.5	0.66	0.78

1. Taula: ebaluaketa-prozesuaren nahasmen-matrizea

Argi ikusten da emozio guztiak ausazko atalasearen gainetik (%17) hautematen direla, corpusa esaldi semantikoki neutroz osatuta egoteak hautemate-prozesua oztopatu badezake ere. Batez besteko hautemate-maila %76.6 da, eta poza da ondoen hautematen den emozioa (%94) eta beldurra, aldiz, gaizkien (%64). Zehaztasun-parametrorik txikiena tristurak dauka, beldurra, nazka edo haserrea ziren estimuluetan aukeratu delako. Nazkaren memoria txikia eta zehaztasun handia azaltzeko, kontuan izan behar da gutxitan aukeratzen dela baina, aukeratzen denean, zehaztasun handiz egiten dela. Era berean, poza askotan hautatzen da; hortaz, memoria handia dauka, baina zehaztasun txikia.

2. eta 3. tauletan aktore bakoitzaren nahasmen matrizeak erakusgai dira. Pozak lortzen ditu emaitzarik onenak kasu bietan (%96 emakumearentzat eta %92 gizonarentzat). Okerren ezagutzen diren emozioak, berriz, ezberdinak dira: Beldurra (%61) Karolinarentzat eta Nazka (%59) Pellorentzat (baina zehaztasun altuarekin). Hala ere, bi emozioen arteko ezberdintasuna estua da eta biak txartoen ezagutzen diren emozioak dira bi datu-baseetan. Bataz besteko ezagutza portzentajea ere osa antzekoa da: %75.83 emakumearentzat eta pixka bat altuagoa gizonaren kasuan (%76.50).

Beldurra eta Tristura dira gehien elkar-nahasten diren emozioak (Beldurra Tristura bezala ezagutzen da %34 kasutan eta kontrako nahastea %20 alditan gertatzen da). Bi aktoreentzat nahasmen ohizkoena Beldurra eta Tristuraren arteko kategoriari dagokio, okerrak %19-%34 arteko tartean mugitzen direlarik. Nahasmen bera, Gaztelaniazko Interfaze datu-basean ere antzeman zen (Nogueiras eta beste., 2001).

Emakumea	Entzuleak							
	A	F	S	D	H	X	P	R
A	75	-	6	15	-	3	0.76	0.75
F	1	61	4	-	1	34	0.73	0.61
S	10	2	68	-	20	-	0.82	0.68
D	13	-	2	75	1	9	0.83	0.75
H	1	-	3	-	96	-	0.81	0.96
X	-	19	-	-	1	80	0.63	0.80

2. Taula: emakumezkoaren nahasmen-matrizea

Gizona	Entzuleak							
	A	F	S	D	H	X	P	R
A	88	4	5	3	-	-	0.81	0.88
F	1	67	2	-	1	29	0.64	0.67
S	2	3	78	2	15	-	0.79	0.78
D	18	8	5	59	4	6	0.91	0.59
H	-	1	7	-	92	-	0.81	0.92
X	-	21	2	1	1	75	0.68	0.75

3. Taula: gizonezkoaren nahasmen-matrizea

4.1. Entzuleen eragina emaitzetan

Entzuleen berezitasunek emozioen identifikazio emaitzetan nolabaiteko eraginik daukaten ikusteko, t-Student froga bat egin zen. Emakumeek %72,78eko identifikazio maila lortu dute (%68,37tik %77,18ko tartea %95eko konfiantzaz) eta gizonezkoek %77,62 (%74,74tik %80,50ko tartea %95eko konfiantzaz). Emaitza hauek adierazgarriak dirudite (t=1,80; p=0,071 > 0,0) baina ez %95eko konfiantza tartean (p=0,05). Euskara lehenengo hizkuntz bezala daukaten entzuleen (A taldea) eta bigarren hizkuntz bezala daukatenen (B taldea) arteko ezberdintasunak ere ez dira adierazgarriak (t=0,858; p=0,39 > 0,05): A taldeak %77,12 batez besteko identifikazio maila dauka eta B taldeak %75. Beraz, badirudi behin esaldiaren mezua ulertutakoan ez daukala eraginik euskara norbere lehenengo ala bigarren hizkuntza izateak.

Entzuleei ez zitzaien entrenamenduzko sesiorik eskaini probak egin aurretik. Beraz, entzuleen zehaztasuna ebaluaketaren hasieran eta amaieran ezberdina zen ala ez aztertu zen. Ebaluaketaren lehenengo erdian identifikazio maila %72,64koa da (%69,26tik %76,07ko tartea %95eko konfiantzaz) eta bigarren erdian &79,70koa (%76,26tik %83,07ko tartea %95eko konfiantzaz). Kasu honetan ebaluaketaren bigarren erdian lortutako emaitzen hobekuntza estatistikoki adierazgarria da %95eko konfiantzaz ($t=2,85$; $p=0,0044 < 0,05$). Bigarren erdialdean %7ko hobekuntza lortzen da identifikazio mailan (ia iraunkorra talde guztietan, gizonezkoan, emakumezkoan, A taldean eta B taldean). Emaitza hau ulergarria da, lehenengo seinaleetan entzulea erantzun bat aukeratzeko behartuta dagoelako, nahiz eta guztiz ziur ez egon, behartutako aukeraketa dela eta. Ebaluaketa aurrera joan ahala, entzuleek hizlariek emozio bakoitza nola espresatzen duten ikasten dute, eta beraz, errazago identifikatzen dituzte.

5. Ondorioak

Froga subjektiboaren emaitzak grabatutako emozio guztiak errekonozitu egiten direla erakusten du, aktore bientzako, eta ausazko atalasearen gainetik. Beraz, datu-base hau baliabide baliogarria da euskara batuan egiten den ahots emozioduna aztertu eta modelatzeko, eta dagoeneko hasita dagoen corpusean oinarritutako emoziodun ahotsa sintesirako sistema garatzeko.

6. Esker onean

Ebaluatutako datu-basea Eusko Jaurlaritzaren diru-laguntzarekin garatu da, ANHITZ programaren (ETORTEK06/114) eta MEC-en diru-laguntzarekin (TEC2006-13694-C03-02/TCM).

Egileek ebaluaketan parte hartu duten entzule guztiak eskertu nahi dituzte.

7. Aipamenak

Bulut, M., Narayanan, S., Syrdal, A. (2002). Expressive speech synthesis using a concatenative synthesizer, In *ICSLP* 2002, pp. 1265--1268

Cowie, R., Cornelius, R.R. (2003) Describing the Emotional States that Are Expressed in Speech, In *Speech Communication* 2003, 40(1,2) pp. 5--32

Heuft, B., Portele, T., Rauth, M. (1996). Emotions in Time Domain Synthesis, In *ICSLP 1996*, pp.1974--1977

Hunt, A., Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech data base, In *ICASSP 1996*, pp. 373-376.Erlbaum Associates, pp. 252--262

IIda, A., Campbell, N. Higuchi, F., Yasumura, M. (2003). A Corpus based speech synthesis system with emotion, In *Speech Communication*, 40, pp. 161--187

Montero, J. M., Gutiérrez-Arriola, J., Colás, J., Enríquez, E., & Pardo, J. M. (1999). Analysis and Modeling of Emotional Speech in Spanish, In *ICPhS* 1999, pp. 957--960

Murray, I.R. and Arnott, J.L. (1996). Synthesising emotions in speech: is it time to get excited?, In *ICSLP* 1996, pp. 1816--1819

Navas, E., Hernáez, I., Luengo, I. (2006). An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS, in *IEEE Transactions on audio, speech and language processing* 2006, vol. 14, n. 4, pp. 1117--1127.

Nogueiras, A., Moreno, A., Bonafonte, A., Mariño, J.B., (2001). Speech Emotion Recognition Using Hidden Markov Models, In *Proceedings of Eurospeech* 2001, pp. 2679--2682

Rank, E., Pirker, H. (2006) Generating Emotional Speech with a Concatenative Synthesizer, In *ICSLP* 98, Vol. 3, pp. 671--674

Saratxaga I, Navas E., Hernaez I., Luengo I., Sanchez, J. (2006). Korpusean oinarritutako sintesirako euskarazko datu base emoziodun baten diseinu eta grabaketa, In *Euskalingua, 9*

Schröder, M. (1999)- Can emotions be synthesized without controlling voice quality?, In *Phonus 4, Research Report of the Institute of Phonetics* 1999, Saarland University, pp. 37--55

Vroomen, J., Collier, R., Mozziconacci, S. J. L. (1993). Duration and Intonation in Emotional Speech, In *Eurospeech* 1993, Vol. 1, pp. 577--580.