Emaitzen post-prozesatze teknikak hizlarien diarizazio-sistema baterako

David Tavarez, Eva Navas, Daniel Erro, Ibon Saratxaga, Inma Hernaez

UPV / EHU
{david,eva,derro,ibon,inma}@aholab.ehu.es

Abstract

This paper presents the post-processing techniques designed to improve the results of a speaker diarization system. Three different techniques are proposed: refinement of speech vs. non speech segmentation, assimilation of short speech segments and fusion of clusters from the same speaker. These techniques have been implemented in a post-processing module that improves the result of the baseline system by 22.3 %. The same module has been applied to another speaker diarization system with a similar architecture to that of the baseline system with a DER improvement of 21% and to another one with a very different architecture where no improvement has been achieved. It has also been used with another database with an improvement of 17 %. These experiments prove the validity of the techniques developed.

Laburpena

Artikulu honetan hizlarien diarizazio sistemen emaitzak hobetzeko diseinatutako post-prozesatze teknikak aurkezten dira. Hiru teknika desberdin proposatu dira: ahostun/ahoskabe segmentazioa fintzea, ahots bolada laburrak asimilatzea, eta hizlari beraren klusterrak bateratzea.Teknika hauek post-prozesatze modulu batean inplementatu dira, oinarrizko sistemaren eginkortasuna %22.3 baten hobetuz. Modulubera beste hizlari diarizazio sistema baten aplikatu da, oinarrizkoaren antzeko arkitektura zuena, %21 DER hobekuntzaz eta arkitektura desberdineko hirugarren sistema baten, azken honetan hobekuntzarik gabe. Beste datu-base batekin erabilita, hobekuntza %17 izan da. Esperimentu hauek garatutako tekniken baliagarritasuna frogatzen dute.

Keywords: Speaker diarization, segmentation, rich transcription

Gako hitzak: Hizlarien diarizazioa, segmentazioa, trankribapen aberaztua

1. Sarrera

Audioaren diarizazioa jario bat bere audio-iturri espezifikoen arabera gune homogeneotan zatitzean datza (Cettolo, Vescovi, eta Rizzi, 2005). Iturrien banaketa audio mota izan daiteke (ahotsa, musika, atzealdeko zarata), hizlariaren nortasuna edo kanalaren ezaugarriak (Reynolds eta Torres-Carrasquillo, 2005). Hizlariaren diarizazioa audioaren diarizazioaren azpi-mota bat dela onar daiteke. Bertan audio grabaketa bat hizlariaren nortasunarekin homogeneoak diren zati desberdinetan zatitu behar da automatikoki, aurrez aurretik hizlari kopurua eta nortasunak ezagutu barik (Tranter eta Reynolds, 2006). Horretaz gain, hizlari bakoitza identifikatzea ere posible da, behar beste informazio edukiz gero. Hau betetzeko algoritmo desberdin batzuk konbinatu behar dira, helburu desberdinak dituztenak. Sistema gehienetan, gainera, sekuentzian abiarazi behar dira, hau da, bakoitza seinale osoan aplikatu behar da ondorengoarekin jarraitu baino lehen (Anguera, 2006). Ataza hauen artean ahotsaren detekzioa, txanda-aldaketen detekzioa, hizlarien taldekatzea eta audioaren birsegmentazioa daukagu (Anguera et al., 2012).

Garatutako algoritmoen baliagarritasuna aztertzeko, ebaluaketa kanpaina-lehiaketak antolatzen dira, NIST Rich Transcription edo Albayzin (Zelenák, Schulz, eta Hernando, 2010) adibidez. Kanpaina lehiaketa hauetan ikerketa talde desberdinek haien algoritmoen testak abiarazten dituzte datu-base berean, sistemen eraginkortasuna konparatu ahal izateko, sistemaren etapa bakoitzeko teknikarik egokiena topatzeko.

Aholab taldeak garatutako sistema Albayzin 2010 kanpainan aurkeztu zen (Luengo et al., 2010). Bere oinarria BIC-en oinarritutako (Chen eta Gopalakrishnan, 1998) txanda aldaketaren detektorearen inplementazio eraginkorrean datza; ahotsaren tarte ahostunak besterik ez dira erabiltzen, eta behetik goranzko metatze-clustering hierarkiko prozesu baten bidezko off-line hizlari taldekatzea (Tavarez et al., 2012). Sistema honek emaitza onak lortu zituen ebaluaketan, audioaren birsegmentazio zatirik ez bazeukan ere. Oinarrizko sistemaren emaitzak hobetzeko erabiltzen den post-prozesatze modulu bat garatzeko, sakonean aztertu dira agertutako erroreak eta estrategia desberdinak planteatu dira mota bakoitza hobetzeko. Artikulu honetan modulu hori aurkezten da, inplementatutako estrategiak eta lortutako emaitzak deskribatuz.

Artikuluaren bigarren atalean sistema garatzeko erabili den datu-basea deskribatzen da. hirugarrenean, agertutako erroreen analisia, eta laugarrenean, emaitzak hobetzeko proposatutako post-prozesatze teknikak. Bosgarren atalean garatutako esperimentuen berri ematen da, eta bukatzeko, seigarrenean, lanaren emaitzak deskribatzen dira.

2. Datu-basea

2010eko Albayzin ebaluaketa kanpainan 3/24 telebista katalanerako kanaletik grabatutako berrien programen ahozko datu-basea eman zen (Zelenák, Schulz, eta Hernando, 2010). Grabaketa UPCko TALP taldeak gauzatu zuen, eta etiketatzea, Verbio Technologies enpresak. Guztira 24 grabaketa edo sesio dira, eta bakoitzeko hizlari kopurua 30etik 250era. Datu basean 87 bat ordu daude, ondorengo banaketarekin: %37a ahots garbia, %5a musika, %15a ahotsa musikarekin, %40a ahotsa zaratarekin eta %3a bestelakorik, hau da, aurreko sailkapenean sartzen ez den guztia, zarata barne

Ebaluaketa kanpainarako datu base osoa bi zatitan banatu zen: 16 sesio entrenamendurako eta 8 frogak egiteko.

3. Erroreen analisia

Oinarrizko sistemak lortutako emaitzak NISTek emandako irizpideen arabera kalkulatu dira. Errorearen neurri nagusia diarizazio errore totala da (DER, overall Diarization Error Rate). Hau ondorengo erroreek osatzen dute: hizlarien ahotsa ez detektatzea (MST, Missed Speaker Time), ahoskabeko segmentuak hizlarien ahotza moduan (FAST, False Alarm Speaker Time) eta hizlariak txarto etiketatzea (SET, Speaker Error Time).

Azkeneko DER parametroa txikitzeko, oinarrizko sistemak egindako erroreen azterketa gauzatu da. Horretarako, lortutako denbora markak Albayzineko antolatzaileek emandakoekin konparatu dira, errore desberdinen izaera identifikatzeko, eta kasu bakoitza behar den moduan kudeatzeko estrategiak diseinatzeko.

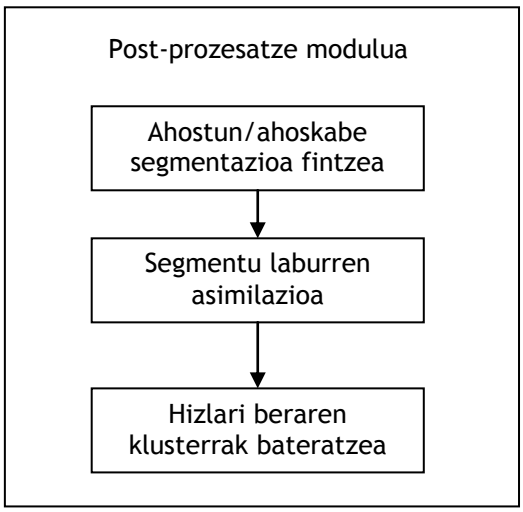
Oinarriko sistemak froga seinaleekin lortutako DER parametroa %30.11 da, eta bertan %2.8 MSTk sortzen du, %2.2 FARTek eta %25.1 SETek. Lortutako emaitzetan SETen garrantzia ikusita, honen motako erroreak sakonean aztertu dira, ondorengo egoerak topatuz:

- hizlarien taldekatze prozedurak okerreko hizlari bati esleitutako segmentu laburrak, segmentazioak benetan gertatu ez den aldaketa bat detektatzen duenean.
- hizlarien taldekatze prozedurak, berea sortu beharrean, aurretik sortutako klusterretan kokatutako agerraldi laburreko hizlarien segmentuak.
- taldekatze prozedurak kluster desberdinetan kokatzen dituen hizlari beraren agerraldi desberdinak, detektatutako hizlari kopurua handituz.

Bai MST eta bai FAST ahots-detekzio moduluaren okerreko portaerak eragiten ditu. Hau dela eta, aldendu behar ziren musika zati batzuk hizlari-aldaketa detekzio-modulura eta hizlari-taldekatze modulura pasatzen dira, normalean hizlari berri baten okerreko etiketa bat hartuz.

4. Post-prosezatze modulua

Oinarrizko diarizazio sistemak egindako erroreak aztertu ondoren, post-prozesatze modulu bat garatu da mota bakoitza era egokian prozesatzeko eta, horrela, azkeneko DER parametroa txikitzeko. Ondoren moduluaren zatiak deskribatzen dira: Ahostun/ahoskabe segmentazioa fintzea, segmentu laburren asimilatzea eta hizlari beraren klusterrak bateratzea, 1. Irudian ikusten denez.



1. Irudia: Post-prozesatze moduluaren eskema.

4.1. Lehenengo etapa: Ahostun/ahoskabe segmentazioa fintzea

Post-prozesatze moduluaren lehenengo etapak ahots-detekzio moduluak sortutako erroreak fintzea du helburu. Horretarako, lehenik eta behin, GMM (Gaussian Mixture Model) modelo bat entrenatzen da oinarrizko sistemak lortutako kluster bakoitzeko, eta beste bat isiluneentzako, 'bestelakoak' marka duten erreferentziazko entrenamendu sesioen datuekin. Ondoren, oinarrizko sistemak ahostun moduan markatutako segmentu bakoitzeko bi GMM modeloko Viterbiren segmentazio bat egiten da: isilunea izateko modeloa eta segmentu horretan jatorrian markatutakoa. Azkenik, kendu egiten dira 750 ms baino laburragoak diren isiluneak.

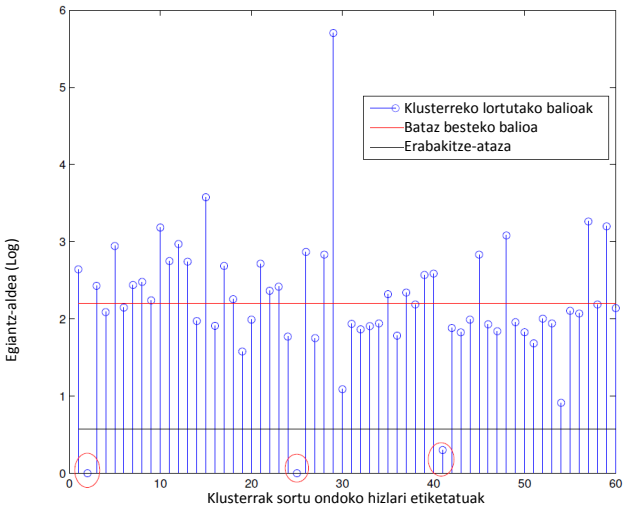
Horrela, isiluneak, musika, zarata, eta beste soinu gertakari batzuk ere 'ahoskabe' markatzen dira, FAST errorea txikituz. Baina era berean, kluster desberdinetan txarto sartutako beste hizlari batzuen agerraldiak, 'ahoskabe' markatuko omen dira. Holan MST handitu badaiteke ere, klusterren kalitatea hobetzen da, eta hau onuragarria izan daiteke ondorengo urratsetan

4.2. Bigarren etapa: Segmentu laburren asimilazioa

Post-prozesatze moduluaren bigarren etaparen helburua hizlari baten agerraldi luzea dagoenean agertzen diren segmentu laburrak kentzea da. Horretarako, lehen, Marka okerra edukitzeko susmagarriak diren segmentuak kokatzen dira, bere iraupena eta aurrekoarena kontuan hartuta. Iraupen horien balioak eskuz jarri dira garapenean etapa honen eginkortasuna optimizatzeko. Ondoren, GMM modelo bat entrenatzen da (G_x), segmentu susmagarriaren etiketaren datu-baseko hizlari beraren eskura dauden datu guztiak erabiliz, segmentu susmagarrikoak izan ezik. Aurreko segmentuan, beste GMM baterako (G_a) antzeko entrenamendua egiten da, datu guztiak erabiliz. Azkenik, segmentu susmagarria hobe modelatzen bada G_a erabiliz G_x erabiliz baino, ondoko hizlariaren klusterrean asimilatzen da.

4.3. Hirugarren etapa: klusterrak bateratzea

Hirugarren etaparen helburua hizlari beraren klusterrak bateratzea da. Horretarako, lehenengoa GMM modeloak sortzen dira oinarrizko sistemak detektatutako kluster guztientzat, bakoitzeko 60s erabiliz. Ondoren, kluster bakoitzeko segmentu bat ateratzen da (GMM modeloa kalkulatzeko erabili ez dan informazioaren beste 60 segundo) eta horren egiantz aldea kalkulatzen da jatorrian markatutako modeloaren eta beste guztiekin. Aldean balio txikiak lortzen badira, klusterren informazioa antzekoa izango da, eta informazio hori erabilita kluster desberdin batzuk bakar baten barruan bateratzea erabaki daiteke. Bateratzeko ataza balioa enpirikoki ezarri da etapa honen emaitzak garapenean optimizatuz.



2. Irudia: 2. klusterrean lortutako egiantz-aldeak

2. Irudian bi klusterrerako lortu diren aldeak ikusten dira, eskala logaritmikoan. Kasu honetan, 25 eta 41 klusterren balioak oso txikiak dira, eta kluster bakar baten bateratuko dira.

5. Gauzatutako esperimentuak

Post-prozesatze modulua frogatzeko esperimentu batzuk gauzatu dira. Lehenengi, hiru etapak aplikatu dira oinarrizko sistemaren emaitzetan, bai konfigurazio parametroak optimizatzeko erabili diren entrenamendu sesioetan, eta baita test sesioetan ere. Ondoren, garatutako moduluaren orokortzeko gaitasuna egiaztatzeko, beste diarizazio-sistema batzuekin aplikatu da, eta baita diarizazio-sistema berarekin baina datu-base desberdin baten aplikatua.

5.1. Esperimentuak oinarrizko sisteman

1. eta 2. Taulan post-prozesatze-modulua oinarrizko sistemak emandako etiketetan aplikatzearen emaitzak erakusten dira. 1.an, entrenamendu sesio bakoitzerako lortutako DER balioa, bai oinarrizko sistemaren marketarako eta baita etapa bakoitzaren irteeran ere. Azken linean datu-basearen entrenamenduan lortutako DER balioa erakusten da.

S	DER	E1	E2	E3
1	22.17%	21.83%	21.54%	29.49%
2	24.58%	24.46%	24.38%	13.10%
3	23.10%	23.01%	22.92%	18.11%
4	27.47%	27.67%	27.50%	27.50%
5	14.15%	12.94%	12.93%	9.89%
6	21.22%	21.40%	21.32%	16.21%
7	24.84%	24.86%	24.89%	27.72%
8	27.26%	27.38%	27.38%	19.90%
9	28.92%	28.28%	28.60%	26.80%
10	34.75%	34.54%	35.26%	26.80%
11	27.94%	27.70%	27.90%	15.91%
12	27.42%	27.22%	27.22%	25.54%
13	31.92%	32.13%	31.86%	32.34%
14	41.16%	40.87%	41.00%	25.84%
15	32.50%	32.73%	32.62%	27.25%
16	32.06%	32.09%	32.02%	24.18%
All	28.25%	28.14%	28.16%	23.33%

1. Taula: Post-prozesatze-etapen emaitzak garapen-sesioetan

2. taulan test sesioetarako lortutako emaitza erakusten da. Aurreko kasuaren moduan, DER balioak agertzen dira, bai oinarrizko sistemaren jatorrizko marketan eta baita etapa bakoitzerako irteeran, sesio bakoitzeko, eta datu-base osorako.

S	DER	E1	E2	E3
1	34.92%	34.69%	34.62%	33.97%
2	31.35%	30.77%	30.82%	19.36%
3	27.14%	27.47%	27.46%	21.05%
4	34.72%	34.57%	34.76%	25.52%
5	34.20%	34.24%	34.14%	18.38%
6	33.06%	33.36%	33.33%	29.81%
7	24.92%	25.05%	25.18%	19.48%
8	22.99%	22.96%	23.11%	17.76%
All	30.11%	30.08%	30.13%	23.40%

2. Taula: Post-prozesatze-etapen emaitzak test-sesioetan

Aurreko tauletan ikusten denez, errorea sesio gehienetan gutxitu da, bai entrenamenduan eta baita testean ere. Post-prozesatze modulua aplikatzean, %17.4ko murrizketa lortu da DER baliorako entrenamenduko sesioetan eta %22.3koa testekoetan, proposatutako etapen baliagarritasuna frogatuz. Lehenengo bi etapetan errorea ez da asko murrizten, baina klusterren kalitatea hobetzen da eta erabilgarria da hori hirugarren etaparen funtzionamendu onerako.

5.2. Esperimentuak beste diarizazio-sistema batzuetan

Oinarrizko sisteman ondo dabilela egiaztatuta, garatutako modulua beste diarizazio-sistema batzuetan erabilgarria den frogatzeko arkitektura desberdinetako sistemetan erabiltzeko esperimentuak gauzatu dira. Horretarako Albayzin 2010 datu-baserako bi sistema desberdinek lortutako markak erabili dira. Kasu guztietan oinarrizko sistemaren entrenamendu-etapan lortutako konfigurazio-parametroak mantendu dira. Lehenik, onlineko antzeko sistema batek kalkulatutako markak erabili dira (Luengo et al. 2010). 3. taulan post-prozesatze-modulua sistema honetan aplikatzean lortutako emaitzak aurkezten dira.

S	DER	E1	E2	E3
E	26.77%	26.72%	26.76%	21.38%
P	27.17%	27.18%	27.32%	21.45%

3. Taula: Post-prozesatze-etapen emaitzak onlineko sisteman

Ikusten denez, emaitzak hobetu dira, oinarrizko diarizazio-sisteman gertatu den antzera. Post-prozesatze-moduluak sistemaren online funtzionamendua deuseztatzen du, baina DER balioa %20.1 hobetuz entrenamenduan eta %21 testean.

Ondoren, Vigoko Unibertsitateko GTM taldeak garatutako diarizazio sistemak (Docio, Lopez, eta Garcia, 2010) emandako marketan aplikatu da post-prozesatze modulua. Esperimentuaren helburua modulua sistema oso desberdin baten frogatzea da, bere errore bereziak aztertu barik. 4. taulan lortutako emaitzak aurkezten dira.

S	DER	E1	E2	E3
E	25.48%	25.54%	25.31%	25.91%
T	25.62%	25.62%	25.26%	27.00%

4. Taula: Post-prozesatze etapen emaitzak GTM sisteman

Ikusten denez, errorea ez da hobetu. Lehenengo bi etapetan, hain espezifikoak ez direnez, aurreko kasuetakoen antzeko emaitzak lortu dira. Baina bateratze etapa sesio batzuetan hobetzeko baliogarria izan bada ere, beste batzuetan handitu egin da errorea. Etapa hau oinarrizko sistemaren erroreak aztertuz garatu denez, posible da arkitektura desberdina duen sistema baten errore berdinak ez agertzea, beraz, errorea handiagoa eginez modulua aplikatzean. Hala ere, konfigurazio parametroak ez dira aldatu, eta horien optimizazio baten ostean errorea murriztea posible litzateke.

5.3. Esperimentuak beste datu-baseekin

Bukatzeko, post-prozesatze-moduluaren erabilitako datu-basearekiko independentzia ikertzea proposatzen da. Horretarako, oinarrizko diarizazio-sistema erabili da Euskal Irrati Telebistaren seinaleez osatutako datu-base txiki bat markatzeko. Seinaleak 2010ean emititutako euskarazko eta gaztelaniazko berrien bilduma batetik hartu dira. Audio fitxategietan, berriak ematen dituzten kazetariez gain (fitxategi desberdinetan agertzen direnak), elkarrizketak ere badaude, eta bikoiztutako hizkera, jatorrizko ahotsaren gainean. Audio klip horietariko batzuk kateatu dira ezaugarri desberdinetako bi sesio osatzeko. Lehenengoa, 20 minutukoa, 9 hizlari ditu agerraldi luzeetan, zarata txikiarekin. Honek hasiera baten lehenengo moduluaren funtzionamendua hobetzen du. Bigarren sesioak, 25 minutukoa, 40 hizlari ditu, zaratarekin segmentu batzuetan eta musikarekin beste batzuetan; horrekin, hirugarren etaparen funtzionamendua da nagusi. Erreferentzia bat edukitzeko, eskuz markatu dira seinaleak. Lortutako emaitza 5. taulan ikus daiteke.

S	DER	E1	E2	E3
	35.65%	34.65%	32.30%	32.30%
	26.83%	26.78%	26.78%	20.53%
All	30.26%	29.84%	28.93%	25.11%

5. Taula: Post-prozesatze etapen emaitzak oinarrizko sisteman, EiTB datu-basearekin

Ikusten denez, emaitzak Albayzin2010 datu-basearekin lortutakoen antzekoak dira. Kasu honetan DERa %17 murriztu da. Emaitza hauekin,

garatutako modulua ezaugarri desberdinetako datu-baseekin erabilgarria dela frogatzen da.

6. Ondorioak

Diarizazio-sistema baten emaitzak hobetzeko post-prozesatze teknika desberdinak deskribatu dira. Hiru teknika proposatu dira sistemak egindako errore mota bakoitza tratatzeko: ahostun/ahoskabe segmentazioa fintzea, segmentu laburren asimilatzea eta hizlari beraren klusterrak bateratzea. Inplementatu egin dira teknika hauek eta optimizatu dira konfigurazio parametroak datu-basearen garapena zatia erabiliz. Post-prozesatze-modulu bat garatu da teknika hauek oinarrizko diarizazio-sisteman aplikatzeko, froga sesioetan %22.3ko hobekuntza lortuz. Garatutako modulua beste diarizazio-sistema baten, oinarrizkoaren antzekoa, aplikatu da, %21eko hobekuntzaz eta sistema desberdin baten, hobekuntza lortu barik, baina konfigurazio parametroak aldatuz hobetzeko aukeraz. Azkenik, garatutako modulua beste datu-base batzuetarako baliogarria den egiaztatu da; %17ko hobekuntza lortu da Euskal Irrati Telebistaren grabaketak erabiliz.

7. Esker onean

Egileok Iker Luengori eskertu nahi diogu oinarrizko diarizazio sistemaren garapena, Vigoko Unibertsitateko GTM taldeari bere diarizazio-sistemaren emaitzak erabiltzea errazteagatik, eta Euskal Irrati Telebistari bere grabaketen erabilera onartzeagatik.

Lan honek Euskal Herriko Unibertsitateko diru-laguntza jaso du (Ayudas para la Formación de Personal Investigador), Eusko Jaurlaritzakoa (BerbaTek proiektua, IE09-262) eta Zientzia eta Berrikuntza Ministeriokoa (Buceador Proiektua, TEC2009-14094-C04-02).

8. Aipamenak

Anguera, X. (2006). *Robust Speaker Diarization for meetings*. Ph.D. tesis. Barcelona: Universitat Politecnica de Catalunya.

Anguera, X.; Bozonnet, S.; Evans, N.; Fredouille, C.; Friedland, G. eta Vinyals, O. (2012). Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech and Language Processing*, 20(2):356-370.

Cettolo, M.; Vescovi, M. eta Rizzi, R. (2005). Evaluation of BIC-based algorithms for audio segmentation. *Computer Speech & Language*, 19(2):147-170.

Chen, S. S. eta Gopalakrishnan, P. S. (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. *DARPA speech recognition workshop*. 6, 127-132.

Docio, L.; Lopez, P. eta Garcia, C. (2010). The uvigo-gtm speaker diarization system for the Albayzin'10 evaluation. *VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop, (FALA 2010)*, 401-404. Vigo.

Luengo, I.; Navas, E.; Saratxaga, I.; Hernáez, I. eta Erro, D. (2010). AhoLab Speaker Diarisation System for Albayzin 2010. *VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop, (FALA 2010)*. 393-396. Vigo.

Reynolds, D. eta Torres-Carrasquillo, P. (2005). Approaches and applications of audio diarization. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 953-956.

Tavarez, D.; Navas, E.; Erro, D. eta Saratxaga, I. (2012). Strategies to Improve a Speaker Diarisation Tool. *LREC 2012*. 4117-4121. Estambul.

Tranter, S. E. eta Reynolds, D. (2006). An overview of automatic speaker diarization systems. *IEEE Trans. on Audio, Speech and Laguage processing*, 14(5):1557-1565.

Zelenák, M.; Schulz, H, eta Hernando, J. (2010). Albayzin 2010 evaluation campaign: Speaker diarization. *VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop, (FALA 2010)*. 301-304. Vigo.