Razvoj postopka diarizacije govorcev z algoritmi strojnega učenja

Marko Katrašnik
Institut „Jožef Stefan“
Jamova cesta 39
1000 Ljubljana, Slovenia
marko.katrasnik@gmail.com

Junoš Lukan, Mitja Luštrek
Institut „Jožef Stefan“
Mednarodna podiplomska šola Jožefa Stefana
Jamova cesta 39
1000 Ljubljana, Slovenia
{junos.lukan, mitja.lustrek}@ijs.si

Vitomir Štruc
Fakulteta za elektrotehniko,
Univerza v Ljubljani
Tržaška cesta 25
1000 Ljubljana, Slovenia
vitomir.struc@fe.uni-lj.si

POVZETEK

Pomemben del konteksta uporabnikov mobilnih telefonov so njihove žive socialne interakcije. Zaznamo jih lahko s pomočjo mikrofona, pri čemer lahko zaznamo prisotnost človeškega govora, ugotavljamo število govorcev in določamo, kdaj je govoril kateri od govorcev, čemur pravimo diarizacija. V članku sta predstavljena detektor govora in širši postopek diarizacije govorcev, za katera smo uporabili že obstoječa orodja in jih prilagodili za široko uporabnost na posnetkih z mobilnih telefonov. Za zaznavo govora smo uporabili logistično regresijo, ki je računsko nezahteven algoritem, hkrati pa je imel visoko točnost, skoraj 90 %, v različnih akustičnih okoljih. Za razvoj ostalih korakov diarizacije smo uporabili že obstoječe zbirke podatkov, posneli pa smo tudi majhno lastno zbirko. Za snemanje smo uporabili telefone različnih proizvajalcev in na ta način preverili robustnost našega postopka v primerjavi z že razvitimi metodami zaznavanja govora in diarizacije govorcev. V kontroliranih pogojih je postopek deloval primerljivo z že obstoječimi, na posnetkih iz vsakdanjega delovnega okolja pa je dosegel izrazito boljše rezultate.

1. UVOD

V okviru projekta na temo stresa na delovnem mestu (angl. Stress At Work project, StrAW) [3] želimo analizirati in opisati odnose med izkušnjami psihosocialnega stresa v delovnem okolju, vsakdanjimi aktivnostmi in dogodki na delu in fiziološkimi signali [16] ter vedenjskimi vzorci, ki jih lahko zaznamo avtomatsko s pomočjo tehnologije. Eden od pomembnih virov podatkov, ki jih lahko uporabimo v ta namen so senzorski podatki in spremljanje interakcije z mobilnim telefonom. S temi podatki je mogoče prepoznati pomemben del uporabnikovega konteksta, kot so socialne interakcije in pogovori. Tega problema smo se lotili v pričujočem delu in že objavljeni diplomski nalogi [11].

V prispevku je predstavljen postopek za diarizacijo govorcev. To je proces označevanja posnetka z informacijo o tem, kateremu govorcu pripadajo določeni segmenti v posnetku. V splošnem zajema tri glavne korake: zaznavanje človeškega govora, iskanje mej med deli posnetka, med katerimi so govorili različni govorci, ter združevanje teh segmentov glede na identiteto govorca. Ker smo želeli večji nadzor nad delovanjem postopka v različnih akustičnih pogojih, smo se prvi komponenti, detekciji govora, posvetili ločeno od ostalih.

Za zaznavanje govora se najpogosteje uporablja pristop z mešanicami Gaussovih porazdelitev (angl. Gaussian mixture models, GMM) [17], za optimizacijo telefonskih klicev preko internetnega protokola pa je standardiziran pristop z uporabo statističnega modeliranja [2]. V zadnjem času se za ta namen pogosto uporabljajo tudi nevronske mreže (npr. [10]).

Diarizacija govorcev se v literaturi (dober pregled je v [1]) največkrat izvaja na posnetkih novic in sestankov, pri čemer imajo uporabljeni posnetki ugodne akustične značilnosti. Pri tem nekateri pristopi izkoristijo snemanje z več mikrofoni in s pomočjo različnih zvočnih tokov izboljšajo točnost določanja govorcev.

Cilj našega dela je bil prilagoditev postopka za uporabo na posnetkih iz vsakdanjega življenja, pri katerih bi za snemanje uporabili mikrofon mobilnega telefona.

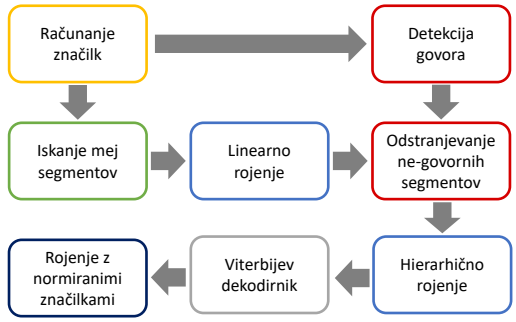
2. METODE

2.1 Postopek diarizacije govorcev

Za zaznavanje govora (angl. voice activity detection, VAD) so uporabne tako značilke v časovni (na primer najvišje in najnižje vrednosti signala) kot v frekvenčni domeni (na primer ploščatost spektra). Poleg teh se za opis človeškega glasu uporabljajo tudi specializirane značilke, kot so kepstralni koeficienti melodičnega spektra (angl. mel-frequency cepstral coefficients, MFCC) in zaznavni kepstralni koeficienti linearne napovedi z relativno spektralno transformacijo (angl. RASTA perceptual linear prediction coefficients, RASTA PLP-CC) [8]. Obe vrsti značilk delujeta v kepstralni domeni, v katero pridemo z inverzno Fourierevo transformacijo logaritma spektra signala, obenem pa z različnimi transformacijami poskušajo posnemati odziv človeškega sluha na zvok.

V našem postopku diarizacije govorcev, prikazanem na sliki 1, smo najprej izračunali značilke, ki smo jih uporabili v ostalih komponentah diarizacije govorcev. Izračunu značilk je sledilo iskanje mej med segmenti govorcev. Združevanje segmentov istega govorca se v procesu diarizacije ponovi večkrat na različne načine (modro obrobljeni koraki na sliki 1). Najprej smo z linearnim rojenjem (angl. linear clustering) združili le zaporedne segmente, ki pripadajo istemu govorcu. Nato smo na podlagi detektorja govora, ki je bil predstavljen v

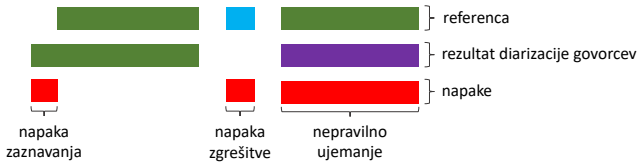
prejšnjem koraku, odstranili ne-govorne segmente in s hierarhičnim rojenjem že združili večino, tudi nezaporednih, segmentov istega govorca. Viterbijev dekodirnik dodatno izboljša začetno segmentacijo s pomočjo prikritih modelov Markova. Zadnji korak pa je bilo hierarhično rojenje z normiranimi značilkami, ki združi še roje istih govorcev, ki so bili do tega koraka ločeni zaradi različnega akustičnega ozadja.



Slika 1: Koraki v postopku diarizacije govorcev.

2.2 Vrednotenje rezultatov

Zaznavanje govora smo vrednotili s pomočjo ustaljenih mer: točnosti (angl. accuracy) in priklica (angl. recall). Poleg teh dveh mer pa smo za vrednotenje diarizacije uporabljali tudi mero napake med ujemanjem referenčnih in samodejno pridobljenih in označenih segmentov, krajše mero DER (angl. diarization error rate). Izračunamo jo kot delež napačno klasificiranih delov posnetka, kjer so možne tri različne napake (slika 2): napaka zaznavanja (ne-govor označen kot govor), napaka zgrešitve (govor označen kot ne-govor) in nepravilno ujemanje (klasificiran napačen govorec). Od običajne točnosti se razlikuje po tem, da delež izračunamo le glede na dele posnetka, v katerih je bil dejansko prisoten govor.



Slika 2: Pri vrednotenju diarizacije upoštevamo tri različne vrste napak. Njihov časovni delež v delih posnetka, v katerih je prisoten govor, je mera DER.

3. REZULTATI

3.1 Orodja

Orodje „openSMILE“ [9] omogoča izračun vseh vrst akustičnih značilk, omenjenih v prejšnjem podpoglavju. V našem delu smo za njihov izračun uporabili okna z dolžino 25 ms in korakom 10 ms. Omeniti velja, da „openSMILE“ ponuja tudi že prednaučeno povratno nevronsko mrežo z dolgim kratkoročnim spominom (angl. long short-term memory recurrent neural network, LSTM-RNN) za detekcijo človeškega glasu. Izhod te nevronske mreže smo pri našem detektorju govora uporabili kot značilko in kot referenco pri primerjavi točnosti drugih algoritmov za zaznavanje govora.

Glavni del diarizacije smo opravili z orodjem „LIUM SpkDiarization“ [14]. Poleg prilagoditve parametrov tega orodja pa smo zamenjali prvo komponento zaznavanje govora, ki smo jo implementirali s pomočjo orodja za strojno učenje „Weka“ [6].

3.2 Podatkovne zbirke

Pri razvoju postopka diarizacije govorcev smo uporabljali tri podatkovne zbirke: eno za učenje in validacijo zaznavanja govora, drugo za nastavljanje hiperparametrov ostalih komponent diarizacije in tretjo za končno testiranje postopka in primerjavo z drugimi metodami.

Kot učno množico za prvi korak zaznavanja govora smo uporabili „VAD-toolkit“ [12]. Gre za pogovore dveh korejskih govorcev, posnetih s telefonom Samsung Galaxy S8 v štirih različnih okoljih z različnim hrupom v ozadju: v sobi, v parku, na avtobusni postaji in na gradbišču. Posnetki so označeni z deli govora in ne-govora (binarne oznake), pri čemer govor vsebuje približno tretjina celotnega trajanja posnetkov.

Za razvoj ostalih komponent diarizacije smo uporabljali podatkovno množico „AMI Corpus“ [5], ki vsebuje označene posnetke sestankov v angleškem jeziku. Izbrali smo tri posnetke, v katerih je sodelovalo od 3 do 5 govorcev s trajanjem od 19 min do 66 min. Da bi posnemali različne akustične pogoje, smo jim umetno dodali šum s pomočjo orodja „Audio Degradation Toolbox“ [13]. Dodali smo jim posnetek z ulice z razmerjem med signalom in šumom 15 dB in posnetek iz gostilne z razmerjem 20 dB.

Kot testno množico za ovrednotenje celotnega postopka diarizacije in primerjavo z drugimi obstoječimi orodji smo uporabili drugo podmnožico zbirke „AMI Corpus“, dodatnih 12 posnetkov. Poleg tega smo posneli še tri lastne posnetke sestankov oziroma pogovora, za kar smo uporabili štiri različne pametne telefone: Huawei P Smart, Motorola Moto X, Samsung Galaxy S6 in Nokia 6.

3.3 Zaznavanje govora

Za zaznavanje govora smo najprej izbrali najbolj primeren algoritem strojnega učenja. Algoritme, ki smo jih preizkusili, prikazuje tabela 1 in so natančneje opisani v [6]. Z vidika točnosti se je najbolje izkazala logistična regresija. Kljub temu, da imajo nekatere druge metode za govor boljši priklic, smo v nadaljevanju izbrali ravno to metodo, saj je tudi relativno računsko nezahtevna.

Tabela 1: Primerjava uspešnosti različnih algoritmov za detekcijo govora pri uporabi vseh izračunanih značilk. Prikazani so priklic za razreda govora (priklic_g) in ne-govora (priklic_ng) ter skupna točnost.

Algoritem	priklic_g	priklic_ng	točnost
Logistična regresija	93,7	97,1	96,1
SVM	94,4	96,8	96,1
Večplastni perceptron	95,6	95,5	95,5
Naključni gozd	93,8	95,4	94,8
AdaBoost	93,0	94,4	94,1
KNN	87,1	93,6	91,4
J48	90,0	91,1	90,7
Naivni Bayes	60,8	93,6	84,1

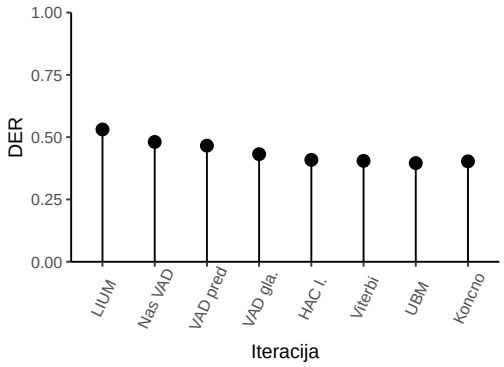
V naslednjem koraku smo izbrali najboljše značilke, saj smo sprva uporabljali vse smiselne, ki jih lahko izračunamo z orod-

jem „openSMILE“. V ta namen smo uporabili več metod: izboljšan postopek Relief (angl. ReliefF), izbor na podlagi korelacij (angl. correlation feature selection, CFS) in metodo z ovojnico (angl. wrapper) [6]. S tem smo nabor značilk zmanjšali s 104 na 54 značilk. Ohranili smo najvišjo, najnižjo in absolutno najvišjo vrednost oknjenega signala, ploščatost spektra, 13 značilk MFCC, 18 RASTA PLP-CC koeficientov in njihove koeficiente delta regresije ter izhod povratne nevronske mreže. Pri tem je točnost logistične regresije ostala enaka, priklic govora pa se je nekoliko izboljšal.

Kot omenjeno, smo izhod nevronske mreže pri našem detektorju govora uporabili kot značilko. Izkazalo se je, da je to v našem postopku najpomembnejša značilka, vendar vključitev ostalih poveča točnost s 86,6 % na 96,1 %.

3.4 Diarizacija govorcev

Komponente diarizacije govorcev smo spreminjali iterativno. Začeli smo z osnovnim postopkom LIUM, zamenjali detektor govora za opisanega v podpoglavju 3.3, nato pa spreminjali še hiperparametre. Slika 3 prikazuje, kako se je spreminjala napaka DER z zaporednimi spremembami postopka diarizacije.



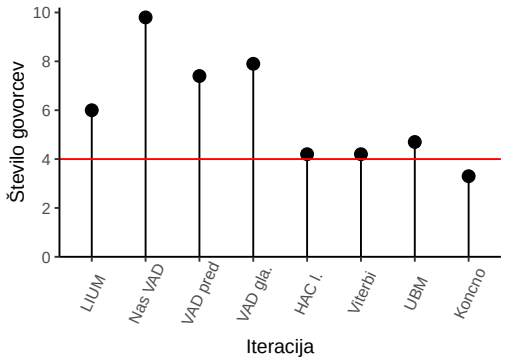
Slika 3: Spreminjanje napake DER skozi iterativne izboljšave postopka diarizacije. Največja, petodstotna, sprememba (označena z „Nas VAD“) je posledica zamenjave privzetega detektorja govora za lastnega.

V splošnem smo izbrali spremembe, ki so postopek izboljšale, tako da so zmanjšale mero DER. Največjo izboljšavo smo dosegli z zamenjavo detektorja govora („Nas VAD“), dodatno smo izboljšali delovanje postopka z odstranjevanjem negovornih segmentov v zgodnejšem koraku („VAD pred“) in nastavitvijo parametrov glajenja rezultatov detektorja govora („VAD gla.“).

Drugi korak (iteracija, označena s „HAC l“.), ki se je izkazal za pomembnega, je bila sprememba parametra v hierarhičnem rojenju (angl. hierarchical agglomerative clustering, HAC). Rojenje smo naredili bolj agresivno, s tem pa zmanjšali končno število rojev oziroma govorcev. Učinek te spremembe je še bolj izrazito opazen na sliki 4, kjer se je število govorcev močno približalo pravemu.

Nato smo prilagodili parametre Viterbijevega dekodirnika, namen katerega je, da izboljša začetno segmentacijo.

V zadnjih dveh korakih, označenih z „UBM“ in „Koncno“, smo spremenili še uporabo prednaučenega splošnega modela



Slika 4: Število zaznanih govorcev se je spreminjalo z vsako iteracijo postopka diarizacije.

govora (angl. universal background model, UBM). Ta model se uporablja za naknadni postopek rojenja, vendar bi ga bilo treba prilagoditi na uporabljeni podatkovni množici. Namesto tega smo komponento nadomestili s preprostejšim hierarhičnim rojenjem, na račun povečanja splošnosti pa se je z zadnjo spremembo nekoliko poslabšala napaka DER (slika 3) in napaka ocene števila govorcev (slika 4).

3.5 Primerjava rezultatov

Končno različico postopka, ki smo ga razvili, smo primerjali z nespremenjeno metodo LIUM ter drugo, v literaturi pogosto uporabljeno metodo diarizacije ALIZE [4].

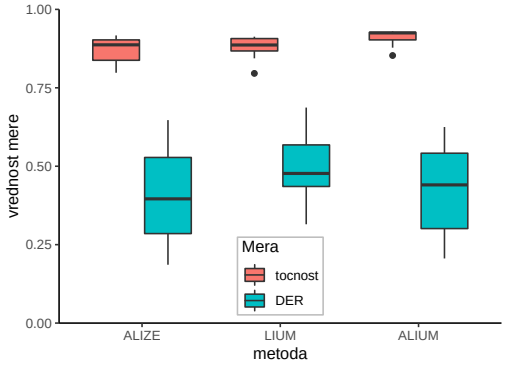
Slika 5 prikazuje točnost detektorja govora in napako DER našega postopka diarizacije (označenega z „ALIUM“) v primerjavi z drugima dvema postopkoma iz literature na posnetkih iz „AMI Corpus“. S spremembami, opisanimi v prejšnjem poglavju, smo izboljšali točnost zaznavanja govora: kot izračunano z analizo variance, se metode razlikujejo statistično pomembno ($\chi^2(2) = 32{,}8$, $p < 0{,}001$). Pri tem se je točnost izboljšala predvsem zaradi priklica razreda govora, medtem ko je priklic razreda ne-govora dejansko slabši kot pri postopku ALIZE. Po drugi strani se metode razlikujejo tudi v napaki DER ($\chi^2(2) = 14{,}8$, $p < 0{,}001$), vendar je bil glede na to mero najboljši postopek ALIZE.

Iste tri metode smo primerjali tudi na treh lastnih posnetkih. V tem primeru se je naš postopek obnesel mnogo bolje tudi z vidika napake DER, saj je napaka v povprečju znašala DER = 0,341, medtem ko je bila pri nespremenjenem postopku LIUM DER = 0,828 in ALIZE DER = 0,794.

Delovanje metode diarizacije smo primerjali tudi preko posnetkov z različnih mobilnih telefonov. Točnost zaznave govora se je med posnetki z različnih telefonov razlikovala za 1 %, napaka DER pa za največ 3,5 %.

4. ZAKLJUČKI

Naš postopek diarizacije govorcev, še posebej pa korak zaznavanja govora, deluje zanesljivo na raznovrstnih posnetkih. Že obstoječe metode v dobro kontroliranih pogojih in sodeč po nekaterih merah sicer dajejo nekoliko boljše rezultate, vendar kaže, da je v našem delu prilagojen postopek širše uporaben. To se je pokazalo v doslednih rezultatih na različnih zbirkah podatkov, in mnogo boljšimi rezultati na zbirki, posneti z mobilnimi telefoni različnih proizvajalcev.



Slika 5: Točnost detektorja govora, opisanega v podpoglavju 3.3, in napaka DER pri diarizaciji s postopkom, opisanim v podpoglavju 3.4 in označenim z „ALIUM“. Ti dve meri smo primerjali z originalnim LIUM postopkom ter postopkom ALIZE na dvanajstih posnetkih „AMI Corpus“.

Druga pomembna prednost razvitega postopka je njena nizka računska zahtevnost. Logistična regresija je preprost algoritem, zaradi česar je detekcijo govora mogoče izvajati v realnem času na sodobnih mobilnih napravah z operacijskim sistemom Android. Večino računske zahtevnosti tako prinesejo nadaljnji koraki diarizacije: čas računanja je linearno naraščal z dolžino posnetka, čas izvedbe celotnega postopka pa je bil na prenosnem računalniku v povprečju 20-krat krajši od trajanja posnetka.

Na posnetkih iz „AMI Corpus“ se je metoda ALIZE brez prilagoditev glede na napako DER izkazala bolje od naše (slika 3). Ti posnetki imajo v primerjavi s posnetki s telefonov iz vsakdanjega življenja ugodno razmerje med signalom in šumom in tudi večjo glasnost. Naš postopek bi tako morda lahko izboljšali s predobdelavo posnetkov, predvsem z normiranjem glasnosti in odstranjevanjem šuma. Druga izboljšava, ki je bila v literaturi že preizkušena [7], je iskanje mej med segmenti govorcev od začetka do konca posnetka in nato še v nasprotni smeri.

Za implementacijo postopka za samodejno zaznavo pogovorov s pomočjo pametnih telefonov pa bomo v prihodnje pozornost poleg točnosti namenili tudi porabi energije. Kljub računski preprostosti detektorja govora je namreč že samo snemanje z mikrofonom veliko breme za baterijo. Govor je mogoče zaznati tudi brez neprestanega snemanja, na primer z uporabo adaptivnega vzorčenja [15].

Literatura

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370, 2012.

[2] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit. ITU-T recommendation G. 729 Annex B: A silence compression scheme for use with G. 729 optimized for V. 70 digital simultaneous voice and data applications. *IEEE Communications Magazine*, 35(9):64–73, 1997.

[3] L. Bolliger, J. Lukan, M. Luštrek, D. D. Bacquer, and E. Clays. Disentangling the sources and context of daily work stress: Study protocol of a comprehensive realtime modelling study using portable devices. 2018. Employability 21.

[4] J.-F. Bonastre, F. Wils, and S. Meignier. ALIZE, a free toolkit for speaker recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I/737–I/740. IEEE, 2005.

[5] J. Carletta. Announcing the AMI meeting corpus. *The ELRA Newsletter*, 11(1):3–5, 2006.

[6] F. Eibe, M. A. Hall, and I. H. Witten. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, fourth edition, 2016.

[7] E. El-Khoury, C. Senac, and J. Pinquier. Improved speaker diarization system for meetings. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4097–4100. IEEE, 2009.

[8] F. Eyben. *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*. Springer International Publishing, 2016.

[9] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM, 2013.

[10] T. Hughes and K. Mierle. Recurrent neural networks for voice activity detection. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7378–7382. IEEE, 2013.

[11] M. Katrašnik. *Razvoj postopka diarizacije govorcev z algoritmi strojnega učenja*. Diplomsko delo, Univerza v Ljubljani, Fakulteta za elektrotehniko, Fakulteta za računalništvo in informatiko, 2019.

[12] J. Kim and M. Hahn. Voice activity detection using an adaptive context attention model. *IEEE Signal Processing Letters*, 25(8):1181–1185, 2018.

[13] M. Mauch and S. Ewert. The Audio Degradation Toolbox and its application to robustness evaluation. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curitiba, Brazil, 2013.

[14] S. Meignier and T. Merlin. LIUM SpkDiarization: An open source toolkit for diarization. In *CMU SPUD Workshop*, Dallas, United States, 2010.

[15] K. K. Rachuri, C. Mascolo, M. Musolesi, and P. J. Rentfrow. SociableSense: Exploring the trade-offs of adaptive sampling and computation offloading for social sensing. In *Proceedings of the 17th annual international conference on Mobile computing and networking - MobiCom '11*. ACM Press, 2011.

[16] M. Trajanoska, M. Katrašnik, J. Lukan, M. Gjoreski, H. Gjoreski, and M. Luštrek. Context-aware stress detection in the aware framework. In M. Luštrek, R. Piltaver, and M. Gams, editors, *Proceedings of the 21st International Multiconference INFORMATION SOCIETY – IS 2018*, volume A, pages 25–28, 2018.

[17] S. E. Tranter and D. A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on audio, speech, and language processing*, 14(5):1557–1565, 2006.