Univerza v Ljubljani
Fakulteta za elektrotehniko

Boštjan Vesnicer

Postopki normalizacije v sistemih za samodejno razpoznavanje govorcev

Doktorska disertacija

Mentor: prof. dr. France Mihelič

Ljubljana 2010

ZAHVALA

Če sem videl dlje, sem zaradi tega, ker sem stal na ramenih velikanov. —Sir Isaac Newton

Ta doktorska disertacija je plod raziskovalnega dela, ki sem ga opravil v Laboratoriju za umetno zaznavanje, sisteme in kibernetiko na Fakulteti za elektrotehniko Univerze v Ljubljani. Da sem postal član tega laboratorija, se moram zahvaliti predstojniku prof. dr. Nikoli Pavešiću, ki me je po opravljeni diplomi dodiplomskega študija povabil v svojo raziskovalno skupino.

V prvi vrsti bi se zahvalil svojemu mentorju prof. dr. Francetu Miheliču, ki me je usmerjal skozi celotno obdobje podiplomskega študija in mi hkrati puščal dovolj akademske svobode, ki je potrebna pri ustvarjalnem delu.

Zahvalil bi se vsem sedanjim in tudi bivšim članom laboratorija: Ankici, Ivu, Jaki, obema Janezoma, Jerneji, Mariu, Meliti, Roku, Sandi, Simonu, Tadeju, Tonetu, Vakilu in Vitu, ki so vedno znali poskrbeti za prijetno vzdušje. Še posebej bi izpostavil Janeza in Simona, v zadnjem obdobju pa tudi Vita, s katerimi smo ob kavi obdelali številne znanstvene, še pogosteje pa filozofske teme.

Pomembno vlogo sta imeli tudi Zoja in Vanja, ki sta mi kazali pot, me vzpodbujali in mi stali ob strani, ko je bilo to najbolj potrebno. Hvaležen sem tudi Vanjinima staršema in sestri, ki so bili vedno pripravljeni priskočiti na pomoč in paziti na Zojo, kadar je bila stiska s časom.

Ob tej priložnosti bi rad izrekel hvaležnost še svojima staršema, ki sta mi omogočila srečno in brezskrbno otroštvo. Posebno mesto pripada mami, ki me je s svojo dobroto in potrpežljivostjo učila o stvareh, ki se jih ne da naučiti iz knjig. Ko bi se le še kdaj mogla veseliti skupaj z mano ...

POVZETEK

V disertaciji se ukvarjamo s problemom besedilno neodvisnega samodejnega razpoznavanja govorcev. Ne ukvarjamo se s problemom ugotavljanja identitete (identifikacija), ampak se posvetimo zgolj splošnejšemu problemu ugotavljanja istovetnosti (verifikacija). Problem ugotavljanja istovetnosti govorcev lahko v najsplošnejši obliki definiramo tako: Denimo, da imamo dva ali več govornih posnetkov. Zanima nas ali je govorec v vseh posnetkih isti?

Celoten postopek razpoznavanja govorcev moremo razdeliti na nekaj korakov. Najprej govorni signal, ki s seboj nosi preveliko količino za nas nepomembne informacije, pretvorimo v parametrični zapis. Tej parametrizaciji signala pravimo v žargonu luščenje značilk. Na ta način dobimo niz vektorjev značilk, vendar obdržimo le tiste vektorje, ki ustrezajo govornim odsekom signala, medtem ko tiste, ki ustrezajo negovornim odsekom signala, zavržemo. Sledi ocenjevanje statističnega modela govorca, čemur pravimo učenje. Osnovni namen učenja je posploševanje, kar z drugimi besedami pomeni, da si mora statistični model ›zapomniti‹ le bistvene govorčeve lastnosti, nebistvene pa izpustiti. Dobro je, če je statistična ocena, ki jo pri tem naredimo, odvisna tudi od količine podatkov, ki je na voljo. Več kot je podatkov, bolj bomo ›verjeli‹ oceni in obratno, manj je podatkov, bolj bo ocena nezanesljiva. Na tak način se izognemo prenaučenju. Tako naučen statistični model uporabimo za ugotavljanje podobnosti med govorci. Bolj kot se testni posnetek prilega na statistični model, bolj sta si govorca v učnem in testnem posnetku med seboj podobna. Če je prileganje dovolj dobro, sprejmemo odločitev, da gre v obeh primerih za istega govorca.

Način človekovega govora je delno pogojem s posameznikovimi fiziološkimi lastnostmi, v veliki meri pa je priučen iz okolja — pravimo, da je govor vedenjska lastnost. Govorni signal, ki ga zajamemo z mikrofonom, ni odvisen le od izgovorjenega besedila, ampak tudi od človekovega psihofizičnega stanja. Poleg tega je znan tudi učinek staranja, do katerega pride, če primerjamo posnetke istega govorca, ki so bili posneti v različnih časovnih obdobjih. Dodatna težava nastane, kadar želimo primerjati govorne signale, ki so zajeti z različnimi mikrofoni in v različnih akustičnih pogojih. Vse našteto povzroči, da isti govorec v dveh različnih posnetkih ›zveni‹ različno, kar s skupnim izrazom imenujemo sejna spremenljivost. Ponavadi vzroke za večino sejne spremenljivosti pripišemo le kanalu, zato včasih sejni spremenljivosti poenostavljeno rečemo kar kanalska spremenljivost.

Prav sejna spremenljivost predstavlja na področju samodejnega razpoznavanja govorcev enega izmed največjih izzivov, zato smo se ji posebej posvetili tudi v naši disertaciji. Vplivom sejne spremenljivosti se lahko skušamo izogniti na različnih nivojih sistema za razpoznavanje govorcev; na nivoju signala oz. značilk, na nivoju statističnega modela ali na nivoju rezultatov prileganja. V disertaciji se ukvarjamo

predvsem s postopki sejne spremenljivosti na nivoju statističnega modela. Posvetimo se dvema že uveljavljenima metodama: postopku projekcije motečih lastnosti in postopku analize vezanih faktorjev ter v povezavi z njima predlagamo alternativne mere podobnosti. Postopek analize vezanih faktorjev je sicer splošnejši od postopka projekcije motečih lastnosti, a je hkrati kompleksnejši z matematičnega vidika in ga je zato teže udejanjiti.

Tako postopek projekcije motečih lastnosti kot postopek analize vezanih faktorjev temeljita na modelu mešanice Gaussovih porazdelitev. Povprečne vektorje posameznih komponent, zložene enega vrh drugega, obravnavata kot točke v visokorazsežnem prostoru supervektorjev. Oba postopka temeljita na predpostavki o linearnosti, ki pravi, da lahko sejni supervektor razstavimo na vsoto govorske in kanalske komponente. Ta dekompozicija je mogoča le, če predpostavimo, da je kanalski vpliv omejen na nižjerazsežni kanalski podprostor. Analiza vezanih faktorjev predpostavko še posploši, saj pravi, da je večji del medgovorske spremenljivosti prav tako omejen na podprostor, ki mu pravimo govorski podprostor in da le majhen del spremenljivosti ostane ›ujet‹ v visokorazsežnem prostoru supervektorjev. Oba podprostora, tako kanalskega kot govorskega, ocenimo iz podatkov v obliki govorne zbirke, v kateri je vsak govorec zastopan s čim večjim številom različnih posnetkov. S tako ocenjenima podprostoroma vnesemo v sistem razpoznavanja apriorno znanje, ki omogoči, da supervektor neznanega govorca, pridobljen iz testnega posnetka, razklenemo na govorsko in kanalsko komponento.

Sistem, ki ga predlagamo v disertaciji, temelji na nekaterih odločitvah, ki so v nasprotju z uveljavljenimi pristopi s področja razpoznavanja govorcev. Znano je, da so rezultati pri ženskih govorkah znatno slabši kot pri moških govorcih, zato je ustaljena praksa, da se za ženske in moške zgradi dva povsem ločena sistema. Sami smo prepričani, da je takšna delitev z znanstvenega stališča neupravičena, zato ženske in moške obravnavamo v skupnem sistemu. Še druga lastnost, po kateri se naš sistem razlikuje od ostalih, je, da ne uporabimo postopka normalizacije na nivoju značilk, ampak le na nivoju modela. Naša teza je, da lahko učinek normalizacije na nivoju značilk v veliki meri nadomestimo z normalizacijo na nivoju modela. Rezultati, ki jih pridobimo na uveljavljeni testni zbirki, kažejo, da je temu res tako. Ugotovitev velja še posebej v primeru, ko za merjenje podobnosti uporabimo metodo podpornih vektorjev.

Znano je, da lahko z združevanjem rezultatov prileganja večih sistemov občutno izboljšamo rezultat razpoznavanja. Zaželjeno je, da so sistemi čimbolj heterogeni, s čimer v postopek združevanja prispevajo zadosti komplementarne informacije. V našem primeru pokažemo, da je združevanje rezultatov prileganja uspešno tudi, če združimo rezultate dveh skoraj enakih sistemov, ki se razlikujeta le v načinu merjenja podobnosti. Rezultat, ki ga dobimo s preprostim seštevkom rezultatov prileganja dveh sistemov, je primerljiv z rezultati najboljših (bolj kompleksnih) sistemov z zadnje uveljavljene prireditve za vrednotenje sistemov razpoznavanja govorcev.

ABSTRACT

The thesis addresses the problem of text-independent speaker recognition. We are particularly interested in the verification problem, which could be defined as follows. Given two speech utterances, decide whether both utterances have been uttered by the same speaker or by different speakers.

In the procedure of recognizing a speaker there are multiple steps involved. First, since the speech signal contains to much redundant information, it has to be transformed into some kind of parametric form. This parametrization is known as feature extraction, which transforms the time-domain signal into a sequence of acoustic feature vectors. Only those feature vectors are kept that correspond to the speech portion of the signal while the non-speech frames are being removed. Next, the feature vectors are used to train the statistical model of the speaker. The statistical model serves us as a compact representation of the intrinsic properties of the speaker's voice. The sole purpose of the training process is that of generalization — by learning we want to get rid of the properties that are specific only to the particular utterance of the speaker but are not actually representative for that speaker in general. If the statistical model exhibits good generalization properties, than it is able to avoid overfitting to the training data and consequently generalizes well to the yet unseen test data of the same speaker. The learned statistical model is used for measuring the similarity between given speaker and the test utterance of the unknown speaker. The more that the test utterances matches to the given model, the more evidence we have that the speaker in the test utterance is the same as the one in the training utterance and vice versa, the less that the test utterances matches to the model, the more confident we can be that the there are actually two different speakers.

Human voice is to some extent subject to the individual's psychophysical characteristics while on the other hand it is also influenced by the environment (in a wide sense) the individual is living in — we account the human voice as being a combination of psychophysical and behavioural characteristics. Unfortunately, the characteristics of the speaker's voice changes to some extent from one utterance to the other. We speak of session variability, which is due to the differences in the speaker's psychophysical condition (e.g., health or emotional state) and to the phonetic variations between different utterances. This type of variability is known as intra-speaker variability. There is also the well known aging phenomenon, which causes the performance of speaker models degrades over time. On the other hand, there is another type of variability, which causes that two utterances of the same person sound different from each other. We speak of channel variability, which is due to differences in acoustic backgrounds and also due to different types of microphones and transmission channels involved while recording the utterances.

The session variability poses one of the main research challenges to the field of speaker recognition and is also in the focus of our interest in this thesis. A lot of research effort has been put into addressing the problem of session variability. To that end, many different techniques that try to decrease the effect of session variability have been proposed. They can be categorized into three different classes: (i) signal- or feature-domain methods, (ii) model-domain methods and (iii) matching-score-domain methods. In this dissertation we are mainly concerned with the model-domain session variability normalization methods. Particularly, we study two well established techniques: *nuisance attribute projection* and *joint factor analysis*. They are both based on the notion of the so called *supervector*, which can be seen as a high-dimensional representation of the speaker characteristics in the parameter-space (usually, only the means are considered) of the Gaussian mixture model. The basic assumption of both methods is that the session supervector can be linearly decomposed into two components: the speaker supervector and the channel supervector. However, if we want the decomposition to be feasible, than the channel supervector has to be confined to a low-dimensional subspace. This requirement seems reasonable since the channel should not be able to transform one speaker into another, otherwise speaker recognition would be an ill-posed problem.

While designing the systems used in our experiments we deliberately made some decisions which disagree with the most common approaches prevailing within the speaker recognition community. Since it is known that the speaker recognition systems perform worse for the female speakers than for the male speakers, it is common practice to treat both genders separately and thus to develop two gender-dependent systems. In our opinion such artificial partition is not grounded from the scientific point of view so we decided to design our systems in a gender-independent fashion. Another property which discern our systems from the prevailing practice is that we don't use any form of the feature-level normalization. We speculate that the feature-level normalization can be made dispensable by performing more powerful normalization in the model domain. The experiments show that this is actually the case, at least if the discriminative method of support vector machines is used as a similarity measure.

It is well known that the score-level fusion of different systems can significantly improve the performance of the individual systems. However, to assure the fusion to be effective, the systems should contribute complimentary information, thus they have to be sufficiently heterogenous. In dissertation we show, that a considerable increase in performance can be achieved by fusing scores of two similar systems that differ only in the decision criterion used for scoring the trials. The results we get on a standard evaluation database are comparable to those obtained by the best performing systems according to the NIST SRE 2008 evaluation campaign.

KAZALO

UVOD 1

1.1 Motivacija

Govor predstavlja človeku enega izmed najbolj naravnih načinov sporazumevanja. Čeprav je osnovna naloga govora prenašanje sporočila med dvema ali večimi sogovorniki, je v vsakem govornem signalu prisotna tudi informacija o lastnostih govorca ter informacija o lastnostih komunikacijskega kanala, preko katerega se govorni signal prenaša.

Ljudje ponavadi z lahkoto »uganemo« identiteto govorca, kakor tudi nimamo težav z razumevanjem v govornem signalu vsebovanega sporočila, zato se laikom nemalokrat zdi, da bo to preprosta naloga tudi za računalnik. A izkaže se ravno nasprotno — na videz zelo preprost problem za človeka predstavlja za računalnik zelo trd oreh.

V disertaciji bomo skušali izpostaviti težave, ki povzročijo, da postane razpoznavanje govorcev zahteven problem. Težave bomo skušali omiliti z uporabo različnih orodij, ki nam jih ponujajo med seboj prepletajoča in dopolnjujoča se raziskovalna področja kot so obdelava signalov, razpoznavanje vzorcev, strojno učenje, statistika ter verjetnostni račun.

Samodejno razpoznavanje govorcev ni zanimivo zgolj s stališča znanosti, ampak ima tudi potencialno široko praktično uporabnost. Stroji, ki bi bili sposobni ugotavljati identiteto govorca na podlagi govornega signala, bi lahko nadomestili oz. dopolnili obstoječe varnostne sisteme za kontrolo vstopa (npr. geslo, PIN), uporabni bi lahko bili za forenzične in obveščevalno-varnostne namene, prav tako pa bi jih lahko s pridom izkoristili pri iskanju po najrazličnejših zvočnih arhivih. Tehnologijo razpoznavanja govorcev je možno tudi zlorabiti, kar v javnosti vzbuja številne etične pomisleke, zato je tehničnemu napredku potrebno slediti tudi na pravnem področju.

1.2 Tema disertacije

Pri razpoznavanju govorcev ločimo med *besedilno odvisnim* (angl. text dependent) in *besedilno neodvisnim* (angl. text independent) razpoznavanjem. Pri besedilno odvisnem razpoznavanju od govorca zahtevamo, da izreče točno določeno besedilo, medtem ko pri besedilno neodvisnem razpoznavanju razpoznavamo identiteto govorca pri poljubnem izrečenem besedilu. Prvi način razpoznavanja je primeren predvsem v aplikacijah, ko govorec s sistemom *sodeluje* (npr. pri kontroli vstopa), medtem ko pri drugem načinu govorec ni nujno seznanjen, da je udeležen v postopku samodejnega razpoznavanja. Govorimo o t.i. *nezavednem* oz. *nekooperativnem* načinu razpoznavanja, ki je primeren v forenzičnih aplikacijah, pri nadziranju telefonskih

pogovorov in pri brskanju po najrazličnejših zvočnih arhivih. Problem besedilno neodvisnega razpoznavanja je bistveno težji in zaradi tega z znanstvenega vidika tudi zanimivejši, zato bo v središču naše pozornosti.

1.2.1 Biometrično razpoznavanje oseb

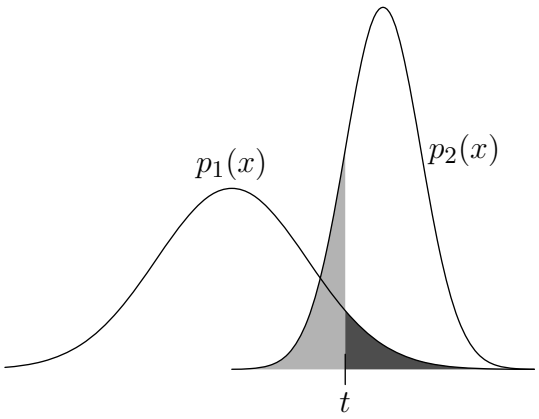
Problem razpoznavanja govorcev uvrščamo na širše področje biometričnega razpoznavanja oseb oz. biometrije. Z besedo biometrija označujemo uporabo *fizioloških* in *vedenjskih* značilnosti za samodejno razpoznavanje oseb. Primeri fizioloških značilnosti so prstni odtisi, obraz, šarenica itd., primeri vedenjskih značilnosti pa hoja, drža, podpis itd. Verjetno nobene od omenjenih značilnosti ne moremo uvrstiti izključno med fiziološke ali med vedenjske, ampak so kombinacija obeh vrst značilnosti. To je še posebej očitno pri govoru, ki je delno določen s človekovimi govornimi organi, delno pa z načinom govora, ki je posledica številnih dejavnikov, kot so poreklo, zdravje, starost, spol, počutje, izobrazba itd.

Pri razpoznavanju govorcev (in drugih biometričnih sistemih) ločimo dva načina razpoznavanja: *identifikacijo* in *verifikacijo*. Pri identifikaciji želi sistem ugotoviti identiteto neznanega govorca, medtem ko želi pri verifikaciji sistem le potrditi ali je govorni signal izgovoril točno določen govorec. Izkaže se, da je problem verifikacije splošnejši od identifikacije, saj lahko problem identifikacije brez težav prevedemo na problem verifikacije.

Vsak biometrični sistem, ki deluje v načinu verifikacije, je podvržen napakam, ki so lahko dvoje vrst. Ko sistem sprejme vsiljivca, govorimo o napaki *napačnega sprejema* (angl. false acceptance, FA), ko pa sistem zavrne klienta, govorimo o napaki *napačne zavrnitve* (angl. false rejection, FR). Napaki sta med seboj odvisni – če želimo doseči majhno napako FR, to nujno pomeni, da povečamo napako FA. Velja tudi obratno. Odvisnost obeh napak pogosto narišemo v obliki ROC (angl. receiver operating point) krivulj oz. njim enakovrednim DET (angl. detection error tradeoff) krivulj (Martin et al., 1997). Poleg omenjenih krivulj ponavadi podamo še vrednosti obeh napak v nekaterih značilnih točkah. Takšni točki sta npr. EER (angl. equal error rate), ki podaja vrednosti napak, ko sta le-ti enaki, in točka, v kateri doseže utežena vsota obeh napak (angl. detection cost function, DCF) najmanjšo vrednost.

Obe vrsti napake lahko nazorno prikažemo grafično (slika **1.1**). Če izvedemo veliko število poskusov, katerih rezultate prileganja razdelimo v dve množici tako, da damo v prvo množico rezultate klientov, v drugo pa rezultate vsiljivcev, dobimo dve porazdelitvi rezultatov prileganja. Porazdelitev $p_1(x)$ ustreza porazdelitvi rezultatov prileganja vsiljivcev, porazdelitev $p_2(x)$ pa porazdelitvi rezultatov prileganja klientov. Verjetnost napake, da bomo pri danem pragu t sprejeli vsiljivca, je enaka ploščini območja, ki je na sliki pobarvano s temnejšo barvo. Verjetnost napake, da bomo pri danem pragu t zavrnili klienta, pa je enaka ploščini območja, ki je na sliki pobarvano s svetlejšo barvo. Zapišemo lahko:

$$P_{\mathrm{FA}}(t) = \int_t^{+\infty} p_1(x)\mathrm{d}x \qquad \text{in} \qquad P_{\mathrm{FR}}(t) = \int_{-\infty}^{t} p_2(x)\mathrm{d}x.$$



Slika 1.1 Dve vrsti napak biometričnega sistema.

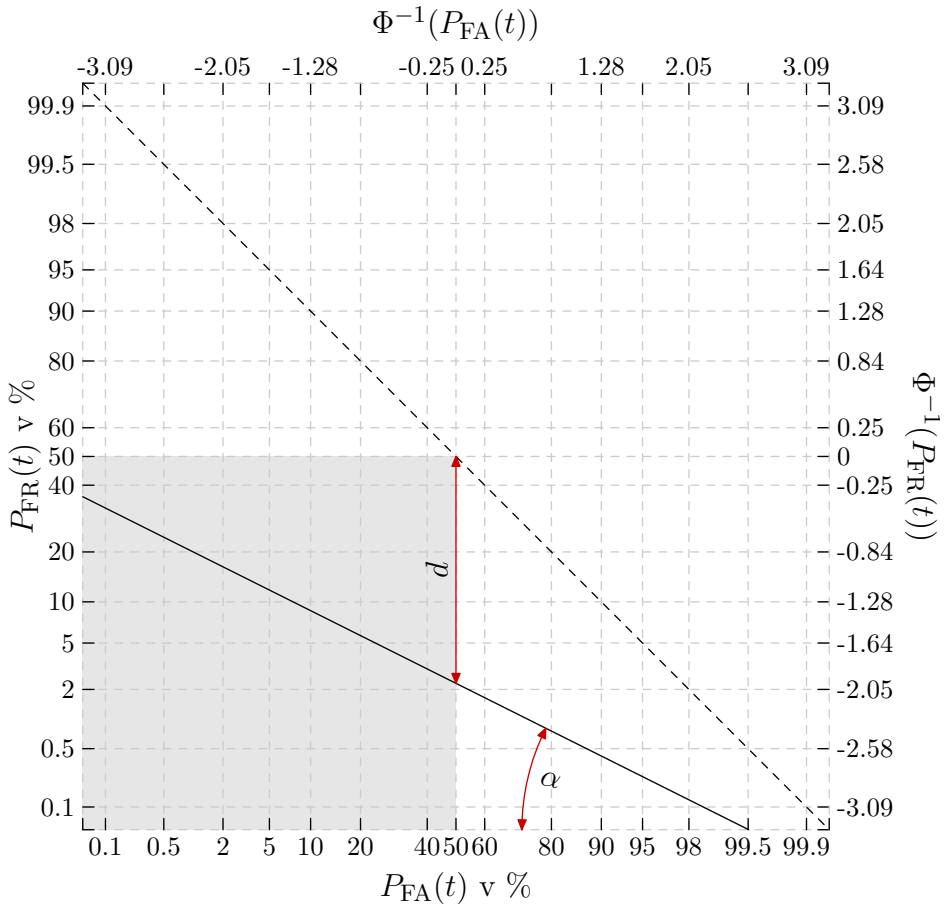
Vidimo lahko, da sta verjetnosti obeh napak odvisni od praga t. Optimalna vrednost praga je odvisna od aplikacije, v kateri želimo uporabiti biometrični sistem. Pri kontroli vstopa je tako zaželeno, da je verjetnost napake napačnega sprejema čim manjša, zato bomo prag postavili na višjo vrednost. Drugačne zahteve imamo pri iskanju govorcev v zvočnih zapisih, kjer smo po navadi pripravljeni tolerirati višjo napako napačnega sprejema, zato smemo prag postaviti nižje. Jasno je torej, da gre pri izbiri pragovne vrednosti t vedno za kompromis med obema vrstama napake. Na skupno napako lahko vplivamo le tako, da spremenimo porazdelitvi $p_1(x)$ in $p_2(x)$. Bolj se bosta porazdelitvi prekrivali, večja bo skupna napaka in obratno, bolj bosta porazdelitvi vsaka k sebi, manjša bo skupna napaka.

Krivuljo DET narišemo tako, da na abscisno in ordinatno os grafa namesto vrednosti $P_{\mathrm{FA}}(t)$ in $P_{\mathrm{FR}}(t)$ nanašamo transformirani vrednosti $\Phi^{-1}(P_{\mathrm{FA}}(t))$ in $\Phi^{-1}(P_{\mathrm{FR}}(t))$, kjer smo s $\Phi^{-1}(\cdot)$ označili inverzno kumulativno porazdelitveno funkcijo standardno-normalne porazdelitve. S tem dosežemo, da bo krivulja DET imela obliko premice, če bosta le porazdelitvi $p_1(x)$ in $p_2(x)$ normalni. Če predpostavimo, da velja: $p_1(x) = \mathcal{N}(x|\mu_1, \sigma_1^2)$ in $p_2(x) = \mathcal{N}(x|\mu_2, \sigma_2^2)$, potem kratek račun pokaže, da bo krivulja DET imela naslednjo obliko:

$$\Phi^{-1}(P_{\mathrm{FR}}(t)) = -\frac{\sigma_1}{\sigma_2}\Phi^{-1}(P_{\mathrm{FA}}(t)) + \frac{\mu_1 - \mu_2}{\sigma_2}.$$

Razmere so skicirane na sliki **1.2**. S črtkano črto je vrisan sistem, ki bi deloval v popolnoma naključnem režimu — to je takrat, ko se porazdelitvi $p_1(x)$ in $p_2(x)$ popolnoma prekrivata. S polno črto je vrisan sistem, za katerega velja (glej oznake na sliki):

$$d = \frac{\mu_1 - \mu_2}{\sigma_2} \quad \text{in} \quad \tan\alpha = \frac{\sigma_1}{\sigma_2}$$



Slika 1.2 Krivulja DET. V praksi omejimo izris na spodnji levi kvadrant, ki je na sliki označen osenčeno.

1.3 Pregled področja disertacije

Naredimo kratek pregled področja razpoznavanja govorcev. Že vnaprej naj opozorimo, da se zaradi obsežnosti in relativno dolge zgodovine področja na tem mestu ne bomo spuščali v vse podrobnosti in še manj naštevali vse relevantne reference. Namen pregleda bo dosežen že, če bo bralec, ki ni nujno ekspert na tem področju, dobil vsaj približen občutek za probleme, ki so predmet naše disertacije. Za natančnejši pregled področja priporočamo ogled preglednih člankov, npr. (Bimbot et al., 2004) ter še posebej aktualnega (Kinnunen in Li, 2009), ki je bil sprejet v objavo ravno v času pisanja te disertacije.

1.3.1 Večnivojskost informacije o identiteti govorca

Informacija o identiteti govorca je v govornem signalu prisotna na večih ravneh, vse od najnižje akustične, preko prozodične, fonetične, idiolektične[1] pa vse do pomenske ravni (Reynolds et al., 2003b). Ni dvoma, da znamo ljudje pri razpoznavanju

identitete govorca to večplastno informacijo zelo dobro združiti, zato je bilo nekaj let nazaj zaslediti veliko poskusov v tej smeri (Shriberg et al., 2005; Campbell et al., 2007). Višjenivojske značilke naj bi v primerjavi z akustičnimi značilkami bile bolj robustne na različne akustične vplive, kar je bil dodaten motiv za njihovo uporabo v sistemih za samodejno razpoznavanje govorcev. Kljub začetnim obetom pa razpoznavanje govorcev na podlagi višjenivojske informacije ni izpolnilo vseh pričakovanj. Rezultati sistemov, ki so upoštevali le višjenivojsko informacijo, se niso približali tistim, ki jih dobimo s sistemi, ki temeljijo na najnižjem, t.j. akustičnem nivoju. Zato se je po začetnem navdušenju večina raziskav ponovno usmerila na ta nivo.

1.3.2 Sejna spremenljivost

Omenili smo že, da ločimo med besedilno odvisnim in besedilno neodvisnim razpoznavanjem. Pri besedilno odvisnem razpoznavanju, ki je primerno tedaj, ko gre za sodelujočega uporabnika, se od uporabnika zahteva, da izgovori vnaprej določeno frazo. Na drugi strani pa sta pri besedilno neodvisnem razpoznavanju izbira besed in besedni red povsem poljubna. Tako se lahko se zgodi, da mora sistem za razpoznavanje govorcev preveriti istovetnost osebe na podlagi dveh povedi (angl. utterance), ki imata popolnoma različno vsebino. Očitno je, da je besedilno neodvisno razpoznavanje veliko zahtevnejša (a zato zanimivejša) naloga kot besedilno odvisno razpoznavanje.

Fonetična spremenljivost je le eden izmed faktorjev, ki vplivajo na zahtevnost naloge in s tem na učinkovitost sistemov za samodejno razpoznavanje govorcev. Med druge faktorje, ki prav tako pomembno vplivajo na robustnost sistemov za razpoznavanje govorcev, štejemo spremembe akustičnega ozadja, vpliv različnih vrst mikrofonov in prenosnih poti, kakor tudi vse vzroke za t.i. *znotrajgovorsko* (angl. within-speaker) spremenljivost (zdravje, počutje, razpoloženje, staranje itd.). Vse razlike med različnimi posnetki istega govorca opišemo s skupnim izrazom *sejna spremenljivost* (angl. session variability) (Kenny et al., 2007b). Med raziskovalci je splošno sprejeto mnenje, da ponuja na področju razpoznavanja govorcev ravno sejna spremenljivost enega izmed najtežjih še ne zadovoljivo rešenih problemov in bo zato ostala v središču pozornosti tudi v prihodnje.

1.3.3 Značilke

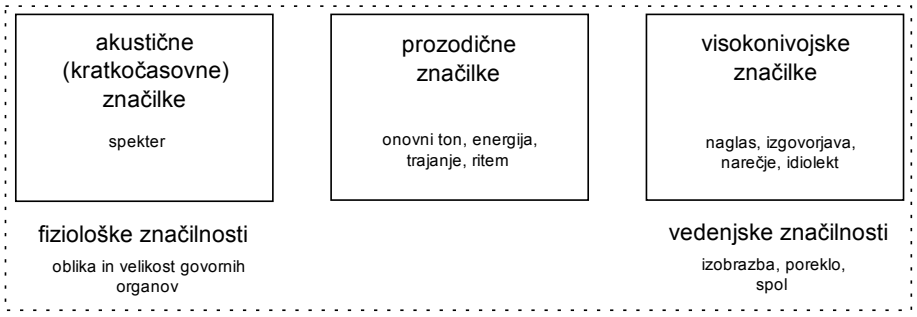
Surov govorni signal vsebuje preveč informacije, ki ne prispeva k boljšemu ločevanju med posameznimi govorci. Da se te odvečne (in moteče) informacije znebimo, iz govornega signala skušamo izluščiti le za nas koristno informacijo v obliki značilk. Idealne značilke bi naj:

- imele veliko medgovorsko in majhno znotrajgovorsko spremenljivost;
- bile odporne na šum in popačitve;

[1] Idiolekt je način govora, ki je lasten vsakemu posamezniku. Pogosto pod idiolekt štejemo govorčev besedni zaklad, posebnosti v izgovorjavi in naglašanju besed, slovnico itd.

- se pogosto pojavljale v govoru;
- se jih dalo enostavno pridobiti iz govora;
- se jih težko posnemalo;
- ne bile odvisne od govorčevega psihofizičnega stanja.

Povsem jasno je, da idealne značilke ne obstajajo oz. jih vsaj še nismo uspeli najti, zato obstaja veliko število različnih vrst značilk, ki jih lahko glede na lastnosti razdelimo v nekaj skupin (slika **1.3**).



Slika 1.3 Delitev značilk v skupine.

Vsaka izmed skupin značilk ima tako dobre kot slabe lastnosti. Izbira vrste značilk za konkreten primer je odvisna predvsem od namena uporabe, računalniških zahtev, količine podatkov, ki je na voljo itd. Trenutno se verjetno kot najboljša izbira ponujajo kratkočasovne značilke, saj se z njimi dosega najboljše rezultate pa še enostavno jih je pridobiti iz signala. Prozodične in višjenivojske značilke bi naj bile sicer bolj robustne na šum, a so hkrati tudi manj diskriminatorne, težje jih pridobimo iz signala (npr. za izpeljavo nekaterih potrebujemo razpoznavalnik govora) in lažje jih je posnemati. Če imamo to možnost, lahko hkrati uporabimo več različnih značilk in tako izkoristimo komplementarnost informacije, ki jo posamezne značilke vsebujejo.

1.3.4 Modeliranje govorčevih značilnosti

Če privzamemo uporabo kratkočasovnih značilk, lahko govorni signal predstavimo v obliki niza vektorjev značilk. Zanima nas, kako izvesti primerjavo med dvema takima nizoma, ki ustrezata dvema različnima govornima signaloma. Najenostavnejša možnost je, da niza med seboj primerjamo kar neposredno. Med bolj znanimi postopki, ki temeljijo na takšni neposredni primerjavi, lahko izpostavimo dinamično ukrivljanje časa (angl. dynamic time warping, DTW) in vektorsko kvantizacijo (angl. vector quantization, VQ). Čeprav lahko takšni postopki v določenih primerih dajo zelo dobre rezultate, jim manjka pomembna lastnost. Ne premorejo namreč lastnosti posploševanja (generalizacije) znanja na nevidene primere, ampak si ›zapomnijo‹ le (natančno) tisto, kar so videli (analogija z učenjem na pamet).

Alternativa tem *šablonskim* (angl. template) postopkom so *statistični* postopki, ki se znajo iz podatkov ›učiti‹. Pogosto jih razvrstimo v dve skupini: (i) *generativni* in (ii) *diskriminatorni* postopki. Večini postopkom iz obeh skupin je skupno, da

znanje, ki ga pridobijo iz podatkov, strnejo v obliki *parametričnega modela*, pri tem pa znajo koristno uporabiti tudi znanje, ki so ga že predhodno pridobili iz prej videnih podatkov (apriorno znanje).

Na področju razpoznavanja govorcev so v zadnjem času popularni predvsem postopki, ki temeljijo na generativnem modelu *mešanice Gaussovih porazdelitev* (angl. Gaussian mixture model, GMM) in *prikritem modelu Markova* (angl. hidden Markov model, HMM) (Rabiner, 1989) ter diskriminatorni *metodi podpornih vektorjev* (angl. support vector machine, SVM) (Vapnik, 1995; Burges, 1998). Model HMM je splošnejši od modela GMM in je sposoben (delno) opisati časovno odvisnost med značilkami, kar se s pridom izkorišča pri razpoznavanju govora. Ta lastnost pri besedilno neodvisnemu razpoznavanju govorcev ni ključnega pomena, saj nas zanima predvsem oblika porazdelitve značilk ne pa tudi časovna odvisnost med njimi[2], zato se zdi, da je model GMM boljša izbira od modela HMM.

1.4 Cilji raziskovalnega dela

V disertaciji bomo skušali predlagati izvirne postopke, s katerimi bomo skušali zmanjšati vpliv sejne spremenljivost na zanesljivost razpoznavanja govora. Predvsem se bomo osredotočili na postopke normalizacije na nivoju modela.

V središču naše pozornosti bosta postopka *projekcije motečih lastnosti* (angl. nuisance attribute projection, NAP) (Solomonoff, 2005; Campbell et al., 2006a) in postopek *analize vezanih faktorjev* (angl. joint factor analysis, JFA). Oba postopka bomo skušali nadgraditi z izvirnimi načini računanja mere podobnosti med dvema govornima signaloma, ki služi kot odločitveni kriterij za razpoznavanje govorcev.

1.5 O izrazoslovju in notaciji

Teoretično ozadje, na katerem temeljijo postopki, s katerimi se ukvarjamo v disertaciji, sega v znatni meri na področje verjetnosti in statistike. Razumljivo je, da se zato na številnih mestih izražamo v jeziku verjetnostnega računa. Čeprav skušamo biti pri tem ves čas konsistentni, se ne trudimo za vsako ceno zadostiti rigoroznim matematičnim standardom. Bolj pomembno se nam zdi, da kar se da jasno predstavimo idejo, od bralca pa mestoma pričakujemo, da natančen pomen določenega izraza razbere iz sobesedila.

V statistiki so osrednjega pomena podatki, s katerimi želimo oceniti parametre statističnega modela. Podatke, ki jih je vedno končno mnogo, zapišemo v obliki niza, npr. $x_1, \ldots, x_N$, ki ga lahko zapišemo tudi krajše kot $x_{1:N}$. Kadar se zgodi, da niz zapišemo le s simbolom x, predpostavljamo, da je razvidno iz sobesedila, da je govor o nizu podatkov in ne le o enem podatku. Kadar želimo poudariti, da podatki niso skalarji, ampak da so večrazsežni (enakorazsežni) vektorji, potem to nakažemo z odebeljeno pisavo, npr. $\boldsymbol{x}$. Podatkom rečemo z drugo besedo tudi *observacije*, za katere predpostavljamo, da jih je porodil *naključni proces*. Pravimo, da je niz

[2] Z drugimi besedami, ne zanima nas, kaj je bilo povedano, temveč le, kako oz. kdo je to povedal.

$x_{1:N}$ uresničitev naključnega procesa, t.j. *zbirke naključnih spremenljivk*. Naključne spremenljivke označimo z velikimi (tiskanimi) neodebeljenimi črkami, npr. X, tudi kadar gre v resnici za naključni vektor. Na ta način jih lahko ločimo od matrik, za katere ›rezerviramo‹ velike odebeljene simbole, npr. $\boldsymbol{A}$.

Porazdelitve verjetnosti, ki nastopajo v različnih enačbah, ponavadi zapišemo s simbolom $p(\cdot)$, ne glede na to ali imamo v mislih funkcijo gostote verjetnosti (angl. probability density function, pdf) ali funkcijo mase verjetnosti (angl. probability mass function, pmf). Takšna notacija sicer odstopa od uveljavljene notacije, ki je prisotna v številnih učbenikih s področja verjetnostnega računa. Tam je meja med zveznimi in diskretnimi naključnimi spremenljivkami zarisana strožje, zato se za funkcije gostote verjetnosti po dogovoru uporablja male črke, npr. $p(\cdot)$, za funkcije mase verjetnosti pa velike črke, npr. $P(\cdot)$. Bi se ne bilo mogoče smiselno držati take notacije? V nekaterih primerih mogoče res, a kaj storiti v primeru, ko imamo vezano porazdelitev dveh naključnih spremenljivk, od katerih je ena zvezna, druga pa diskretna? Da ne gre pretirano razlikovati med diskretnimi in zveznimi spremenljivkami (in posledično med velikimi in malimi simboli za maso in gostoto verjetnosti), sledi tudi iz *teorije mere*, ki obravnavo diskretnih in zveznih naključnih spremenljivk povsem poenoti[3].

Kadar želimo eksplicitno poudariti, da gre za porazdelitev točno določene naključne spremenljivke, to nakažemo tako, da simbol $p(\cdot)$ podpišemo s simbolom, s katerim smo označili konkretno naključno spremenljivko, npr. $p_X(x)$. V večini primerov lahko indeks izpustimo, saj sledi iz argumenta. Operator matematičnega upanja označimo s simbolom $\mathbb{E}[\cdot]$, npr. $\mathbb{E}[X]$. Ponavadi lahko porazdelitveno funkcijo, na katero se matematično upanje nanaša, razberemo iz konteksta. Kadar temu ni tako, uporabimo zapis z indeksom. Naredimo primer: pogojno matematično upanje funkcije dveh naključnih spremenljivk X in Y z ozirom na pogojno porazdelitev $p_{X|Y}(x|y)$ zapišemo kot $\mathbb{E}_X[f(X,Y)|Y]$.

1.6 O bayesovski interpretaciji verjetnostnega računa

Vsi teoretični rezultati, ki jih bomo podali oz. izpeljali v disertaciji in ki se dotikajo verjetnostnega računa, sledijo iz dveh osnovnih pravil verjetnosti:

$$p(x,y) = p(x|y)p(y) \tag{1.1}$$

$$p(x) = \int_{\Omega_Y} p(x,y)\mathrm{d}y \tag{1.2}$$

[3] Če verjetnostni račun predstavimo v jeziku teorije mere, zmoremo obravnavati tudi naključne spremenljivke, ki jih ne moremo opisati ne s funkcijo mase verjetnosti niti s funkcijo gostote verjetnosti — tak primer je Cantorjeva porazdelitev. Prav tako je mersko-teoretična obravnava verjetnosti neizbežna v primerih, ko imamo opravka z neskončnodimenzionalnimi vzorčnimi prostori.

Enačbi (1.1) pravimo *pravilo produkta*, enačbi (1.2) pa *pravilo vsote*. Integral v enačbi (1.2) je izveden po celotnem vzorčnem prostoru[4] (tega označimo z Ω_Y) naključne spremenljivke Y. V kolikor je vzorčni prostor spremenljivke Y diskreten, integracijo zamenjamo s seštevanjem. Izraz $p(x, y)$ imenujemo *vezana* porazdelitev spremenljivk X in Y, izraz $p(x|y)$ pa *pogojna* porazdelitev spremenljivke X pri dani vrednosti spremenljivke Y. Ob upoštevanju obeh osnovnih pravil verjetnosti in lastnosti simetričnosti: $p(x, y) = p(y, x)$, lahko pokažemo, da velja zveza

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}, \tag{1.3}$$

ki ji pravimo *Bayesov izrek*. Kljub temu, da je veljavnost Bayesovega izreka matematično dejstvo, pa je s statističnega vidika možnih več različnih interpretacij (in z njimi povezanih načinov uporabe) tega izreka. Nam je še posebej blizu naslednja interpretacija: Denimo, da imamo parametrični statistični model, pri čemer parameter modela obravnavamo kot naključno spremenljivko, ki jo označimo s Θ. Predpostavke, ki jih imamo o parametrih, še preden smo opazili podatke, opišemo z apriorno porazdelitvijo $p(\theta)$. Brž ko opazimo podatke, ki jih označimo z $\mathcal{D}$, lahko izpeljemo aposteriorno porazdelitev $p(\theta|\mathcal{D})$:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}, \tag{1.4}$$

ki sledi iz Bayesove formule. Količinam, ki nastopajo v gornji enačbi, nadenemo naslednja imena: $p(\theta)$ je prior, $p(\theta|\mathcal{D})$ je posterior, $p(\mathcal{D}|\theta)$ je *verjetje* (angl. likelihood), normalizacijska konstanta $p(\mathcal{D})$ pa popolno verjetje oz. v angl. model evidence. Vse količine obravnavamo kot funkcije parametrov θ pri konstantni vrednosti podatkov $\mathcal{D}$. Normalizacijsko konstanto $p(\mathcal{D})$ izračunamo tako, da iz vezane porazdelitve $p(\mathcal{D}, \theta)$ izintegriramo spremenljivko θ:

$$p(\mathcal{D}) = \int_{\Omega_\Theta} p(\mathcal{D}|\theta)p(\theta)\mathrm{d}\theta,$$

kar ustreza matematičnemu upanju funkcije verjetja $p(\mathcal{D}|\Theta)$ glede na apriorno porazdelitev $p(\theta)$, t.j.:

$$p(\mathcal{D}) = \mathbb{E}_\Theta[p(\mathcal{D}|\Theta)].$$

Gornja interpretacija Bayesovega izreka je lastna bayesovski interpretaciji verjetnosti, po kateri lahko *nedoločenost* (angl. uncertainty) vrednosti parametrov statističnega modela, ki izhaja iz pomanjkljivega znanja o tem parametru, opišemo z apriorno porazdelitvijo. Ortodoksni statistik se seveda s takšno interpretacijo, ki obravnava parametre kot naključne spremenljivke, ne strinja. Zanj so parametri le konstantne (čeprav morda neznane) vrednosti, ki jih ocenimo iz podatkov, a jim nikakor ne moremo pripisati verjetnostnih porazdelitev[5]. Naj opozorimo, da sta oba

[4] Kadar ›področje‹, po katerem integriramo, ne zapišemo eksplicitno, implicitno predpostavljamo, da poteka integracija po celotnem vzorčnem prostoru ustrezne naključne spremenljivke.

pogleda na statistiko v skladu z aksiomatsko definicijo verjetnostnega računa, le njuni interpretaciji sta konceptualno povsem različni.

Bayesovska interpretacija verjetnostnega računa ima s stališča razpoznavanja vzorcev in strojnega učenja nekaj prednosti pred frekvencionistično interpretacijo. Proces ›izpeljave‹ posteriorne porazdelitve lahko interpretiramo kot ›učenje‹, s katerim apriorno znanje (podano z apriorno porazdelitvijo) ›osvežimo‹ z znanjem, ki izvira iz podatkov. Rezultat bo posteriorna porazdelitev, ki jo lahko v primeru, da opazimo nove podatke, spet obravnavamo kot apriorno porazdelitev. Proces učenja nastopi na ta način naravno — skozi osveževanje posteriorne porazdelitve. Pri majhni količini podatkov bo imela posteriorna porazdelitev široko varianco oz. bo precej nedoločena, z večanjem količine podatkov pa se bo porazdelitev začela gostiti okoli določene vrednosti parametrov. V limiti, ko bo količina podatkov šla proti neskončnosti, bo porazdelitev degenerirala v točko; z drugimi besedami, nedoločenost vrednosti parametrov bo izginila.

Popolno verjetje $p(\mathcal{D})$, ki se nahaja v imenovalcu izraza (**1.4**), lahko uporabimo v primeru, ko med seboj primerjamo več modelov in nas zanima, kateri med njimi najbolje opiše podatke $\mathcal{D}$. Kadar imamo podatke razdeljene na več delov (npr. učni in testni podatki), lahko izračunamo popolno verjetje pod posteriorno porazdelitvijo:

$$p(x|\mathcal{D}) = \int_{\Omega_\Theta} p(x|\theta)p(\theta|\mathcal{D})\mathrm{d}\theta,$$

kjer smo predpostavili, da so učni ($\mathcal{D}$) in testni (x) podatki pogojno neodvisni (pri dani vrednosti parametrov θ). Količini $p(x|\mathcal{D})$ pravimo *prediktivna* porazdelitev.

Pogosto se zgodi, da natančne posteriorne porazdelitve ne znamo izračunati po analitični poti. Takrat si lahko pomagamo z metodami Monte Carlo, ki temeljijo na naključnem vzorčenju, ali pa se zadovoljimo s kakšno izmed metod deterministične aproksimacije (Laplaceova metoda, variacijska metoda). Še radikalneje je, če celotno posteriorno porazdelitev opišemo z le eno značilno točko — pravimo, da smo naredili točkovno oceno. Najbolj logična izbira je točka, v kateri doseže posteriorna porazdelitev največjo vrednost[6] (angl. maximum a posteriori, MAP) — označimo jo z $\hat{\theta}_{\mathrm{MAP}}$. Kadar predpostavljamo, da so vse vrednosti parametrov a priori enako verjetne[7], sovpade ocena $\hat{\theta}_{\mathrm{MAP}}$ z oceno, pri kateri doseže verjetje $p(x|\theta)$ najvišjo

[5] Filozofi govorijo o dveh vrstah nedoločenosti — prva izhaja iz naključnosti (aleatorna nedoločenost), druga pa iz neznanja (epistemološka nedoločenost). Ortodoksni (frekvencionistični) statistik je mnenja, da epistemiološke nedoločenosti ne moremo opisati oz. meriti z verjetnostmi, medtem ko bayesovski statistik uporablja verjetnosti za opis obeh vrst nedoločenosti. Zgodovinsko gledano je močnejši vpliv na ›uradno‹ interpretacijo verjetnosti imela frekvencionistična šola na čelu z vplivnim zagovornikom R. A. Fisherjem, zato je bayesovska interpretacija verjetnosti manj znana. Bralcu, ki se prvič srečuje z bayesovsko interpretacijo, priporočamo ogled zanimivega in poučnega pogleda na dve vrsti nedoločenosti, statistike in statistikov (O'Hagan, 2004). Za poglobljenejšo razlago bayesovske statistike je potrebno poseči po dodatni literaturi, npr. (Jaynes, 2003; de Finetti, 1974, 1975).

[6] Vrednosti naključne spremenljivke, kjer doseže njena porazdelitev največjo vrednost, rečemo *modus* oz. *vrh*.

[7] Priorju, pri katerem so vse vrednosti enako verjetne, pravimo neinformativni prior.

vrednost. To točko označimo z $\hat{\theta}_{\mathrm{ML}}$ in ji pravimo ocena po kriteriju največjega verjetja (angl. maximum likelihood, ML). Takšno točkovno oceno lahko predstavimo z Dirac-delta porazdelitvijo. Takrat lahko prediktivno porazdelitev $p(x|\mathcal{D})$ zapišemo kot:

$$p(x|\mathcal{D}) = \int_{\Omega_\Theta} p(x|\theta)\delta(\theta - \hat{\theta}_{\mathrm{ML}})\mathrm{d}\theta,$$

kar se zaradi vzorčne lastnosti Diracove porazdelitve poenostavi v $p(x|\hat{\theta}_{\mathrm{ML}})$. Vidimo, da v primeru točkovne ocene parametrov preide prediktivna porazdelitev v verjetje.

Slabost ML ocene je, da ni imuna na prenaučitev, kar pomeni, da bodo kompleksnejši modeli, t.j. modeli z več parametri, v splošnem dosegali višje vrednosti verjetja kot manj kompleksni modeli. To ni v skladu s splošno sprejeto znanstveno paradigmo, ki pravi, da če imamo na voljo dve teoriji (modela), ki enako dobro opišeta eksperiment (podatke), potem je bolj verjetna enostavnejša izmed obeh teorij[8]. To pomanjkljivost se poskuša omiliti tako, da uvedemo t.i. kazenski člen, s katerim ›kaznujemo‹ kompleksnost modela. V primeru, ko imamo namesto točkovne ocene na voljo pravo porazdelitev, takšen ad hoc popravek ni potreben, saj je implicitno vsebovan pri integraciji preko posteriorne porazdelitve.

1.7 O učenju in razpoznavanju

Poglejmo si splošni teoretični okvir, v katerem bomo v disertaciji obravnavali problem samodejnega razpoznavanja oseb na podlagi govora. Imejmo parametrični statistični model $\mathcal{M}$, s katerim želimo opisati značilnosti posameznega govorca. Naša naloga je, da iz učnih podatkov $\mathcal{D}$, t.j. govornega posnetka danega govorca, ocenimo parametre statističnega modela. V okviru bayesovskega učenja to storimo tako, da apriorno porazdelitev parametrov $p(\theta|\mathcal{M})$ modela $\mathcal{M}$ preslikamo v posteriorno porazdelitev $p(\theta|\mathcal{D}, \mathcal{M})$ tako, da upoštevamo Bayesov izrek, ki pravi:

$$p(\theta|\mathcal{D}, \mathcal{M}) = \frac{p(\mathcal{D}|\theta, \mathcal{M})p(\theta|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})}. \tag{1.5}$$

Kadar imamo na voljo več učnih posnetkov danega govorca, imamo dve možnosti, ki pripeljeta do identičnih rezultatov, t.j. posteriornih porazdelitev. Prva možnost je, da vse posnetke združimo in izračunamo posteriorno porazdelitev le enkrat. Takemu načinu rečemo *paketno* (angl. batch) učenje. Druga možnost pa je, da za vsak posnetek posebej izpeljemo posteriorno porazdelitev, pri čemer vzamemo posteriorno porazdelitev iz prejšnjega koraka kot apriorno porazdelitev naslednjega koraka. Takemu načinu ocenjevanja pravimo *postopno* (angl. offline) učenje, ki je še posebej primerno v realno-časovnih sistemih, ki so sposobni sprotnega prilagajanja na spremembe okolja (npr. avtonomni roboti).

Razpoznavanje najlaže predstavimo na problemu razvrščanja, kjer želimo testni posnetek neznanega govorca pripisati natanko enemu izmed nam znanih govorcev.

[8] Temu principu pravimo Okamova britev.

Tako definiranemu razvrščanju na področju biometrije pravimo problem identifikacije. Ta problem lahko predstavimo v okviru bayesovskega testiranja hipotez, čemur pravimo tudi *primerjava modelov* (angl. model comparison).

Denimo, da imamo v množici govorcev M različnih govorcev, za katere smo naučili M različnih statističnih modelov $\{\mathcal{M}_i\}_{i=1}^{M}$. Zanima nas, kakšna je verjetnost, da je testni posnetek x izgovoril i-ti govorec, kar zapišemo s pogojno verjetnostjo $p(\mathcal{M}_i|x)$. Teh verjetnosti ne moremo izračunati neposredno, lahko pa si pomagamo z Bayesovim izrekom:

$$p(\mathcal{M}_i|x) = \frac{p(x|\mathcal{M}_i)p(\mathcal{M}_i)}{p(x)}$$

Količine $p(\mathcal{M}_i)$ so apriorne verjetnosti in jih interpretiramo kot verjetnosti posameznih modelov, še preden je govorec v testnem posnetku spregovoril. Količina $p(x)$ je normalizacijska konstanta, ki zagotovi, da izraz $p(\mathcal{M}_i|x)$ zadošča lastnosti verjetnostne porazdelitve, t.j. $\sum_i p(\mathcal{M}_i|x) = 1$. Pogosto nas sama vrednost $p(x)$ ne zanima in je ni potrebno izračunati, saj vemo, da velja $p(\mathcal{M}_i|x) \propto p(x|\mathcal{M}_i)p(\mathcal{M}_i)$. Če predpostavimo še, da so vsi modeli a priori enako verjetni, sledi, da so posteriorne verjetnosti modelov premo sorazmerne verjetju $p(x|\mathcal{M}_i)$, t.j. $p(\mathcal{M}_i|x) \propto p(x|\mathcal{M}_i)$. Vidimo torej, da je verjetje naravna mera podobnosti, ki jo lahko uporabimo za oceno kakovosti prileganja podatkov (testnega posnetka) na govorski model.

Opazimo lahko, da smo verjetje modela že srečali v imenovalcu izraza za posteriorno porazdelitev parametrov (**1.5**) pri učenju modela govorca, le da smo takrat verjetje namesto pri testnih podatkih x izračunali pri učnih podatkih $\mathcal{D}$. Takrat je verjetje služilo le kot normalizacijska konstanta, sedaj pa ga potrebujemo za primerjavo modelov. Verjetje $p(x|\mathcal{M}_i)$ modela $\mathcal{M}_i$ izračunamo tako, da seštejemo posamezne prispevke preko celotnega prostora parametrov:

$$p(x|\mathcal{M}_i) = \int_{\Omega_\Theta} p(x, \theta|\mathcal{M}_i)\mathrm{d}\theta.$$

Z upoštevanjem pravila produkta verjetnosti lahko gornji izraz zapišemo še drugače:

$$p(x|\mathcal{M}_i) = \int_{\Omega_\Theta} p(x|\theta, \mathcal{M}_i)p(\theta|\mathcal{M}_i)\mathrm{d}\theta,$$

čemur pravimo (posteriorna) prediktivna porazdelitev.

Velikokrat zaradi enostavnosti ne ocenimo celotne porazdelitve $p(\theta|\mathcal{M}_i)$, ampak izračunamo le točkovno oceno parametrov, npr. $\hat{\theta}_{\mathrm{ML}}$. Takrat velja:

$$p(x|\mathcal{M}_i) = p(x|\hat{\theta}_{\mathrm{ML}}, \mathcal{M}_i)$$

Če moramo testni posnetek pripisati enemu izmed govorcev, bomo to storili tako, da bomo med seboj primerjali vrednosti verjetij $p(x|\mathcal{M}_i)$ in sprejeli odločitev, da je pravilna tista hipoteza, pri kateri je vrednost verjetja največja: $\arg\max_i p(x|\mathcal{M}_i)$.

Postopek razpoznavanja lahko posplošimo tudi na primer, kjer govorec v testnem posnetku ni nujno eden izmed nam znanih govorcev. Ker za neznanega govorca nimamo učnih podatkov, s pomočjo katerih bi ocenili posteriorno porazdelitev,

je najbolj naravna izbira, da hipotezo, da govorec ne spada v množico nam znanih govorcev, ovrednotimo kar z apriorno porazdelitvijo. Tak primer srečamo tudi, ko se ukvarjamo z verifikacijo govorcev, kjer za vsak testni posnetek tehtamo med dvema hipotezama: (i) govorec v testnem posnetku je enak govorcu iz učnega posnetka in (ii) govorec v testnem posnetku ni enak govorcu iz učnega posnetka.

1.8 Pregled vsebine disertacije

V prvem poglavju smo predstavili temo in izpostavili glavne cilje doktorske disertacije. V drugem poglavju bomo našteli in na kratko povzeli izvirne prispevke doktorske disertacije. V tretjem poglavju bomo opisali referenčni sistem za razpoznavanje govora, ki bo služil kot izhodišče za primerjavo z vsemi nadaljnimi postopki, s katerimi se bomo ukvarjali v disertaciji. V tem poglavju bomo predstavili vse podsklope, ki jih je treba udejanjiti pri gradnji razpoznavalnika govorcev. V četrtem poglavju bomo podrobno obdelali statistični model mešanice Gaussovih porazdelitev, ki ima osrednje mesto v disertaciji. Tema petega poglavja bodo mere, s katerimi merimo podobnost med govornimi posnetki oz. govorci. V šestem poglavju se bomo posvetili postopku projekcije motečih lastnosti, prvi izmed dveh metod, uporabljenih za zmanjšanje vpliva neenakih razmer v učnem in testnem posnetku. Drugo metodo normalizacije, analizo vezanih faktorjev, bomo predstavili v sedmem poglavju. V osmem poglavju bomo opisali opravljene eksperimente in podali dobljene rezultate. Zaključne misli bomo podali v devetem poglavju. V dodatku bomo predstavili splošen postopek maksimizacije verjetja. Obdelali bomo model faktorske analize in s pomočjo variacijske metode izpeljali enačbe analize vezanih faktorjev.

IZVIRNI PRISPEVKI K ZNANOSTI 2

(i) Metoda računanja verjetja modela mešanice Gaussovih porazdelitev na podlagi zadostne statistike
(ii) Odločitveni kriterij temelječ na Kullback-Leiblerjevi divergenci
(iii) Kriterij razmerja verjetij za normalizacijo kanala z uporabo postopka projekcije motečih lastnosti
(iv) Kriterij metode podpornih vektorjev za normalizacijo kanala s postopkom analize vezanih faktorjev
(v) Postopek fuzije kriterijev razmerja verjetij in metode podpornih vektorjev pri razpoznavanju govorcev

(i) *Metoda računanja verjetja modela mešanice Gaussovih porazdelitev na podlagi zadostne statistike*

Najbolj pogosto uporabljena mera podobnosti na področju razpoznavanja govorcev je verjetje, ki je hkrati tudi najbolj naravna mera podobnosti za statistični model mešanice Gaussovih porazdelitev. Slabost verjetja je, da je izračun vrednosti verjetja časovno zelo potraten, še posebej, če sta razsežnost značilk in število komponent mešanice veliki. Če želimo v sistem vključiti še postopek normalizacije rezultatov prileganja, postane potreben čas za izvedbo eksperimenta nezanemarljiv.

Pogost prijem za pohitritev izračuna verjetja, ki se je uveljavil na področju razpoznavanja govorcev, je, da pri izračunu verjetja namesto prispevkov vseh komponent mešanice upoštevamo le nekaj najbolj verjetnih komponent, ki jih za vsak testni posnetek poiščemo le enkrat. V ta namen uporabimo splošni model govorca, katerega parametre ocenimo iz obsežne govorne zbirke. Predpostavka, na kateri temelji ta metoda, je, da je vrstni red komponent, urejenih po (posteriornih) verjetnostih, neodvisen od govorca.

Sami predlagamo postopek, ki je računsko še precej bolj učinkovit kot metoda najverjetnejših komponent. Ob predpostavki, da so pri danem posnetku posteriorne verjetnosti komponent (približno) neodvisne od govorca, lahko verjetje zapišemo kot funkcijo zadostne statistike. Za izračun verjetij poljubnega števila govorcev je tako zadostno statistiko za vsak testni posnetek potrebno izračunati le enkrat, za kar ponovno uporabimo splošni model govorca.

(ii) *Odločitveni kriterij temelječ na Kullback-Leiblerjevi divergenci*

Verjetje je po svoji naravi nesimetričen postopek, saj prilega podatke na model. To pomeni, da se vlogi učnega in testnega posnetka razlikujeta. Medtem ko učni posnetek uporabimo za učenje statističnega modela, testni posnetek uporabimo tako,

da ga prilegamo na model. V principu ni nobenega posebnega razloga, da ne bi mogli vlogi obeh posnetkov obrniti in uporabiti testni posnetek za oceno modela, na katerega bi prilegali učni posnetek[9].

V disertaciji za merjenje podobnosti med dvema posnetkoma predlagamo metodo, ki temelji na merjenju divergence med dvema porazdelitvama. Pri tej metodi, za razliko od verjetja, ocenimo model mešanice Gaussovih porazdelitev tako za učni kot tudi za testni posnetek. Rezultat prileganja izračunamo tako, da ocenimo vrednost divergence med modeloma učnega in testnega posnetka. Natančne vrednosti Kullback-Leiblerjeve divergence med dvema modeloma mešanice Gaussovih porazdelitev analitično ne znamo izračunati, ampak se zadovoljimo z oceno v obliki zgornje meje. Ker je divergenca mera različnosti (manjša pri bolj podobnih porazdelitvah in večja pri manj podobnih), moramo vrednosti divergence spremeniti predznak.

(iii) *Kriterij razmerja verjetij za normalizacijo kanala z uporabo postopka projekcije motečih lastnosti*

Postopek projekcije motečih lastnosti, ki obravnava povprečne vektorje modela mešanice Gaussovih porazdelitev kot supervektorje, je bil predlagan v kombinaciji z mero podobnosti, ki temelji na metodi podpornih vektorjev. V disertaciji predlagamo način, kako metodo projekcije motečih lastnosti uporabiti tudi v navezi z mero podobnosti, ki temelji na razmerju verjetij. To storimo tako, da najprej oba posnetka, tako učnega kot tudi testnega, preslikamo v prostor supervektorjev. Z metodo projekcije motečih lastnosti nato oba supervektorja razčlenimo na govorski in kanalski komponenti. Govorski komponenti učnega supervektorja nato prištejemo kanalsko komponento testnega supervektorja in dobljeni supervektor pretvorimo ponovno v model mešanice Gaussovih porazdelitev. Na ta način dobimo govorski model, ki je prilagojen akustičnim razmeram v testnem posnetku. Tako prilagojen model nato uporabimo za izračun verjetja na običajen način.

(iv) *Kriterij metode podpornih vektorjev za normalizacijo kanala s postopkom analize vezanih faktorjev*

Podobno kot je bil postopek projekcije motečih lastnosti predlagan v navezi s kriterijem podpornih vektorjev, tako je bil postopek analize vezanih faktorjev predlagan v kombinaciji s kriterijem razmerja verjetij. Po našem mnenju ni nikakršnega tehtnejšega razloga, da bi postopek normalizacije vezali na točno določeno metodo merjenja podobnosti, ampak lahko ta kriterij poljubno izberemo. Tako v disertaciji k postopku analize vezanih faktorjev predlagamo odločitveni kriterij na podlagi metode podpornih vektorjev, kot alternativo že uveljavljenemu kriteriju razmerja verjetij. Učinkovitost predlaganega kriterija prikažemo v eksperimentih.

[9] To je popolnoma veljavna strategija tudi s stališča NIST-ovega protokola. Npr., možna strategija bi bila, da bi za učenje izbrali daljšega od obeh posnetkov ali da bi celo upoštevali obe možnosti in izračunali navzkrižno verjetje.

(v) *Postopek fuzije kriterijev razmerja verjetij in metode podpornih vektorjev pri razpoznavanju govorcev*

Znano je, da lahko z združevanjem rezultatov prileganja heterogenih sistemov občutno izboljšamo rezultate razpoznavanja. Izboljšanje rezultatov bo tem večje, tem bolj bodo sistemi med seboj komplementarni.

V disertaciji pokažemo, da se da rezultate razpoznavanja izboljšati tudi z združevanjem rezultatov prileganja dveh bolj ali manj enakih sistemov, ki se razlikujeta le v kriteriju, na podlagi katerega merita podobnost med dvema posnetkoma oz. govorcema.

REFERENČNI SISTEM 3

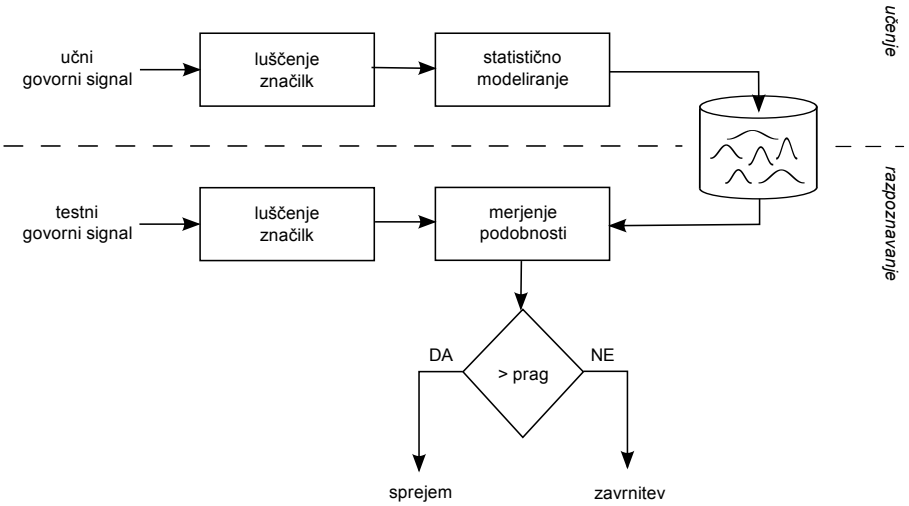
V tem poglavju bomo predstavili referenčni sistem, ki nam bo služil kot izhodišče pri razvoju naprednejših pristopov na področju razpoznavanja govorcev. Referenčni sistem, ki ga bomo zgradili, bo približno tak, kot je bil ›state-of-the-art‹ sistem v času, ko smo se sami začeli intenzivneje ukvarjati s problemom razpoznavanja govorcev (Reynolds et al., 2000).

Podrobno si bomo ogledali vse korake, ki so potrebni za implementacijo sistema za samodejno razpoznavanje govorcev. Med opisom bomo skušali opozoriti na ključne pomanjkljivosti uporabljenih pristopov in nakazati možnosti za njihovo morebitno kasnejšo izboljšavo.

3.1 Shema sistema za verifikacijo

Sistem za razpoznavanje govorcev lahko deluje v režimu ugotavljanja (angl. identification) ali v režimu preverjanja identitete (angl. verification). Ker je postopek preverjanja identitete splošnejši od postopka ugotavljanja identite, bomo predstavili le prvega.

Ponavadi postopek preverjanja identitete osebe na podlagi govora (ali kakšne druge biometrične lastnosti) predstavimo v dveh korakih. V prvem koraku (faza *učenja*) govorni signal, za katerega vemo, kateremu govorcu pripada, najprej pretvorimo v niz vektorjev značilk, s pomočjo katerega naučimo statistični model govorca (zgornji del slike **3.1**).



Slika 3.1 Shema biometričnega sistema za verifikacijo govorcev.

V drugem delu, ki je predstavljen na spodnjem delu slike **3.1**, pa sistemu na vhodu predstavimo dva podatka: testni govorni signal in identiteto izbranega govorca. Signal pretvorimo v niz vektorjev značilk na enak način, kot smo to storili v postopku učenja statističnega modela izbranega govorca. V naslednjem koraku izmerimo podobnost med nizom vektorjev značilk in modelom. Izmerjeni podobnosti pogosto rečemo rezultat prileganja (angl. matching score). Ta rezultat prileganja nato primerjamo z vnaprej določenim pragom. Če je rezultat prileganja večji od tega praga, potem sistem testni posnetek sprejme, v nasprotnem primeru pa ga zavrne.

3.2 Predobdelava govornega signala

3.2.1 Parametrizacija

Surov govorni signal ni neposredno primeren za razpoznavanje identitete govorcev, ampak ga moramo najprej ustrezno obdelati. To storimo tako, da zaporedne prekrivajoče se kratke časovne odseke (tipično 20 do 25 ms) pretvorimo v niz vektorjev značilk. Kot značilke so se uveljavili nekateri parametri, ki se uporabljajo na sorodnem področju razpoznavanja govora, kot so koeficienti melodičnega kepstra (angl. mel frequency cepstral coefficients, MFCC) in koeficienti zaznavne linearne predikcije (angl. perceptual linear prediction, PLP). Čeprav je včasih veljalo prepričanje, da naj bi bile značilke MFCC do neke mere od govorca neodvisne, je uporaba istih značilk na obeh področjih v prid tezi, da značilke, ki bi nosile informacijo izključno o identiteti govorca ali o vsebini sporočila, ne obstajajo.

Postopek luščenja značilk MFCC lahko strnemo v nekaj korakov (glej sliko **3.2**). Govorni signal najprej razkosamo na prekrivajoče se kratke časovne odseke dolžine od 20 do 25 ms. Za vsak odsek nato po predhodnem oknjenju s Hammingovim oknom izračunamo amplitudni frekvenčni spekter. Tega nato spustimo skozi vrsto filtrov, ki jih razporedimo enakomerno po melodični skali. Na koncu sledi še logaritmiranje izhodov filtrov in diskretna kosinusna transformacija, za katero je bilo ugotovljeno, da učinkovito dekorelira posamezne komponente kepstra.

Slika 3.2 Shema pridobivanja značilk iz govornega signala.

Postopek za izračun značilk MFCC je predlagal že (Pols, 1977). Pri tem se je zgledoval po fizioloških značilnostih slušnih organov. Tako pasovnih filtrov, ki prepuščajo le določen frekvenčni pas spektra in so namenjeni glajenju amplitudnega spektra, ni razporedil linearno po frekvenčnem območju, ampak jih je razporedil po melodični skali. S tem je želel posnemati frekvenčno občutljivost bazilarne membrane kohlee v

notranjem ušesu. Podobno velja za logaritmiranje izhodov filtrov, s čimer se je želel približati logaritemski jakostni občutljivosti ušesa.

Zanimivo je, da so bile značilke MFCC najprej predlagane za (samodejno) tvorjenje govora kot alternativa značilkam linearne predikcije (angl. linear prediction coefficients, LPC) in so se šele kasneje izkazale uporabne tudi pri razpoznavanju govora. Številni poskusi, da bi jih nadomestili z značilkami, ki bi bile bolj primerne za razpoznavanje govora, so se izkazali za več ali manj jalove. Nekaj uspeha je imel Hermansky[10], ki je predlagal značilke *zaznavne linearne predikcije* (angl. perceptual linear prediction, PLP) (Hermansky, 1990), za katere je bilo ugotovljeno, da so nekoliko robustnejše v šumnih razmerah.

Če izračunamo značilke melodičnega kepstra na nekoliko drugačen način (Tokuda et al., 1995a), lahko iz njih ponovno rekonstruiramo prvotni signal v časovnem prostoru (Imai, 1983). To lastnost izkoriščajo vse bolj popularni statistični sistemi za umetno tvorjenje govora (Tokuda et al., 1995b; Vesnicer in Mihelič, 2004; Zen et al., 2009), ki zaradi številnih dobrih lastnostih vse pogosteje nadomeščajo bolj klasične (nestatistične) difonske sisteme.

Pogosto osnovnim (statičnim) značilkam dodamo tudi dinamične značilke oz. regresijske koeficiente, ki opisujejo dinamiko spreminjanja osnovnih značilk. Ta poteza je s stališča kasnejšega statističnega modeliranja sporna, saj tako med sosednje značilke neposredno vnesemo analitično odvisnost, čeprav statistični model ponavadi predpostavlja statistično neodvisnost. Rešitev iz te zagate so pred kratkim predlagali (Zen et al., 2006), kjer so pokazali, da lahko model HMM redefiniramo kot trajektorni model tako, da upoštevamo eksplicitne odvisnosti med statičnimi in dinamičnimi značilkami.

3.2.2 Normalizacija značilk

Neenakost akustičnih razmer med učnim in testnim posnetkom lahko privede do poslabšanja zanesljivosti razpoznavanja. Da bi zmanjšali občutljivost razpoznavalnika na različne akustične vplive, je bilo predlaganih vrsta postopkov, ki skušajo signal ali značilke transformirati tako, da bi izboljšali robustnost razpoznavalnika. Med te metode prištevamo odštevanje kepstralnega povprečja (angl. cepstral mean subtraction, CMS), normalizacija povprečja in variance kepstra (angl. mean variance normalization, MVN), filtriranje RASTA (Hermansky in Morgan, 1994), na področju razpoznavanja govorcev pa se je v zadnjem času še posebej uveljavila metoda ukrivljanja značilk (angl. feature warping), ki temelji na normalizaciji histograma značilk (Pelecanos in Sridharan, 2001). Podobno, a splošnejšo metodo, srečamo tudi pod imenom kratkočasovne Gaussianizacije (angl. short-time Gaussianization) (Xiang et al., 2002).

[10] Profesor Hynek Hermansky je znan raziskovalec s področja razpoznavanja in kodiranja govora. Med številnimi dosežki je objavil tudi članek (Hermansky, 1998) s pomenljivim naslovom, katerega slovenski prevod se glasi: ›Naj imajo razpoznavalniki govora ušesa?‹, v katerem špekulira, da mogoče razpoznavalniki govora ne bi smeli preveč posnemati človeškega sluha, saj tudi letala ne mahajo s krili, pa vseeno letijo.

Morda velja opozoriti na dejstvo, da med raziskovalci ni enotnega mnenja, kdaj izvesti normalizacijo značilk (nekateri jo izvedejo pred izračunom dinamičnih značilk, medtem ko drugi šele po izračunu le-teh) in ali to storiti pred ali po odstranjevanju negovornih delov.

Večino metod normalizacije značilk je mogoče izvesti na dva načina; bodisi globalno, preko celotnega posnetka[11], bodisi lokalno, z uporabo drsečega okna (Ouellet in Kenny, 2005). Pri metodi ukrivljanja značilk se je prijela uporaba drsečega okna dolžine 3 s.

Nobena izmed metod normalizacije značilk seveda ni brez ›stranskih učinkov‹. To pomeni, da se z normalizacijo značilk sicer do neke mere znebimo neželenih vplivov, a po drugi strani poslabšamo tudi ločevanje med različnimi govorci. Hitro se lahko prepričamo, da se normalizacija značilk splača le pri ne dovolj čistih govornih signalih, medtem ko pri čistih rezultate razpoznavanja ponavadi poslabša.

3.2.3 Izločanje negovornih odsekov

Eden izmed zelo pomembnih korakov predobdelave je, da iz govornega signala izločimo negovorne odseke. Če tega ne bi storili, bi se lahko zgodilo, da bi namesto govorca razpoznavali akustično ozadje, v katerem je bil signal zajet. Če si predstavljamo analogijo z razpoznavanjem oseb na osnovi slik obrazov, so negovorni deli v signalu podobni ozadju na sliki, ki ga je pred izvedbo razpoznavanja obrazov potrebno odstraniti.

Na videz je izločanje negovornih delov trivialen problem, a se izkaže, da generičnega postopka, ki bi z dovolj veliko zanesljivostjo ločil govor od negovora, ne poznamo. Zato se ponavadi zatečemo k raznim hevurističnim postopkom, ki temeljijo na merjenju kratkočasovne energije signala.

Druga skupina postopkov izločanja negovornih delov temelji na informaciji, ki jo dobimo s pomočjo razpoznavalnika govora. Ker razpoznavalnik govora kot rezultat vrne niz razpoznanih besed skupaj s časovnimi mejami, lahko le-te uporabimo za izločanje negovornih odsekov. Čeprav se zdi, da je v osnovi takšen način ločevanja govora od negovora ›pravilnejši‹, pa lahko ima tak pristop tudi pomanjkljivosti. Če zanemarimo dejstvo, da razpoznavalnika govora nimamo vedno pri roki, se moramo zavedati, da je razpoznavanje govora (še posebej spontanega, ki je posnet v različnih akustičnih ozadjih) nujno podvrženo napakam, zaradi katerega lahko postane določanje govornih odsekov nezanesljivo. Druga, a ne manj pomembna ovira, na katero naletimo, če želimo uporabiti splošen razpoznavalnik govora, je, da je razpoznavanje besed računsko zahteven in s tem časovno potraten proces. Tako bi lahko čas, ki bi ga potrebovali za izločanje negovornih delov, pomenil velik delež skupnega časa, ki bi ga potrebovali za celoten postopek razpoznavanja govorcev.

V izogib tem oviram je bila predlagana rešitev za ločevanje govornih in negovornih delov, ki namesto splošnega razpoznavalnika besed uporablja precej preprostejši

[11] Ta način ni primeren za realnočasovne sisteme oz. sisteme, kjer si ne moremo privoščiti daljših časovnih zamikov. Tak primer je sprotno podnaslavljanje s samodejnim razpoznavalnikom govora.

razpoznavalnik glasov (Žibert et al., 2006). Prednost takega pristopa je, da je časovno manj potraten in da je praktično neodvisen od jezika, kar je bilo potrjeno z eksperimenti.

Pri našem delu smo eksperimentirali z različnimi načini izločanja negovornih delov, a smo se nazadnje odločili, da bomo uporabili kar besedne prepise, ki so kot del zbirke na voljo skupaj z zvočnimi posnetki. Opozoriti velja, da so te oznake bile pridobljene z razpoznavalnikom angleškega govora, za katerega navajajo tipično napako besed (angl. word error rate, WER) med 15 in 30 %.

3.3 Statistično modeliranje akustičnih značilnosti govorca

Čeprav so si vsi postopki, ki spadajo v skupino biometričnega razpoznavanja oseb, v osnovi med seboj podobni, pa obstaja med njimi tudi nekaj pomembnih razlik. Tako se razpoznavanje govorcev od nekaterih drugih biometričnih postopkov loči po časovni komponenti, ki je prisotna v govornem signalu. Poskušajmo osvetliti nekatere posledice, ki jih povzroči ta ›časovna‹ narava govornega signala.

Spoznali smo, da lahko govorni signal predstavimo v obliki niza vektorjev značilk. Predstavljamo si lahko, da vsak vektor značilk, ki je element tega niza, nosi informacijo o stanju govoril v danem trenutku. Jasno je, da le eden vektor značilk ne nosi prav veliko informacije o načinu govora posameznega govorca. Z vsakim dodatnim vektorjem značilk izvemo nekaj več o tem govorcu. Popolno ›sliko‹ govorca bomo v teoriji dobili le, ko bo dolžina niza vektorjev značilk šla preko vseh meja. V praksi pa bomo imeli na razpolago vedno le niz končne dolžine, zato bomo razpolagali le z nepopolno ›sliko‹ govorca.

Po tej lastnosti se razpoznavanje oseb na podlagi govora razlikuje od nekaterih drugih biometričnih postopkov, ki temeljijo predvsem na časovno-statičnih lastnostih. Med takšne spada večina biometričnih metod, ki za razpoznavanje oseb uporabljajo slikovno informacijo, npr. razpoznavanje obrazov, prstnih odtisov, šarenice itd.

Razliko med dinamično in statično informacijo lahko nazorno ponazorimo z naslednjim miselnim eksperimentom. Denimo, da imamo fotografski aparat z zelo omejenim zornim kotom, ki se v vsakem trenutku usmeri na naključno mesto in posname majhno sličico. Če to snemanje ponavljamo nekaj časa, dobimo niz sličic, iz katerih želimo sestaviti sliko obraza. Pri tem se moramo zavedati, da zaradi končnega časa snemanja najverjetneje nismo zajeli dovolj sličic, da bi lahko sestavili sliko celotnega obraza in da za posamezno sličico ne vemo a priori, iz katerega dela obraza (ali celo ozadja) izhaja oz. povedano drugače, sličice niso opremljene s podatkom, na katero mesto je bil usmerjen aparat, ko je zajel sličico. Opisani miselni eksperiment je zelo podoben realni situaciji, s katero se soočamo pri besedilno neodvisnemu razpoznavanju govorcev.

Za razpoznavanje časovnih vrst oz. soodvisnih vzorcev, ki so predstavljeni v obliki niza vektorjev značilk, se pogosto uporabljajo prikriti Markovovi modeli (angl. hidden Markov models, HMM). Modeli HMM predstavljajo zelo vsestransko uporabno in prilagodljivo orodje. Širše znani so postali z razmahom raziskav na področju

samodejnega razpoznavanja govora v osemdesetih in devetdesetih letih prejšnjega stoletja. Čeprav je od razvoja teorije in prvih poskusov uporabe modelov HMM preteklo že precej časa, še danes praktično vsi najboljši razpoznavalniki govora temeljijo ravno na njih.

Zaradi podobnosti med razpoznavanjem govora in razpoznavanjem govorcev se zdi logično pričakovati, da bodo modeli HMM prava izbira tudi za razpoznavanje govorcev. To seveda drži za tekstovno odvisno razpoznavanje govorcev, ki se npr. uporablja za kontrolo dostopa, kjer od govorca zahtevamo, da izreče nek vnaprej pripravljen tekst. Pri tekstovno neodvisnem razpoznavanju govorcev pa nas ne zanima, kaj je bilo povedano, ampak le, kdo je to povedal. Zato se je izkazalo, da je za tekstovno neodvisno (angl. text independent) razpoznavanje bolje vzeti enostavnejši model, ki ga bomo predstavili v naslednjem razdelku.

3.3.1 Model mešanice Gaussovih porazdelitev

Na področju (akustičnega) razpoznavanja govorcev se je kot model, s katerim opišemo značilnosti govorca, uveljavil *model mešanice Gaussovih porazdelitev* (angl. Gaussian mixture model, GMM). Ta model temelji na predpostavki, da so naključne spremenljivke (akustični vektorji) med seboj *neodvisni in enako porazdeljeni*[12] (angl. independent and identically distributed, i.i.d.). Ne glede na to, da je jasno, da pri govoru ta predpostavka še zdaleč ni izpolnjena, se je v praksi ta model izkazal za zelo uspešnega.

Parametrični model GMM nam omogoča, da z njim opišemo poljubno porazdelitev, če le uporabimo dovolj komponent v mešanici. Posamezne komponente imajo v splošnem polne kovariančne matrike, te pa za zanesljivo oceno zahtevajo zadostno število podatkov. V praksi se je izkazalo (Reynolds et al., 2000), da je bolje, če manjše število komponent s polnimi kovariančnimi matrikami nadomestimo z večjim številom komponent z diagonalnimi kovariančnimi matrikami. Hkrati se na ta način izognemo netrivialnemu invertiranju polnih kovariančnih matrik, s čimer privarčujemo precej dragocenega časa.

3.3.2 Ocenjevanje parametrov modela GMM

V modelu GMM, ki je sestavljen iz K komponent, kot parametri nastopajo uteži $\{\pi_k\}_{k=1}^K$, povprečni vektorji $\{\boldsymbol{\mu}_k\}_{k=1}^K$ in kovariančne matrike $\{\boldsymbol{\Sigma}_k\}_{k=1}^K$, ki jih iz podatkov, ki jih imamo na voljo, ocenimo po kriteriju največjega verjetja (angl. maximum likelihood, ML). Žal analitična formula, po kateri bi lahko v enem koraku izračunali te parametre, ne obstaja. Pomagati si moramo z iterativnim postopkom *maksimizacije upanja* (angl. expectation maximization, EM) (Dempster et al., 1977), ki nam zagotavlja monotono naraščanje verjetja in s tem konvergenco k lokalnemu maksimumu (glej dodatek **A**).

[12] Splošnejša in manj stroga je predpostavka o *zamenljivosti* (angl. exchangeability), ki igra pomembno vlogo v moderni teoriji verjetnosti.

V tem razdelku opisujemo le izkustveno interpretacijo modela GMM, bolj teoretično obarvano obravnavo pa podajamo v poglavju 4.

3.3.3 Verjetje modela GMM

Mera podobnosti, s katero ocenimo kakovost prileganja testnih podatkov $\boldsymbol{x}$ na model GMM, je verjetje $\mathcal{L}(\theta|\boldsymbol{x})$. Ponavadi zaradi numeričnih razlogov izračunamo logaritem verjetja:

$$\begin{aligned}\log \mathcal{L}(\theta|\boldsymbol{x}) = \log p(\boldsymbol{x}|\theta) &= \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta) \\ &= \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),\end{aligned}$$

kjer smo s θ označili vse parametre, ki nastopajo v modelu GMM: $\theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$.

Vidimo, da logaritem verjetja pri testnem posnetku $\boldsymbol{x}$ izračunamo kot vsoto prispevkov preko vseh elementov $\boldsymbol{x}_n$ posnetka $\boldsymbol{x}_{1:N}$. Ta vsota ima pomembno posledico, saj nam centralni limitni izrek zagotavlja, da bo vsota N i.i.d. naključnih spremenljivk limitirala proti normalni porazdelitvi, ko bo N dovolj velik. Tako bo vsaj v teoriji verjetje modela pri večjem številu različnih testnih posnetkih posnemalo normalno porazdelitev, če bodo le izpolnjeni potrebni pogoji. Podobno bodo normalno porazdeljena tudi verjetja različnih modelov pri istem testnem posnetku.

Vrednosti verjetij pri različnih testnih podatkih same po sebi niso primerne za neposredno primerjavo, zato moramo opazovati *razmerje verjetij* (angl. likelihood ratio, LR), kjer verjetje normiramo glede na referenčno vrednost. Pri verifikaciji referenčna vrednost verjetja ustreza hipotezi, da v testnem posnetku ne nastopa isti govorec kot v učnem posnetku. To hipotezo lahko predstavimo s končno množico modelov govorcev — kohorto. Tak način je bil v veljavi, vse dokler ni (Reynolds et al., 2000) predlagal alternativnega pristopa, kjer kohorto zamenjamo z enim samim modelom, v učenje katerega vključimo govorne posnetke velikega števila govorcev. Takemu – od govorca neodvisnemu (angl. speaker independent, SI) modelu – rečemo *splošni model govorca* (angl. universal background model, UBM) ali tudi model sveta (angl. world model). Izkazalo se je, da je ravno model UBM temeljni gradnik vseh trenutno najbolj konkurenčnih sistemov za razpoznavanje govora.

Če namesto produkta verjetij izračunamo geometrijsko sredino verjetij, se znebimo odvisnosti od dolžine govornega posnetka. Pogosto namesto razmerja verjetij izračunamo logaritem tega razmerja. Takrat geometrijska sredina preide v aritmetično sredino, kar lahko zapišemo s formulo:

$$\frac{1}{N} \log \frac{\mathcal{L}(\theta_\mathrm{s}|\boldsymbol{x})}{\mathcal{L}(\theta_0|\boldsymbol{x})},$$

kjer smo s θ_s označili parametre govorskega modela, s θ_0 pa parametre modela UBM.

3.4 Učenje modela UBM

Učenje UBM modela izvedemo z iterativnim postopkom *maksimizacije upanja*, ki ga podrobneje predstavimo v poglavju 4. Ker nam le-ta zagotavlja konvergenco verjetja k lokalnemu maksimumu, je končna ocena parametrov zelo odvisna od inicializacije, ki jo lahko izvedemo na več načinov. Eden izmed njih je inicializacija s postopkom rojenja k-tih povprečij (angl. k-means clustering), drug pa s hierarhičnim postopkom Linde-Buzo-Gray (LBG) (Linde et al., 1980).

Imejmo vektor značilk $\boldsymbol{x}_{1:N}$. Označimo z $\pi_j^{(t-1)}$, $\boldsymbol{\mu}_j^{(t-1)}$ in $\boldsymbol{\Sigma}_j^{(t-1)}$ trenutne vrednosti parametrov j-te komponente modela GMM, izračunane v $(t-1)$-ti iteraciji postopka EM. Potem nove ocene teh parametrov, ki jih označimo z $\pi_j^{(t)}$, $\boldsymbol{\mu}_j^{(t)}$ in $\boldsymbol{\Sigma}_j^{(t)}$, določimo v dveh korakih postopka EM, ki jima pravimo e-korak in m-korak:

i) *e-korak*: izračun aposteriornih verjetnosti γ_{nj} pri trenutnih vrednostih parametrov

$$\gamma_{nj} = \frac{\pi_j^{(t-1)}\,\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j^{(t-1)}, \boldsymbol{\Sigma}_j^{(t-1)})}{\sum_{k=1}^{K} \pi_k^{(t-1)}\,\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)})}$$

ii) *m-korak*: izračun novih ocen parametrov

$$\pi_j^{(t)} = \frac{\sum_{n=1}^{N} \gamma_{nj}}{\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nj}}$$

$$\boldsymbol{\mu}_j^{(t)} = \frac{\sum_{n=1}^{N} \gamma_{nj}\boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma_{nj}}$$

$$\boldsymbol{\Sigma}_j^{(t)} = \frac{\sum_{n=1}^{N} \gamma_{nj}(\boldsymbol{x}_n - \boldsymbol{\mu}_j^{(t)})(\boldsymbol{x}_n - \boldsymbol{\mu}_j^{(t)})^T}{\sum_{n=1}^{N} \gamma_{nj}}$$

Izkaže se, da je za dosego dobrih rezultatov razpoznavanja potrebno uporabiti več sto ur govornih posnetkov čim večjega števila različnih govorcev in da je treba v modelu GMM uporabiti veliko število Gaussovih komponent (v praksi se najpogosteje uporablja 512 ali 2048 komponent). To predstavlja težavo, saj je EM postopek računsko in podatkovno izredno intenziven. Ker bi v našem primeru učenje na enem osebnem računalniku trajalo predolgo, smo razvili lastno rešitev porazdeljenega računanja, s katero smo uspeli čas, ki je bil potreben za ocenitev parametrov 2048 komponent, skrajšati na okoli en mesec[13]. Kljub temu je čas, potreben za učenje modela UBM predolg, da bi si lahko privoščili bolj obsežno eksperimentiranje z različnimi nastavitvami. Velikokrat se tako zgodi, da se v raziskovalni skupini učenje modela UBM izvede le enkrat, vse raziskave pa so osredotočene na kasnejše faze v razvoju razpoznavalnika govorcev. Zato je po našem mnenju v postopku učenja

[13] Če bi imeli na razpolago večje število osebnih računalnikov, bi lahko ta čas skrajšali še precej bolj.

modela UBM odprtih še veliko vprašanj, na katere še nismo poiskali zadovoljivih odgovorov. Nekatera izmed teh vprašanj so:

i) Kakšno je najprimernejše število komponent v modelu UBM?
ii) Kako na uspešnost razpoznavanja vpliva količina podatkov, ki jih uporabimo za učenje modela UBM?
iii) Kako je uspešnost razpoznavanja odvisna od inicializacije oz. od različnih lokalnih optimumov, ki jih dosežemo pri ocenjevanju parametrov modela UBM s postopkom EM?

Pri implementaciji učenja modela UBM smo primorani izbrati nekatere hevristične parametre kot so končno število komponent v mešanici, število iteracij oz. konvergenčni kriterij, najmanjše in največje dovoljene vrednosti elementov v kovariančnih matrikah itd. Ker traja učenje modela UBM predolgo, da bi lahko empirično izbrali najbolj optimalno kombinacijo odprtih parametrov, se moramo zanesti na (velikokrat skope) podatke iz literature in lastno intuicijo.

Na področju razpoznavanja govorcev se je – v nasprotju z ustaljeno prakso s področja razpoznavanja govora – uveljavilo prepričanje, da dosežemo boljše rezultate, če naučimo model UBM ločeno za ženske in moške govorce. Delno se je ločevanje na moške in ženske govorke uveljavilo zaradi dolgoletne tradicije NIST-ovih vrednotenj razpoznavalnikov govorcev, kjer je spol govorca v testnem posnetku vnaprej podan. Kljub temu, da velja problem določevanja spola (angl. gender detection) za bolj ali manj rešen problem, je po našem mnenju delitev glede na spol z znanstvenega stališča neupravičena. Ker takšna delitev hkrati pomeni podvojeno delo za izvedbo eksperimentov, smo se odločili, da v eksperimentih, katerih rezultate bomo podali v disertaciji, uporabimo spolno neodvisen (angl. gender independent, GI) model UBM.

3.5 Učenje govorskega modela

Na prvi pogled se zdi, da bi lahko učenje modela posameznega govorca izvedli na enak način kot učenje modela UBM. Vendar se hitro izkaže, da moramo v primeru od govorca odvisnega (angl. speaker dependent, SD) modela postopati drugače, saj imamo na voljo bistveno manj govornega materiala (tipično od deset sekund do nekaj minut). Zato kot izhodišče vzamemo model UBM, iz katerega ›izpeljemo‹ model govorca. To izpeljavo izvedemo s pomočjo postopka maksimizacije vrha posteriorne porazdelitve (angl. maximum a posteriori, MAP) (Gauvain in Lee, 1994), ki zagotavlja, da ocenimo samo tiste parametre modela, ki so bili ›videni‹ v učnih podatkih. V nasprotnem primeru bi premajhna količina podatkov privedla do singularnosti oz. do prenaučenega modela.

Eksperimenti so pokazali (Reynolds et al., 2000), da dobimo boljše rezultate, če namesto adaptiranja vseh parametrov, ki nastopajo v modelu, ocenimo le povprečne vektorje. Te adaptiramo po sledeči formuli:

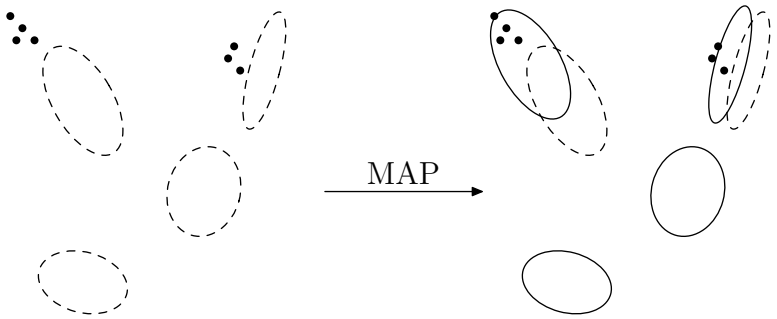
$$\boldsymbol{\mu}_j^{\mathrm{MAP}} = \alpha_j \, \boldsymbol{\mu}_j^{\mathrm{ML}} + (1 - \alpha_j) \, \boldsymbol{\mu}_j^{\mathrm{UBM}}. \tag{3.1}$$

Vidimo, da je novo ocenjena vrednost povprečnega vektorja $\boldsymbol{\mu}_j^{\mathrm{MAP}}$ enaka linearni kombinaciji ML ocene $\boldsymbol{\mu}_j^{\mathrm{ML}}$ in povprečnega vektorja $\boldsymbol{\mu}_j^{\mathrm{UBM}}$. Razmerje njunih prispevkov je določeno s parametrom α_j, ki ga izračunamo po formuli

$$\alpha_j = \frac{\sum_{n=1}^{N} \gamma_{nj}}{\tau + \sum_{n=1}^{N} \gamma_{nj}},$$

kjer je skalar τ eksperimentalno določen *utežni faktor* (angl. relevance factor). Z njim uravnavamo vpliv količine podatkov na MAP oceno parametrov. Pri manjšem τ bo tako ocena parametrov (pri isti količini podatkov) bolj odvisna od podatkov in manj od ocene UBM. Velja tudi obratno, da bo MAP ocena pri večjem τ bolj odvisna od vrednosti UBM in manj od podatkov. V literaturi je zaslediti vrednosti τ med 8 in 16.

Na sliki **3.3** vidimo 2D prikaz adaptacije povprečnih vektorjev modela UBM na učne (govorčeve) podatke (označene z majhnimi polnimi krožci). Rezultat adaptacije so nove ocene le tistih povprečnih vektorjev, ki so bili ›videni‹ v učnih podatkih.



Slika 3.3 Prikaz MAP adaptacije povprečij modela UBM. Prikazane so štiri komponente mešanice v dvorazsežnem prostoru značilk. Črtkano so narisane komponente modela UBM, s polnimi črtami pa komponente po izvedeni adaptaciji. Vidimo, da adaptacija vpliva le na tiste komponente, ki ›opazijo‹ podatke (narisane z majhnimi polnimi krožci).

Ker ML ocene, ki nastopa v enačbi **3.1**, ne moremo izračunati analitično, se moramo ponovno zateči k iterativnemu postopku EM. Za razliko od učenja modela UBM je tukaj konvergenca postopka hitrejša, zato zadošča že nekaj iteracij. V praksi pogosto izvedemo le eno.

3.6 Razpoznavanje in razvrščanje

Razvrščanje testnega posnetka poteka tako, da pri testnih podatkih izračunamo razmerje med verjetjem govorskega modela in verjetjem modela UBM. Če je to razmerje – rezultat prileganja – večje od nekega vnaprej predpisanega praga, testni

posnetek razvrstimo v razred klientov, v nasprotnem primeru pa testni posnetek razvrstimo v razred vsiljivcev.

3.7 Normalizacija rezultatov prileganja

Še pred razvrščanjem lahko rezultat prileganja normiramo. Na področju razpoznavanja govorcev sta se uveljavili dve vrsti normalizacije rezultatov razpoznavanja. To sta z-norm (Reynolds, 1997) in t-norm (Auckenthaler et al., 2000) ter njuna kombinacija zt-norm oz. tz-norm.

3.7.1 Normalizacija z-norm

Motivacija za normalizacijo rezultatov prileganja izvira iz dejstva, da je optimalni prag, ki bi vodil do najboljših rezultatov prileganja, v splošnem za vsakega govorca drugačen. Z drugimi besedami to pomeni, da če bi narisali porazdelitve (histograme) rezultatov prileganja več različnih govorcev pri enakih testnih posnetkih, bi se ti histogrami med seboj razlikovali. Če torej želimo izbrati enoten (globalen) prag za vse učne govorce, je dobro, če rezultate prileganja predhodno normiramo. To storimo tako, da izberemo množico testnih posnetkov (rečemo jim z-norm posnetki), ki jo uporabimo za izračun porazdelitve rezultatov prileganj za vsakega učnega govorca posebej. Če to porazdelitev opišemo z dvema parametroma, povprečjem μ in standardnim odklonom σ, lahko rezultat prileganja s normaliziramo v skladu s sledečo enačbo:

$$s_z = \frac{s - \mu}{\sigma}.$$

V kolikor je bila prvotna porazdelitev rezultatov razpoznavanja normalna, t.j. $\mathcal{N}(\mu, \sigma^2)$, bo po transformaciji postala standardno-normalna, t.j. $\mathcal{N}(0, 1)$.

Slaba plat vključitve z-norm normalizacije v sistem za samodejno razpoznavanje govorcev je povečana časovna zahtevnost izvedbe eksperimenta. Če imamo v množici z-norm N_z posnetkov, moramo namreč za vsakega učnega govorca izračunati še dodatnih N_z rezultatov prileganj. Na srečo se povečanost časovne zahtevnosti pozna le v fazi predstavitve (angl. enrollment) novega govorca, ki jo lahko izvedemo nesprotno (angl. off-line) in zato nima vpliva na čas, ki je potreben za izračun testnega rezultata prileganja.

3.7.2 Normalizacija t-norm

Podobna ideja se skriva za normalizacijo, ki ji pravimo t-norm. Razlika v primerjavi z z-norm je ta, da tukaj v ›središče‹ postavimo testni posnetek. Če bi za vsak tesni posnetek izračunali rezultate prileganja pri enakih učnih posnetkih (modelih), bi v splošnem dobili različne porazdelitve. Da se temu izognemo, lahko za vsak testni posnetek ocenimo parametra μ in σ na vnaprej pripravljeni množici t-norm modelov.

Rezultat prileganja s za konkreten testni posnetek nato pretvorimo na enak način kot pri z-norm:

$$s_t = \frac{s - \mu}{\sigma}.$$

V nasprotju z normalizacijo z-norm, normalizacije t-norm ne moremo izvesti nesprotno, kar pomeni, da njena uporaba znatno poveča čas, ki ga potrebujemo za izračun rezultata prileganja pri danem testnem posnetku. Pozorni moramo biti tudi pri izbiranju z-norm in t-norm posnetkov. Čeprav je v literaturi (Mclaren et al., 2009) moč najti izsledke, s katerimi lahko to izbiro opravimo bolj selektivno, se ponavadi zadovoljimo kar z naključno izbiro.

3.7.3 Normalizacija zt-norm in tz-norm

Izkaže se, da lahko oba načina normalizacije rezultatov prileganj izvedemo hkrati. Imamo dve možnosti:

i) najprej izvedemo z-norm, nato t-norm (zt-norm);
ii) najprej izvedemo t-norm, nato z-norm (tz-norm).

Zanimivo je, da je učinkovitost posamezne normalizacije rezultatov prileganj zelo odvisna od postopka, ki ga uporabimo za razpoznavanje govorcev. Splošnega recepta, kateri način normalizacije bo najbolj primeren za katero vrsto razpoznavalnika, ne poznamo. Zato je potrebno primerno vrsto normalizacije izbrati eksperimentalno. Izkaže se, da se pri nekaterih razpoznavalnikih rezultat prileganja z uporabo normalizacije znatno izboljša, spet pri drugih pa normalizacija nima nobenega vpliva.

3.8 Združevanje rezultatov prileganja

Ker vemo, da popoln razvrščevalnik ne obstaja in da različni razpoznavalniki dajejo različne rezultate, se ponuja možnost, da realiziramo več različnih razpoznavalnikov in pred razvrščanjem posamezne rezultate teh razvrščevalnikov med seboj na primeren način združimo. Združevanju rezultatov prileganja večih razpoznavalnikov pravimo v angleščini *score fusion* ali tudi *score combination*.

Poraja se vprašanje, kakšni naj bodo ti posamezni razpoznavalniki, da bo združevanje rezultatov prileganja uspešno. Izkaže se, da bo združevanje tem bolj uspešno, čim bolj bodo razpoznavalniki med seboj neodvisni. Povedano drugače, posamezni rezultati prileganja morajo vsebovati *komplementarno* informacijo.

V sistemih za razpoznavanje govorcev se za združevanje rezultatov prileganja najpogosteje uporablja kar linearna kombinacija (utežena vsota) posameznih rezultatov. Izkaže se, da združevanje večjega števila heterogenih razpoznavalnikov znatno zmanjša napako razpoznavanja (Brummer et al., 2007) — še posebej, če so v združevanje vključeni razpoznavalniki, ki skušajo informacijo o govorčevih značilnosti

zajeti na različnih nivojih — vendar si gradnjo večjega števila razpoznavalnikov lahko privoščijo le številčnejše raziskovalne skupine.

3.9 Komentar

Opisali smo vse pomembnejše korake, ki jih je potrebno izvesti, ko želimo udejanjiti sistem za samodejno razpoznavanje oseb na osnovi govora. Razložili smo pomen in izvedbo postopka parametrizacije govornega signala ter postopek za izločanje negovornih odsekov. Še posebej smo se osredotočili na postopek statističnega modeliranja, konkretno na model mešanice Gaussovih porazdelitev, ki igra na področju razpoznavanja govorcev pomembno vlogo. Opisali smo postopek učenja modela UBM, za učenje katerega potrebujemo obsežno govorno zbirko, in izpostavili časovno zahtevnost tega postopka. Pokazali smo, kako model UBM uporabimo za izpeljavo modelov posameznih govorcev, za katere imamo na voljo le omejeno količino govornih podatkov, kar storimo z adaptacijo MAP.

MODEL MEŠANICE GAUSSOVIH PORAZDELITEV

4

Model GMM — ki igra v naši disertaciji osrednjo vlogo, zato je prav, da mu posvetimo posebno poglavje — sestavimo kot linearno kombinacijo večih Gaussovih funkcij:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \sigma_k^2). \tag{4.1}$$

Čeprav je Gaussova funkcija[14] sama po sebi preprosta, se izkaže, da smo na ta način dobili novo funkcijo, s katero lahko aproksimiramo skoraj katerokoli (zvezno) funkcijo gostote verjetnosti do poljubne natančnosti. Dobljeni linearni kombinaciji pravimo model mešanice Gaussovih porazdelitev[15]. Posameznim Gaussovim funkcijam pravimo *komponente* mešanice, parametrom π_k pa *mešalni koeficienti*. Ti morajo ustrezati pogoju $\sum_k \pi_k = 1$. Če zahtevamo še, da so vsi koeficienti nenegativni, potem jih lahko obravnavamo kot verjetnosti.

4.1 Model GMM kot končni avtomat

Model GMM si lahko predstavljamo kot polno povezan avtomat, ki prehaja med K stanji (slika 4.1) in oddaja[16] observacije x_n. Verjetnosti prehodov med stanji lahko podamo s tranzicijsko matriko:

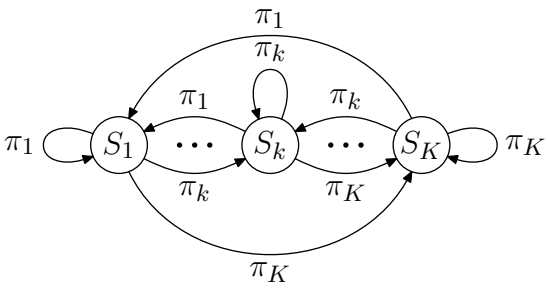
$$\boldsymbol{T} = \begin{pmatrix} \pi_1 & \dots & \pi_k & \dots & \pi_K \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \pi_1 & \dots & \pi_k & \cdots & \pi_K \end{pmatrix},$$

kjer (i, j)-ti element matrike $\boldsymbol{T}$ pomeni verjetnost prehoda iz stanja S_i v stanje S_j. Vidimo, da verjetnost prehodov ni odvisna od začetnega ampak le od končnega stanja (vrstice matrike $\boldsymbol{T}$ so enake). Pri prehodu v stanje S_k GMM odda vrednost x_n v skladu s porazdelitvijo $\mathcal{N}(x_n|\mu_k, \sigma_k^2)$.

[14] Gaussova funkcija spada v družino t.i. eksponencialnih porazdelitev, ki si delijo precej pomembnih lastnosti; glej poglavje 2.4 v (Bishop, 2007).

[15] V splošnem lahko modeli mešanic vsebujejo tudi druge funkcije. Ugodno je, če te funkcije spadajo v družino eksponencialnih funkcij. Tak primer je tudi npr. mešanica Bernoullijevih porazdelitev.

[16] V tem primeru je avtomat oddajnik. Možna je tudi interpretacija, kjer GMM obravnavamo kot sprejemnik.



Slika 4.1 Model GMM upodobljen končnega avtomata. Predstavljamo si lahko, da vsako stanje vsebuje Gaussovo porazdelitev, v skladu s katero avtomat oddaja observacije.

Obravnava verjetnostnih modelov kot končnih avtomatov pride zelo prav v primeru, ko smo že ocenili parametre modela in želimo le še udejanjiti kompleksen sistem kot je npr. razpoznavalnik govora. Razumljivo je, da je tak način obravnave verjetnostnih modelov zelo prisoten v klasični literaturi s področja razpoznavanja govora (Rabiner, 1989; Young et al., 2009), kjer je v središču pozornosti modelu GMM soroden prikriti model Markova.

4.2 Pojem *opažene* in *prikrite* spremenljivke

Naj bo $Z = (Z_1, \ldots, Z_K)^{\mathrm{T}}$ K-razsežna kategoriþMna naključna spremenljivka, ki lahko zavzame le take vrednosti $\boldsymbol{z} = (z_1, \ldots, z_K)^{\mathrm{T}}$, za katere velja, da je med njimi le ena vrednost z_k enaka 1, vse ostale pa so enake 0[17]. In naj bo njena porazdelitvena funkcija $p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$. Če interpretiramo realizacijo naključne spremenljivke Z kot naključni proces izbiranja komponent iz modela GMM, lahko pogojno porazdelitev $p(x|\boldsymbol{z})$ zapišemo kot $p(x|\boldsymbol{z}) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \sigma_k^2)^{z_k}$.

Iz osnovnih pravil verjetnosti vemo, da lahko vezano porazdelitev $p(x, \boldsymbol{z})$ zapišemo kot produkt $p(\boldsymbol{z})$ in $p(x|\boldsymbol{z})$ in da lahko marginalno porazdelitev $p(x)$ dobimo s seštevanje preko vseh možnih vrednosti spremenljivke Z:

$$\begin{aligned} p(x) &= \sum_{\boldsymbol{z} \in \Omega_Z} p(\boldsymbol{z}) p(x|\boldsymbol{z}) = \sum \prod_{k=1}^{K} \pi_k^{z_k} \mathcal{N}(x|\mu_k, \sigma_k^2)^{z_k} \\ &= \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \sigma_k^2) \end{aligned} \tag{4.2}$$

Tako smo pokazali, da marginalna porazdelitev naključne spremenljivke X ustreza modelu mešanice Gaussovih porazdelitev iz enačbe (**4.1**).

[17] Primer: trirazsežna kategoriþMna naključna spremenljivka lahko zavzame le tri različne vrednosti: $(1, 0, 0)^{\mathrm{T}}$, $(0, 1, 0)^{\mathrm{T}}$ in $(0, 0, 1)^{\mathrm{T}}$.

Vpeljava naključne spremenljivke Z nam omogoča, da lahko naključni proces generiranja vrednosti x v skladu s porazdelitvijo $p(x)$ ekvivalentno predstavimo z dvostopenjskim naključnim procesom, kjer v prvem koraku iz mešanice izberemo eno izmed komponent v skladu s porazdelitvijo $p(\boldsymbol{z})$, iz katere nato v drugem koraku generiramo vrednost x v skladu s pogojno porazdelitvijo $p(x|\boldsymbol{z})$. Vidimo, da je *opažena* (angl. observed) le spremenljivka x, medtem ko spremenljivka $\boldsymbol{z}$ ostaja *prikrita* (angl. hidden, tudi angl. latent).

Čeprav je naključna spremenljivka Z neopažena, pa lahko o njeni vrednosti vseeno sklepamo posredno preko vrednosti x. Ker imamo podano apriorno porazdelitev $p(\boldsymbol{z})$, lahko po Bayesu izpeljemo njeno posteriorno porazdelitev $p(\boldsymbol{z}|x)$:

$$\begin{aligned} p(\boldsymbol{z}|x) &= \frac{p(x|\boldsymbol{z})p(\boldsymbol{z})}{p(x)} \\ &= \frac{\prod_k \pi_k^{z_k} \mathcal{N}(x|\mu_k, \sigma_k^2)^{z_k}}{\sum_k \pi_k \mathcal{N}(x|\mu_k, \sigma_k^2)} = \frac{\pi_j \mathcal{N}(x|\mu_j, \sigma_j^2)}{\sum_k \pi_k \mathcal{N}(x|\mu_k, \sigma_k^2)}. \end{aligned}$$

Ta posteriorna porazdelitev nam govori, kako dobro j-ta komponenta mešanice pojasni observacijo x. Pogosto jo označimo z γ_j in ji pravimo *odgovornost* (angl. responsibility).

Na prvi pogled se sicer zdi, da z vpeljavo prikrite spremenljivke Z nismo dosti pridobili, a bomo kmalu videli, da uporaba vezane porazdelitve $p(x, \boldsymbol{z})$ namesto marginalne porazdelitve $p(x)$ privede do poenostavitve nadaljnjih izpeljav.

4.3 Model GMM kot grafični model

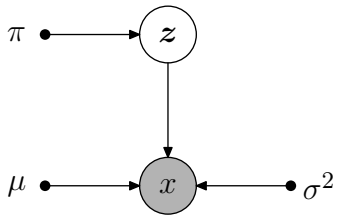
Na področju strojnega učenja in med statistiki je v zadnjem času zelo popularna drugačna predstavitev verjetnostnih modelov. Odvisnosti med naključnimi spremenljivkami, ki nastopajo v modelu, nazorno upodobimo v obliki diagrama, za katerega se je prijelo ime *verjetnosti grafični model*[18] (angl. probabilistic graphical model, PGM) ali krajše *grafični model* (angl. graphical model, GM) (Bishop (2007), poglavje 8).

Pridevnik grafični izraža dvoje: i) model lahko narišemo in ii) upodobitev ima obliko (matematičnega) grafa. Prednosti predstavitve verjetnostnih modelov v obliki grafičnega modela so naslednje:

i) vizualizacija strukture verjetnostnega modela motivira in omogoči načrtovanje novih modelov;
ii) z opazovanjem grafa dobimo vpogled v lastnosti verjetnostnega modela (vključno z lastnostjo pogojne neodvisnosti med spremenljivkami);

[18] Tukaj se omejimo na *usmerjene* (angl. directed) grafične modele, ki jim pravimo tudi bayesovske mreže (angl. Bayesian networks, BN). Obstajajo tudi neusmerjeni grafični modeli, ki jim pravimo *verjetnostna Markovova polja* (angl. Markov random fields). Obe vrsti GM lahko pretvorimo v *faktorske grafe* (angl. factor graphs), ki pridejo prav pri izpeljavi učnih postopkov.

iii) zapletene matematične izračune, ki jih je potrebno narediti v postopkih učenja kompleksnih modelov, lahko implicitno izrazimo v obliki grafičnih operacij nad grafom (podobnost s Feynmanovimi diagrami);
iv) vodi k učinkovitejši programski implementaciji.



Slika 4.2 Prikaz modela GMM v obliki grafičnega modela. Naključne spremenljivke so prikazane z večjimi vozlišči, parametri pa z manjšimi polnimi vozlišči. Vozlišče spremenljivke X je pobarvano s sivo barvo, kar pomeni, da je ta spremenljivka opažena.

Grafični model na učinkovit način izraža tudi, kako se vezana porazdelitev vseh naključnih spremenljivk, ki nastopajo v verjetnostnem modelu, razcepi na produkt posameznih faktorjev. S pogledom na sliko 4.2 lahko za model GMM hitro zapišemo:

$$p(x, \boldsymbol{z}) = p(x|\boldsymbol{z})p(\boldsymbol{z}).$$

4.4 Predpostavka o neodvisnosti in enaki porazdeljenosti

Kadar obravnavamo naključni proces, imamo namesto z eno naključno spremenljivko opravka z nizom naključnih spremenljivk, ki jih lahko obravnavamo z vezano porazdelitvijo. Označimo z $X = \{X_1, \ldots, X_N\}$ množico naključnih spremenljivk, katerih vezano porazdelitev zapišemo kot:

$$p_X(x) = p_{X_1,\ldots,X_N}(x_1, \ldots, x_N).$$

Vezano porazdelitev verjetnosti lahko z upoštevanjem pravila produkta verjetnosti vedno zapišemo kot produkt pogojnih porazdelitev verjetnosti:

$$p_X(x) = \prod_{n=1}^{N} p_{X_n|X_1,\ldots,X_{n-1}}(x_n|x_1, \ldots, x_{n-1})$$

Če predpostavimo, da so vse naključne spremenljivke X_n med seboj *neodvisne*, se vezana porazdelitev poenostavi v produkt posameznih porazdelitev:

$$p_X(x) = \prod_{n=1}^{N} p_{X_n}(x_n).$$

Če dodatno predpostavimo, da so spremenljivke X_n enako porazdeljene, lahko pri zapisu p_{X_n} izpustimo indeks X_n in zapišemo

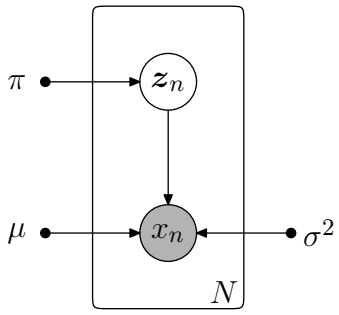
$$p_X(x) = \prod_{n=1}^{N} p(x_n),$$

Vidimo, da smo tako ob upoštevanju predpostavke o *neodvisnosti in enaki porazdeljenosti* vezano porazdelitev prevedli na produkt (enakih) porazdelitev.

Čeprav je v praksi predpostavka i.i.d. velikokrat neupravičena[19], pa na njen temelji veliko teoretičnih rezultatov, saj znatno olajša matematično obravnavo.

4.5 Največje verjetje

Vzemimo, da smo opazili niz $x = \{x_1, \ldots, x_N\}$, ki ga je porodil model GMM s parametri θ. V tem primeru imamo N podatkov $\{x_n\}_{n=1}^N$ in N prikritih spremenljivk $\{Z_n\}_{n=1}^N$. Razmere lahko ponovno predstavimo grafično (slika 4.3).



Slika 4.3 Prikaz modela GMM v obliki (dinamičnega) grafičnega modela. Opazimo lahko, da smo spremenljivki X_n in Z_n postavili na *pladenj* (angl. plate), označen s simbolom N. Tako smo na zgoščen način prikazali N i.i.d. podatkov x_n skupaj s pripadajočimi prikritimi spremenljivkami Z_n. Opazimo lahko, da so parametri π, μ in σ^2 narisani izven pladnja. To pomeni, da si vse spremenljivke $\{X_n, Z_n\}_{n=1}^N$ parametre delijo med seboj.

Grafični model s slike 4.3 izraža naslednjo faktorizacijo vezane porazdelitve vseh spremenljivk, ki nastopajo v modelu GMM:

$$p(\{x_n, \boldsymbol{z}_n\}_{n=1}^N) = \prod_{n=1}^{N} p(x_n|\boldsymbol{z}_n)p(\boldsymbol{z}_n).$$

Logaritem funkcije verjetja lahko, upoštevajoč rezultat (4.2), zapišemo kot:

[19] Predpostavka i.i.d. je strožja od predpostavke o zamenljivosti, na kateri temelji de Finettijev izrek, ki ima ključno vlogo v bayesovski statistiki (Fox (2009), poglavje 2.1.1).

$$\log p(x|\theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \sigma_k^2). \tag{4.3}$$

Zanimalo nas bo, pri katerih vrednostih parametrov θ modela GMM bo funkcija verjetja zavzela največjo vrednost. Preden preidemo na postopek iskanja maksimuma funkcije verjetja, opozorimo na težave, na katere lahko pri tem naletimo in ki se jih je dobro zavedati.

4.5.1 Singularnost

Maksimizacija verjetja je v osnovi slabo definiran problem, saj lahko vedno najdemo take vrednosti parametrov modela GMM, kjer bo imelo verjetje neskončno vrednost[20]. To se zgodi takrat, ko povprečje ene izmed komponent sovpade z vrednostjo ene izmed vrednosti točk x_n, varianca te komponent pa limitira proti vrednosti 0.

Problem singularnosti je še en obraz *prenaučenosti*, do katerega lahko pride pri ocenjevanju parametrov (učenju) po kriteriju največjega verjetja. Proti singularnostim se sicer lahko borimo do neke mere s hevristikami, kot je npr. omejevanje najmanjše vrednosti variance (angl. variance flooring), a zares opravimo s problemom singularnosti šele z uporabo bayesovskega pristopa.

4.5.2 Identifikabilnost

Pri iskanju parametrov, pri katerih bo funkcija verjetja dosegla največjo vrednost, trčimo na še eno težavo. Ta se skriva v dejstvu, da je mešanica določena le do vrstnega reda komponent natančno. Z drugimi besedami, katerakoli izmed $K!$ permutacij komponent bo vodila do iste porazdelitve. Temu dejstvu pravimo *identifikabilnost* (angl. identifiability) in ima pomembno vlogo, ko želimo interpretirati vrednosti parametrov modela mešanice. Kadar je naša naloga, da samo najdemo parametre modela, ki bodo dovolj dobro opisali opažene podatke, se nam s to potencialno težavo ni potrebno ukvarjati, saj je v tem primeru katerakoli izmed ekvivalentnih rešitev enako dobra.

4.6 Maksimizacija upanja

Čeprav vemo, da za oceno parametrov Gaussove funkcije po kriteriju ML obstaja enostavna analitična rešitev, se izkaže, da je iskanje največje vrednosti verjetja modela GMM mnogo kompleksnejši problem. Težavnost se skriva v dejstvu, da logaritem ne ›prijemlje‹ neposredno na Gaussovi funkciji, temveč na vsoti Gaussovih funkcij. Posledica tega je, da analitične rešitve ne znamo poiskati, ampak se moramo zateči k iterativnemu postopku. V ta namen bi lahko uporabili postopek, ki bi temeljil na računanju gradienta funkcije verjetja, a na srečo poznamo še elegantnejši in učinkovitejši postopek, ki mu pravimo *maksimizacija upanja*.

[20] To velja za mešanice z dvema ali več komponentami.

Pokazati se da, da če poiščemo maksimum funkcije verjetja tako, da (parcialno) odvajamo enačbo (4.3) po vseh parametrih μ_k, dobimo:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk}\, x_n, \tag{4.4}$$

kjer smo N_k definirali kot $N_k = \sum_{n=1}^{N} \gamma_{nk}$. Podobno, če odvajamo po parametrih σ_k^2, dobimo

$$\sigma_k^2 = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk}(x_n - \mu_k)^2. \tag{4.5}$$

Pri iskanju maksimuma verjetja v odvisnosti od mešalnih koeficientov π_k, je potrebno upoštevati še pogoj, da znaša njihova vsota 1, kar zaobjamemo z Lagrangeovim multiplikatorjem $\lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$, ki ga prištejemo desni strani enačbe (4.3). Po odvajanju in enačenju z 0 dobimo

$$\pi_k = \frac{N_k}{N}. \tag{4.6}$$

Na ta način smo dobili enačbe za izračun vrednosti parametrov, pri katerih bo imela funkcija verjetja največjo vrednost. Vidimo, da v teh enačbah nastopajo odgovornosti $\gamma(z_{nk})$, ki pa so ponovno odvisne od parametrov π_k, μ_k in σ_k^2. Ne glede na to, da dobljene rešitve torej niso analitične, pa nam ponujajo preprosto iterativno shemo, s katero lahko najdemo rešitev našega problema. To shemo lahko strnemo v tri korake:

i) *inicializacija*
   (Naključno) izberemo začetne vrednosti parametrov θ.
ii) *e-korak*
   Trenutne vrednosti parametrov uporabimo za izračun odgovornosti γ_{nk}.
iii) *m-korak*
   Dobljene vrednosti odgovornosti in trenutne vrednosti parametrov uporabimo za izračun novih vrednosti parametrov v skladu z enačbami (4.4), (4.5) in (4.6). Izračunamo razliko vrednosti logaritma verjetja pri prejšnjih vrednostih parametrov in novih vrednostih parametrov. Če je razlika manjša od izbrane vrednosti, postopek končamo, sicer ponovno preidemo na e-korak.

Izkaže se, da je ta iteracijska shema poseben primer postopka maksimizacije upanja, ki ga v vsej splošnosti predstavimo v dodatku. Omenimo še, da postopek EM zagotavlja monotono naraščanje vrednosti verjetja proti lokalnemu maksimumu. Maksimum, ki ga pri tem dosežemo, je odvisen od izbranih začetnih vrednosti parametrov.

4.7 Postopek EM za model s prikritimi spremenljivkami

Nepoznavalci postopek EM velikokrat dojemajo kot sinonim za postopek ocenjevanja parametrov modela GMM. V resnici je EM optimizacijski postopek, ki je primeren za poljuben verjetnostni model, ki vsebuje prikrite spremenljivke.

Označimo z oznako X množico vseh opaženih spremenljivk, z oznako Z množico vseh prikritih spremenljivk in z oznako θ množico vrednosti vseh parametrov modela. Vrednost logaritma funkcije verjetja lahko zapišemo kot:

$$\log p(x|\theta) = \log \sum_z p(x, z|\theta). \tag{4.7}$$

Tukaj se nismo omejili le na diskretne prikrite spremenljivke Z — enako velja tudi za zvezne prikrite spremenljivke, le vsoto v enačbi (**4.7**) moramo zamenjati z integralom.

Ugotovimo lahko, da je v enačbi (**4.7**) vsota preko vseh možnih vrednosti spremenljivke Z izvedena znotraj logaritma, zaradi česar postane iskanje največje vrednosti verjetja zapleteno. Če bi bil vrstni red logaritma in vsote zamenjan, bi logaritem prijemal neposredno na vezani porazdelitvi $p(x, z|\theta)$, kar bi bilo ugodno, v kolikor bi vezana porazdelitev spadala v družino eksponencialnih funkcij.

Predpostavimo, da je maksimizacijo izraza $\log p(x, z|\theta)$ izvesti mnogo enostavneje kot maksimizacijo izraza $\log p(x|\theta)$. Potem bi bilo koristno, če bi poleg vrednosti spremenljivk X poznali tudi vrednosti spremenljivk Z. V tem primeru bi govorili o polnem naboru podatkov (angl. complete data set), za razliko od nepolnega nabora podatkov (angl. incomplete data set) v primeru, ko je množica spremenljivk Z prikrita. Čeprav vrednosti množice spremenljivk Z v resnici ne poznamo, pa lahko o njihovih vrednostih sklepamo preko posteriorne porazdelitve $p(z|x, \theta)$. Zato lahko vrednost logaritma verjetja pri polnem naboru podatkov nadomestimo s pričakovano vrednostjo logaritma verjetja pod posteriorno porazdelitvijo, kar ustreza e-koraku v postopku EM. Za pričakovano vrednost logaritma verjetja pri polnem naboru podatkov (angl. expected value of the complete-data log-likelihood), ki nastopa v številnih izpeljavah postopka EM, se je prijelo ime pomožna (angl. auxiliary) funkcija in jo pogosto označimo s črko $\mathcal{Q}$:

$$\mathcal{Q}(\theta, \theta_t) = \sum_z p(z|x, \theta_t) \log p(x, z|\theta), \tag{4.8}$$

kjer smo z oznako θ_t označili množico trenutnih vrednosti parametrov. Novo oceno parametrov θ_{t+1} dobimo tako, da poiščemo maksimum pomožne funkcije, kar ustreza m-koraku v postopku EM:

$$\theta_{t+1} = \arg\max_\theta \mathcal{Q}(\theta, \theta_t)$$

Vidimo, da logaritem v enačbi (**4.8**) deluje neposredno na vezani porazdelitvi $p(x, z|\theta)$, zato bo m-korak — po predpostavki iz začetka odstavka — izvedljiv.

Povzemimo postopek EM v nekaj korakih. Naj bo $p(x, z|\theta)$ vezana porazdelitev množice opaženih naključnih spremenljivk x in množice prikritih naključnih spremenljivk z, parametrizirana glede na množico parametrov θ. Naša naloga je poiskati take vrednosti parametrov θ, pri katerih bo vrednost verjetja $p(x|\theta)$ največja:

i) Izberi začetne vrednosti parametrov θ_t.
ii) e-korak: Oceni $p(z|x, \theta_t)$.
iii) m-korak: Določi nove vrednosti parametrov θ_{t+1}, tako da rešiš enačbo $\theta_{t+1} = \arg\max \mathcal{Q}(\theta, \theta_t)$, kjer je $\mathcal{Q}(\theta, \theta_t) = \sum_z p(z|x, \theta_t) \log p(x, z|\theta)$.
iv) Preveri konvergenco bodisi vrednosti logaritma verjetja bodisi vrednosti parametrov. Če pogoj za konvergenco ni izpolnjen, postavi $\theta_t := \theta_{t+1}$ in se vrni na drugi korak.

4.7.1 Maksimizacija modusa posteriorne porazdelitve

Postopek EM lahko uporabimo tudi za iskanje rešitev po kriteriju *največje vrednosti modusa posteriorne porazdelitve* (angl. maximum a posteriori, MAP), kadar razpolagamo z apriorno porazdelitvijo parametrov $p(\theta)$. V tem primeru ostane e-korak enak kot v primeru iskanja ML rešitve, medtem ko se v m-koraku pomožna funkcija (**4.8**) spremeni tako, da ji prištejemo vrednost $\log p(\theta)$. Ena izmed odlik kriterija MAP je ta, da se z njegovo pomočjo izognemo potencialnim singularnostim, na katere bi trčili pri kriteriju ML, če le izberemo smiselno a priorno porazdelitev parametrov. Z drugimi besedami, uporaba kriterija MAP nas zavaruje pred prenaučitvijo, kar s pridom izkoriščamo v postopkih učenja, kadar je za oceno parametrov na voljo le majhna količina podatkov.

4.8 Postopek EM za model GMM

Aplicirajmo splošni postopek EM, ki smo ga predstavili v prejšnjem razdelku, za konkretni primer modela GMM. Naša naloga je poiskati maksimum logaritma funkcije verjetja pri nepopolnem naboru podatkov (**4.3**). Videli smo, da ta naloga ni enostavna, saj se vsota preko vseh komponent mešanice nahaja znotraj operacije logaritmiranja. Če bi poleg vrednosti opaženih naključnih spremenljivk X izvedeli tudi vrednosti skritih naključnih spremenljivk Z — z drugimi besedami, za vsak podatek x_n bi izvedeli, katera komponenta ga je porodila — bi lahko poiskali maksimum funkcije verjetja pri popolnem naboru podatkov:

$$p(x, z|\theta) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{kn}} \mathcal{N}(x_n|\mu_k, \sigma_k^2)^{z_{kn}},$$

kjer smo z z_{kn} označili k-to komponento spremenljivke z_n. Z logaritmiranjem gornjo enačbo prevedemo v:

$$\log p(x, z|\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{kn} \left[\log \pi_k + \log \mathcal{N}(x_n|\mu_k, \sigma_k^2)\right].$$

Vidimo, da sedaj logaritem deluje direktno na Gaussovi funkciji. Ne preseneča, da to vodi do enostavnejše rešitve. Če dobro pogledamo, ugotovimo, da je iskanje maksimuma gornje funkcije glede na povprečje in kovarianco določene komponente enako iskanju maksimuma verjetja pri Gaussovi funkciji, le da moramo upoštevati le tiste podatke, ki jih je porodila dotična komponenta. Pri iskanju maksimuma glede na posamezne mešalne koeficiente je potrebno uvesti Lagrangeove multiplikatorje in tako upoštevati njihovo medsebojno sklopljenost, ki je posledica dejstva, da mora vsota vseh mešalnih koeficientov znašati 1.

Pokazali smo, da je problem iskanja maksimuma funkcije verjetja pri popolnem naboru podatkov analitično rešljiv. Žal v praksi informacije o konkretnih vrednostih množice prikritih spremenljivk Z nimamo na voljo, zato pa lahko o njihovih vrednostih sklepamo preko posteriorne porazdelitve. To nam omogoča, da obravnavamo pričakovano vrednost funkcije logaritma verjetja pri popolnem naboru podatkov glede na to posteriorno porazdelitev.

Za posteriorno porazdelitev $p(z|x,\theta)$ velja:

$$p(z|x,\theta) \propto \prod_{n=1}^{N}\prod_{k=1}^{K}\left[\pi_k \mathcal{N}(x_n|\mu_k,\sigma_k^2)\right]^{z_{kn}},$$

iz česar lahko razberemo, da so posamezne naključne spremenljivke Z_n med seboj neodvisne, saj smo vezano posteriorno porazdelitev zapisali kot produkt posameznih porazdelitev.

Ker se da pokazati, da je pričakovana vrednost spremenljivke Z_{kn} pod to posteriorno porazdelitvijo enaka odgovornosti $\gamma(z_{kn})$, lahko pričakovano vrednost funkcije logaritma verjetja pri popolnem naboru podatkov zapišemo kot:

$$\mathbb{E}_Z[\log p(X,Z|\theta)|X=x] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{kn})\left[\log\pi_k + \log\mathcal{N}(x_n|\mu_k,\sigma_k^2)\right].$$

Ta enačba nam zopet ponuja že znano iteracijsko shemo postopka EM.

MERJENJE PODOBNOSTI 5

V tem poglavju bomo pobliže spoznali mere podobnosti, ki so še posebej primerne za računanje izidov prileganja, kadar za razvrščanje uporabljamo model GMM. Razdelimo jih lahko v tri skupine:

- mere podobnosti na osnovi verjetja;
- mere podobnosti na osnovi razdalje med porazdelitvami;
- mere podobnosti na osnovi metode podpornih vektorjev.

5.1 Verjetje

Najbolj naravna mera podobnosti, s katero lahko ugotavljamo kakovost prileganja niza značilk na model GMM, je verjetje. Če imamo niz $\boldsymbol{x} = \{\boldsymbol{x}_n\}_{n=1}^N$ in model GMM s parametri $\theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$, vrednost verjetja parametrov θ pri podatkih $\boldsymbol{x}$ izračunamo na sledeč način:

$$\mathcal{L}(\theta|\boldsymbol{x}) = p(\boldsymbol{x}|\theta) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{5.1}$$

V praksi zaradi numeričnih razlogov namesto verjetja izračunavamo logaritem verjetja, ki produkt prevede na vsoto:

$$\log \mathcal{L}(\theta|\boldsymbol{x}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

5.1.1 Približni izračun verjetja

Izračun verjetja je računsko zahtevna operacija, še posebej, v kolikor imamo v modelu GMM večje število mešanic. Izračun lahko pohitrimo, če namesto prave vrednosti verjetja izračunamo njegov približek. V sistemih za razpoznavanje govorcev se je tako uveljavila metoda, ki pri izračunu verjetja za vsak vektor $\boldsymbol{x}_n$ upošteva le nekaj najbolj verjetnih komponent modela GMM.

Izračun verjetja z upoštevanjem najbolj verjetnih komponent

Za točen rezultat verjetja bi morali za vsak $\boldsymbol{x}_n$ sešteti prispevke vseh komponent v modelu GMM. Na srečo se izkaže, da velik delež k skupni vsoti prispeva le nekaj tistih komponent modela, ki ležijo najbližje vektorju $\boldsymbol{x}_n$, prispevke ostalih komponent pa

lahko mirno zanemarimo. Če bi torej za vsak $\boldsymbol{x}_n$ vnaprej poznali njemu najbližje komponente, bi lahko potreben čas za izračun verjetja znatno skrajšali.

To je ideja, ki stoji za metodo računanja verjetja na osnovi najbolj verjetnih komponent. Predpostavimo, da je vrstni red komponent, ki jih uredimo glede na vrednosti aposteriornih verjetnosti, neodvisen od govorca. Potem najbolj verjetnih komponent ni potrebno poiskati za vsakega govorca posebej, ampak je dovolj, če to storimo za vsak $\boldsymbol{x}_n$ le enkrat. V ta namen uporabimo model UBM.

Na prvi pogled se mogoče zdi, da z opisanim aproksimativnim izračunom verjetja postopek pohitrimo le, ko razpoznavalnik deluje v načinu identifikacije. Takrat namreč vsak testni posnetek primerjamo z večimi modeli govorcev. Izkaže pa se, da je pohitritev še posebej izrazita tudi v načinu verifikacije, kadar rezultate prileganja normiramo s postopkom t-norm (pri tej normalizaciji se tipično uporablja med 300 in 1000 t-norm govorcev).

Računanje verjetja na podlagi zadostne statistike

Sami v disertaciji predlagamo drugačen način približnega računanja verjetja, ki je računsko še precej učinkovitejši kot računanje verjetja z upoštevanjem najbolj verjetnih komponent.

Vpeljimo spremenljivko γ_{nk}:

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}, \tag{5.2}$$

ki podaja (aposteriorno) verjetnost, da je observacijo $\boldsymbol{x}_n$ generirala komponenta k in ji pogosto pravimo *odgovornost* (angl. responsibility). Pokazati se da, da lahko logaritem verjetja $\log p(\boldsymbol{x}|\theta)$ zapišemo kot:

$$\log p(\boldsymbol{x}|\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left(\log \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) - \log \gamma_{nk} \right), \tag{5.3}$$

s čimer smo zamenjali operaciji logaritma in seštevanja[21]. Posledično se gornji izraz poenostavi v:

$$\begin{aligned} \log p(\boldsymbol{x}|\theta) &= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left(\log \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) - \log \gamma_{nk} \right) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left(\log \pi_k - \frac{D}{2} \log(2\pi) - \log \gamma_{nk} - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| \cdots \right. \\ &\qquad \left. \cdots - \frac{1}{2} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) \right). \end{aligned}$$

[21] Pozoren bralec bo opazil, da je enačba (**5.3**) tesno povezana s postopkom EM, ki ga v vsej splošnosti predstavimo v dodatku **A**.

Po zamenjavi vrstnega reda seštevanja in preureditvi dobimo:

$$\begin{aligned}\log p(\boldsymbol{x}|\theta) = \sum_{k=1}^{K} & -\frac{1}{2}\mathrm{Tr}(\boldsymbol{S}_k\boldsymbol{\Sigma}_k^{-1}) + \boldsymbol{F}_k^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}N_k\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k \cdots \\ & \cdots + N_k\left(\log\pi_k - \frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_k|\right) \\ & - \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}\log\gamma_{nk}, \end{aligned} \tag{5.4}$$

kjer smo vpeljali spremenljivke N_k, $\boldsymbol{F}_k$, $\boldsymbol{S}_k$:

$$N_k = \sum_{n=1}^{N}\gamma_{nk} \qquad \boldsymbol{F}_k = \sum_{n=1}^{N}\gamma_{nk}\boldsymbol{x}_n \qquad \boldsymbol{S}_k = \sum_{n=1}^{N}\gamma_{nk}\boldsymbol{x}_n\boldsymbol{x}_n^{\mathrm{T}},$$

ki jih s skupnim imenom imenujemo *zadostna statistika* (angl. sufficient statistics). Vidimo, da smo logaritem verjetja zapisali v odvisnosti od zadostne statistike in ne več neposredno od podatkov $\boldsymbol{x}$. Pokazati se da, da je vrednost tako ocenjenega verjetja manjša ali kvečjemu enaka pravi vrednosti verjetja, t.j. spodnja meja verjetja. (Enakost velja le pri tistih vrednostih parametrov, ki smo jih uporabili pri izračunu zadostne statistike.)

Če predpostavimo, da je zadostna statistika (približno) enaka pri vseh modelih govorcev, zadostuje, da jo izračunamo le enkrat (za to uporabimo UBM). Na ta način znatno skrajšamo čas, ki ga potrebujemo za izračun vrednosti verjetij. Pohitritev se še posebej očitno pozna v primeru, kadar testni posnetek $\boldsymbol{x}$ prilegamo na večje število različnih modelov.

Ker nas pri razpoznavanju ne zanima absolutna vrednost verjetja ampak le razmerje verjetij, je nujno izračunati le tiste člene v enačbi (5.4), ki so različni pri različnih modelih. V našem primeru moramo izračunati le oba člena, v katerih nastopajo povprečni vektorji $\boldsymbol{\mu}_k$, saj so uteži π_k in kovariančne matrike $\boldsymbol{\Sigma}_k$ enake pri vseh modelih (jih ne adaptiramo).

Če iz skalarjev N_k, vektorjev $\boldsymbol{F}_k$ in matrik $\boldsymbol{S}_k$ sestavimo diagonalno matriko $\boldsymbol{N}$, vektor $\boldsymbol{F}$ ter bločno diagonalno matriko $\boldsymbol{S}$:

$$\boldsymbol{N} = \begin{pmatrix} N_1\boldsymbol{I} & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & N_2\boldsymbol{I} & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & N_K\boldsymbol{I} \end{pmatrix} \quad \boldsymbol{F} = \begin{pmatrix} \boldsymbol{F}_1 \\ \boldsymbol{F}_2 \\ \vdots \\ \boldsymbol{F}_K \end{pmatrix} \quad \boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_1 & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_2 & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{S}_K \end{pmatrix}$$

ter iz vektorjev $\boldsymbol{\mu}_k$ in matrik $\boldsymbol{\Sigma}_k$ vektorja $\boldsymbol{\mu}$ in bločno matriko $\boldsymbol{\Sigma}$:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_K \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Sigma}_K \end{pmatrix}$$

lahko enačbo (5.4) zapišemo v matrični obliki:

$$\log p(\boldsymbol{x}|\theta) = \boldsymbol{F}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{N}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathrm{const},$$

kjer smo vse člene, ki niso odvisni od povprečnih vektorjev $\boldsymbol{\mu}_k$ (in so zato enaki pri vseh modelih), vsrkali v konstanto const.

5.2 Mere različnosti med porazdelitvami

Težava obeh opisanih metod, ki temeljita na računanju aproksimativne vrednosti verjetja, je ta, da temeljita na predpostavki, da je poravnava med vektorji značilk testnega posnetka in komponentami modela GMM pri modelu UBM in govorčevemu modelu enaka. Izkazalo se je, da v praksi ta predpostavka ne vzdrži, saj se je z uporabo teh dveh metod rezultat razpoznavanja občutno poslabšal. Poslabšanje je bilo še izrazitejše pri metodi z uporabo zadostne statistike. Menimo, da bi razlog za to poslabšanje lahko tičal v precejšnjem neujemanju akustičnih razmer med učnimi in testnimi posnetki. Še posebej, ker se pri naših eksperimentih nismo odločili izvesti normalizacije na nivoju značilk, ki bi do neke mere razliko v akustičnih razmerah omilila.

Odločili smo se, da poskušamo poiskati metodo, ki ne bi imela omenjenih slabosti in bi hkrati bila računsko bolj sprejemljiva kot metoda klasičnega (natančnega) izračunavanja verjetij.

Ena izmed metod za računanje rezultata prileganja, ki bi lahko ustrezala tem zahtevam, je računanje različnosti[22] med porazdelitvami. Osnovna ideja je ta, da bi primerjanje učnega modela in testnega posnetka izvedli v dveh korakih. Najprej bi iz testnega posnetka izpeljavi testni model na enak način, kot smo to storili za učni posnetek. V drugem koraku pa bi ta dva modela med seboj primerjali z eno izmed mer različnosti med porazdelitvami.

Prva takšna mera, ki se nam ponuja kar sama od sebe, je *divergenca Kullbacka in Leiblerja* (angl. Kullback-Leibler divergence), krajše divergenca KL. Včasih ji pravimo tudi relativna entropija in ima pomembno vlogo v teoriji informacij. Definirana je kot:

$$\mathbb{D}_{\mathrm{KL}}[p \mid\mid q] = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x, \tag{5.5}$$

[22] Spomni se, da lahko mero različnosti vedno preslikamo v mero podobnosti s spremembo predznaka.

kjer sta $p(x)$ in $q(x)$ verjetnostni porazdelitvi dveh zveznih naključnih spremenljivk. (Podobna formula velja tudi za diskretne naključne spremenljivke, le operacijo integriranja moramo zamenjati z operacijo seštevanja.) Z uporabo Jensenove neenakosti lahko pokažemo, da je divergenca KL nenegativna, pri čemer znaša 0 le v primeru, ko sta porazdelitvi p in q enaki. V splošnem je divergenca KL nesimetrična: $\mathbb{D}_{\mathrm{KL}}[p \mid\mid q] \neq \mathbb{D}_{\mathrm{KL}}[q \mid\mid p]$, zato se jo pogosto simetrizira tako, da izračunamo povprečje divergenc $\mathbb{D}_{\mathrm{KL}}[p \mid\mid q]$ in $\mathbb{D}_{\mathrm{KL}}[q \mid\mid p]$.

Divergenca KL je povezana (Bishop, 2007) z *vzajemno informacijo* (angl. mutual information, MI), ki jo poznamo iz teorije informacij. Vzajemna informacija, ki jo označimo z $\mathbb{I}[\cdot, \cdot]$, je med zveznima naključnima spremenljivkama X in Y definirana kot:

$$\begin{aligned}\mathbb{I}[X, Y] = & -\iint p(x, y) \log\left(\frac{p(x)p(y)}{p(x, y)}\right) \mathrm{d}x\, \mathrm{d}y \\ & = \mathbb{D}_{\mathrm{KL}}[p(x, y) \mid\mid p(x)p(y)].\end{aligned}$$

Vidimo, da lahko MI interpretiramo kot (nenegativno) mero odvisnosti med dvema naključnima spremenljivkama. Bolj bosta spremenljivki med seboj odvisni, višja bo njuna medsebojna informacija; ta bo znašala 0 le v primeru, ko bosta spremenljivki med seboj neodvisni.

Kratka izpeljava pokaže, da je MI povezana s pogojno entropijo preko izraza:

$$\mathbb{I}[X, Y] = \mathbb{H}[X] - \mathbb{H}[X|Y] = \mathbb{H}[Y] - \mathbb{H}[Y|X],$$

kar omogoči, da lahko vidimo MI kot mero zmanjšanja nedoločenosti spremenljivke X, ko izvemo vrednost spremenljivke Y (ali obratno). Če si nadenemo bayesovska ›očala‹, vidimo $p(x)$ kot apriorno porazdelitev, $p(x|y)$ pa kot posteriorno porazdelitev. Medsebojna informacija torej predstavlja zmanjšanje nedoločenosti spremenljivke X kot posledico observacij $Y = y$.

V primeru dveh (d-razsežnih) normalnih porazdelitev $p = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ in $q = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, ko je integral v enačbi (**5.5**) analitično rešljiv, lahko divergenco KL izračunamo po sledeči formuli:

$$\mathbb{D}_{\mathrm{KL}}[p \mid\mid q] = \frac{1}{2}\left(\log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} + \mathrm{Tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0\right) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - d\right).$$

Vidimo, da bo divergenca KL med dvema Gaussovima porazdelitvama simetrična le v primeru, ko bosta porazdelitvi imeli enaki kovariančni matriki. Žal v primeru modelov mešanic takšne lepe analitične rešitve za divergenco KL ne znamo poiskati. Edina metoda, ki lahko v tem primeru oceni divergenco s poljubno natančnostjo, je simulacija Monte Carlo. A se izkaže, da je za mnogorazsežne porazdelitvene funkcije računsko nedopustno zahtevna, zato se (moramo) ponovno zateči k aproksimativni rešitvi.

Že (Do, 2003) je pokazal, da če imamo dva modela GMM (p in q) z enakim številom komponent, potem lahko zgornjo mejo divergence KL ocenimo z izrazom

$$\mathbb{D}_{\mathrm{KL}}[p \mid\mid q] \leq \mathbb{D}_{\mathrm{KL}}[\pi \mid\mid \omega] + \sum_k \pi_k \mathbb{D}_{\mathrm{KL}}[p_k \mid\mid q_k]$$
$$= \sum_k \pi_k \left(\log \frac{\pi_k}{\omega_k} + \mathbb{D}_{\mathrm{KL}}[p_k \mid\mid q_k] \right),$$

kjer sta p_k in q_k k-ti komponenti modelov p in q, v tem vrstnem redu. Enačba se še nekoliko poenostavi, če imata oba modela enaki porazdelitvi apriornih verjetnosti: $\pi = \omega$. Takrat zgornjo mejo divergence KL zapišemo kot:

$$\mathbb{D}_{\mathrm{KL}}[p \mid\mid q] \leq \sum_k \pi_k \mathbb{D}_{\mathrm{KL}}[p_k \mid\mid q_k], \tag{5.6}$$

kar lahko interpretiramo kot uteženo vsoto divergenc KL paroma poravnanih komponent.

Obstaja še več možnih načinov aproksimacije divergence KL med dvema modeloma GMM. Primerjava različnih aproksimacij je podana v (Hershey in Olsen, 2007).

Čeprav smo se v disertaciji omejili na mero KL, bi prav lahko izbrali tudi kakšno drugo izmed številnih mer različnosti med porazdelitvami, npr. Hellingerjevo razdaljo ali razdaljo Bhattacharyya.

5.3 Metoda podpornih vektorjev

V zadnjem času se je na področju razpoznavanja govorcev prijel odločitveni kriterij, ki temelji na *metodi podpornih vektorjev* (angl. support vector machine, SVM). Postopek SVM spada med diskriminatorne razvrščevalnike in ima številne privlačne lastnosti, tako teoretične kot tudi praktične (Burges, 1998). Eden izmed glavnih razlogov za popularnost razvrščevalnika SVM je v dobri sposobnosti posploševanja iz videnih na nevidene podatke.

Razvrščevalnik SVM je binarni razvrščevalnik, ki poišče tako *ločilno mejo* med dvema razredoma, ki ima največji rob (angl. maximal margin). Izkaže se, da lahko postopek določitve ločilne meje z največjim robom prevedemo na optimizacijski problem kvadratičnega programiranja. Matematično lahko ločilno funkcijo izrazimo kot:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i \, t_i \, K(\boldsymbol{x}, \boldsymbol{x}_i) + d, \tag{5.7}$$

kjer vektorje $\{\boldsymbol{x}_i\}_{i=1}^N$, ki določajo ločilno hiperravnino, imenujemo *podporni vektorji*. Z oznako $t_i \in \{+1, -1\}$ je določena razredna pripadnost i-tega podpornega vektorja, α_i pa so nenegativne uteži, ki morajo zadoščati pogoju: $\sum_i \alpha_i \, t_i = 0$. Jedro $K(\boldsymbol{x}, \boldsymbol{y})$ ustreza skalarnemu produktu $\Phi(\boldsymbol{x})^{\mathrm{T}} \Phi(\boldsymbol{x})$, kjer je $\Phi(\boldsymbol{x})$ funkcija, ki preslika vektor $\boldsymbol{x}$ iz vhodnega prostora (angl. input space) v visokorazsežen (lahko tudi neskončnorazsežen) prostor značilk (angl. feature space.). Ideja je ta, da problem, ki v vhodnem prostoru ni linearno ločljiv, preslikamo v višjedimenzionalen prostor,

kjer problem običajno postane linearno ločljiv[23]. Ker pri reševanju problema (iskanju ločilne meje) podatki (vektorji $\boldsymbol{x}$) vstopajo le v obliki skalarnega produkta, se lahko dejanski preslikavi v visokorazsežen prostor značilk izognemo[24] tako, da v vhodnem prostoru izračunamo le vrednosti jedrne funkcije $K(\boldsymbol{x}, \boldsymbol{y})$ za vse pare vektorjev $(\boldsymbol{x}, \boldsymbol{y})$. Najdena ločilna meja, ki je v prostoru značilk linearna, je v splošnem v vhodnem prostoru nelinearna.

Če želimo metodo SVM uporabiti kot mero podobnosti v primeru razpoznavanja govorcev, moramo vse govorne signale oz. nize vektorjev značilk, ki imajo lahko poljubno in različno dolžino, preslikati v vektorje enakih razsežnosti. Izkaže se, da lahko za to preslikavo izkoristimo postopek adaptacije MAP[25], s katerim ocenimo vrednosti povprečnih vektorjev posameznih Gaussovih komponent. Če te povprečne vektorje zložimo enega vrh drugega, dobimo primerno predstavitev, ki jo lahko neposredno uporabimo v postopku SVM. Poleg opisane možnosti je bilo predlagano še več načinov, kako preslikati različno dolge posnetke (nize vektorjev značilk) v vektorje s točno določenim številom komponent. Med njimi je vredno izpostaviti postopka MLLR (Stolcke et al., 2009; Leggetter in Woodland, 1995) in GLDS (Campbell et al., 2006a).

Znano je, da metoda SVM ni neodvisna od linearne transformacije prostora značilk, zato je vektorje, ki vstopajo v postopek SVM, potrebno predhodno normirati. S tem želimo preprečiti, da komponente vektorja, ki imajo večji razpon, ne bi pri računanju skalarnega produkta ›zasenčile‹ komponent z manjšim dinamičnim razponom. Avtorji programskega orodja LibSVM v svojem priročniku (Chang in Lin, 2001) za začetek predlagajo normiranje razpona vsake komponente vektorja na interval $[-1, 1]$ ali $[0, 1]$. Na področju razpoznavanja govorcev se je prijela neparametrična metoda normalizacije ranga, kjer vrednost vsake značilke zamenjamo s pripadajočo relativno lokacijo (rangom) v podatkih ›ozadja‹ (Stolcke et al., 2008).

5.4 Komentar

V razdelku smo predstavili različne mere podobnosti, ki se uporabljajo v sistemih za samodejno razpoznavanje govorcev in smo jih preizkusili tudi sami. Mere podobnosti smo razdelili v tri skupine: mere, ki temeljijo na računanju verjetja; mere, ki temeljijo na računanju razdalje med porazdelitvami in mere, ki temeljijo na metodi SVM. Ugotovili smo, da je računanje verjetja računsko zahteven postopek, zato smo predlagali veliko hitrejšo aproksimativno metodo, pri kateri verjetje izračunamo na podlagi zadostne statistike. Kot alternativo verjetju smo predlagali tudi mero podobnosti, ki temelji na merjenju razdalje med porazdelitvami.

---

[23] Postopek SVM je posplošen tudi na primere, kjer meja, ki bi brez napake ločila vzorce dveh razredov, ne obstaja. Govorimo o t.i. razvrščevalniku z mehkim robom (angl. soft margin).

[24] Tej implicitni preslikavi problema v visokorazsežen prostor značilk pravimo *jedrni trik* (angl. kernel trick).

[25] Adaptacija MAP povprečnih vrednosti komponent modela GMM ohranja urejenost med komponentami.

PROJEKCIJA MOTEČIH LASTNOSTI 6

Pri samodejnem razpoznavanju govorcev ločimo dve vrsti spremenljivosti. Prva je medgovorska spremenljivost (angl. intra-speaker variability), ki je za razpoznavanje govorcev ključna, saj brez nje ne bi mogli ločevati med različnimi govorci. Druga vrsta spremenljivost je medsejna spremenljivost (angl. inter-session variability), ki je odgovorna za to, da isti govorec zveni drugače v različnih posnetkih. Ta vrsta spremenljivosti je za nas moteča, saj ne prispeva k boljšemu ločevanju med govorci.

Medsejno ali krajše sejno spremenljivost po navadi pripišemo različnim akustičnim razmeram in različnim lastnostim mikrofona, s katerimi smo govorni signal zajeli. Poleg teh vplivov (ki jim krajše rečemo kar kanalski vplivi), pa obstaja tudi vrsta bolj subtilnih dejavnikov kot sta psihofizično stanje govorca ter vpliv staranja govorca (posnetki so namreč lahko pridobljeni v različno oddaljenih časovnih trenutkih). K sejni spremenljivost lahko štejemo tudi vpliv izgovorjenega besedila in celo trajanje samega posnetka.

Izkaže se, da ima sejna spremenljivost velik vpliv na robustnost postopkov za samodejno razpoznavanje govorcev, zato predstavlja enega izmed najbolj zanimivih izzivov na tem področju. V literaturi je bila predlagana vrsta metod, s katerimi se skuša omiliti kvarni vpliv sejne spremenljivost na rezultate razpoznavanja. Razdelimo jih lahko v tri skupine:

- postopki normalizacije na nivoju značilk;
- postopki normalizacije na nivoju modela;
- postopki normalizacije na nivoju rezultatov prileganja.

S postopki normalizacije na nivoju značilk in na nivoju rezultatov prileganja, ki smo jih predstavili že v razdelkih **3.7** in **3.2.2**, se v disertaciji nismo posebej ukvarjali. Bolj smo se osredotočili na postopke, ki skušajo kompenzirati vpliv sejne variabilnosti na modelskem nivoju. V zadnjem času sta se še posebej uveljavila dva postopka; projekcija motečih lastnosti in analiza vezanih faktorjev. Prvemu postopku se bomo posvetili v tem poglavju, drugega pa bomo obdelali v naslednjem.

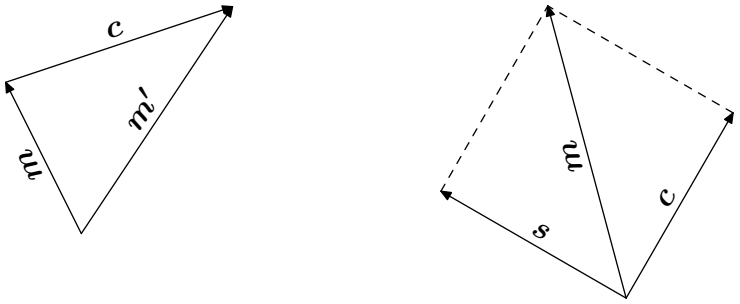
6.1 Modeliranje sejne spremenljivosti

Na področju samodejnega razpoznavanja govorcev se je v zadnjem obdobju med raziskovalci izoblikovali prepričanje, da se bo za izboljšanje robustnosti razpoznavalnikov potrebno intenzivno spopasti s problemom sejne spremenljivosti (angl. session variability). Izkazalo se je, da sejne spremenljivosti ne bo mogoče zadovoljivo izničiti le na nivoju signala. Tako je že pred iznajdbo postopka NAP bila predlagana vrsta postopkov, ki skušajo normalizirati sejno spremenljivost na nivoju modela. Med njimi se zdi koristno izpostaviti predvsem dva — to sta *sinteza govorskih modelov*

(angl. speaker model synthesis, SMS) (Teunen et al., 2000) in *postopek preslikave značilk* (angl. feature mapping, FM) (Reynolds, 2003).

Tako SMS kot FM temeljita na predpostavki, da so kanalski vplivi diskretne narave in da je kanal mogoče detektirati v postopku predobdelave. Diskretnost kanala pomeni, da je kanal mogoče razvrstiti v končno število razredov. (V praksi je število razredov pogosto odvisno od števila različnih tipov uporabljenih mikrofonov oz. prenosnih poti.)

Oba postopka obravnavata govorne posnetke kot točke v evklidskem prostoru. Osnovna ideja postopka SMS je, da obstaja za vsak par (učni kanal,testni kanal) od govorca neodvisen vektor $\boldsymbol{c}$, ki preseli govorca iz učnega v testni kanal (slika **6.1a**). Še korak dlje gre postopek FM, ki predpostavlja, da je vsakega govorca mogoče preseliti v nevtralen kanal in ga od tam po potrebi preseliti v katerikoli drug kanal (slika **6.1b**).



(a) (b)

Slika 6.1 Ideja normalizacije sejne variabilnosti. Vektor $\boldsymbol{m}$ predstavlja govorca v učnem kanalu, vektor $\boldsymbol{m}'$ pa tega istega govorca v testnem kanalu (a). Vektor $\boldsymbol{s}$ predstavlja govorca v nevtralnem kanalu, ki ga po potrebi preselimo v poljuben kanal $\boldsymbol{c}$ (b).

Kljub precejšnji učinkovitosti imata oba postopka določene pomanjkljivosti. Pri obeh je namreč v učnem delu potrebno ročno označiti vrsto kanala za vsak posnetek posebej. Tako označevanje je lahko precej zamudno in podvrženo napakam. Zato se je pojavila potreba po nadgraditvi teh postopkov. Skoraj istočasno sta bila predlagana dva postopka, ki predpostavko o diskretnem kanalu posplošita na zvezne kanale in s tem za vselej opravita s potrebo po ročnem označevanju vrste kanala v posnetkih. Eden izmed teh dveh postopkov je postopek NAP, s katerim se posebej ukvarjamo v tem poglavju.

6.2 Linearni model govorske in kanalske komponente

Postopek NAP predpostavlja, da lahko vsak govorni signal obravnavamo kot točko v visokorazsežnem vektorskem prostoru in da lahko v tem prostoru govorni signal, ki ga označimo s črko $\boldsymbol{m}$, linearno razstavimo na dve komponenti; govorsko komponento $\boldsymbol{s}$ in kanalsko komponento $\boldsymbol{c}$:

$$\boldsymbol{m} = \boldsymbol{s} + \boldsymbol{c} \tag{6.1}$$

Na prvi pogled se zdi predpostavka o linearni dekompoziciji privzeta precej *ad hoc* in je po vsej verjetnosti bolj posledica matematične prikladnosti kot pa odsev realnih razmer. Z nekaj truda se da pokazati, da temu ni povsem tako (Vesnicer in Mihelič, 2008).

Najprej predpostavimo, da je kanalski vpliv v večji meri posledica konvolutivnega šuma, povzročenega zaradi uporabe različnih tipov mikrofonov in prenosnih poti. Konvolutivni šum se zaradi uporabe logaritmiranja v prostoru akustičnih značilk odraža kot aditivni šum. Vsak vektor značilk lahko zato obravnavamo kot vsoto dveh neodvisnih naključnih spremenljivk; ena pripada govoru, druga pa kanalu. Iz verjetnostnega računa vemo, da je porazdelitvena funkcija vsote neodvisnih naključnih spremenljivk podana s konvolucijo posameznih porazdelitev. Brez izgube splošnosti[26] lahko predpostavimo, da je porazdelitvena funkcija obeh naključnih spremenljivk podana z Gaussovo mešanico. Pokažemo lahko (Vesnicer in Mihelič (2008), dodatek A), da je konvolucija dveh Gaussovih mešanic spet Gaussova mešanica. Natančneje, če ima prva mešanica K_1 komponent in druga K_2 komponent in označimo uteži, povprečja in variance prve mešanice s črkami α_{1i}, μ_{1i} in σ_{1i}, druge mešanice pa z α_{2i}, μ_{2i} in σ_{2i}, je konvolucija obeh sestavljena iz K_1K_2 komponent z utežmi $\alpha_{1i}\alpha_{2j}$, povprečji $\mu_{1i}+\mu_{2j}$ in variancami $\sigma_{1i}+\sigma_{2j}$. S tem smo uspeli upravičiti predpostavko o linearnosti.

V resnici so zadeve malenkost bolj zakomplicirane, saj imamo poleg konvolutivnega šuma primešan tudi aditivni šum, zaradi katerega postane ›funkcija‹ kanalske in govorske naključne spremenljivke nelinearna.

6.3 Preslikava v prostor supervektorjev

Ena izmed možnosti, kako predstaviti govorni posnetek kot točko v visokorazsežnem vektorskem prostoru, sloni na uporabi modela GMM. V razdelku **3.5** smo si pogledali, kako iz modela UBM z adaptacijo MAP izpeljemo model govorca. Če adaptiramo le povprečne vrednosti, ne pa tudi kovariančnih matrik in mešalnih koeficientov, dobimo za vsakega govorca oz. posnetek model GMM, ki se od vseh ostalih razlikuje le v vektorjih povprečnih vrednosti posameznih komponent. Adaptacija MAP igra pri tem ključno vlogo, saj se izkaže, da ohranja urejenost komponent in nas s tem zaščiti pred problemom identifikabilnosti (glej razdelek **4.5.2**). S tem nam omogoča, da povprečne vektorje posameznih komponent staknemo v en sam vektor, za katerega se je uveljavilo ime *supervektor*.

6.4 Predpostavka o kanalskem podprostoru

Postopek adaptacije MAP nam torej omogoča, da govorni signal, ki je predstavljen v obliki niza vektorjev značilk, preslikamo v prostor supervektorjev. V idealnem primeru bi vse posnetke, ki bi jih izgovoril isti govorec, z opisanim postopkom preslikali

[26] Pokazati se da, da lahko z Gaussovo mešanico aproksimiramo kakršnokoli porazdelitveno funkcijo do poljubne natančnosti, če le uporabimo dovolj veliko število komponent v mešanici.

v isti supervektor, medtem ko bi posnetke različnih govorcev preslikali v različne supervektorje. Z drugimi besedami, želimo si, da bi bile slike posnetkov istega govorca čimbolj podobne ena drugi, slike posnetkov istih govorcev pa med seboj čimbolj različne. Še drugače, znotrajgovorska spremenljivost naj bo čim manjša, izvengovorska spremenljivost pa čim večja. V praksi temu ni vedno tako. Največkrat vzroke za to pripisujemo neenakim akustičnim razmeram v različnih posnetkih. Pomembno vlogo pa lahko igrajo tudi nekateri manj očitni dejavniki. Naštejmo jih nekaj:

- izgovorjeno besedilo;
- govorčevo psihofizično stanje;
- učinek staranja;
- uporaba različnih mikrofonov, telefonskih aparatov in prenosnih poti;
- šumno akustično ozadje.

Če se želimo znebiti omenjenih vplivov, moramo poiskati način, kako sejni supervektor $\boldsymbol{m}$ razkleniti na vsoto govorskega supervektorja $\boldsymbol{s}$ in kanalskega supervektorja $\boldsymbol{c}$. V splošnem je ta problem nerešljiv oz. ima neskončno rešitev, saj imamo eno enačbo in kar dve neznanki. Na srečo postane problem rešljiv, če predpostavimo, da je vpliv kanala omejen na nižjedimenzionalni podprostor osnovnega prostora supervektorjev. Ta predpostavka se zdi smiselna, saj pričakujemo, da kanal ne more povzročiti, da bi dva različna govorca ›zvenela‹ enako. (V nasprotnem bi bilo razpoznavanje govorcev slabo definiran problem.)

Predpostavko o omejenosti kanalskih vplivov na kanalski podprostor lahko formalno zapišemo z enačbo:

$$\boldsymbol{c} = \boldsymbol{U}\boldsymbol{x},$$

kjer je $\boldsymbol{U}$ *matrika lastnih kanalov* (angl. eigenchannel matrix), vektor $\boldsymbol{x}$ pa predstavitev kanalske komponente $\boldsymbol{c}$ v kanalskem podprostoru.

6.5 Ocena kanalskega podprostora

Opišimo postopek, kako poiščemo bazo kanalskega podprostora. Na voljo moramo imeti govorno zbirko s čim večjim številom govorcev, vsak med njimi pa mora biti posnet v čim več različnih sejah.

Naj bo N_i število posnetkov oz. sej i-tega govorca. Označimo z $\boldsymbol{m}_{ij}$ j-ti posnetek i-tega govorca. Za vsakega govorca najprej izračunamo povprečni supervektor preko vseh sej:

$$\boldsymbol{\mu}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \boldsymbol{m}_{ij}.$$

S povprečenjem smo se — ob predpostavki, da je povprečna vrednost kanala preko večjega števila sej istega govorca (približno) enaka ničelnemu vektorju — kanala (vsaj delno) znebili. To nam omogoča, da lahko iz vsakega sejnega vektorja izluščimo kanalski supervektor tako, da danemu sejnemu vektorju odštejemo pripadajoči

povprečni supervektor. Tako dobljene supervektorje združimo v matriko kanalskih supervektorjev, ki jo označimo z $\boldsymbol{A}$. Bazo prostora, ki ga kanalski supervektorji oklepajo, v splošnem poiščemo z razcepom singularnih vrednosti (angl. singular value decomposition, SVD). Za poljubno $m \times n$-razsežno matriko $\boldsymbol{A}$ z realnimi koeficienti velja:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T,$$

kjer je $\boldsymbol{U}$ $(m \times m)$-razsežna ortonormalna matrika, ki ustreza lastnim vektorjem produkta $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$, $\boldsymbol{V}$ je $(n \times n)$-razsežna ortonormalna matrika, ki ustreza lastnim vektorjem produkta $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ in $\boldsymbol{\Sigma}$ je $(m \times n)$-razsežna diagonalna matrika, ki ima po diagonali razporejena nenegativna realna števila — singularne vrednosti matrike $\boldsymbol{A}$. Za natančnejšo razlago razcepa singularnih vrednosti — enega izmed pomembnejših rezultatov linearne algebre — kakor tudi ostalih pojmov s tega področja priporočamo ogled odličnega učbenika (Strang, 2009).

Poleg metode razcepa singularnih vrednosti lahko bazo kanalskega prostora določimo tudi z razcepom lastnih vrednosti (angl. eigenvalue decomposition):

$$\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}},$$

kjer je matrika $\boldsymbol{D}$ $(m \times m)$-razsežna diagonalna matrika lastnih vrednosti.

Med obema razcepoma obstaja naslednja zveza:

$$\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\left(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\right)^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}.$$

Po predpostavki, da je kanal omejen na nizkodimenzionalen podprostor osnovnega prostora supervektorjev, pričakujemo, da bomo lahko večino kanalske spremenljivosti zajeli z relativno majhnim številom lastnih vektorjev. Z drugimi besedami, če uredimo lastne vrednosti po velikosti, se nadejamo, da bodo le-te padale zelo hitro proti 0.

Pri implementaciji se lahko stvari zapletejo, saj je lahko dimenzija matrike $\boldsymbol{A}$ prevelika, da bi jo naenkrat prebrali v pomnilnik računalnika. Kratek izračun pokaže, da bi v našem primeru matrika $\boldsymbol{A}$, če bi njene elemente zapisali z dvojno natančnostjo, zasedala skoraj 15 GB pomnilnika. Ker takšne količine pomnilnika večina današnjih (osebnih) računalnikov ne premore[27], smo bili pri našem delu velikokrat primorani opustiti uporabo priročnih orodij kot je MATLAB™ in razviti lastne programske rešitve.

Četudi bi imeli na voljo računalnik, ki bi imel dovolj pomnilnika, da bi vanj lahko zapisali matriko $\boldsymbol{A}$, pa bi se znalo zaplesti pri izračunu kovariančne matrike $\boldsymbol{A}\boldsymbol{A}^T$, ki jo potrebujemo pri iskanju baze prostora s pomočjo razcepa lastnih vrednosti.

Na prvi pogled se zdi, da je edina možnost, da namesto razcepa lastnih vrednosti uporabimo razcep SVD. A ker je število stolpcev matrike $\boldsymbol{A}$ (v našem primeru 16252) veliko manjše od števila vrstic (v našem primeru 122880), se nam ponuja še

[27] Veliko današnjih računalnikov (in operacijskih sistemov) je še 32-bitnih, zato lahko naslovijo največ 4 GB pomnilnika.

učinkovitejša rešitev. Vemo namreč, da je število od nič različnih lastnih vrednosti enako rangu matrike $\boldsymbol{A}$ in je torej v našem primeru manjše ali kvečjemu enako številu stolpcev, kar lahko s pridom izkoristimo. Izkaže se namreč, da so lastne vrednosti matrike $\boldsymbol{A}^T\boldsymbol{A}$ enake neničelnim lastnim vrednostim matrike $\boldsymbol{A}\boldsymbol{A}^T$. To lahko pokažemo tako, da obe strani enačbe

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{D}$$

pomnožimo z leve strani z matriko $\boldsymbol{A}$, kar da

$$\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{U} = \boldsymbol{A}\boldsymbol{U}\boldsymbol{D}.$$

S primerjavo obeh enačb opazimo, da so lastni vektorji produkta $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ enaki lastnim vektorjem produkta $\boldsymbol{A}^T\boldsymbol{A}$, pomnoženimi z matriko $\boldsymbol{A}$.

Kanalski podprostor definiramo kot podprostor osnovnega prostora supervektorjev, ki ga oklepajo lastni vektorji, ki ustrezajo prvim nekaj po vrednosti urejenim lastnim vrednostim. Dimenzijo kanalskega podprostora ponavadi določimo izkustveno.

6.6 Razčlenitev sejnega supervektorja

Ko smo ocenili kanalsko matriko $\boldsymbol{V}$, lahko vsak sejni supervektor $\boldsymbol{m}$ razčlenimo na vsoto kanalskega supervektorja $\boldsymbol{c}$ in govorskega supervektorja $\boldsymbol{s}$. Očitno je, da lahko kanalsko komponento iz sejnega supervektorja izoliramo tako, da sejnemu supervektorju najprej odštejemo povprečni supervektor $\boldsymbol{\mu}$ in ga nato projiciramo v kanalski podprostor ter takoj spet nazaj v osnovni prostor. Ko imamo enkrat kanalsko komponento, nas do rekonstrukcije govorske komponente loči le še trivialno odštevanje. Opisano razčlenitev lahko precizneje podamo v jeziku linearne algebre:

$$\begin{aligned}
\boldsymbol{x} &= \boldsymbol{V}^T(\boldsymbol{m} - \boldsymbol{\mu}) \\
\boldsymbol{c} &= \boldsymbol{V}\boldsymbol{x} = \boldsymbol{V}\boldsymbol{V}^T(\boldsymbol{m} - \boldsymbol{\mu}) \\
\boldsymbol{s} &= \boldsymbol{m} - \boldsymbol{c} = (\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^T)(\boldsymbol{m} - \boldsymbol{\mu}) + \boldsymbol{\mu}
\end{aligned}$$

Z oznako $\boldsymbol{\mu}$ smo označili povprečni vektor preko vseh sej vseh govorcev, ki jih imamo na razpolago. Izkaže se, da lahko največkrat shajamo kar brez odštevanja povprečnega supervektorja $\boldsymbol{\mu}$ in se zadovoljimo s približkom, saj pričakujemo, da bo povprečni supervektor ležal ›v bližnji okolici‹ *ničelnega prostora* matrike $\boldsymbol{A}^T$. Ob tej predpostavki velja:

$$\begin{aligned}
\boldsymbol{x} &\approx \boldsymbol{V}^T\boldsymbol{m} \\
\boldsymbol{c} &\approx \boldsymbol{V}\boldsymbol{V}^T\boldsymbol{m} \\
\boldsymbol{s} &\approx (\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^T)\boldsymbol{m}
\end{aligned}$$

6.7 Razpoznavanje govorcev z uporabo postopka NAP

Postopek NAP je bil razvit za uporabo v razpoznavalnikih govorcev, ki za razvrščanje uporabljajo metodo podpornih vektorjev. Celoten postopek razpoznavanja lahko povzamemo v nekaj korakih. Na že opisan način najprej preslikamo tako učni kot tudi testni posnetek v prostor supervektorjev. Nato izvedemo razčlenitev obeh supervektorjev na govorski in kanalski komponenti. Ker nas kanalski komponenti ne zanimata, ju lahko zavržemo, obdržimo pa obe govorski komponenti. Govorsko komponento učnega posnetka uporabimo za oceno modela SVM. Rezultat razpoznavanja izračunamo kot razdaljo med govorsko komponento testnega posnetka in modelom SVM, ki smo ga ocenili na podlagi govorske komponente učnega supervektorja.

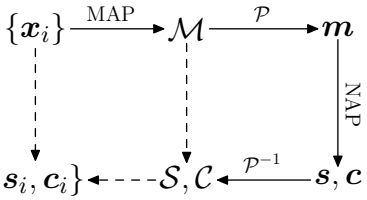
6.8 Postopek NAP in kriterij razmerja verjetij

Kot alternativo uveljavljenemu načinu uporabe postopka NAP v razpoznavalnikih govorcev, katerih razvrščanje temelji na metodi SVM, v disertaciji predlagamo izvirno rešitev, kako postopek NAP uporabiti tudi v razpoznavalnikih, katerih razvrščanje temelji na kriteriju razmerja verjetij. Preden opišemo predlagani postopek, je koristno, če se zavedamo določenih razlik med obema kriterijema.

Glavna razlika med obema kriterijema izhaja iz različnih učnih postopkov. Medtem ko postopek SVM temelji na diskriminatornem učenju, je model GMM in z njim povezan kriterij LR generativne narave. To pomeni, da za razliko od učenja modela GMM, za učenje modela SVM nujno potrebujemo tako pozitiven (učni) primer kot tudi negativne primere. Druga, za nas bistvenejša razlika je ta, da poteka razpoznavanje v primeru kriterija SVM v prostoru supervektorjev, v primeru kriterija LR pa v prostoru akustičnih vektorjev značilk. Tukaj naletimo na oviro, saj je projekcija NAP definirana v prostoru supervektorjev. Če želimo uporabiti kriterij LR, moramo torej poiskati način, kako supervektorje preslikati nazaj v akustični prostor.

Ker vemo, da smo supervektor dobili iz modela GMM tako, da smo povprečne vektorje posameznih komponent staknili v en sam vektor, ga lahko brez težav preslikamo spet nazaj v model GMM. Od tod ideja, da po adaptaciji MAP dobljeni GMM preslikamo v prostor supervektorjev, tam s pomočjo postopka NAP opravimo dekompozicijo na kanalsko in govorsko komponento in nato govorsko komponento preslikamo spet nazaj v model GMM. Tako lahko postopamo za učni, ne pa tudi za testni posnetek. Pri računanju verjetja mora namreč testni posnetek biti predstavljen v obliki niza vektorjev akustičnih značilk. Kanalskega vpliva pa iz niza vektorjev značilk ne znamo enolično izločiti, saj preslikava niza vektorjev značilk v model GMM ni injektivna.

Na sliki 6.2 je prikazan diagram prehodov med različnimi predstavitvami govornega signala oz. govorca. S črtkanimi povezavami so označene preslikave, ki jih ne znamo izvesti. Diagram ponazarja v matematiki pogosto uporabljano taktiko, ko problem, ki ga v osnovnem prostoru ne znamo rešiti, preslikamo v drug prostor, v katerem rešitev znamo poiskati.



Slika 6.2 Diagram prehodov.

Lahko bi se zadovoljili s podoptimalno rešitvijo in bi kanal odstranili le iz učnega posnetka, medtem ko bi testni niz vektorjev značilk pustili ›okužen‹ s kanalom. Vendar je bolje, če učni model GMM spremenimo tako, da ga prilagodimo kanalu testnega posnetka. To dosežemo tako, da učnemu sejnemu supervektorju odštejemo učni kanalski supervektor in mu prištejemo testni kanalski supervektor. Tako spremenjen učni supervektor nato le še preslikamo nazaj v model GMM. Tako smo dobili model GMM, ki je prilagojen testnemu kanalu. Pomembno je, da pred izračunom verjetja ne pozabimo ustrezno spremeniti tudi modela UBM.

6.9 Komentar

Predstavili smo postopek NAP, ki se je izkazal za zelo učinkovitega pri normalizaciji kanalskih vplivov. Postopek predpostavlja, da je mogoče vsak govorni posnetek predstaviti v obliki supervektorja. Preslikava iz prostora akustičnih značilk v prostor supervektorjev je izvedena posredno preko modela GMM, ki ga ocenimo s postopkom MAP. Postopek NAP je bil razvit za uporabo v razpoznavalnikih, ki za razvrščanje uporabljajo metodo SVM, v disertaciji pa smo postopek prilagodili tako, da ga je mogoče uporabiti tudi v navezi s kriterijem razmerja verjetij. Omenimo še, da postopek NAP daje dobre rezultate tudi pri sistemih, ki za preslikavo v prostor supervektorjev namesto adaptacije MAP uporabljajo adaptacijo MLLR (angl. maximum likelihood linear regression, MLLR) (Leggetter in Woodland, 1995), vendar to transformacijo ne ocenijo na modelu GMM, temveč jo pridobijo iz razpoznavalnika govora, ki temelji na modelih HMM (Stolcke et al., 2009).

Osnovna ideja postopka NAP je precej preprosta a hkrati dovolj splošna, da jo je mogoče prenesti tudi izven konteksta razpoznavanja govorcev. Tako smo ga uspešno uporabili za normalizacijo osvetlitve slik pri razpoznavanju oseb na osnovi slik obrazov (Štruc et al., 2009). Verjamemo, da bi ga bilo mogoče s pridom uporabiti tudi na področju razpoznavanja čustev, a tega zaenkrat še nismo imeli časa eksperimentalno potrditi.

ANALIZA VEZANIH FAKTORJEV 7

V tem poglavju se bomo ukvarjali s še enim postopkom za normalizacijo sejne variabilnosti, ki so mu avtorji nadeli ime *analiza vezanih faktorjev* (angl. joint factor analysis, JFA) (Kenny, 2005). Kot bomo videli, je postopek JFA v marsičem zelo podoben postopku NAP. Oba slonita na predpostavki, da lahko govorni signal predstavimo kot točko v prostoru supervektorjev in da je kanalska spremenljivost omejena na nižjedimenzionalni podprostor tega prostora. Čeprav je osrednji problem obeh postopkov problem lastnih vrednosti, pa se tega problema lotita na različna načina. Medtem ko se postopek NAP problema razčlenitve sejnega supervektorja loti izključno z algebrajskimi orodji, pa postopek JFA omenjeni problem rešuje v okviru verjetnostnega računa. Kot bomo videli, je za razumevanje sejne spremenljivosti v luči verjetnostnega računa potrebno izvesti ›miselni preskok‹, ki pa nam odpre celovitejši vpogled na problematiko, s katero se ukvarjamo.

7.1 Predpostavke postopka JFA

Prav tako kot NAP, tudi postopek JFA temelji na domnevi, da lahko porazdeljevanje vektorjev značilk v akustičnem prostoru predstavimo s parametričnim modelom GMM. Če imamo na voljo veliko število različnih posnetkov večjega števila govorcev, lahko s postopkom EM ocenimo model GMM, ki ga krajše označimo s kratico UBM. (Postopek ocene modela UBM smo že predstavili v poglavju **3.4**.)

Naj bo K število Gaussovih komponent modela UBM in D razsežnost akustičnega prostora. Definirajmo supervektor kot KD-razsežni vektor, ki ga dobimo, če zložimo povprečne vektorje posameznih komponent enega nad drugega. Predpostavimo, da lahko od govorca in kanala odvisni sejni supervektor $\boldsymbol{m}$ razstavimo na vsoto dveh *statistično neodvisnih in normalno porazdeljenih*[28] supervektorjev, govorski supervektor $\boldsymbol{s}$ in kanalski supervektor $\boldsymbol{c}$:

$$\boldsymbol{m} = \boldsymbol{s} + \boldsymbol{c}.$$

Nadalje predpostavimo, da lahko porazdelitev supervektorja $\boldsymbol{s}$ zapišemo v obliki

$$\boldsymbol{s} = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{D}\boldsymbol{z}. \tag{7.1}$$

Oznake v enačbi (**7.1**) pomenijo naslednje: $\boldsymbol{\mu}$ je povprečni supervektor; $\boldsymbol{V}$ je pravokotna matrika nižjega ranga; $\boldsymbol{y}$ je standardno normalno porazdeljen naključni vektor; $\boldsymbol{D}$ je diagonalna matrika[29] in $\boldsymbol{z}$ je standardno normalno porazdeljen supervektor

[28] Vidimo, da supervektorje obravnavamo kot naključne spremenljivke.
[29] Z nekaj truda se da model JFA posplošiti tudi na primer, ko je $\boldsymbol{D}$ bločno diagonalna matrika.

dimenzije KD. Stolpcem matrike $\boldsymbol{V}$ pravimo *lastni glasovi* (angl. eigenvoices), komponentam vektorja $\boldsymbol{y}$ pa *govorski faktorji* (angl. speaker factors)[30].

Podobno predpostavimo, da lahko porazdelitev supervektorja $\boldsymbol{c}$ zapišemo v obliki

$$\boldsymbol{c} = \boldsymbol{U}\boldsymbol{x}, \tag{7.2}$$

kjer je $\boldsymbol{U}$ pravokotna matrika nižjega ranga in $\boldsymbol{x}$ standardno normalno porazdeljen naključni vektor. Komponentam vektorja $\boldsymbol{x}$ pravimo *kanalski faktorji* (angl. channel factors), stolpcem matrike $\boldsymbol{U}$ pa lastni kanali (angl. eigenchannels).

Nazadnje definirajmo za vsako komponento Gaussove mešanice še kovariančno matriko $\boldsymbol{\Sigma}_k$, s katerimi opišemo vso preostalo spremenljivost v akustičnem prostoru, ki ni zajeta niti s strani govorskega modela (7.1) niti s strani kanalskega modela (7.2). Podobno kot smo vpeljali supervektor, lahko tvorimo superkovariančno matriko $\boldsymbol{\Sigma}$ tako, da njene diagonalne bloke sestavimo iz posameznih matrik $\boldsymbol{\Sigma}_k$.

Predstavljeni model JFA ima veliko skupnega z modelom GMM, a se od njega bistveno razlikuje po konceptualni plati. V klasičnem modelu GMM kot naključne spremenljivke obravnavamo le opažene vektorje akustičnih spremenljivk (skupaj s pripadajočimi prikritimi spremenljivkami, glej razdelek 4.2), medtem ko tukaj kot naključne spremenljivke obravnavamo tudi povprečne vektorje posameznih komponent, ki so v modelu GMM zgolj točkasti parametri. Ideja, da lahko parametrom modela pripišemo porazdelitve — in jih tako obravnavamo kot naključne spremenljivke — je v statistični skupnosti še vedno precej kontroverzna in med statistiki ni splošno sprejeta. Statistikom, ki zagovarjajo mnenje, da lahko parametre modela in celo hipoteze obravnavamo kot naključne količine, pravimo Bayesovci, statistikom, ki tako obravnavo zavračajo, pa frekvencionisti.

Obravnava, ki jo predstavljamo v tem razdelku, zahteva določen miselni preskok glede na klasičen GMM-UBM pristop, saj iz porazdeljevanja vektorjev značilk v akustičnem prostoru prehajamo na porazdeljevanje parametrov (povprečnih vektorjev) v prostoru supervektorjev.

7.2 Povezava s faktorsko analizo

Opazimo lahko, da je model JFA splošnejši od modela NAP. Ne samo, da smo sedaj predpostavke o kanalskem prostoru postavili v verjetnostni okvir, z enačbo (7.1) smo dodatno vpeljali še govorski podprostor. Ta enačba je tudi glavni razlog za ime postopka JFA, saj vključuje dva vezana faktorja — spremenljivki $\boldsymbol{y}$ in $\boldsymbol{z}$. Če v enačbi izpustimo člen $\boldsymbol{V}\boldsymbol{y}$, dobimo nekoliko manj splošen model, ki mu pravimo kar *analiza faktorjev* (angl. factor analysis, FA). V nadaljevanju bomo videli, da je model FA soroden modelu NAP, le da je zavit v verjetnostni okvir.

Če enačbi (7.1) in (7.2) združimo, dobimo:

$$\boldsymbol{m} = \boldsymbol{\mu} + \boldsymbol{U}\boldsymbol{x} + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{D}\boldsymbol{z}.$$

[30] V statistični literaturi se običajno uporablja drugačna terminologija; stolpcem matrike $\boldsymbol{V}$ pravijo govorski faktorji, komponentam vektorja $\boldsymbol{y}$ pa faktorske uteži.

Če uvedemo vektor $\boldsymbol{w}$, matriko $\boldsymbol{W}$ in vektor $\boldsymbol{\epsilon}$:

$$\boldsymbol{w} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}, \quad \boldsymbol{W} = \begin{pmatrix} \boldsymbol{U} & \boldsymbol{V} \end{pmatrix} \quad \text{in} \quad \boldsymbol{\epsilon} = \boldsymbol{D}\boldsymbol{z},$$

lahko gornjo enačbo zapišemo kot:

$$\boldsymbol{m} = \boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{w} + \boldsymbol{\epsilon}, \tag{7.3}$$

kar prepoznamo kot standardni zapis *faktorske analize* (B.1). Vektorju $\boldsymbol{\epsilon}$, ki nastopa v faktorski enačbi, pogosto pravimo šum. Ker je $\boldsymbol{z}$ standardno normalno porazdeljen, sledi: $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}\boldsymbol{D}^{\mathrm{T}})$.

7.3 Dvonivojski naključni proces

Zdi se, da bomo lahko rezultate, ki smo jih izpeljali v razdelku B, kjer smo obdelali teorijo faktorske analize, kar brez sprememb prepisali za naš primer. Izkaže se, da temu ni čisto tako. Klasična FA namreč predpostavlja, da je naključna spremenljivka $\boldsymbol{m}$, ki nastopa na desni strani enačbe, opažena. To v našem primeru na žalost ne drži. O vrednosti te spremenljivke lahko le sklepamo na podlagi niza vektorjev akustičnih značilk, ki pa so opaženi. Vidimo, da je zahtevnost našega problema mnogo večja od zahtevnosti klasične FA, saj imamo pred seboj dvonivojski naključni proces.

7.4 Izpeljava enačb

Pri izpeljavi enačb, ki jih potrebujemo za implementacijo analize vezanih faktorjev, bomo sledili izvajanjem v (Kenny, 2005). Privzemimo, da imamo na voljo H_s posnetkov govorca s.

Ker imamo namesto enega posnetka H_s posnetkov, moremo zapisati prav toliko faktorskih enačb:

$$\boldsymbol{m}_h = \boldsymbol{\mu} + \boldsymbol{U}\boldsymbol{x}_h + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{D}\boldsymbol{z}, \qquad h = 1, \ldots, H_s,$$

kjer smo zaradi preglednosti pri vseh naključnih spremenljivkah izpustili indeks s. Po predpostavki je vseh H_s posnetkov izgovoril govorec s, zato si vseh H_s enačb deli govorski spremenljivki $\boldsymbol{y}$ in $\boldsymbol{z}$, medtem ko je kanalskih spremenljivk $\boldsymbol{x}_h$ toliko, kolikor je različnih posnetkov. Če uvedemo spremenljivki $\boldsymbol{W}$ in $\boldsymbol{w}$:

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{U} & \cdots & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{V} & \boldsymbol{D} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{U} & \cdots & \boldsymbol{0} & \boldsymbol{V} & \boldsymbol{D} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \cdots & \boldsymbol{U} & \boldsymbol{V} & \boldsymbol{D} \end{pmatrix} \quad \text{in} \quad \boldsymbol{w} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_h \\ \vdots \\ \boldsymbol{x}_{H_s} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix}$$

lahko vseh H_s enačb preoblikujemo v eno enačbo:

$$\begin{pmatrix} \boldsymbol{m}_1 \\ \vdots \\ \boldsymbol{m}_h \\ \vdots \\ \boldsymbol{m}_{H_s} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{\mu} \end{pmatrix} + \boldsymbol{W}\boldsymbol{w}.$$

Denimo, da razpolagamo s polnim naborom podatkov, t.j. za vsak akustični vektor značilk vemo, katera komponenta modela GMM ga je porodila. Podatke za vseh H_s posnetkov danega govorca združimo v niz o. Zanimalo nas bo:

i) Kakšna je pogojna porazdelitev opaženih spremenljivk o pri dani vrednosti prikritih spremenljivk $\boldsymbol{w}$; $p(o|\boldsymbol{w})$?
ii) Kakšna je posteriorna porazdelitev prikritih spremenljivk $\boldsymbol{w}$ pri dani vrednosti opaženih spremenljivk o; $p(\boldsymbol{w}|o)$?
iii) Kakšna je robna porazdelitev opaženih spremenljivk o; $p(o)$?
iv) Kako določiti hiperparametre $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{U}$, $\boldsymbol{V}$ in $\boldsymbol{D}$?
v) Kako uporabiti model JFA za razpoznavanje govorcev?

Da bomo lahko odgovorili na ta vprašanja, najprej definirajmo nekaj statistik. Naj bo N_{hk} število vektorjev značilk posnetka $h \in \{1, \ldots, H_s\}$, ki pripadajo komponenti k. Podobno za vsak posnetek h in vsako komponento k definirajmo še statistike prvega in drugega reda: $\boldsymbol{F}_{hk} = \sum_n (\boldsymbol{o}_{hn} - \boldsymbol{\mu}_k)$ in $\boldsymbol{S}_{hk} = \sum_n (\boldsymbol{o}_{hn} - \boldsymbol{\mu}_k)(\boldsymbol{o}_{hn} - \boldsymbol{\mu}_k)^{\mathrm{T}}$, kjer vsota teče preko vseh vektorjev značilk $\boldsymbol{o}_{hn}$, ki so poravnani s komponento k. Iz teh statistik na že znan način (glej razdelek **5.1.1**, str. **45**) tvorimo matrike $\boldsymbol{N}_h$, vektorje $\boldsymbol{F}_h$ in matrike $\boldsymbol{S}_h$. Zaradi prikladnejše notacije definirajmo še matriko $\boldsymbol{N}$, ki jo tvorimo tako, da po njeni diagonali razporedimo vse matrike $\boldsymbol{N}_h$ ter vektor $\boldsymbol{F}$, ki ga dobimo tako, da postavimo posamezne vektorje $\boldsymbol{F}_h$ enega nad drugega.

7.4.1 Pogojna porazdelitev opaženih spremenljivk

Če predpostavimo, da so vrednosti prikritih spremenljivk znane, lahko z nekaj truda zapišemo izraz za pogojno porazdelitev opaženih spremenljivk kot vsoto dveh členov, od katerih je le eden odvisen od $\boldsymbol{w}$:

$$\log p(o|\boldsymbol{w}) = G + H(\boldsymbol{w}), \tag{7.4}$$

kjer je člen G:

$$G = \sum_{h=1}^{H_s} \sum_{k=1}^{K} N_{hk} \left(\log \pi_k - \frac{D}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| \right) - \frac{1}{2} \mathrm{Tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{S}_h),$$

člen $H(\boldsymbol{w})$ pa:

$$H = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{F} - \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{N}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}\boldsymbol{w}.$$

V kolikor obravnavamo izraz $\log p(o|\boldsymbol{w})$ kot funkcijo spremenljivke $\boldsymbol{w}$, imenujemo ta izraz logaritem verjetja. Iz lastnosti Gaussove funkcije sledi, da je funkcija verjetja (**7.4**) Gaussova funkcija.

7.4.2 Posteriorna porazdelitev prikritih spremenljivk

Pri izpeljavi posteriorne porazdelitve prikritih spremenljivk pri danih vrednostih opaženih spremenljivk, si pomagamo z Bayesovim izrekom:

$$p(\boldsymbol{w}|o) = \frac{p(o|\boldsymbol{w})p(\boldsymbol{w})}{p(o)} \propto p(o|\boldsymbol{w})p(\boldsymbol{w}).$$

Ker je po definiciji prior $p(\boldsymbol{w})$ porazdeljen normalno in ker ima tudi funkcija verjetja Gaussovo obliko, sledi, da bo tudi posterior porazdeljen normalno. (Vemo namreč, da je družina Gaussovih funkcij konjugirana sama sebi.)

Izkaže se, da lahko izraz za posteriorno porazdelitev $p(\boldsymbol{w}|o)$ zapišemo kot:

$$p(\boldsymbol{w}|o) = \mathcal{N}\left(\boldsymbol{w}\middle|\boldsymbol{L}^{-1}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{F}, \boldsymbol{L}^{-1}\right), \tag{7.5}$$

kjer velja:

$$\boldsymbol{L} = \boldsymbol{I} + \boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{W}.$$

7.4.3 Robna porazdelitev opaženih spremenljivk

Robno porazdelitev opaženih spremenljivk $p(o)$ izpeljemo iz vezane porazdelitve $p(o, \boldsymbol{w})$ z upoštevanjem osnovnih pravil verjetnosti:

$$p(o) = \int p(o, \boldsymbol{w})\mathrm{d}\boldsymbol{w} = \int p(o|\boldsymbol{w})p(\boldsymbol{w})\mathrm{d}\boldsymbol{w}.$$

Na srečo je gornji integral v našem primeru analitično rešljiv. Kot rezultat dobimo sledeč izraz:

$$\begin{aligned}\log p(o) &= G - \frac{1}{2}\log|\boldsymbol{L}| + \frac{1}{2}\mathbb{E}\left[\boldsymbol{w}^{\mathrm{T}}\middle|o\right]\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{F} \\ &= G - \frac{1}{2}\log|\boldsymbol{L}| + \frac{1}{2}\left|\left|\boldsymbol{L}^{-1/2}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{F}\right|\right|^2,\end{aligned} \tag{7.6}$$

kjer je $\boldsymbol{L}^{-1/2}$ zgornje trikotna matrika, ki jo dobimo z razcepom Cholesky matrike $\boldsymbol{L}^{-1}$.

7.4.4 Sočasno ocenjevanje hiperparametrov po kriteriju največjega verjetja

Za določitev hiperparametrov, ki nastopajo v modelu analize vezanih faktorjev, potrebujemo govorno zbirko z večjim številom govorcev, pri čemer mora vsak izmed govorcev biti posnet v čim več različnih sejah.

Postopek EM za *sočasno* ocenjevanje hiperparametrov $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{U}$, $\boldsymbol{V}$ in $\boldsymbol{D}$ po kriteriju največjega verjetja[31] je izpeljal (Kenny, 2005) v razdelku IV. Kljub temu, da je postopek dokaj zapleten (še težje pa ga je pravilno implementirati), zaradi celovitosti podajmo formule, ki jih potrebujemo pri implementaciji postopka.

Ker nas zanimajo od govorca neodvisne ocene hiperparametrov modela JFA, je smiselno, če za povprečni vektor $\boldsymbol{\mu}$ vzamemo kar oceno iz modela UBM in na novo ocenimo le ostale štiri hiperparametre. Predpostavimo, da smo v množico θ_0 zbrali začetne vrednosti hiperparametrov: $\theta_0 = \{\boldsymbol{\mu}, \boldsymbol{U}_0, \boldsymbol{V}_0, \boldsymbol{D}_0, \boldsymbol{\Sigma}_0\}$. V enem prehodu preko ustrezne zbirke izluščimo naslednjo statistiko (indeks s teče preko vseh govorcev, indeks h preko vseh posnetkov danega govorca, indeks k pa preko vseh komponent modela GMM):

$$
\begin{aligned}
N_k &= \sum_s \sum_h N_{shk}, \quad k \in \{1, \ldots, K\} \\
\boldsymbol{N}_s &= \sum_h \boldsymbol{N}_{sh}, \quad s \in \{1, \ldots, S\} \\
\boldsymbol{F}_s &= \sum_h \boldsymbol{F}_{sh}, \quad s \in \{1, \ldots, S\} \\
\boldsymbol{A}_k &= \sum_s \sum_h N_{shk}\, \mathbb{E}\left[\begin{pmatrix} \boldsymbol{x}_{sh} \\ \boldsymbol{y}_s \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_{sh} \\ \boldsymbol{y}_s \end{pmatrix}^{\mathrm{T}} \right], \quad k \in \{1, \ldots, K\} \\
\boldsymbol{B} &= \sum_s \sum_h \boldsymbol{N}_{sh}\, \mathbb{E}\left[\boldsymbol{z}_s \begin{pmatrix} \boldsymbol{x}_{sh} \\ \boldsymbol{y}_s \end{pmatrix}^{\mathrm{T}} \right] \\
\boldsymbol{C} &= \sum_s \sum_h \boldsymbol{F}_{sh}\, \mathbb{E}\left[\begin{pmatrix} \boldsymbol{x}_{sh} \\ \boldsymbol{y}_s \end{pmatrix}^{\mathrm{T}} \right] \\
\boldsymbol{a} &= \sum_s \operatorname{diag}\left(\boldsymbol{N}_s\, \mathbb{E}[\boldsymbol{z}_s \boldsymbol{z}_s^{\mathrm{T}}] \right) \\
\boldsymbol{b} &= \sum_s \operatorname{diag}\left(\boldsymbol{F}_s\, \mathbb{E}[\boldsymbol{z}_s^{\mathrm{T}}] \right),
\end{aligned}
$$

[31] Poleg kriterija največjega verjetja Kenny et al. predlaga tudi postopek določanja hiperparametrov po kriteriju najmanjše divergence, ki je uporaben pri adaptaciji modela JFA na novo populacijo govorcev ter pri izpeljavi od govorca odvisnih hiperparametrov (Kenny (2005), razdelki V, VI in VII).

kjer vsa matematična upanja izračunamo z upoštevanjem trenutnih vrednostih hiperparametrov. To je e-korak postopka EM.

V m-koraku določimo novo oceno parametrov $\boldsymbol{U}$ in $\boldsymbol{V}$ tako, da najprej definiramo $\boldsymbol{W} = \begin{pmatrix}\boldsymbol{U} & \boldsymbol{V}\end{pmatrix}$. Potem i-to vrstico podmatrike $\boldsymbol{W}$, ki ustreza k-ti komponenti modela GMM, in i-ti element diagonalne matrike $\boldsymbol{D}$, ki prav tako ustreza k-ti komponenti modela GMM, določimo tako, da rešimo naslednje enačbe:

$$\begin{pmatrix}\boldsymbol{W}_{ki} & \boldsymbol{d}_{ki}\end{pmatrix}\begin{pmatrix}\boldsymbol{A}_k & \boldsymbol{B}_{ki}^{\mathrm{T}} \\ \boldsymbol{B}_{ki} & \boldsymbol{a}_{ki}\end{pmatrix} = \begin{pmatrix}\boldsymbol{C}_{ki} & \boldsymbol{b}_{ki}\end{pmatrix},$$

kjer indeks k teče preko vseh komponent modela GMM ($k \in \{1, \ldots, K\}$), indeks i pa preko vseh dimenzij akustičnega prostora ($i \in \{1, \ldots, D\}$).

Matriko $\boldsymbol{\Sigma}$ določimo po naslednji formuli:

$$\boldsymbol{\Sigma} = \boldsymbol{N}^{-1}\left(\sum_s \boldsymbol{S}_s - \operatorname{diag}\left(\boldsymbol{C}\boldsymbol{W}^{\mathrm{T}} + \boldsymbol{B}\boldsymbol{D}\right)\right).$$

Izkaže se, da je najbolj kompliciran del postopka izračun posteriornih porazdelitev vezanih faktorjev $\boldsymbol{x}$, $\boldsymbol{y}$ in $\boldsymbol{z}$. Učinkovita implementacija tega izračuna predstavlja zaradi velikanske razsežnosti vektorja $\boldsymbol{w}$ in matrike $\boldsymbol{W}$ poseben izziv. Razsežnost vektorja $\boldsymbol{w}$ je namreč enaka $HD_u + D_v + KD$, kjer je H število posnetkov določenega govorca, K število komponent modela UBM, D dimenzija akustičnega prostora, D_u dimenzija kanalskega podprostora in D_v dimenzija govorskega podprostora. Tako je pri standardni konfiguraciji modela UBM in desetih posnetkih razsežnost precizijske matrike $\boldsymbol{L}$ kar 124180×124180. Čeprav je kovariančna matrika priorja redka, to ne velja za kovariančno matriko posteriorja (inverz redke matrike namreč v splošnem ni redka matrika)[32]. Zato je bilo potrebno domisliti način, kako narediti postopek praktično izvedljiv. Celotno izvajanje (glej Kenny (2005), razdelek III.C) je precej zapleteno in ga na tem mestu ne bomo povzemali, omenimo le, da si lahko delno pomagamo z Woodburyjevo enakostjo. Izkaže se, da bi bilo veliko lažje, če bi lahko člen $\boldsymbol{D}\boldsymbol{z}$ v celoti izpustili, saj ostane v tem primeru kovariančna matrika posteriorne porazdelitve redka. To bi nam omogočilo, da bi lahko uporabili veliko učinkovitejši postopek za izračun posteriorne porazdelitve (glej Kenny (2005), razdelek III.D).

7.4.5 Ločeno ocenjevanje hiperparametrov po kriteriju največjega verjetja

Ocenjevanje parametrov po kriteriju ML ima poleg težavne implementacije še druge slabosti. Posledica uporabe kriterija ML je namreč, da bo model kar se da natančno opisal učne podatke. To pa ni vedno nujno ugodno, saj naš primarni cilj ni, da bi čim natančneje opisali podatke, ampak le, da bi celotni prostor supervektorjev čimbolje

[32] Povedano drugače, naključne spremenljivke, ki so pod priorjem neodvisne — in zato tudi nekorelirane — postanejo pod posteriorjem odvisne (statistična neodvisnost preide v pogojno odvisnost). Ta pojav je znan pod imenom pojasnjevanje (angl. explaining away) in se nanaša na dejstvo, da če obstaja več možnih vzrokov za neko posledico, potem opaženje enega izmed vzrokov poveča naše vedenje tudi o ostalih vzrokih.

ločili na govorski in sejni oz. kanalski podprostor. Žal se izkaže, da tega znotraj predstavljenega ML okvira ne moremo zagotoviti. V primeru premajhnega ranga matrike $\boldsymbol{V}$ se namreč lahko zgodi, da bo del medgovorske spremenljivosti pokrila tudi matrika $\boldsymbol{U}$, ki naj bi po predpostavki pokrivala kanalsko spremenljivost. Še večja težava nastane, ker sočasno ocenjevanje vseh parametrov vodi do premajhne ocene za diagonalno matriko $\boldsymbol{D}$, kar povzroči, da igra matrika $\boldsymbol{D}$ v modelu JFA zgolj obrobno vlogo. Vzrok za premajhno oceno se skriva v dejstvu, da imata matriki $\boldsymbol{U}$ in $\boldsymbol{V}$ veliko večje število parametrov kot matrika $\boldsymbol{D}$ in zaradi omejene učne množice ›pobereta‹ večino spremenljivosti, ki je prisotna v podatkih. V izogib tej težavi je (Kenny et al., 2008) predlagal postopek *ločenega* (angl. decoupled) ocenjevanja hiperparametrov. Celotnega postopka na tem mestu ne bomo povzemali, nakažimo le osnovno idejo predlaganega postopka. V grobem je ideja ta, da učno množico, ki jo uporabljamo za ocenitev hiperparametrov, razdelimo na več delov in vsak del nato uporabimo za ocenjevanje le določenega hiperparametra. Enačbe se pri ločenem ocenjevanju hiperparametrov nekoliko poenostavijo, še bolj pa se poenostavi implementacija, saj postanejo razsežnosti matrik, ki nastopajo v postopku, obvladljive.

7.4.6 Ocenjevanje hiperparametrov po kriteriju najmanjše divergence

Kadar sta orientaciji govorskega in kanalskega podprostora znani, lahko namesto kriterija ML za oceno hiperparametrov uporabimo kriterij najmanjše divergence. Tak način ocenjevanja hiperparametrov je primeren v primeru, ko smo že pridobili ocene hiperparametrov po kriteriju ML, sedaj pa želimo te ocene prilagoditi novi populaciji govorcev (Kenny (2005), razdelek V) ali pa v primeru, ko želimo pridobiti ocene hiperparametrov, ki bi bile odvisne od govorca (Kenny (2005), razdelek VI). (Takšna strategija pride v upoštev npr. pri NIST-ovi nalogi nenadzorovane adaptacije, kjer lahko za učenje govorskega modela uporabimo vse posnetke, za katere smo ugotovili, da pripadajo temu govorcu (Nist, 2008).)

Pri enkratnem prehodu skozi govorno zbirko izluščimo naslednjo statistiko:

$$\boldsymbol{\mu}_y = \frac{1}{S}\sum_s \mathbb{E}\left[\boldsymbol{y}_s\right]$$

$$\boldsymbol{\mu}_z = \frac{1}{S}\sum_s \mathbb{E}\left[\boldsymbol{z}_s\right]$$

$$\boldsymbol{K}_{xx} = \frac{1}{H}\sum_s\sum_h \mathbb{E}\left[\boldsymbol{x}_{sh}\boldsymbol{x}_{sh}^{\mathrm{T}}\right]$$

$$\boldsymbol{K}_{xx} = \frac{1}{S}\sum_s \mathbb{E}\left[\boldsymbol{y}_s\boldsymbol{y}_s^{\mathrm{T}}\right] - \boldsymbol{\mu}_y\boldsymbol{\mu}_y^{\mathrm{T}}$$

$$\boldsymbol{K}_{zz} = \mathrm{diag}\left(\frac{1}{S}\sum_s \mathbb{E}\left[\boldsymbol{z}_s\boldsymbol{z}_s^{\mathrm{T}}\right] - \boldsymbol{\mu}_z\boldsymbol{\mu}_z^{\mathrm{T}}\right)$$

$$\boldsymbol{N} = \sum_s \sum_h \boldsymbol{N}_{sh}$$
$$H = \sum_s H_s,$$

kjer je S število govorcev, H_s pa število posnetkov s-tega govorca. Matematična upanja so izračunana pri trenutnih vrednostih hiperparametrov. Nove vrednosti hiperparametrov določimo po naslednjih enačbah:

$$\begin{aligned}
\boldsymbol{\mu} &= \boldsymbol{\mu}_0 + \boldsymbol{V}_0\boldsymbol{\mu}_y + \boldsymbol{D}_0\boldsymbol{\mu}_z \\
\boldsymbol{U} &= \boldsymbol{U}_0\boldsymbol{K}_{xx}^{1/2} \\
\boldsymbol{V} &= \boldsymbol{V}_0\boldsymbol{K}_{yy}^{1/2} \\
\boldsymbol{D} &= \boldsymbol{D}_0\boldsymbol{K}_{zz}^{1/2} \\
\boldsymbol{\Sigma} &= \boldsymbol{N}^{-1} \sum_s \sum_h \Big(\boldsymbol{S}_{sh} - 2\,\mathrm{diag}\left(\boldsymbol{F}_{sh}\,\mathbb{E}\left[\boldsymbol{m}_{sh} - \boldsymbol{m}_0\right]\right) \\
&\qquad + \mathrm{diag}\left(\mathbb{E}\left[(\boldsymbol{m}_{sh} - \boldsymbol{m}_0)(\boldsymbol{m}_{sh} - \boldsymbol{m}_0)^{\mathrm{T}}\right] \boldsymbol{N}_{hs}\right)\Big)
\end{aligned}$$

Čeprav je postopek ocenjevanja hiperparametrov po kriteriju najmanjše divergence po svoje delno omejen (omogoča le rotacijo lastnih vektorjev in skaliranje pripadajočih lastnih vrednosti), se ga splača kombinirati s kriterijem ML, saj se izkaže, da ocenjevanje hiperparametrov po kriteriju ML vodi do ocen lastnih vrednosti, ki jih je težko interpretirati (Kenny, 2005).

7.4.7 Odločitveni kriterij na osnovi funkcije verjetja

Pokazali smo, kako na podlagi primerne govorne zbirke ocenimo hiperparametre, ki nastopajo v modelu JFA, nismo pa še odgovorili na vprašanje, kako model JFA uporabimo za razpoznavanje govorcev.

Odločitveni kriterij lahko skonstruiramo na zelo podoben način kot pri klasičnem postopku razpoznavanja govorcev. Učni govorni posnetek[33] o_1 uporabimo za oceno govorskega supervektorja $\boldsymbol{s}$. V ta namen izračunamo povprečno vrednost $\mathbb{E}[\boldsymbol{w}|o_1]$ posteriorne porazdelitve $p(\boldsymbol{w}|o_1)$, na podlagi katere izračunamo govorski supervektor v skladu z enačbo (7.1). Testni govorni posnetek o_2 uporabimo tako, da izračunamo razmerje verjetij med modelom, ki je določen z govorskim supervektorjem $\boldsymbol{s}$ in modelom, ki je določen s povprečnim (UBM) supervektorjem $\boldsymbol{\mu}$:

$$\frac{p(o_2|\boldsymbol{s})}{p(o_2|\boldsymbol{\mu})}.$$

[33] V splošnem je lahko učnih posnetkov več. V terminologiji, ki jo uporablja NIST, testu, pri katerem je na voljo več učnih posnetkov (navadno 3, 8 ali 16), v angleščini pravimo ›extended data condition‹.

Na ta način smo vpliv kanala odstranili le iz učnega posnetka. Ker želimo, da bo rezultat razpoznavanja čim bolj neodvisen tudi od testnega kanala, moramo vrednost verjetja izračunati nekoliko drugače. Najprej iz testnega posnetka ocenimo testni kanal $\boldsymbol{c}$. To storimo tako, da najprej izračunamo pričakovano vrednost posteriorne porazdelitve $p(\boldsymbol{x}|o_2)$, iz katere določimo kanal $\boldsymbol{c}$ po enačbi (7.2). Ta kanal nato upoštevamo tako, da izračunamo razmerje verjetij na sledeč način:

$$\frac{p(o_2|\boldsymbol{s},\boldsymbol{c})}{p(o_2|\boldsymbol{\mu},\boldsymbol{c})}.$$

Opisani način, kako pri izračunu verjetja upoštevati tako razmere v učnem kot tudi razmere v testnem posnetku, je zelo podoben tistemu, ki smo ga predlagali pri postopku NAP (glej razdelek 6.8). Sedaj, ko obravnavamo supervektorje kot naključne spremenljivke, pa imamo še drugačno možnost. Namesto da bi testni posnetek uporabili za izračun točkovne ocene kanala in izračunali verjetje le pri tej vrednosti kanala, lahko verjetje izračunamo s seštevanjem preko vseh možnih konfiguracij kanala:

$$p(o_2|\boldsymbol{s}) = \int p(o_2,\boldsymbol{x}|\boldsymbol{s})\mathrm{d}\boldsymbol{x} = \int p(o_2|\boldsymbol{s},\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \tag{7.7}$$

kjer je $p(\boldsymbol{x})$ po predpostavki normalna apriorna porazdelitev naključne spremenljivke $\boldsymbol{x}$. Za izračun gornjega integrala lahko uporabimo rezultat (7.6), vendar moramo poprej usrediščiti zadostno statistiko na sledeč način:

$$\widetilde{\boldsymbol{F}} = \boldsymbol{F} - \boldsymbol{N}(\boldsymbol{s}-\boldsymbol{\mu}) \tag{7.8}$$

$$\widetilde{\boldsymbol{S}} = \boldsymbol{S} - 2\,\mathrm{diag}\left(\boldsymbol{F}(\boldsymbol{s}-\boldsymbol{\mu})^{\mathrm{T}}\right) + \mathrm{diag}\left(\boldsymbol{N}(\boldsymbol{s}-\boldsymbol{\mu})(\boldsymbol{s}-\boldsymbol{\mu})^{\mathrm{T}}\right). \tag{7.9}$$

Rezultat integracije (7.7) lahko v tem primeru zapišemo kot:

$$\log p(o_2|\boldsymbol{s}) = \widetilde{G} - \frac{1}{2}\log|\boldsymbol{L}| + \frac{1}{2}\left|\left|\boldsymbol{L}^{-1/2}\boldsymbol{W}^{T}\boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{F}}\right|\right|^2, \tag{7.10}$$

kjer je matrika $\boldsymbol{L} = \boldsymbol{I} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{U}$. Z oznako $\widetilde{G}$ smo označili člen G, v katerem smo namesto statistike $\boldsymbol{S}$ uporabili usrediščeno statistiko $\widetilde{\boldsymbol{S}}$.

Če upoštevamo, da supervektor $\boldsymbol{s}$ ni podan le s točkovno oceno ampak z aposteriorno porazdelitvijo, ki izraža nezanesljivost oz. nedoločenost ocene, lahko to nedoločenost upoštevamo pri središčenju statistike tako, da $\widetilde{\boldsymbol{F}}$ in $\widetilde{\boldsymbol{S}}$ nadomestimo z njunima aposteriornima pričakovanima vrednostima $\mathbb{E}[\widetilde{\boldsymbol{F}}]$ in $\mathbb{E}[\widetilde{\boldsymbol{S}}]$:

$$\begin{aligned}
\mathbb{E}[\widetilde{\boldsymbol{F}}] &= \boldsymbol{F} - \boldsymbol{N}(\mathbb{E}[\boldsymbol{s}] - \boldsymbol{\mu}) \\
\mathbb{E}[\widetilde{\boldsymbol{S}}] &= \boldsymbol{S} - 2\,\mathrm{diag}\left(\boldsymbol{F}(\mathbb{E}[\boldsymbol{s}] - \boldsymbol{\mu})^{\mathrm{T}}\right) \\
&\quad + \mathrm{diag}\left(\boldsymbol{N}\left((\mathbb{E}[\boldsymbol{s}] - \boldsymbol{\mu})(\mathbb{E}[\boldsymbol{s}] - \boldsymbol{\mu})^{\mathrm{T}} + \mathrm{Cov}\,(\boldsymbol{s},\boldsymbol{s})\right)\right).
\end{aligned}$$

Nedoločenost ocene govorskega vektorja lahko na podoben način upoštevamo tudi pri izračunu verjetja $p(o|\boldsymbol{\mu})$, le da se zdi v tem primeru za povprečno vrednost in kovarianco bolje uporabiti

$$\mathbb{E}[\boldsymbol{s}] = \boldsymbol{\mu}$$
$$\mathrm{Cov}(\boldsymbol{s}, \boldsymbol{s}) = \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} + \boldsymbol{D}\boldsymbol{D}^{\mathrm{T}},$$

kar ustreza apriorni porazdelitvi govorskega faktorja $\boldsymbol{s}$. (Glej Kenny et al. (2007a) za podrobnosti.)

Postopek JFA nam nudi še drugačen način računanja razmerja verjetij, ki temelji na konceptu, ki mu v statistični literaturi pravijo *primerjanje modelov* (angl. model comparison). Predpostavimo, da imamo na voljo vsaj en učni posnetek govorca s in en testni posnetek neznanega govorca. Postavimo lahko dve hipotezi:

i) govorec v testnem posnetku *je* govorec s;
ii) govorec v testnem posnetku *ni* govorec s.

Zanima nas, katera izmed teh dveh hipotez je bolj verjetna. Kot odgovor na to vprašanje nam (Kenny (2005), razdelek III) ponudi odgovor v obliki dveh načinov izračuna razmerja verjetij. *Paketno* (angl. batch) razmerje verjetij je najlaže razložiti v primeru, ko imamo samo en učni posnetek (t.i. problem primerjave govorcev), čeprav je posplošitev na poljubno število učnih posnetkov enostavna. Če predpostavimo, da je resnična hipoteza (i) — govorec v učnem in testnem posnetku je isti — potem lahko verjetje para posnetkov izračunamo tako, da izračunamo totalno verjetje (**7.6**), pri čemer postavimo $H_s = 2$. Na drugi strani, ker pravilnost hipoteze (ii) implicira, da sta posnetka med seboj statistično neodvisna, lahko verjetje v tem primeru izračunamo kot produkt totalnih verjetij obeh posnetkov,[34] le da sedaj pri tem postavimo $H_s = 1$. Paketno razmerje verjetij definiramo kot kvocient obeh verjetij:

$$\frac{p(o_1, o_2)}{p(o_1)\, p(o_2)}.$$

Večje kot bo razmerje, bolj bomo prepričani o resničnosti hipoteze (i) in obratno, manjše kot bo razmerje, bolj bomo prepričani o resničnosti hipoteze (ii).

Opazimo lahko, da se paketno razmerje verjetij razlikuje od ostalih načinov izračuna razmerja verjetij, saj se v nasprotju z drugimi, kjer striktno ločimo učni in testni del postopka, izvede v enem koraku. Vse kar je potrebno storiti je ovrednotiti integral (**7.6**). Izkaže se, da je izračun tega integrala, ki ga moramo izvesti za vsak par posnetkov posebej, časovno zelo potraten, zaradi česar je njegova praktična uporabnost vprašljiva. Omenimo naj, da je kriterij paketnega razmerja verjetij zelo

[34] Formalno gledano ne bi smeli govoriti o verjetju posnetka, marveč o verjetju modela ali o verjetju parametrov, saj je verjetje funkcija parametrov in ne funkcija podatkov (MacKay (2003), str. 28–29).

podoben bayesovskemu informacijskemu kriteriju (angl. Bayesian information criterion, BIC), ki je popularen na področju segmentacije zvočnih posnetkov in rojenja govorcev (Tranter in Reynolds, 2006).

Časovno bolj učinkovit je *zaporedni* (angl. sequential) način izračuna razmerja verjetij, ki je bolj podoben klasičnemu načinu. Pri tem načinu uporabimo učni posnetek znanega govorca s za oceno od govorca odvisnih hiperparametrov (Kenny (2005), razdelek II.E) modela JFA, s katerimi opišemo posteriorno porazdelitev govorskega supervektorja danega govorca. To nam omogoči, da izračunamo verjetje na dva načina; prvič izračunamo verjetje od govorca odvisnih parametrov, drugič pa verjetje od govorca neodvisnih parametrov. Razmerje obeh tako definiranih verjetij nam lahko služi za še en način preverjanja resničnosti postavljenih hipotez (i) in (ii).

7.4.8 Odločitveni kriterij na osnovi metode podpornih vektorjev

Poleg generativnega razmerja verjetij je mero podobnosti mogoče zasnovati tudi na diskriminatorni metodi podpornih vektorjev. Pri razmerju verjetij smo posteriorno porazdelitev govorskih faktorjev izračunali le za učni posnetek, tukaj pa enako ponovimo še za testni posnetek. Pri tem obdržimo le povprečno vrednost porazdelitve (supervektor), kovariančno matriko pa zavržemo. Učni supervektor uporabimo za učenje modela SVM, testnega pa prilegamo na ta model. Oba supervektorja poprej še ustrezno normiramo.

7.5 Povezava med postopkom JFA in MAP

Pokazali bomo, da lahko postopek MAP vidimo kot poseben primer modela JFA. Faktorska enačba (**7.3**) določa obliko porazdelitve, ki jo na nek način vsilimo supervektorjem $\boldsymbol{m}$. Naša predpostavka je, da se naključna spremenljivka $\boldsymbol{m}$ porazdeljuje normalno, pri čemer kovariančna matrika te porazdelitve ni poljubna, ampak o njeni strukturi odločajo hiperparametri $\boldsymbol{U}$, $\boldsymbol{V}$ in $\boldsymbol{D}$.

Vprašamo se lahko, kaj se zgodi, če postavimo $\boldsymbol{U} = \boldsymbol{0}$ in $\boldsymbol{V} = \boldsymbol{0}$. Tedaj se faktorska enačba poenostavi v: $\boldsymbol{m} = \boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{z}$. Ker je to poseben primer splošnejšega modela, ostanejo vse izpeljave veljavne tudi v tem primeru, le da lahko v rezultatih simbol $\boldsymbol{W}$ zamenjamo s simbolom $\boldsymbol{D}$. Na prvi pogled se zdi, da s tem nismo nič pridobili. A se izkaže, da ima ta sprememba dramatične posledice za časovno kompleksnost izračuna posteriorne porazdelitve, saj postane v tem primeru precizijska matrika $\boldsymbol{L}$ diagonalna.

V razdelku **7.4.2** smo pokazali, da je posteriorna porazdelitev prikrite spremenljivke $\boldsymbol{w}$ normalna. Če v izrazu za povprečno vrednost te spremenljivke zamenjamo simbol $\boldsymbol{w}$ s simbolom $\boldsymbol{z}$ in simbol $\boldsymbol{W}$ s simbolom $\boldsymbol{D}$, dobimo:

$$\mathbb{E}[\boldsymbol{z}|o] = \boldsymbol{L}^{-1}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{F},$$

kjer je $\boldsymbol{L}$ podan z izrazom $\boldsymbol{L} = \boldsymbol{I} + \boldsymbol{D}\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{D}$. Ker vemo, da povprečna vrednost naključne spremenljivke, ki je porazdeljena po Gaussu, sovpada z modusom Gaussove porazdelitve, vidimo, da je gornji rezultat enak tistemu, ki ga dobimo pri

klasični adaptaciji MAP (Gauvain in Lee, 1994). Prednost obravnavanega pristopa pred adaptacijo MAP je v tem, da MAP vrne le točkovno oceno (angl. point estimate), mi pa izračunamo celotno posteriorno porazdelitev. (Ta razlika je posebej pomembna v primeru, ko imamo na voljo le majhno količino podatkov. Pri večanju količine podatkov postaja vrh aposteriorne porazdelitve vedno bolj izrazit in v limiti degenerira v točko.)

Povprečno vrednost naključne spremenljivke $\boldsymbol{m}$ oz. oceno MAP lahko zapišemo z izrazom:

$$\boldsymbol{m}_{\mathrm{MAP}} = \boldsymbol{\mu} + \boldsymbol{D}\mathbb{E}[\boldsymbol{z}|o].$$

Vrednost hiperparametra $\boldsymbol{D}$ lahko ocenimo po kriteriju ML na enak način, kot smo to storili v splošnejšem modelu JFA, s to razliko, da sta sedaj implementacija in računska zahtevnost postopka veliko enostavnejša. Po drugi strani pa lahko vrednost matrike $\boldsymbol{D}$ izberemo tudi ad hoc.

Če postavimo $\boldsymbol{D} = \tau^{-1/2}\boldsymbol{\Sigma}^{1/2}$, kjer je $\boldsymbol{\Sigma}^{1/2}$ spodnje trikotna matrika, potem se izraz za povprečno vrednost naključne spremenljivke $\boldsymbol{m}$ poenostavi v:

$$\boldsymbol{m}_{\mathrm{MAP}} = \boldsymbol{L}^{-1}\boldsymbol{F},$$

kjer je $\boldsymbol{L} = \boldsymbol{I} + \tau^{-1}\boldsymbol{N}$. Če namesto usrediščene statistike $\boldsymbol{F}$ uporabimo neusrediščeno statistiko $\widetilde{\boldsymbol{F}} = \boldsymbol{F} + \boldsymbol{N}\boldsymbol{\mu}$, lahko gornjo enačbo preoblikujemo v sledeč zapis:

$$\boldsymbol{m}_{\mathrm{MAP}} = (\boldsymbol{I} - \boldsymbol{L}^{-1}\boldsymbol{N})\boldsymbol{\mu} + \boldsymbol{L}^{-1}\boldsymbol{N}\boldsymbol{m}_{\mathrm{ML}}.$$

Vrednost $\boldsymbol{m}_{\mathrm{ML}}$ je podana z izrazom $\boldsymbol{m}_{\mathrm{ML}} = \boldsymbol{N}^{-1}\widetilde{\boldsymbol{F}}$ in pomeni oceno, ki bi jo dobili, če bi za ocenjevanje uporabili kriterij ML. Gornja enačba ima zanimivo interpretacijo. Vidimo, da je ocena $\boldsymbol{M}_{\mathrm{MAP}}$ dobljena kot utežena vsota ocene ML in apriorne vrednosti $\boldsymbol{\mu}$. Več kot bomo imeli na voljo podatkov, večji bo $\boldsymbol{N}$ in bolj bo prevladala ocena ML. Velja tudi obratno, manj kot bo na voljo podatkov, manj bomo ›verjeli‹ oceni ML. Na tak način se izognemo problemu singularnosti. Ker je matrika $\boldsymbol{D}$ (bločno) diagonalna, bo adaptacija MAP zmogla oceniti povprečne vrednosti le tistih komponent modela GMM, ki bodo ›videne‹ v podatkih.

Če smo bili pri branju pozorni, se lahko spomnimo, da smo do podobne ugotovitve prišli že v razdelku **3.5**, ko smo adaptacijo MAP predstavili na nekoliko bolj inženirski način. Sedaj smo do adaptacije MAP prišli po drugi poti. Najprej smo zapisali predpostavke, na podlagi katerih smo izpeljali splošnejši postopek JFA, katerega poenostavitev je vodila do postopka MAP. Videli smo, da vrača adaptacija MAP kot rezultat točkasto oceno, medtem ko postopek JFA vrne celotno posteriorno porazdelitev.

7.6 Eigenvoice MAP

Podobno kot smo se vprašali, kaj se zgodi, če v faktorski enačbi obdržimo le hiperparameter $\boldsymbol{D}$, se lahko vprašamo, kaj dobimo, če postavimo $\boldsymbol{U} = \boldsymbol{0}$ in $\boldsymbol{D} = \boldsymbol{0}$ in obdržimo le hiperparameter $\boldsymbol{V}$. Izkaže se, da v tem primeru dobimo *eigenvoice*

MAP (EMAP). Njegova glavna prednost pred klasično adaptacijo MAP je v tem, da je dosti bolj učinkovit v primeru majhne količine podatkov, ki jih imamo na voljo za oceno posteriorne porazdelitve. Razlog za večjo učinkovitost se skriva v dejstvu, da EMAP predvideva koreliranost posameznih komponent modela GMM, medtem ko navadni MAP predpostavlja, da so komponente med seboj nekorelirane. Po drugi strani pa je slabost adaptacije EMAP ta, da EMAP ocena pri neomejeni količine podatkov ni identična oceni ML. (EMAP je namreč omejen na podprostor, ki ga oklepajo lastni vektorji matrike $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}$.) Tukaj je v prednosti navadni MAP, saj se izkaže, da je le-ta asimptotično ekvivalenten postopku ML. Vidimo da se navadni in eigenvoice MAP dopolnjujeta, zato ju je smiselno združiti, kar je storjeno v modelu JFA.

7.7 Komentar

Predstavili smo postopek analize vezanih faktorjev, s katerim modeliramo govorsko in sejno spremenljivost v prostoru parametrov (povprečnih vektorjev) modela GMM. Osnovna predpostavka modela JFA je, da je mogoče sejni supervektor razstaviti na vsoto govorskega in kanalskega supervektorja, pri čemer je kanalski supervektor omejen na nižjerazsežni podprostor. Postopek je s teoretičnega vidika vse prej kot enostaven, pa tudi implementacija je relativno zahtevna.

Pokazali smo, da je postopek JFA v tesni povezavi s postopkom faktorske analize, iz katere izhaja predpostavka, da se supervektorji porazdeljujejo normalno. V upravičenost te predpostavke je podvomil (Kenny et al., 2006), ki je predpostavko omilil z uporabo mešanice faktorjev (angl. mixture of factor analyzers, MFA) (Ghahramani in Hinton, 1996). Eksperimenti so pokazali, da je predpostavka o normalnosti več ali manj upravičena, vendar bi za zanesljivejši odgovor nujno potrebovali obsežnejše govorne zbirke, ki jih trenutno še ni na voljo.

V navezi s postopkom JFA smo predstavili nekaj možnih mer podobnosti, ki temeljijo na funkciji verjetja in jih lahko uporabimo kot odločitveni kriterij pri razpoznavanju govorcev. Poleg tega smo predlagali tudi odločitveni kriterij, ki temelji na metodi podpornih vektorjev.

EKSPERIMENTI IN REZULTATI 8

8.1 Govorne zbirke

Ko razvijemo sistem za biometrično razpoznavanje, ga moramo tudi preizkusiti na realnih podatkih. Če je le mogoče, izberemo takšno podatkovno zbirko, ki je med raziskovalci že uveljavljena. Na tak način dosežemo, da bomo učinkovitost našega sistema lahko pošteno primerjali z učinkovitostjo vseh drugih sistemov, za katere lahko najdemo poročila o rezultatih razpoznavanja na tej isti zbirki.

Pri eksperimentalnem delu smo največ uporabljali govorne zbirke, ki jih je mogoče kupiti pri ameriškemu združenju LDC (Linguistic Data Consortium)[35]. To združenje se ukvarja s pridobivanjem in trženjem različnih podatkovnih virov, ki so primarno namenjeni raziskavam na področju govornih tehnologij.

Do uporabe govornih zbirk, ki so primerne za raziskave na področju razpoznavanja govorcev, so upravičeni tudi vsi udeleženci tekmovanj oz. vrednotenj sistemov za razpoznavaje govorcev (angl. speaker recognition evaluation, SRE), ki jih običajno enkrat letno priredi ameriški nacionalni inštitut za standarde in tehnologijo — NIST (National Institute for Standards and Technology).

Pred pričetkom vsakokratne evaluacije NIST predpiše poseben protokol, ki se ga morajo držati vse (raziskovalne) skupine, ki želijo sodelovati na tekmovanju. V tem protokolu je definiranih več različnih nalog, ki se med seboj v glavnem razlikujejo od količine govornega materiala, ki je na voljo v učnem in testnem delu poskusa. Udeleženci ne potrebujejo reševati vseh nalog, obvezno je rešiti le t.i. glavno nalogo, v kateri je v učnem in testnem delu vsakega poskusa (angl. trial) na voljo približno 5 minut dolg telefonski posnetek, v katerem pa je lahko prisoten precejšen del negovornih segmentov. Udeležencem sta za vsak posnetek na voljo le podatka o spolu in jeziku govorca, prepovedano pa je kakršnokoli ročno analiziranje posnetkov (npr. poslušanje, vizualni pregled). Polega tega lahko udeleženci, če želijo, uporabijo besedne prepise posnetkov, ki jih NIST pridobi s sistemom za samodejno razpoznavanje besed. Zavedati se je treba, da prepisi niso brez napak — napaka besed (angl. word error rate, WER) naj bi znašala med 15 in 30 %.

Pomembno a pogosto spregledano pravilo uradnega protokola je, da pri obravnavi danega poskusa, ki se nanaša na primerjavo enega (ali večih, odvisno od naloge) učnega in enega testnega posnetka, ni dovoljeno uporabiti nobene informacije, ki bi jo pridobili iz ostalih učnih ali testnih posnetkov. Vsak poskus je treba analizirati torej neodvisno od vseh ostalih poskusov. Izkaže se, da upoštevanje tega pravila znatno poveča zahtevnost zadane naloge.

Izjema, kjer je to pravilo malenkost spremenjeno, je naloga, poimenovana *nenadzorovana adaptacija* (angl. unsupervised adaptation). Pri tej nalogi je testne

[35] Domača stran združenja LDC se nahaja na http://www.ldc.upenn.edu.

posnetke, za katere ugotovimo, da pripadajo govorcu iz učnega posnetka, dovoljeno uporabiti za ›doučevanje‹ modela govorca.

Večina govorcev, ki so sodelovali pri snemanju teh zbirk, je angleško govorečih, čeprav nekaterim angleščina ne predstavlja materinega jezika. Določen delež posnetkov je posnet v drugih jezikih (kitajščina, arabščina, španščina itd.). Za izčrpnejši opis zbirk in natančnejši opis NIST-ovega protokola[36] svetujemo ogled ustrezne literature (Leeuwen et al., 2006; Przybocki et al., 2007; Cieri et al., 2007; Brandschain et al., 2008; Martin in Greenberg, 2009).

Glavna težava, s katero se v zadnjem času raziskovalci največ ukvarjajo, je v neenakih akustičnih razmerah v učnem in testnem delu poskusa. Izkazalo se je namreč, da uporaba različnih tipov telefonskih aparatov, mikrofonov in telefonskih linij znatno poslabša rezultate razpoznavanja govorcev. Ravno zato so naloge v okviru vrednotenj zastavljene tako, da omenjeni problem še dodatno izpostavijo in s tem raziskovalce ›prisilijo‹, da se temu problemu še posebej posvetijo. Poleg omenjenega problema neenakih razmer se skuša najti odgovore tudi na nekatera druga vprašanja, npr. kako je razpoznavanje govorcev odvisno od jezika, kako je uspešnost razpoznavanja odvisna od količine govornega materiala, ki ga imamo na voljo ipd.

Na zadnji evaluaciji (2008) so poleg telefonskih posnetkov v glavno nalogo vključili tudi znatno število mikrofonskih posnetkov telefonskih pogovorov in mikrofonskih posnetkov intervjujev. Vsak mikrofonski posnetek je posnet z enim izmed večih različnih tipov mikrofonov, a podatek, za kateri tip mikrofona gre v konkretnem primeru, udeležencem vrednotenj ni na voljo.

V naših poskusih smo uporabljali podatke (glej tabelo **8.1**), ki so služili vrednotenju razpoznavalnikov govorcev v letih 2004, 2005, 2006 in 2008. Od tega smo podatke iz let 2004-2006 uporabljali kot razvojno množico (angl. development set), medtem ko smo podatke iz leta 2008 uporabljali kot testno množico. Tukaj je vredno poudariti, da se govorci, ki so prisotni v zbirki za leto 2008, ne pojavijo v nobenem posnetku iz zbirk prejšnjih let.

zbirka	št. govorcev[37]			št. posnetkov			št. ur in minut[38]		
	Ž	skupaj	M	Ž	skupaj	M	Ž	skupaj	M
2004	184	306	122	2603	4455	1852	84:20	143:58	59:38
2005	278	461	183	3392	5758	2366	117:24	200:15	82:51
2006	320	563	243	3440	6039	2599	114:11	200:26	86:15
2008	844	1336	492	2818	4259	1541	94:16	143:26	49:10
skupaj	1532	2514	982	12253	20511	8358	410:11	688:5	277:54

Tabela 8.1 Podatki o uporabljenih govornih zbirkah.

[36] Protokol za leto 2008 je dostopen na naslovu http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan/release4.pdf.

[37] Nekateri govorci so prisotni v večih zbirkah, zato skupna številka vseh različnih govorcev ni enaka vsoti različnih govorcev po posameznih zbirkah.

[38] Podatek se nanaša na skupno dolžino govornih odsekov po izločitvi negovornih odsekov.

Z vključitvijo mikrofonskih posnetkov se je kompleksnost izvedbe glavne naloge za leto 2008 še dodatno povečala, zato smo se odločili, da se bomo v eksperimentih, katerih rezultate bomo podali v pričujoči disertaciji, omejili le na del naloge, ki se nanaša na telefonske posnetke. Skupaj gre za 37001 poskusov (angl. trials), od tega je 24128 ženskih poskusov in 12873 moških poskusov, kar je skoraj 2 proti 1 v prid ženskam. Od 37001 poskusov jih je 3826 takšnih, kjer je govorec v učnem in testnem posnetku isti, 33175 poskusov pa takšnih, kjer sta govorca v učnem in testnem posnetku različna.

8.2 Eksperimenti in rezultati

Eksperimente, ki smo jih opravili v okviru disertacije, lahko razdelimo v štiri sklope:

- Referenčni sistem (GMM-UBM)
- Sistem NAP
- Sistem JFA
- Združevanje rezultatov razpoznavanja

Model UBM

Izhodišče vseh sistemov je model UBM, ki smo ga ocenili na veliki količini podatkov (glej razdelek **3.4**). Za parametrizacijo govornih signalov smo uporabili značilke MFCC. Vsakih 10 ms smo iz 25 ms dolgega oknjenega izseka signala izračunali 20 koeficientov MFCC (ničti koeficient smo zamenjali z vrednostjo logaritma energije), pri čemer smo uporabili 24 filtrov, ki smo jih razporedili po melodični skali na frekvenčnem območju od 300 do 3400 Hz. Statične značilke smo skupaj s pripadajočimi regresijskimi koeficienti prvega in drugega reda zbrali v 60-razsežne vektorje značilk.

Model UBM smo začeli graditi tako, da smo na celotnem materialu najprej izračunali globalni povprečni vektor in diagonalno kovariančno matriko. Tako smo dobili enokomponentni model GMM. Nato smo iz enokomponentnega tvorili dvokomponentni model GMM tako, da smo začetne vrednosti uteži (π), povprečij ($\boldsymbol{\mu}$) in kovariančnih matrik ($\boldsymbol{\Sigma}$) obeh komponent določili po naslednjem pravilu:

$$\pi_{11} = \frac{1}{2}\pi_1 \qquad \boldsymbol{\mu}_{11} = \boldsymbol{\mu}_1 - \frac{1}{5}\boldsymbol{\Sigma}_1^{1/2} \qquad \boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_1$$
$$\pi_{12} = \frac{1}{2}\pi_1 \qquad \boldsymbol{\mu}_{12} = \boldsymbol{\mu}_1 + \frac{1}{5}\boldsymbol{\Sigma}_1^{1/2} \qquad \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_1$$

Te vrednosti smo nato ›doučili‹ z 10 iteracijami postopka EM. Nato smo na zgoraj opisan način nadaljevali s podvojevanjem[39] števila komponent modela GMM. Ustavili smo se, ko je število komponent doseglo vrednost 2048. Takrat smo namesto 10

[39] Tovrstno podvojevanje je znano tudi pri postopku vektorske kvantizacije, kjer mu pravijo postopek LBG (po začetnicah avtorjev (Linde et al., 1980)).

izvedli 20 iteracij postopka EM. Tako naučen model GMM smo uporabili pri vseh eksperimentih, ki bodo opisani v nadaljevanju.

Učenje modela UBM je časovno najbolj zamuden del eksperimenta. Da bi pohitrili učenje, smo izdelali lastno rešitev za porazdeljeno računanje zadostne statistike. Kljub temu, da smo vzporedno uporabljali več kot 10 osebnih računalnikov, je celotno učenje trajalo nekaj tednov. Eksperimentiranje z različnimi nastavitvami pri učenju modela UBM je zato skrajno oteženo. Posledično večina raziskovalcev privzame ugotovitve, do katerih so se dokopali drugi raziskovalci in pri svojem delu uporabi nastavitve, ki jih je mogoče izbrskati iz literature. Po našem mnenju obstaja v zvezi z učenjem modela UBM še dosti vprašanj, na katerih odgovore bo potrebno počakati še nekaj časa. Naštejmo nekaj takšnih vprašanj:

- Kako vpliva količina podatkov (govorcev, posnetkov itd.) na učinkovitost razpoznavanja?
- Kako izbrati optimalno število komponent modela?
- Katere so najbolj primerne akustične značilke?
- Kakšna je optimalna dimenzija vektorjev značilk?

Opomnimo, da smo pri učenju modela UBM uporabljali tako ženske kot tudi moške posnetke. S tem smo postopali v nasprotju z uveljavljeno prakso, ki pravi, da je bolje imeti ločen model za ženske in moške govorce. (V veliki meri je za ločeno obravnavo moških in žensk odgovoren NIST, ki v svojem protokolu predvideva, da je spol govorca vnaprej znan.) Čeprav preverjeno od spola odvisni (angl. gender dependent, GD) model v primerjavi z od spola neodvisnim (angl. gender independent, GI) modelom dosega boljše rezultate, je po našem mnenju takšno *drobljenje* problema na enostavnejše probleme z znanstvenega stališča najmanj vprašljivo če že ne neupravičeno.

Mere podobnosti

Pri eksperimentih smo vseskozi uporabljali različne mere podobnosti, s katerimi smo računali rezultate prileganja. Uporabljene mere podobnosti lahko razvrstimo v tri skupine:

- mera podobnosti na osnovi razmerja verjetij (LR);
- mera podobnosti na osnovi razdalje (divergence) med porazdelitvami (KL);
- mera podobnosti na osnovi metode podpornih vektorjev (SVM).

Razmerje verjetij

Razmerje verjetij je zaradi velikega števila poskusov, ki jih je potrebno opraviti pri posameznem eksperimentu, časovno zelo potratna mera. Zato smo namesto točne formule za izračun verjetja uporabili približek v obliki metode, ki izračuna verjetje tako, da upošteva le nekaj (v našem primeru 10) najboljših komponent mešanice. Te najboljše komponente izberemo za vsak vektor značilk posebej na modelu UBM.

(V nadaljevanju uporabljamo za tak način računanja verjetja oznako LRa). Sami v disertaciji predlagamo še veliko hitrejši način, pri katerem verjetje izračunamo na podlagi zadostne statistike, ki jo pridobimo s pomočjo modela UBM. (V nadaljevanju za takšen izračun verjetja uporabljamo oznako LRb).

Divergenčna mera KL

Pri metodi merjenja podobnosti na osnovi razdalje med porazdelitvami, smo uporabili divergenčno mero KL, izpeljano za primerjavo dveh modelov GMM. Tudi ta izračun divergence med modeloma GMM ni natančen, saj izraza za divergenco med dvema modeloma GMM ni mogoče zapisati v obliki analitične formule. Zato se namesto točne vrednosti zadovoljimo z oceno zgornje meje divergence. Izraz, s katerim izračunamo to divergenco, je podan v razdelku **5.2**. (V nadaljevanju za tako definirano mero podobnosti uporabljamo izraz KL).

Metoda SVM

Mera podobnosti na osnovi metode podpornih vektorjev se bistveno razlikuje od obeh prej opisanih metod v smislu, da potrebujemo pri ocenjevanju modela SVM za učnega govorca poleg vsaj enega govorčeve posnetka še posnetke, ki jih pri učenju uporabimo kot negativne primere. Čeprav bi za negativne primere lahko izbrali posnetke iz katerekoli govorne zbirke, smo se odločili kar za tiste posnetke, ki smo jih že uporabljali tudi pri učenju modela UBM. S tem smo želeli doseči pošteno primerjavo med različnimi postopki. Ker smo uporabili vseh 16252 posnetkov, smo bili ponovno prisiljeni v pisanje lastne programske opreme, s čimer smo se izognili težavam, povezanimi z omejenim pomnilnikom. Pri metodi SVM lahko izbiramo med različnimi *jedri* (angl. kernel)[40]. Med bolj znanimi je Gaussovo oz. RBF jedro, ki smo ga preizkušali tudi sami. Na koncu smo se vendarle odločili za najenostavnejše linearno jedro, saj smo z njim dosegli primerljive rezultate s tistimi, ki smo jih dobili z Gaussovim jedrom. Zdi se, da je linearno jedro ustrezno takrat, ko je razsežnost vhodnega prostora (angl. input space) dovolj velika, kar v našem primeru gotovo velja. Linearno jedro ima v nasprotju z ostalimi jedri to dobro lastnost, da je brez odprtega parametra, ki bi ga bilo potrebno nastavljati.

Znano je, da je metoda SVM občutljiva na različen razpon vrednosti posameznih koeficientov vektorjev (atributov), zato je atribute potrebno predhodno normirati. Običajna praksa je, da se jih normira na interval $[-1, 1]$ ali $[0, 1]$. V našem primeru smo se odločili za nekoliko drugačno rešitev — uporabili smo normalizacijo ranga. To je neparametrična metoda, ki vsako vrednost atributa nadomesti z njegovim rangom (normiranim na interval $[0, 1]$), ki ustreza percentilu porazdelitve ›ozadja‹. Porazdelitev tako transformiranih vrednosti atributov je približno enakomerna. Normalizacija ranga je sicer računsko in izvedbeno zahtevnejša (sploh če je razsežnost vektorjev ogromna) od preprostega skaliranja, a se izkaže, da vodi do konsistentno

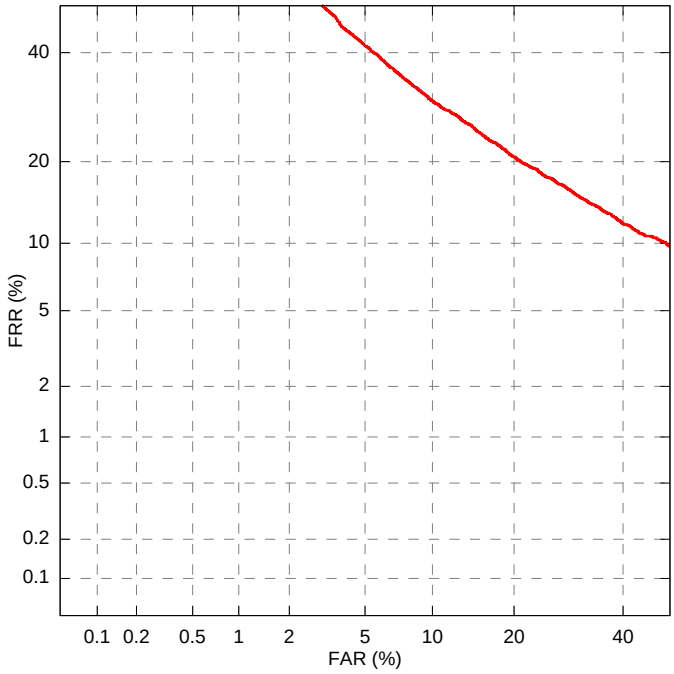
[40] Jedro je posplošen skalarni produkt, ki s pomočjo t.i. *jedrnega trika* (angl. kernel trick) linearen SVM posploši na nelinearnega.

boljših rezultatov (Stolcke et al., 2008), zato smo jo z veseljem dodali v naš ›arzenal orodij‹.

8.3 Referenčni sistem

Referenčni sistem smo zasnovali kar se da podobno sistemu, ki je opisan v (Reynolds et al., 2000) in pomeni nekakšno prelomnico na področju raziskav samodejnega razpoznavanja govorcev, saj prvič uvede model UBM, s katerim opišemo ›povprečnega govorca‹.

Modele učnih govorcev smo izpeljali iz zgoraj opisanega modela UBM. Izpeljavo posameznega modela smo izvedli z eno iteracijo postopka MAP (parameter τ smo postavili na vrednost 16). Za mero podobnosti smo uporabili verjetje, pri izračunu katerega smo za vsak akustični vektor iz učnega posnetka upoštevali le 10 najbolj verjetnih komponent modela UBM. Rezultati referenčnega sistema so predstavljeni na sliki **8.1**.



Slika 8.1 Krivulja DET referenčnega sistema GMM-LRa pri nenormiranih rezultatih prileganja.

Vidimo, da smo z referenčnim sistemom dosegli EER $\approx$ 20% (vsak peti poskus naredimo napako), kar ni najboljši rezultat. Nekoliko slabši rezultat[41] od pričakovanega lahko delno upravičimo z našima odločitvama, ki sta v nasprotju z uveljavljeno prakso:

[41] Slab rezultat ima to koristno lastnost, da ga je lažje popraviti.

- odločili smo se, da bomo uporabili isti model UBM za ženske in moške govorce;
- odločili smo se, da ne bomo izvajali normalizacije značilk.

Poleg rezultatov v obliki krivulj DET bomo rezultat vsakega sistema opremili tudi z najmanjšo vrednostjo DCF (angl. decision cost function), ki se izračuna kot utežena vsota napak P_{FA} in P_{FR}:

$$\mathrm{DCF} = C_{\mathrm{FA}} P_{\mathrm{FA}} + C_{\mathrm{FR}} P_{\mathrm{FR}},$$

kjer sta vrednosti C_{FA} in C_{FR} določeni s strani NIST-a in znašata $0,99$ in $0,1$, v tem vrstnem redu. Ker sta napaki P_{FA} in P_{FR} odvisni od praga, bo od praga odvisna tudi vrednost DCF. Pri rezultatih podamo vrednost DCF pri tistem pragu, pri katerem je vrednost DCF najmanjša. Vrednost DCF ponavadi še normiramo tako, da jo podelimo z vrednostjo DCF, ki bi jo dosegel naiven sistem, ki bi vse poskuse bodisi sprejel bodisi zavrnil.

Rezultate smo najprej poskusili popraviti z normalizacijo rezultatov prileganja. Dobljene rezultate smo zbrali v tabeli **8.2**. Opazimo lahko, da normalizacija rezultatov prileganja pri sistemih GMM-LRa in GMM-LRb ne pripomore bistveno k izboljšanju rezultatov. Če primerjamo oba sistema med seboj, vidimo, da da sistem GMM-LRb nekoliko slabši rezultat v okolici točke EER, medtem ko v okolici točke najnižje vrednosti DCF doseže približno enak rezultat kot sistem LRa.

			z-norm		t-norm		zt-norm		tz-norm	
	EER	DCF	EER	DCF	EER	DCF	EER	DCF	EER	DCF
GMM-LRa	20,4	0,75	20,4	0,69	20,6	0,65	20,3	0,63	20,4	0,63
GMM-LRb	23,5	0,75	22,9	0,72	23,9	0,69	22,9	0,69	23,4	0,69
GMM-KL	35,7	0,10	30,1	0,96	30,6	0,95	19,7	0,69	19,1	0,72
GMM-SVM	9,82	0,43	10,0	0,45	9,51	0,41	9,51	0,42	9,62	0,43

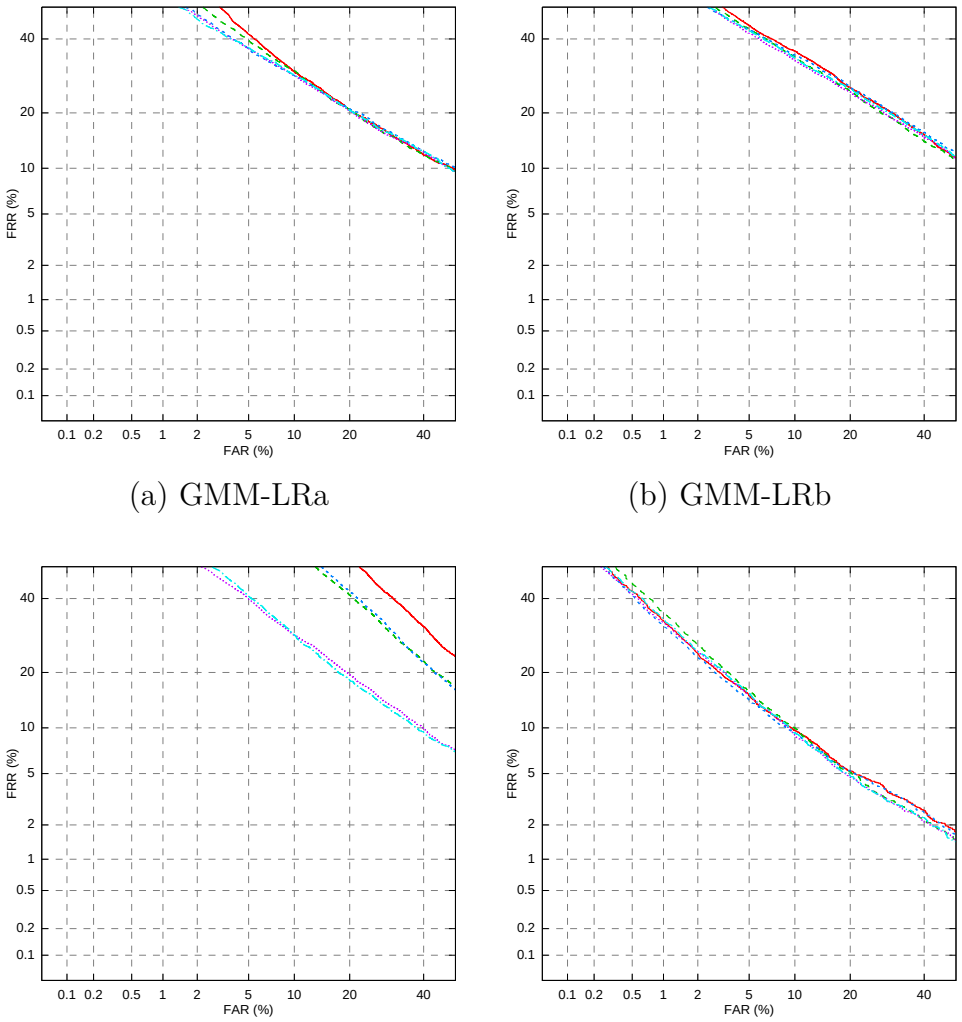
Legenda:

GMM-LRa	referenčni sistem (verjetje; 10 najverjetnejših komponent)
GMM-LRb	referenčni sistem (verjetje; zadostna statistika)
GMM-KL	referenčni sistem (divergenca KL)
GMM-SVM	referenčni sistem (metoda SVM)

Tabela 8.2 Rezultati referenčnega sistema z uporabo različnih mer prileganja podani v obliki EER in DCF brez in z izvedbo različnih normalizacij rezultatov prileganja.

Najbolj zanimivi rezultati se skrivajo v tretji vrstici tabele, ki povzema rezultate sistema, pri katerem smo kakovost prileganja merili z divergenčno mero KL. Vidimo, da so nenormirani rezultati znatno slabši kot pri sistemih GMM-LRa in GMM-LRb,

a se le-ti popravijo, če jih normiramo. Najboljše rezultate dobimo, če izvedemo normiranje na način zt-norm ali tz-norm — takrat celo presežejo tiste, ki jih dobimo pri sistemih GMM-LRa in GMM-LRb. Učinkovitost normiranja rezultatov prileganja je najbolje razvidna iz krivulje DET (slika **8.2c**).



(a) GMM-LRa (b) GMM-LRb

(c) GMM-KL (d) GMM-SVM

Slika 8.2 Krivulje DET referenčnega sistema v navezi s štirimi različnimi merami podobnosti. Različne metode normalizacije rezultatov prileganja so vrisane z različnimi barvami: ■ (brez normalizacije), ■ (z-norm), ■ (t-norm), ■ (zt-norm) in ■ (tz-norm).

Poglejmo si še, kakšne rezultate dobimo, če namesto verjetja za računanje rezultatov prileganja izberemo metodo SVM. Pričakovali bi, da bodo rezultati primerljivi s tistimi, ki smo jih dobili s kriterijem razmerja verjetij. Presenetljivo pa se izkaže, da se rezultati občutno popravijo (slika **8.2d**). Tako se EER zmanjša za polovico iz približno 20 % na zavidljivih 10 %. Prav tako se minimalni dosežen DCF popravi relativno za dobrih 40 %. Vprašamo se lahko, kaj je vzrok takemu občutnemu

izboljšanju rezultata? Po našem mnenju je prednost kriterija SVM pred kriterijema LR in KL v *diskriminatorni* naravi modela SVM. Sklepamo lahko, da metoda SVM postavi ločilno mejo med učnim vektorjem in vektorji ozadja tako, da je le-ta ›pravokotna‹ na smeri, ki ne prispevajo k boljši ločljivosti oz. celo kvarno vplivajo na ločevanje med govorci. V nekem smislu tako že sama metoda SVM poišče podprostor, ki je do neke mere ›imun‹ na sejno spremenljivost. Do tako velike razlike med kriterijem SVM in ostalimi verjetnostnimi kriteriji je po našem mnenju prišlo zaradi tega, ker nismo opravili normalizacije na nivoju značilk, s katero bi se delno lahko znebili vpliva različnih akustičnih razmer. Če bi izvedli eno izmed normalizacij značilk (npr. ukrivljanje značilk), bi verjetno razlika med kriteriji bila manjša.

Če pogledamo številke v četrti vrstici tabele **8.2**, opazimo, da vpliv normalizacije rezultatov prileganja pri kriteriju SVM ni zelo izrazit. Kljub temu lahko trdimo, da se najbolj splača izvesti normalizacijo t-norm, kar se sklada tudi z ugotovitvami v literaturi.

8.4 Sistem NAP

V tem razdelku bomo predstavili rezultate, ki smo jih dobili, ko smo v postopek samodejnega razpoznavanja govorcev vključili metodo za normalizacijo sejne spremenljivosti — NAP.

Najprej je potrebno oceniti kanalski podprostor v obliki matrike $\boldsymbol{V}$. Kako se to stori, smo podrobno predstavili v razdelku **6.5**. Tukaj na kratko opišimo le praktično izvedbo ocene te matrike.

Osnova so bili govorni posnetki, ki smo jih uporabljali že pri učenju modela UBM. Skupno število posnetkov je 16252, število različnih govorcev pa 1330 (782 žensk in 548 moških), kar znese v povprečju približno 12 posnetkov na enega govorca. Postopali smo na sledeč način. Najprej smo za vsak posnetek z eno iteracijo adaptacije MAP izpeljali sejni model GMM, ki smo ga pretvorili v supervektor. Nato smo vsem supervektorjem posameznega govorca odšteli (od govorca odvisni) povprečni supervektor[42]. Na ta način smo dobili množico supervektorjev, v katerih glavni delež spremenljivosti pripada kanalski komponenti. Na podlagi teh supervektorjev lahko ocenimo kanalski podprostor z eno izmed metod redukcije razsežnosti. Ena izmed preprostejših je metoda PCA, ki smo jo uporabili tudi sami. Razsežnost podprostora smo določili eksperimentalno. Izbrali smo dimenzijo 100, kar je vrednost, ki jo je pogosto zaslediti v literaturi.

Metoda NAP je bila predlagana v navezi z mero podobnosti na osnovi metode SVM. Razpoznavanje poteka enako kot pri referenčnemu sistemu, le da pred ocenjevanjem modela SVM za posameznega govorca iz govorčevega supervektorja odstranimo kanalsko komponento. To storimo tako, da najprej poiščemo kanalsko komponento supervektorja tako, da izvedemo preslikavo sejnega supervektorja v kanalski podprostor in ga takoj zatem spet preslikamo v prostor supervektorjev. Dobljeni

[42] Upravičeno lahko pričakujemo, da s povprečenjem preko dovolj velikega števila sejnih vektorjev odstranimo večino kanalske spremenljivosti.

kanalski supervektor odštejemo sejnemu supervektorju in kar ostane, je govorski supervektor. Matematično to zgoščeno zapišemo z izrazom $\boldsymbol{s} = (\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}})\boldsymbol{x}$.

Enako ponovimo tudi za vsak testni supervektor. Nadaljni postopek razpoznavanja je čisto enak tistemu, ki smo ga uporabili pri referenčnemu sistemu. Učni supervektor uporabimo za učenje modela SVM, na katerega prilegamo testne supervektorje.

Poleg metode SVM smo preizkusili tudi kriterija na osnovi razmerja verjetij (LRb) in kriterija na osnovi divergence KL, ki smo ju predlagali v disertaciji. Če naša predpostavka o kompenzacijski moči vplivov kanala metode SVM drži, potem bi morala ta dva kriterija po eksplicitni normalizaciji kanalske spremenljivosti postati konkurenčnejša metodi SVM.

Medtem ko je kriterij KL trivialno uporabiti v navezi z metodo NAP, pa se situacija nekoliko zaplete, če želimo uporabiti kriterij LR. Poiskati moramo način, kako izkompenzirati vpliv testnega posnetka, ki ga pri kriteriju LR za razliko od kriterijev SVM in KL ne pretvorimo v supervektor, preden ga prilegamo na učni model. V disertaciji predlagamo eno izmed možnih rešitev. Postopamo tako, da iz testnega posnetka najprej pridobimo kanalsko komponento testnega supervektorja, ki ga prištejemo govorski komponenti učnega supervektorja. Na ta način smo govorski supervektor prilagodili kanalu, ki je prisoten v testnem posnetku. Bolj natančen opis postopka se nahaja v razdelku **6.8**.

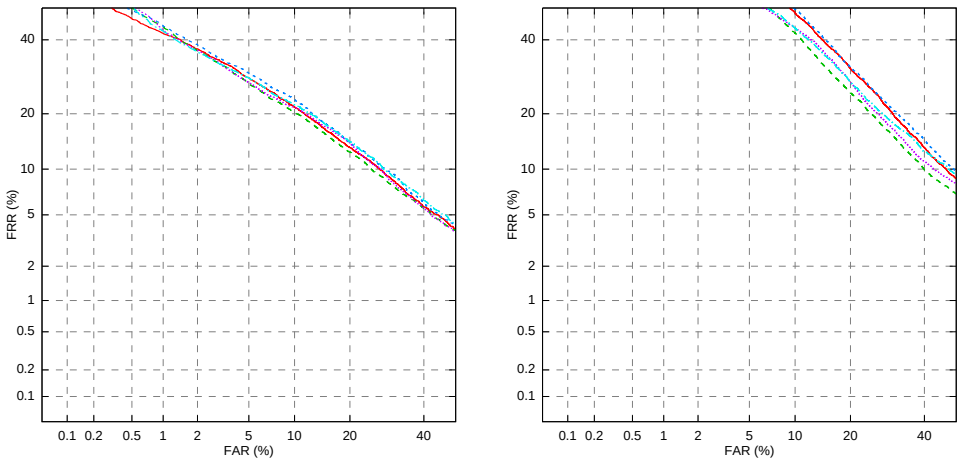
Če pogledamo krivulje na sliki **8.3**, ugotovimo, da kompenzacija kanalske spremenljivosti z metodo NAP ni preveč blagodejno vplivala na kriterij LR. Rezultati kriterija LRa so se v primerjavi s sistemom GMM sicer nekoliko izboljšali, rezultati kriterija LRb pa so se celo še nekoliko poslabšali. Povsem drugače ugotovimo za kriterija SVM in KL. Kriterij SVM se je po pričakovanju odrezal zelo dobro. Napaka EER, ki je pri sistemu GMM-SVM znašala približno 9,5 %, se je zmanjšala še za dobro polovico odstotka. Še bolj zanimivi so rezultati kriterija KL. Opazimo podobno lastnost, da so nenormirani rezultati zelo slabi (EER $\approx$ 30%), vendar se občutno izboljšajo po zt-norm in tz-norm normalizaciji. Učinek normalizacije rezultatov prileganja je še bolj učinkovit kot pri sistemu GMM in znaša zavidljivih 8,5 % EER ter 0,40 DCF, s čimer se zelo približa kriteriju SVM (glej tabelo **8.3**).

8.5 Sistem JFA

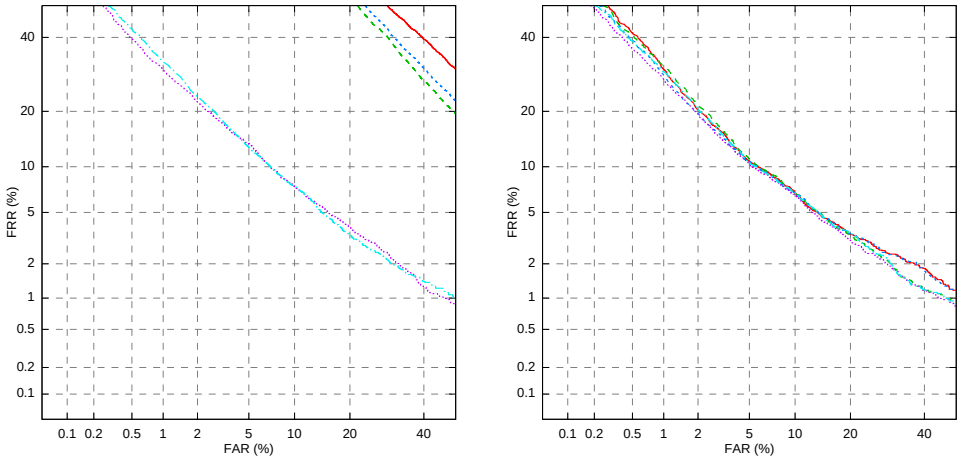
Pri postopku JFA je potrebno najprej oceniti hiperparametre modela. Za to potrebujemo govorno zbirko, ki naj vsebuje čim večje število govorcev v čimveč različnih posnetkih. Hiperparametrov nismo ocenjevali sočasno ampak ločeno. Pri tem smo sledili napotkom, predstavljenimi v (Kenny et al., 2008).

Ponovno smo izhajali iz podatkov, ki smo jih uporabljali za učenje modela UBM. Vse posnetke smo razdelili v dve množici. V manjšo množico B smo dali vse govorce, ki so imeli največ osem različnih posnetkov, v večjo množico A pa vse ostale. (Natančne številke so zbrane v tabeli **8.4**.)

Večjo množico smo uporabili za učenje parametrov govorskega podprostora (matrika $\boldsymbol{V}$), kanalskega podprostora (matrika $\boldsymbol{U}$) in povprečnega vektorja $\boldsymbol{\mu}$, manjšo



(a) NAP-LRa (b) NAP-LRb



(c) NAP-KL (d) NAP-SVM

Slika 8.3 Krivulje DET sistemov NAP z uporabo različnih odločitvenih kriterijev in različnih tehnikah normalizacije rezultatov prileganja. Različne metode normalizacije rezultatov prileganja so vrisane z različnimi barvami: ■ (brez normalizacije), ■ (z-norm), ■ (t-norm), ■ (zt-norm) in ■ (tz-norm).

množico pa za oceno matrike $\boldsymbol{D}$ in matrike $\boldsymbol{\Sigma}$. Celoten postopek učenja lahko razdelimo v štiri korake, ki jih izvedemo v naslednjem vrstnem redu:

- določitev začetnih vrednosti hiperparametrov
- učenje modela lastnih govorcev
- učenje modela lastnih kanalov
- učenje diagonalnega modela

Določitev začetnih vrednosti hiperparametrov

			z-norm		t-norm		zt-norm		tz-norm	
	EER	DCF	EER	DCF	EER	DCF	EER	DCF	EER	DCF
NAP-LRa	16,0	0,51	15,4	0,54	16,7	0,54	16,3	0,53	16,6	0,52
NAP-LRb	25,5	0,87	22,6	0,86	25,6	0,88	23,2	0,73	23,6	0,76
NAP-KL	39,7	1,00	33,9	1,00	35,4	1,00	8,49	0,40	8,52	0,42
NAP-SVM	8,10	0,39	8,21	0,40	8,02	0,37	7,91	0,37	7,95	0,39

Legenda:

NAP-LRa	referenčni sistem (verjetje; 10 najverjetnejših komponent)
NAP-LRb	referenčni sistem (verjetje; zadostna statistika)
NAP-KL	referenčni sistem (divergenca KL)
NAP-SVM	referenčni sistem (metoda SVM)

Tabela 8.3 Rezultati sistema NAP z uporabo različnih mer prileganja podani v obliki EER in DCF brez in z izvedbo različnih normalizacij rezultatov prileganja.

množica	št. govorcev			št. posnetkov		
	Ž	skupaj	M	Ž	skupaj	M
A	613	1047	434	8492	14711	6219
B	137	224	87	943	1541	598
skupaj	9435	16252	6817	750	1271	521

Tabela 8.4 Razdelitev podatkov na dve množici pri ločenem ocenjevanju hiperparametrov modela JFA.

Najprej smo izbrali razsežnosti govorskega in kanalskega podprostora. Z upoštevanjem izsledkov iz literature smo se odločili, da naj bo razsežnost govorskega podprostora 300, razsežnost kanalskega podprostora pa 100. Začetne vrednosti parametrov $\boldsymbol{\mu}$ in $\boldsymbol{\Sigma}$ smo ›prepisali‹ iz modela UBM, začetne vrednosti ostalih parametrov ($\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{D}$) pa smo izbrali naključno.

Učenje modela lastnih govorcev

Za učenje modela lastnih govorcev smo uporabili posnetke iz množice A. Najprej smo združili vse posnetke posameznega govorca. Z združevanjem smo želeli doseči, da se kanalski vplivi čimbolj izpovprečijo. (To je upravičeno pričakovati, saj imamo na voljo veliko število posnetkov vsakega govorca.) Nato smo izvedli deset iteracij postopka EM, pri čemer smo postavili $\boldsymbol{U} = 0$ in $\boldsymbol{D} = 0$, z drugimi besedami, ocenili smo model lastnih govorcev (matriki $\boldsymbol{V}$ in $\boldsymbol{\Sigma}$). Sledile so še tri iteracije postopka EM po kriteriju MD, kjer smo ocenili vektor $\boldsymbol{\mu}$ in popravili oceno matrike $\boldsymbol{V}$.

Učenje modela lastnih kanalov

Za učenje modela lastnih kanalov smo uporabili posnetke iz množice A. Najprej smo s pomočjo modela lastnih govorcev izračunali točkovno oceno posteriorne porazdelitve govorskih faktorjev $\boldsymbol{y}$. Na ta način smo dobili oceno govorskega supervektorja $\boldsymbol{s}$:

$$\boldsymbol{s} = \boldsymbol{\mu} + \boldsymbol{V}\mathbb{E}[\boldsymbol{y}]$$

Nato smo zadostno statistiko prvega in drugega reda vsakega posnetka usrediščili tako, kot je opisano v razdelku **7.4.7** (glej enačbi (**7.8**) in (**7.9**) na str. **68**). Ker smo na ta način iz statistik izločili govorsko komponento, je v statistikah preostala le še sejna oz. kanalska spremenljivost.

Sledil je analogen postopek tistemu, s katerim smo učili model lastnih govorcev, le da smo sedaj posnetke obravnavali ločeno in jih nismo združili po govorcih. Po desetih iteracijah postopka EM po kriteriju ML in tremi iteracijami po kriteriju MD smo dobili končno oceno matrike $\boldsymbol{U}$, hkrati pa smo popravili še oceni matrike $\boldsymbol{\Sigma}$ in vektorja $\boldsymbol{\mu}$.

Učenje diagonalnega modela

V zadnjem koraku smo ocenili še parametre diagonalnega modela. Pri tem smo uporabili posnetke iz množice B. Najprej smo vse posnetke enega govorca združili (s tem smo izločili kanalsko spremenljivost), nato pa smo, podobno kot pri učenju modela lastnih kanalov, izpeljali točkovno oceno posteriorne porazdelitve govorskih faktorjev $\boldsymbol{s}$. Ko smo s to oceno usrediščili statistiko prvega in drugega reda, smo iz statistik izločili še večino govorske spremenljivosti. V statistiki je ostala le preostala spremenljivost, ki je nismo mogli opisati ne z modelom lastnih govorcev ne z modelom lastnih kanalov. Izvedli smo deset iteracij postopka EM (kriterij ML) in še dodatne tri iteracije po kriteriju MD, pri čemer smo zahtevali, da je $\boldsymbol{\mu} = 0$. (Ta zahteva izhaja iz dejstva, da smo iz statistike skupaj z govorsko komponento $\boldsymbol{s}$ izločili tudi povprečni vektor $\boldsymbol{\mu}$). Kot rezultat smo dobili oceno matrike $\boldsymbol{D}$ in popravljeno oceno matrike $\boldsymbol{\Sigma}$.

Razpoznavanje

Preizkusili smo tri različne odločitvene kriterije: razmerje verjetij, divergenco KL in metodo SVM. Razmerje verjetij smo izračunali na dva različna načina; z metodo na podlagi zadostne statistike in z metodo, ki upošteva nedoločenost kanala.

Verjetje na osnovi zadostne statistike smo izračunali tako, da smo za učni posnetek izpeljali točkovno oceno govorskega supervektorja $\boldsymbol{s}$:

$$\boldsymbol{s} = \boldsymbol{\mu} + \boldsymbol{V}\mathbb{E}[\boldsymbol{y}] + \boldsymbol{D}\mathbb{E}[\boldsymbol{z}],$$

za testni vektor pa točkovno oceno kanalskega supervektorja c:

$$\boldsymbol{c} = \boldsymbol{U}\mathbb{E}[\boldsymbol{x}].$$

(Količine $\boldsymbol{x}$, $\boldsymbol{y}$ in $\boldsymbol{z}$ sledijo iz rezultata (7.5), ki podaja posteriorno porazdelitev prikritih spremenljivk modela JFA.) Nato smo učni govorski supervektor prilagodili na kanalske razmere testnega posnetka tako, da smo mu prišteli testni kanalski supervektor. (Enako smo naredili tudi za supervektor ›povprečnega‹ govorca, s katerim izračunamo verjetje, ki nastopa v imenovalcu izraza za razmerje verjetij.) Razmerje verjetij smo izračunali tako, da smo ocenili kakovost prileganja testnega posnetka na adaptirani učni supervektor (števec) oz. na adaptirani povprečni supervektor (imenovalec):

$$\frac{p(o|\boldsymbol{s}+\boldsymbol{c})}{p(o|\boldsymbol{\mu}+\boldsymbol{c})}.$$

Konceptualno precej drugače smo postopali pri metodi računanja verjetja z upoštevanjem nedoločenosti kanala. Tudi tukaj smo najprej za vsak učni posnetek izpeljali točkovno oceno govorskega supervektorja $\boldsymbol{s}$:

$$\boldsymbol{s} = \boldsymbol{\mu} + \boldsymbol{V}\mathbb{E}[\boldsymbol{y}] + \boldsymbol{D}\mathbb{E}[\boldsymbol{z}],$$

računanje točkovne ocene kanalskega supervektorja v testnem posnetku pa smo izpustili. Kakovost prileganja testnega posnetka na govorski supervektor smo ocenili tako, da smo izračunali verjetje po enačbi (7.10). Pri računu verjetja na ta način predpostavljamo, da kanalski supervektor v testnem posnetku ni znan, vse kar vemo je, da se porazdeljuje normalno: $\boldsymbol{c} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}})$. Enačbo (7.10) lahko interpretiramo kot (utežen) seštevek verjetij po vseh možnih konfiguracijah kanala $\boldsymbol{c}$, t.j.:

$$p(o|\boldsymbol{s}) = \int_{\mathbb{R}^{D_u}} p(o|\boldsymbol{s}+\boldsymbol{U}\boldsymbol{x})\,\mathcal{N}(\boldsymbol{x}|\boldsymbol{0},\boldsymbol{I})\mathrm{d}\boldsymbol{x}.$$

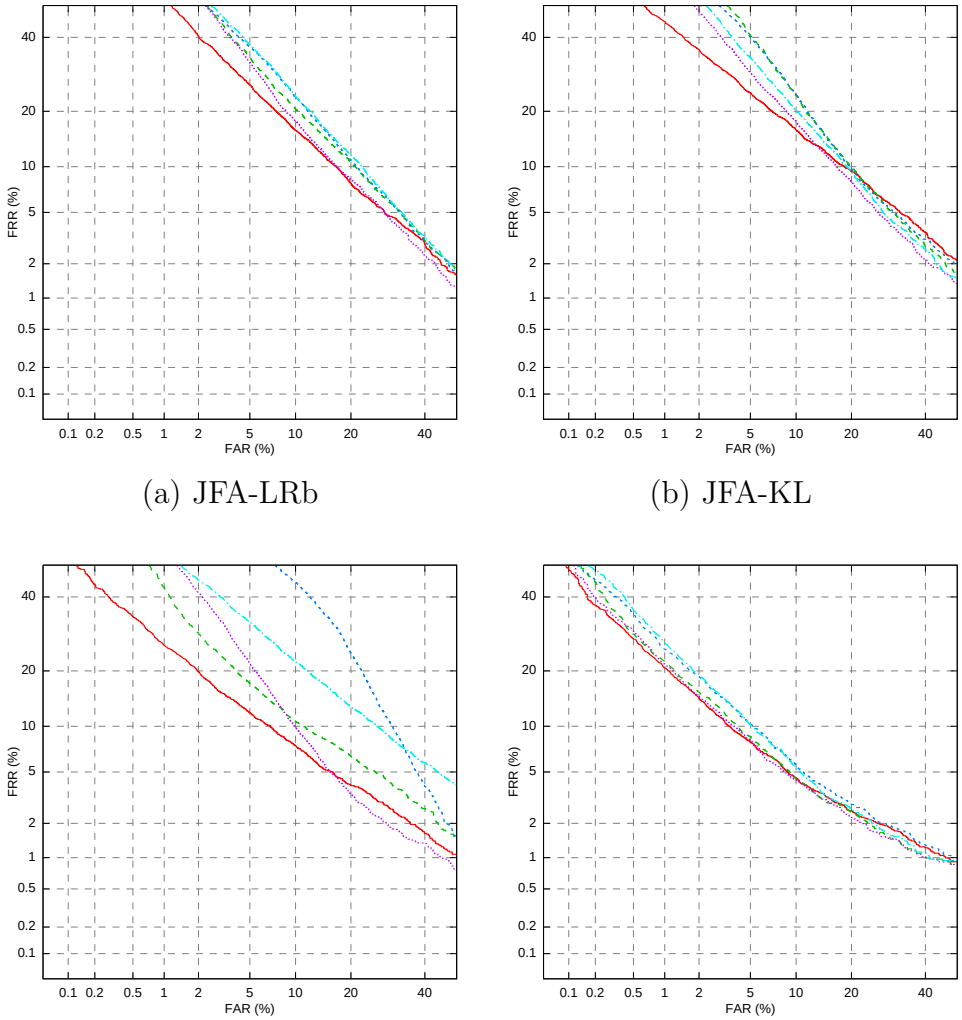
V primeru divergence KL smo ocenili govorski supervektor tako za učni kot za testni posnetek, njuno prileganje pa smo ocenili tako, da smo izračunali divergenco KL po enačbi (5.6).

Podobno smo storili v primeru metode SVM, kjer smo govorski supervektor učnega posnetka uporabili za učenje modela, govorski supervektor testnega posnetka pa za izračun prileganja po enačbi (5.7). Predhodno smo oba supervektorja ›podvrgli‹ normalizaciji ranga na enak način, kot smo to storili pri sistemih GMM in NAP.

Rezultati

Rezultati sistema JFA z uporabo različnih mer podobnosti so zbrani v tabeli 8.5 in prikazani v obliki krivulj DET na sliki 8.4. Ugotovimo lahko, da so rezultati, ki smo jih dosegli s kriterijema LRb in KL daleč pod pričakovanimi. Rezultati sistema JFA-LRb so sicer nekoliko boljši od rezultatov primerljivega sistema NAP-LRb, a so zato rezultati sistema JFA-KL precej slabši od rezultatov sistema NAP-KL. Veliko bolje se odreže kriterij LRc, ki upošteva nedoločenost kanala v testnem posnetku. Najboljše rezultate dosežemo ponovno s kriterijem SVM, za katerega lahko trdimo, da daje v vseh pogledih najboljše in tudi najbolj konsistentne rezultate.

Velika razlika v učinkovitosti kriterijev LRb in LRc daje slutiti, da je nekaj narobe s točkovno oceno kanala, ki jo naredimo pri kriteriju LRb. Po našem mnenju se skriva vzrok v privzetku, da lahko obravnavamo zadostno statistiko od govorca in kanala neodvisno. Zdi se, da ta predpostavka ni povsem upravičena, še posebej zato, ker smo izpustili normalizacijo na nivoju značilk. K sreči se izkaže, da s povprečenjem verjetja preko vseh možnih konfiguracij testnega kanala (kriterij LRc) to predpostavko dovolj ›zrahljamo‹, saj so rezultati sistema JFA-LRc veliko boljši. Podobno moč ima tudi metoda SVM, ki prav tako uspešno nadoknadi zaostanek, ki izvira iz naše odločitve, da skušamo odpraviti potrebo po normalizaciji značilk.



(a) JFA-LRb (b) JFA-KL

(b) JFA-LRc (c) JFA-SVM

Slika 8.4 Krivulje DET sistema JFA v kombinaciji z različnimi merami podobnostmi in normalizacijskimi tehnikami rezultatov prileganja. Različne metode normalizacije rezultatov prileganja so vrisane z različnimi barvami: ■ (brez normalizacije), ■ (z-norm), ■ (t-norm), ■ (zt-norm) in ■ (tz-norm).

			z-norm		t-norm		zt-norm		tz-norm	
	EER	DCF	EER	DCF	EER	DCF	EER	DCF	EER	DCF
JFA-LRb	13,0	0,60	14,7	0,72	15,3	0,71	13,4	0,71	15,5	0,73
JFA-KL	13,28	0,54	14,95	0,77	14,76	0,74	13,25	0,68	14,20	0,72
JFA-LRc	8,52	0,36	10,43	0,48	21,52	0,73	9,93	0,61	15,89	0,64
JFA-SVM	6,54	0,30	6,72	0,32	7,32	0,35	6,27	0,31	7,45	0,36

Legenda:

JFA-LRc	sistem JFA (verjetje z upoštevanjem nedoločenosti kanala)
JFA-KL	sistem JFA (divergenca KL)
JFA-SVM	sistem JFA (metoda SVM)

Tabela 8.5 Rezultati sistema JFA z uporabo različnih mer prileganja podani v obliki EER in DCF brez in z izvedbo različnih normalizacij rezultatov prileganja.

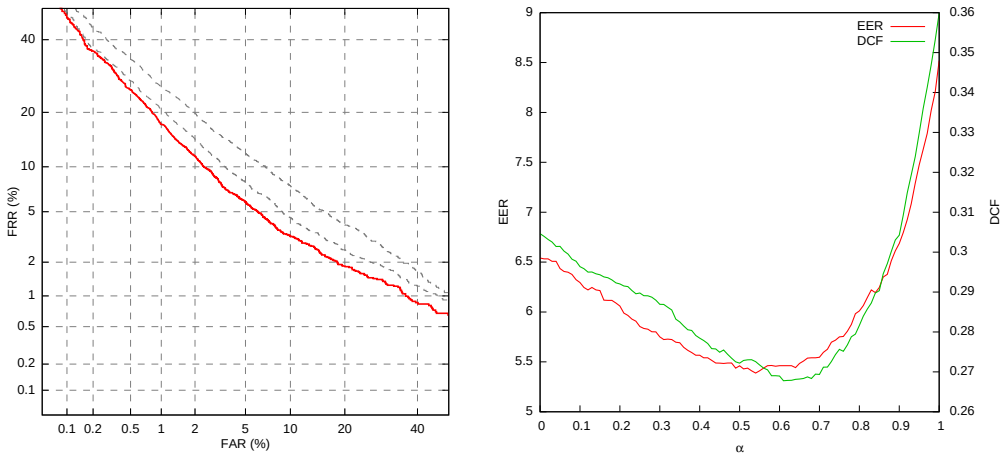
Do zanimivih ugotovitev pridemo tudi, če preučimo vpliv različnih metod rezultatov prileganja. V oči bode dejstvo, da se rezultati razpoznavanja v primeru, ko smo izvedli normalizacijo rezultatov prileganja, niso izboljšali niti v enem primeru, ampak so se celo poslabšali. Še posebej izrazit negativen učinek je imela normalizacija rezultatov prileganja pri kriteriju LRc, kar je v velikem nasprotju z ugotovitvami iz literature, kjer ponavadi poročajo o sinergijskem učinku med metodo JFA in normalizacijo zt-norm (Kenny et al., 2007a). Vzrok za tako nasprotujoče si ugotovitve bi lahko bil v tem, da smo posnetke, ki smo jih uporabili za normalizacijo rezultatov prileganja, izbrali v od spola neodvisni maniri. V to nas navajajo izsledki (Burget et al., 2009), kjer so prišli do podobnih ugotovitev, namreč da od spola neodvisna normalizacija rezultatov prileganja vodi do znatnega poslabšanja rezultatov, medtem ko od spola odvisna normalizacija rezultate občutno izboljša (relativno glede na nenormalizirane rezultate).

8.6 Združevanje rezultatov razpoznavanja

Iz literature vemo, da lahko z združevanjem rezultatov prileganja različnih razpoznavalnikov govorcev znatno izboljšamo rezultate razpoznavanja (Brummer et al., 2007). Zato je uveljavljena praksa vseh večjih raziskovalnih skupin, da v okviru NIST-ovih evaluacij predstavijo rezultate sistema, v katerem združijo rezultate velikega števila (tipično do deset) heterogenih sistemov. Nas je zanimalo, če se da rezultate izboljšati že s fuzijo dveh enakih sistemov, ki se razlikujeta le v uporabljenem odločitvenem kriteriju. Tako smo združili rezultate prileganja dveh sistemov, JFA-LRc in JFA-SVM. Rezultate smo združili tako, da smo za vsak poskus izračunali uteženo vsoto prispevkov obeh sistemov po enačbi:

$$s = \alpha s_1 + (1 - \alpha) s_2.$$

Rezultati so predstavljeni na sliki 8.5. Vidimo, da smo rezultat kar precej popravili tako glede na doseženo vrednost DCF (iz 0,30 pri JFA-SVM in 0,36 pri JFA-LRc na 0,27) kot tudi glede na doseženo vrednost EER (iz 6,54 % pri JFA-SVM in 8,52 % pri JFA-LRc na 5,5 %).



(a) (b)

Slika 8.5 Na sliki (a) je prikazana krivulja DET, ki jo dobimo z združevanjem rezultatov prileganja sistemov JFA-LRc in JFA-SVM pri $\alpha = 0,5$. Za primerjavo sta dorisani še krivulji DET obeh posameznih sistemov. Na sliki (b) sta prikazani količini EER in najmanjši DCF v odvisnosti od uteži α, ki določa razmerje med prispevkoma obeh sistemov k skupnemu rezultatu prileganja.

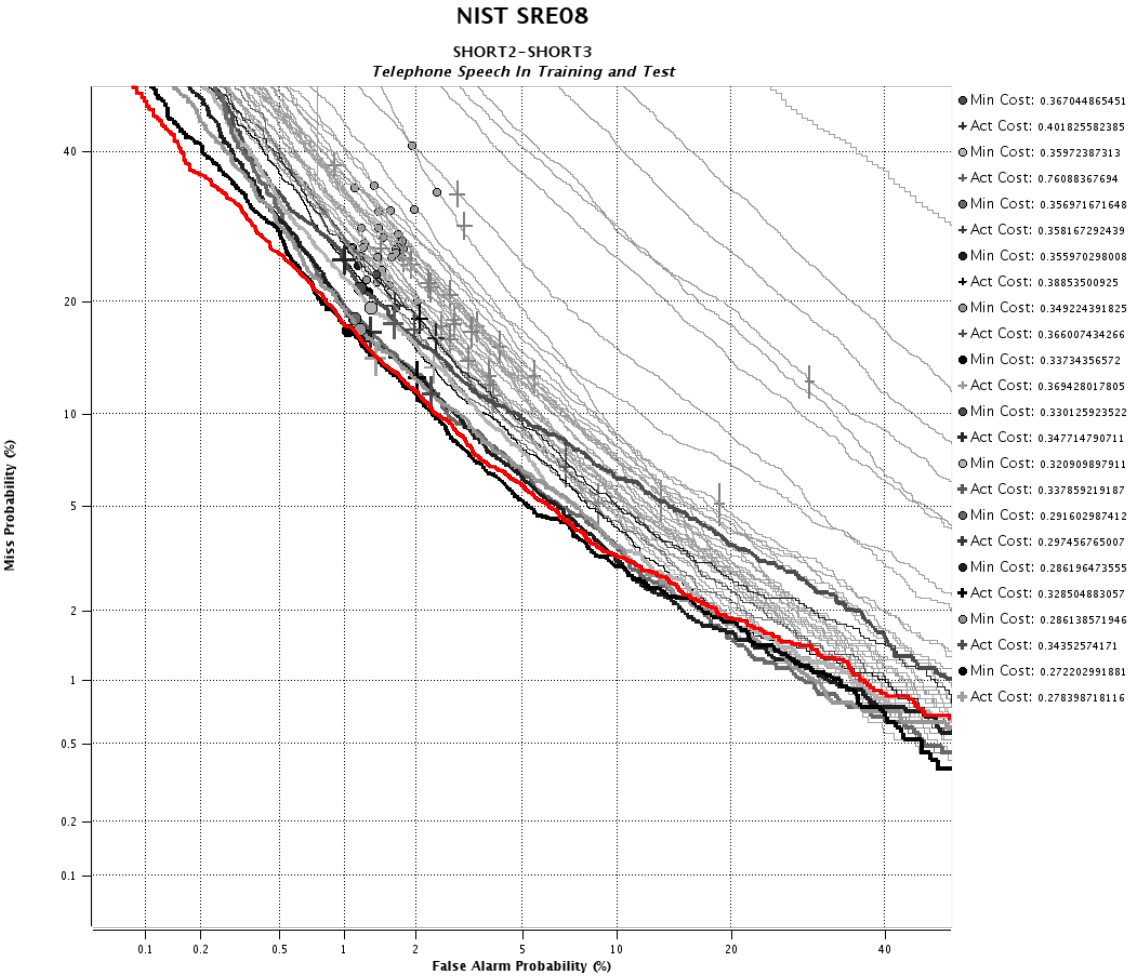
8.7 Uradni rezultati NIST SRE 2008

Za konec prikažimo še uradne rezultate zadnjega NIST-ovega vrednotenja sistemov za samodejno razpoznavanje govorcev (NIST SRE 2008). Za lažjo primerjavo smo k uradnim rezultatom (slika 8.6) z rdečo barvo dorisali še krivuljo DET našega sistema s slike 8.5. Ugotovimo lahko, da je naš sistem povsem konkurenčen najboljšim sistemom z NIST-ovega tekmovanja.

8.8 Komentar

Z dobljenimi rezultati, ki smo jih pridobili na telefonskem delu osnovnega testa NIST SRE 2008, smo lahko več kot zadovoljni. S sistemom, ki smo ga tvorili z združevanjem rezultatov prileganja sistemov JFA-SVM in JFA-LRc, smo dosegli EER $= 5,5\%$ in DCF $= 0,27$, s čimer smo se postavili ob bok najboljšim sistemom z NIST-ovega tekmovanja 2008. Še pomembneje, za razliko od večine najboljših sistemov, ki po navadi

[43] Rezultati so dostopni na internetnem naslovu http://www.itl.nist.gov/iad/mig/tests/sre/2008/official_results/index.html.



Slika 8.6 Primerjava rezultatov, doseženih z združevanjem rezultatov prileganja sistemov JFA-LRc in JFA-SVM (krivulja DET v rdeči barvi), z uradnimi rezultati[43] NIST-ovega vrednotenja sistemov za samodejno razpoznavanje govorcev NIST SRE 2008 (krivulje DET v sivih odtenkih).

združujejo rezultate prileganja večjega števila različnih (in karseda komplementarnih) podsistemov, smo tak rezultat dosegli z relativno nekompleksnim[44] sistemom. Ponavadi se združevanje rezultatov — čemur s tujko pravimo tudi fuzija — opravi s kakšno bolj sofisticirano metodo, npr. z večslojnim perceptronom ali logistično regresijo (Brummer et al., 2007), mi pa smo rezultate prileganja (pod)sistemov[45] JFA-SVM in JFA-LRc združili z ›navadnim‹ seštevanjem.

Čeprav je osnovni sistem JFA zelo podoben tistemu, ki ga je v seriji člankov predlagal (Kenny et al., 2004, 2005, 2007a, 2007b, 2008), pa se od večine podobnih

[44] S kompleksnostjo tukaj merimo predvsem na število različnih podsistemov.
[45] Lahko bi celo rekli, da je sistem le eden, dve pa sta meri podobnosti, s katerima smo izračunali rezultate prileganja.

sistemov razlikuje po nekaterih pomembnih lastnostih. Vredno je, da te razlike še enkrat poudarimo:

- Od spola odvisni UBM (ženski in moški) smo nadomestili z od spola neodvisnim modelom UBM. Hkrati s tem smo zavrgli informacijo o spolu govorca v vseh učnih in testnih posnetkih, ki jo priskrbi NIST in jo je po protokolu dovoljeno uporabiti.
- Nismo izvedli normalizacije na nivoju značilk, kar je v nasprotju z uveljavljeno prakso. Praktično ni zaslediti sistema, ki ne bi uporabljal ene izmed metod za normalizacijo značilk, med katerimi so bolj pogoste metoda ukrivljanja značilk, normalizacija povprečja in variance in filtriranje RASTA.

Odločitev, da ne bomo razlikovali med ženskimi govorkami in moškimi govorci[46] in zaradi tega gradili dveh enakih a ločenih sistemov, izvira iz našega prepričanja, da (umetno) drobljenje osnovnega problema na manjše podprobleme ni v skladu z znanstvenim pristopom, ampak je bolj plod inženirske pragmatičnosti v smislu, če dobimo s tem boljše rezultate, potem je odločitev upravičena[47].

Prav tako smo prekršili tudi drugo nenapisano pravilo, ki pravi, da je v primeru neenakih akustičnih razmer v učnih in testnih posnetkih nujno potrebno opraviti normalizacijo značilk. Naš argument, da te normalizacije ne opravimo, izvira iz dejstva, da pri vsaki normalizaciji izgubimo tudi vsaj nekaj koristne informacije in gre v bistvu za iskanje kompromisa med tem, koliko z normalizacijo pridobimo in koliko izgubimo. Naša ideja je bila, da normalizacijo povsem prepustimo eni izmed metod normalizacije sejne variabilnosti, ki to normalizacijo opravi na modelskem nivoju. Glede na rezultate lahko zaključimo, da normalizacija na modelskem nivoju bolj ali manj uspešno nadomesti tudi normalizacijo, ki jo drugače zagotovi normalizacija značilk. To je še posebej očitno pri metodi SVM, ki doseže dobre rezultate že pri referenčnem sistemu GMM, in do neke mere tudi pri metodi KL, medtem ko je kriterij LR precej manj robusten na nenormalizirane značilke. Kriterij LR se lahko po rezultatih kosa s kriterijem SVM le v primeru, ko verjetje izračunamo z upoštevanjem nedoločenosti kanala (kriterij LRc). Sklepamo lahko, da nenormalizirane značilke povročijo, da je točkovna ocena kanala nenatančna. Razlago, zakaj bi temu lahko bilo tako, smo poskusili podati že v razdelku **5.2**.

Glede na izsledke, objavljene v literaturi (Burget et al., 2009), smo pričakovali, da bomo rezultate razpoznavanja dodatno izboljšali z normalizacijo rezultatov prileganja. Žal so se ta pričakovanja uresničila le v primeru odločitvenega kriterija KL, medtem ko smo pri nekaterih rezultatih opazili celo občutno poslabšanje (to velja še posebej za kriterij LRc). Vzrok za neuspešnost normalizacije rezultatov prileganja bi se po našem mnenju lahko skrival v dejstvu, da porazdelitve rezultatov prileganja precej odstopajo od normalne porazdelitve. Mogoče bi lahko na te porazdelitve vplivali tako, da posnetkov, ki jih uporabimo za normalizacijo rezultatov prileganja, ne bi izbrali naključno, ampak bi jih izbrali na bolj selektiven način.

[46] V tem pogledu se obnašamo povsem nediskriminatorno.
[47] Z drugimi besedami, cilj posvečuje sredstva.

ZAKLJUČEK

9

V disertaciji smo se ukvarjali s problemom samodejnega razpoznavanja govorcev, ki ga lahko definiramo na naslednji način: Denimo, da imamo dva govorna posnetka. Zanima nas ali je oba posnetka izgovoril isti govorec? V splošnem je besedilo v obeh posnetkih različno. Takrat govorimo o besedilno neodvisnemu razpoznavanju, ki pride v poštev predvsem v aplikacijah s t.i. nesodelujočim govorcem, kjer se govorec ponavadi ne zaveda, da je udeležen v postopku razpoznavanja. Tehnologija razpoznavanja govorcev je uporabna v številnih aplikacijah. Izpostavimo jih le nekaj: sledenje telefonskim pogovorom, forenzične preiskave in po našem mnenju najzanimivejša — iskanje po avdio in video vsebinah.

Čeprav vemo, da se identiteta govorca v govornem signalu ›skriva‹ na različnih nivojih, smo se v disertaciji omejili le na najnižji, t.j. akustični nivo. Izkazalo se je namreč, da dosežejo sistemi, ki temeljijo na akustičnem nivoju, daleč boljše rezultate od tistih, ki temeljijo na višjenivojski informaciji.

Problem samodejnega razpoznavanja govorcev ponuja z znanstvenega stališča številne izzive. Eden izmed ključnih izzivov, ki je bil v disertaciji v središču naše pozornosti, je problem sejne spremenljivosti. Pojem sejne spremenljivosti se nanaša na vse spremembe, ki zaradi kakršnegakoli vzroka nastanejo med različnimi posnetki istega govorca. Čeprav imamo največkrat v mislih kanalsko spremenljivost, do katere pride zaradi uporabe različnih vrst mikrofonov, prenosnih poti in različnih akustičnih razmer, pa imajo vzroki za sejno spremenljivost lahko tudi bolj neotipljivo ozadje. Tako lahko k vzrokom za sejno spremenljivost štejemo tudi psihofizično stanje govorca, vpliv staranja ter tudi izgovorjeno besedilo.

Problem samodejnega razpoznavanja govorcev smo obravnavali v okviru statistične teorije. Glavni gradnik, na katerem je temeljila večina nadaljnih postopkov, s katerimi smo se ukvarjali v disertaciji, je bil model mešanice Gaussovih porazdelitev. Ta enostaven, a hkrati dovolj prožen model, s katerim lahko opišemo tako rekoč poljubno funkcijo gostote verjetnosti, nam je omogočil, da smo iz velikega števila govornih posnetkov velikega števila različnih govorcev ocenili splošni model govorca. Ta splošni model govorca smo uporabili na dva načina: (i) služil nam je za izpeljavo modelov posameznih govorcev in (ii) uporabili smo ga za normalizacijo verjetja v imenovalcu izraza za razmerje verjetij.

Model mešanice Gaussovih porazdelitev smo nadgradili tako, da smo iz opazovanja porazdeljevanja vektorjev značilk v akustičnem prostoru prešli na opazovanje porazdeljevanja parametrov modela v prostoru parametrov. Pokazali smo, da lahko povprečne vektorje modela mešanice obravnavamo kot supervektorje. Predpostavili smo, da lahko supervektor razstavimo na vsoto: (i) govorske komponente in (ii) kanalske komponente. Ugotovili smo, da lahko ti dve komponenti ločimo, če predpostavimo, da je kanalska komponenta omejena na nizkorazsežen podprostor supervektorskega prostora. Podrobno smo obdelali dva postopka za normalizacijo

sejne variabilnosti, ki slonita na tej predpostavki. To sta postopek *projekcije motečih lastnosti* in *analiza vezanih faktorjev*. Večji poudarek smo namenili slednji, ki je splošnejša od obeh metod. Pokazali smo, da je analiza vezanih faktorjev hierarhični model, ki vključuje model mešanice Gaussovih porazdelitev in faktorsko analizo. Predstavitev modela analize vezanih faktorjev v obliki verjetnostnega grafičnega modela nam je omogočila, da smo izpeljali iterativni postopek za oceno porazdelitve prikritih spremenljivk, ki nastopajo v modelu.

Velik poudarek smo namenili primerjavi različnih mer podobnosti, ki so primerne v sistemih za samodejno razpoznavanje govorcev, s katerimi smo se ukvarjali v disertaciji. Najnaravnejša mera podobnosti statističnih modelov je verjetje, katerega izračun je v primeru večjega števila komponent v modelu mešanice Gaussovih porazdelitev časovno precej požrešno opravilo. Ponavadi se zato namesto natančne vrednosti z metodo najverjetnejših komponent izračuna približno vrednost verjetja. Sami smo predlagali še učinkovitejšo aproksimativno metodo, pri kateri izračunamo verjetje na podlagi zadostne statistike. Poleg verjetja smo predlagali še mero podobnosti, ki temelji na merjenju razdalje med porazdelitvami. V navezi s postopkom projekcije motečih lastnosti smo podali postopek za izračun verjetja, za metodo analize vezanih faktorjev pa smo predlagali mero podobnosti, ki temelji na metodi podpornih vektorjev.

Pri eksperimentalnem delu smo uporabljali govorne zbirke iz preteklih NIST-ovih vrednotenj sistemov za samodejno razpoznavanje govorcev. Natančneje, v razvojne namene smo uporabili zbirke iz let 2004, 2005 in 2006, eksperimente pa smo izvajali na telefonskem delu zbirke iz leta 2008. Sistem, ki smo ga preizkušali, smo zasnovali na podlagi nekaterih odločitev, ki se razlikujejo od uveljavljene prakse s področja razpoznavanja govorcev. Običajno je, da se ženske in moške govorce obravnava ločeno in se posledično razvije dva ločena sistema, mi pa med ženskami in moškimi nismo delali razlik, ampak smo vse dele sistema zasnovali od spola neodvisno. Druga pomembna razlika do večine sistemov, predstavljenih v literaturi, je, da smo uporabljali nenormalizirane značilke. Za takšen korak smo se odločili, ker vemo, da normalizacija iz značilk vedno odstrani tudi nekaj koristne informacije. Predvidevali smo, da bomo normalizacijo značilk lahko nadomestili z normalizacijo na nivoju modela, bodisi v postopku projekcije motečih lastnosti bodisi v postopku analize vezanih faktorjev.

Dobljeni rezultati pričajo, da je uporaba enega izmed postopkov normalizacije na nivoju modela ključna za dosego dobrih rezultatov. Izkazalo se je, da je odločitveni kriterij, ki temelji na razmerju verjetij, precej občutljiv na nenormalizirane značilke. Metoda razmerja verjetij je dala vzpodbudne rezultate šele v primeru, ko smo v izraz za izračun razmerja verjetij vključili nedoločenost testnega kanala. Nasprotno se je metoda SVM izkazala za izredno robustno in v veliki meri neobčutljivo na nenormalizirane značilke, saj smo z njo konsistentno dosegali najboljše rezultate. Po našem mnenju lahko učinkovitost metode SVM pripišemo njeni diskriminatorni naravi. Z metodami za normalizacijo rezultatov razpoznavanja smo rezultate uspeli izboljšati le v redkih primerih, medtem ko smo pri nekaj postopkih rezultate celo poslabšali. Na občutno poslabšanje rezultatov smo naleteli pri metodi JFA-LRc, kar

lahko po našem mnenju pripišemo naši od govorca neodvisni drži in nenormaliziranim značilkam. Tukaj bi bilo vredno poskusiti z bolj selektivno izbranimi posnetki za normalizacijo rezultatov prileganja.

Čeprav smo se v disertaciji posvečali samodejnemu razpoznavanju govorcev, smo pri tem naleteli na probleme, ki so skupni številnim problemom s področja razpoznavanja vzorcev in strojnega učenja. Po našem mnenju je ključna lastnost vsakega razpoznavalnika vzorcev sposobnost posploševanja. Posploševanje relativno enostavno dosežemo, kadar imamo na voljo veliko podatkov za učenje. Kadar pa so podatki redki (angl. sparse), je posploševanje težje doseči — pojavi se problem prenaučitve. Prenaučitve se lahko izognemo le tako, da v proces učenja vključimo apriorno znanje, kar lahko storimo na več načinov. Konceptualno še posebej eleganten način za vključitev apriornega znanja nam ponuja bayesovska interpretacija verjetnosti, ki proces ocenjevanja parametrov statističnega modela ›vidi‹ v luči manjšanja nedoločenosti porazdelitve parametrov.

Brez zanašanja na tako podano apriorno znanje bi imeli težave pri ocenjevanju parametrov govorskih modelov, za oceno katerih imamo na voljo le omejeno količino podatkov. V disertaciji smo preizkušali dve metodi, s katerima smo ocenili parametre modela na osnovi predhodnega znanja, ki smo ga pridobili iz obsežne govorne zbirke. Najprej smo predstavili enostavnejšo metodo maksimizacije vrha posteriorne porazdelitve, kasneje pa še splošnejši postopek analize vezanih faktorjev.

V disertaciji smo uporabljali le majhen nabor vseh orodij, ki nam jih ponuja bayesovska interpretacija verjetnosti. V veliki meri je temu vzrok velika zahtevnost obravnavanega problema s stališča računalniških kapacitet, ki jih pri reševanju problema potrebujemo. Čeprav je lahko neka metoda konceptualno in teoretično elegantna, pa je mogoče zato njena implementacija vse prej kot trivialna. V teoriji je bilo predlaganih že veliko obetavnih metod, ki dajejo vzpodbudne rezultate na podatkovno in računsko manj zahtevnih problemih, in bi jih bilo zanimivo preizkusiti tudi na področju razpoznavanja govorcev, a so zaenkrat še prenerodne za prenos na tako kompleksne probleme. To se, upamo, utegne že kmalu spremeniti. Vse več se namreč pojavlja učinkovitih orodij[48], ki omogočajo relativno enostavno uporabo splošnih postopkov učenja v verjetnostnih grafičnih modelih.

V zadnjem času potekajo še posebej živahne raziskave na področju *neparametričnih bayesovskih metod* (angl. nonparametric bayesian), ki utegnejo po našem mnenju v bodoče postreči z odgovori na nekatera vprašanja, na katera obstoječe metode ne znajo ponuditi zadovoljivih odgovorov. Vendar bo na to priložnost potrebno počakati še nekaj časa.

[48] Eno izmed takih orodij je *Infer.NET*, razvito v Microsoftovem razvojnem oddelku v Cambridgeu in dostopno na naslovu http://research.microsoft.com/en-us/um/cambridge/projects/infernet.

BIBLIOGRAFIJA

1 Au, C. in Tam, J. (1999). Transforming Variables Using the Dirac Generalized Function. *The American Statistician, 53*(3), 270–272.

2 Auckenthaler, R., Carey, M. in Lloyd-Thomas, H. (2000). Score normalization for text-independent speaker verification systems. *Digital Signal Processing, 10*(1–3), 42–54.

3 Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G. in Magrin-Chagnolleau, I. et al. (2004). A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing, 2004*(1), 430–451.

4 Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. 2 izdaja Springer.

5 Brandschain, L., Cieri, C., Graff, D., Neely, A. in Walker, K. (2008). Speaker Recognition: Building the Mixer 4 and 5 Corpora. V *Proc. LREC*, str. 3551–3554. Marrakech, Morocco.

6 Brummer, N., Burget, L., Cernocky, J., Glembek, O. in Grezl, F. et al. (2007). Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. *Audio, Speech, and Language Processing, IEEE Transactions on, 15*(7), 2072-2084.

7 Burges, C. J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery, 2*, 121–167.

8 Burget, L., Matejka, P., Hubeika, V. in Cernocky, J. (2009). Investigation into variants of Joint Factor Analysis for speaker recognition. V *Proc. Interspeech 2009*, str. 1263–1266.

9 Campbell, W., Campbell, J., Gleason, R., Reynolds, D. in Shen, W. (2007). Speaker verification using support vector machines and high-level features. *IEEE Transactions On Audio, Speech, and Language Processing, 15*(7), 2085–2094.

10 Campbell, W., Campbell, J., Reynolds, D., Singer, E. in Torres-Carrasquillo, P. (2006a). Support vector machines for speaker and language recognition. *Computer, Speech and Language, 2–3*(20), 210–229.

11 Campbell, W. M., Sturim, D. E. in Reynolds, D. A. (2006a). Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters, 13*(5), 308–311.

12 Chang, C.-C. in Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. . Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

13 Cieri, C., Corson, L., Graff, D. in Walker, K. (2007). Resources for New Research Directions in Speaker Recognition: the Mixer 3, 4 and 5 Corpora. V *Proc. Interspeech*, str. 950–953. Antwerp, Belgium.

14 de Finetti, B. (1974). *Theory of Probability*, volume 1. New York: John Wiley and Sons.

15 de Finetti, B. (1975). *Theory of Probability*, volume 2. New York: John Wiley and Sons.

16 Dempster, A. P., Laird, N. M. in Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1–38.

17 Do, M. N. (2003). Fast Approximation of Kullback-Leibler Distance for Dependence Trees and Hidden Markov Models. *IEEE Signal Processing Letters, 10*(4), 115–118.

18 Fox, E. B. (2009). *Bayesian Nonparametric Learning of Complex Dynamical Phenomena*. PhD thesis, Massachusetts Institute of Technology.

19 Gauvain, J.-L. in Lee, C.-H. (1994). Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing, 2*(2), 291–298.

20 Ghahramani, Z. in Hinton, G. E. (1996). The EM Algorithm for Mixtures of Factor Analyzers. Technical Report Department of Computer Science, University of Toronto.

21 Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America, 87*(4), 1738–1752.

22 Hermansky, H. (1998). Should recognizers have ears?. *Speech Communication, 25*(1–3), 3–27.

23 Hermansky, H. in Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing, 2*(4), 578-589.

24 Hershey, J. in Olsen, P. (2007). Approximating the kullback-leibler divergence between gaussian mixture models. V *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07)*, str. 317-320.

25 Imai, S. (1983). Cepstral analysis synthesis on the mel frequency scale. V *Proc. ICASSP*, str. 93–96.

26 Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

27 Kenny, P. (2005). Joint factor analysis of speaker and session variability : Theory and algorithms. Technical Report Centre de recherche informatique de Montreal.

28 Kenny, P., Boulianne, G. in Dumouchel, P. (2005). Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing, 13*(3), 345–354.

29 Kenny, P., Boulianne, G., Ouellet, P. in Dumouchel, P. (2004). Speaker adaptation using an eigenphone basis. *IEEE Transactions on Speech and Audio Processing, 12*(6), 579–589.

30 Kenny, P., Boulianne, G., Ouellet, P. in Dumouchel, P. (2006). The Geometry of the Channel Space in GMM-Based Speaker Recognition. V *Proc. IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, str. 1–5.

31 Kenny, P., Boulianne, G., Ouellet, P. in Dumouchel, P. (2007a). Joint Factor Analysis versus Eigenchannels in Speaker Recognition. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(4), 1435–1447.

32 Kenny, P., Boulianne, G., Ouellet, P. in Dumouchel, P. (2007b). Speaker and Session Variability in GMM-Based Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(4), 1448–1460.

33 Kenny, P., Ouellet, P., Dehak, N., Gupta, V. in Dumouchel, P. (2008). A study of interspeaker variability in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing, 16*(5), 980–988.

34 Kinnunen, T. in Li, H. (2009). An Overview of Text-Independent Speaker Recognition: from Features to Supervectors. *Speech Communication*.

35 Leeuwen, D., Martin, A., Przybocki, M. in Bouten, J. (2006). NIST and NFI-TNO evaluations of automatic speaker recognition. *Computer Speech and Language, 20*(2–3), 128–158.

36 Leggetter, C. in Woodland, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer, Speech and Language, 9*(2), 171–185.

37 Linde, Y., Buzo, A. in Gray, R. (1980). An Algorithm for Vector Quantizer Design. *IEEE Transactions on Communications, 28*(1), 84–95.

38 MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. 2 izdaja Cambridge University Press.

39 Martin, A., Doddington, G., Kamm, T., Ordowski, M. in Przybocki, M. (1997). The DET Curve In Assessment Of Detection Task Performance. V *Proc. Eurospeech*, str. 1895–1898.

40 Martin, A. F. in Greenberg, C. S. (2009). NIST 2008 Speaker Recognition Evaluation: Performance Across Telephone and Room Microphone Channels. V *Proc. Interspeech*, str. 2579–2582. Brighton, UK.

41 Mclaren, M., Vogt, R., Baker, B. in Sridharan, S. (2009). Data-driven impostor selection for t-norm score normalisation and the background dataset in svm-based speaker verification. V *ICB '09: Proceedings of the Third International Conference on Advances in Biometrics*, str. 474–483. Berlin, Heidelberg.

42 Nist (2008). *The NIST Year 2008 Speaker Recognition Evaluation Plan.* . Available at http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan/release4.pdf.

43 O'Hagan, T. (2004). Dicing with the unknown. *Significance, 1*(3), 132–133.

44 Ouellet P. Boulianne, G. in Kenny, P. (2005). Flavors of gaussian warping. V *Proc. of European Conference on Speech Communication and Technology (Interspeech '05)*, str. 2957–2960. Lisboa, Portugal.

45 Pelecanos, J. in Sridharan, S. (2001). Feature warping for robust speaker verification. V *Proc. Odyssey*, str. 213–218. Crete, Greece.

46 Pols, L. C. W. (1977). *Spectral Analysis and Identification of Dutch Vowels in Monosyllabic Words*. PhD thesis, Amsterdam: Vrije Universiteit te Amsterdam.

47 Przybocki, M., Martin, A. in Le, A. (2007). NIST Speaker Recognition Evaluations Utilizing the Mixer Corpora — 2004, 2005, 2006. *IEEE Trans. Audio, Speech Language Process., 15*(7), 1951–1959.

48 Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

49 Reynolds, D. (1997). Comparison of background normalization methods for text-independent speaker verification. V *Proc. Eurospeech*, str. 963–966.

50 Reynolds, D., Andrews, W., Campbell, J., Navratil, J. in Peskin, B. et al. (2003b). The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition. V *Proc. IEEE ICASSP*, str. 784–787.

51 Reynolds, D. A. (2003). Channel Robust Speaker Verification via Feature Mapping. V *Proc. ICASSP*, str. 53–56.

52 Reynolds, D. A., Quatieri, T. F. in Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models,. *Digital Signal Processing, 10*(1–3), 19–41.

53 Rubin, D. in Thayer, D. (1983). EM Algorithms for ML Factor Analysis. *Psychometrika, 47*(1), 69–76.

54 Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A. in Stolcke, A. (2005). Speaker verification using support vector machines and high-level features. *Speech Communication, 46*(3–4), 455–472.

55 Solomonoff A., C. W. B. I. (2005). Advances in channel compensation for SVM speaker recognition. V *Proc. ICASSP*, str. 629–632. Philadelphia, USA.

56 Stolcke, A., Kajarekar, S. in Ferrer, L. (2008). Nonparametric Feature Normalization for SVM-based Speaker Verification. V *Proc. IEEE ICASSP*, str. 1577–1580.

57 Stolcke, A., Kajarekar, S., Ferrer, L. in Shriberg, E. (2009). Speaker recognition with session variability normalization based on MLLR adaptation transforms. *IEEE Transactions on Audio, Speech and Language Processing, 7*(15).

58 Strang, G. (2009). *Introduction to Linear Algebra*. 4 izdaja Wellesley-Cambridge Press.

59 Teunen, R., Shahshahani, B. in Heck, L. (2000). A model-based transformational approach to robust speaker recognition. V *Proc. ICSLP*, str. 495–498.

60 Tokuda, K., Kobayashi, T. in Imai, S. (1995a). Adaptive cepstral analysis of speech. *IEEE Transactions on Speech and Audio Processing, 3*(6), 481–489.

61 Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T. in Imai, S. (1995b). An Algorithm for Speech Parameter Generation from Continuous Mixture HMMs with Dynamic Features. V *Proc. EUROSPEECH*, str. 757–760.

62 Tranter, S. E. in Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(5), 1557–1565.

63 Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc..

64 Vesnicer, B. in Mihelič, F. (2004). Evaluation of the Slovenian HMM-based speech synthesis system. V *Proc. Text, Speech and Dialog*, str. 513–520.

65 Vesnicer, B. in Mihelič, F. (2008). The likelihood ratio decision criterion for nuisance attribute projection in GMM speaker verification. *EURASIP J. Appl. Signal Process., 2008*(3), 1–11.

66 Štruc, V., Vesnicer, B., Mihelič, F. in Pavešić, N. (2009). Nuisance attribute projection in the logarithm domain for face recognition under severe illumination changes. V *Proc. ERK*, str. 279–282.

67 Žibert, J., Pavešić, N. in Mihelič, F. (2006). Speech/non-speech segmentation based on phoneme recognition features. *EURASIP Journal on Applied Signal Processing, 2006*.

68 Xiang, B., Chaudhari, U., Navratil, J., Ramaswamy, G. in Gopinath, R. (2002). Short-time gaussianization for robust speaker verification. V *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '02)*, str. 681–684.

69 Young, S., Evermann, G., Gales, M., Hain, T. in Kershaw, D. et al. (2009). *The HTK Book (for HTK Version 3.4)*. . Available at `http://htk.eng.cam.ac.uk`.

70 Zen, H., Tokuda, K. in Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication, 51*(11), 1039–1064.

71 Zen, H., Tokuda, K. in Kitamura, T. (2006). Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. *Comput. Speech Language, 21*(1), 153–173.

POSTOPEK MAKSIMIZACIJE UPANJA A

Postopek maksimizacije upanja (angl. expectation maximization, EM) je splošni optimizacijski postopek za določitev vrednosti parametrov verjetnostnega modela po kriteriju največjega verjetja. Postopek je primeren za verjetnostne modele, ki vsebujejo prikrite spremenljivke. Pomembna lastnost postopka EM je, da ima zagotovljeno konvergenco k lokalnemu maksimumu (Dempster et al., 1977).

Imejmo verjetnosti model, v katerem nastopajo opažene in prikrite spremenljivke. Vse opažene spremenljivke združimo v množico X, vse prikrite pa v množico Z. Vezana porazdelitev videnih in prikritih spremenljivk naj bo parametrična — množico parametrov označimo s črko Θ. Naš cilj bo poiskati takšne vrednosti parametrov $\theta \in \Theta$, pri katerem bo dosegla funkcija verjetja največjo vrednost. Funkcijo verjetja zapišemo z izrazom

$$p(x|\theta) = \sum_z p(x, z|\theta),$$

kjer smo predpostavili, da je naključna spremenljivka Z diskretna, čeprav bi enako veljalo tudi v primeru, ko bi bila Z zvezna, le operacijo seštevanja bi morali na ustreznih mestih zamenjati z operacijo integriranja.

Predpostavimo, da je neposredna optimizacija verjetja $p(x|\theta)$ težavna, optimizacija verjetja polnega nabora podatkov $p(x, z|\theta)$ pa veliko enostavnejša. Če vpeljemo porazdelitev $q(z)$ prikritih spremenljivk, lahko zapišemo:

$$\log p(x|\theta) = \mathcal{L}(q, \theta) + \mathbb{D}_{\mathrm{KL}}[q \,||\, p], \tag{A.1}$$

kjer je:

$$\mathcal{L}(q, \theta) = \sum_z q(z) \log \frac{p(x, z|\theta)}{q(z)} \tag{A.2}$$

$$\mathbb{D}_{\mathrm{KL}}[q \,||\, p] = -\sum_z q(z) \log \frac{p(z|x, \theta)}{q(z)}. \tag{A.3}$$

Iz enačbe (**A.3**) vidimo, da ustreza izraz $\mathbb{D}_{\mathrm{KL}}[q \,||\, p]$ divergenci Kullbacka in Leiblerja med porazdelitvijo $q(z)$ in posteriorno porazdelitvijo $p(z|x, \theta)$. Ker vemo, da je divergenca KL nenegativna — nič bo le v primeru, ko $q(z) = p(z|x, \theta)$ — sledi, da je $\mathcal{L}(q, \theta) \leq p(x|\theta)$, z drugimi besedami $\mathcal{L}(q, \theta)$ je spodnja meja verjetja.

Do izraza (**A.2**) za spodnjo mejo verjetja lahko pridemo tudi tako, da zapišemo:

$$\begin{aligned}\log p(x|\theta) &= \log \sum_z p(x,z|\theta) \\ &= \log \sum_z q(z) \frac{p(x,z|\theta)}{q(z)} \\ &\geq \sum_z q(z) \log \frac{p(x,z|\theta)}{q(z)} \\ &= \mathcal{L}(q,\theta),\end{aligned}$$

kjer neenakost sledi iz konkavnosti funkcije logaritma (neenakost je znana pod imenom Jensenova neenakost.)

Dekompozicija (**A.1**) je ključna za izpeljavo postopka EM. Označimo s $\theta^{(t-1)}$ trenutne vrednosti parametrov. V e-koraku poiščemo tak $q^{(t)}(z)$, pri katerem bo vrednost $\mathcal{L}(q,\theta^{(t-1)})$ maksimalna. Očitno je, da se bo to zgodilo takrat, ko bo $q(z)$ enaka posteriorni porazdelitvi $p(z|x,\theta)$. V m-koraku fiksiramo $q^{(t)}(z)$ in poiščemo tak $\theta^{(t)}$, pri katerem bo $\mathcal{L}(q^{(t)},\theta)$ maksimalna. Na ta način bomo povečali vrednost izraza $\mathcal{L}$, razen takrat, ko se bomo že nahajali v lokalnem maksimumu.

Če v enačbo (**A.2**) vstavimo $q(z) = p(z|x,\theta^{(t-1)})$, dobimo:

$$\begin{aligned}\mathcal{L}(q,\theta) &= \sum_z p(z|x,\theta^{(t-1)}) \log p(x,z|\theta) - \sum_z p(z|x,\theta^{(t-1)}) \log p(z|x,\theta^{(t-1)}) \\ &= Q(\theta,\theta^{(t-1)}) + \mathbb{H}[q],\end{aligned}$$

kjer je $\mathbb{H}[q]$ entropija porazdelitve q, ki ni odvisna od parametrov θ.

Izraz $Q(\theta,\theta^{(t-1)})$, katerega maksimum iščemo v m-koraku, je enak matematičnemu upanju verjetja polnega nabora podatkov.

Postopek EM lahko skoraj brez spremembe uporabimo tudi v primeru, ko želimo namesto ML ocene poiskati MAP oceno parametrov. Prav tako obstaja tudi bayesovska posplošitev postopka EM (Bishop (2007), poglavje 10).

FAKTORSKA ANALIZA

B

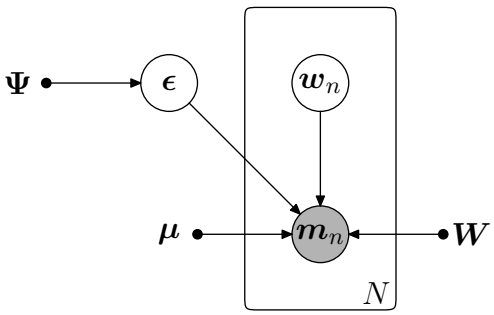
Na tem mestu obdelajmo teorijo faktorske analize. Enačba faktorske analize ima naslednjo obliko:

$$\boldsymbol{m} = \boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{w} + \boldsymbol{\epsilon}. \tag{B.1}$$

V enačbi nastopajo naključne spremenljivke $\boldsymbol{m}$, $\boldsymbol{w}$ in $\boldsymbol{\epsilon}$ ter parametra $\boldsymbol{\mu}$ in $\boldsymbol{W}$. Njihov pomen je sledeč:

- $\boldsymbol{m}$, opažena spremenljivka;
- $\boldsymbol{\mu}$, povprečna vrednost opažene spremenljivke;
- $\boldsymbol{W}$, matrika faktorskih uteži;
- $\boldsymbol{w}$, prikrita spremenljivka porazdeljena $\mathcal{N}(\cdot\,|\mathbf{0}, \boldsymbol{I})$;
- $\boldsymbol{\epsilon}$, neodvisen $\mathcal{N}(\cdot\,|\mathbf{0}, \boldsymbol{\Psi})$ porazdeljen šum.

Predpostavljene statistične neodvisnosti med posameznimi naključnimi spremenljivkami lahko nazorno predstavimo v obliki grafičnega modela na sliki B.1.



Slika B.1 Grafični model faktorske analize.

V skladu s predpostavkami lahko vezano porazdelitev verjetnosti $p(\boldsymbol{m}, \boldsymbol{w}, \boldsymbol{\epsilon})$ faktoriziramo v:

$$p(\boldsymbol{m}, \boldsymbol{w}, \boldsymbol{\epsilon}) = p(\boldsymbol{m}|\boldsymbol{w}, \boldsymbol{\epsilon})p(\boldsymbol{w})p(\boldsymbol{\epsilon}).$$

Od tod sledi, da je pogojna porazdelitev $p(\boldsymbol{m}, \boldsymbol{\epsilon}|\boldsymbol{w})$ enaka:

$$p(\boldsymbol{m}, \boldsymbol{\epsilon}|\boldsymbol{w}) = \frac{p(\boldsymbol{m}|\boldsymbol{w}, \boldsymbol{\epsilon})p(\boldsymbol{w})p(\boldsymbol{\epsilon})}{p(\boldsymbol{w})} = p(\boldsymbol{m}|\boldsymbol{w}, \boldsymbol{\epsilon})p(\boldsymbol{\epsilon}).$$

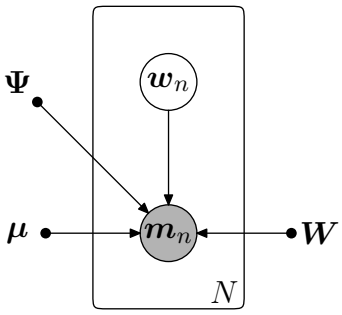
Robno pogojno porazdelitev $p(\boldsymbol{m}|\boldsymbol{w})$ dobimo tako, da iz zgornje enačbe izintegriramo spremenljivko $\boldsymbol{\epsilon}$:

$$p(\boldsymbol{m}|\boldsymbol{w}) = \int p(\boldsymbol{m}|\boldsymbol{w}, \boldsymbol{\epsilon})p(\boldsymbol{\epsilon})d\boldsymbol{\epsilon}, \tag{B.2}$$

kar da[49] pričakovan rezultat

$$p(\boldsymbol{m}|\boldsymbol{w}) = \mathcal{N}(\boldsymbol{m}|\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{w}, \boldsymbol{\Psi}). \tag{B.3}$$

Z integriranjem smo izločili odvisnost od spremenljivke $\boldsymbol{\epsilon}$, ki nas ponavadi ne zanima. Sedaj izgleda grafični model nekoliko drugače (slika **B.2**):



Slika B.2 Grafični model faktorske analize po tem, ko smo izločili spremenljivko $\boldsymbol{\epsilon}$.

Robno porazdelitev opažene spremenljivke dobimo tako, da integriramo vezano porazdelitev $p(\boldsymbol{m}, \boldsymbol{w}) = p(\boldsymbol{m}|\boldsymbol{w})p(\boldsymbol{w})$:

$$p(\boldsymbol{m}) = \int p(\boldsymbol{m}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} = \mathcal{N}(\boldsymbol{m}|\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}} + \boldsymbol{\Psi}).$$

Zanima nas še pogojna porazdelitev prikrite naključne spremenljivke $\boldsymbol{w}$ pri dani opaženi spremenljivki $\boldsymbol{m}$, ki jo dobim s pomočjo Bayesove formule:

$$p(\boldsymbol{w}|\boldsymbol{m}) = \frac{p(\boldsymbol{m}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{m})} = \mathcal{N}(\boldsymbol{w}|\boldsymbol{L}^{-1}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}(\boldsymbol{m} - \boldsymbol{\mu}), \boldsymbol{L}^{-1}),$$

kjer smo zaradi krajšega zapisa vpeljali $\boldsymbol{L} = \boldsymbol{I} + \boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{W}$.

B.1 Ocenjevanje parametrov FA po kriteriju ML

Kot smo že navajeni, nas spet zanima, kako iz podatkov $\mathcal{D} = \{\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N\}$ oceniti parametre $\boldsymbol{\mu}$, $\boldsymbol{W}$ in $\boldsymbol{\Psi}$, ki nastopajo v faktorskem modelu. Žal za razliko od podobnega problema *analize glavnih komponent* (angl. principal component analysis,

[49] Integral na desni strani enačbe **B.2** je bolj zanimiv kot se mogoče zdi na prvi pogled, saj se da hitro videti, da je porazdelitev $p(\boldsymbol{m}|\boldsymbol{w}, \boldsymbol{\epsilon})$ ›točkasta‹ in gre v bistvu za konvolucijo porazdelitve $p(\boldsymbol{\epsilon})$ z Diracovim impulzom $\delta(\boldsymbol{m} - \boldsymbol{W}\boldsymbol{w} - \boldsymbol{\epsilon})$. Pazljiv bralec bo ugotovil, da lahko do enake rešitve pridemo še po drugačni poti — z uporabo izreka o zamenjavi spremenljivk neposredno iz enačbe (**B.1**), kar nakazuje na uporabnost Diracove posplošene funkcije kot alternativno tehniko pri določanju porazdelitvene funkcije transformiranih spremenljivk (Au in Tam, 1999).

PCA) rešitve ne znamo poiskati po analitični poti, zato nam preostane le iterativni postopek — postopek EM, ki ga je za faktorsko analizo izpeljal že (Rubin in Thayer, 1983).

Algoritem EM za faktorsko analizo lahko izpeljemo tako, da sledimo navodilom splošnega okvira postopka EM. Tako najprej zapišemo logaritem verjetja popolnega nabora podatkov in izpeljemo izraz za pričakovano vrednost tega verjetja glede na posteriorno porazdelitev prikrite spremenljivke z uporabo ›starih‹ vrednosti parametrov. Nove vrednosti parametrov nato izračunamo tako, da poiščemo maksimum tega izraza.

Ker predpostavljamo, da so podatki $\mathcal{D}$ med seboj neodvisni, lahko logaritem verjetja popolnega nabora podatkov zapišemo kot:

$$\log p(\{\boldsymbol{m}_n\}, \{\boldsymbol{w}_n\}|\boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{\Psi}) = \sum_{n=1}^{N} \log p(\boldsymbol{m}_n|\boldsymbol{w}_n) + \log p(\boldsymbol{w}_n).$$

Ker $p(\boldsymbol{w}_n)$ ni odvisen od parametrov, ga lahko izpustimo iz obravnave in se posvetimo le prvemu členu. Z upoštevanjem rezultata (**B.3**) lahko zapišemo:

$$\begin{aligned}\log p(\{\boldsymbol{m}_n\}|\{\boldsymbol{w}_n\}, \boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{\Psi}) = &-\frac{1}{2}\sum_{i=1}^{N} D\log\pi + |\boldsymbol{\Psi}|^{1/2} \\ &-\frac{1}{2}\sum_{i=1}^{N}\left(\boldsymbol{m}_n - \boldsymbol{W}\boldsymbol{w}_n\right)^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\left(\boldsymbol{m}_n - \boldsymbol{W}\boldsymbol{w}_n\right).\end{aligned}$$

Ko izpeljemo izraz za matematično upanje gornjega izraza glede na porazdelitev $p(\boldsymbol{w_n}|\boldsymbol{m}_n, \boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{\Psi})$, nam preostane le še, da poiščemo maksimum verjetja z odvajanjem po posameznih parametrih. Vse to se da strniti v naslednji postopek EM. V e-koraku z upoštevanjem starih vrednosti parametrov izračunamo zadostno statistiko prvega in drugega reda:

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{w}_n|\boldsymbol{m}_n\right] &= \boldsymbol{L}_{t-1}^{-1}\boldsymbol{W}_{t-1}^{T}\boldsymbol{\Psi}_{t-1}^{-1}(\boldsymbol{m}_n - \overline{\boldsymbol{m}}) \\ \mathbb{E}\left[\boldsymbol{w}_n\boldsymbol{w}_n^{T}|\boldsymbol{m}_n\right] &= \boldsymbol{L}_{t-1}^{-1} + \mathbb{E}[\boldsymbol{w}_n|\boldsymbol{m}_n]\mathbb{E}[\boldsymbol{w}_n|\boldsymbol{m}_n]^{T},\end{aligned}$$

kjer smo z $\overline{\boldsymbol{m}}$ označili vzorčno povprečje: $\overline{\boldsymbol{m}} = \frac{1}{N}\sum_n \boldsymbol{m}_n$. V m-koraku izračunamo nove ocene parametrov:

$$\begin{aligned}\boldsymbol{\mu}_t &= \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{m}_n \\ \boldsymbol{W}_t &= \left[\sum_{n=1}^{N}(\boldsymbol{m}_n - \overline{\boldsymbol{m}})\mathbb{E}[\boldsymbol{w}_n]^{T}\right]\left[\sum_{n=1}^{N}\mathbb{E}\left[\boldsymbol{w}_n\boldsymbol{w}_n^{T}|\boldsymbol{m}_n\right]\right]^{-1} \\ \boldsymbol{\Psi}_t &= \frac{1}{N}\operatorname{diag}\left[\sum_{n=1}^{N}\boldsymbol{m}_n\boldsymbol{m}_n^{T} - \boldsymbol{W}_t\sum_{i=1}^{N}\mathbb{E}[\boldsymbol{w}_n|\boldsymbol{m}_n](\boldsymbol{m}_n - \overline{\boldsymbol{x}})^{T}\right]\end{aligned}$$

Opazimo lahko, da se v skladu s pričakovanjem povprečna vrednost $\boldsymbol{\mu}_t$ iz iteracije v iteracijo ne spreminja in je enaka vzorčnemu povprečju.

B.2 Povezava z analizo glavnih komponent

Pokazati se da (Bishop, 2007, poglavje 12), da je verjetnostna analiza glavnih komponent (angl. probabilistic component analysis, PPCA) le poseben primer faktorske analize. Faktorska analiza preide v PPCA takrat, ko imajo komponente spremenljivke $\boldsymbol{\epsilon}$ enake variance: $\boldsymbol{\Psi} = \sigma \boldsymbol{I}$ (takrat govorimo o izotropnem šumu). V primeru, ko spremenljivko $\boldsymbol{\epsilon}$ iz modela popolnoma izpustimo: $\boldsymbol{\Psi} = \mathbf{0}$, pa verjetnostni PCA preide v navadni PCA.

VARIACIJSKA OBRAVNAVA MODELA JFA

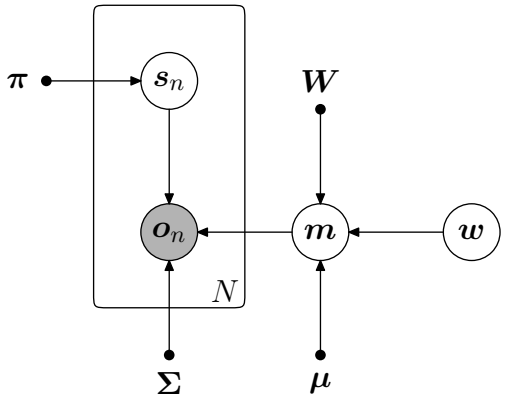
C

V poglavju 7 smo obravnavali model analize vezanih faktorjev. V tem modelu nastopajo štiri vrste spremenljivk:

- opažene spremenljivke: $o = \{\boldsymbol{o}_n\}_{n=1}^N$;
- prikrite spremenljivke: $s = \{\boldsymbol{s}_n\}_{n=1}^N$;
- parametri: $\boldsymbol{m}$, $\boldsymbol{x}$, $\boldsymbol{y}$ in $\boldsymbol{z}$;
- hiperparametri: $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{U}$, $\boldsymbol{V}$ in $\boldsymbol{D}$.

Zaradi lažje obravnave združimo spremenljivke $\boldsymbol{x}$, $\boldsymbol{y}$ in $\boldsymbol{z}$ v spremenljivko $\boldsymbol{w}$, matrike $\boldsymbol{U}$, $\boldsymbol{V}$ in $\boldsymbol{D}$ pa v matriko $\boldsymbol{W}$ (glej razdelek 7.2).

Statistične odvisnosti med posameznimi spremenljivkami lahko nazorno predstavimo v obliki grafičnega modela (slika C.1).



Slika C.1 Usmerjeni grafični model modela JFA. Naključne spremenljivke so v grafu prikazane z večjimi vozlišči, (hiper)parametri modela pa z manjšimi (polnimi) vozlišči. Statistične odvisnosti med spremenljivkami so nakazane z usmerjenimi povezavami. Za zgoščen prikaz N spremenljivk $\boldsymbol{x}_n$ in $\boldsymbol{z}_n$ je uporabljen *pladenj* z oznako N.

Razlika med parametri in hiperparametri modela JFA je ta, da parametre obravnavamo kot naključne spremenljivke, medtem ko hiperparametre obravnavamo kot konstantne vrednosti, ki jih določimo po kriteriju ML. Videli smo, da je osnovna naloga, ki jo je potrebno rešiti, izpeljava posteriorne porazdelitve parametrov $\boldsymbol{x}$, $\boldsymbol{y}$ in $\boldsymbol{z}$. Da bi znali poiskati rešitev te naloge, smo predpostavili, da za vsak akustični vektor $\boldsymbol{o}_n \in o$ vemo, katera komponenta modela GMM ga je porodila. Z drugimi

besedami, razpolagamo s popolnim naborom podatkov, kar pomeni, da so spremenljivke s v resnici opažene. To predpostavko smo sprejeli ad hoc in se je nismo posebej trudili utemeljiti.

Spomnimo se, da smo pri izpeljavi postopka EM za model GMM prišli do podobne ugotovitve; če bi razpolagali s popolnim naborom podatkov, bi problem ocenjevanja parametrov modela GMM bil trivialno rešljiv.

V tem razdelku bomo izpeljali postopek za izračun *približka*[50] aposteriorne porazdelitve parametrov modela JFA, pri čemer bomo predpostavili, da imamo na voljo le nepopoln nabor podatkov. Izpeljavo bomo izvedli pod okriljem teorije variacijskih metod, ki jo v tuji literaturi srečamo pod mnogimi imeni, med njimi tudi variational Bayes (VB), učenje zbirk (angl. ensemble learning) in teorija povprečnega polja (angl. mean-field theory), ki jo poznamo iz statistične fizike. Pri izpeljavi bomo uporabili teorijo variacijskega učenja, a se vanjo ne bomo pregloboko poglabljali. Bralec, ki bi ga tema podrobneje zanimala, naj raje poseže po ustrezni literaturi, npr. poglavje 10 v (Bishop, 2007).

C.1 Osnovna ideja variacijskega učenja

Imejmo splošen model. Vse parametre modela in vse prikrite spremenljivke združimo v množico Z. Verjetnostni model je podan z vezano porazdelitvijo $p(x, z)$ in naš cilj je poiskati približek k pravi posteriorni porazdelitvi $p(z|x)$. Logaritem robne porazdelitve lahko razstavimo na vsoto dveh členov:

$$\log p(x) = \mathcal{L}(q) + \mathbb{D}_{\mathrm{KL}}[q \,||\, p],$$

kjer sta:

$$\mathcal{L}(q) = \int q(z) \log \frac{p(x, z)}{q(z)} \mathrm{d}z$$

$$\mathbb{D}_{\mathrm{KL}}[q \,||\, p] = -\int q(z) \log \frac{p(z|x)}{q(z)} \mathrm{d}z.$$

Naša naloga je določiti porazdelitev $q(z)$ tako, da bo spodnja meja $\mathcal{L}(q)$ čim večja. Maksimizacija izraza $\mathcal{L}(q)$ je ekvivalentna minimizaciji KL divergence $\mathbb{D}_{\mathrm{KL}}[q \,||\, p]$, saj vemo, da je divergenca nenegativna funkcija. Če dopuščamo kakršnokoli obliko funkcije $q(z)$, bo maksimum spodnje meje nastopil takrat, ko bo KL divergenca izginila, kar se zgodi le v primeru, ko je $q(z)$ enaka posteriorni porazdelitvi $p(z|x)$.

Kadar je prava posteriorna porazdelitev $p(z|x)$ takšna, da je ne znamo poiskati, jo lahko aproksimiramo s faktorizirano obliko porazdelitve $q(z)$, ki jo lahko zapišemo kot:

[50] Obstajajo tudi metode, ki vsaj v teoriji dajo natančen rezultat. Ker temeljijo na naključnem (Monte Carlo) vzorčenju, so (trenutno, t. j. v času pisanja disertacije) primerne le za ›šolske‹ probleme.

$$q(z) = \prod_{m=1}^{M} q_i(z_i),$$

kjer smo množico Z razdelili na M neprekrivajočih se podmnožic z_i.

S takšno faktorizirano porazdelitvijo smo predpostavili, da so podmnožice z_i pod posteriorno porazdelitvijo statistično neodvisne.

Izkaže se, da lahko splošen izraz za optimalno rešitev $q_j^*(z_j)$, pri kateri bo spodnja meja $\mathcal{L}(q)$ največja, zapišemo v obliki sistema enačb:

$$\log q_j^*(z_j) = \mathbb{E}_{i \neq j}[\log p(x, z)] + \text{const}, \qquad j \in \{1, \ldots, M\}. \tag{C.1}$$

Sistem enačb predstavlja konsistenčne pogoje maksimuma spodnje meje pri predpostavljeni faktorizirani porazdelitvi q. Vendar rešitev ni podana eksplicitno, saj je desna stran enačbe, po kateri se izračuna $q_j^*(Z_j)$, odvisna od ostalih faktorjev $q_i(Z_i)$. To pomeni, da bomo morali rešitev poiskati tako, da bomo najprej določili začetne približke faktorjev, ki jih bomo nato iterativno izboljševali vse dokler ne bo izpolnjen konvergenčni pogoj, podobno kot je to navada pri postopku EM.

C.2 Model JFA kot generativni model

Model JFA lahko vidimo tudi kot generativni model, ki generira vektorje $\boldsymbol{o}$ v skladu z vezano porazdelitvijo $p(\boldsymbol{o}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{s})$, ki se faktorizira na sledeč način:

$$p(o, s, \boldsymbol{m}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = p(o|s, \boldsymbol{m})p(\boldsymbol{m}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})p(s)p(\boldsymbol{x})p(\boldsymbol{y})p(\boldsymbol{z}).$$

Vidimo, da če želimo uporabiti model JFA kot naključni generator, s katerim tvorimo niz vektorjev značilk (govorni posnetek), moramo najprej v skladu z ustreznimi porazdelitvami izbrati vrednosti spremenljivk $\boldsymbol{x}$, $\boldsymbol{y}$ in $\boldsymbol{z}$. Tako posredno izberemo tudi supervektor $\boldsymbol{m} = \boldsymbol{\mu} + \boldsymbol{U}\boldsymbol{x} + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{D}\boldsymbol{z}$, ki določa povprečne vektorje posameznih komponent modela GMM. To storimo le enkrat za celoten posnetek. V naslednjem koraku glede na porazdelitev $p(\boldsymbol{s}_n)$ naključno izberemo eno izmed komponent mešanice, iz katere nato v skladu s porazdelitvijo $\mathcal{N}(\boldsymbol{o}_n|\boldsymbol{m}_k, \boldsymbol{\Sigma}_k)$ generiramo vektor značilk $\boldsymbol{o}_n$. Ta korak lahko ponavljamo poljubno dolgo, pač odvisno od željene dolžine posnetka.

C.3 Variacijsko ocenjevanje posteriornih porazdelitev parametrov in prikritih spremenljivk modela JFA

Najprej zapišimo izraz za vezano porazdelitev vseh naključnih spremenljivk, pri čemer združimo spremenljivke $\boldsymbol{x}$, $\boldsymbol{y}$ in $\boldsymbol{z}$ v novo spremenljivko $\boldsymbol{w}$ na način, kot je že bil podan v razdelku **7.2**:

$$p(o, s, \boldsymbol{m}, \boldsymbol{w}) = p(s)p(o|s, \boldsymbol{m})p(\boldsymbol{m}|\boldsymbol{w})p(\boldsymbol{w}).$$

Ker je spremenljivka $\boldsymbol{m}$ podana analitično: $\boldsymbol{m} = \boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{w}$, je pogojna porazdelitev $p(\boldsymbol{m}|\boldsymbol{w})$ degenerirana (Diracova delta) porazdelitev. V jeziku verjetnosti bi lahko

rekli, da brž ko izvemo vrednost naključne spremenljivke $\boldsymbol{w}$, nedoločenost vrednosti naključne spremenljivke $\boldsymbol{m}$ izgine. V tej luči lahko porazdelitev $\boldsymbol{m}$ izpustimo iz obravnave in zapišemo vezano porazdelitev kot produkt treh faktorjev:

$$p(o, s, \boldsymbol{w}) = p(s)p(o|s, \boldsymbol{w})p(\boldsymbol{w}). \tag{C.2}$$

Posvetimo se posameznim faktorjem. Porazdelitev prikritih spremenljivk s lahko zapišemo kot:

$$p(s) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{s_{nk}},$$

kar smo spoznali že pri izpeljavi postopka EM za model GMM. Porazdelitev $p(o|s, \boldsymbol{w})$ je (zaradi predpostavke iid) enaka produktu Gaussovih porazdelitev:

$$p(o|s, \boldsymbol{w}) = \prod_{n=1}^{N} \mathcal{N}(o_n|\boldsymbol{\mu_k} + \boldsymbol{W}_k\boldsymbol{w}, \boldsymbol{\Sigma}_k),$$

kjer smo z $\boldsymbol{W}_k$ označili podmatriko matrike $\boldsymbol{W}$, ki ustreza k-ti komponenti Gaussove mešanice.

Poiskati želimo posteriorno verjetnost prikritih spremenljivk s in parametra $\boldsymbol{w}$. Predpostavimo, da lahko vezano posteriorno porazdelitev razcepimo na produkt posameznih porazdelitev:

$$q(s, \boldsymbol{w}) = q(s)q(\boldsymbol{w}).$$

Za izpeljavo enačb variacijskega učenja uporabimo splošen rezultat (**C.1**).

Faktor $q^*(s)$

Najprej se posvetimo faktorju $q(s)$. Logaritem optimiziranega faktorja je podan z izrazom

$$\log q^*(s) = \mathbb{E}_{\boldsymbol{w}}[\log p(o, s, \boldsymbol{w})] + \text{const},$$

kjer je const normalizacijska konstanta, ki zagotavlja, da je $q^*(s)$ veljavna porazdelitev.

Ob upoštevanju razcepa (**C.2**) lahko gornji izraz zapišemo kot:

$$\log q^*(s) = \mathbb{E}_{\boldsymbol{w}}[\log p(o|s, \boldsymbol{w})] + \mathbb{E}_{\boldsymbol{w}}[\log p(s)] + \text{const},$$

kjer smo člen $\log p(\boldsymbol{w})$ — ker ni odvisen od spremenljivke s — vsrkali v normalizacijski konstanto. Če vstavimo v gornjo enačbo izraza za porazdelitvi $p(o|s, \boldsymbol{w})$ in $p(s)$, dobimo:

$$\log q^*(s) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \log \rho_{nk} + \text{const},$$

kjer smo ρ_{nk} definirali kot:

$$\begin{aligned}\log \rho_{nk} = \; & \log \pi_k - \frac{1}{2} \log |\Sigma_k| - \frac{D}{2} \log \pi \\ & - \frac{1}{2} \mathbb{E}_{\boldsymbol{w}} \left[(\boldsymbol{x}_n - \boldsymbol{\mu}_k - \boldsymbol{W}_k \boldsymbol{w})^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k - \boldsymbol{W}_k \boldsymbol{w}) \right].\end{aligned}$$

Če obe strani enačbe eksponenciramo in zahtevamo, da ustreza $q^*(s)$ lastnostim porazdelitve, dobimo

$$q^*(s) = \prod_{i=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}},$$

kjer je

$$r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}}.$$

Vidimo, da ima optimalna rešitev za faktor $q(s)$ enako obliko kot prior $p(s)$. Rešitev je odvisna od matematičnega upanja, ki ga izračunamo glede na porazdelitev spremenljivke $\boldsymbol{w}$, kar pomeni, da bodo variacijske enačbe sklopljene in jih bomo morali reševati iterativno.

Z nekaj truda lahko pokažemo, da lahko matematično upanje, ki nastopa v rešitvi za $q^*(s)$, zapišemo kot:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{w}} \left[(\boldsymbol{x}_n - \boldsymbol{\mu}_k - \boldsymbol{W}_k \boldsymbol{w})^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k - \boldsymbol{W}_k \boldsymbol{w}) \right] = & \\ (\boldsymbol{x}_n - \boldsymbol{\mu}_k - \boldsymbol{W}_k \boldsymbol{\mu}_{\boldsymbol{w}})^{\mathrm{T}} \, \boldsymbol{\Sigma}_k^{-1} \, (\boldsymbol{x}_n - \boldsymbol{\mu}_k - \boldsymbol{W}_k \boldsymbol{\mu}_{\boldsymbol{w}}) & \\ + \operatorname{Tr} (\boldsymbol{W}_k^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{W}_k \boldsymbol{\Sigma}_{\boldsymbol{w}}), &\end{aligned}$$

kjer smo z $\boldsymbol{\mu}_{\boldsymbol{w}}$ in $\boldsymbol{\Sigma}_{\boldsymbol{w}}$ označili trenutno povprečno vrednost in kovariančno matriko porazdelitve $q^*(\boldsymbol{w})$.

Z upoštevanjem vsega naštetega, lahko izraz za r_{nk}, ki nastopa v rešitvi za faktor $q^*(s)$, zapišemo z naslednjim izrazom:

$$\begin{aligned}r_{nk} \propto \pi_k |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} \exp \Bigg(& -\frac{1}{2} (\boldsymbol{x}_n - \boldsymbol{\mu}_k - \boldsymbol{W}_k \boldsymbol{\mu}_{\boldsymbol{w}})^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k - \boldsymbol{W}_k \boldsymbol{\mu}_{\boldsymbol{w}}) \cdots \\ & \cdots - \frac{1}{2} \operatorname{Tr} (\boldsymbol{W}_k^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{W}_k \boldsymbol{\Sigma}_{\boldsymbol{w}}) \Bigg)\end{aligned}$$

ki je podoben rešitvi, kot jo dobimo pri izpeljavi postopka EM za model GMM, le da imamo sedaj prisotno še sled $\operatorname{Tr} (\boldsymbol{W}_k^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{W}_k \boldsymbol{\Sigma}_{\boldsymbol{w}})$, ki izraža ›nezaupanje‹ v oceno spremenljivke $\boldsymbol{w}$.

Faktor $q^*(\boldsymbol{w})$

Poiskali smo enačbo za izračun optimalne porazdelitve $q^*(s)$, sedaj pa storimo enako še za faktor $q^*(\boldsymbol{w})$.

Z upoštevanjem statističnih neodvisnosti med spremenljivkami modela JFA in z uporabo splošnega rezultata (C.1) lahko zapišemo:

$$\log q^*(\boldsymbol{w}) = \mathbb{E}_s[\log p(o|s, \boldsymbol{w}) + \log p(\boldsymbol{w})] + \text{const},$$

kjer smo člen $p(s)$, ker ni odvisen od spremenljivke $\boldsymbol{w}$, vsrkali v normalizacijsko konstanto. Če v gornjo enačbo vstavimo izraza za $p(o|s, \boldsymbol{w})$ in $p(\boldsymbol{w})$ ter vse od $\boldsymbol{w}$ neodvisne člene vsrkamo v normalizacijsko konstanto, dobimo:

$$\begin{aligned}\log q^*(\boldsymbol{w}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}[s_{nk}] \left(-\frac{1}{2}(\boldsymbol{o}_n - \boldsymbol{\mu}_k - \boldsymbol{W}_k\boldsymbol{w})^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{o}_n - \boldsymbol{\mu}_k - \boldsymbol{W}_k\boldsymbol{w})\right) \\ - \frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu}_{\boldsymbol{w}})^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1}(\boldsymbol{w} - \boldsymbol{\mu}_{\boldsymbol{w}}) + \text{const}\end{aligned}$$

Za matematično upanje $\mathbb{E}[s_{nk}]$ kategorične naključne spremenljivke poznamo standarden rezultat: $\mathbb{E}[s_{nk}] = r_{nk}$. Vrednost r_{nk} lahko interpretiramo kot odgovornost k-te komponente za generiranje observacije $\boldsymbol{o}_n$. Če uvedemo statistiko ničtega (N_k), prvega ($\boldsymbol{F}_k$) in drugega ($\boldsymbol{S}_k$) reda:

$$\begin{aligned}N_k &= \sum_{n=1}^{N} r_{nk}, \\ \boldsymbol{F}_k &= \sum_{n=1}^{N} r_{nk}(\boldsymbol{o}_n - \boldsymbol{\mu}_k), \\ \boldsymbol{S}_k &= \sum_{n=1}^{N} r_{nk}(\boldsymbol{o}_n - \boldsymbol{\mu}_k)(\boldsymbol{o}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\end{aligned}$$

in jih zložimo v matrike $\boldsymbol{N}$, $\boldsymbol{F}$ in $\boldsymbol{S}$ (glej stran 45), se da z nekaj truda dvojno vsoto, ki nastopa v izrazu za $q^*(\boldsymbol{w})$, zapisati v matrični obliki:

$$\operatorname{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}) - 2(\boldsymbol{W}\boldsymbol{w})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{F} + (\boldsymbol{W}\boldsymbol{w})^{\mathrm{T}}\boldsymbol{N}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}\boldsymbol{w}.$$

Če v izrazu za $\log q^*(\boldsymbol{w})$ obdržimo le člene, ki so odvisni od spremenljivke $\boldsymbol{w}$, vse ostale pa vsrkamo v normalizacijsko konstanto, lahko ob upoštevanju gornjega rezultata $q^*(\boldsymbol{w})$ zapišemo kot:

$$\begin{aligned}\log q^*(\boldsymbol{w}) = &-\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} + \boldsymbol{W}^{\mathrm{T}}\boldsymbol{N}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}\right)\boldsymbol{w} \\ &+ \boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1}\boldsymbol{\mu}_{\boldsymbol{w}} + \boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{F}\right) + \text{const}.\end{aligned}$$

Z metodo natančnega pogleda lahko v gornjem izrazu opazimo Gaussovo porazdelitev. Z uporabo metode *dopolnitve kvadrata* (angl. completing the square) moremo razbrati pripadajočo povprečno vrednost in kovariančno matriko:

$$\begin{aligned}\mathbb{E}[\boldsymbol{w}] &= \boldsymbol{L}^{-1}\left(\boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1}\boldsymbol{\mu}_{\boldsymbol{w}} + \boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{F}\right)\\ \mathrm{Var}\,[\boldsymbol{w}] &= \boldsymbol{L}^{-1},\end{aligned}$$

kjer je precizijska matrika $\boldsymbol{L}$ podana z izrazom:

$$\boldsymbol{L} = \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} + \boldsymbol{W}^{\mathrm{T}}\boldsymbol{N}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}.$$

S tem je naša variacijska izpeljava enačb za iterativno ocenjevanje posteriorne porazdelitve naključne spremenljivke $\boldsymbol{w}$ končana. Dobljena izraza za povprečno vrednost in kovariančno matriko sta zelo podobna tistima, ki smo ju dobili v razdelku **7.4.2** na str. **63**, ko smo predpostavljali, da imamo na voljo poln nabor podatkov. Identična sta le v prvem koraku iteracije, če za začetni vrednosti $\boldsymbol{\mu}_{\boldsymbol{w}}$ in $\boldsymbol{\Sigma}_{\boldsymbol{w}}$ izberemo ničelni vektor in enotsko matriko.

Strnimo izpeljane rezultate še v obliki iteracijskega postopka:

1. Izberi začetne vrednosti $\boldsymbol{\mu}_{\boldsymbol{w}} = \mathbf{0}$ in $\boldsymbol{\Sigma}_{\boldsymbol{w}} = \boldsymbol{I}$.

2. Izračunaj odgovornosti r_{nk} posameznih komponent pri trenutni porazdelitvi $q^*(\boldsymbol{w})$. Izračunaj tudi zadostno statistiko N_k, $\boldsymbol{F}_k$ in $\boldsymbol{S}_k$.

3. Izračunaj novi vrednosti povprečnega vektorja in kovariančne matrike.

4. Preveri konvergenčni pogoj (npr. ocene med zaporednima iteracijama so se le malenkost spremenile). Če je konvergenčni pogoj izpolnjen, končaj postopek, sicer se vrni na korak 2.

Opozorimo še, da postane v splošnem v naslednji ponovitvi postopka kovariančna matrika $\boldsymbol{\Sigma}_{\boldsymbol{w}}$ polna, tudi če je bila pred tem diagonalna. To predstavlja dodatno težavo, saj je v praksi razsežnost te matrike glede na zmogljivosti trenutnih računalnikov pogosto prevelika, zato smo prisiljeni poiskati kompromisno rešitev. Ena izmed možnosti, ki so nam na voljo, je, da poiščemo približek porazdelitve naključne spremenljivke $\boldsymbol{w}$ tako, da predpostavimo neodvisnost med spremenljivko $\boldsymbol{z}$ in spremenljivkama $\boldsymbol{x}$ in $\boldsymbol{y}$.